Unifying Perception and Creation with Generative Models

Zhipeng Bao

CMU-RI-TR-25-99

November 3, 2025

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Martial Hebert, *chair*
Deva Ramanan
Jun-Yan Zhu
Alexei Efros, *University of California, Berkeley*
Yu-Xiong Wang, *University of Illinois Urbana-Champaign*
Pavel Tokmakov, *Toyota Research Institute*

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Robotics.

To my wife and my family.

Abstract

The field of generative models has made remarkable progress in synthesizing photorealistic visual content, encompassing images, videos, and even text. However, the potential of these powerful generative models, such as diffusion-based and auto-regressive transformers, for visual perception and understanding remains underexplored. This thesis investigates how generative models can serve as powerful visual learners – bridging the long-standing divide between generative and discriminative paradigms.

In the first work, we introduce a unified, versatile, diffusion-based framework, Diff-2-in-1, that can simultaneously handle both multi-modal data generation and dense visual perception, through a unique exploitation of the *diffusion-denoising process*. Within this framework, we further enhance discriminative visual perception via multi-modal generation by utilizing the denoising network to create multi-modal data that mirrors the distribution of the original training set. Importantly, Diff-2-in-1 optimizes the utilization of the created diverse and faithful data by leveraging a novel self-improving learning mechanism.

In the second work, we extend such a diffusion-based framework from image understanding to video understanding. Specifically, we explore the potential of diffusion models for video understanding by analyzing the feature representations learned by both image- and video-based diffusion models, alongside non-generative, self-supervised approaches. Our findings reveal that video diffusion models consistently rank among the top performers, particularly excelling at modeling temporal dynamics and scene structure. This observation not only sets them apart from image-based diffusion models but also opens a new direction for advancing video understanding, offering a fresh alternative to traditional discriminative pre-training objectives.

In the third work, beyond merely leveraging video diffusion models as feature extractors, we present REM, a framework for segmenting a wide range of concepts in video that can be described through natural language, by repurposing text-to-video diffusion models for referral video segmentation. A key insight of our approach is preserving as much of the generative model's original representation as possible, while fine-tuning it on narrow-domain Referral Object Segmentation datasets. As a result, our framework can accurately segment and track rare and unseen objects, despite being trained on object masks from a limited set of categories.

Our experiments show that REM performs on par with state-of-the-art approaches on in-domain datasets, like Ref-DAVIS, while outperforming them by up to 12 points in terms of region similarity on out-of-domain data, leveraging the power of Internet-scale pre-training.

Finally, building on this foundation, we introduce a unified perceptual–generative framework that repurposes a generative model, comprising both diffusion architecture and visual auto-regressive (VAR) architecture, across a broad spectrum of perception, restoration, and editing tasks. For diffusion-based models, we show that diffusion-based models deliver superior or comparable performance to discriminative counterparts, revealing their intrinsic ability to encode rich, multi-modal world representations. For the VAR variant, we present the first unified architecture capable of efficiently solving 15 tasks within a single framework. We show that latent-variable designs, particularly those leveraging variational autoencoders, are key to achieving coherent multi-modal understanding and consistent generation. Compared with diffusion counterparts, VAR-based models offer substantial gains in latency, scalability, and output consistency.

Collectively, these studies offer a cohesive perspective on unifying perception and synthesis through generative modeling, charting a path toward general-purpose visual foundation models that seamlessly integrate understanding, reasoning, and creation.

Acknowledgments

Looking back on my PhD journey at RI, I am overwhelmed with gratitude.

My deepest, most profound gratitude goes to my advisor, Martial Hebert. I feel incredibly fortunate to have had him as my advisor. Despite his illustrious position as the Dean, Martial's door was always open, and he always found the time to meet with me nearly every single week. He taught me the very essence of how to think like a researcher. I will always carry with me his guiding principle: "Make things as simple as possible." He taught me to cut through the complexity and focus on the fundamental "IO flow" of any problem. He also fundamentally changed how I communicate my work: a great talk isn't about showing every single project, but about weaving them into a single, compelling story – a lesson that was foundational for this very thesis.

I am immensely indebted to mentors – Yu-Xiong Wang and Pavel Tokmakov, with whom I worked so closely. Yu-Xiong taught me the craft of research, from brainstorming to the rigorous process of writing. I can still picture him guiding me, almost word-by-word, through the first draft of my first paper. His strong, level-headed advice at every major decision point in my academic life has consistently proven to be correct. I also want to thank Pavel. I must admit, even after reading countless texts generated by today's advanced LLMs, I still believe the papers he revised are the most beautiful ones I have ever read. He embodies a perfect balance: chill in life (I still and will always remember every TRI party he has organized), yet remarkably meticulous at work, catching every small detail while elevating the high-level vision. They have been more than mentors; they have been my cherished friends.

My sincere thanks also go to my other committee members, Jun-Yan Zhu, Deva Ramanan, and Alyosha Efros, for their sharp insights and for reviewing this thesis and discussing my research. I was always in awe during my meetings with Jun-Yan and Deva; they could effortlessly grasp the most complex technical details, always offering new perspectives that pushed my work forward. I also extend my deep respect and gratitude to Alyosha Efros for his time and perspective on my research.

A PhD can be a long road, and I would not have survived it without my friends. To my friends at CMU in those first few years, especially Zifan

Wang, Yi Hu, Yuting Bu, Yuhang Yao, Yuding Wang, Weihong Wang, and Yangfan Ran, thank you for your companionship. Your presence was a lifeline that saved me from despair, especially during the deep isolation of the COVID pandemic. To my friends in my final years, particularly Jiashun Wang, Tianyun Zou, Oliver Wang, Stella Chen, Rui Chen, Zhixuan Liu, Yaqi Xie, Yujia Zheng, Bowen Li, Qinglin feng, Zhiqiu Lin, Tiffany Ling, Zihan Wang, Qitao Zhao, Siyuan Ceng, you brought so much light and joy back into my life; every meal we shared, every party, is an unforgettable memory. To my research colleagues and friends at CMU, including my Anurag Bagchi, Anurag Ghosh, Sriram Narayanan, Nupur Kumari, Mihir Prabhudesai, Kangle Deng, Khiem Vuong, thank you for the fun and insightful discussions during meetings and cafe breaks. My gratitude also extends to my friends beyond CMU, particularly a lot of friends at UIUC, Zhen Zhu, my previous roommate, future roommate, hommie, and collaborator, Nan Liu, Ziqi Pang, Yunze Man, Shengcao Cao. It's truly amazing how what started as research discussions blossomed into fruitful collaborations and, more importantly, deep and lasting friendships.

I am also incredibly thankful for our administrative staff: Nichole Merritt and Lynnetta Miller, who make the entire journey much smoother.

To my dearest family, thank you for your unconditional love and unwavering support. Across this long journey and from an ocean away, your constant belief gave me the strength to persevere through every single challenge. I know my family will always be my greatest source of support.

Finally, my heart belongs to my wonderful wife, Yifei Yang. Our story has been the most cherished part of my PhD. We met, built a life together, and celebrated our marriage – all while navigating the intensity of this final chapter. Yifei has taught me to always prepare for the worst and hope for the best. Because of her, I have learned to think further ahead – to see beyond the immediate and consider the long term. This mindset not only helped me endure the challenges of my PhD but also made me a better person. Whenever I struggled, whether in research or simply in life, she was always there, accompanying me through the difficult moments, offering comfort, strength, and understanding. Your unwavering love, patience, and faith in me have been my anchor and my harbor. This achievement is as much yours as it is mine, and I am endlessly grateful to walk through life by your side. And, of course, thanks to our lovely pets, Lucky the dog and Cola the cat, for their simple, joyful company and for always knowing when it was time to step away from the terminal.

Funding

This work was supported by Toyota Research Institute.

Contents

When this dissertation is viewed as a PDF, the page header is a link to this Table of Contents.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The field of generative modeling has witnessed transformative progress in recent years, leading to models capable of synthesizing highly realistic and diverse visual content across modalities. Among these, diffusion models [79, 185, 203, 279] have emerged as a dominant family for image and video generation, setting new standards in fidelity and controllability. In parallel, large-scale auto-regressive (AR) transformers have demonstrated exceptional performance in sequential prediction tasks, including text generation, visual synthesis, and multimodal reasoning. Together, these advances have redefined what it means for a model to "understand" and "create" visual data.

While the generative capabilities of such models are well established, their potential for *visual perception and understanding* – tasks traditionally framed as discriminative problems – remains comparatively underexplored. Diffusion models, in particular, possess an inherently structured *denoising process* that gradually refines noisy signals into semantically coherent images. This process implicitly encodes rich spatial and semantic information across multiple noise scales, suggesting that diffusion models are not only powerful generators but also expressive representation learners. Recent works have begun to uncover this property, leveraging diffusion features for dense perception tasks such as semantic segmentation [14, 247, 254] and depth estimation [192, 284]. However, most approaches treat diffusion models as *standalone* modules – either using them for synthetic data augmentation [26] or as frozen backbones for feature extraction [93, 191, 254, 284]. Such isolated use disregards the unique iterative refinement mechanism of diffusion models, limiting their representational and perceptual potential.

This thesis aims to bridge this gap by exploring how generative models, both diffusion-based and auto-regressive, can serve as *powerful visual learners*. Specifically, it investigates how the generative process itself can be repurposed to advance visual understanding, enabling models to perceive the world as effectively as they can synthesize it. Across four major studies, the thesis establishes a unified perspective that connects generation and perception through shared representations, objectives, and architectures.

In the first study, we introduce **Diff-2-in-1**, a unified diffusion-based framework that simultaneously performs multi-modal generation and dense visual perception. By explicitly exploiting the diffusion-denoising process, this framework allows the denoising network to synthesize diverse, faithful training samples while improving its own discriminative capability through a self-improving learning mechanism.

In the second study, we extend this perspective from images to videos, systematically evaluating both image- and video-based diffusion models against non-generative, self-supervised counterparts. Through a unified probing framework covering four key video understanding tasks: action recognition, object discovery, scene understanding, and label propagation. We reveal that video diffusion models excel at capturing temporal dynamics and scene structure, achieving top-tier performance across diverse domains.

In the third study, we move beyond perception to **ReferEverything (REM)**, a framework that repurposes text-to-video diffusion models for referring video segmentation. REM preserves the full generative structure of the model while adapting it to localize and segment objects and processes described by natural language. This approach demonstrates strong generalization to unseen concepts and dynamic, non-object entities, such as natural phenomena, through Internet-scale pretraining and minimal task-specific tuning.

Finally, building upon these insights, we propose a unified perceptual–generative paradigm that integrates both diffusion and visual auto-regressive (VAR) architectures under a common framework. We show that diffusion-based models inherently encode structured, multimodal representations beneficial for understanding and restoration, while VAR-based architectures offer superior efficiency and scalability. Notably, our unified VAR model achieves strong performance across fifteen diverse perception, restoration, and editing tasks within a single system, highlighting the potential of generative architectures as general-purpose visual learners.

Collectively, these works advance a central thesis: generative models can serve as the foundation for unifying perception and creation. By reinterpreting the denoising and autoregressive processes as forms of structured visual reasoning, this dissertation charts a path toward general-purpose visual foundation models that seamlessly integrate understanding, reasoning, and generation.

1.2 Thesis Overview

In Chapter 2, we show that beyond high-fidelity image synthesis, diffusion models have recently exhibited promising results in dense visual perception tasks. However, most existing work treats diffusion models as a standalone component for perception tasks, employing them either solely for off-the-shelf data augmentation or as mere feature extractors. In contrast to these isolated and thus sub-optimal efforts, we introduce a unified, versatile, diffusion-based framework, Diff-2-in-1, that can simultaneously handle both multi-modal data generation and dense visual perception, through a unique exploitation of the *diffusion-denoising process*. Within this framework, we further enhance discriminative visual perception via multi-modal generation by utilizing the denoising network to create multi-modal data that mmirrorss the distribution of the original training set. Importantly, Diff-2-in-1 optimizes the utilization of the created diverse and faithful data by leveraging a novel self-improving learning mechanism. Comprehensive experimental evaluations validate the effectiveness of our framework, showcasing consistent performance improvements across various discriminative backbones and high-quality multi-modal data generation characterized by both realism and usefulness.

In Chapter 3, we explore the potential of diffusion models for video understanding by analyzing the feature representations learned by both image- and video-based diffusion models, alongside non-generative, self-supervised approaches. We propose a unified probing framework to evaluate seven models across four core video understanding tasks: action recognition, object discovery, scene understanding, and label propagation. Our findings reveal that video diffusion models consistently rank among the top performers, particularly excelling at modeling temporal dynamics and scene structure. This observation not only sets them apart from image-based diffusion models but also opens a new direction for advancing video understanding, offering a fresh alternative to traditional discriminative pre-training objectives. Interestingly, we demonstrate that higher-generation performance does not always correlate with improved performance in downstream tasks, highlighting the importance of careful representation selection. Overall, our results suggest that video diffusion models hold substantial promise for video understanding by effectively capturing both spatial and temporal information, positioning them as strong competitors in this evolving domain.

In Chapter 4, we present REM, a framework for segmenting a wide range of concepts in video that can be described through natural language. Our method capitalizes on visual-language representations learned by video diffusion models on Internet-scale datasets. A key insight of our approach is preserving as much of the generative model's original representation as possible, while fine-tuning it on narrow-domain Referral Object Segmentation datasets. As a result, our framework can accurately segment and track rare and unseen objects, despite being trained on object masks from a limited set of categories. Additionally, it can generalize to non-object dynamic concepts, such as waves crashing in the ocean, as demonstrated in our newly introduced benchmark for Referral Video Process Segmentation (Ref-VPS). Our experiments show that REM performs on par with state-of-the-art approaches on in-domain datasets, like Ref-DAVIS, while outperforming them by up to 12 points in terms of region similarity on out-of-domain data, leveraging the power of Internet-scale pre-training.

In Chapter 5, we briefly discuss how to repurpose a diffusion model from a single task performer to a generalist expert for a wide range of tasks, without additional finetuning. This chapter serves as a transition and motivation between Chapter 4 and Chapter 6, revealing the success of the design of multimodal large language models bundled with a visual decoder for unified visual generation. We also demonstrate the limitation of such a design – the latency and objective mismatch issue, which motivates us to find an alternative design choice – a pure autoressive system for unified visual generation tasks.

In Chapter 6, we present UniGen-AR, a framework that pairs a general-purpose multi-modal language model (MLLM) with an efficient visual auto-regressive (VAR) decoder. This design retains the flexibility of MLLM-based conditioning while leveraging the sampling efficiency and latent unification properties of VAR models. In our framework, the MLLM encodes free-form instructions and control signals into a unified sequence, which guides the VAR decoder to generate image-valued outputs for 12 tasks spanning three families. Empirically, UniGen-AR achieves up to 5× lower inference latency than diffusion-based baselines while maintaining or improving output quality. Our ablations further reveal that VQ-VAE tokenizer design, particularly codebook size and hierarchy, is a critical factor for VAR scalability. These results establish visual auto-regressive modeling as a compelling and efficient backbone for unified visual generation.

Finally, in Chapter 7, we discuss the proposed research projects for the thesis research and show the detailed plans for them.

1.3 Thesis contributions

In this thesis, we introduce four research projects that aim to leverage generative models – either use their representations or directly repurpose them – for downstream tasks. We cover both diffusion models and autoregressive models, and we summarize the primary contributions of the papers as follows;

Diff-2-in-1: Bridging Generation and Dense Perception with Diffusion Models

- We propose Diff-2-in-1, an integrated framework that seamlessly performs multi-modal generation and dense visual perception based on diffusion models;
- We introduce a novel self-improving mechanism that progressively enhances multi-modal generation in a self-directed manner, thereby effectively boosting the discriminative performance via generative learning;
- Our method demonstrates consistent performance improvements across various discriminative backbones and high-quality multi-modal generation under both realism and usefulness.

Video Diffusion Models Learn the *Structure* of the *Dynamic* World

Through comprehensive evaluation, we bring the following insights for video diffusion models:

- Video diffusion models excel at capturing motion dynamics while maintaining a high-level understanding of the structure of the visual world, which supports their consistently strong performance;
- These models encode different information at various layers: early layers focus on abstract, high-level features, while later layers capture finer details;
- Progress in video generation does not always correlate with performance on downstream tasks. However, proxy tasks can help systematically identify the best-performing checkpoint.

ReferEverything: Towards Segmenting Everything We Can Speak of in Videos

- We demonstrate that Web-scale video diffusion models have learned universal visual-language mapping that can be repurposed for open-world referring video segmentation.
- We further introduce a new benchmark for Referring Video Process Segmentation called Ref-VPS, expanding the focus of RVS beyond conventional object tracking.
- Finally, we provide a detailed analysis of our approach, demonstrating that retaining the full architecture of the generative model, rather than isolating the de-noising network as a feature extractor, is key to unlocking the strongest generalization in RVS.

UniGen-AR: Unifying Visual Generation with Auto-Regressive Modeling

- We present, to our knowledge, the first framework that scales visual auto-regressive modeling to the full UVG setting, unifying open-ended synthesis, restoration, and visual perception within a single image-out backbone.
- We demonstrate a compelling latency–quality trade-off compared to diffusion-based systems, achieving consistent speedups while maintaining or improving performance.
- We identify and validate the importance of VQ-VAE tokenizer design, including codebook size and hierarchy, as a key driver of VAR scalability and effectiveness.
- We study the bidirectional connection between multimodal understanding and

generation, showcasing how understanding-enhanced MLLM front-ends improve control and quality in image synthesis.

1.4 Additional papers

Here is a list of papers I contributed to but omitted from this thesis to maintain a clear focus and consistent narrative.

- **Lexicon3d: Probing visual foundation models for complex 3d scene understanding**. Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. NeurIPS 2024.
- **Separate-and-enhance: Compositional finetuning for text-to-image diffusion models**. Zhipeng Bao, Yijun Li, Krishna Kumar Singh, Yu-Xiong Wang, and Martial Hebert. SIGGRAPH 2024.
- **Multi-task view synthesis with neural radiance fields**. Shuhong Zheng*, Zhipeng Bao*, Martial Hebert, and Yu-Xiong Wang. ICCV 2023.
- **Object discovery from motion-guided tokens**, Zhipeng Bao, Pavel Tokmakov, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. CVPR 2023.
- **Beyond rgb: Scene-property synthesis with neural radiance fields**. Mingtong Zhang*, Shuhong Zheng*, Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. WACV 2023.
- **Generative Modeling for Multi-task Visual Learning**. Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. ICML 2022.
- **Discovering Objects that Can Move**. Zhipeng Bao*, Pavel Tokmakov*, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. CVPR 2022.
- **Bowtie networks: Generative modeling for joint few-shot recognition and novel-view synthesis**. Zhipeng Bao, Yu-Xiong Wang, and Martial Hebert. ICLR 2021.

Chapter 2

Diff-2-in-1: Bridging Generation and Dense Perception with Diffusion Models

2.1 Introduction

Diffusion models have emerged as powerful generative modeling tools for various high-fidelity image synthesis tasks [79, 185, 203, 279]. Beyond their primary synthesis capabilities, diffusion models are increasingly recognized for their expressive representation abilities. This has spurred interest in leveraging them for dense pixel-level visual perception tasks, such as semantic segmentation [14, 247, 254] and depth estimation [192, 284]. Nonetheless, most existing approaches treat diffusion models as a *standalone* component for perception tasks, either employing them for off-the-shelf data augmentation [26], or utilizing the diffusion network as feature extraction backbone [93, 191, 254, 284]. These efforts overlook the *unique* diffusion-denoising process inherent in diffusion models, thus limiting their potential for discriminative dense visual perception tasks.

Inspired by foundational studies that explore the interplay between generative and discriminative learning [153, 178, 188, 222], we argue that the diffusion-denoising process plays a critical role in unleashing the capability of diffusion models for the discriminative visual perception tasks. The diffusion process corrupts the visual input with noise, enabling the *generation* of abundant new data with diversity. Subsequently, the denoising process removes the noise from noisy images to create high-fidelity data, thus obtaining informative features for *discriminative* tasks at the same time. As a result, the diffusion-denoising process naturally connects the generative process with discriminative learning.

Interestingly, this synergy further motivates us to propose a novel *unified* diffusion modeling framework that integrates both discriminative and generative learning within a single, coherent paradigm. From the generative perspective, we focus on synthesizing photo-realistic *multi-modal* paired data (*i.e.*, RGB images and their associated pixel-level visual attributes) that accurately capture various types of visual information. Simultaneously, the unified diffusion model can achieve promising results in different visual prediction tasks from the discriminative standpoint. As an example illustrated in Figure 2.1, when considering RGB and depth interactions, if the model receives an RGB image as input, its function is to predict an accurate depth map. Meanwhile, the model is equipped to produce photo-realistic and coherent RGB-depth pairs sampled from noise. Despite its conceptual simplicity, fully operationalizing the unified framework – acquiring enhanced performance for both multi-modal generation and dense perception such as by effectively leveraging generated samples for discriminative tasks – presents non-trivial challenges. In particular, the generation process inevitably produces data of relatively inferior quality compared to real data. Additionally, generated samples may exhibit considerable data distribution gaps from the target domain.

To address these challenges, we introduce Diff-2-in-1, a diffusion framework bridging multi-modal generation and discriminative dense visual perception within one unified diffusion model. The core design within our Diff-2-in-1 is a self-improving learning mechanism, featuring two sets of parameters for our unified diffusion model *during the training process*. Specifically, the *creation parameters* are tailored to generate additional multi-modal data for discriminative learning, while the *exploitation parameters* are employed for utilizing both the original and synthetic data to learn the discriminative dense visual perception task. Meanwhile, the creation parameters continuously undergo *self-improvement* based on the weights of the exploitation parameters via exponential moving average (EMA). With our novel design of two sets

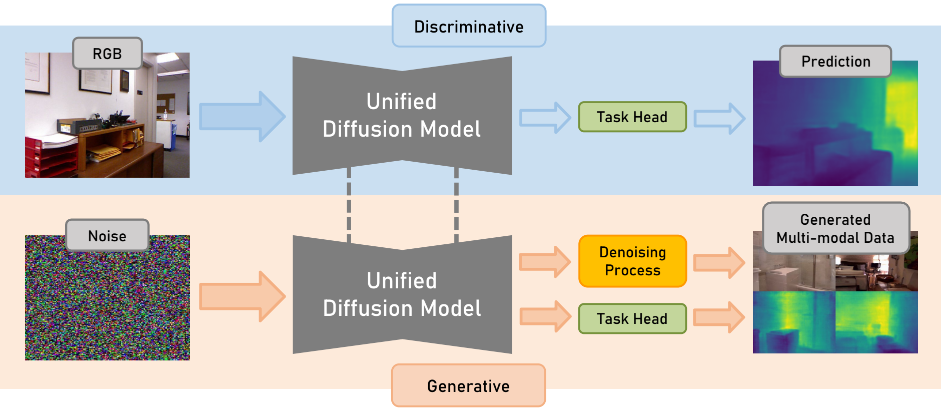


Figure 2.1: **A single, unified diffusion-based model for both generative and discriminative learning**. If the model receives an RGB image as input, its function is to predict an accurate visual attribute map. Simultaneously, the model is equipped to produce photo-realistic and coherent multi-modal data sampled from Gaussian noise. We use depth as an example here for illustration, and the framework is also applicable to other visual attributes such as segmentation, surface normal, *etc.*

of parameters interplaying with each other, the discriminative learning process can benefit from the synthetic samples generated by the model itself, while the quality of the generated data is iteratively refined at the same time.

We validate the effectiveness of Diff-2-in-1 through extensive and multi-faceted experimental evaluations. We start with the evaluation of the discriminative perspective, demonstrating its superiority over state-of-the-art discriminative baselines across various tasks in both single-task and multi-task settings. We additionally show that Diff-2-in-1 is generally applicable to different backbones and consistently boosts performance. Next, we ablate the experimental settings such as different training data sizes, to gain a comprehensive understanding of our method. Finally, we demonstrate the realism and usefulness of the multi-modal data generated by our Diff-2-in-1.

Our contributions include: **(1)** We propose Diff-2-in-1, a unified framework that seamlessly integrates multi-modal generation and discriminative dense visual perception based on diffusion models. **(2)** We introduce a novel self-improving mechanism that progressively enhances multi-modal generation in a self-directed manner, thereby effectively boosting the discriminative visual perception performance via generative learning. **(3)** Our method demonstrates consistent performance improvements across various discriminative backbones and high-quality multi-modal data generation under both realism and usefulness.

2.2 Related Work

Pixel-level dense visual perception covers a broad range of discriminative computer vision tasks including depth estimation [49, 60, 64, 119], segmentation [129, 252], surface normal prediction [4, 233], keypoint detection [28, 29, 220, 291], etc. After the convolutional neural network (CNN) [104, 199, 210, 273] shows great success in ImageNet classification [45] even outperforming humans [71], adopting CNN for dense prediction tasks [48, 49, 233] becomes a prototype for model design. With Vision Transformer (ViT) [47] later becoming a revolutionary advance in architecture for vision models, an increasing number of visual perception models [180, 181, 252] start to adopt ViT as their backbones, benefiting from the scalability and global perception capability brought by ViT.

Generative modeling for discriminative tasks. The primary objective of generative models has traditionally been synthesizing photo-realistic images. However, recent advancements have expanded their utility to the generation of "useful" images for downstream visual tasks [1, 11, 167, 275, 280, 286, 294, 295]. This is typically accomplished by generating images and corresponding annotations off-the-shelf, subsequently using them for data augmentation in specific visual tasks.

Nowadays, with the emergence of powerful diffusion models in high-fidelity synthesis tasks [31, 79, 185, 203, 231, 279], there has been a growing interest in applying them to discriminative tasks. Among them, ODISE [254] and VPD [284] extract features using the stable diffusion model [185] to perform discriminative tasks such as segmentation and depth estimation. DIFT [212] and its concurrent work [74, 135, 277] utilize diffusion features for identifying semantic correspondence. DDVM [191] solves depth and optical flow estimation tasks by denoising from Gaussian noise with RGB images as a condition. Diffusion Classifier [111] utilizes diffusion models to enhance the confidence of zero-shot image classification. Other studies [26, 51, 221] have explored using diffusion models to augment training data for image classification. Different from them, we propose a *unified* diffusion-based model that can directly work for discriminative dense visual perception tasks, and simultaneously utilize its generative process to facilitate discriminative learning through the proposed novel self-improving algorithm.

2.3 Unified Diffusion Model: Diff-2-in-1

2.3.1 Preliminary: Latent Diffusion Models

Diffusion models [79] are latent variable models that learn the data distribution with the inverse of a Markov noising process. Instead of leveraging the diffusion models in the RGB color space [79, 203], we build our method upon the state-of-the-art latent diffusion model (LDM) [185]. First, an encoder $\mathcal{E}$ is trained to map an input image $x \in \mathcal{X}$ into a spatial latent code $z = \mathcal{E}(x)$. A decoder $\mathcal{D}$ is then tasked with reconstructing the input image such that $\mathcal{D}(\mathcal{E}(x)) \approx x$.

Considering the clean latent $z_0 \sim q(z_0)$, where $q(z_0)$ is the posterior distribution of z_0, LDM gradually adds Gaussian noise to z_0 in the *diffusion process*:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I}), \tag{2.1}$$

where β_t is a variance schedule that controls the strength of the noise added in each timestep. We can derive a closed-form process from Equation 2.1 to convert a clean latent z_0 to a noisy latent z_T of arbitrary timestep T:

$$z_T \sim q(z_T|z_0) = \mathcal{N}(z_T; \sqrt{\bar{\alpha}_T} z_0, (1-\bar{\alpha}_T)\mathbf{I}), \tag{2.2}$$

where the notation $\alpha_T = 1 - \beta_T$ and $\bar{\alpha}_T = \prod_{s=1}^{T} \alpha_s$ makes the formulation concise. When $T \to \infty$, z_T is nearly equivalent to sampling from an isotropic Gaussian distribution.

The denoising process takes inverse operations from the diffusion process. We estimate the denoised latent at timestep $t-1$ from t by:

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \mathbf{\Sigma}_\theta(z_t, t)), \tag{2.3}$$

where the parameters $\mu_\theta(z_t, t), \mathbf{\Sigma}_\theta(z_t, t)$ of the Gaussian distribution are estimated from the model.

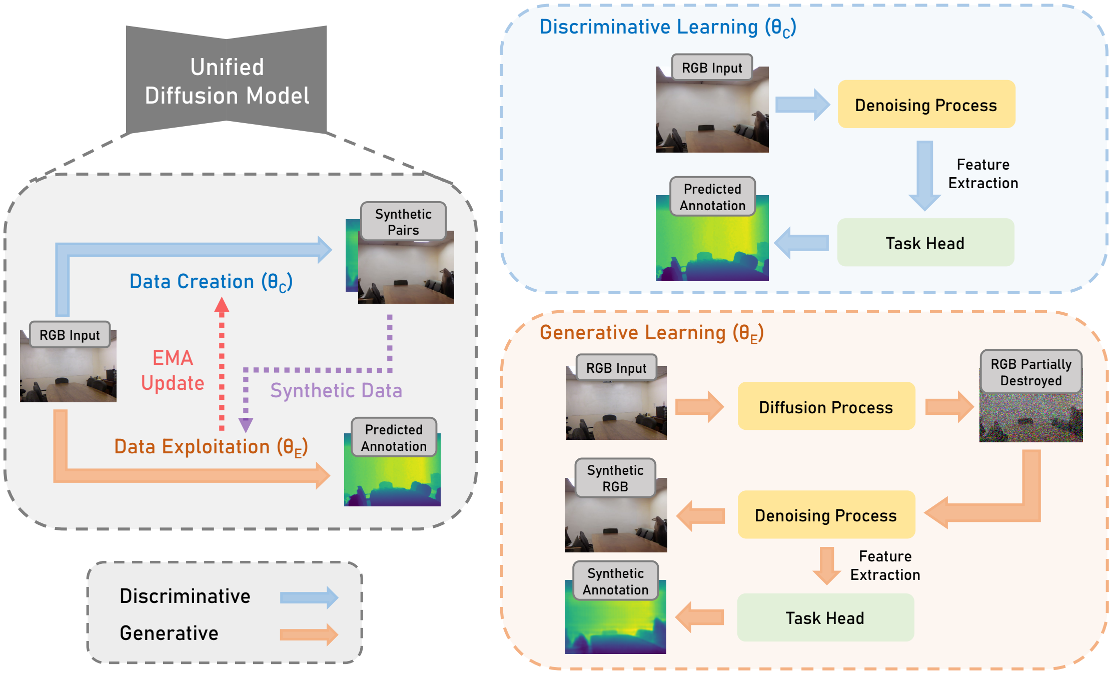


Figure 2.2: **Our self-improving learning paradigm with two sets of interplayed parameters during training.** The data creation parameter θ_{C} generates samples serving as additional training data for the data exploitation parameter θ_{E}, while θ_{E} performs discriminative learning and provides guidance to update θ_{C} through exponential moving average. Finally, θ_{C} performs both discriminative and generative tasks during inference.

As revealed by Ho *et al.* [79], $\mathbf{\Sigma}_\theta(z_t, t)$ has few effects on the results experimentally, therefore estimating $\mu_\theta(z_t, t)$ becomes the main objective. A reparameterization is introduced to estimate it:

$$\mu_\theta(z_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t)\right), \tag{2.4}$$

where $\epsilon_\theta(z_t, t)$ is a denoising network to predict the additive noise ϵ for z_t at timestep t. The final objective is:

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{\mathcal{E}(x), \epsilon\sim\mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t\right)\|_2^2\right]. \tag{2.5}$$

2.3.2 A Unified Model Beyond RGB Generation

In this section, we use diffusion-based models for both discriminative and generative tasks to form our Diff-2-in-1 framework. Concretely, for a diffusion-based unified model Φ, we want it to predict task label $\hat{\mathbf{y}} = \Phi^{\text{dis}}(\mathbf{x})$ given input image $\mathbf{x}$; meanwhile, after training, it can generate multi-modal paired data from Gaussian: $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \Phi^{\text{gen}}(\epsilon)$. We describe how we achieve this below.

Discriminative perspective. Previous work [254, 284] has demonstrated the possibility of using diffusion models for perceptual tasks. Following VPD [284], with the latent code $\mathbf{z} = \mathcal{E}(\mathbf{x})$ from given image $\mathbf{x}$, we perform one-step denoising on $\mathbf{z}$ through the denoising U-Net [186] to produce multi-scale features. Afterward, we rescale and concatenate those features and further pass them to a task head for downstream prediction.

Generative perspective. To generate multi-modal data consisting of paired RGB and visual attributes, we first produce a latent vector $\tilde{z_0}$ by denoising from Gaussian with conditional text. Next, we directly generate the color image $\tilde{x}$ by passing it to the LDM decoder; meanwhile, we perform another one-step denoising with $\tilde{z_0}$ and send the resulting multi-scale features to the task head to obtain the corresponding label $\tilde{y}$.

The two perspectives reflect different usages of the unified diffusion model while they are *not* fully separated: performing generation can be treated as a process of denoise-and-predict for a noisy image at timestep $t = T$; while predicting labels can be treated as a process of data generation conditioned on a given latent vector z_0. This special connection motivates the design of our Diff-2-in-1.

2.4 Learning Mechanism of Diff-2-in-1

To effectively leverage the generated multi-modal data for dense visual perception, we propose a *self-improving* mechanism for our Diff-2-in-1 framework to make the discriminative and generative processes interact with each other, as shown in Figure 2.2. The details are described as below.

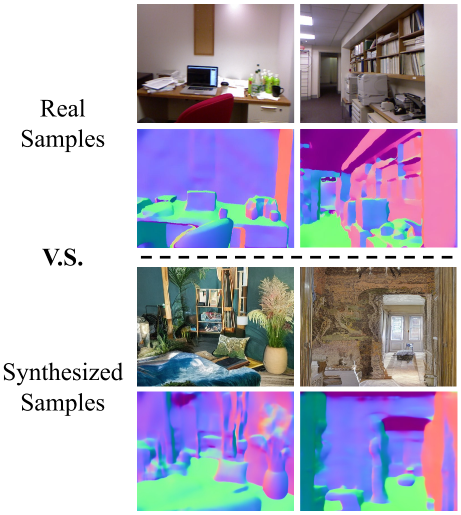


Figure 2.3: Real data samples from NYUv2 and synthesized samples generated from Gaussian noise. The distribution of the generated data varies from the real data distribution.

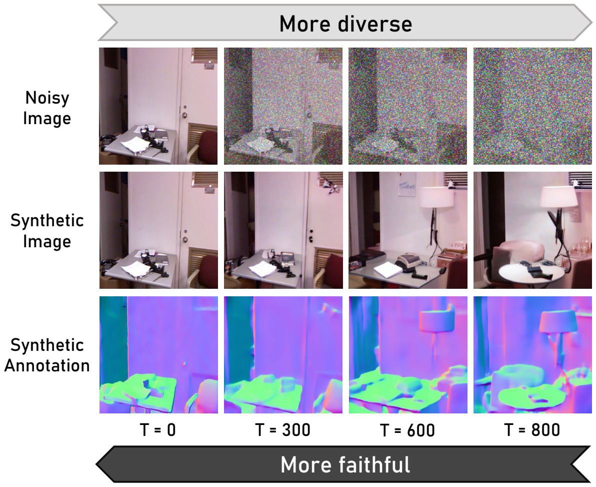


Figure 2.4: In-distribution data generation using partial noise. We generate in-distribution data by denoising from a noisy image at timestep T with $0 < T < T_{\text{max}}$. A larger T leads to greater diversity, whereas a smaller T enhances the resemblance to the original distribution.

2.4.1 Warm-up Stage

Since pretrained diffusion models are only designed for RGB generation, we need a warm-up stage to activate the task head in Figure 2.2 for additional tasks. To achieve this, we train our unified diffusion model using its discriminative learning pipeline with all the original training data with loss

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_{\text{sup}}(f_{\theta_{\text{W}}}(\mathbf{x}_i), \mathbf{y}_i), \tag{2.6}$$

where $\mathcal{L}_{\text{sup}}$ is the supervised loss for our chosen discriminative task on the original paired training data $D_{\text{train}} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$. We obtain a set of parameter weights θ_{W} after this warm-up stage.

2.4.2 Data Generation

Many approaches [26, 51] that use diffusion models for data augmentation generate data from Gaussian noise as discussed in Section 2.3.2. However, as shown in Figure 2.3, the synthetic samples generated from Gaussian noise have a non-negligible distribution shift from the original training data, posing huge obstacles to utilizing the generated data for boosting the discriminative task performance. To narrow down the domain gap between the generated data and original data, inspired by SDEdit [145] and DA-Fusion [221], we use the inherent diffusion-denoising mechanism to control the data generation process.

Concretely, we add noise to the latent z_i of an image $\mathbf{x}_i$ from the training set using Equation 2.2 at a timestep T satisfying $0 < T < T_{\max}$, where $T_{\max}$ is the maximum timestep in the training process of diffusion models ($T_{\max} = 1000$ for all our experiments). This process partially corrupts the image with noise, yet maintains a degree of the original content, as depicted in the first row of Figure 2.4. After denoising the noisy image with Equation 2.3 and decoding with the variational autoencoder, we obtain the synthetic image $\tilde{\mathbf{x}}_i$ with different content but a relatively small domain gap, as shown in the second row of Figure 2.4. At the same time, we can obtain the prediction $\tilde{\mathbf{y}}_i$ which is decoded from the task head of the unified diffusion model. As shown in the third row of Figure 2.4, the generated annotations (surface normal as an example) well match the generated RGB images. The timestep T, representing the noise level, acts as a modulator, balancing the diversity of the generated samples and the fidelity to the in-distribution data: higher noise levels lead to greater diversity, whereas lower levels enhance the resemblance to the original distribution.

2.4.3 Self-improving Stage

To more effectively utilize the generated multi-modal data, we propose a self-improving mechanism inspired by the mean teacher learning system [215]. As shown in Figure 2.2, our self-improving mechanism introduces the following two sets of parameters, both are initialized with θ_{W}, to iteratively perform the self-improvement for both generative and discriminative learning.

Data creation network ($\theta_{\mathbf{C}}$) is used to create samples through the generative

Model	Training Samples	11.25° (↑)	22.5° (↑)	30° (↑)	Mean (↓)	Median (↓)	RMSE (↓)
SkipNet [10]	795	47.9	70.0	77.8	19.8	12.0	28.2
GeoNet [173]	30,816	48.4	71.5	79.5	19.0	11.8	26.9
PAP [282]	12,795	48.8	72.2	79.8	18.6	11.7	25.5
GeoNet++ [174]	30,816	50.2	73.2	80.7	18.5	11.2	26.7
Bae *et al.* [4]	30,816	62.2	79.3	85.2	14.9	7.5	23.5
Bae *et al.* [4]	795	56.6	76.8	83.0	17.2	9.3	26.6
GNA on Bae *et al.*	795	56.4	76.7	83.0	17.3	9.3	26.7
DA-Fusion [221] on Bae *et al.*	795	58.1	77.5	83.6	16.8	8.9	26.1
Diff-2-in-1 on Bae *et al.*(Ours)	795	67.4	83.4	88.2	13.2	6.5	22.0
iDisc [166]	30,816	63.8	79.8	85.6	14.6	7.3	22.8
iDisc [166]	795	57.3	76.4	82.9	17.8	8.8	26.4
GNA on iDisc	795	56.9	76.2	82.4	18.1	8.9	26.7
DA-Fusion [221] on iDisc	795	58.7	78.3	83.4	17.3	8.6	26.2
Diff-2-in-1 on iDisc (Ours)	795	**68.7**	**83.7**	**88.4**	**12.7**	**6.0**	**21.6**

Table 2.1: Surface normal evaluation on NYUv2 [108, 197]. When applying our Diff-2-in-1 on top of state-of-the-art baselines, we achieve consistently and significantly better performance with notably fewer training data, demonstrating the advantages of data efficiency from our unified diffusion model. Additionally, Diff-2-in-1 outperforms augmentation methods GNA and DA-Fusion, proving the usefulness of the multi-modal data generated by our pipeline, and the effectiveness of our self-improving mechanism in utilizing synthetic data.

process within our unified diffusion model. During every iteration, for a batch of m real paired data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{m}$, we additionally generate n paired samples $\{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^{n}$ with θ_{C} following the data creation scheme described in Section 2.4.2. Both real and synthetic data are used for data exploitation.

Data exploitation network ($\theta_{\mathbf{E}}$) is used for exploring the parameter space by exploiting both the original and the synthetic data samples to learn the discriminative task. With those $m + n$ samples, θ_{E} is updated via the discriminative loss:

$$\mathcal{L} = \sum_{i=1}^{m} \mathcal{L}_{\text{sup}}(f_{\theta_{\text{E}}}(\mathbf{x}_i), \mathbf{y}_i) + \sum_{i=1}^{n} \mathcal{L}_{\text{syn}}(f_{\theta_{\text{E}}}(\tilde{\mathbf{x}}_i), \tilde{\mathbf{y}}_i), \tag{2.7}$$

where $\mathcal{L}_{\text{syn}}$ is the loss term for synthetic data for which we regard the generated annotation $\tilde{\mathbf{y}}_i$ as the ground truth. It has the same format as the supervised loss $\mathcal{L}_{\text{sup}}$.

Feedback from data exploitation: EMA optimization. The additional generated data from θ_{C} naturally facilitate the training of θ_{E}. In response, we apply a feedback from θ_{E} to θ_{C} to update its weights every few iterations via the exponential

Model	mIoU (↑)
Swin-L [126]	52.1
ConvNeXt-L [127]	53.2
ConvNeXt-XL [127]	53.6
MAE-ViT-L/16 [72]	53.6
CLIP-ViT-B [182]	50.6
VPD [284]	53.7
DA-Fusion [221] on VPD	54.0
Diff-2-in-1 on VPD (Ours)	**54.5**

Table 2.2: Comparison with diffusion-based segmentation method VPD [284]. The other baselines follow the setting of VPD, which utilize features from supervised pretraining [126, 127], self-supervised pretraining [72], and visual-language pretraining [182] combined with a learnable segmentation head [251]. Our proposed Diff-2-in-1 further improves the performance of the diffusion-based VPD model.

moving average (EMA) strategy:

$$\theta_{\mathrm{C}} \leftarrow \alpha\theta_{\mathrm{C}} + (1-\alpha)\theta_{\mathrm{E}}, \tag{2.8}$$

where $\alpha \in [0, 1)$ is a momentum hyperparameter that is usually set to close to 1. As discussed in Section 2.3.2, the data generation process essentially follows a denoise-and-predict manner, so Equation 2.8 ensures θ_{C} to have a better task prediction head thereby producing higher-quality multi-modal data. A large α maintains the overall quality of the generated data, preventing θ_{C} from getting distracted by the inevitable inferior data.

After the self-improvement, only one set of parameter, θ_{C}, is used to perform both generative and discriminative tasks during inference. The capability of this final model, regarding both discriminative learning and multi-modal data generation, gets promoted simultaneously.

2.5 Experimental Evaluation

2.5.1 Evaluation Setup

We first evaluate our proposed Diff-2-in-1 in the single-task settings with surface normal estimation and semantic segmentation as targets. Next, we apply Diff-2-in-1 in multi-task settings of NYUD-MT [197] and PASCAL-Context [149] to show that

Model	**Semseg** mIoU (↑)	**Depth** RMSE (↓)	**Normal** mErr (↓)
Cross-stitch [147]	36.34	0.6290	20.88
PAP [282]	36.72	0.6178	20.82
PSD [290]	36.69	0.6246	20.87
PAD-Net [253]	36.61	0.6270	20.85
NDDR-CNN [55]	36.72	0.6288	20.89
MTI-Net [224]	45.97	0.5365	20.27
ATRC [25]	46.33	0.5363	20.18
DeMT [258]	51.50	0.5474	20.02
MQTransformer [257]	49.18	0.5785	20.81
DeMT [258]	51.50	0.5474	20.02
InvPT [262]	53.56	0.5183	19.04
DA-Fusion [221] on InvPT	53.70	0.5167	18.81
Diff-2-in-1 on InvPT (Ours)	54.71	**0.5015**	18.60
TaskPrompter [264]	55.30	0.5152	18.47
DA-Fusion [221] on TaskPrompter	55.13	0.5065	18.15
Diff-2-in-1 on TaskPrompter (Ours)	**55.73**	0.5041	**17.91**

Table 2.3: Comparison with state-of-the-art methods on the multi-task NYUD-MT [197] benchmark. Our Diff-2-in-1 brings additional performance gain to the state-of-the-arts.

it can provide universal benefit for more tasks simultaneously.

Datasets and metrics. We evaluate surface normal estimation on the **NYUv2** [108, 197] dataset. Different from previous methods that leverage additional raw data for training, we only use the 795 training samples. We include the number of training samples for each method in Table 2.1 for reference. Following Bae *et al.* [4] and iDisc [166], we adopt $11.25°, 22.5°, 30°$ to measure the percentage of pixels with lower angle error than the corresponding thresholds. We also report the mean/median angle error and the root mean square error (RMSE) of all pixels. We evaluate semantic segmentation on the **ADE20K** [289] dataset and use mean Intersection-over-Union (mIoU) as the metric. For multi-task evaluations, **NYUD-MT** spans across three tasks including semantic segmentation, monocular depth estimation, and surface normal estimation; **PASCAL-Context** takes semantic segmentation, human parsing, saliency detection, and surface normal estimation for evaluation. We adopt mIoU for semantic segmentation and human parsing, RMSE for monocular depth estimation, maximal F-measure (maxF) for saliency detection, and mean error (mErr) for surface normal estimation, following the same standard evaluation schemes [25, 55, 141, 147, 224, 253, 257, 258, 262, 264, 282, 290].

Key implementation details. To speed up training, instead of creating the paired

Model	**Semseg** mIoU (↑)	**Parsing** mIoU (↑)	**Saliency** maxF (↑)	**Normal** mErr (↓)
ASTMT [141]	68.00	61.10	65.70	14.70
PAD-Net [253]	53.60	59.60	65.80	15.30
MTI-Net [224]	61.70	60.18	84.78	14.73
ATRC-ASPP [25]	63.60	60.23	83.91	14.30
ATRC-BMTAS [25]	67.67	62.93	82.29	14.24
MQTransformer [257]	71.25	60.11	84.05	14.74
DeMT [258]	75.33	63.11	83.42	14.54
InvPT [262]	79.03	67.61	84.81	14.15
DA-Fusion [221] on InvPT	79.33	68.45	84.45	14.04
Diff-2-in-1 on InvPT (Ours)	80.36	69.55	84.64	13.89
TaskPrompter [264]	80.89	68.89	**84.83**	13.72
DA-Fusion [221] on TaskPrompter	80.81	69.23	84.47	13.70
Diff-2-in-1 on TaskPrompter (Ours)	**80.93**	**69.73**	84.35	**13.64**

Table 2.4: Comparison on the multi-task PASCAL-Context [149] benchmark. Equipped with our Diff-2-in-1, the state-of-the-art methods reach an overall better performance.

Model	T	11.25° (↑)	22.5° (↑)	30° (↑)	Mean (↓)	Median (↓)	RMSE (↓)
Diff-2-in-1 on Bae *et al.* [4]	300	67.2	83.3	88.1	13.3	6.6	22.1
	600	**67.4**	**83.4**	**88.2**	**13.2**	**6.5**	**22.0**
	800	67.3	83.3	88.1	13.3	6.6	22.1
Diff-2-in-1 on iDisc [166]	300	68.6	83.6	**88.4**	12.8	**6.0**	**21.6**
	600	**68.7**	**83.7**	**88.4**	**12.7**	**6.0**	**21.6**
	800	68.5	83.6	88.3	12.8	**6.0**	**21.6**

Table 2.5: Ablation study on different timesteps T during the data generation process within Diff-2-in-1. A medium timestep $T = 600$ achieves the best performance, but overall Diff-2-in-1 is robust to different choices of T.

data on the fly which takes significantly longer time due to denoising, we pre-synthesize a certain number of RGB images and later use θ_C to produce corresponding labels during the self-improving stage.

2.5.2 Downstream Task Evaluation

Surface normal estimation. We build our Diff-2-in-1 on two state-of-the-art surface normal prediction frameworks:Bae *et al.* [4] and iDisc [166]. Our Diff-2-in-1 creates 500 synthetic pairs with timestep $T = 600$ (refer to Section 2.4.2). Besides conventional methods, we include two additional baselines with diffusion-based data augmentation. *DA-Fusion* [221] generates in-distribution RGB images with labels sharing a similar spirit as us, but only focuses on improving image classification task. To adapt it for dense pixel prediction, we adopt an off-the-shelf captioning strategy [113] to replace its textual inversion and apply the pretrained instantiated model to get the pixelwise

Model	Source → Target	11.25° (↑)	22.5° (↑)	30° (↑)	Mean (↓)	Median (↓)	RMSE (↓)
Bae *et al.* [4]	ScanNet → NYUv2	59.0	77.5	83.7	16.0	8.4	24.7
	NYUv2 → NYUv2	62.2	79.3	85.2	14.9	7.5	23.5
Diff-2-in-1 on Bae *et al.* [4](Ours)	ScanNet → NYUv2	**63.0**	**80.4**	**86.0**	**14.6**	**7.3**	**23.3**

Table 2.6: Cross-domain evaluation on the surface normal estimation task of NYUv2 [108, 197]. The performance of our method trained on ScanNet even outperforms the baseline Bae *et al.* trained on NYUv2, suggesting our generalizability to unseen datasets.

annotations for the generated images. Afterward, the generated RGB-annotation pairs are utilized in the same way as DA-Fusion originally uses RGB-class pairs to boost the performance. *Gaussian Noise Augmentation (GNA)* is a self-constructed baseline that generates additional data by denoising from Gaussian noise, then applies the self-improving strategy to utilize the generated data.

With the results shown in Table 2.1, we observe: **(1)** When applying our Diff-2-in-1 on top of the state-of-the-art baselines, we achieve significantly better performance with notably fewer training data, demonstrating the great advantages of data efficiency from a unified diffusion model. **(2)** Our Diff-2-in-1 has better performance than other augmentation methods like GNA and DA-Fusion, showcasing the usefulness of the multi-modal data generated by our pipeline, and the effectiveness of synthetic data utilization with our self-improving mechanism. **(3)** Our Diff-2-in-1 is a general design that can universally bring benefits to different discriminative backbones.

Semantic segmentation. We instantiate our Diff-2-in-1 on VPD [284], a diffusion-based segmentation model. For self-improving, we synthesize one sample for each image in the training set. With the results shown in Table 2.2, we observe that the diffusion-based VPD can benefit from our paradigm by effectively performing self-improvement to leverage the generated samples.

Multi-task evaluations. We apply our Diff-2-in-1 on two state-of-the-art multi-task methods, InvPT [262] and TaskPrompter [264]. A total of 500 synthetic samples are generated for NYUD-MT following the surface normal evaluation. For PASCAL-Context, one sample is synthesized for each image in the training set with our Diff-2-in-1. The comparisons on NYUD-MT and PASCAL-Context are shown in Table 2.3 and Table 2.4, respectively. The results validate that our Diff-2-in-1 is a versatile design that can elevate the performance of a wide variety of vision tasks.

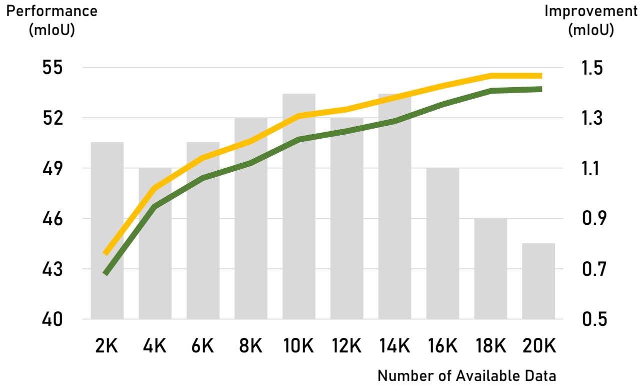


Figure 2.5: Ablation study on different data settings with our Diff-2-in-1. *Green line*: Performance of the baseline VPD. *Yellow line*: Performance with our Diff-2-in-1. *Gray bars*: Improvement in each data setting. Our Diff-2-in-1 could consistently bring performance gain for all different data settings with more benefits in mid-range data settings.

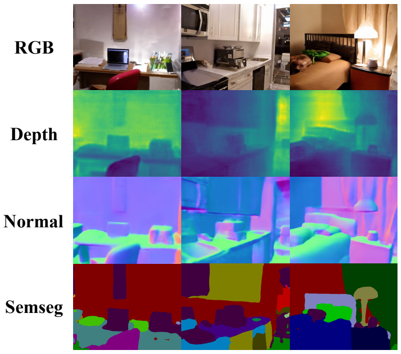


Figure 2.6: Multi-modal samples generated by Diff-2-in-1 on NYUD-MT [197]. Our method can generate high-quality RGB images and precise multi-modal annotations, further facilitating discriminative learning via our self-improvement.

2.5.3 Ablation Study

In this section, we offer a better understanding of the superiority of our Diff-2-in-1 by answering the three primary questions.

How does timestep T in data creation affect final performance? As illustrated in Figure 2.4, the timestep T balances the trade-off between the content variation and domain shift of the generated data. We ablate different timesteps $T \in \{300, 600, 800\}$ in the experiments on surface normal instantiated on Bae *et al.* [4] and iDisc [166]. The results in Table 2.5 indicate that we achieve the best performance when $T = 600$, with a balance of data diversity and quality. Nevertheless, it is noteworthy that our performance is generally robust to different choices of T.

How robust is Diff-2-in-1 for domain shift? We perform the cross-domain evaluation to show that our Diff-2-in-1 has strong generalizability. We train both the baseline Bae *et al.* [4] and our Diff-2-in-1 on the ScanNet [42] dataset for the surface normal estimation task, and evaluate the performance on the test set of NYUv2 [108, 197]. Interestingly, with the results shown in Table 2.6, we find that the performance of our method trained on ScanNet even outperforms the baseline

Bae *et al.* trained on NYUv2, suggesting the generalizability of our method to unseen datasets and its great potential in real practice.

How Diff-2-in-1 is helpful in different data settings? We ablate different settings when the number of available training samples for Diff-2-in-1 varies to investigate whether it is more helpful in data abundance or data shortage scenarios. We run this ablation for semantic segmentation on the ADE20K dataset: we randomly select 10% (2K) to 90% (18K) samples with 10% (2K) intervals in between, assuming that Diff-2-in-1 only gets access to partial data. In each setting, one additional sample for each image is generated using our data generation scheme.

With the results shown in Figure 2.5, we offer the following observations: (1) Diff-2-in-1 consistently boosts the performance under all settings, with improvement ranging from 0.8 to 1.4 in mIoU, indicating the effectiveness and robustness of our method. (2) Diff-2-in-1 provides more benefits in the data settings from 40% (8K) to 70% (14K). We analyze the reasons including that when the data are scarce, it is relatively hard to train a good model via Equation 2.6 to provide high-quality multi-modal synthetic data for self-improvement. On the other hand, when the data are already adequate, there is less demand for more diverse data. Under both scenarios, the benefit of our method is still noticeable yet less significant.

2.5.4 Synthetic Data Evaluation

In addition to Figure 2.4, we visualize samples generated by our method on NYUD-MT [197] in Figure 2.6. Diff-2-in-1 is able to generate high-quality RGB images and precise multi-modal annotations, further facilitating discriminative learning via our self-improvement.

Generated samples serving as data augmentation. We select surface normal estimation as the target task and train an external discriminative model, Bae *et al.* [4], under the following two settings: **(1)** only use the original 795 samples to train the model until convergence *(GT Only)*; and **(2)** finetune the converged model in *GT Only* using the mixture of original samples and generated samples from our Diff-2-in-1 before the self-improving stage *(GT + Syn)*. For (2), we generate 500 synthetic samples with $T = 600$ and naively merge them together with the original samples. We report two variants of setting (2) with generated samples before or after

Setting	11.25° (↑)	22.5° (↑)	30° (↑)	Mean (↓)	Median (↓)	RMSE (↓)
GT Only	56.6	76.8	83.0	17.2	9.3	26.6
GT + Syn (Before Self-improving)	57.5	**77.1**	**83.3**	17.1	9.1	**26.5**
GT + Syn (After Self-improving)	**57.8**	**77.1**	**83.3**	**17.0**	**9.0**	**26.5**

Table 2.7: Comparison between two data settings. *GT Only*: Use real samples to train Bae *et al.* [4] until converges. *GT + Syn*: Further finetune the converged model with real and synthetic samples. Synthetic data further boost the performance of a converged model, demonstrating their realism.

Backbone	Setting	11.25° (↑)	22.5° (↑)	30° (↑)	Mean (↓)	Median (↓)	RMSE (↓)
Bae *et al.*[4]	Synthetic	67.4	83.4	**88.2**	**13.2**	**6.5**	**22.0**
	Real	**67.5**	**83.5**	**88.2**	**13.2**	**6.5**	**22.0**
iDisc [166]	Synthetic	**68.7**	**83.7**	**88.4**	**12.7**	**6.0**	21.6
	Real	**68.7**	**83.7**	**88.4**	12.8	**6.0**	**21.5**

Table 2.8: Comparison between using generated samples and unlabeled real images in NYUv2 surface normal estimation. Comparable performance proves the premium quality of our generated data.

the self-improving stage in Table 2.7. We have the following observations: firstly, the synthetic samples are capable of boosting the performance of a converged model, indicating that the generated RGB and annotation maps are consistent. Moreover, the generated multi-modal data get refined during the self-improving stage, verifying the effectiveness of our method towards generation.

Synthetic data V.S. real data. In the surface normal task, we replace the 500 generated samples with 500 additional real captured images from NYUv2 raw video clips. The annotations of them are produced by our Diff-2-in-1 on the fly. Then, we use the same training strategy to train Diff-2-in-1. As shown in Table 2.8, using our generated data achieves comparable performance to using the real captured data, proving the premium quality of the synthetic data.

2.6 Conclusion

In this paper, we bridge generative and discriminative learning by proposing a unified diffusion-based framework Diff-2-in-1. It enhances discriminative learning through the generative process by creating diverse while faithful data, and gets the discriminative and generative processes to interplay with each other using a self-improving learning mechanism. Extensive experiments demonstrate its superiority in various settings of discriminative tasks, and its ability to generate high-quality multi-modal data characterized by both realism and usefulness.

2.7 Implementation Details

2.7.1 Architecture Details

Feature extraction from diffusion models. We first describe how we extract features for downstream dense prediction tasks from the pretrained stable diffusion model [185] in our framework, which is generally applicable to all the model instantiations discussed below. We take the latent vector obtained from the VAE encoder in stable diffusion as input for the denoising network, followed by a one-step denoising to obtain the features. Since the denoising operation in stable diffusion is realized by a U-Net [186] module, multi-scale features can be obtained through the one-step denoising process for a given image. As we use the publicly released stable diffusion pretrained weight `Stable Diffusion v1-5` which is finetuned on 512×512 resolution, the input images are also resized to 512×512 before being processed by our model. Therefore, the raw multi-scale features $\{f_i^{\text{raw}}\}_{i=0}^3$ extracted from our model are in the spatial resolutions of 8×8, 16×16, 32×32, and 64×64. Following Li et al. [117], for each pair of features $f_{i-1}^{\text{raw}}, f_i^{\text{raw}} (1 \leq i \leq 3)$ with adjacent resolutions, we upsample the lower-resolution feature to the higher-resolution one, concatenating them, and processing with a convolutional layer:

$$f_i^{\text{proc}} = \text{Conv}(\text{Up}(f_{i-1}^{\text{raw}}), f_i^{\text{raw}}). \tag{2.9}$$

Then, we get the processed multi-scale features $\{f_i^{\text{proc}}\}_{i=1}^3$ which are further used for fitting into the specific network architectures when we build our Diff-2-in-1 on existing works.

Surface normal estimation. For both Bae et al. [4] and iDisc [166], the surface normal maps are decoded from multi-scale features extracted by their original encoder. When instantiating our Diff-2-in-1 upon them, we replace their original encoders with the unified model described above. If the decoder requires a feature map with a spatial resolution unavailable in $\{f_i^{\text{proc}}\}_{i=1}^3$, we use a similar strategy as Equation 2.9 to obtain the feature of a new spatial resolution. If the features required are of higher resolution than the existing features, then we increase the resolution range of the features by

$$f_{i+1}^{\text{proc}} = \text{Conv}(\text{Up}(f_i^{\text{proc}}), \text{Deconv}(f_i^{\text{proc}})), \tag{2.10}$$

where the upsampling and deconvolutional [157] layers increase the feature size by the same ratio. For obtaining lower resolution features, we simply replace the upsampling and deconvolutional layers in Equation 2.10 with downsampling and convolutional layers. The upsampling or downsampling factor in Equation 2.10 is set to 2. Moreover, we can iteratively perform Equation 2.10 multiple times if the required features are more than twice larger or smaller than the features $\{f_i^{\text{proc}}\}_{i=1}^{3}$ from the pretrained stable diffusion model.

Semantic segmentation. As VPD [284] also builds upon stable diffusion [185], we directly apply the self-improving algorithm in our Diff-2-in-1 on VPD to boost its performance.

Multi-task learning. The decoder of InvPT [262] requires multi-scale features. Therefore, we use the same strategy as the surface normal estimation methods [4, 166] to provide the decoder with the required features. The decoder of TaskPrompter [264] only requires single-scale features. Therefore, we use Equation 2.10 to resize all the features in $\{f_i^{\text{proc}}\}_{i=1}^{3}$ to this specific scale. As a result, the multi-scale knowledge extracted from stable diffusion can be injected into the TaskPrompter framework. Additionally, both InvPT and TaskPrompter adopt pretrained ViT [47] or Swin Transformer [126] as their encoders. To better utilize the prior knowledge within the original encoders, we merge the knowledge from the two sources by adding the features from stable diffusion to their original encoders.

Summary. From the instantiations above, we have the following guidelines for converting existing methods to the unified diffusion-based models in our Diff-2-in-1: **(1)** By default, we replace the encoders in the original models with the stable diffusion feature extractor; **(2)** If the features required by the original decoder is unavailable in the multi-scale features, we can use Equation 2.10 to expand the range of the multi-scale features; **(3)** If the original model design contains a pretrained encoder, we consider merging the knowledge of the stable diffusion model and the pretrained encoder.

2.7.2 Text Prompts

Our Diff-2-in-1 uses the generative nature of diffusion models to create samples, which requires text prompts as conditions during the denoising process to generate high-quality samples. However, the text prompts are not always available in our target datasets. To solve this challenge, we use the off-the-shelf image captioning model BLIP-2 [113] to generate text descriptions for each image. The generated text descriptions serve as conditions when performing denoising to generate new data samples with our Diff-2-in-1. We further show in the ablation study in Section 2.8 that the choice of the image captioning model has little influence on the performance.

2.7.3 Additional Training Details

In the warm-up stage, we follow the same hyperparameters of the learning rate, optimizer, and training epochs of the original works that our Diff-2-in-1 builds on. In the self-improving stage, the exploitation parameter θ_E continues the same training scheme in the warm-up stage, while the creation parameter θ_C updates once when θ_E consumes 40 samples. Thus, the interval of the EMA update for θ_C depends on the batch size used in the self-improving stage. For the surface normal estimation and semantic segmentation tasks, we adopt a batch size of 4, so the EMA update happens every 10 iterations. For the multi-task frameworks, the batch size is 1, so we perform the EMA update every 40 iterations. The momentum hyperparameter α for the EMA update is set as 0.999 for multi-task learning on PASCAL-Context [149], and 0.998 for the rest of the task settings.

2.8 Additional Ablation Study

What text prompts to use for the unified diffusion model? As mentioned in Section 2.7.2, we adopt BLIP-2 to generate text prompts for creating new samples based on the reference images. *What if the text prompters are less powerful?* We show that different choices of image captioning models have a marginal influence on the performance of our Diff-2-in-1. We first show the captions generated by BLIP-2 and another relatively weaker model ClipCap [148] in Figure 2.7. The captions generated

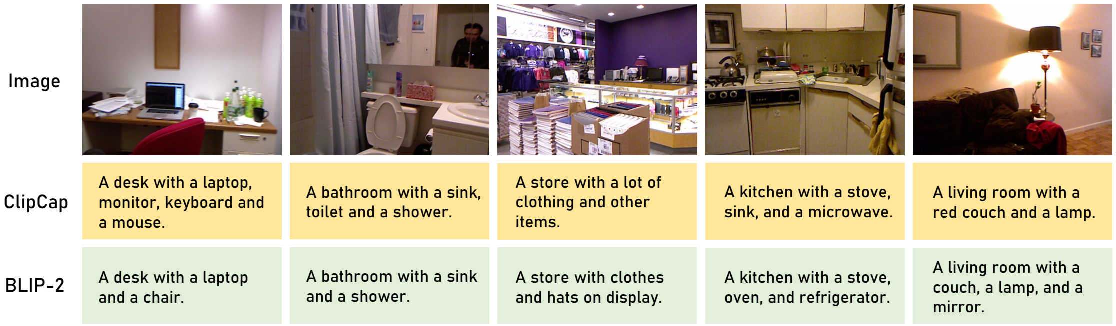


Image					
ClipCap	A desk with a laptop, monitor, keyboard and a mouse.	A bathroom with a sink, toilet and a shower.	A store with a lot of clothing and other items.	A kitchen with a stove, sink, and a microwave.	A living room with a red couch and a lamp.
BLIP-2	A desk with a laptop and a chair.	A bathroom with a sink and a shower.	A store with clothes and hats on display.	A kitchen with a stove, oven, and refrigerator.	A living room with a couch, a lamp, and a mirror.

Figure 2.7: Captions generated by ClipCap [148] and BLIP-2 [113] on the NYUv2 [197] dataset. The generated captions using these two off-the-shelf image captioning models not only have similar semantic meanings, but also share similar text formats.

Model	Caption	11.25° (↑)	22.5° (↑)	30° (↑)	Mean (↓)	Median (↓)	RMSE (↓)
Diff-2-in-1 on Bae et al. [4]	None	66.0	83.0	88.0	13.6	7.0	**22.0**
	ClipCap [148]	67.3	**83.4**	**88.2**	**13.2**	**6.5**	**22.0**
	BLIP-2 [113]	**67.4**	**83.4**	**88.2**	**13.2**	**6.5**	**22.0**
Diff-2-in-1 on iDisc [166]	None	67.2	83.4	88.1	13.0	6.6	21.7
	ClipCap [148]	**68.7**	**83.7**	**88.4**	**12.7**	**6.0**	**21.6**
	BLIP-2 [113]	**68.7**	**83.7**	**88.4**	**12.7**	**6.0**	**21.6**

Table 2.9: Ablation study on using text prompts from different off-the-shelf image captioning models ClipCap [148] and BLIP-2 [113] to generate samples with Diff-2-in-1. The evaluation is conducted on the surface normal estimation task on the NYUv2 [108, 197] dataset. Our Diff-2-in-1 is robust to different choices of image captioning models. Nevertheless, it is necessary to have an image captioning model to provide text prompts in the denoising process during data generation.

by these two off-the-shelf models have similar semantic meanings, as well as sharing similar formats of *"A [Place] with [Object 1], [Object 2], ..., [Object N-1], and [Object N]."* We further evaluate the performance of using the text prompts from ClipCap and BLIP-2 to generate synthetic samples for the self-improving learning system in Diff-2-in-1. The results are shown in Table 2.9. We can observe that once again there is no large difference between the two variants and both of them greatly outperform the baseline, demonstrating that our Diff-2-in-1 is robust to different text prompters used during the denoising process for data generation. Nonetheless, it does not indicate that the image captioning model is dispensable. If we completely get rid of the image captioning model and do not use text as the condition during denoising (*None* for *Caption* in Table 2.9), we could observe an evident drop in the performance on discriminative tasks.

Should we finetune the diffusion backbone? As shown in Figure 2.3, if the generation process of our unified diffusion model starts from Gaussian noise, the generated samples will have an evident domain shift from the original distribution. Therefore, we adopt the halfway diffusion-denoising mechanism to synthesize in-distribution data. Another potential solution to overcome the domain shift issue is to finetune the stable diffusion backbone. We test this setting with two finetuning strategies for a comprehensive ablation: (1) directly finetune all the parameters of the denoising U-Net *(Direct Finetuning)*; (2) adopt parameter-efficient finetuning strategy Low-Rank Adaptation (LoRA) [83] on the denoising modules of stable diffusion *(LoRA Finetuning)*. We conduct the experiments on the surface normal task on the NYUv2 dataset with Bae et al. [4] as the task head. The results are shown in Table 2.10. The inferior performance of using the finetuned stable diffusion indicates that the diffusion-denoising data generation scheme and the self-improving learning system in our Diff-2-in-1 are essential. One factor for the unsatisfactory performance of using finetuning is that the finetuning process incurs a loss in the generalization capability, especially during finetuning with limited data (*e.g.*, 795 samples on NYUv2), making the features extracted from the stable diffusion model less informative for decoding visual task predictions. In comparison, our proposed diffusion-denoising data generation scheme injects external knowledge from the pretrained stable diffusion model to the samples in the training data, without risks of knowledge forgetting with respect to its discriminative ability.

What timestep T to choose for discriminative feature extraction? In our current experiments, we follow existing works ODISE [254] and VPD [284] to adopt $T = 0$ as the timestep for feature extraction from the pretrained stable diffusion model. We ablate different timesteps T for extracting features from stable diffusion in Table 2.11. The performance is generally satisfactory with relatively small timesteps T, which add little noise to the clean latents before extracting features from denoising U-Net. We do not attentively optimize for the best T and it is likely that a better T may exist in other settings which can further improve the performance of our Diff-2-in-1. We leave the exploration of optimal T for different tasks as future work.

How to choose hyperparameters for the EMA update? We ablate the choice of $\alpha \in [0.99, 0.999]$ for the EMA update according to guidelines in Liu et al. [125].

Setting	11.25° (↑)	22.5° (↑)	30° (↑)	Mean (↓)	Median (↓)	RMSE (↓)
Direct Finetuning	58.0	76.5	82.4	16.9	8.7	26.5
LoRA Finetuning	64.8	82.0	87.4	14.1	7.3	22.8
Diff-2-in-1 (Ours)	**67.4**	**83.4**	**88.2**	**13.2**	**6.5**	**22.0**

Table 2.10: Ablation study on strategies to finetune the diffusion backbone. *Direct Finetuning*: Directly finetune the denoising U-Net. *LoRA Finetuning*: Adopt LoRA [83] to finetune the U-Net. Their unsatisfactory results indicate that the features extracted from the finetuned network are less informative and have worse generalizability. The information loss introduced by finetuning is inevitable even if using the parameter-efficient finetuning technique LoRA to mitigate forgetting. In contrast, our diffusion-denoising strategy injects external knowledge from the pretrained stable diffusion to the samples, without risks of forgetting the discriminative ability of diffusion models.

T	11.25° (↑)	22.5° (↑)	30° (↑)	Mean (↓)	Median (↓)	RMSE (↓)
0	67.4	**83.4**	**88.2**	**13.2**	**6.5**	**22.0**
50	**67.5**	83.3	88.1	**13.2**	**6.5**	**22.0**
100	66.9	82.6	87.5	13.5	**6.5**	22.4
150	65.5	81.6	86.7	14.0	6.8	23.0

Table 2.11: Ablation study on extracting features from the pretrained stable diffusion model with different timesteps T on NYUv2 surface normal evaluation. Our Diff-2-in-1 achieves better performance with smaller T in this task setting.

The results with Bae et al. [4] on the NYUv2 [108, 197] surface normal task are shown in Table 2.12 where $\alpha = 0.998$ achieves the best performance. Nevertheless, the performance of our Diff-2-in-1 is robust to different choices of α within a broad range.

2.9 More Visualizations

We provide more qualitative results from the following two aspects: **(1)** performance comparison with state-of-the-art methods on discriminative tasks and **(2)** multi-modal data generation quality of the synthetic samples from our Diff-2-in-1.

2.9.1 Comparisons on Discriminative Tasks

The qualitative comparisons of our Diff-2-in-1 and the baselines are shown in Figures 2.8, 2.9 (surface normal prediction) and 2.10 (multi-task). Our Diff-2-in-1

α	11.25° (↑)	22.5° (↑)	30° (↑)	Mean (↓)	Median (↓)	RMSE (↓)
N/A (Baseline)	62.2	79.3	85.2	14.9	7.5	23.5
0.99	67.1	83.2	88.1	13.4	6.6	22.1
0.993	67.3	**83.4**	**88.2**	13.3	6.6	**22.0**
0.996	67.3	**83.4**	**88.2**	13.3	6.6	**22.0**
0.998	**67.4**	**83.4**	**88.2**	**13.2**	**6.5**	**22.0**
0.999	67.1	83.3	88.1	13.3	6.7	22.1

Table 2.12: Ablation study on different α for the EMA update within Diff-2-in-1. $\alpha = 0.998$ reaches the best performance in this setting of surface normal prediction with Bae et al. [4] on NYUv2. Nonetheless, our Diff-2-in-1 is robust to different α within a broad range.

outperforms the baselines, demonstrating the competence of our unified diffusion-based model in the discriminative perspective.

2.9.2 Data Generation Quality

We display the synthetic multi-modal data from our Diff-2-in-1 data creation framework in Figures 2.11, 2.12 (RGB-normal pairs) and 2.13, 2.14 (RGB and multiple annotations) to show that Diff-2-in-1 has powerful generation ability that is capable of generating high-quality and consistent samples.

2.10 Discussions and Future Work

Limitation. One major limitation of this work is that adopting diffusion models for data generation is relatively time-consuming as diffusion models typically need multi-step denoising to produce samples. To alleviate this shortcoming, current advancement on accelerating the inference process of diffusion models [124, 132, 266, 285] can be adopted to speed up the data generation process.

Future work. Looking ahead, the potential applications of this unified approach are vast. Future research directions include extending this methodology to other types of tasks, such as 3D detection, and refining and optimizing the Diff-2-in-1 framework such as a more efficient data creation scheme and knowledge transfer to a new domain.

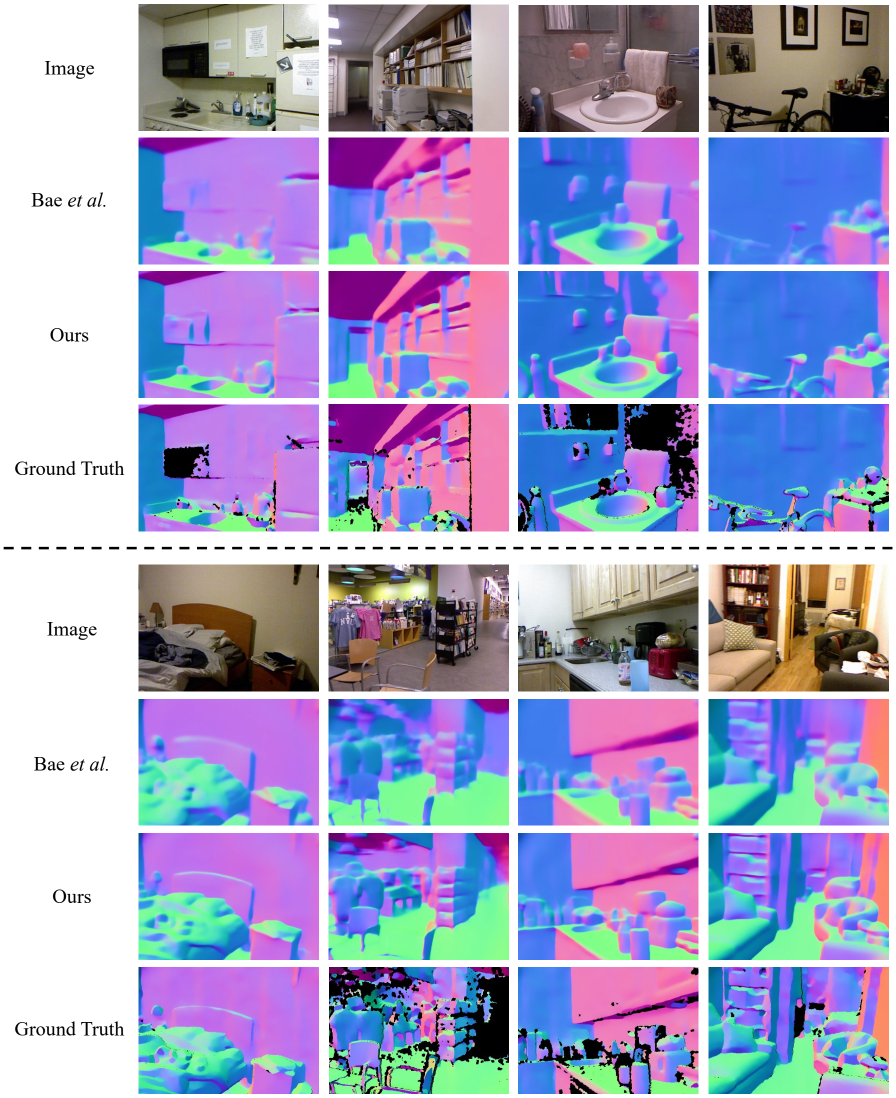


Figure 2.8: Qualitative results on the surface normal prediction task of NYUv2 [108, 197]. Our proposed Diff-2-in-1 outperforms the baseline with more accurate surface normal estimations, indicating that our unified diffusion-based models excel at handling discriminative tasks. The black regions in the ground truth visualizations are invalid regions.

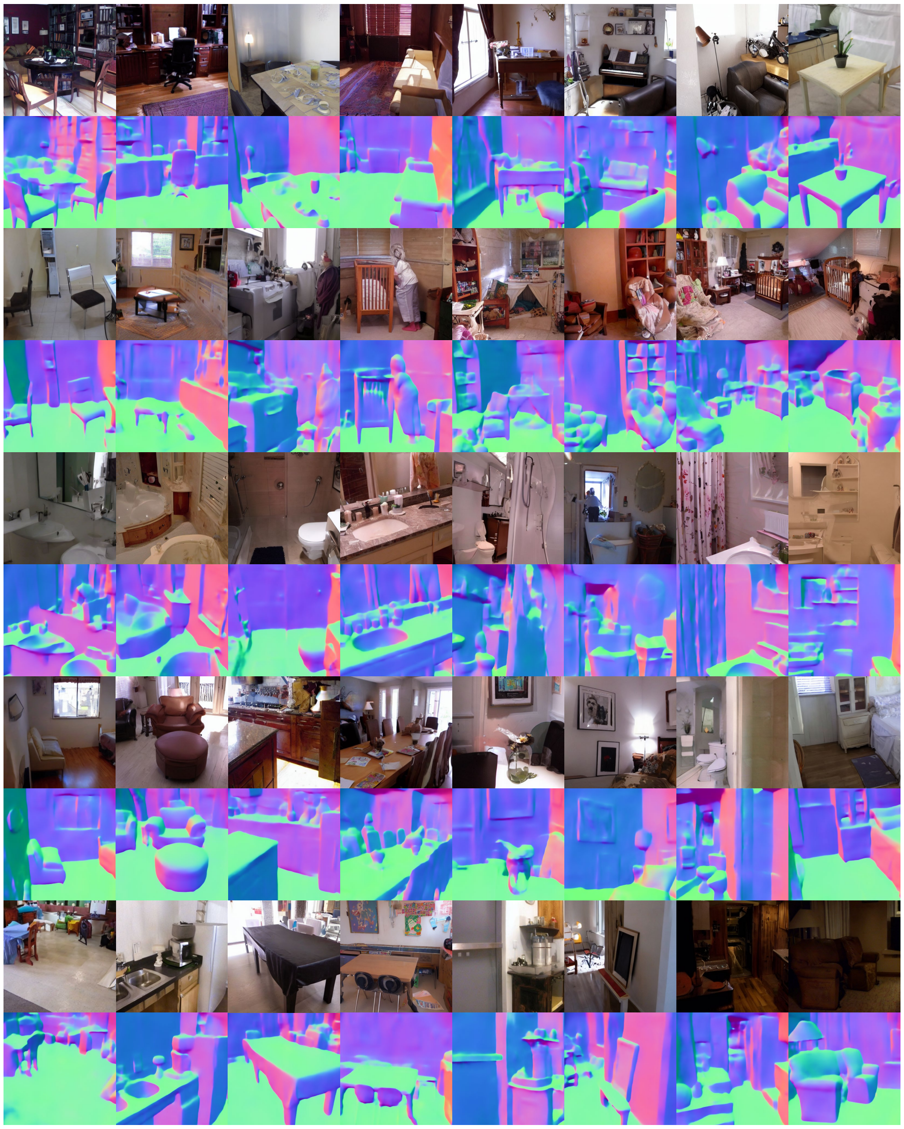
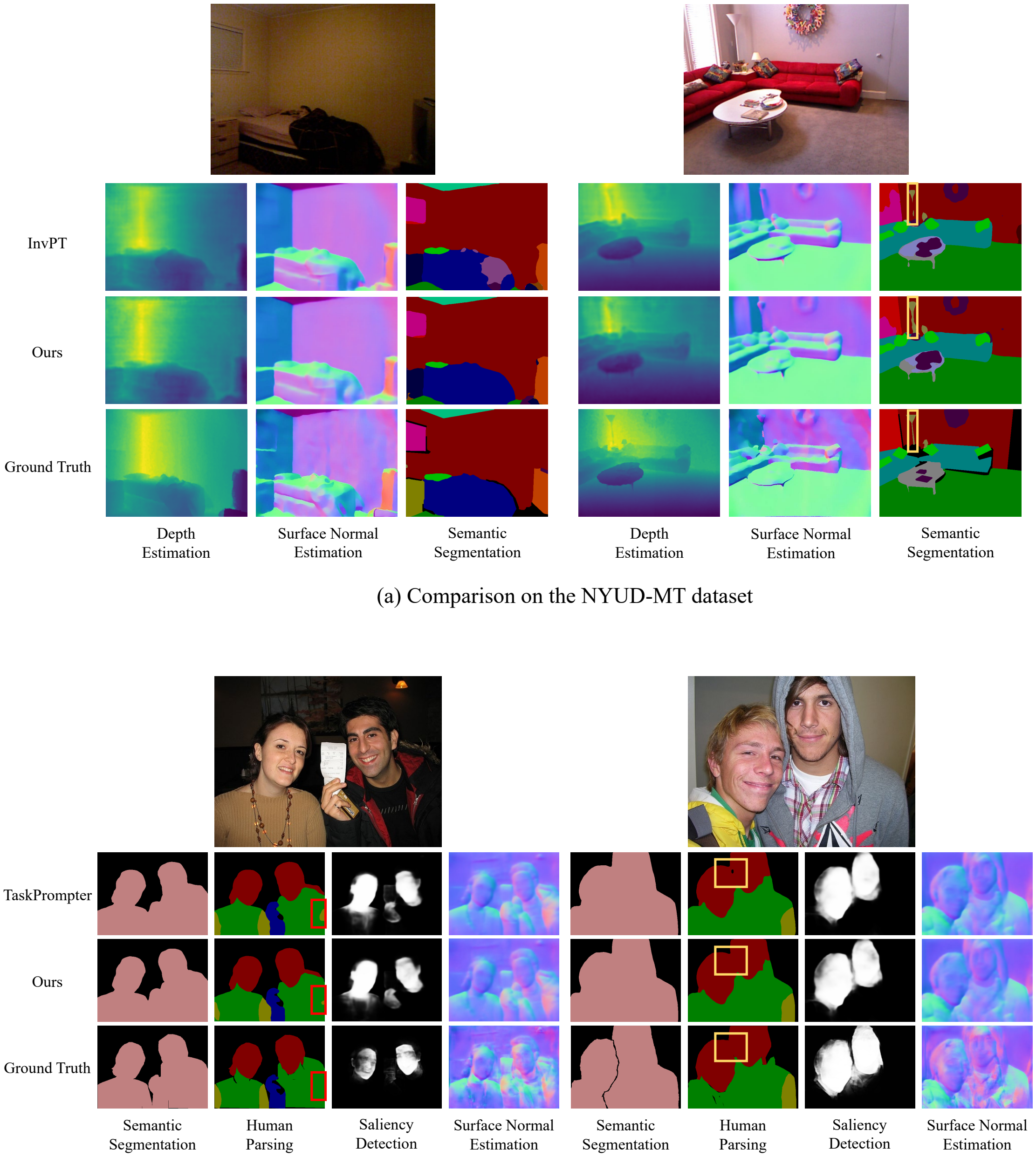


Figure 2.9: Qualitative results on the surface normal task of NYUv2 [108, 197]. Our proposed Diff-2-in-1 outperforms the baseline with more accurate surface normal estimations, indicating that our unified diffusion-based models excel at handling discriminative tasks. The black regions in the ground truth visualizations are invalid regions.



(a) Comparison on the NYUD-MT dataset

(b) Comparison on the PASCAL-Context dataset

Figure 2.10: Qualitative results on the multi-task datasets NYUD-MT [197] and PASCAL-Context [149]. Diff-2-in-1 has superior performance compared to the baselines, demonstrating the effectiveness of our unified diffusion-based model design. Zoom in for the regions with bounding boxes to better see the comparison.

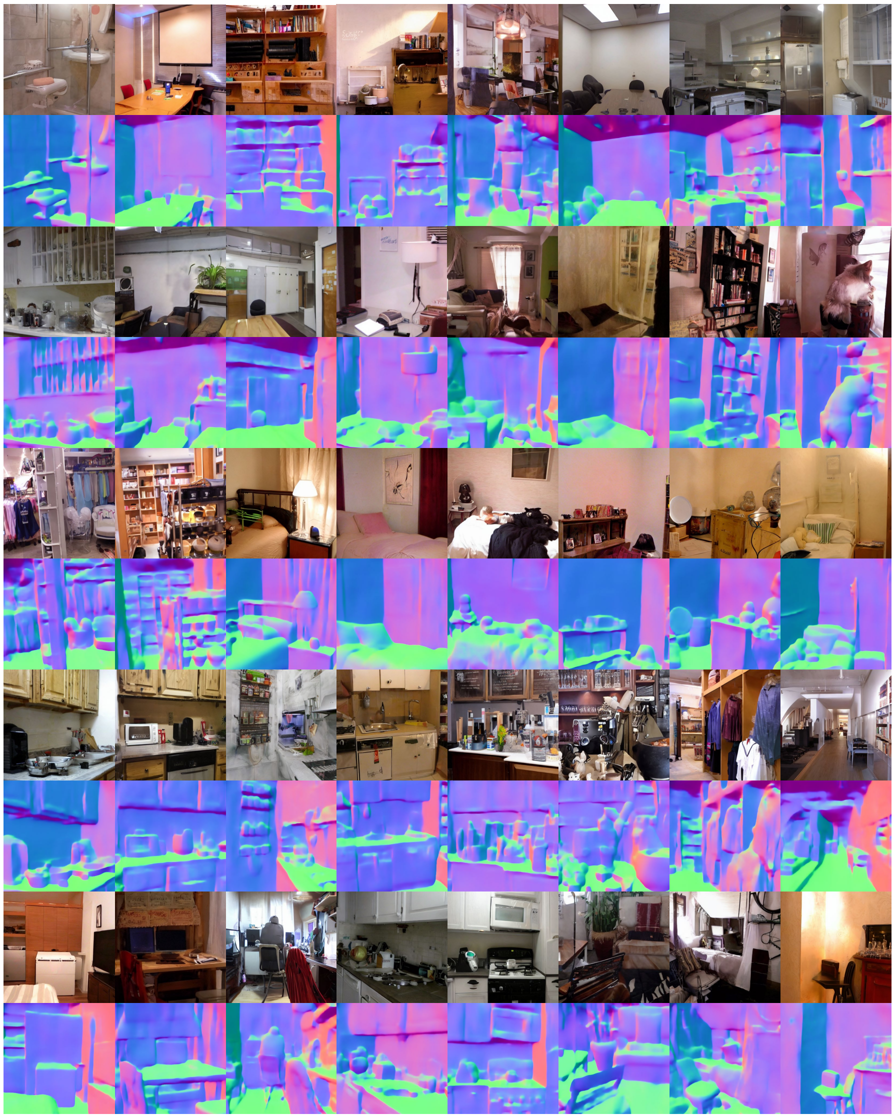


Figure 2.11: Synthetic samples from our method after the Diff-2-in-1 framework is trained on the surface normal task of NYUv2 [108, 197]. The odd rows are the generated RGB images while the even rows are the generated surface normal maps. The model is capable of generating diverse and high-fidelity images with the corresponding surface normal maps matching the generated RGB images.

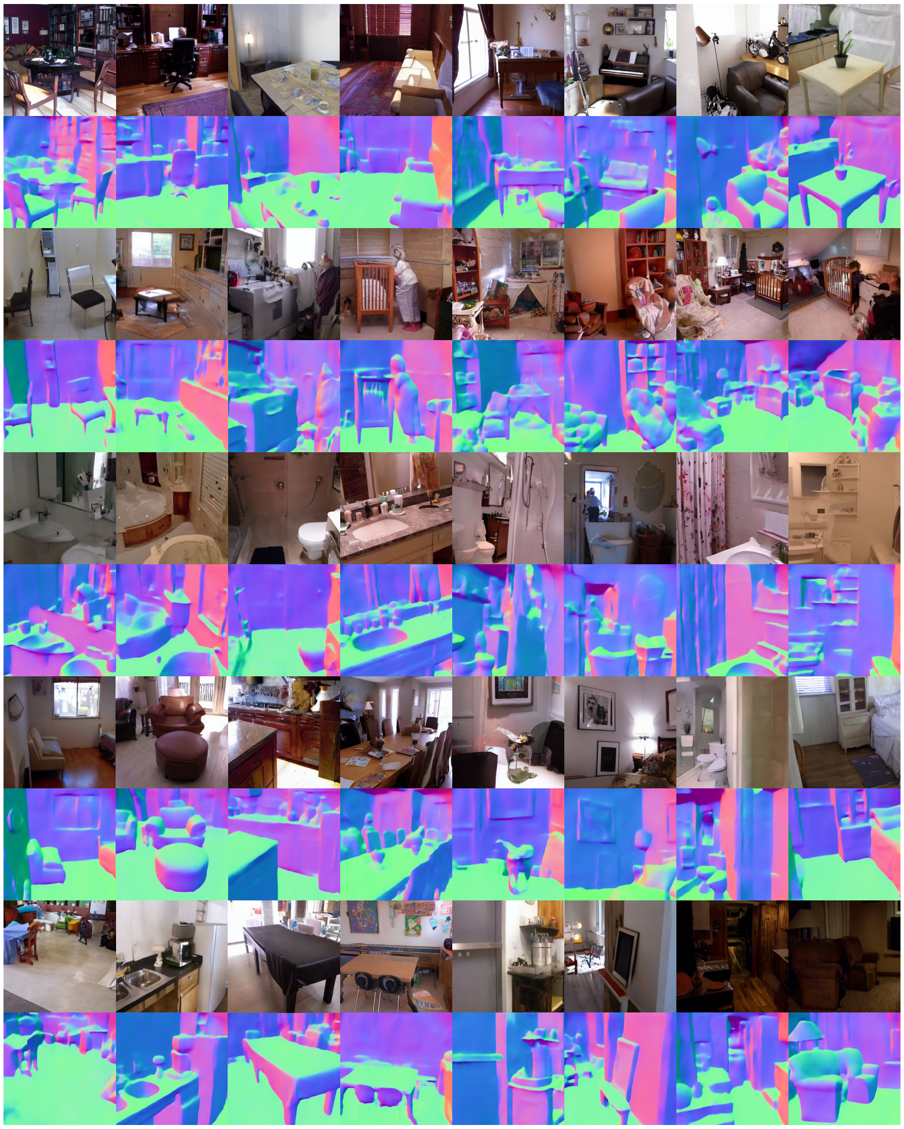


Figure 2.12: Synthetic samples from our method after the Diff-2-in-1 framework is trained on the surface normal task of NYUv2 [108, 197]. The odd rows are the generated RGB images while the even rows are the generated surface normal maps. The model is capable of generating diverse and high-fidelity images with the corresponding surface normal maps matching the generated RGB images.

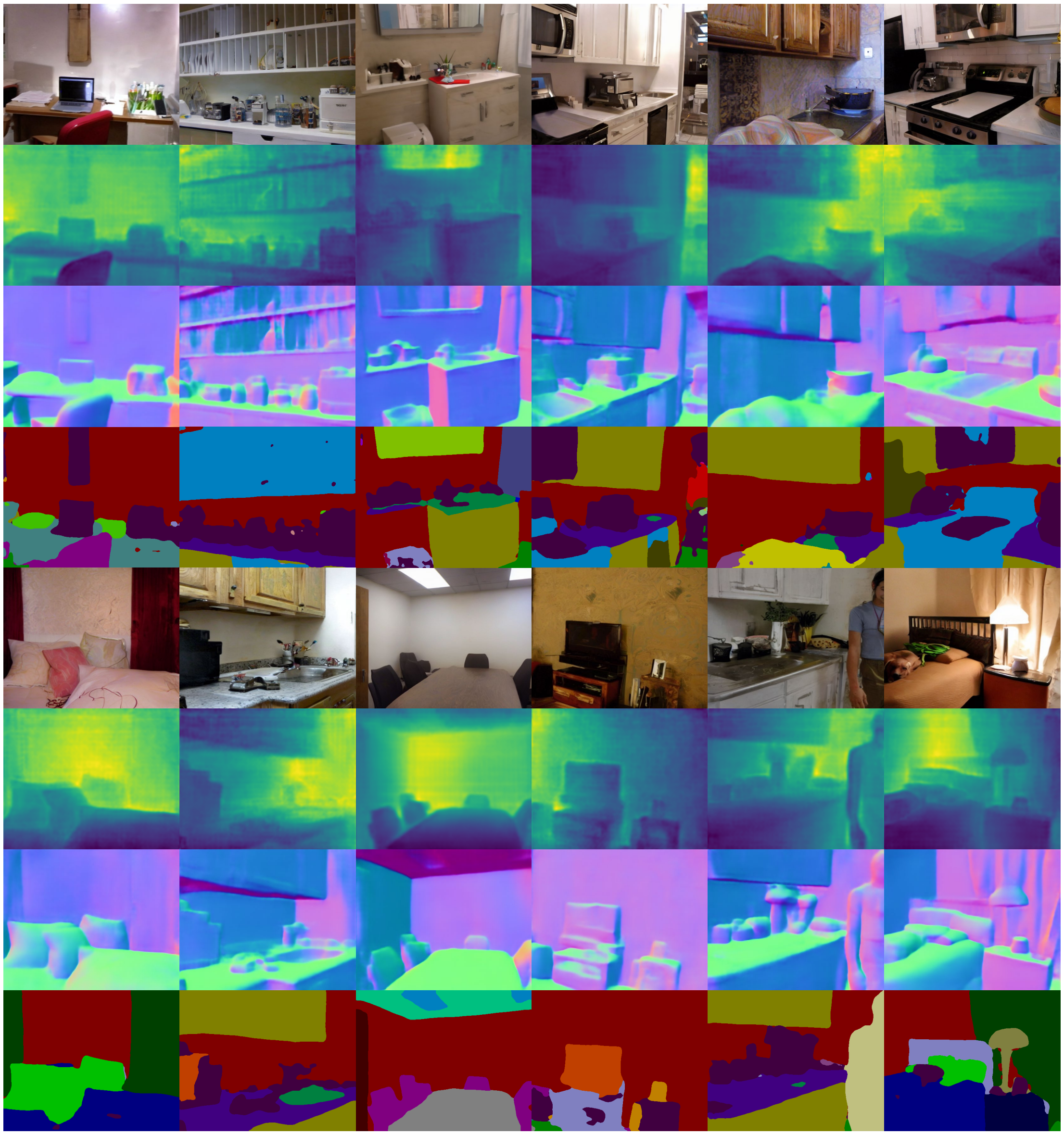


Figure 2.13: Synthetic samples from our method after the Diff-2-in-1 framework is trained on the multi-task setting of NYUD-MT [197]. Each batch of samples contains four rows: RGB, depth map, surface normal map, and semantic labels *(from top to bottom)*. The generated samples are of high quality with their multi-task annotations.

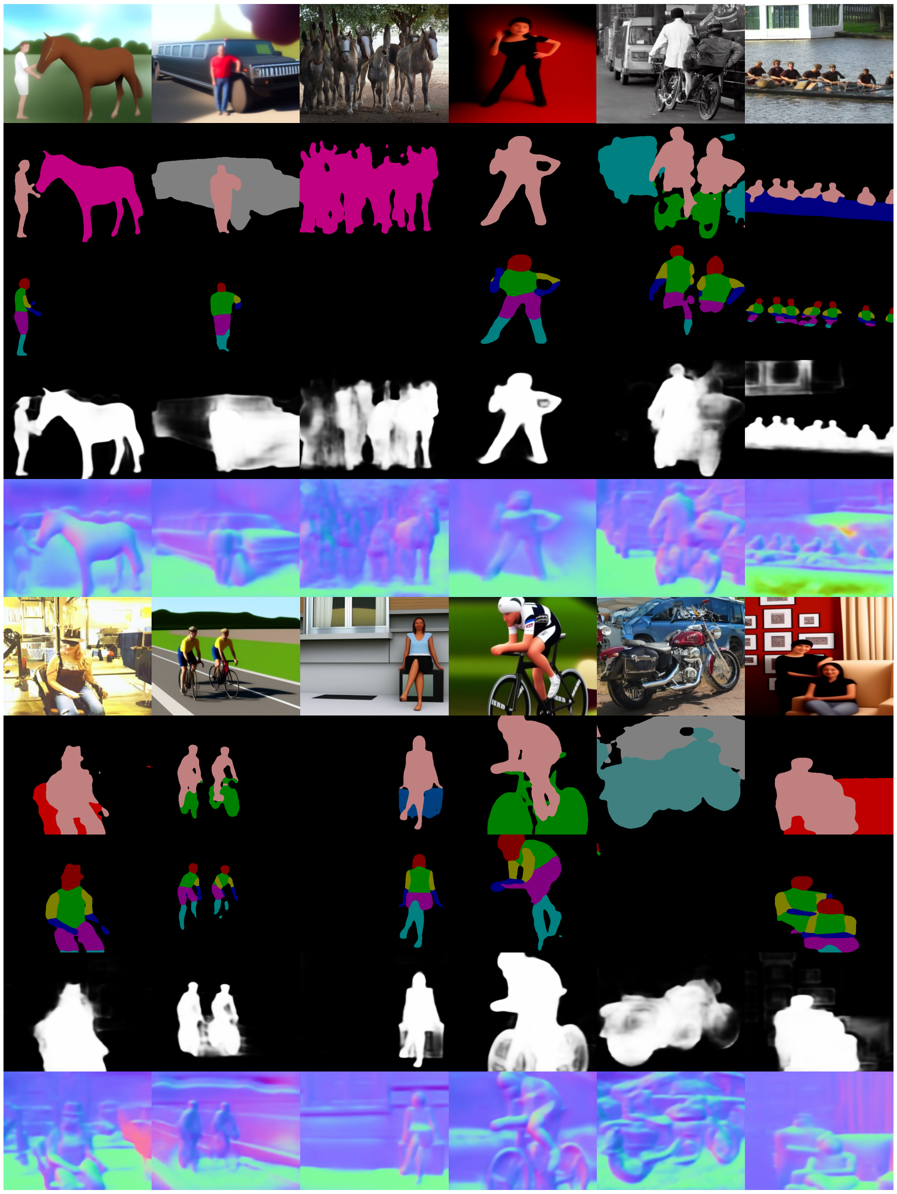


Figure 2.14: Synthetic samples from our method after the Diff-2-in-1 framework is trained on the multi-task setting of PASCAL-Context [149]. Each batch of samples contains five rows: RGB, semantic labels, human parsing labels, saliency map, and surface normal map *(from top to bottom)*. If the human parsing labels are all black, it means that there is no human in the generated image. The generated samples are of high quality with their multi-task annotations.

Chapter 3

Video Diffusion Models Learn the *Structure* of the *Dynamic* World

3.1 Introduction

Beyond generating high-fidelity images, diffusion models have achieved significant breakthroughs in visual perception. Their success is largely attributed to the large-scale vision-language pretaining, which allows them to capture detailed, object-centric features, and positions them as strong candidates for tasks such as image segmentation [254, 284] and classification [111]. Naturally, this raises a question: *Can diffusion models' success in images extend to the more complex domain of video understanding?*

Video understanding presents unique challenges absent in the image domain, particularly in capturing *temporal dynamics and motion patterns*. Unlike image diffusion models, video diffusion models [20, 230] are inherently designed to capture such spatial-temporal dynamics, making them far better suited for these tasks. As illustrated in Figure 3.1, where we visualize video representations using K-Means clustering and three-channel PCA for several widely used visual foundation models, video diffusion models excel at capturing motion dynamics – a critical capability that sets them apart from their image-based counterparts. Additionally, they retain a high-level structured representation of the visual world, further enhancing their implicit understanding of object relationships and environmental context. This dual capability of modeling both *motion* and *structure* makes them strong candidates for video understanding tasks.

To further investigate the effectiveness of video diffusion models in video understanding, we introduce a unified probing framework to systematically analyze feature representations from diffusion models across a range of video understanding tasks. This framework enables a detailed examination of the relative strengths and limitations of video diffusion models, providing practical insights for their optimal use. To ensure a comprehensive analysis, our evaluation spans seven models, including both image- and video-based architectures, as well as non-diffusion [15, 159, 219] and diffusion-based approaches. In the diffusion category, we further evaluate both UNet-based [20, 185, 230] and diffusion-transformer-based techniques [50, 163, 288].

Our study focuses on four key tasks that highlight different aspects of video understanding: (1) *action recognition*, a supervised classification task for assessing global video-level representations; (2) *object discovery*, an unsupervised segmentation task measuring dense feature quality; (3) *scene understanding*, a supervised task to test the semantic and geometrical awareness; and (4) *label propagation*, a training-free task evaluating the temporal consistency of features. These tasks provide a comprehensive examination of the strengths and weaknesses of each model across various facets of video understanding.

Key insights from our study include:

- Video diffusion models excel at capturing motion dynamics while maintaining a high-level understanding of the structure of the visual world, which supports their consistently strong performance.
- These models encode different information at various layers: early layers focus on abstract, high-level features, while later layers capture finer details. Fine-tuning only the most relevant layers enhances adaptation efficiency with minimal performance loss.
- Surprisingly, greater generative capacity does not always improve performance in visual perception tasks—earlier model versions sometimes outperform newer ones in downstream applications.

Overall, video diffusion models show significant promise for video understanding,

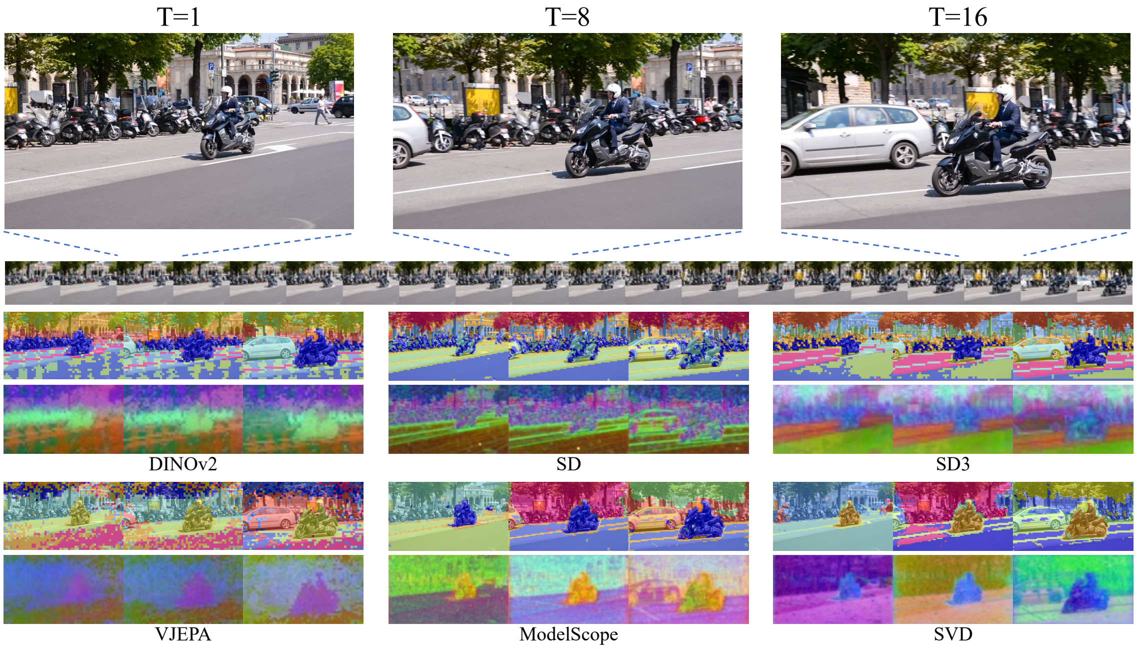


Figure 3.1: Video feature visualizations on DAVIS17 [170] dataset. Row 1: K-Means clusters (K=10); Row 2: three-channel PCA visualizations. Compared to image diffusion, or discriminatively trained models, video diffusion models excel at capturing motion dynamics while retaining a higher-level structured representation of the video input. These unique characteristics position them as strong candidates for video understanding.

excelling at capturing the dynamic structure of the visual world and emerging as competitive solutions in this field.

3.2 Related Work

Diffusion Models. Inspired by principles of heat and anisotropic diffusion, diffusion models have emerged as a powerful class of generative models for image and video synthesis [165, 238]. Recent advancements have positioned diffusion models as state-of-the-art across unconditional [22, 46, 79, 203, 204] and conditional image synthesis tasks [62, 78, 155, 179, 185, 190, 231, 268, 279]. Notably, Denoising Diffusion Probabilistic Models (DDPMs) [79] introduced the use of neural networks for modeling the denoising process, optimizing with a weighted variational bound. The Denoising Diffusion Implicit Model (DDIM) [79] enhanced this by incorporating a non-Markov sampling strategy to accelerate inference. Stable Diffusion [185] extended the diffusion-denoising process into the latent space of a pre-trained autoencoder [101], enabling more efficient large-scale model training. More recently, Transformer-based models have been introduced to further scale up training, achieving superior performance [50, 163].

The extension of diffusion models from image to video generation [73, 81, 138] gains remarkable achievements, encompassing both text-to-video (T2V)[21, 90, 98, 175, 232] and image-to-video (I2V) generation[65, 154, 227, 281].These efforts largely build upon pre-trained image-level diffusion models, such as Stable Diffusion [185], by training the additional video backbone with extra video data [20, 32, 33, 59, 66, 80, 230]. Some approaches avoid retraining entirely by utilizing training-free algorithms for video generation from image models [200, 243, 271]. Most recently, Sora [24] and its open-sourced couterparts [107, 288] demonstrated leading video generation capabilities with the more advanced architecture of diffusion transformer [163]. Among them, ModelscopeT2V [230], Stable Video Diffusion (SVD) [20], and OpenSora [281] have open-sourced their large-scale pre-trained model which serves as our backbones for this study.

Diffusion Models for Visual Perception. Diffusion models have also demonstrated strong semantic correspondence in their feature spaces [76, 213, 278]. This has spurred a line of research that utilizes diffusion models for visual perceptual tasks, through either training diffusion-based models for specific tasks such as segmentation [160, 254, 284], depth estimation [63, 191, 192] or open-world novel view synthesis [122]. Other work leverages pre-trained *frozen* diffusion models for perceptual learning [75, 99, 136, 152, 213, 278], or explores their use in data augmentation for discriminative tasks [26, 51, 144, 221].

Among them, DIFT [213] proposes a general pipeline to extract features from real images with diffusion models, which we adopt in our evaluation pipeline. Other awesome works [35, 150] leverage diffusion models for video-related tasks, but they *do not* leverage a video diffusion model with spatial-temporal reasoning modules. GenRec [239] proposes a joint optimization for video generation and recognition to better facilitate the learning of each other. VD-IT [296] and REM [6] leverage video diffusion models specifically for referring object segmentation. Lexicon3D [140] conducted a comprehensive study of visual foundation models, including diffusion-

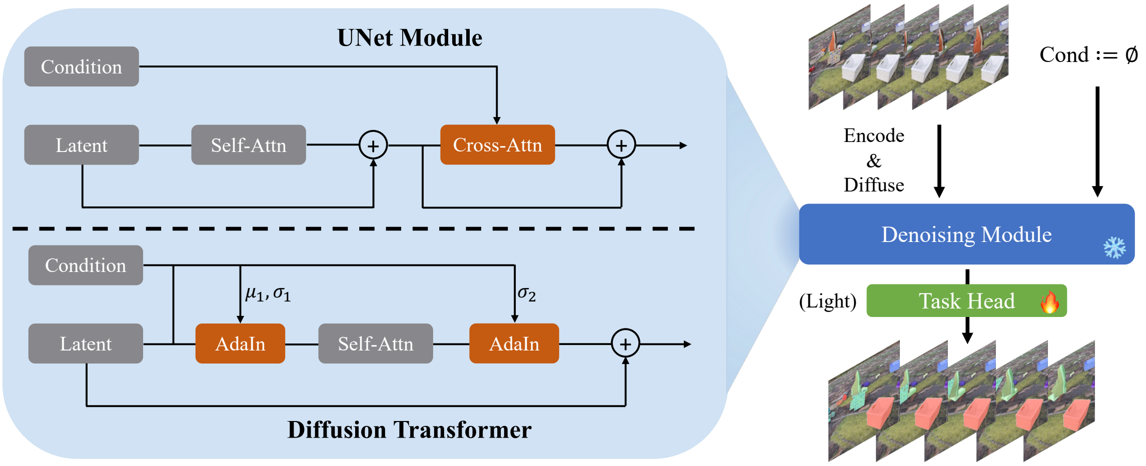


Figure 3.2: The architecture of our probing framework for video understanding using diffusion models. Video feature representations are extracted from the denoising module, followed by a lightweight task head to produce task-specific annotations. The process of feature extraction from UNet or DiT models (SD3 [50]) is illustrated on the left. Notice that we ignore the timestep input for simplification.

based ones, on 3D scene understanding. Unlike previous work, this study addresses the general video understanding with diffusion models across multiple tasks, each with a distinct focus.

3.3 Probing Video Understanding with Diffusion Models

3.3.1 Preliminaries

Latent Diffusion Models. Diffusion models [79] are latent variable models that learn the data distribution with the inverse of a Markov noise process. Latent diffusion models (LDM) [185] further switch the diffusion-denoising mechanism from RGB space to latent space, which improves the scalability and enables large-scale training. Concretely, an encoder $\mathcal{E}$ is trained to map a given image $x \in \mathcal{X}$ into a spatial latent code $z = \mathcal{E}(x)$. A decoder $\mathcal{D}$ is then tasked with reconstructing the input image such that $\mathcal{D}(\mathcal{E}(x)) \approx x$.

Considering the clean latent $z_0 \sim q(z_0)$, where $q(z_0)$ is the posterior distribution of z_0, LDM gradually adds Gaussian noise to z_0 in the *diffusion process*:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I}), \tag{3.1}$$

The denoising process takes inverse operations from the diffusion. The denoised latent at timestep $t-1$ is estimated via:

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t,t), \mathbf{\Sigma}_\theta(z_t,t)), \tag{3.2}$$

where the parameters $\mu_\theta(z_t,t), \mathbf{\Sigma}_\theta(z_t,t)$ of the Gaussian distribution are learned by the denoising network Σ_θ. As shown in [79], $\mathbf{\Sigma}_\theta(z_t,t)$ has only a marginal effect on the results, therefore estimating $\mu_\theta(z_t,t)$ becomes the main objective. A reparameterization is introduced to estimate it:

$$\mu_\theta(z_t,t) = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t,t)\right), \tag{3.3}$$

where $\epsilon_\theta(z_t,t)$ is typically a denoising UNet module [186] or diffusion transformer [163] module. $\epsilon_\theta(z_t,t)$ is usually conditioned on additional inputs, such as texts or image embeddings, to steer the denoising trajectory. In Figure 3.2 (left), we demonstrate how the extra modality is fused to the latent space: for UNet-based models, cross-attention modules are utilized to fuse the features while for DiT-based models, the additional embedding is fused via AdaIn [85] modules together with the broadcasted self-attention. The final objective of latent diffusion models is:

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t,t\right)\|_2^2\right]. \tag{3.4}$$

Video Diffusion Models. Video diffusion models generally share a similar architecture to the 2D diffusion models. Given a video $\mathbf{v} = [x^1, x^2, \cdots, x^N]$, a spatial encoder $\mathcal{E}^v$ is applied to each frame to map them to the latent code $z^i = \mathcal{E}^v(x^i)$, where i is the frame index. We use the notation $\mathbf{z} = [z^1, z^2, \cdots, z^N]$ for convenience. For the decoder, usually, a spatio-temporal decoder is applied to enforce the temporal consistency $\mathcal{D}^v(\mathbf{z}) \approx \mathbf{v}$.

One crucial distinction for video diffusion models is that they explicitly model spatio-temporal information with the denoising network, denoted as ϵ_θ^v. This network is extended to 3D by either introducing additional temporal attention modules [20, 225], or replacing the spatial attention modules with spatio-temporal ones [230, 281].

3.3.2 Video Understanding Probing Framework

Figure 3.2 illustrates our unified probing framework. We extract video representations from the denoising module and subsequently apply a lightweight task-specific head for various tasks.

Diffusion Features

We extract video features with diffusion models following DIFT [213]. The process begins by adding noise at timestep T to the real video latent (Equation 3.1), moving it into the $\mathbf{z}_T$ distribution. This noisy video latent, along with T, is then passed to ϵ_θ^v. Instead of using the final output of ϵ_θ^v, which predicts the noise, we extract features from intermediate layer activations that effectively capture the video's underlying representations:

$$\mathbf{z}_{\text{feature}} = \epsilon_\theta^{v(n)}(\mathbf{z}_T, T), \tag{3.5}$$

where (n) indicates the block index. Following DIFT, we extract the intermediate representations from upsampling blocks, forming the diffusion features. For features from image diffusion models, we follow a nearly identical process, except that we process the videos frame by frame. Additionally, during feature extraction, we introduce a fixed "null-embedding" as the condition for ϵ_θ^v. For language-based models, this embedding is obtained by passing an empty prompt to the text encoder. For image-based models, we use an all-zero conditional image.

Adaptation for Downstream Tasks

After extracting features from diffusion models, we use a lightweight task head (fewer than 1% of the backbone's parameters) to adapt these features for the target tasks, as demonstrated by the object discovery task in Figure 3.2. We detail the specific task heads for our evaluated tasks below.

Action Recognition is the task that aims to predict an action label for a given video. Following previous practice [15, 219], we take the averaged feature map and

Model	Type	Architecture	Dataset	Feature Dim	Downsample
DINOv2 [159]	Image	ViT-L	LVD-142M	1024	14
VideoMAE [219]	Video	ViT-L	Kinetics400-240k	1024	16
VJEPA [15]	Video	ViT-L	VideoMix2M	1024	16
SD [185]	Image	UNet	LAION-5B	1280/640	8/16
SD3 [50]	Image	DiT	PublicImgs-1B	1536	16
ModelScope [230]	Video	UNet	WebVid-10M	1280/640	8/16
SVD [20]	Video	UNet	LVD-152M	1280/640	8/16
Open-Sora [288]	Video	DiT	Mix-210M	288	8

Table 3.1: Details of the pretrained visual foundation models we used for our video understanding evaluation.

apply a two-layer MLP, where the hidden dimension is the same as the input features, to predict the final label.

Object Discovery identifies and tracks dynamic objects from videos in a self-supervised manner. We adopt the architecture from MoTok [12] where cross-attention layers with learnable queries, called slots [128], are trained to group foreground regions in video with feature-level reconstruction as the learning signal.

Scene Understanding aims to predict pixel-wise scene properties, *e.g.*semantic labels and depth values, for the given video. Following DINOv2 [159], we directly apply a two-layer MLP on top of the feature map and interpolate them to the original resolution to predict the labels.

Label Propagation is a training-free task where instance masks or keypoints from an initial frame are propagated to each subsequent frame in a video. Rather than predicting new labels, label propagation transfers the initial labels frame-by-frame, leveraging the continuity of appearance across frames. As in prior methods [89, 213], we achieve this by using a k-nearest neighbors (k-NN) search across a feature queue containing the initial frame and the most recent m frames, thus no specialized task head is required.

Backbone	UCF101		HMDB51		MOVi-C		MOVi-E		CityScape		DAVIS17			JHMDB	
	Top1 Acc	Top 5 Acc	Top1 Acc	Top 5 Acc	FG.ARI	mBO	FG.ARI	mBO	mIoU(SS)	mErr(DE)	$\mathcal{J}_m$	$\mathcal{F}_m$	$\mathcal{J}\&F_m$	PCK@0.1	PCK@0.2
DINOv2	89.8	97.8	61.6	89.6	55.6	29.2	71.9	26.3	53.6	4.30	64.8	69.1	67.0	50.42	78.71
VideoMAE	87.9	97.9	55.4	83.4	24.5	14.3	32.7	14.1	37.8	5.73	30.5	37.5	34.0	32.51	59.30
VJEPA	92.1	98.5	66.5	92.3	31.8	18.6	49.9	18.0	41.3	5.27	52.3	58.0	55.1	37.55	70.31
SD	63.5	86.1	33.0	68.1	40.6	24.8	63.4	26.9	44.5	4.97	67.8	74.6	71.2	60.48	80.77
SD3	60.9	85.8	32.4	62.1	43.3	26.3	65.1	28.6	46.0	5.09	48.5	54.8	51.6	38.17	65.89
ModelScope	80.6	94.9	50.7	80.2	41.3	25.1	63.7	27.5	49.3	3.98	65.3	72.4	68.4	60.90	82.83
SVD	92.3	98.6	63.8	89.7	44.2	26.7	65.4	29.4	48.1	4.68	59.8	67.7	63.8	60.52	81.84
Open-Sora	47.3	75.9	22.1	54.8	-	-	-	-	-	-	-	-	-	-	-

Table 3.2: Quantitative evaluations on the four evaluated tasks. The top two results are marked in green and yellow respectively. Video diffusion models provide semantic- and geometric-aware representations that contain both high-level abstractions and detailed information, positioning them as unique and competitive candidates for video understanding.

3.4 Experimental Evaluations

3.4.1 Evaluation Settings

Baseline Models. We perform our video understanding analysis with seven visual foundation models. **DINOv2** [159] is a contrastive learning-based image-level foundation model. **VJEPA** [15] and **VideoMAE** [219] learn comprehensive video representations by reconstructing from masked video patches. **Stable Diffusion (SD)**[185] and **Stable Diffusion 3 (SD3)**[50] are text-to-image diffusion model with UNet [186] and DiT [163] as denoising backbones. **ModelScopeT2V** [230] and **Stable Video Diffusion (SVD)** are video diffusion models that take SD as the initialization and further fine-tune on large-scale video data. Additionally, we include the DiT-based video diffusion model, Open-Sora [288], in the action recognition evaluation but exclude it from other tasks due to its inability to produce precise patch-wise representations. Detailed configurations of these feature extractors are provided in Table 3.1.

Datasets and Metrics. We evaluate *action recognition* recognition with top 1 and top 5 accuracy on UCF101 [205] and HMDB51 [105]. We study the object discovery task on MOVi-C and MOVi-E [61], and take foreground adjust random index (FG. ARI) and video mean best overlap (mBO) as metrics. We evaluate the scene understanding task with semantic segmentation and depth estimation on CityScapes [41], and take mean interaction over unions (mIoU) and mean L_2 error (mErr), for the two tasks respectively. We conduct the label propagation for video object segmentation on DAVIS17 [170] and keypoint estimation on JHMDB [91] following the same setup as DIFT [213]. We report region-based similarity $\mathcal{J}$ and contour-based accuracy $\mathcal{F}$ [164] for DAVIS17, and percentage of correct keypoints (PCK) for JHMDB.

Key Implementation Details. We use the noise level 50 by default, with a corresponding timestep T=50 (for SD, ModelScope, and SVD) or T=16 (for SD3 and Open-Sora). For the layer index, we design the use of block index 1 (for SD, ModelScope, and SVD) and layer index 12 (for SD3 and Open-Sora) for action recognition. For the other tasks, we use block index 2 and layer index 24 respectively. We use batch size 12 with 4 NVIDIA-A100 GPUs running in parallel for all the backbones except ModelScope. We use batch size 6 with 8 GPUs in parallel for ModelScipe to fit its CUDA requirement.

More details about datasets, model implementation, and training configurations are included in the supplementary material.

3.4.2 Main Results

We show the quantitative results for our main evaluation of the four tasks in Table 3.2, and representative visual comparisons in Figure 3.3. More visualizations are included in the supplementary material.

Comparisons between ModelScope and SVD. For the following discussions, we treat ModelScope and SVD as variants of the same "video diffusion model" category, despite differences in their model type (Text-to-Video *v.s.* Image-to-Video), for which we use unconditional versions to minimize conditioning effects. While their performance varies across tasks – likely due to differences in training data and fine-tuning strategies (ModelScope fine-tunes only temporal modules, while SVD uses full fine-tuning) – these variations make it challenging to draw universal conclusions based on specific tasks. Given the lack of a standardized training strategy, we focus on their shared foundations instead: both models are based on SD with additional video training, which enables us to discuss their common strengths and limitations in video understanding.

Overall Conclusions. Across all four tasks, video diffusion models consistently rank among the top performers, highlighting their robustness and adaptability in video understanding. As illustrated by the visual comparisons in Figure 3.3, video diffusion models capture motion dynamics more effectively than image-based models and demonstrate a stronger understanding of world structure compared to conventional video foundation models. This balance of dynamic and structural comprehension enables them to consistently perform at a high level.

Action Recognition. Surprisingly, SVD achieves the highest performance on UCF101 and ranks second on HMDB51, consistently outperforming both image diffusion models and the conventional DINOv2 and VideoMAE encoders. This result highlights the ability of well-trained video diffusion models to capture global-level video representations effectively. However, Open-Sora and SD3, which use DiT architectures, exhibit suboptimal performance. A potential reason may lie in how DiT models fuse multi-modal features, suggesting an open research challenge: developing improved feature extraction techniques tailored for DiT-based diffusion models.

Object Discovery. Overall DINOv2 achieves the highest performance among all models, demonstrating its superior object-awareness. However, it is worth noticing that video diffusion models outperform in terms of mBO on the MOVi-E dataset which involves more complex ego and object motion. This suggests that diffusion models are particularly effective at identifying and tracking objects in challenging motion scenarios, making them especially useful for tasks requiring precise localization and tracking. Visual comparisons in Figure 3.3 provide further evidence where SVD precisely tracks objects with complicated motion.

Scene Understanding. ModelScope and DINOv2 emerge as the top performers in these tasks, with DINOv2 excelling in semantic understanding and ModelScope showing superior performance in depth estimation. For ModelScope, we hypothesize that its success stems from its ability to leverage motion information, which inherently aids in understanding depth.

Label Propagation. On DAVIS17, video diffusion models generally lag behind their image-based counterparts. We hypothesize that this is because video diffusion models learn detailed representations of moving objects (refer to Figure 3.1) but struggle to differentiate static objects from the background, a key challenge in video object segmentation (VOS). In contrast, on the JHMDB dataset, where pose estimation focuses solely on a single moving object, video diffusion models demonstrate their strengths.

In summary, video diffusion models provide semantic- and geometric-aware representations that contain both high-level abstractions and detailed information, positioning them as unique and competitive candidates for video understanding.

3.4.3 Guidelines for Video Diffusion Adoption

Optimal Use of Video Diffusion Models

In our main evaluation, we use frozen video diffusion representations with fixed noise levels and layer indices. In this section, we investigate how to better adapt these representations for video understanding by providing guidance on layer selection and fine-tuning strategies.

Noise Levels and Block Indices. We examine the effects of noise level selection and block indices in SVD for action recognition on HMDB51 and label propagation on DAVIS17, as summarized in Table 3.3. The results suggest that noise level plays a relatively smaller and task-specific role compared to block indices. Generally, a small amount of noise (*e.g.*, corresponding to $T = 50$) yields strong results. In contrast, block indices significantly influence downstream task performance: features from earlier blocks encode abstract, high-level information, making them ideal for classification tasks, while features from later blocks capture finer details, benefiting dense prediction tasks. These findings are consistent with observations from image diffusion models, as reported by DIFT [213].

Fine-Tuning Video Diffusion Models. For certain video understanding tasks, fine-tuning the backbone is essential and typically results in improved performance. To explore the impact and strategies of fine-tuning video diffusion models for perception tasks, we fine-tune the SVD denoising UNet on HMDB51 and MOVi-E. The results are summarized in Table 3.4 (first two rows). Notably, for object discovery, we slightly modify the baseline architecture [12], with details provided in the supplementary material. As a result, the reported FG.ARI score for the frozen model differs from that in Table 3.2.

Notably, by comparing the change of parameters of all the modules, we find that the last in-use upsampling block (*i.e.*block index 1 for action recognition and block

Noise Level	Block Index	HMDB51		DAVIS17		
		Top1 Acc	Top5 Acc	$\mathcal{J}_m$	$\mathcal{F}_m$	$\mathcal{J}\&F_m$
0	1	60.3	88.0	52.1	44.9	48.5
50	1	63.8	**89.7**	51.1	42.6	46.9
100	1	**63.9**	89.4	50.3	41.6	46.0
200	1	62.6	88.7	50.2	41.3	45.8
0	2	31.1	64.0	**60.8**	**68.0**	**64.4**
50	2	33.7	66.9	59.8	67.7	63.8
100	2	35.4	68.0	59.6	67.2	63.4
200	2	32.8	66.8	59.1	64.5	62.8

Table 3.3: Ablation on noise level selection and block index of SVD on HMDB51 and DAVIS17. Compared to noise level, the block index has a significant impact on downstream task performance. Features from earlier blocks capture more abstract, high-level information, while features from later blocks are more object-oriented.

index 2 for object discovery) exhibits the highest sensitivity to parameter changes, highlighting their critical role in enhancing task performance. Inspired by previous efficient diffusion fine-tuning approaches [13, 106, 189], we construct two fine-tuning variants: one incorporates LoRA [82] adaptation layers in all cross-attention blocks, while the other fine-tune only the most sensitive upsampling block. The results for these two variants are reported in Table 3.4 (last two rows). These findings demonstrate that efficient fine-tuning strategies can significantly enhance performance while keeping training costs reasonable, offering practical guidance for optimizing video diffusion models.

Generation V.S. Perception

In this section, we explore an intriguing question: *does a diffusion model with superior generation capacity inherently perform better in visual perception tasks?* While we could evaluate the generative capacity of different diffusion models directly, this approach is challenging due to their diverse conditioning mechanisms –some are text-conditioned, others image-conditioned – and their application across both image and video generation. Instead, we adopt an alternative strategy: comparing the

Strategy	HMDB51			MOVi-E		
	Top1 Acc	Mem.	Time	FG.ARI	Mem.	Time
Frozen	63.8	1.0 ×	1.0 ×	66.1	1.0 ×	1.0 ×
Full	68.3	2.6 ×	2.3 ×	69.2	2.7 ×	2.5 ×
LoRA	66.9	1.1 ×	1.7 ×	67.0	1.2 ×	1.7×
Sensitive	67.1	1.3 ×	1.8 ×	68.1	1.4 ×	1.9×

Table 3.4: Performance and training cost for finetuning SVD UNet. "Sensitive" denotes only fine-tuning the most sensitive UNet block (the last in-use upsampling block). While finetuning the diffusion backbone yields performance improvements, it comes with significantly higher computational costs. Using an efficient finetuning strategy by only tweaking the most sensitive layers leads to an effective.

performance of different checkpoints of the same model, under the assumption that later versions exhibit improved generative capacity.

We use SD and SVD as backbone models and evaluate four versions of SD (v1.4, 1.5, 2.0, and 2.1) alongside two versions of SVD (v1 and v1.1) across the four tasks, with results summarized in Figure 3.4. For SVD, the later checkpoint consistently improves performance across all tasks, aligning with its enhanced generative capacity. However, for SD, the 1-series models generally outperform the 2-series models, though the optimal version varies by task. This discrepancy may stem from differences in the scale and composition of training data across versions.

Overall, these results suggest that greater generative capacity does not necessarily translate to improved performance in visual perception tasks, indicating that there is no universal metric for selecting a representation exists as of yet.

3.4.4 Discussions

Inference Cost. We report the inference time and memory usage for a single batch of size $[6, 256, 256]$ on the MOVi-E dataset, using an NVIDIA A100 GPU in Table 3.5. The baseline model, DINOv2, has an inference time of 0.224 seconds and consumes 2.6 GB of GPU memory. Notably, the memory consumption for ModelScope is an outlier, due to the lack of optimization in its public implementation. In general, diffusion-based and video-based models require more computational resources, though these costs

Model	DINOv2	VideoMAE	VJEPA	SD	SD3	ModelScope	SVD
Memory	1.0×	1.7 ×	1.1×	1.8×	4.6×	8.3×	2.7×
Inf. time	1.0×	2.1 ×	1.7×	1.1×	3.3×	2.0×	2.1×

Table 3.5: Time and Memory Consumptions for all the compared models. Results are tested on the MOVi-E dataset with a single batch with dimensions $[6, 256, 256]$. Diffusion-based and video-based models require more computational resources but the costs remain acceptable.

remain acceptable. The exception is SD3, which employs a DiT-based architecture. This observation is consistent with our earlier conclusions and highlights the need to develop more efficient and effective feature extraction methods for DiT-based models.

Limitations of Video Diffusion Representations. We show two typical limitations for video diffusion representations on label propagation in Figure 3.5: difficulty in handling occlusion among instances of the same semantic category, and challenges with distinguishing nearby objects that share similar motion.

3.5 Conclusion and Future Work

This paper showcases that video diffusion models offer a powerful approach to video understanding, excelling in capturing motion dynamics and high-level structural representations. By systematically analyzing their performance across multiple tasks, we highlight their robustness, adaptability, and the distinct advantages they bring to video perception. These models stand out for their unique balance of dynamic and structural comprehension, positioning them as promising tools for advancing video understanding. Moreover, our findings provide actionable insights into how their representations can be optimized through careful layer selection and fine-tuning strategies, paving the way for more efficient and effective utilization of video diffusion models in various applications.

Two feasible **future work** of this study include: (1) designing a more advanced feature extraction pipeline with newly introduced DiT-based models. (2) Exploring other ways of leveraging video diffusion models beyond merely using them as encoders [6, 239].

Social Impact. By pushing the boundaries of what is possible with video diffusion models, the findings in this paper can further inspire future explorations with video diffusion models in both generative and video analysis aspects.

3.6 Implementation Details

3.6.1 Backbone Model Implementations

We implement our backbone models including both diffusion-based ones and discriminative ones with public implementations of DIFT[1] [213], ModelScopeT2V[2] [230], Stable Video Diffusion[3] [20], Stable Diffusion 3[4] [50], DINOv2[5] [159], and VJEPA[6] [15]. We will release our code for reproduction upon publication.

3.6.2 Action Recognition

Datasets. We evaluate action recognition with top 1 and top 5 accuracy on two widely used datasets, UCF101 [205] and HMDB51 [105]. Both datasets are relatively small-scaled datasets containing 101 and 51 action categories respectively. UCF101 and HMDB51 contain around 9.5k/3.5k train/val videos and 3.5k/1.5k train/val videos, respectively. We center-cropped the video frames to 224 × 224 for evaluation and uniformly sampled 16 frames for each video for training. We build all the compared models by applying a dense layer on the averaged representation for all the video tokens to make the final prediction. Following previous works [15, 219], we report the averaged results among the 3 test splits for both datasets.

Training details. We remove the data augmentation for training. Otherwise, we follow the same training procedure as VideoMAE [219]. We train all the models for 100 epochs for UCF101 and 50 epochs for HMDB51. We use AdamW for optimization with a maximum learning rate of 5e-4. We use the same learning rate scheduler as the object discovery task. It takes about 1 day to train our model for UCF101 and 6 hours for HMDB51 with NVIDIA A-100 GPUS.

[1] https://github.com/Tsingularity/dift

[2] https://github.com/ali-vilab/VGen

[3] https://github.com/Stability-AI/generative-models

[4] https://huggingface.co/stabilityai/stable-diffusion-3-medium-diffusers

[5] https://github.com/facebookresearch/dinov2

[6] https://github.com/facebookresearch/jepa

3.6.3 Object Discovery

Datasets. We evaluate the object discovery task on two widely-used photo-realistic synthetic datasets, MOVi-C and MOVi-E [61]. Both datasets feature multiple objects exhibiting rigid motion. MOVi-C solely focuses on moving objects without camera movement, whereas MOVi-E includes a mix of moving and static objects, complemented by linear camera motion. Both datasets contain 9,750 videos for training and 250 videos for validation. Each video contains 24 frames under the resolution of 256 × 256. We use the original resolution as the input but the mask evaluations are conducted under the resolution of 32 × 32, which is consistent with previous object discovery models [3, 12, 102, 201, 272]. We take video foreground adjust random index (FG. ARI) and video mean best overlap (mBO) as metrics.

Baselines. We build our main model and the additional three variants with different pre-trained feature extraction backbones upon the public implementation of MoTok[7] [12]. We replace their Resnet [70] feature extractor backbone with the other pre-trained feature extractors. For VJEPA, we repeat each frame twice to match the shape.

Training details. We build all the baseline models with 15 slots following the setting of VideoSAUR [272]. We train all the baseline models for 500 epochs with a batch size of 48. We use AdamW [130] for optimization with a gradient clip with norm 0.1. We apply a Cosine annealing learning rate scheduler with the largest learning rate as 5e-5 and warm-up steps as 3000. It takes about 3 days to train our model with NVIDIA A-100 GPUS.

3.6.4 Scene Understanding

Datasets. We conduct scene understanding on CityScape [41] datasets. We select video semantic segmentation and monocular depth estimation as the target task. CityScape dataset contains 5,000 labeled frames with a train, val, and test split. However, the labels for the test set are not of the same quality as the other two. Therefore, we evaluate on the val set, following previous work [137]. The original resolution of CityScape is 1024 × 2048, we downsample them to 256 × 512 to run the

[7] https://github.com/zpbao/MoTok

evaluation. We train each model with a video clip of 16 frames.

Training details. We train each model for 100 epochs with a batch size of 24. We use AdamW [130] for optimization with a gradient clip with norm 0.1. We apply a Cosine annealing learning rate scheduler with the largest learning rate as 5e-5 and warm-up steps as 3000. It takes about 1 day to train our model with NVIDIA A-100 GPUS. We did not include any data augmentation in our training. We randomly select a video clip that contain the labeled one during training, while in inference, we start from a fixed frame where the labeled one is in the middle.

3.6.5 Label Propagation

Datasets. We conduct the label propagation for video object segmentation on DAVIS17 [170] and keypoint estimation on JHMDB [91] following the same setup as DIFT [213]. DAVIS17 is a multi-object segmentation dataset with unfixed lengths from around 40 frames to 110 frames. We evaluate our model on the resolution of 512 × 896 by resizing the original 480p frames. JHMDB is a keypoint estimation dataset. We follow the implementation of CRW [89], we resize each video frame's smaller side to 320 and keep the original aspect ratio. We report region-based similarity $\mathcal{J}$ and contour-based accuracy $\mathcal{F}$ [164] for DAVIS17, and percentage of correct keypoints (PCK) for JHMDB.

Hyperparameters. For use the same evaluation pipeline as DIFT [213]. The hyperparameters are listed in Table 3.6. We cite the results for all the other methods from DIFT [213].

Dataset	Time step t	Block index n	Temperature for softmax	Propagation radius	k for top-k	Number of prev. frames
DAVIS-2017	25	2	0.1	10	15	28
JHMDB	25	2	0.1	5	15	14

Table 3.6: Hyperparatemers for the label propagation tasks.

3.7 Additional Visualizations

We show additional visualizations with backbone SVD on object discovery, and ModelScope on label propagation on Figs. 3.6, 3.7, respectively, showing the promising results of video diffusion features for video understanding.

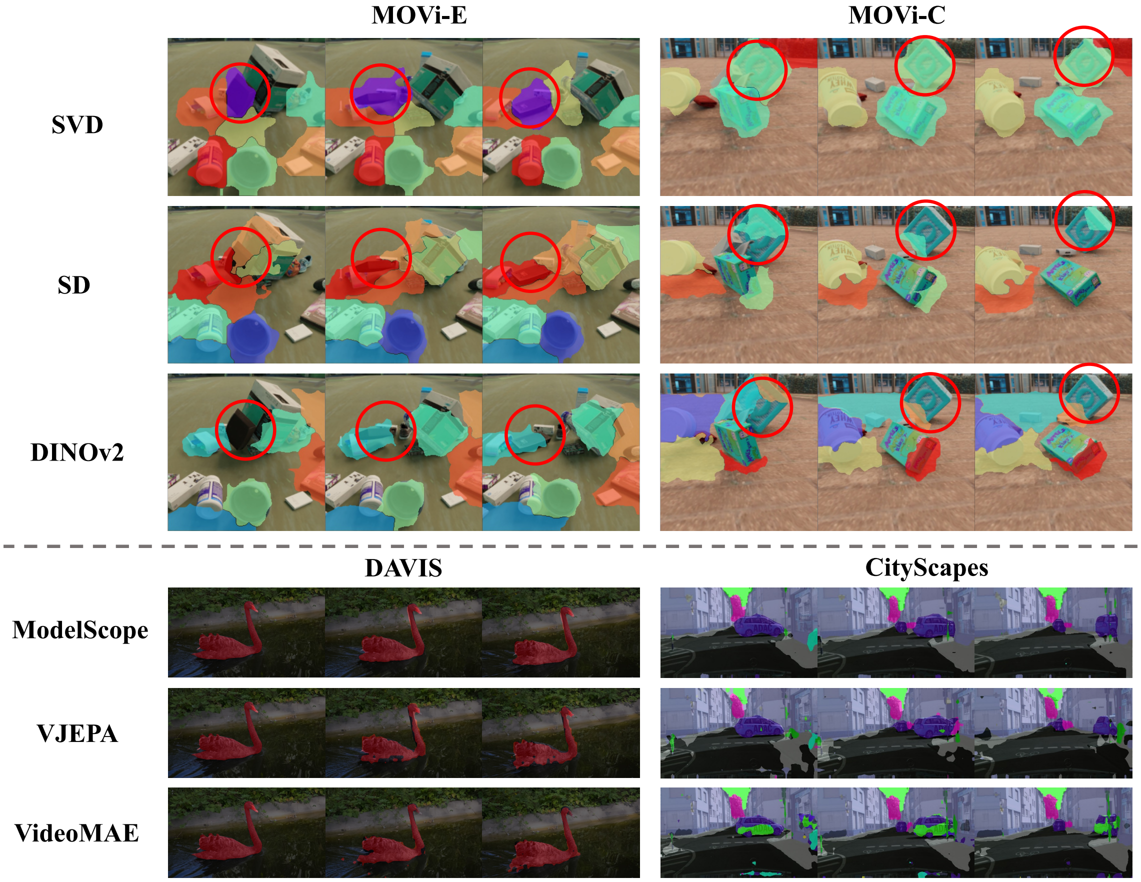


Figure 3.3: Representative visual comparisons between the results of video diffusion models and other foundation models. **Top:** Video diffusion models capture motion dynamics more effectively than image-based models; **Bottom:** Video diffusion models demonstrate a stronger understanding of world structure compared to conventional video foundation models. This balance of dynamic and structural comprehension enables them to consistently perform at a high level.

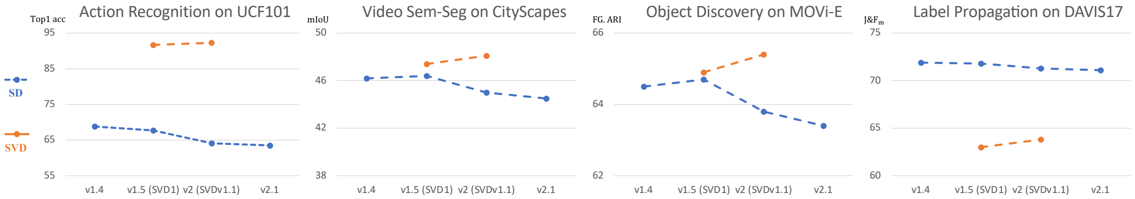


Figure 3.4: Comparison between generation ability and downstream task performance on SD and SVD series. The later SVD checkpoint consistently improves performance across all tasks while the 1- series SD models generally outperform the 2- series models. These results indicate that greater generative capacity does not necessarily translate to improved performance in visual perception tasks.

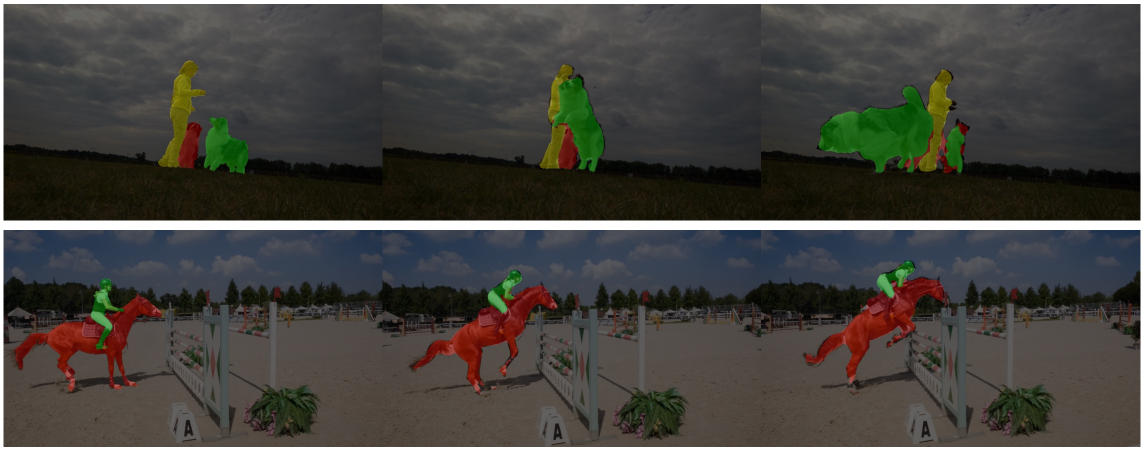


Figure 3.5: Limitations of Video Diffusion Representations: difficulty in handling occlusion among instances of the same semantic category, and challenges with distinguishing nearby objects that share similar motion.

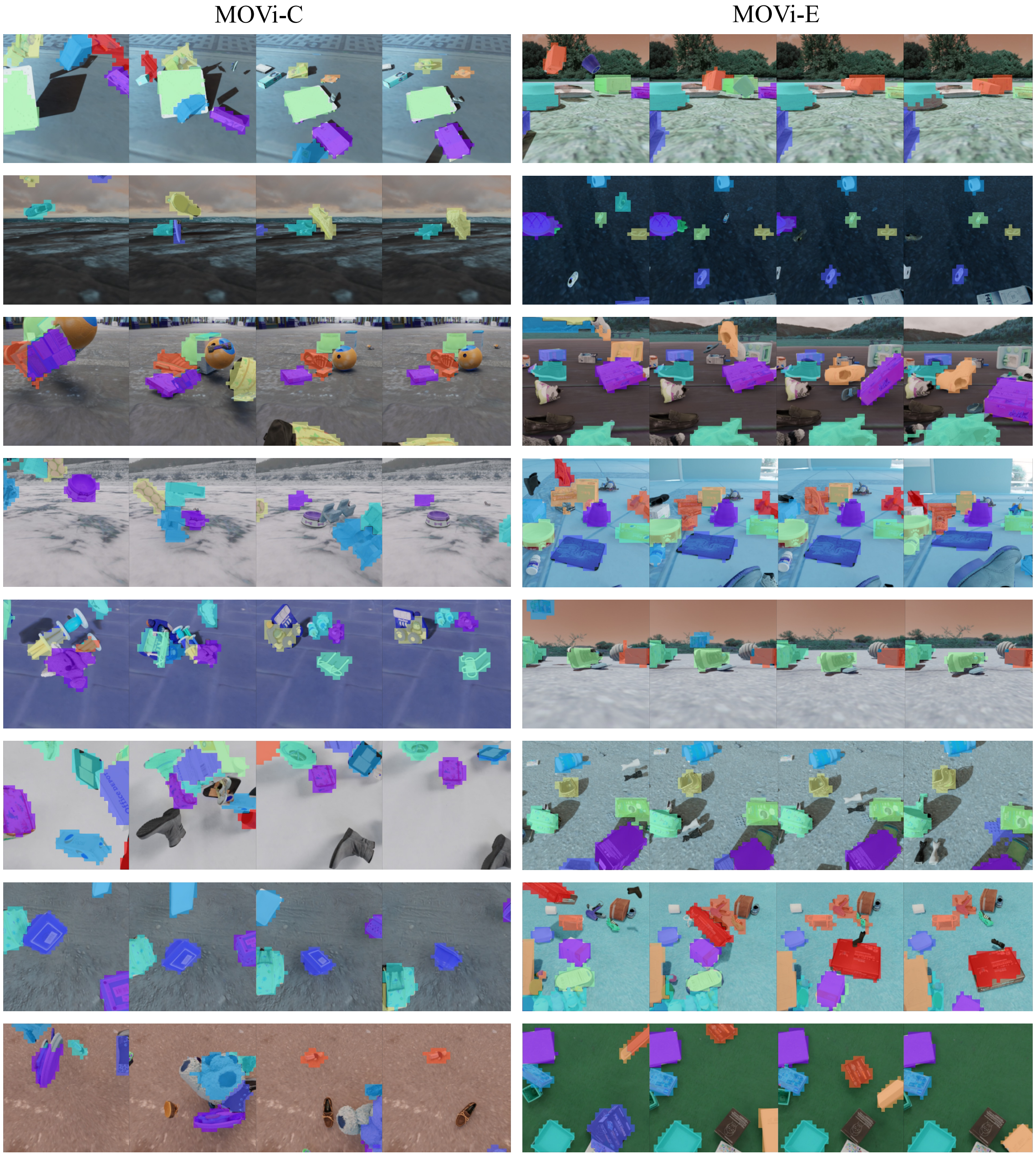


Figure 3.6: Additional visualizations for object discovery on MOVi-C and MOVi-E datasets [61] with SVD [20] as backbone. We show the Top 10 object masks for each method and ignore the background masks for better visualizations. Our model achieves promising results for object discovery tasks.

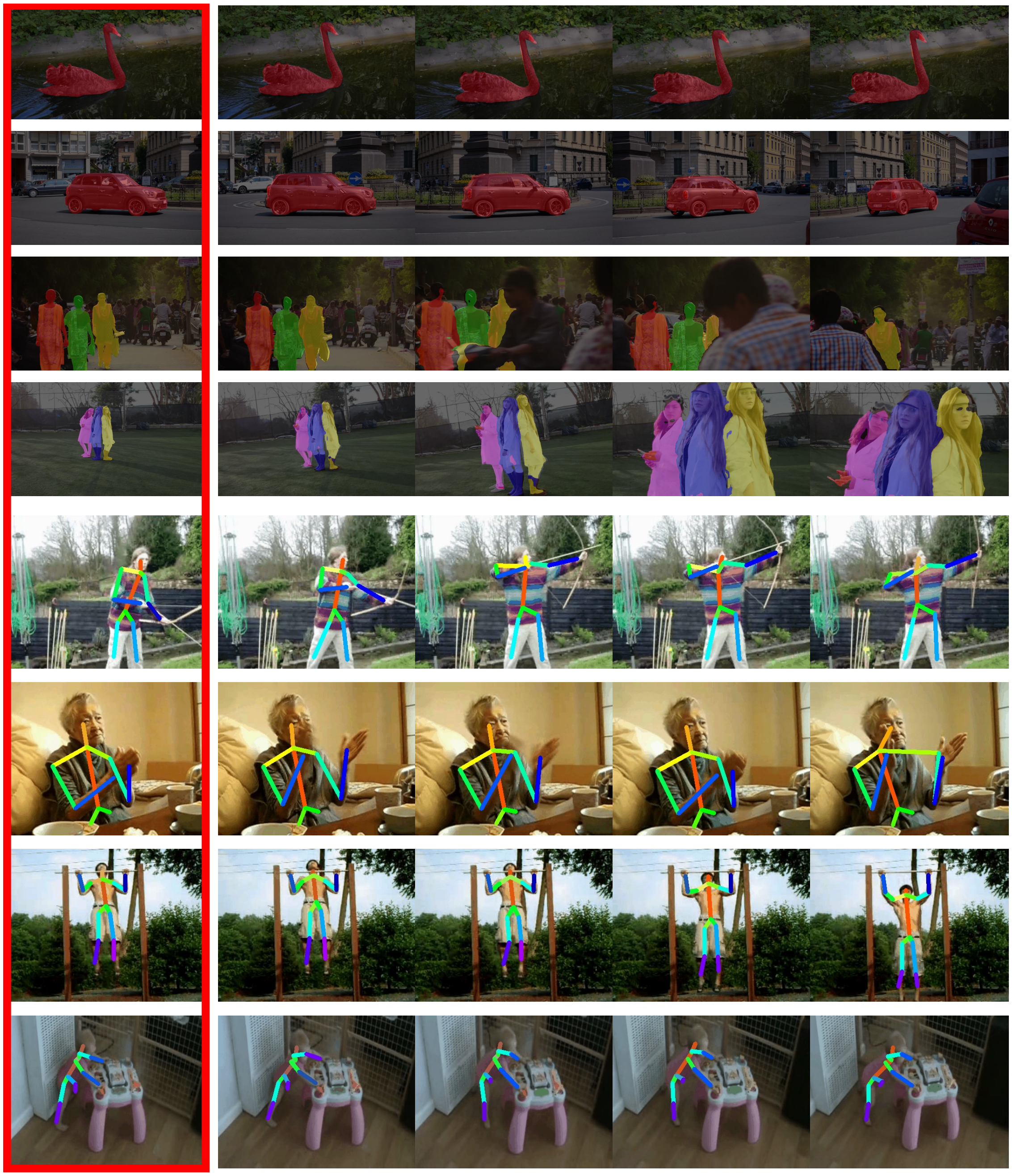


Figure 3.7: Additional visualizations for label propagation tasks with ModelScope [230] as the backbone. Top 4 rows: video object segmentation task on DAVIS17 [170]; Bottom 4 rows: keypoint estimation on JHMDB [91].

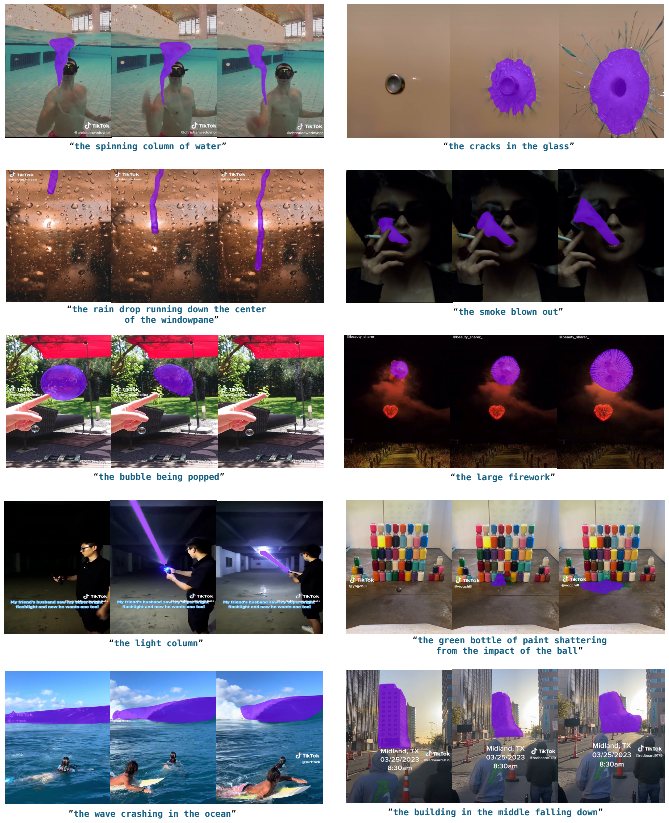
Chapter 4

ReferEverything: Towards Segmenting Everything We Can Speak of in Videos

4.1 Introduction

One of the most remarkable features of natural language is its ability to describe human visual experience in all of its richness and complexity. Whether capturing fleeting moments, like raindrops rolling down the window, or smoke dissipating from a cigarette (see row 2 in Figure 4.1), or describing dynamic processes, such as a glass shattering or a whirlpool forming in the water (row 1 in Figure 4.1), if we can utter them, we can also accurately localize them in space and time. This universal mapping between the discrete, symbolic realm of language and the continuous, ever-changing visual world is developed through a lifetime of visual-linguistic interaction [16, 172].

The corresponding problem in computer vision - Referral Video Segmentation (RVS) [56, 84], is defined as the task of segmenting a specific region in a video based on a natural language description. However, virtually all existing benchmarks and methods focus on a specific subset of RVS - Referral Video Object Segmentation (RVOS) [195, 244], where the goal is to track and segment the *object* referenced by a given expression. Why has the field concentrated so narrowly on this task?


"the spinning column of water"

"the cracks in the glass"

"the rain drop running down the center of the windowpane"

"the smoke blown out"

"the bubble being popped"

"the large firework"

"the light column"

"the green bottle of paint shattering from the impact of the ball"

"the wave crashing in the ocean"

"the building in the middle falling down"

Figure 4.1: We present REM, a framework for segmenting a wide range of concepts in video that can be described through natural language by capitalizing on powerful visual-language representations learned by video diffusion models. REM generalizes with ease to challenging, dynamic concepts, such as raindrops or shattering glass, shown above. Video visualizations are available here.

Although multiple factors contribute, we argue that the primary reason lies in the data. Historically, RVOS datasets have been developed by adding referral expression annotations to existing object tracking benchmarks [171, 256], which are inherently object-centric and limited in scale.

At the same time, recent advances in Internet-scale datasets with billions of paired image- and video-language samples [9, 194] have opened new possibilities. These datasets have been used to train powerful denoising diffusion models [185, 230], and provide excellent representations of the natural visual-language space. In the image domain, numerous studies have shown that re-purposing diffusion models can yield highly generalizable representations of object shapes [160, 284]. Very recently, Zhu *et al.* [296] explored the application of video diffusion models for referral segmentation, but their approach exhibited limited generalization capabilities.

In this work, we introduce a novel approach to RVS that leverages large-scale video-language representations learned by diffusion models. Our method, shown in Figure 4.3, enables spatio-temporal localization of a wide range of concepts in video that can be described through natural language, beyond conventional object tracking. A key factor behind our approach's success is preserving the rich representations learned by the generative model (see Figure 4.2). To this end, we retain the original model architecture and fine-tune it on existing referral image- and video-segmentation datasets, adjusting the output to generate object masks instead of Gaussian noise. As shown in Section 4.5.1, our model demonstrates competitive performance against state-of-the-art models on RVOS benchmarks. More significantly, it exhibits a much stronger generalization.

To quantify this effect, we report results on the open-world object tracking benchmark - BURST [2], and the non-object 'Stuff' segmentation dataset - VSPW [146], as well as collect a new benchmark that focuses on dynamic process in Section 4.4. We define the latter as temporally evolving *events*, where the subjects undergo continuous changes in state, shape, or appearance (see examples in Figure 4.1). Our new benchmark, which we call Ref-VPS for Referral Video Process Segmentation, consists of 141 videos that are labeled with referral expressions and masks at 24 fps and span 39 unique concepts. Our experiments in Section 4.5.2 demonstrate that existing approaches fail to generalize to outside of the narrow training distribution, whereas our method effortlessly segments a wide spectrum of targets as shown in Figures 4.1 and 4.4.

Crucially, our approach exhibits a relative improvement of up to 32% compared to the very recent method of Zhu *et al.* [296], which is also based on a video-diffusion representation. We investigate this in Section 4.5.3 and demonstrate that preserving as much of the representation learned during generative pre-training as possible is key to achieving the highest degree of generalization in referral video segmentation. We will release the code, models, and data for reproducing our results.

"Explode colorful smoke coming out"

"Time-lapse of a blooming flower on a stem "

Figure 4.2: Through Internet-scale pre-training, video diffusion models can generate realistic videos capturing the entire diversity of the dynamic visual world (generated samples shown above). We leverage their powerful visual-language representation for open-world referral video segmentation .

4.2 Related Work

Referring Video Segmentation (RVS) involves segmenting specific regions in a video based on a natural language description [56, 100, 195]. Most benchmarks for this task were developed by adding referral expression annotations to existing Video Object Segmentation (VOS) datasets, such as DAVIS'17 [171] or YouTube-VOS [256]. Consequently, the role of language in these benchmarks is limited to providing an interface for user-initialized object tracking [164, 249]. While this specific task — Referral Video Object Segmentation (RVOS) — is valuable, it addresses only a narrow subset of the possible interactions between language and the space-time continuum of videos. Equally important is the ability of RVS methods to segment video concepts beyond common object categories. To address this gap, we introduce a new benchmark focused on segmenting dynamic processes, which we term Referral Video Process Segmentation (Ref-VPS).

Earlier RVOS approaches [18, 87, 156] generally employed a *bottom-up* strategy: first, image-level methods [30, 168, 187, 265] were applied to obtain frame-level masks, followed by spatio-temporal reasoning, such as mask propagation [195], to refine the segmentation across frames. More recently, with the success of cross-attention-based methods [143, 225, 274] in object segmentation and tracking, query-based architectures have been introduced to RVOS, leading to significant improvements, with ReferFormer [244] and MUTR [260] being notable examples. The limited scale of paired video-language data with segmentation annotations has always been a major limitation in RVOS, causing most methods to train jointly on video and image samples [92, 96]. The latest approaches go even further and unify all object localization datasets and tasks in a single framework to maximize the amount of training data [36, 246, 259]. However, while these models excel in object tracking, they struggle to generalize to more dynamic concepts. In contrast, we demonstrate that generative video-language pre-training on Internet-scale data [9, 194] results in a universal mapping between the space of language and the ever-changing visual world.

Diffusion Models have become the de-fact standard for generative learning in computer vision [79, 202] and beyond [37]. Among them, the Denoising Diffusion Probabilistic Model (DDPM) [79] leverages neural network components to model the denoising process and builds a weighted variational bound for optimization. Stable Diffusion (SD) [185] shifts the denoising process into the latent space of a pre-trained autoencoder [101], allowing for model scaling. Expanding from images to videos, diffusion models have seen success in text-to-video (T2V) generation [20, 32, 33, 230, 288]. In addition to the capacity to generate high-fidelity images based on text prompts, the T2V diffusion models implicitly learn the mapping from linguistic descriptions to video regions, providing an opportunity to repurpose them for RVOS. Among current T2V methods, ModelScope [230] and VideoCrafter [32, 33] stand out for their open-source implementations, forming the backbone of our research.

Visual-language Pre-training for Perception: in addition to being highly effective in image and video generation, diffusion models have been shown to learn a strong representation of the natural image manifold. Several works have demonstrated that these representations can be re-purposed for classical computer vision problems, including semantic segmentation [254, 278, 284] and pixel-level correspondence [213], achieving an impressive degree of generalization. Others have shown that image diffusion models learn powerful representations of objects, enabling open-world novel view synthesis [122] and amodal segmentation [160]. Most recently, Zhu *et al.* [296] also leverages pretrained T2V models for RVOS, however, our analysis shows that their approach fails to fully capitalize on the universal visual-language mapping learned in generative pre-training. In this work, we explore the application of video diffusion models to RVS, demonstrating how to maintain a high-level generalizability during fine-tuning.

In a separate line of work, visual-language representations learned with contrastive objectives [11, 177] have been adapted for referring image [110, 183, 255, 267] and video segmentation [292]. However, their performance remains limited, compared to both generative models, as well as classical referral segmentation approaches.

4.3 Method

4.3.1 Learning the visual-language manifold via video denoising

Text-to-Video (T2V) diffusion models [33, 230, 288] generate videos that align with a given language description, starting from Gaussian noise. The process can be formalized as:

$$\hat{x} = f_{\mathrm{vdm}}(x_T, c, T), \tag{4.1}$$

where $\hat{x}$ is the generated video, T denotes the maximum timestep specified by the video diffusion model f_{vdm}, x_T is a sample drawn from a Gaussian distribution $\mathcal{N}(\mu_T, \sigma_T^2)$ predefined by the video diffusion model, and c is the conditioning prompt. To reduce computational complexity, these models often perform denoising in the latent space [185]. Specifically, a pretrained Variational Autoencoder (VAE) [101] is employed to map the video x from pixel space into latent space, denoted as $\mathcal{E}(x) = z$, while a decoder reconstructs it from the latent space, $\mathcal{D}(z) \approx x$. Thus, the generation process becomes:

$$\hat{x} = \mathcal{D}\left(f_{\mathrm{vdm}}(z_T, c, T)\right). \tag{4.2}$$

During training, rather than denoising from pure Gaussian latents, T2V models denoise from partially noisy video latents and optimize the following latent diffusion objective:

$$\min_{\theta} \mathbb{E}_{z \sim \mathcal{E}(x), t, \epsilon \sim \mathcal{N}(0,1)} \left\| \epsilon - \epsilon_\theta(z_t, e_c, t) \right\|_2^2, \tag{4.3}$$

where ϵ is the Gaussian noise added to the clean video latent, z_t represents noisy video latent at timestep t derived by the diffusion forward pass [79, 185], and e_c is the conditional embedding generated from c using a text encoder, such as CLIP [177]. The denoising network $\epsilon_\theta(z_t, e_c, t)$, typically a U-Net [186], is tasked with predicting the noise ϵ. In this network, the conditional embedding e_c interacts with the latent representations through cross-attention mechanisms, guiding the model to generate diverse, semantically accurate videos based on the provided text descriptions.

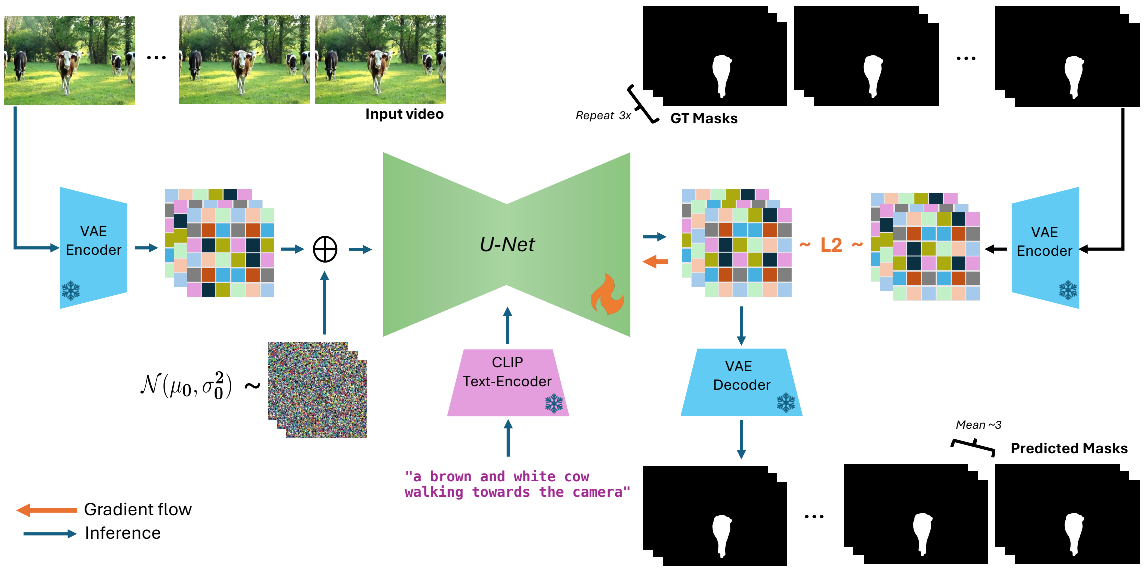


Figure 4.3: The model architecture of Refer Everything with Diffusion Models (REM). Like a video diffusion model it is based on, our approach takes video frames with added noise and a language expression as input. Our key insight is preserving as much of the diffusion representation intact as possible by supervising segmentation masks in the latent space of the frozen VAE.

4.3.2 From language-conditioned denoising to referral video segmentation

Referral Video Segmentation (RVS) involves segmenting an entity in a video across spatial and temporal dimensions, guided by a natural language description. Formally, the task is defined as:

$$\hat{m} = f_{\text{RVS}}(x, c), \tag{4.4}$$

where f_{RVS} is the RVS model, x represents a video sequence, c denotes a referral text prompt, and $\hat{m}$ corresponds to the binary masks produced as output. This task aligns naturally with T2V models, which establish a robust mapping between the entities described in the text and the corresponding spatial-temporal regions in the video by optimizing the denoising objective in Equation 4.3.

Several prior works have explored the alignment of diffusion models with referral segmentation [254, 284, 296], typically employing these models as *feature extractors*. Specifically, they adjust the input format of a referral segmentation model to match that of the denoising network ϵ_θ, and pass the resulting features to a task-specific decoder f_{dec} (*e.g.*, a convolutional network) to predict the target masks:

$$\hat{m} = f_{\text{dec}}(\epsilon_\theta^{(n)}(z_t, e_c, t)), \tag{4.5}$$

where z_t is the noisy latent representation of the input images at timestep t, e_c is a feature embedding of the referral expression c, and $\epsilon_\theta^{(n)}$ denotes the intermediate feature at the n^{th} layer. In practice, t is usually set to a small value (*e.g.*, 50), and n is set to the last layer indexes to obtain the optimal performance. The entire model is then trained in a conventional discriminative learning setup. However, replacing parts of the generative model with newly initialized layers can disrupt alignment between the model's representation from pre-training and the new features learned on narrow-domain datasets, leading to a substantial loss of generalization capabilities.

In our approach, shown in Figure 4.3, we propose to instead preserve the architecture (and thus the representation) of the video diffusion model in its entirety. Specifically, rather than using intermediate features $\epsilon_\theta^{(n)}$, REM repurposes the whole denoising network ϵ_θ (together with the VAE) by shifting its objective from predicting noise to predicting mask *latents* (shown on the right in Figure 4.3):

$$\hat{m} = \mathcal{D}(\epsilon_\theta(z_t, e_c, t)), \tag{4.6}$$

where $\mathcal{D}$ denotes the (frozen) VAE decoder used to produce the actual binary segmentation masks from the predicted latents. That is, instead of learning the decoder network f_{dec} from scratch, we reuse the VAE from the video diffusion model. This subtle yet powerful modification allows the model to better preserve its representation learned on Internet-scale data during generative pretraining while adapting to the task of RVS.

Training and Optimization. During training, to encode the ground-truth segmentation masks with the VAE, we broadcast the single-channel mask into three channels by simply duplicating it (shown in the top right of Figure 4.3). For simplicity, we still denote this three-channel mask representation as m. The pretrained VAE can then map the mask sequence into the latent space via $\mathcal{E}(m) = z^m$ and decode the masks back from predicted latents via $\mathcal{D}(z^m) \approx m$. For the noisy latent z_t and timestep t, we prioritize using latents that remain as clean as possible. Therefore, we always set the timestep to its minimum value, $t = 0$. To train the model, we supervise the predicted mask latents using an $\mathcal{L}_2$ loss (shown in the center-right of Figure 4.3) by minimizing:

$$\min_{\theta} \mathbb{E}_{z^m \sim \mathcal{E}(m), t=0} \left\| z^m - \epsilon_\theta(z_t, e_c, t) \right\|_2^2 . \tag{4.7}$$

Model Inference. During inference, we follow Equation 4.6, with $t = 0$, to decode the predicted mask latent and generate three-channel mask predictions. We then compute the single-channel masks by averaging the pixel values of the three channels and applying a constant threshold of 0.5 to binarize the result (as shown in the bottom right of Figure 4.3).

4.4 Benchmark design and collection

Existing datasets only allow to quantify the generalization of RVS models to rare *object* categories [2] or *static* ‘Stuff’ [146]. To present a more comprehensive generalization evaluation, we now discuss our approach to collecting a new benchmark. As covering the entire spectrum of concepts that can be spoken of in videos would be extremely costly, we seek to identify the most salient subset of the problem that requires joint modeling of language and temporal dynamics. Specifically, we target **dynamic processes**, defined as temporally evolving *events*, where the subjects undergo continuous changes in state, shape, or appearance. Crucially, the subjects in this context are not limited to objects, but include all concepts that are spatio-temporally localizable in videos, such as light or fire. The key steps for collecting this new benchmark, which we call Referral Video Process Segmentation (Ref-VPS), include selecting representative videos and annotating them with referring expressions and segmentation masks.

4.4.1 Video Selection

To source the videos for our benchmark we require a large, public, and diverse database that is queriable with natural language and allows re-distribution of content for research purposes. Based on these requirements, we choose the TikTok social media platform which has over 1 billion active users across the world and receives tens of millions of video uploads daily, capturing a wide range of dynamic visual content. TikTok's policies generally allow for free redistribution of content, with individual users having the option to opt-out.

To search for videos that capture dynamic processes, as defined above, we first identify a non-exhaustive list of six broad and possibly overlapping concepts (*e.g.*, 'object transformations', or 'entities with dynamic boundaries', the full list together with definitions provided in the appendix). Then, for each concept we ask ChatGPT [158] to provide a list of concrete examples together with multiple text queries for search on TikTok (*e.g.*, 'a wax candle melting' for 'object transformations'), resulting in 120 individual concepts. We retrieved over 1,000 samples based on these queries, however, a majority of the queries did not yield suitable videos because of the physical nature of the event or ambiguity of the search query not lending itself to being accurately captured on TikTok. After removing irrelevant videos, the retrieved set is reduced to 342 samples.

We then manually filter these videos based on the following criteria: (1) videos that do not feature significant dynamic changes of the subject (*e.g.*, mostly stationary clouds in the sky); (2) dynamic processes that occur too rapidly to allow for the labeling of a sufficient number of non-empty frames (*e.g.*, flashes of lightning); (3) video with frequent shot changes, which make it impossible to extract an interrupted clip capturing the event of interest. Additionally, for videos that represent compilations of similar events, we split them into individual clips and treat each one independently. The resulting dataset contains 141 video clips representing 39 dynamic process concepts. The entire dataset is intended for zero-shot evaluation, so we do not define any additional splits. Representative samples of the videos are shown in Figure 4.1.

4.4.2 Annotation collection and evaluation

To label the videos selected above, we begin by adjusting the temporal boundaries of each clip to focus on the event of interest and avoid shot changes. We also make sure that the event is captured in its entirety whenever possible, including some context before and after it. The clips are then exported at 24 FPS as frames. If a video contains irrelevant frames, such as the TikTok logo at the end, we crop the frames accordingly to remove the padding.

To collect referral expressions, we first manually identify the entity of interest in each clip. The selected entity is then labeled with referral expressions by two independent annotators. Each annotator provides two expressions for the target, resulting in a total of four expressions per clip, capturing different ways to describe the same phenomenon. Following the standard protocol [100, 195], models are evaluated on all queries and the results are averaged.

Finally, we densely label the targets identified above with segmentation masks at 24 FPS. To this end, we employ a semi-automatic pipeline, capitalizing on the recently introduced SAM2 [184] foundational model for interactive video segmentation. In particular, we provide positive and negative click annotation in the middle frame of a video first to ensure accurate boundary segmentation. SAM2 then automatically segments the entity of interest in the frame, as well as propagates the mask across the entire clip. We interactively improve segmentation quality by providing additional clicks as needed. In the end, we manually refine the masks in frames where SAM2 fails. We report the statistics of our benchmark in the appendix. For evaluation, we follow Tokmakov *et al.* [218] and only report region similarity $\mathcal{J}$ as contour accuracy $\mathcal{F}$ is often not well defined for the entities like smoke or light which are frequent in Ref-VPS.

4.5 Experiments

Datasets and Evaluation. We evaluate our method on five benchmarks in total. Ref-YTB [195] and Ref-DAVIS [100] are standard RVOS benchmarks for evaluating our model's performance on object tracking. The evaluation on Ref-YTB is done on the official challenge server, and Ref-DAVIS is evaluated using the official evaluation code. For evaluating generalization to rare objects and 'Stuff' categories, we use the BURST [2] and VSPW [146] datasets respectively. Finally, we evaluate REM and the strongest baselines on our newly introduced Ref-VPS benchmark that focuses on dynamic process segmentation (detailed in Section 4.4). All these datasets except Ref-YTB are only used for evaluation (*i.e.*, the results are zero-shot).

For Ref-YTB [195] and Ref-DAVIS [100] we use the standard evaluation metrics -

Region Similarity ($\mathcal{J}$), Contour accuracy ($\mathcal{F}$) and their mean ($\mathcal{J}\&\mathcal{F}$). For all other evaluations we use the Region Similarity ($\mathcal{J}$) metric.

Implementation details. We build our method upon ModelScopeT2V [230]. We adopt a two-stage training strategy following prior work [296]. In the first stage, we fine-tune only the spatial weights using image-text samples from Ref-COCO [270] for 1 epoch and then fine-tune all weights for 40 epoch using Ref-YTB [195] video-text training samples and 12k samples from Ref-COCO jointly. In the second stage, the image samples from Ref-COCO are converted to pseudo videos through augmentations following previous works [245]. We freeze the CLIP encoder and VAE during training.

4.5.1 Referral Video Object Segmentation Results

In this section, we compare REM to the state of the art on the standard RVOS benchmarks. We report results on the validation set of Ref-DAVIS [100] and the test set of Ref-YTB [195] in Table 4.1. Our method outperforms state of the art in terms of $\mathcal{J}$ on Ref-DAVIS and is only second to UNINEXT [259] on Ref-YTB. Note that this approach is specifically designed for object segmentation and utilizes more than 10 datasets with localization annotations like bounding boxes and masks for training. In contrast, REM adopts an architecture of a video generation model and is only fine-tuned on one image- and one video-segmentation dataset. Despite this, our method is competitive with UNINEXT on standard RVOS benchmarks, and we will show next, outperforms it by up to 21 points out-of-domain in terms of $\mathcal{J}$.

Another notable observation is that REM also outperforms VD-IT [296], which is built on top of the same video diffusion backbone of Wang et al. [230], on both datasets. This result demonstrates the effectiveness of our approach to preserving the visual-language representations learned on the Internet data, which will become even more evident in out-of-domain evaluation.

4.5.2 Out-of-domain generalization

We begin by performing a generalization study on existing open-world tracking BURST dataset [2] as well as on the 'Stuff' categories [27] from VSPW [146] in Table 4.2. BURST is an open-world video object segmentation benchmark, whereas

Method	Pretraining Data	Mask/Box Supervision	Ref-DAVIS			Ref-YTB		
			$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Referformer [244]	ImageNet + Kinetics + SSv2	Ref-COCO/+/g + Ref-YTB	61.1	58.1	64.1	62.9	61.3	64.6
MUTR [260]	ImageNet + Kinetics + SSv2	Ref-YTB + AVS	68.0	64.8	71.3	68.4	66.4	70.4
VLMO-L [292]	Unknown	Ref-COCO/+/g + Ref-YTB	70.2	66.3	74.1	67.6	65.3	69.8
UNINEXT [259]	Object365	10+ Image/Video datasets	72.5	68.2	**76.8**	**70.1**	**67.6**	**72.7**
VD-IT [296]	LAION5B + WebVid	Ref-COCO/+/g + Ref-YTB	69.4	66.2	72.6	66.5	64.4	68.5
REM (Ours)	LAION5B + WebVid	Ref-COCO/+/g + Ref-YTB	**72.6**	**69.9**	75.29	68.4	67.05	69.73

Table 4.1: Comparison to the state of the art on the validation set of the Ref-DAVIS and the test set of Ref-YTB benchmarks using the standard metrics. Our method performs on par with the strong UNINEXT approach, despite not being specifically designed for object localization and having access to only a fraction of the localization labels used by that method.

Benchmark	MUTR	UNINEXT	VD-IT	REM (Ours)
VSPW	10.5	10.1	12.7	**15.2**
BURST	27.9	30.2	30.9	**40.4**

Table 4.2: $\mathcal{J}$ Comparison to state of the art on the 'Stuff' categories in the eval set of VSPW and on the joint val and test sets of BURST. Our approach demonstrates much stronger generalization, notably, outperforming VD-IT which is based on the same diffusion backbone.

Benchmark	MUTR	UNINEXT	VD-IT	REM (Ours)
Ref-VPS ($\mathcal{J}$)	26.22	28.36	37.58	**49.56**

Table 4.3: Comparison to the state of the art on our new Ref-VPS benchmark. REM shows much stronger generalization to challenging, dynamic concepts in this dataset compared to the baselines by effectively capitalizing on Internet-scale visual-language pre-training.

VSPW tests the ability to generalize to non-object categories. We report zero-shot evaluation results on the validation set of VSPW and combined validation and test sets of BURST and compare to the top performing methods from Table 4.1 that have public models.

Firstly, we observe that on both out-of-domain challenges our method outperforms all the baselines by significant margins. The improvements are especially noticeable on BURST, demonstrating that REM successfully preserves the strong object representation learned by Internet-scale pre-training of the diffusion backbone. In contrast, VD-IT loses this generalization capacity during fine-tuning and only performs on par with UNINEXT. On the 'Stuff' categories all the methods do relatively poorly, reflecting the challenge of generalizing to more amorphous 'Stuff'. Here VD-IT maintains a lead over entirely object-centric UNINEXT but REM still outperforms both baselines.

Finally, we compare REM to the top-performing RVS baselines on our Ref-VPS benchmark in Table 4.3. Here the differences between the methods are more pronounced, highlighting the value of Ref-VPS in assessing video-language understanding

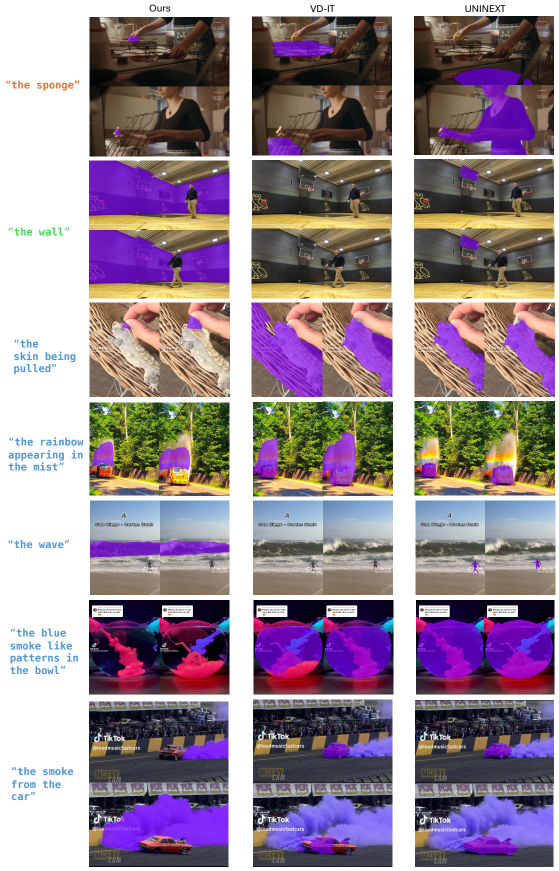


Figure 4.4: Qualitative results of REM and state of the art baselines on BURST, VSPW and Ref-VPS benchmarks. Our method demonstrates both superior coverage of rare, dynamic concepts and higher segmentation precision. Video comparisons are available here.

capabilities of neural representations. Our approach outperforms all baselines by up to 12 points in Region Similarity (32% relative improvement), and notably surpasses the top RVOS method, UNINEXT, by 21.2 points (74.8% relative improvement). While generative pre-training enhances VD-IT's generalization ability compared to UNINEXT, it struggles to preserve its representations as effectively as REM.

A qualitative comparison of REM with VD-IT and UNINEXT is provided in Figure 4.4. Firstly, we can see that our approach can successfully track the sponge (which was never seen in training) in a challenging sequence from BURST, whereas other methods focus on foreground objects. In the second sequence from VSPW REM successfully generalizes to the non-object ‘wall’ category, whereas UNINEXT focuses on a nearby *object* and VD-IT fails entirely. The following examples from our

Ref-VPS illustrate that both baselines exhibit object-centric bias, as in the examples with the lizard skin in row 3 and blue smoke in row 6. While VD-IT shows better generalization, it often simply segments the dominant region in the video (see rows 4 and 7 in Figure 4.4). In contrast, REM demonstrates both good coverage of rare concepts and high precision with respect to the language prompt.

4.5.3 Ablation analysis

In this section, we analyze our proposed approach of transferring generative representations to the task of RVS. We report results on one representative RVOS benchmark (Ref-YTB) and on our new Ref-VPS. Note that for efficiency we fine-tune all the models on a subset of image and video data (12000 samples) so the numbers are lower than those reported in the previous section.

Generative pre-training. We begin by evaluating the effect of the generative pre-training strategy in Table 4.4. Firstly, we design a frame-level baseline that fine-tunes StableDiffusion [20] on every frame individually (row 1 in the table). While this variant has no temporal modeling capacity, its architecture is similar to UNINEXT [259] - the state-of-the-art approach for RVOS. Interestingly, it strongly under-performs compared to our best video-based variant not only on our Ref-VPS but also on the object-centric Ref-YTB benchmark. These results demonstrate that, despite the fact that images are the dominant data source in generative pre-training, fine-tuning StableDiffsuion for video generation is crucial for learning an effective representation for tracking.

Next, we compare several strategies for learning video diffusion models. We begin by studying two variants of the VideoCrafter model [32, 33] (denoted as VideoCrafter-1 and VideoCrafter-2 in Table 4.4). They are both trained on 600M images from LAION [194] and 10-20M Internet videos. However, VideoCrafter-2 is further tuned to increase the quality of the generated samples. Our findings indicate that this fine-tuning step leads to significant performance gains across both benchmarks. This suggests that improving the quality of video generation models can directly translate to enhanced performance in our video segmentation framework.

Finally, we evaluate ModelScope [230], which is trained on larger LAION 2B and a comparable amount of video samples (last row in Table 4.4). This model

Backbone	Decoder	Ref-YTB	Ref-VPS
		($\mathcal{J}$&$\mathcal{F}$)	$\mathcal{J}$
Stable Diffusion 2.1	Frozen VAE	59.38	30.69
VideoCrafter-1		59.10	29.22
VideoCrafter-2		**65.00**	37.28
ModelScope T2V		**64.57**	**39.05**
ModelScope T2V	CNN	59.6	29.93

Table 4.4: Analysis of the effects of generative pre-training and discriminate fine-tuning strategies on Ref-YTB and Ref-VPS. The key to the success of REM is capitalizing on Internet-scale image and video pre-training and preserving as much of this representation as possible.

delivers performance comparable to the best version of VideoCrafter on the Ref-YTB benchmark, while demonstrating superior generalization to more challenging concepts in Ref-VPS. These results further highlight that both large-scale pre-training on image data as well as learning to model video-language interactions are crucial components for robust RVS representation learning.

Fine-tuning strategy. We now ablate the effectiveness of our design decision to re-use a frozen VAE decoder for mask prediction, rather than replacing it with a dedicated mask prediction module, as was done in some of the prior work [284, 296]. To this end, we replace the VAE with a CNN mask decoder adopted from [284] and train it jointly with the rest of the model (last row in Table 4.4). Removing the pre-trained VAE decoder has a moderate negative effect on performance on Ref-YTB, but, notably, destroys the model's ability to generalize our challenging Ref-VPS benchmark. This result underscores the main message of our paper - preserving as much of the representation learned during generative pre-training is key for achieving generalization in referral video segmentation.

4.6 Discussion

In this paper, we proposed REM, a framework that capitalizes on Internet-scale video-language representations learned by diffusion models to segment a wide range of concepts in video that can be described through natural language. Our key insight is that changing as little as possible in the representation is key to preserving its universal mapping between language and visual concepts during fine-tuning. We have evaluated the generalization capabilities of our approach on existing datasets for open-world object and stuff segmentation, and also collected Ref-VPS - a new benchmark for referral segmentation of dynamic processes in videos. Our extensive experimental evaluation demonstrates that, despite only being trained on object masks, REM successfully generalizes to unseen object and non-object concepts, outperforming all prior work by up to 12 points in terms of region similarity.

Despite REM's impressive generalization abilities, the problem of RVS is far from being solved. REM still exhibits some object-centric bias and struggles with extremely fast changes. Exploring ways to preserve even more of the representation learned during generative pre-training, *e.g.*, via low-rank adaptation [82] of the visual backbone, is a very promising direction to address some of these issues. In addition, note that REM is a generic framework where the backbone of Wang *et al.* Wang et al. [230] can be replaced with a more advanced representation, tracing the progress of language-conditioned video generative models.

In this appendix, we first include the dataset details for your Ref-VPS dataset in Section 4.7. Next, we include additional experimental evaluations, including comparisons with state-of-the-art methods on more challenging fighting scenes, failure case analysis, and concept coverage comparisons in Section 4.8. Finally, in Section 4.9, we show the full implementation details.

4.7 Dataset Details

During our dataset collection, we first identify a non-exhaustive list of six broad and possibly overlapping concepts. The list of the concepts and their definitions are included below:

- **Temporal object changes:** Concepts involving changes over time (*e.g.*, object deformation, melting)
- **Motion Patterns:** Concepts involving movement and displacement of non-object regions (*e.g.*, water ripples, flickering flames)
- **Dynamic environmental changes:** Changes in the environment that affect spatial regions over time (*e.g.*, clouds moving across the sky, waves rising)
- **Interaction Sequences:** Concepts involving interactions between objects (*e.g.*, bullet hitting glass, object collisions)
- **Pattern evolution:** Concepts where patterns or textures evolve or change

dynamically (*e.g.*, changing patterns of smoke dispersion, fluctuating light levels)

The final dataset contains 141 video clips representing 39 dynamic process concepts. We report a comprehensive list of key statistics in Table 4.5. Most of our samples are around 2.5 to 5 secs in length but can go up to more than 20 seconds. The distribution of our sample lengths is reported in Figure 4.5.

4.8 Additional Experimental Evaluations

4.8.1 Evaluation on Challenging Fight Scenes

Fight sequences in movies, television, and animated shows pose a unique set of challenges. Typically fight scenes are characterized by objects/characters undergoing severe and frequent occlusions and leaving the frame entirely, coupled with frequent pose changes of the camera. This leads to drastic changes in the appearance of the object and requires high levels of temporal and semantic consistency to accurately track, re-identify, and segment the referred entity. Our diffusion fine-tuning method excels in this domain of extremely challenging samples as illustrated in Figure 4.6. It is clear to find that UNINEXT and VD-IT both fail whenever there is a large occlusion causing the referred entity to become invisible. Even though VD-IT uses frozen Video diffusion features, their method is unable to leverage the temporal consistency learned during Video Diffusion pre-training as well as our method.

4.8.2 Failure Cases

A few representative failure cases of REM on Ref-VPS are shown in Figure 4.7. Our method suffers from object-centric bias in the most challenging scenarios and struggles with extremely fast processes.

4.8.3 Concept Coverage Plot on BURST Dataset

We additionally show the concept coverage on BURST [2] dataset among VD-IT [296], MUTR [260], UNINEXT [259], and ours in Figure 4.8. In general, our model has significantly better coverage of different object concepts with a performance

Clips	141
FPS	24
Frames	22831
Concepts	39
Avg length (s)	6.75
Annotation FPS	24
Min-resolution	712 × 576
Max-resolution	1024 × 576

Table 4.5: Statistics of our Ref-VPS benchmark. Our dataset contains 141 video clips covering 39 concepts for dynamic processes.

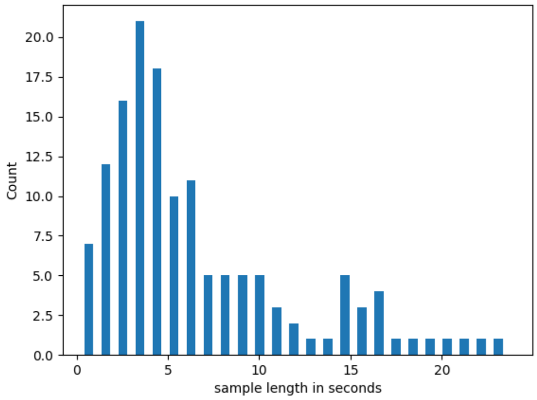


Figure 4.5: Distribution of sample lengths in Ref-VPS.

Benchmark	Type	Training Samples	Testing Samples
Ref-COCO [270]	Image	320K	-
Ref-YTB [195]	Video	12,913	2,096
Ref-DAVIS [100]	Video	-	90
BURST [2]	Video	-	2,049
VSPW [146]	Video	-	343

Table 4.6: Details about the benchmarks we used for training and evaluation.

improvement of at least 9.5 %. Moreover, our REM is more robust in the long-tail regions, indicating the promising generalization capacity of our method by repurposing video diffusion models. The performance improvement compared to VD-IT, which uses frozen video diffusion features, indicates that simply freezing the diffusion model as a backbone does not guarantee the transfer of the diffusion knowledge to downstream tasks and verifies the design of our approach.

4.9 Implementation Details

4.9.1 Dataset Details

We report the details about the benchmarks we used in Table 4.6.

4.9.2 Stuff Category Evaluation

Neither BURST [2] nor VSPW [146] contains referral text for the segmented entities. For them, we automatically generate referral expressions using only the category information of the mask entity as ”the <class>” (*e.g.*, ”the hat”). For VSPW we conduct our evaluation on the validation set which has 66 different stuff categories. In the case of BURST, we evaluate the combined validation and test set which contains 454 classes and a total of 2049 sequences.

4.9.3 Baseline Models

In addition to Table 4.1 in the main paper, we report the comprehensive list of bounding/mask supervision used by all methods in Table 4.7. We quote the results of all the baseline models on Ref-YTB and Ref-DAVIS from their original paper. For the evaluation of BURST, VSPW, and our dataset, we report the numbers by running the official checkpoints of MUTR[1], UNINEXT[2], and VD-IT[3].

4.9.4 Training Details

We build our method upon ModelScopeT2V [230]. We adopt a two-stage training strategy following prior work [244, 260]. In the first stage, we fine-tune only the spatial weights using image-text samples from Ref-COCO [270] for 1 epoch and then fine-tune all weights for 40 epoch using Ref-YTB [195] video-text samples and 12k samples from Ref-COCO jointly. In the second stage, the image samples from Ref-COCO are converted to pseudo videos through augmentations following Wu et al. [245]. We resize all the training data to 512 × 512 for training and evaluate under the original resolution. We freeze the CLIP encoder and VAE during training. We use AdamW [131] for optimization with a constant learning rate of 1e-6 in both stages. The training batch size is 4, and for each sample, we randomly load a 8-frame video clip. We train our model using 4 NVIDIA 80GB A100 GPUs, and it takes about 1 week to finish the whole training process.

[1] https://github.com/OpenGVLab/MUTR
[2] https://github.com/MasterBin-IIAU/UNINEXT
[3] https://github.com/buxiangzhiren/VD-IT

Method	Mask/Box annotations	Ref-Davis ($\mathcal{J}$&$\mathcal{F}$)	Ref-YTB ($\mathcal{J}$&$\mathcal{F}$)
Referformer	RefCOCO/g/+, Ref-Youtube-VOS	61.1	62.9
MUTR	Ref-Youtube-VOS, AVS	68.0	68.4
VLMO-L	RefCOCO/g/+, Ref-Youtube-VOS	70.2	67.6
UNINEXT	Objects365, COCO, RefCOCO/g/+, GOT-10K, LaSOT, TrackingNet, Youtube-VOS, BDD100K, VIS19, OVIS, Ref-Youtube-VOS	72.5	**70.1**
VD-IT	RefCOCO/g/+, Ref-Youtube-VOS	69.4	66.5
REM (Ours)	RefCOCO/g/+, Ref-Youtube-VOS	**72.6**	68.4

Table 4.7: Comprehensive list of bounding/mask supervision used by all methods.

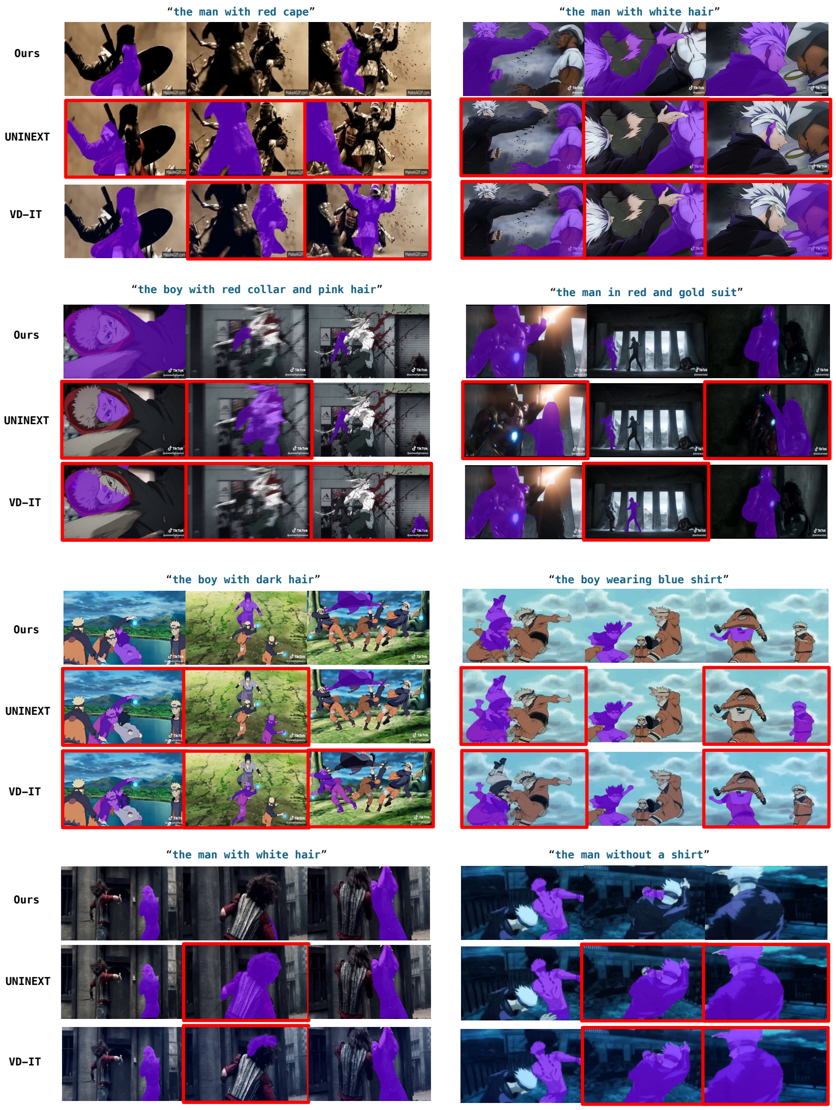
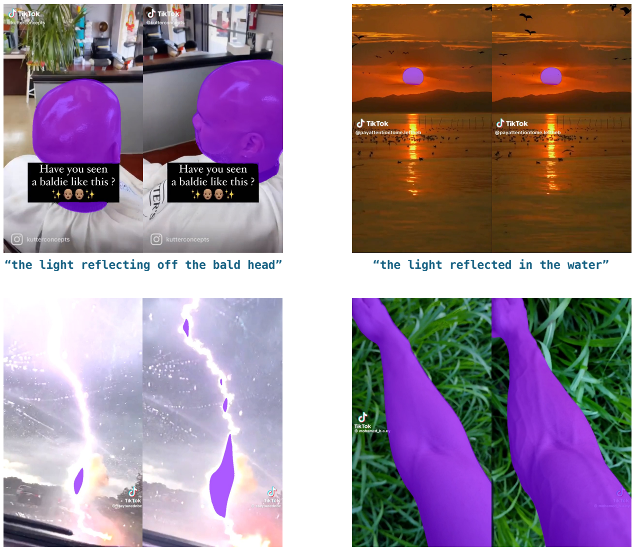

"the man with red cape"
"the man with white hair"
Ours
UNINEXT
VD-IT
"the boy with red collar and pink hair"
"the man in red and gold suit"
Ours
UNINEXT
VD-IT
"the boy with dark hair"
"the boy wearing blue shirt"
Ours
UNINEXT
VD-IT
"the man with white hair"
"the man without a shirt"
Ours
UNINEXT
VD-IT

Figure 4.6: Qualitative comparison of REM with state-of-the-art baselines on dynamic and challenging fight scenes. The incorrectly labeled frames are outlined in red. Our method outperforms the other methods in handling frequent occlusions and POV changes. For a better illustration of the differences, please watch the full videos here.


Have you seen a baldie like this ?
Have you seen a baldie like this ?

“the light reflecting off the bald head”

“the light reflected in the water”

“the lightning strike”

“the veins on the arm”

Figure 4.7: Failure cases of REM on Ref-VPS. Our model still exhibits some object-centric bias and struggles with extremely dynamic entities such as lightning.

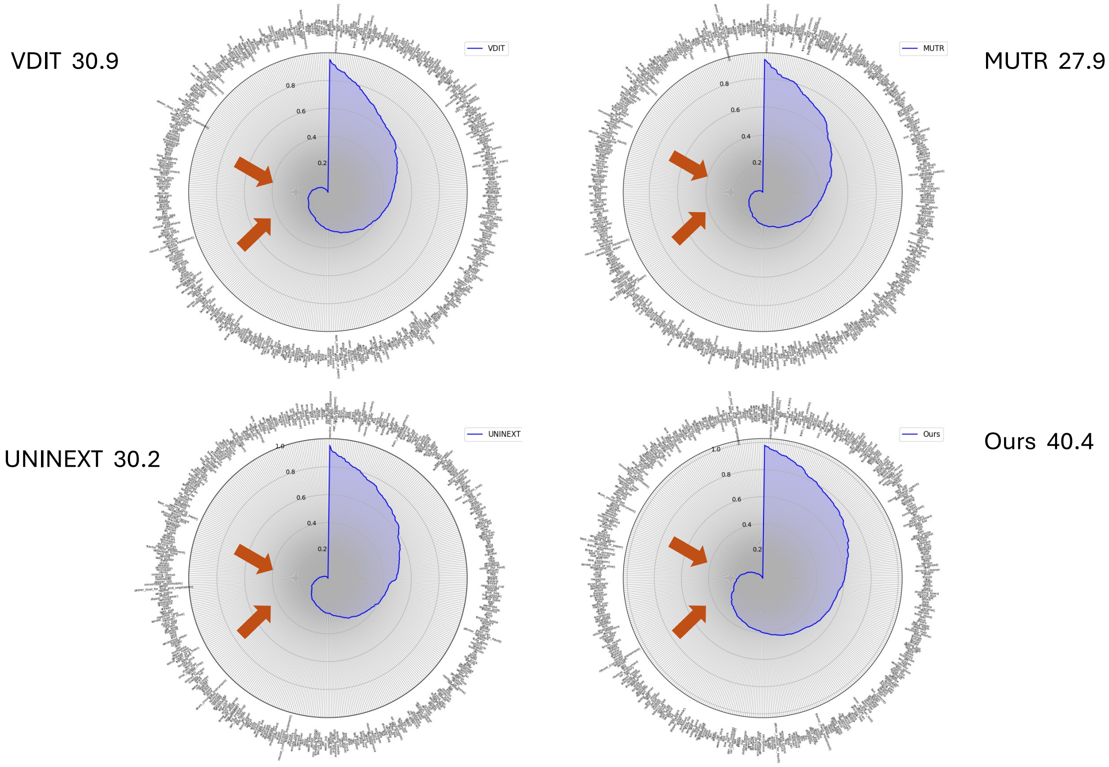


Figure 4.8: Class-wise $\mathcal{J}$ scores (mIoU) concept coverage on BURST. As indicated by the arrows, Our method is robust on the long-tail region compared to other methods.

Chapter 5

Extending Diffusion Models from a Single Task Performance to a Generalist Model

5.1 Introduction

A central theme across this dissertation is unifying perception and generative modeling under a single, coherent framework. In earlier chapters, we demonstrated that Refer Everything Models (REM) benefit significantly from using pretrained visual autoencoders. In particular, REM leverages a high-quality VAE to decode segmentation masks, enabling it to generalize across challenging referring video segmentation benchmarks. This design choice highlights a deeper insight: many structured visual modalities—masks, edges, depth, surface normals, and more—can be represented in an image-like form, as shown in Figure 5.1. Once cast into this shared latent image space, they become compatible with powerful pretrained generative backbones, especially diffusion models.

This observation motivates the broader question explored in this chapter: Can we train a single diffusion-based model that serves as a generalist engine for diverse visual tasks? Instead of building separate models for segmentation, depth estimation, inpainting, or stylization, we aim to learn a unified model that handles all of them

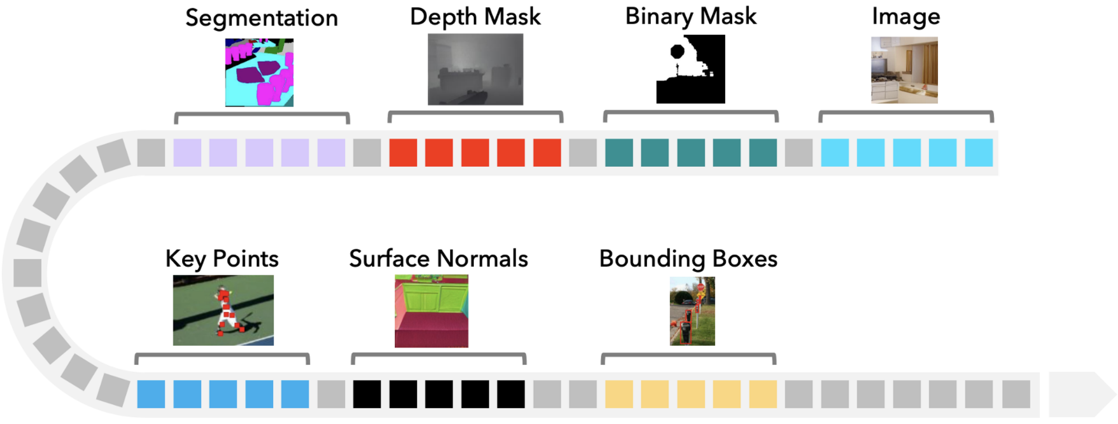


Figure 5.1: Taskwise, many structured visual modalities—masks, edges, depth, surface normals, and more—can be represented in an image-like form, making it possible to learn a unified generalist model for all these tasks.

within one architecture, one training pipeline, and one inference API. Such a generalist model would greatly simplify system design, improve reusability, and reveal deeper connections across seemingly different visual tasks.

In this chapter, we present a diffusion-based generalist model that bridges the REM line of work and the Qwen-VAR approach. First, we demonstrate that a single model can share a diffusion backbone across a wide range of core computer vision tasks. These tasks include segmentation-derived modalities, geometry prediction (e.g., depth, surface normals), and image manipulation tasks formulated as conditioned generation.

Beyond traditional CV tasks, we push further to investigate whether the same system can extend to non-CV tasks, including image restoration, enhancement, or local edits. Many of these tasks require richer multimodal reasoning: users may specify structural constraints, provide exemplars for reference, or describe fine-grained editing instructions. We show that while diffusion models remain effective in this larger space, purely text-conditioned diffusion backbones struggle to interpret complex instructions. This naturally motivates integrating a modern multimodal large language model (MLLM) to process and understand more expressive user inputs.

Finally, although our unified diffusion framework provides broad task coverage and strong qualitative performance, it suffers from a major limitation: latency. Even with a reasonably lightweight configuration such as Qwen-VL + Flux, producing a single output typically takes around **20 seconds**. For a state-of-the-art production-grade model such as Qwen-Image, generation time often approaches one minute per image—far too slow for interactive use or downstream pipelines requiring high throughput. This motivates the next chapter, where we transition from diffusion-based decoders to autoregressive generative modeling, culminating in the design of UniGen-AR as a more efficient, scalable, and general-purpose generative backbone.

5.2 Diffusion Generalist Models

5.2.1 Task Formulation

We follow a task formulation analogous to modern multimodal large language models, but with visual outputs.

Let the model receive an input sequence

$$\mathcal{X} = \{x^{\text{text}}, x^{\text{img}}\}, \tag{5.1}$$

where x^{text} is a natural-language instruction and x^{img} is an optional input image (e.g., a degraded observation, a reference exemplar, or an empty placeholder in text-to-image generation). These inputs are encoded into a shared embedding space using a multimodal encoder:

$$E(\mathcal{X}) = \mathbf{h} = \{h_1, \ldots, h_L\}. \tag{5.2}$$

The diffusion model is then tasked with generating a visual output y, represented in the latent space of the image decoder. Depending on the task, y may correspond to RGB images, structural modalities such as depth or normals, segmentation masks, or restoration outputs. Formally, the model learns the conditional distribution

$$p_\theta(y \mid \mathbf{h}), \tag{5.3}$$

and is optimized using a standard denoising objective over latent variables z_t sampled

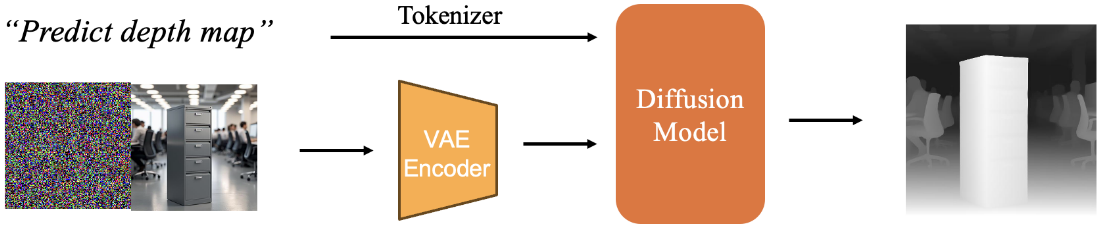


Figure 5.2: Architecture for our diffusion-based model for multiple visual tasks. We concatenate clean reference tokens with the noisy target latent for the denoising network to provide context for the model.

from a predefined noise schedule:

$$\mathcal{L} = \mathbb{E}_{y,t,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(z_t, t, \mathbf{h})\right\|_2^2\right]. \tag{5.4}$$

This formulation unifies diverse visual tasks by expressing them as conditional generation problems, where all inputs—textual or visual—are embedded into a single conditioning sequence and all outputs are synthesized as images through the diffusion decoder.

5.2.2 From a Single Task to Multiple Perception Tasks

We begin by describing the text-only conditioning pathway, which forms the foundation of our generalist model. The core idea is simple: for any visual task, the model receives a textual instruction describing the task and the desired output. This instruction can range from basic prompts ("generate corresponding depth map" "produce segmentation mask") to more detailed formatting instructions used in restoration or local editing.

To operationalize this, we adopt a unified text-conditioning interface inspired by existing instruction-driven diffusion models. The text prompt is tokenized using a lightweight language encoder, producing a sequence of contextualized embeddings. Meanwhile, to enable editing-like operations, we further concatenate the clean reference image latent to the noisy target latent for the denoising network, following prior approaches [97]. The model design of this architecture is shown in Figure 5.2.

This interface provides several advantages:

- Task unification – By treating every task as text-conditioned generation, the model does not require architecture-specific branches.
- Extensibility – New tasks can be introduced simply by defining new task descriptors or embeddings, without modifying the backbone.
- Interpretable control – Users can change task behavior at inference time through natural language alone.

For structured modalities (*e.g.*, depth, normals, masks), the model learns to produce outputs in the visual domain. These are later decoded or post-processed to obtain their native formats. Because all modalities share the same latent space, the model benefits from cross-task transfer, often producing more stable and spatially consistent predictions.

5.2.3 Towards Multimodal Reasoning: MLLMs Meet Diffusion

While text-only conditioning works well for standard CV tasks, if we further extend our task set from pure perception to unified visual generation, including tasks like image editing or reference-guided modifications, texts alone are not enough for these tasks. For example:

- Local edits require precise localization cues ("remove the object on the right," "brighten only the sky region").
- Pose- or identity-preserving edits require interpreting a reference image and maintaining structure.
- Restoration tasks must reason about image semantics while dealing with degradation.

To address these challenges, we introduce an MLLM-based multimodal encoder, similar in design to modern multimodal generation systems. The MLLM processes both the instruction text and an optional set of input images (*e.g.*, the degraded image, a reference exemplar, or a control signal). Its output is a set of joint visual-text embeddings that represent the task semantics, image context, and editing constraints in a unified token space. We show our model architecture design in Figure 5.3.

These embeddings are injected into the diffusion backbone via cross-attention

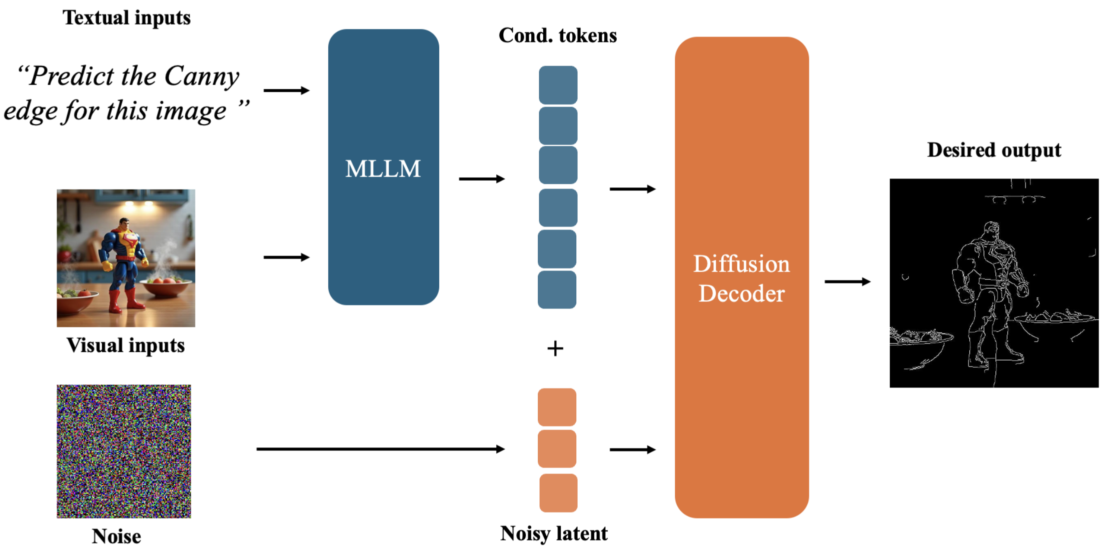


Figure 5.3: Architecture for our diffusion generalist models. A multimodal large language model is used to encode multimodal information to provide strong reasoning capability and understanding for the diffusion backbone.

layers at all resolutions. The key elements are:

Joint representation learning: The MLLM fuses textual instructions with visual content before conditioning the generative model.

Fine-grained cross-attention control: Embeddings corresponding to visual regions, reference identities, or object attributes steer the denoising trajectory.

Unified conditional interface: Whether the user wants to segment, restore, edit, or stylize, the model receives one consistent API: a mixed sequence of tokens summarizing all input modalities.

This design substantially expands the expressiveness of the diffusion model, enabling complex editing and restoration tasks that require contextual understanding.

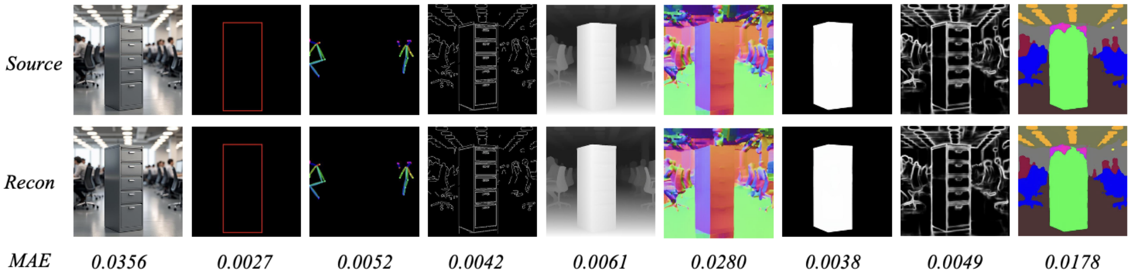


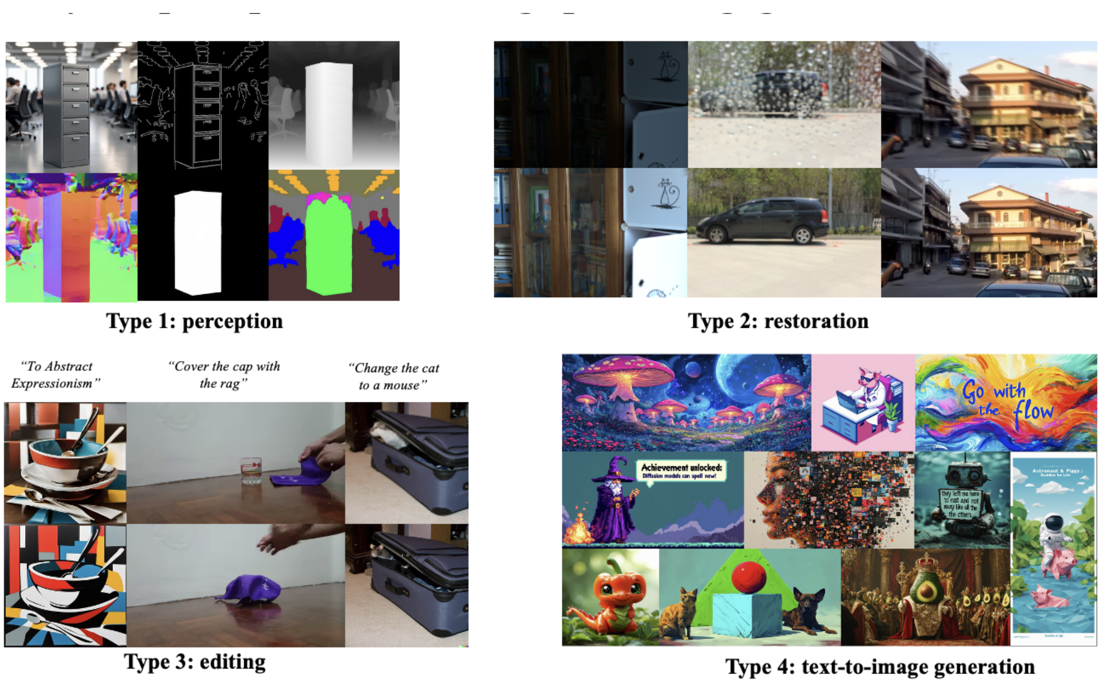
Figure 5.4: Reconstruction visualizations and error measurement for VAE regarding different visual modalities. A pretrained VAE is able to provide a robust latent space for a variety of visual attributes.

5.3 Evaluation

5.3.1 Sanity Check: Can VAE Provide a Unified Latent Space?

One of our motivations is that a well-trained VAE can provide a shared and robust latent space for different visual attributes. However, the distribution of these visual modalities varies significantly. We first conduct a sanity check to study whether the VAE itself is able to provide such a space. Specifically, we measure the reconstruction error for different visual attributes and provide the reconstruction visualizations and corresponding numbers in Figure 5.4.

Interestingly, we find that the reconstruction metric, mean absolution error (MAE), for other visual attributes is much lower than in the RGB space. For example, 0.0038 in binary mask reconstruction versus 0.0356 in RGB space – a degree of magnitude smaller. The reason we think is that, for most cases, visual attributes contain less information or structure compared to the corresponding RGB images, which makes the reconstruction task itself much easier. Nevertheless, this result confirms that the pretrained VAE is able to provide a robust shared latent space for multiple visual attributes, therefore enabling our design of a unified visual generation model.


Type 1: perception

Type 2: restoration

Type 3: editing

Type 4: text-to-image generation

"To Abstract Expressionism"
"Cover the cap with the rag"
"Change the cat to a mouse"
Go with the flow
Achievement unlocked: Diffusion models can spell now!

Figure 5.5: Visualizations of the four types of tasks we considered: text-to-image generation, instruction-guided editing, perception tasks, and image restoration tasks

5.3.2 Task Groups

In this section and the next, we consider a group of four types of tasks: text-to-image generation, instruction-guided editing, perception tasks, and image restoration tasks. We show visualizations for them in Figure 5.5 and detail the source and volume of them in Table 5.1.

5.3.3 Qualitative Results on Computer Vision Tasks

We evaluate our unified model on a variety of computer vision tasks that can be represented in the latent image domain. We consider the same group of perception tasks detailed in Table 5.1.

Visualizations are shown in Figure 5.6 Across these tasks, the model produces coherent outputs with strong spatial consistency. Modalities with global structure, such as depth or normals, benefit from the diffusion backbone's holistic reasoning, while tasks with localized boundaries (*e.g.*, masks) leverage cross-task generalization

Task	Dataset	Data Volume	Annotation
T2I	Laion-coco-aesthetics	4.1m	CLIP captions
Depth estimation	Graph200k	205k	Depth anything
Pose estimation	Graph200k	205k	Open-pose
Normal prediction	Graph200k	205k	DSINE
Edge detection	Graph200k	205k	Canny
Referring segmentation	RefCOCO	320k	Human label
Style transfer	Stylebooth	11k	De-stylize and restyle
Long-prompt editing	OmniEdit	740k	Mixed expert models
Short-prompt editing	OmniEdit	740k	Mixed expert models
Editing - change	OmniEdit	410k	Mixed expert models
Colorization	Graph200k	205k	RGB2Gray
Denoising	Graph200k	205k	10 random noise
Inpainting	Graph200k	205k	Image masking
Derain	Graph200k	205k	Raindrop synthesis
Low-light enhancement	Graph200k	205k	Physics-based methods

Table 5.1: Details of the data we used for training a diffusion generalist model.

built during training.

Importantly, all tasks share a single model, illustrating that diverse vision problems can indeed be cast as conditional image-generation tasks within one unified architecture.

5.3.4 Results on Editing, Restoration, and Non-CV Tasks

We further evaluate the same diffusion backbone on restoration tasks (denoising, deblurring, recoloring, relighting), as well as local and global image editing. Visualizations are provided in Figures 5.7 and 5.8.

With the support of the MLLM encoder, the model successfully interprets complex instructions and produces visually aligned outputs. Tasks that require both semantic understanding and fine-grained pixel manipulation—such as removing objects or modifying specific regions—show substantial improvement compared to text-only conditioning.

These results highlight the generalization capability of the unified multimodal pipeline across both traditional CV tasks and more expressive editing tasks.

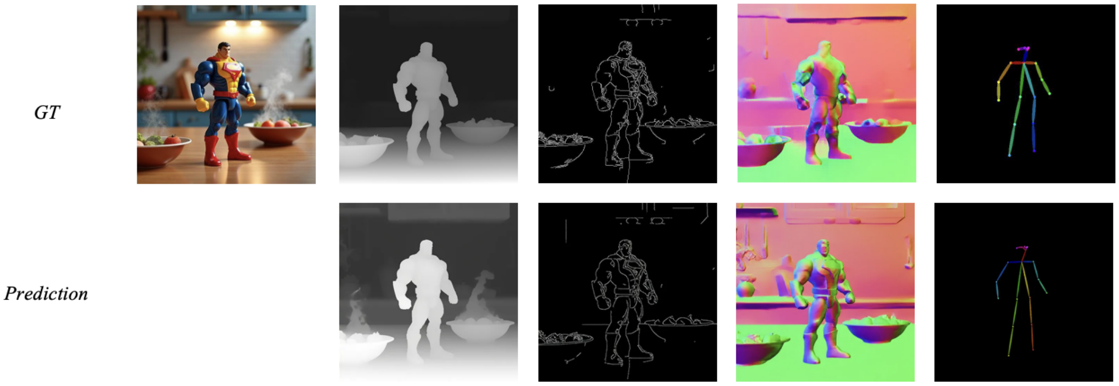


Figure 5.6: Visualizations for our model on perception tasks. Across these tasks, the model produces coherent outputs with strong spatial consistency.

5.4 Latency Analysis

Despite the broad task coverage and high-quality outputs, the major limitation of this unified diffusion model is inference latency. Empirically, using a configuration combining Qwen-VL-3B and a Flux diffusion model, generating a single image typically requires approximately 20 seconds. For a production-grade MLLM-driven system similar to state-of-the-art image generation models, the inference time can exceed **one minute** per image.

Such latency renders diffusion-based generalist models impractical for interactive applications, robotics pipelines, or high-throughput multimodal systems. This limitation motivates the next chapter, where we explore autoregressive generative modeling as a more scalable, efficient, and flexible alternative, culminating in the UniGen-AR design.

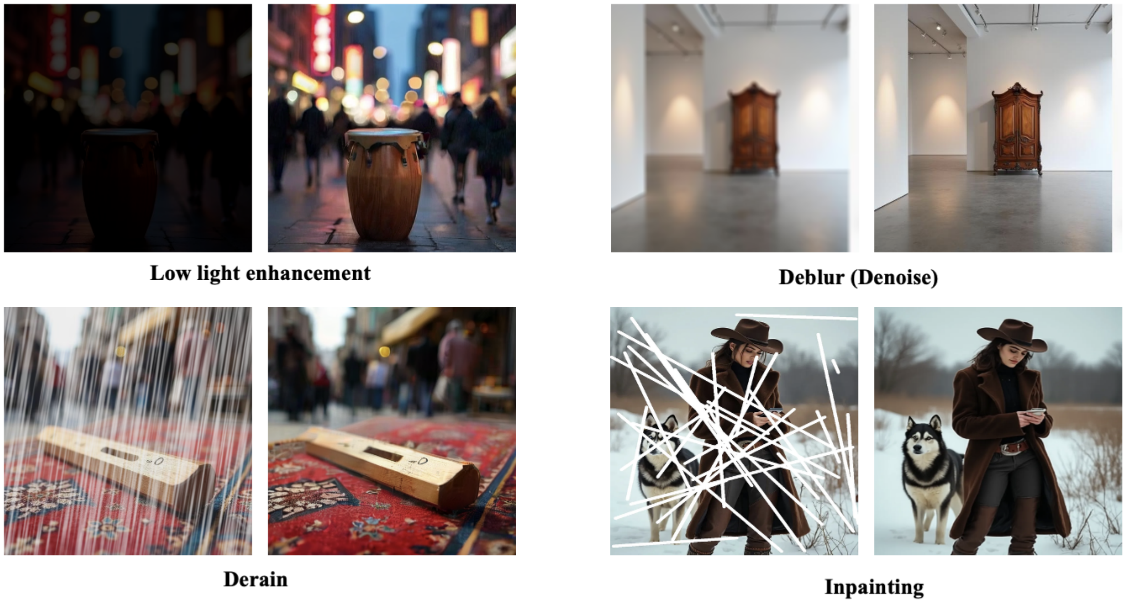


Figure 5.7: Visualizations for our model on restoration tasks. Our model successfully restores the destroyed details of the input image.

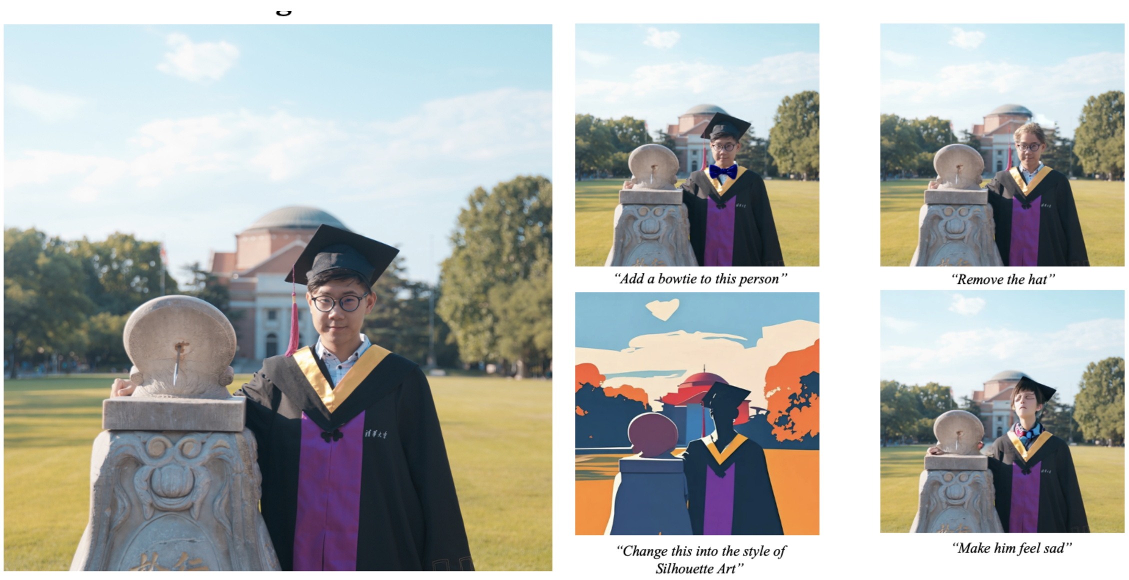

"Add a bowtie to this person"
"Remove the hat"
"Change this into the style of Silhouette Art"
"Make him feel sad"

Figure 5.8: Visualizations for our model on restoration tasks. With the support of the MLLM encoder, the model successfully interprets complex instructions and produces visually aligned outputs.

Chapter 6

UniGen-AR: Unifying Visual Generation with Auto-Regressive Modeling

6.1 Introduction

Modern visual generation pipelines remain fragmented: text-to-image synthesis, local and global editing, restoration, and classical vision tasks – such as depth and surface normal estimation and semantic segmentation – are typically handled by separate models or specialized heads [6, 123, 185, 287, 293]. In this work, we study *Unified Visual Generation* (UVG) [88, 116, 240], a setting in which a single model produces diverse *image-valued outputs* under a shared interface. UVG promises shared representations, consistent controllability, and operational simplicity. However, scaling a single model to support many output types and control modalities is non-trivial: as the catalog of tasks and outputs grows, the model must learn to handle heterogeneous visual domains and control signals while maintaining coherent behavior across all their combinations. This imposes a combinatorial burden and severe capacity and optimization challenges, making scalability the central obstacle for effective UVG.

To tackle this scaling challenge, current approaches to UVG are dominated by diffusion-based architectures [79] paired with powerful multimodal encoders [7]. This

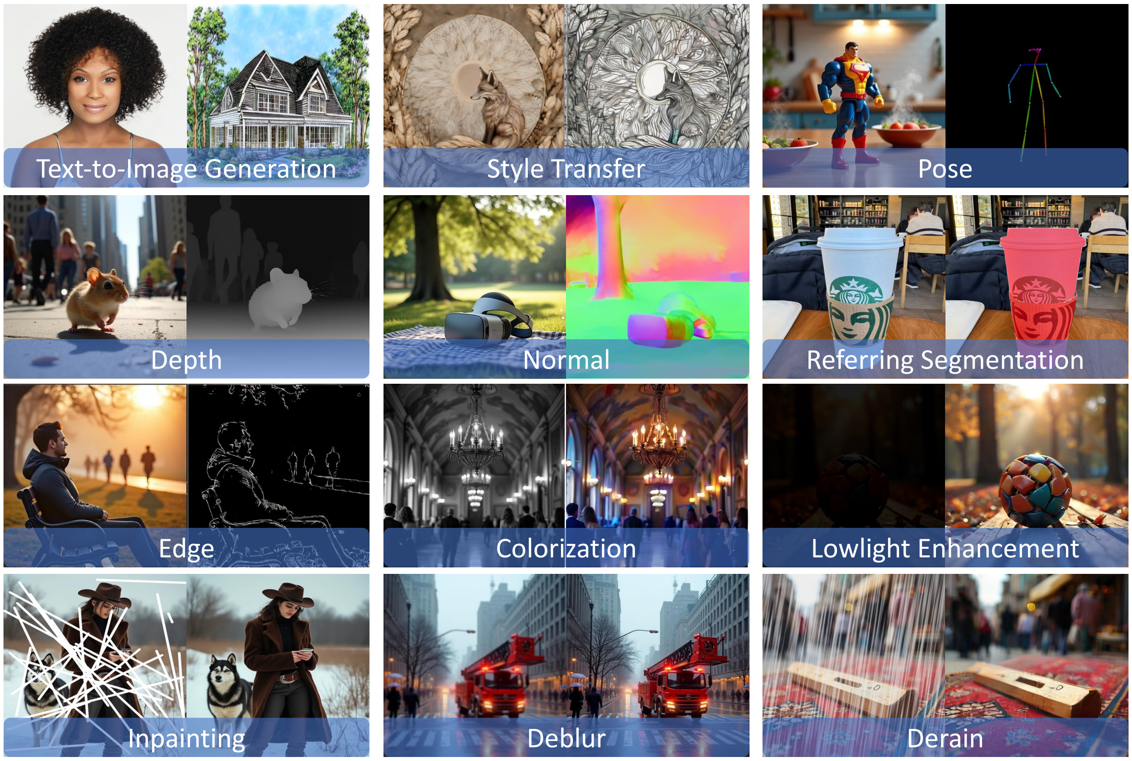


Figure 6.1: **UniGen-AR: A Single Model for Unified Visual Generation.** Our framework jointly handles 12 diverse tasks, spanning text-to-image synthesis, restoration, and classical perception. All outputs are generated by a single, MLLM-conditioned auto-regressive backbone, using a unified prompting interface and a single set of model weights without any task-specific heads.

bundled design directly addresses the combinatorial challenge: the MLLM acts as a universal interface to interpret heterogeneous control signals (*e.g.*, text, reference images), while the diffusion backbone learns a shared generative representation for the diverse image-valued outputs. These models, *e.g.*, Batifol et al. [17], Comanici et al. [39], Wu et al. [240], set a high standard for output quality and task coverage. However, they incur significant computational costs: iterative denoising, bundled with MLLM inference, yields significant inference latency, and the global nature of the denoising process often leads to unintended spurious edits [142]. These limitations motivate the search for alternative architectures that retain rich conditioning while offering a better **latency–quality trade-off**.

To this end, we revisit *auto-regressive* (AR) modeling as an alternative backbone for UVG. While naïve AR over coarse image tokens often struggles with fine-grained detail, recent work on *visual auto-regressive* (VAR) modeling demonstrates that *next-scale prediction* over discrete visual tokens can achieve high-fidelity synthesis, stable likelihood-based training, and efficient sampling [217]. In the context of UVG, VAR presents two key advantages: (i) *latent unification* – similar to diffusion, VAR decoders can produce both natural images and structured predictions, like depth maps, within a shared token space; and (ii) *sampling efficiency* – AR models typically require significantly fewer steps than diffusion to achieve comparable perceptual quality [67, 217].

The VAR backbone provides sampling efficiency, but to achieve the rich, instruction-based controllability required for UVG [116, 240], it must be paired with a powerful multimodal front-end. We therefore propose **UniGen-AR**, a framework that couples a general-purpose multi-modal language model (MLLM) with a visual auto-regressive decoder. This architecture is designed to retain diffusion-style flexibility while benefiting from AR efficiency. In our framework, the MLLM encodes free-form instructions and diverse control signals (*e.g.*, text, reference images) into a unified conditioning sequence. The auto-regressive decoder then predicts discrete visual tokens conditioned on this sequence, which are decoded into the final image or dense map output via a VQ-VAE [62].

To evaluate our proposed framework, we instantiate UniGen-AR by re-purposing a powerful pre-trained *text-to-image* VAR model [67] for the full UVG setting. We train a single backbone jointly across *12 tasks* spanning three families: text-to-image generation, classic perception, and restoration. Typical visualizations are shown in Figure 6.1. Training is performed with a unified likelihood objective under the consistent MLLM-conditioned interface. Empirically, UniGen-AR achieves strong performance across all 12 tasks, outperforming prior AR-based systems and exhibiting especially strong results on restoration and classical perception. Compared to diffusion-based UVG models under matched conditioning, UniGen-AR presents a favorable latency–quality Pareto frontier, achieving up to $\sim 5\times$ lower inference latency while maintaining or improving output quality on representative benchmarks. Ablation studies further reveal the design of the VQ-VAE tokenizer, particularly codebook size and hierarchy, as a critical factor influencing performance at scale.

Our contributions are summarized as follows:

- We present, to our knowledge, the first framework that scales visual auto-regressive modeling to the full UVG setting, unifying open-ended synthesis, restoration, and visual perception within a single image-out backbone.
- We demonstrate a compelling latency–quality trade-off compared to diffusion-based systems, achieving consistent speedups while maintaining or improving performance.
- We identify and validate the importance of VQ-VAE tokenizer design, including codebook size and hierarchy, as a key driver of VAR scalability and effectiveness.
- We study the bidirectional connection between multimodal understanding and generation, showcasing how understanding-enhanced MLLM front-ends improve control and quality in image synthesis.

6.2 Related Work

Unified visual generation aims to support tasks such as text-to-image generation, editing, restoration, and perception within a single model. Early methods explored shared latent spaces using variational autoencoders (VAEs) and vision transformers [101, 133, 134]. More recent approaches have adopted diffusion-based models, pretrained on large-scale data, to address a wide range of generative tasks under a unified interface [23, 52, 68, 115, 118, 242, 250, 283]. As a follow-up, another line of work builds MLLM-mediated pipelines for controllable image generation [17, 116, 123, 162, 240]. For example, Qwen-Image [240], Step1X-Edit [123], and MetaQueries [162] pair powerful multimodal front-ends with diffusion decoders to support instruction-based rendering and precise editing. Flux-Kontext [17] and VisualCloze [116] focus specifically on in-context learning, enabling models to follow few-shot examples for visual tasks.

A second strand builds MLLM-mediated pipelines for controllable generation [17, 116, 123, 162, 240]. Among them, Qwen-Image [240], Step1X-Edit [123], and Metaquryries [162] pair powerful multimodal front-ends with diffusion decoders for text rendering and precise editing. Flux-Kontext [17] and Visualcloze [116] specifically focus on in-context learning to enable models with the capability to learn from few-shot examples.

A smaller yet growing body of work investigates AR modeling for image generation [8, 57, 109, 207, 208, 209, 214]. These models typically adopt the standard "next-token prediction" formulation. In contrast, our work builds on the "next-scale prediction" paradigm introduced in visual autoregressive (VAR) models [217], which is better suited for UVG due to its coarse-to-fine decoding strategy.

Visual auto-regressive models factorize image generation into scale-wise predictions over discrete visual tokens, allowing coarse-to-fine decoding. The foundational work of Tian et al. [217] demonstrates favorable scaling laws and superior latency–quality trade-offs compared to diffusion models. Subsequent studies extend this formulation to conditional image generation in different domains beyond ImageNet [34, 114, 139, 176, 196, 229, 297].

Among them, two recent approaches adapt next-scale VAR to text-to-image generation [67, 226]. Both adopt cross-attention modules to inject text signals into the visual decoder, following a design similar to Stable Diffusion [185]. Switti [226] introduces a refined attention masking strategy that restricts each token to attend only to spatially local neighbors within the current scale, improving inference speed. Infinity [67] identifies large codebook sizes in VQ-VAE as key to achieving high-quality synthesis. A few recent works have begun extending VAR beyond text-to-image to support editing [142, 228, 235]. However, EditInfinity [228] and related methods [235] do not support direct instruction-guided UVG. Instead, they rely on indirect mechanisms such as modifying attention maps or text embeddings to steer edits. The concurrent VAREdit [142] adapts VAR models for editing, but focuses solely on this task and does not incorporate MLLMs or address broader UVG settings. In contrast, we present the first framework that combines next-scale VAR modeling with MLLM-based conditioning to support full-spectrum UVG tasks.

Unified models aim to handle both understanding (text-out) and generation (image-out) in a single architecture. One direction pursues tightly-coupled token-based models [94, 95, 209, 211, 216, 234, 248]. Chameleon [216] pioneered early-fusion, any-order modeling over text and image tokens in a single Transformer. Emu3 [234] extends this approach to support both understanding and generation over images and videos, using next-token prediction on discrete tokens. LaVIT [94] and its successors integrate LLMs with discrete visual tokenizers to perform both perception and synthesis tasks under a unified generative interface.

A second direction explores hybrid designs that decouple encoding and decoding while maintaining a single interface. These approaches pair MLLM front-ends with task-specific decoders [43, 86, 103, 161, 241, 276], often sharing a central autoregressive core to support both instruction following (text-out) and controllable generation or editing (image-out). This design balances task flexibility with operational simplicity. Our framework follows this hybrid philosophy. By coupling an MLLM front-end with a next-scale VAR decoder, we enable instruction-conditioned image generation across a wide task spectrum. Preliminary results also indicate that jointly fine-tuning the MLLM and visual decoder improves alignment between vision and language representations, suggesting a promising path toward fully unified multimodal models, which we leave for future exploration.

6.3 Method

This section presents the design of UniGen-AR (illustrated in Figure 6.2). We begin by reviewing VAR modeling as the core generative backbone. We then detail how we perform UVG based on existing T2I VAR models.

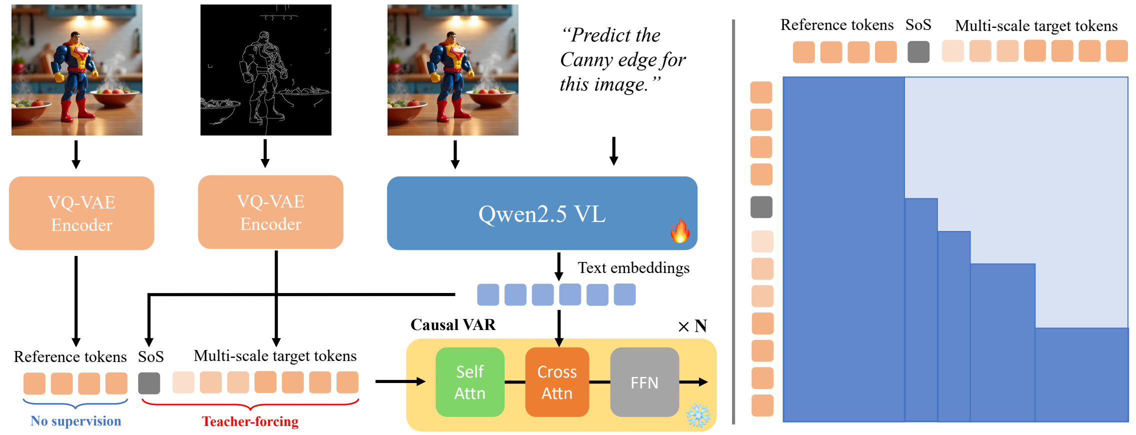


Figure 6.2: **Architecture of UniGen-AR.** Left: We extend an Infinity-style VAR backbone with a multimodal language model encoder to support Unified Visual Generation with reference images. The MLLM encodes the instruction and reference image, whose text embeddings seed a learnable [SoS] token and provide keys/values for cross-attention, while finest-scale reference tokens are prepended as a non-predictive context prefix; Right: A block-wise causal mask lets all target tokens attend to reference and text context while preserving the standard coarse-to-fine VAR schedule.

6.3.1 Preliminary: Visual Auto-Regressive Modeling

Vanilla VAR models. VAR [217] generates high-fidelity images by autoregressively predicting discrete visual tokens across multiple *spatial scales*. It operates in a latent space defined by a multi-scale vector-quantized tokenizer, typically implemented via a VQ-VAE [101, 223].

Given an image I, the encoder $\mathcal{E}$ produces K token maps:$R = \mathcal{E}(I) = (r_1, r_2, \dots, r_K), \quad r_k \in [V]^{h_k \times w_k}$, where V is the codebook size, and spatial resolution increases with scale index k (*i.e.*, $h_1 w_1 \leq \cdots \leq h_K w_K$). The decoder $\mathcal{D}$ reconstructs the image via: $\hat{I} = \mathcal{D}(r_1, \dots, r_K)$.

With the training multi-scale token sequence R, VAR defines an autoregressive factorization across scales:

$$p_\theta(R) = \prod_{k=1}^{K} p_\theta(r_k | r_{<k}), \tag{6.1}$$

where $r_{<k}$ denotes tokens from coarser scales. Each r_k is predicted in parallel, conditioned on $r_{<k}$ and scale-specific position embeddings. A block-wise causal mask

ensures each token in r_k only attends to tokens in $r_{\leq k}$. Inference proceeds sequentially from $k = 1$ to K, with key-value caching for efficiency.

The model is trained with teacher forcing using the sum of cross-entropy losses over all token positions and scales:

$$\mathcal{L}_{\text{VAR}} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \text{CE}\Big(p_\theta(\cdot \mid r_{<k}),\ r_k^{(i)}\Big)\,, \tag{6.2}$$

where $n_k = h_k w_k$ is the number of tokens at scale k, and $r_k^{(i)}$ is the ground-truth token (from $\mathcal{E}$) at position i.

Text-conditioned next-scale prediction. To enable text-to-image generation, the VAR factorization in Equation (6.1) is extended to condition on a prompt t, encoded by a frozen language model $\psi(t)$:

$$p_\theta(R) = \prod_{k=1}^{K} p_\theta(r_k|_{<k}, \psi(t)). \tag{6.3}$$

Following prior work [67, 226], the text embedding $\psi(t)$ is injected at each transformer layer via cross-attention. A learnable start-of-sequence ([SoS]) token, derived from a projection of $\psi(t)$, is prepended at the coarsest scale to bootstrap generation. Training remains unchanged, using block-wise causal masking and the loss in Equation (6.2).

Infinity [67], a state-of-the-art T2I VAR model serving as our backbone VAR model in our main experiments, replaces the standard categorical tokenizer with a *bitwise* tokenizer that encodes each visual token as a binary vector. Instead of predicting a single index, Infinity predicts the bits of this code, so that increasing the bit-width enlarges the effective vocabulary exponentially while only mildly growing the classifier head. This binary indexing enables extremely large visual vocabularies and thus higher-fidelity reconstruction and richer visual details. The bitwise design remains compatible with cross-entropy-style training, implemented as independent binary cross-entropy losses over bits. At inference time, Infinity conditions on the text prompt t and autoregressively generates the multi-scale residual token maps $(r_1, \ldots, r_K)$ with cached key–value attention states, following the VAR coarse-to-fine schedule.

6.3.2 From Text-to-Image to Unified Visual Generation

We use Infinity [67] as our T2I backbone and extend its architecture to UVG with reference images. To enable image-to-image transformations, the model must first understand the input images; we therefore introduce a multimodal encoder that processes reference images (and text) into a shared token space. Our overall design is illustrated in Figure 6.2.

Reference- and target-tokenization. Given a reference image I^{ref} and a target image I^{tar}, we first encode them with the same multi-scale VQ-VAE encoder $\mathcal{E}$:

$$R^{\text{ref}} = \mathcal{E}(I^{\text{ref}}), \quad R^{\text{tar}} = \mathcal{E}(I^{\text{tar}}). \tag{6.4}$$

For the target image, we keep all scales $R^{\text{tar}} = (r_1^{\text{tar}}, \ldots, r_K^{\text{tar}})$ as in vanilla Infinity. For the reference image, we only retain the finest scale r_K^{ref}, and discard coarser scales:

$$R^{\text{ref}} = (r_K^{\text{ref}}), \quad r_K^{\text{ref}} \in [V]^{h_K \times w_K}. \tag{6.5}$$

The concurrent work, EditVAR [142], also demonstrates that finest-scale tokens are sufficient for UVG tasks. This design, prepending the finest reference tokens, leaves the original VAR scale schedule unchanged: the reference tokens are never traversed by the next-scale generation process and are excluded from the loss.

Multimodal encoder. We replace the Infinity's text encoder, T5 [38], with a multimodal language model ψ (Qwen2.5-VL [7]) to encode the input instruction t, together with the reference image: $Z_t = \phi(I^{\text{ref}}, t)$. Similar to prior efforts in the Diffusion regime [240], Z_t are the *pure text embeddings* from the last self-attention layer of the MLLM encoder to remove the redundancy. These text embeddings are used in two ways, following Infinity [67]: a learnable [SoS] token is obtained by projecting the text embedding and prepended at the coarsest scale to bootstrap generation, and the text embedding also serves as the key and value sequence for the cross-attention layers that modulate the visual tokens.

Unified token sequence and causal masking. We unify reference and target tokens into a single autoregressive sequence

$$\mathbf{s} = \left(r^{\text{ref}},\ [\text{SoS}],\ r_1^{\text{tar}}, \ldots, r_K^{\text{tar}}\right), \tag{6.6}$$

where the reference tokens are *prepended* before the [SoS] token. Conceptually, r^{ref} acts as a non-predictive context prefix, the [SoS] token marks the autoregressive start, and the multi-scale target tokens follow the standard VAR schedule.

We implement a block-wise causal mask $M \in \{0,1\}^{|\mathbf{s}|\times|\mathbf{s}|}$, visualized on the right side of Figure 6.2, that enforces:

- reference tokens are visible to all subsequent tokens ([SoS] and target tokens) but are never used as prediction targets;
- the [SoS] token and all target tokens respect the original next-scale causal ordering: tokens in scale k can attend to reference tokens, [SoS], and all tokens in $r^{\text{tar}}_{<k}$, but not to future scales.

Training and inference. Training follows teacher forcing as in Equation (6.2). At inference time, we perform iterative next-scale prediction as in standard VAR: we first sample r^{tar}_1 conditioned on $\mathbf{r}^{\text{ref}}$ and t, then proceed to finer scales until r^{tar}_K is obtained. Due to the causal mask, tokens at each step can freely attend to the entire reference sequence and the text context while respecting the multi-scale ordering. When no reference image is provided, the same decoding procedure reduces to conventional T2I generation.

6.4 Experimental Evaluations

6.4.1 Experimental Setup

Training data. Following prior work [116, 237, 250], we train UniGen-AR using publicly available paired datasets. For text-to-image (T2I) generation, we use the LAION-COCO-Aesthetic subset [193, 250], containing approximately 4M images. For perception tasks, we adopt the Graph200K dataset from VisualCloze [116], which provides 200K images paired with annotations for depth estimation, surface normals, edge detection, and human pose estimation. For image restoration, we consider five tasks: deblurring, deraining, colorization, inpainting, and low-light enhancement. We follow the processing scheme of VisualCloze to generate the noisy version of these labels. For referring image segmentation, we use RefCOCO [269], which contributes roughly 320K (image, mask, text) triplets. For style transfer, we

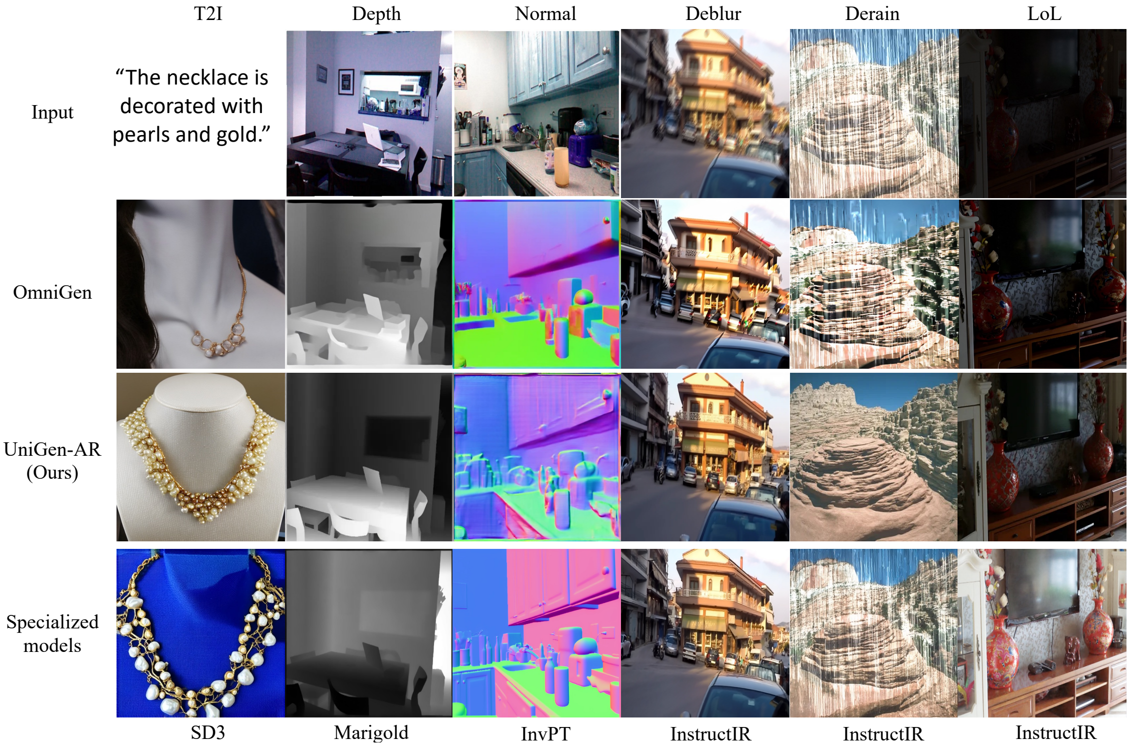


Figure 6.3: **Qualitative comparisons across UVG tasks.** UniGen-AR consistently produces higher-quality results than the UVG baseline OmniGen and achieves visually comparable or superior outputs to specialized models, notably outperforming InstructIR [40] on deraining.

train on StyleBooth [69], containing around 11K images. In total, our training corpus comprises roughly 6M paired examples across 12 UVG tasks.

Implementation details. We initialize our system from the pretrained Infinity-2B model [67]. We replace its original T5 [38] text encoder with Qwen2.5-VL (3B) [7] for multimodal conditioning. Following [240], we extract only the textual embeddings from the MLLM and omit visual embeddings. Training proceeds in two stages: **Stage I (alignment).** We freeze the entire Qwen2.5-VL and Infinity backbones, except for the text-normalization layer, text-projection layer, and unconditional embeddings used for classifier-free guidance. This aligns the pretrained Infinity decoder with the new Qwen2.5-VL conditioning. In this stage, we only train the model with the T2I data. **Stage II (UVG training).** We jointly train the Infinity backbone components on the full mixture of 12 tasks. Each batch is either a pure T2I batch or a mixed batch drawn from all other tasks, with the T2I sampling probability set to 0.25. All

Model	# Params	GenEval			
		Two Obj.	Position	Color Attr.	Overall
SDv1.5 [185]	0.9B	0.38	0.04	0.06	0.43
SDv2.1 [185]	0.9B	0.51	0.07	0.17	0.50
DALL-E 2 [179]	6.5B	0.66	0.10	0.19	0.52
DALL-E 3 [19]	-	-	-	-	0.67
SDXL [169]	2.6B	0.74	0.15	0.23	0.55
SD3 (d=21) [50]	2B	0.74	0.34	0.36	0.62
LlamaGen [206]	0.8B	0.34	0.07	0.04	0.32
Chameleon [216]	7B	-	-	-	0.39
Emu3 [234]	8.5B	0.81	**0.49**	0.45	0.66
Infinity [67]	2B	**0.85**	**0.49**	**0.57**	**0.73**
UniGen-AR (Ours)	2B	0.76	0.41	0.45	0.68

Table 6.1: **GenEval Text-to-Image Results.** Comparison with diffusion-based models (top) and autoregressive models (bottom). UniGen-AR achieves competitive performance while using only public data and supporting 12 unified visual generation tasks.

outputs are generated at a fixed resolution of 512 × 512.

Stage I is trained for 2 epochs with an effective batch size of 256; Stage II is trained for 100K steps with an effective batch size of 128. We use AdamW [130] with learning rates of 1e-4 (Stage I) and 5e-6 (Stage II). Training requires approximately 3 days for Stage I and 7 days for Stage II on a single NVIDIA H100 node.

Evaluation benchmarks. We evaluate on three groups of tasks. **T2I generation:** GenEval benchmark [58], following Infinity [67]. **Perception tasks:** depth estimation and surface normals on NYUv2 [198]. **Restoration tasks:** low-light enhancement on LOL [236], deblurring on GoPro [151], and deraining on Rain-13K [53].

Qualitative comparisons for all tasks appear in Figures 6.1 and 6.3, with additional examples provided in the supplementary. More details about our training data and implementation are also included in the supplementary.

6.4.2 Main results

We report results on T2I generation (Table 6.1), perception tasks, and image restoration (Table 6.2). Representative outputs are shown in Figure 6.3. For T2I, we

Model	NYUv2-Depth	NYUv2-Normal	LOL-Lowlight		GoPro-Deblur		Rain100L-Derain	
	RMSE (↓)	Mean Angle Err ↓	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)
Depth Anything [261]	**0.206**	-	-	-	-	-	-	-
Marigold [97]	0.224	-	-	-	-	-	-	-
Bae *et al.* [5]	-	**14.90**	-	-	-	-	-	-
InvPT [263]	-	19.04	-	-	-	-	-	-
AirNet [112]	-	-	18.18	0.735	24.35	0.781	32.98	0.951
InstructIR [40]	-	-	**23.00**	**0.836**	**29.40**	**0.886**	**36.84**	**0.937**
InstructCV [54]	0.297	-	-	-	-	-	-	-
UnifiedIO [133]	0.387	-	-	-	-	-	-	-
OmniGen [250]	0.480	-	13.38	0.392	13.39	0.321	12.02	0.233
X-Prompt [209]	0.277	19.17	19.71	0.810	21.04	0.761	25.53	0.843
UniGen-AR (Ours)	**0.245**	**18.76**	**21.03**	**0.825**	**22.99**	**0.774**	**33.71**	**0.926**

Table 6.2: **Results on Perception and Restoration Tasks.** UniGen-AR significantly outperforms AR-based UVG prior work (X-Prompt) and demonstrates competitive performance on image restoration tasks.

compare against both diffusion-based and autoregressive models; for perception and restoration tasks, we compare with specialized task models and recent UVG systems.

Text-to-image generation. Table 6.1 shows that: (1) UniGen-AR achieves strong performance on GenEval, outperforming larger diffusion-based models such as DALL-E 2 [179], demonstrating that next-scale VAR remains competitive even in the unified setting. (2) Compared with the Infinity checkpoint, our model shows a mild drop in performance, likely due to Infinity's use of large-scale proprietary training data. (3) Despite using only public data, our model surpasses diffusion counterparts with similar model sizes, including SD3 (2B) [50]. This suggests that VAR-based backbones retain strong prior knowledge during finetuning and remain a compelling alternative to diffusion for controllable image generation.

Perception and image restoration tasks. Based on the results in Table 6.2, we offer the following observations: (1) UniGen-AR consistently outperforms the strongest AR-based UVG model X-Prompt [209] across all evaluated tasks, highlighting the advantage of coarse-to-fine refinement in next-scale prediction. (2) Compared with specialized task-specific models, a performance gap remains—reflecting the inherent challenge of UVG, where a single model must master diverse, heterogeneous objectives. (3) Notably, our model showcases superior performance on the restoration tasks. In particular, for low-light enhancement and derain tasks, our model surpasses a dedicated restoration model (AirNet [112]), suggesting that the bitwise VQ-VAE used in Infinity provides a favorable structure for correcting token-level degradations.

Qualitative comparisons. Figure 6.3 illustrates that UniGen-AR produces higher-

Model	Two Obj.	Position	Color Attr.	Overall	inf. time (s/img)
SD3 (d=21)	0.74	0.34	0.36	0.62	3.18
Infinity	0.85	0.49	0.57	0.73	0.92
Qwen+SD3	0.68	0.36	0.29	0.52	5.23
Qwen+Infinity	0.75	0.45	0.43	0.64	1.05

Table 6.3: **Diffusion v.s. VAR decoders.** We compare SD3 (diffusion) and Infinity (VAR) under identical finetuning settings. VAR provides better generation accuracy and is substantially faster at inference.

quality results than the UVG baseline OmniGen [250] across all evaluated tasks. Moreover, for most tasks, our outputs are visually comparable to those of specialized models. Notably, on the deraining task, UniGen-AR achieves cleaner rain removal than the state-of-the-art dedicated model InstructIR [40]. These qualitative results further highlight the strength of our MLLM–VAR architecture for UVG and suggest promising potential for real-world applications.

6.4.3 Ablation Study

Diffusion v.s. VAR. We first compare the impact of the decoder architecture by replacing the Infinity VAR decoder with the SD3 diffusion decoder, while keeping all other training settings (*i.e.*, data, steps, resolution) identical. For efficiency, all variants are finetuned for two epochs on the T2I subset only; therefore, absolute numbers differ from Table 6.1. Results are summarized in Table 6.3.

Both models exhibit performance drops relative to their original checkpoints, primarily due to the substantially smaller and purely public finetuning data. Nevertheless, under this controlled setting, the VAR-based Infinity decoder consistently outperforms the diffusion-based SD3 decoder across all GenEval categories. This highlights next-scale prediction as a robust alternative to diffusion when finetuned jointly with a multimodal encoder. Interestingly, integrating Qwen2.5-VL improves the spatial grounding capability of the SD3 variant – its Position score increases from 0.34 to 0.36 with poorer training data, showcasing the benefit of replacing a text-only encoder with an MLLM for UVG. Finally, the Infinity variants achieve approximately **5× faster** inference than their SD3 counterparts, underscoring the practical appeal

Model	LOL-Lowlight PSNR (↑)	GoPro-Deblur PSNR (↑)	Rain100L-Derain PSNR (↑)
Infinity + T5	20.10	22.17	29.32
Infinity + Qwen	**21.03**	**22.99**	**29.71**

Table 6.4: **Effect of multimodal encoder.** Replacing T5 with Qwen2.5-VL leads to significant improvements on all restoration tasks, indicating the benefit of multimodal grounding.

of VAR for real-time or interactive generation workloads.

Choice of multimodal encoder. To evaluate the value of multimodal conditioning, we compare Qwen2.5-VL [7] with the original T5 [38] encoder used in Infinity. Both variants are trained on the same data and with the same schedule. Table 6.4 reports results on three image restoration tasks. The Qwen-based model achieves notably higher PSNR across all tasks. This suggests that sending both the reference image and the instruction prompt into an MLLM yields text embeddings that implicitly encode object- and region-level semantics, which are more informative than the purely linguistic embeddings produced by T5. As UVG tasks often require localized reasoning, this multimodal grounding becomes particularly beneficial.

Impact of visual tokenizer. We further study the influence of the discrete visual tokenizer by training three variants using VQ-VAEs from VAR [217], Switti [226], and Infinity [67]. Different from the core experiment, here we finetune the checkpoint from the **vanilla VAR model** All models are trained in a single-stage T2I setting for two epochs. Figure 6.4 shows reconstruction FID (rFID) on ImageNet [44, 77] and GenEval [58] performance. We observe a strong inverse correlation between rFID and generation quality: tokenizers with lower reconstruction error yield higher GenEval scores. This underscores the visual tokenizer as a central bottleneck in VAR-style architectures. Improving token expressiveness and reconstruction fidelity remains a promising direction for advancing VAR generation.

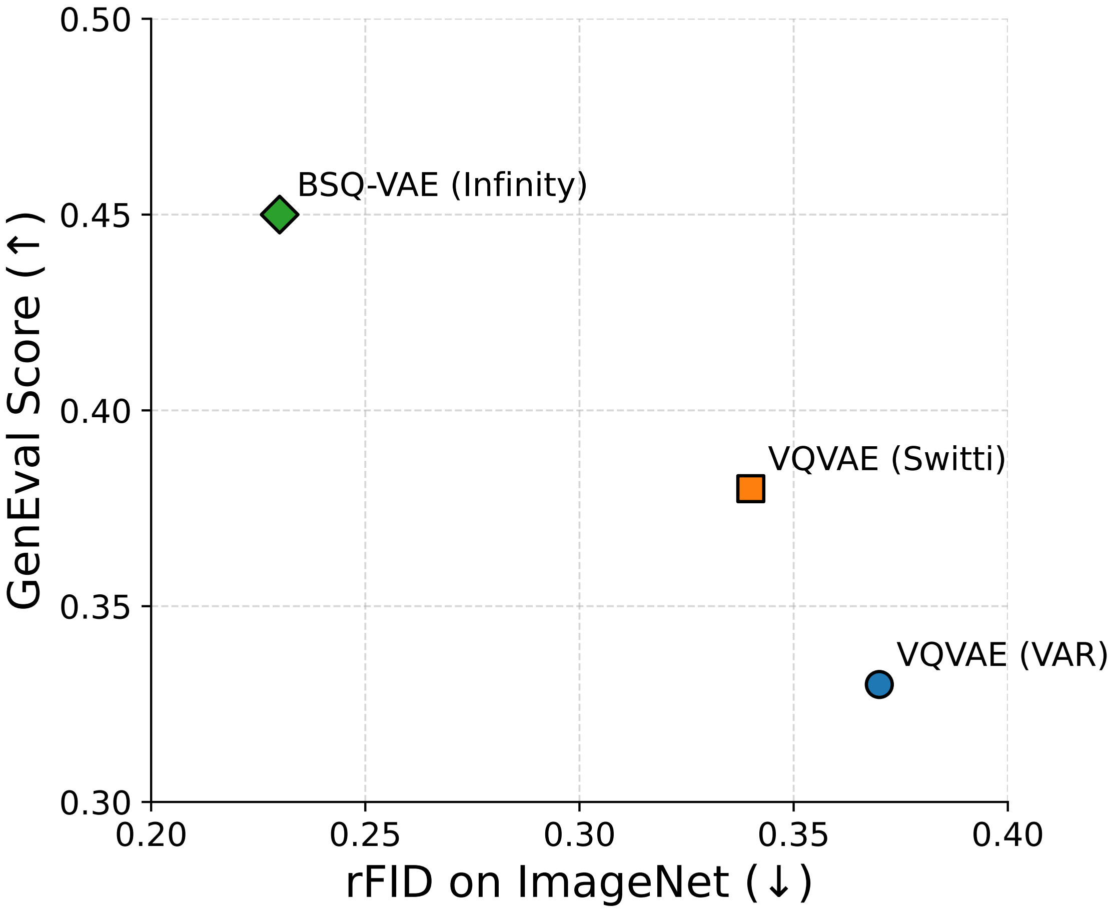


Figure 6.4: **Impact of visual tokenizers.** Better reconstruction fidelity leads to substantially improved generation quality, highlighting the tokenizer as a key design factor in VAR-based models.

6.4.4 Multimodel Understanding and Unified Visual Generation

Thus far, we have focused on how an MLLM can provide stronger conditioning signals for the VAR backbone. In this experiment, we take a slightly different perspective: treating UniGen-AR as a visual-output branch of an MLLM, we investigate whether improving the MLLM's understanding ability can, in turn, enhance its visual generation.

To control for data and isolate the effect, we exclusively reuse the same T2I subset employed for training the generator. Each T2I sample is repurposed into a VQA-style

Setting	Two Obj.	Position	Color Attr.	Overall
Ours ((UVG only))	0.75	0.45	0.43	0.64
Ours (Joint MMU+UVG)	**0.82**	**0.47**	**0.46**	**0.69**

Table 6.5: **Multimodal understanding improves unified visual generation.** Jointly finetuning Qwen2.5-VL for multimodal understanding leads to consistent performance gains on GenEval, especially for multi-object reasoning, highlighting the practicality of coupling understanding with generation.

instance by converting the caption into an answer and assigning a fixed question template: *"Generate a caption for this image."* To increase linguistic diversity, we follow and pre-sample 50 paraphrased variants of this question via Qwen2.5-VL itself for training. During joint training, we finetune the last 10 layers of Qwen2.5-VL together with the Infinity decoder. The multimodal-understanding loss updates only the Qwen layers, whereas the UVG loss updates both Qwen and Infinity in a coupled manner. All variants are trained for two epochs on T2I data for fair comparison.

Table 6.5 reports the results. We observe that jointly training for multimodal understanding consistently improves T2I generation quality. The gains are especially pronounced for the *Two Objects* category in GenEval, which requires resolving relationships across multiple entities—an ability naturally strengthened by the auxiliary understanding objective. These findings suggest that multimodal understanding and multimodal generation are mutually beneficial: enhancing the semantic reasoning capability of the MLLM leads to improved visual generation fidelity. This synergy points toward a promising direction for future unified architectures that treat understanding and generation as tightly coupled objectives rather than isolated tasks.

6.5 Limitation and Future Work

Limitation. A key limitation of our current design lies in its fixed output resolution of 512 × 512. While this choice simplifies training across heterogeneous tasks, it prevents UniGen-AR from flexibly adapting to inputs of arbitrary size. Adopting the dynamic-resolution strategies used in Stable Diffusion [185] and Infinity [67], *e.g.*, multiple groups of resolution choices and spatial padding, represents a practical next step toward broad deployment.

We have three **future research directions** following the current work: First, extending UniGen-AR to additionally handle editing tasks remains an important avenue, especially those requiring fine-grained, spatially localized modifications. Second, inspired by recent advances in MLLMs [120, 121] and diffusion transformers [17, 50], we aim to move from cross-attention conditioning toward a unified self-attention architecture, which we believe offers stronger coupling between modalities and improved controllability. Finally, our preliminary findings suggest that joint training of the MLLM and VAR decoder benefits multimodal alignment; we therefore see fully unified modeling – capable of both multimodal understanding (text-out) and generation (image-out) – as an exciting long-term goal.

Chapter 7

Conclusions

7.1 Summary

Generative models have achieved remarkable success in synthesizing realistic visual content across images, videos, and text. Yet their potential for visual perception and understanding remains underexplored. This thesis investigates how generative models, particularly diffusion and auto-regressive transformers, can serve as powerful visual learners, bridging the divide between generative and discriminative paradigms. Through a series of studies, we demonstrate that the same architectures enabling high-fidelity generation can also yield rich, transferable representations for visual understanding.

We first introduce Diff-2-in-1, a unified diffusion-based framework that jointly handles multi-modal generation and dense perception via the denoising process and a self-improving learning loop. Extending this idea to videos, we show that video diffusion models inherently capture temporal and structural cues, outperforming non-generative and image-based models in representation quality. Building upon these insights, REM repurposes text-to-video diffusion models for referring video segmentation, achieving superior generalization to unseen objects and scenes. Finally, we present a unified perceptual–generative framework encompassing both diffusion and visual auto-regressive (VAR) architectures, enabling diverse tasks in perception, restoration, and editing. Together, these works chart a coherent path toward general-purpose visual foundation models that integrate understanding and creation within a single generative paradigm.

7.2 Future work

Looking ahead, a key direction of this research lies in further advancing auto-regressive (AR) architectures as scalable and unified solutions for visual understanding and generation. While diffusion models have demonstrated exceptional representational capacity, their iterative denoising process remains computationally demanding, limiting large-scale scaling and deployment. In contrast, AR models naturally align with transformer-based training pipelines and are easier to parallelize, making them more suitable for future large-scale multimodal learning. In particular, I plan to transition from the current cross-attention–based VAR architecture to a purely self-attention–based design, analogous to the evolution from Flamingo to LLaVA and from UNet to DiT. Such a design shift would enable tighter integration between visual and linguistic tokens, more efficient information sharing across modalities, and better scalability for long-context reasoning. The main challenge lies in the computational cost of large-scale text-to-image pretraining, but as new public or open-sourced self-attention–based T2I models emerge, we aim to repurpose them as unified backbones for both perception and synthesis tasks.

Beyond architectural evolution, another promising avenue is to fully unify multimodal generation and understanding within a single AR-driven framework. This thesis has shown that generative and discriminative paradigms can benefit each other: multimodal generation improves representation learning, and multimodal perception enhances controllable generation. Building on this insight, our goal is to develop a model that can produce structured outputs spanning text, image, and video domains, using a single sequence modeling formulation. Such a model would accept arbitrary combinations of input modalities, such as text prompts, sketches, or partial visual observations, and autoregressively predict consistent, multi-scale outputs. This unified treatment would not only streamline the learning of complex visual–linguistic correspondences but also support a wide range of applications, including text-to-video generation, visual reasoning, cross-modal editing, and interactive world modeling.

Ultimately, this research aims to contribute toward a new generation of multimodal foundation models that seamlessly integrate perception, reasoning, and generation.

As AR architectures continue to mature and computational resources become more accessible, I envision models that can learn from Internet-scale multimodal data while maintaining compositionality, interpretability, and controllability. Such models could serve as general-purpose visual agents, capable of perceiving, understanding, and creating in open-world environments, paving the way toward a unified foundation for intelligence that learns from and interacts with the world through both understanding and imagination.

Bibliography

[1] Filippo Aleotti, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Generative adversarial networks for unsupervised monocular depth prediction. In *ECCV Workshops*, 2018.

[2] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. BURST: A benchmark for unifying object recognition, segmentation and tracking in video. In *WACV*, 2023.

[3] Görkay Aydemir, Weidi Xie, and Fatma Guney. Self-supervised object-centric learning for videos. *NeurIPS*, 2023.

[4] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *ICCV*, 2021.

[5] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *ICCV*, 2021.

[6] Anurag Bagchi, Zhipeng Bao, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. Refereverything: Towards segmenting everything we can speak of in videos. *arXiv preprint arXiv:2410.23287*, 2024.

[7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[8] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *CVPR*, 2024.

[9] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

[10] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr Revisited: 2D-3D model alignment via surface normal prediction. In *CVPR*, 2016.

[11] Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Generative modeling for multi-task visual learning. In *ICML*, 2022.

[12] Zhipeng Bao, Pavel Tokmakov, Yu-Xiong Wang, Adrien Gaidon, and Martial

Hebert. Object discovery from motion-guided tokens. In *CVPR*, 2023.

[13] Zhipeng Bao, Yijun Li, Krishna Kumar Singh, Yu-Xiong Wang, and Martial Hebert. Separate-and-enhance: Compositional finetuning for text-to-image diffusion models. In *SIGGRAPH*, 2024.

[14] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2022.

[15] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-JEPA: Latent video prediction for visual representation learning, 2024. URL https://openreview.net/forum?id=WFYbBOEOtv.

[16] Lawrence W Barsalou. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660, 1999.

[17] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, 2025.

[18] M Bellver, C Ventura, C Silberer, I Kazakos, J Torres, and X Giro-i Nieto. Refvos: a closer look at referring expressions for video object segmentation, 2020.

[19] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2023.

[20] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[21] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023.

[22] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *ECCV*, 2022.

[23] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, pages 18392–18402, 2023.

[24] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

[25] David Bruggemann, Menelaos Kanakis, Anton Obukhov, Stamatios Georgoulis, and Luc Van Gool. Exploring relational context for multi-task dense prediction. In *ICCV*, 2021.

[26] Max F. Burg, Florian Wenzel, Dominik Zietlow, Max Horn, Osama Makansi, Francesco Locatello, and Chris Russell. Image retrieval outperforms diffusion models on data augmentation. *TMLR*, 2023.

[27] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018.

[28] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017.

[29] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *TPAMI*, 2021.

[30] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[31] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2Tex: Text-driven texture synthesis via diffusion models. In *ICCV*, 2023.

[32] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[33] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024.

[34] Huayu Chen, Kai Jiang, Kaiwen Zheng, Jianfei Chen, Hang Su, and Jun Zhu. Visual generation without guidance. *arXiv preprint arXiv:2501.15420*, 2025.

[35] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. In *ICCV*,

2023.

[36] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *CVPR*, 2023.

[37] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *RSS*, 2023.

[38] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *JMLR*, 2024.

[39] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[40] Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *ECCV*, 2024.

[41] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[42] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017.

[43] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[44] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[45] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[46] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021.

[47] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[48] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015.

[49] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014.

[50] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[51] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *ICCV*, 2023.

[52] Tsu-Jui Fu, Yusu Qian, Chen Chen, Wenze Hu, Zhe Gan, and Yinfei Yang. Univg: A generalist diffusion model for unified image generation and editing. *arXiv preprint arXiv:2503.12652*, 2025.

[53] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, 2017.

[54] Yulu Gan, Sungwoo Park, Alexander Schubert, Anthony Philippakis, and Ahmed M Alaa. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. *arXiv preprint arXiv:2310.00390*, 2023.

[55] Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, and Alan L. Yuille. NDDR-CNN: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *CVPR*, 2019.

[56] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, 2018.

[57] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

[58] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 2023.

[59] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

[60] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow.

Digging into self-supervised monocular depth estimation. In *ICCV*, 2019.

[61] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. In *CVPR*, 2022.

[62] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022.

[63] Vitor Guizilini, Pavel Tokmakov, Achal Dave, and Rares Ambrus. Grin: Zero-shot metric depth with pixel-level diffusion. *arXiv preprint arXiv:2409.09896*, 2024.

[64] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *ECCV*, 2018.

[65] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Chongyang Ma, Weiming Hu, Zhengjun Zha, Haibin Huang, Pengfei Wan, et al. I2v-adapter: A general image-to-video adapter for video diffusion models. *arXiv preprint arXiv:2312.16693*, 2023.

[66] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[67] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, 2025.

[68] Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chenwei Xie, Yu Liu, and Jingren Zhou. Ace: All-round creator and editor following instructions via diffusion transformer. *arXiv preprint arXiv:2410.00086*, 2024.

[69] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. In *ICCV*, 2025.

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[72] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[73] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022.

[74] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. In *NeurIPS*, 2023.

[75] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *NeurIPS*, 2023.

[76] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[77] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. In *NeurIPS*, 2017.

[78] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[79] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[80] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[81] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[82] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[83] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[84] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016.

[85] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with

adaptive instance normalization. In *ICCV*, 2017.

[86] Zhipeng Huang, Shaobin Zhuang, Canmiao Fu, Binxin Yang, Ying Zhang, Chong Sun, Zhizheng Zhang, Yali Wang, Chen Li, and Zheng-Jun Zha. Wegen: A unified model for interactive multimodal generation as we chat. In *CVPR*, 2025.

[87] Tianrui Hui, Shaofei Huang, Si Liu, Zihan Ding, Guanbin Li, Wenguan Wang, Jizhong Han, and Fei Wang. Collaborative spatial-temporal modeling for language-queried video actor segmentation. In *CVPR*, 2021.

[88] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[89] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *NeurIPS*, 2020.

[90] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. *arXiv preprint arXiv:2312.07509*, 2023.

[91] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, 2013.

[92] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, 2013.

[93] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. DDP: Diffusion model for dense visual prediction. In *ICCV*, 2023.

[94] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023.

[95] Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv preprint arXiv:2402.03161*, 2024.

[96] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

[97] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024.

[98] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.

[99] Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. Slime: Segment like me. *arXiv preprint arXiv:2309.03179*, 2023.

[100] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, 2019.

[101] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[102] Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional Object-Centric Learning from Video. In *ICLR*, 2022.

[103] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *NeurIPS*, 2023.

[104] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[105] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011.

[106] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023.

[107] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. URL https://doi.org/10.5281/zenodo.10948109.

[108] Lubor Ladicky, Bernhard Zeisl, and Marc Pollefeys. Discriminatively trained dense surface normal estimation. In *ECCV*, 2014.

[109] Bolin Lai, Felix Juefei-Xu, Miao Liu, Xiaoliang Dai, Nikhil Mehta, Chenguang Zhu, Zeyi Huang, James M Rehg, Sangmin Lee, Ning Zhang, et al. Unleashing in-context learning of autoregressive models for few-shot image manipulation. In *CVPR*, 2025.

[110] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024.

[111] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, 2023.

[112] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng.

All-in-one image restoration for unknown corruption. In *CVPR*, 2022.

[113] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[114] Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. *arXiv preprint arXiv:2410.01756*, 2024.

[115] Yaowei Li, Yuxuan Bian, Xuan Ju, Zhaoyang Zhang, Junhao Zhuang, Ying Shan, Yuexian Zou, and Qiang Xu. Brushedit: All-in-one image inpainting and editing. *arXiv preprint arXiv:2412.10316*, 2024.

[116] Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. Visualcloze: A universal image generation framework via visual in-context learning. In *ICCV*, 2025.

[117] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *ICCV*, 2023.

[118] Yijing Lin, Mengqi Huang, Shuhan Zhuang, and Zhendong Mao. Realgeneral: Unifying visual generation via temporal in-context learning with video models. *arXiv preprint arXiv:2503.10406*, 2025.

[119] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, 2015.

[120] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

[121] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024.

[122] Ruoshi Liu, Rundi We, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023.

[123] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

[124] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In *ICLR*, 2024.

[125] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021.

[126] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[127] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, 2022.

[128] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *NeurIPS*, 2020.

[129] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[130] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[131] Ilya Loshchilov, Frank Hutter, et al. Decoupled weight decay regularization. In *ICLR*, 2019.

[132] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022.

[133] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

[134] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *CVPR*, 2024.

[135] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *NeurIPS*, 2023.

[136] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *NeurIPS*, 2023.

[137] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *SIGGRAPH*, 2020.

[138] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023.

[139] Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing autoregressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024.

[140] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *arXiv preprint arXiv:2409.03757*, 2024.

[141] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In *CVPR*, 2019.

[142] Qingyang Mao, Qi Cai, Yehao Li, Yingwei Pan, Mingyue Cheng, Ting Yao, Qi Liu, and Tao Mei. Visual autoregressive modeling for instruction-guided image editing. *arXiv preprint arXiv:2508.15772*, 2025.

[143] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, 2022.

[144] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[145] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022.

[146] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. VSPW: A large-scale dataset for video scene parsing in the wild. In *CVPR*, 2021.

[147] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, 2016.

[148] Ron Mokady, Amir Hertz, and Amit H. Bermano. ClipCap: CLIP prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[149] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan L. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

[150] Sauradip Nag, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, and Tao Xiang. Difftad: Temporal action detection with proposal denoising diffusion. *arXiv preprint arXiv:2303.14863*, 2023.

[151] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017.

[152] Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models. *arXiv*

preprint arXiv:2401.11739, 2024.

[153] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *NeurIPS*, 2001.

[154] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *CVPR*, 2023.

[155] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[156] Ke Ning, Lingxi Xie, Fei Wu, and Qi Tian. Polar relative positional encoding for video-language segmentation. In *IJCAI*, 2020.

[157] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

[158] OpenAI. Chatgpt, 2023. URL https://chat.openai.com.

[159] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023.

[160] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *CVPR*, 2024.

[161] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023.

[162] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

[163] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[164] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

[165] Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *PAMI*, 1990.

[166] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. iDisc: Internal discretization

for monocular depth estimation. In *CVPR*, 2023.

[167] Andrea Pilzer, Dan Xu, Mihai Puscas, Elisa Ricci, and Nicu Sebe. Unsupervised adversarial depth estimation using cycled generative networks. In *3DV*, 2018.

[168] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

[169] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[170] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[171] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[172] Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, 24(11):1628–1636, 2021.

[173] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. GeoNet: Geometric neural network for joint depth and surface normal estimation. In *CVPR*, 2018.

[174] Xiaojuan Qi, Zhengzhe Liu, Renjie Liao, Philip H.S. Torr, Raquel Urtasun, and Jiaya Jia. GeoNet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. *TPAMI*, 2022.

[175] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. *arXiv preprint arXiv:2312.04483*, 2023.

[176] Yunpeng Qu, Kun Yuan, Jinhua Hao, Kai Zhao, Qizhi Xie, Ming Sun, and Chao Zhou. Visual autoregressive modeling for image super-resolution. *arXiv preprint arXiv:2501.18993*, 2025.

[177] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[178] Rajat Raina, Yirong Shen, Andrew Y. Ng, and Andrew McCallum. Classification

with hybrid generative/discriminative models. In *NeurIPS*, 2003.

[179] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[180] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

[181] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2022.

[182] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022.

[183] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024.

[184] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[185] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[186] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[187] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM TOG*, 2004.

[188] Y. Dan Rubinstein and Trevor Hastie. Discriminative vs. informative learning. In *KDD*, 1997.

[189] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023.

[190] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[191] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J. Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. In *NeurIPS*, 2023.

[192] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J. Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023.

[193] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[194] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022.

[195] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, 2020.

[196] Chenze Shao, Fandong Meng, and Jie Zhou. Continuous visual autoregressive generation via score maximization. *arXiv preprint arXiv:2505.07812*, 2025.

[197] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012.

[198] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[199] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[200] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[201] Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. *NeurIPS*, 2022.

[202] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

[203] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[204] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[205] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[206] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

[207] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

[208] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024.

[209] Zeyi Sun, Ziyang Chu, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. X-prompt: Towards universal in-context image generation in auto-regressive vision language foundation models. *arXiv preprint arXiv:2412.01824*, 2024.

[210] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[211] Hongxuan Tang, Hao Liu, and Xinyan Xiao. Ugen: Unified autoregressive multimodal model with progressive vocabulary learning. *arXiv preprint arXiv:2503.21193*, 2025.

[212] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *NeurIPS*, 2023.

[213] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *NeurIPS*, 2023.

[214] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context interleaved and interactive any-to-any generation. In *CVPR*, 2024.

[215] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

[216] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[217] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *NeurIPS*, 2024.

[218] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the" object" in video object segmentation. In *CVPR*, 2023.

[219] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022.

[220] Alexander Toshev and Christian Szegedy. DeepPose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

[221] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *ICLR*, 2024.

[222] Ilkay Ulusoy and Christopher M. Bishop. Generative versus discriminative methods for object recognition. In *CVPR*, 2005.

[223] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017.

[224] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. MTI-Net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, 2020.

[225] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[226] Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024.

[227] Cong Wang, Jiaxi Gu, Panwen Hu, Songcen Xu, Hang Xu, and Xiaodan Liang. Dreamvideo: High-fidelity image-to-video generation with image retention and text guidance. *arXiv preprint arXiv:2312.03018*, 2023.

[228] Jiahuan Wang, Yuxin Chen, Jun Yu, Guangming Lu, and Wenjie Pei. Editinfinity: Image editing with binary-quantized generative models. *arXiv preprint arXiv:2510.20217*, 2025.

[229] Jinhong Wang, Jian Liu, Dongqi Tang, Weiqiang Wang, Wentong Li, Danny Chen, Jintai Chen, and Jian Wu. Scalable autoregressive monocular depth estimation. In *CVPR*, 2025.

[230] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[231] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen,

Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022.

[232] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. *arXiv preprint arXiv:2312.15770*, 2023.

[233] Xiaolong Wang, David Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *CVPR*, 2015.

[234] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[235] Yufei Wang, Lanqing Guo, Zhihao Li, Jiaxing Huang, Pichao Wang, Bihan Wen, and Jian Wang. Training-free text-guided image editing with visual autoregressive model. *arXiv preprint arXiv:2503.23897*, 2025.

[236] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[237] Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *ICLR*, 2024.

[238] Joachim Weickert et al. *Anisotropic diffusion in image processing*. Teubner Stuttgart, 1998.

[239] Zejia Weng, Xitong Yang, Zhen Xing, Zuxuan Wu, and Yu-Gang Jiang. Genrec: Unifying video generation and recognition with diffusion models. *arXiv preprint arXiv:2408.15241*, 2024.

[240] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[241] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *CVPR*, 2025.

[242] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

[243] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video:

One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023.

[244] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, 2022.

[245] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation, 2022. URL https://arxiv.org/abs/2201.00487.

[246] Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *CVPR*, 2024.

[247] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *ICCV*, 2023.

[248] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

[249] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, 2013.

[250] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *CVPR*, 2025.

[251] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018.

[252] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

[253] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. PAD-Net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, 2018.

[254] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. ODISE: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023.

[255] Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. Pixel-aligned language model. In *CVPR*, 2024.

[256] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao

Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

[257] Yangyang Xu, Xiangtai Li, Haobo Yuan, Yibo Yang, and Lefei Zhang. Multi-task learning with multi-query transformer for dense prediction. *TCSVT*, 2023.

[258] Yangyang Xu, Yibo Yang, and Lefei Zhang. DeMT: Deformable mixer transformer for multi-task learning of dense prediction. In *AAAI*, 2023.

[259] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023.

[260] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Hao Dong, Zhongjiang He, and Peng Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. In *AAAI*, 2024.

[261] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

[262] Hanrong Ye and Dan Xu. InvPT: Inverted pyramid multi-task transformer for dense scene understanding. In *ECCV*, 2022.

[263] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *ECCV*, 2022.

[264] Hanrong Ye and Dan Xu. TaskPrompter: Spatial-channel multi-task prompting for dense scene understanding. In *ICLR*, 2023.

[265] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, 2019.

[266] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024.

[267] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

[268] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[269] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016.

[270] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L.

Berg. Modeling context in referring expressions, 2016. URL https://arxiv.org/abs/1608.00272.

[271] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. *arXiv preprint arXiv:2312.12490*, 2023.

[272] Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. *NeurIPS*, 2023.

[273] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[274] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *ECCV*, 2022.

[275] Fangneng Zhan, Shijian Lu, and Chuhui Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *ECCV*, 2018.

[276] Hong Zhang, Zhongjie Duan, Xingjun Wang, Yuze Zhao, Weiyi Lu, Zhipeng Di, Yixuan Xu, Yingda Chen, and Yu Zhang. Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*, 2025.

[277] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements DINO for zero-shot semantic correspondence. In *NeurIPS*, 2023.

[278] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *NeurIPS*, 2023.

[279] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

[280] Mingtong Zhang, Shuhong Zheng, Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Beyond RGB: Scene-property synthesis with neural radiance fields. In *WACV*, 2023.

[281] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

[282] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, 2019.

[283] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *NeurIPS*, 2023.

[284] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023.

[285] Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In *ICML*, 2023.

[286] Shuhong Zheng, Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Multi-task view synthesis with neural radiance fields. In *ICCV*, 2023.

[287] Shuhong Zheng, Zhipeng Bao, Ruoyu Zhao, Martial Hebert, and Yu-Xiong Wang. Diff-2-in-1: Bridging generation and dense perception with diffusion models. In *ICLR*, 2025.

[288] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL `https://github.com/hpcaitech/Open-Sora`.

[289] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017.

[290] Ling Zhou, Zhen Cui, Chunyan Xu, Zhenyu Zhang, Chaoqun Wang, Tong Zhang, and Jian Yang. Pattern-structure diffusion for multi-task learning. In *CVPR*, 2020.

[291] Xingyi Zhou, Arjun Karpur, Linjie Luo, and Qixing Huang. StarMap for category-agnostic keypoint and viewpoint estimation. In *ECCV*, 2018.

[292] Zikun Zhou, Wentao Xiong, Li Zhou, Xin Li, Zhenyu He, and Yaowei Wang. Driving referring video object segmentation with vision-language pre-trained models. *arXiv preprint arXiv:2405.10610*, 2024.

[293] Hanshen Zhu, Zhen Zhu, Kaile Zhang, Yiming Gong, Yuliang Liu, and Xiang Bai. Training-free geometric image editing on diffusion models. In *ICCV*, 2025.

[294] Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. Emotion classification with data augmentation using generative adversarial networks. In *PAKDD*, 2018.

[295] Zhen Zhu, Yijun Li, Weijie Lyu, Krishna Kumar Singh, Zhixin Shu, Soeren

Pirk, and Derek Hoiem. Consistent multimodal generation via a unified GAN framework. In *WACV*, 2024.

[296] Zixin Zhu, Xuelu Feng, Dongdong Chen, Junsong Yuan, Chunming Qiao, and Gang Hua. Exploring pre-trained text-to-video diffusion models for referring video object segmentation. In *ECCV*, 2024.

[297] Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. *arXiv preprint arXiv:2501.12327*, 2025.