

Towards Natural Language-Driven Shape Arrangement Synthesis with Semantically-Aware Geometric Constraint Systems

Vihaan Misra
CMU-RI-TR-25-19

April 15, 2025

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

Thesis Committee:

Jean Oh, *chair*
Jun-Yan Zhu
Zackory Erickson
Peter Schaldenbrand

*Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Robotics.*

To my brother, Ashwin Misra.

Abstract

While diffusion-based models excel at generating photorealistic images from text, a more nuanced challenge emerges when constrained to using only a fixed set of rigid shapes—akin to solving tangram puzzles or arranging real-world objects to match semantic descriptions. We formalize this problem as *shape-based image generation*, a new natural language-guided image-to-image translation task that requires rearranging the input set of rigid shapes into non-overlapping configurations and visually communicating the target concept.

Unlike pixel-manipulation approaches, our method explicitly parameterizes each shape within a differentiable vector graphics pipeline, iteratively optimizing placement and orientation through score distillation sampling from pretrained diffusion models. To preserve arrangement clarity, we introduce a semantically-aware collision resolution mechanism that applies minimal contextually coherent adjustments when overlaps occur, ensuring smooth convergence toward physically valid configurations. By bridging diffusion-based semantic guidance with explicit geometric constraint systems, our approach yields interpretable compositions where spatial relationships clearly embody the natural language prompt. Extensive experiments demonstrate compelling results across diverse scenarios, with quantitative and qualitative advantages over alternative techniques.

Acknowledgments

The universe has been beyond generous to keep me in the company of wonderful human beings throughout my academic journey. I would like to start by wholeheartedly thanking Jean for taking me under her wing as an undergrad and helping me shape my understanding of research. Being your student showed me that mentorship extends beyond just academic guidance. Thank you for being patient, supportive, and kind, and for your unwavering efforts in shaping my path. I would also like to thank Jun-Yan, whose projects and research have helped me much more than he probably realizes. I am grateful to be starting a collaboration with you and very excited about the project and its potential scope. I am also thankful to Zackory for stepping in at a crucial time and for his guidance. I look forward to furthering this field through systems that improve human and robot creative and positive interaction.

I am deeply grateful to Peter for being the most amazing mentor I could have asked for. A lot of what I know today has come from taking inspiration from your insights, curiosity, and patience. Thank you also for being an incredible conference and research trip buddy. I want to deeply thank my lab colleagues and the wonderful people in it - Alonso, Pablo, Arthur, Elliot, Uksang, Tanmay, Jon, Zhixuan, Ingrid, Ben, Beverly, Andrew, Hyun, and Gavin. Your support, feedback, and camaraderie have made this journey both intellectually stimulating and personally rewarding.

My time at CMU and in Pittsburgh has been enriched by friendships that have made this experience far better than I could have imagined. While there are too many friends to name in this limited space, I am profoundly grateful for the friends both here and back home in India. Your encouragement, laughter, and support have been invaluable. My heartfelt thanks to Mononito, whose support began even before I arrived at CMU. His guidance in helping me get into CMU and throughout my time here has been invaluable.

Last but certainly not least, I am thankful to my family - Varsha Misra, Anupam Misra, and my brother Ashwin Misra who has been both an inspiration and a mentor throughout my life. He showed me the ropes when I didn't even know there were ropes to be shown and always kept me motivated. Thank you all for instilling the importance of education in me, and more importantly, for teaching me to strive to be a better person.

Thank you for helping me understand empathy, love, and compassion, and for providing unwavering support through every challenge and celebration. Your belief in me has been my greatest strength.

Funding

This work was in part supported by the Technology Innovation Program (20018295, Meta-human: a virtual cooperation platform for specialized industrial services) funded by the Ministry of Trade, Industry & Energy(MOTIE, Korea) and NSF IIS-2112633.

Contents

1	Introduction	1
2	Related Work	5
3	Method: ShapeShift	7
3.0.1	Preliminaries	7
3.0.2	Minimal Shape Parameterization	9
3.0.3	Multi-Scale Rendering and Semantic Guidance	9
3.0.4	Content-Aware Collision Resolution	10
3.0.5	Overall Optimization Process and Weighting	12
4	Results	14
4.0.1	Adding Physical Constraints to Image Generators	14
4.0.2	Improving Semantic-Guidance in Vision-Language Model Plan- ning	16
4.0.3	Ablation Study	17
4.0.4	Human Rater Evaluation	18
5	Conclusion	20
6	Discussion and Limitations	21
7	Future Work	22
7.0.1	Robotic Implementation of 2D Arrangements	22
7.0.2	Extension to 3D Shape Assemblies	23
A	Appendix	25
A.1	Effect of Shape Count on Semantic Alignment	27
	Bibliography	29

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

List of Figures

1.1	Various shape arrangements produced by our pipeline. Given a goal concept and an input image containing a set of arbitrary shapes, the task is to generate an image of the same set of shapes rearranged to match the textual concept, e.g., “Horse,” “Tree,” or “Crown,” without pieces overlapping each other.	1
1.2	A Lack of Physical Constraints in Image Generators. People cleverly use a small set of blocks to create semantically-rich arrangements with Tangram puzzles. While Stable Diffusion [2] generates semantically-rich Tangram arrangements, these are invalid due to not using the available blocks or having physically impossible overlaps. This illustrates the gap between pixel-based generation and physically constrained arrangement tasks.	3
3.1	ShapeShift Overview Our method iteratively optimizes the positions and orientations of a given arrangement of objects to match a given language description and obey physical constraints. The process begins by extracting shape parameters P (position x_k, y_k and orientation α_k) from the input arrangement using SAM [22]. These parameters are used to render the current canvas via DiffVG. The goal concept is encoded through CLIP, enabling Multi-Scale Score Distillation Sampling that generates gradients to update shape parameters (ΔP). Simultaneously, our Content-Aware Collision Resolution Module utilizes GPT to identify semantic regions in the arrangement and employs geometric adjustments with SAT (Separating Axis Theorem) to ensure shapes remain physically valid while respecting semantic relationships. . . .	8
3.2	Multi-Scale Rendering. (a) High-resolution render capturing fine details and precise edge alignments;(b) Low-resolution render emphasizing global layout and overall shape placement; This provides for a robust semantic guidance and avoids overfitting to pixel-level noise. We use aggregated SDS loss computed by averaging losses across scales, ensuring that both coarse and fine semantic features guide the optimization.	10

3.3	Content-Aware Collision Resolution Process for the goal concept “<i>Sword</i>” (a) Plain collision resolution relies solely on geometric penetration, often producing unnatural displacements. (b) The initial render state before collision detection. (c) Our content-aware method integrates extracted semantic concepts and dynamic attention weights to preserve object orientations and contextual relationships (e.g., maintaining a sword’s natural alignment).	12
4.1	Comparison with VLM Planning. The figure shows arrangements of everyday objects into three semantic concepts (Windmill, Dog, Shoe). The top row displays arrangements following the VLM’s coordinate and orientation specifications, while the bottom row shows our method’s results for the same prompts.	16
4.2	Human-Rater Study. Example arrangements of the “rabbit” concept generated using three methods (columns) and shown with complete or partial (30% removed) object sets (rows).	19
A.1	Relationship between number of shapes and semantic alignment (measured by CLIP score). With fewer than 6 shapes, arrangements struggle to capture sufficient detail for strong semantic alignment. Between 10-20 shapes, semantic clarity plateaus, suggesting an optimal range for balancing expressivity and arrangement simplicity. Beyond 20 shapes, we observe further improvements as more complex concepts become representable.	27

List of Tables

4.1	Qualitative Comparison with Baselines. We evaluate how different generative approaches respond to the task of arranging specific objects and basic shapes according to semantic prompts. While Stable Diffusion Img2Img [24] introduces extraneous elements and distortions, InstructPix2Pix [1] sacrifices discrete shape boundaries for photorealism, and SORA [18] exhibits good quality objects but inconsistent spatial organization despite its temporal capabilities. ShapeShift (ours) preserves both the integrity of each input primitive and their semantic arrangement, demonstrating the advantage of explicit geometric parameterization over pixel-space transformations when physical constraints must be maintained.	15
4.2	Ablation Study. Comparison of three variants of our pipeline across different shape counts. We report the average overlap area percentage (Overlap %), number of collisions (# Coll.), and CLIP Score.	17
4.3	Human-Rater Results. We report CLIP Scores for survey images (example in Figure 4.2) alongside participant accuracy in identifying target concepts, along with standard error. Comparing accuracy before and after random 30% image removal reveals that the only significant difference occurred with our proposed method.	19
A.1	Additional arrangement examples generated by ShapeShift. Each arrangement uses the exact same set of shapes from the input image, rearranged to match the specified concept. The numbers in filenames indicate the experimental batch identifier.	26
A.2	Detailed analysis of shape count effect on arrangement quality and semantic alignment	28

Chapter 1

Introduction

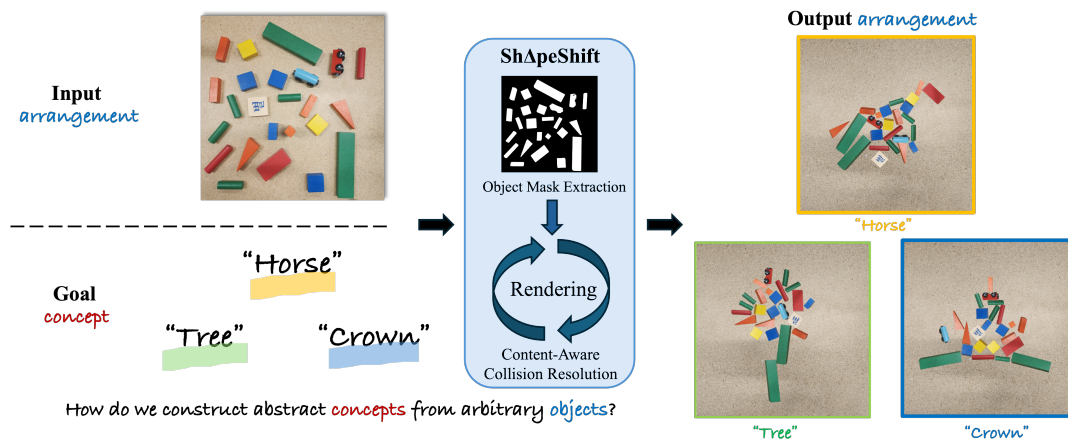


Figure 1.1: **Various shape arrangements produced by our pipeline.** Given a goal concept and an input image containing a set of arbitrary shapes, the task is to generate an image of the same set of shapes rearranged to match the textual concept, e.g., “Horse,” “Tree,” or “Crown,” without pieces overlapping each other.

The human ability to create meaning from simple arrangements is remarkable, as shown by tangram puzzles or Lego structures, where a small set of geometric pieces can create countless meaningful configurations. Creating such arrangements requires understanding both semantic intent and physical constraints.

Through this work, we tackle the problem of semantic shape arrangement, where a given set of visual objects must be rearranged to fit a given text description. Importantly, this arrangement must also be physically possible, meaning that the

shapes cannot overlap or violate other geometric constraints.

In this context, we introduce *shape-based image generation* as a new text-guided image-to-image translation task where the input set of rigid shapes must be rearranged into a collision-free configuration that visually represents the semantic goal.

Despite the impressive advancements in AI-generated imagery and rearrangement problems, bridging the gap between abstract textual concepts and physically constrained visual representations remains challenging. Previous work on language-guided object rearrangement often focuses on moving a single object into a place via a language command [12, 16]. Works that rearrange many objects often only work on a subset of objects in a dataset [5] or have limited semantic goal specifications [16, 17].

Modern text-to-image models [2, 24, 25] have improved greatly in their ability to generate images of diverse subjects with very general language inputs. However, these models operate in a continuous pixel or latent space and do not have a firm understanding of real-world constraints. Making things in the real world invariably involves working within material constraints: crafting an image with charcoal differs fundamentally from working with stained glass or geometric tiles. As seen in Figure 1.2, Stable Diffusion can generate high-fidelity images of tangram blocks arranged in shapes. However, they frequently introduce shapes beyond the standard tangram set, and they do not obey other physical constraints such as overlapping. Our aim is to **ground pixel-generating diffusion models with real-world, geometric constraints**. In this paper, we specifically focus on object arrangement in 2D with two types of constraints: geometric shape preservation—i.e., the generated image must use the same set of shapes in the input image—and collision avoidance—i.e., objects cannot overlap with each other.

We introduce ShapeShift, a pipeline that shifts the emphasis from pixel-based generation to *parameterized 2D primitives*, each defined solely by its *position* and *orientation*. By imposing strict but semantically grounded collision avoidance, we ensure that shapes never obscure one another, facilitating neat, legible arrangements. Rather than requiring domain-specific training data for shape arrangements, our approach leverages the rich semantic priors embedded in pretrained diffusion models through Score Distillation Sampling (SDS) [10, 19], aligning the final layout with the input textual cues while maintaining physical plausibility. Unlike purely pixel-oriented methods, our approach produces compositions in which each shape stands out as

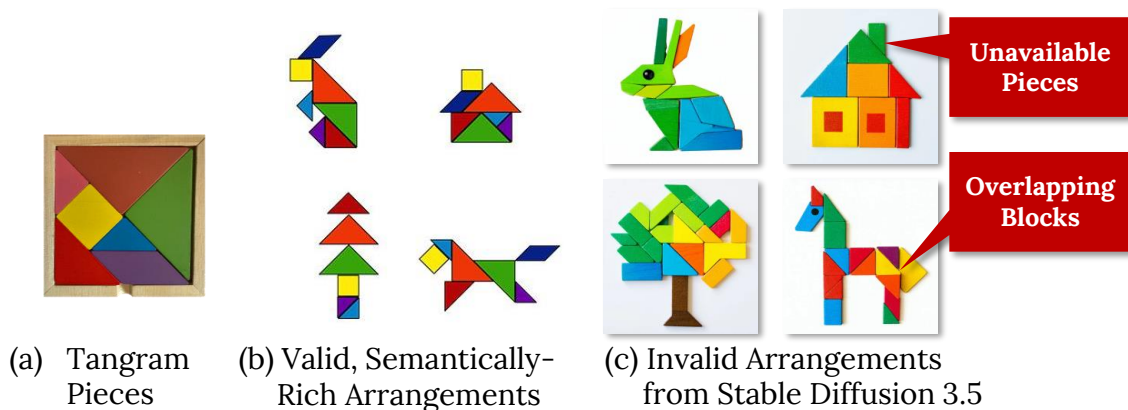


Figure 1.2: **A Lack of Physical Constraints in Image Generators.** People cleverly use a small set of blocks to create semantically-rich arrangements with Tangram puzzles. While Stable Diffusion [2] generates semantically-rich Tangram arrangements, these are invalid due to not using the available blocks or having physically impossible overlaps. This illustrates the gap between pixel-based generation and physically constrained arrangement tasks.

distinct and unoccluded, bridging the gap between abstract semantic guidance and concrete physical constraints. The ability to transform abstract text descriptions into physically valid object arrangements opens new possibilities across design, education, and human-computer interaction.

Focusing exclusively on *position* and *orientation* simplifies the geometric search space, allowing the diffusion model’s gradients to guide the global arrangement of shapes rather than inducing local deformations. This simplification facilitates the attainment of a stable, collision-free solution, while limiting shape complexity helps the diffusion model to better capture the overall structure in the final rendering. Our contributions are threefold:

1. We propose **ShapeShift**, a pipeline that integrates multi-scale SDS with collision-free shape arrangement, leveraging minimal shape parameters to maintain physical validity while preserving semantic alignment.
2. We introduce a *content-aware* and *lightweight* collision resolution strategy that, following each text-driven gradient update, gently adjusts shapes through minimal and semantically coherent translations, ensuring a stable, overlap-free arrangement over time.

1. Introduction

3. We demonstrate results on a variety of conceptual prompts, illustrating that constraining geometry to simple primitives—while leveraging a diffusion model to determine optimal placement—produces coherent, interpretable compositions for semantic illustration tasks.

Chapter 2

Related Work

Differentiable Vector Graphics. Scalable Vector Graphics (SVGs) represent images using elements such as Bézier curves, lines, and shapes combined with color information. Differentiable rasterizers like DiffVG [13] enable gradient-based optimization of these elements—strokes, transformations, and colors—bridging the gap between vector and raster representations. Several works have incorporated text-driven objectives [4, 26] and combined text-to-image diffusion models [24] with differentiable rendering [13] for SVG synthesis [8, 9, 10]. While these approaches excel at expressive vector generation, our work enforces collision-free geometry through precise position and orientation control, ensuring both semantic coherence and spatial clarity.

Text-to-Image Diffusion. Diffusion models have transformed text-to-image generation by iteratively denoising latent representations [7, 28], superseding earlier adversarial [23, 31, 33] and autoregressive approaches [21, 32]. Modern systems like Stable Diffusion [24] operate in latent space using UNet-based denoising networks conditioned on CLIP text embeddings [20] and trained on massive datasets [27]. Our work extends these advances by leveraging Stable Diffusion’s semantic priors while preserving explicit geometric primitives, bridging textual semantics with spatial control.

2. Related Work

Arrangement Prediction. Recent advances in arrangement prediction leverage large language models (LLMs) and vision-language models (VLMs) to plan spatial configurations based on high-level semantic cues. For example, Dream2Real [12] and StructDiffusion [16] focus on predicting goal poses or placements in simple scenarios, such as aligning objects along a line or positioning a single item within a scene. Other object arrangement works can rearrange multiple objects but the semantic inputs are often limited to goals such as “set the table” [16] or “put objects in a line” [17].

For more semantically flexible inputs, Dall-E-Bot [11] uses a pretrained image generator along with descriptions of each object in the image to generate new images of the objects in a desired arrangement. Although this can work with scenes with distinct items (e.g., fork, knife, plate), it fails with items that are less identifiable with language, such as the blocks in Figure 1.2. Blox-Net [5] uses a VLM to rearrange 3D objects into a language-described goal; however, this work only operates on cuboids and cylinders. In our work, we strive to rearrange arbitrarily shaped objects without requiring language descriptions or priors for each.

Collision Avoidance Traditional approaches to collision detection and resolution in computer graphics rely on geometric primitives through methods like the Separating Axis Theorem (SAT) [6, 14], which efficiently determines overlap between convex polygons. Recent arrangement generation systems have incorporated collision avoidance mechanisms but primarily focus on geometric validity without semantic context. LayoutVLM [29] exemplifies this approach by using vision-language models to guide 3D arrangement while optimizing object positions through differentiable graphics to prevent intersections. However, their collision resolution operates purely in geometric space, applying corrections without considering semantic relationships between objects. Our work fundamentally differs by introducing content-aware collision resolution that integrates semantic understanding directly into the geometric constraint satisfaction process. Unlike existing approaches that default to physical validity at the expense of semantic coherence, our method leverages cross-modal embeddings to inform how objects should be displaced when overlaps occur. This preserves meaningful spatial relationships while maintaining physical plausibility.

Chapter 3

Method: ShapeShift

Given an input image of a set of objects and a text description, our goal is to generate a new image such that the same set of objects are rearranged into a configuration that is both physically feasible and semantically relevant of the text description. Our proposed method, ShapeShift, first segments out the object shapes using SAM2 [22] and GroundingDINO [15]. We can differentiably render these shapes into the scene with new positions and orientations using DiffVG [13]. We optimize the positions and orientations of the shapes so that the new rendered image of the arrangement (1) decreases the Score Distillation Sampling (SDS) [19] loss, which compares the input prompt to the rendered image, and (2) avoids overlapping shapes using our proposed Content-Aware Collision Resolution algorithm.

3.0.1 Preliminaries

Diffusion Models and Score Distillation Sampling. Diffusion models generate images by progressively denoising a Gaussian sample, where a noisy version of an image x_0 at timestep t is defined as $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ representing cumulative noise schedule coefficients. During denoising, a noise predictor $\epsilon_\theta(x_t; y, t)$ estimates noise given the current sample, conditioning text y , and timestep. Latent Diffusion Models [24] enhance efficiency by operating in a compact latent space via an encoder-decoder architecture.

SDS [19] extends this framework to optimize non-rasterized representations by

3. Method: ShapeShift

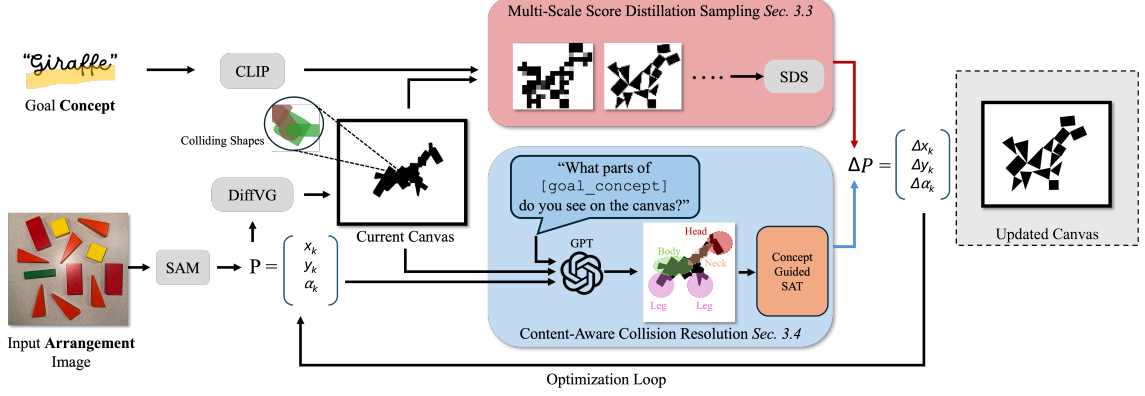


Figure 3.1: **ShapeShift Overview** Our method iteratively optimizes the positions and orientations of a given arrangement of objects to match a given language description and obey physical constraints. The process begins by extracting shape parameters P (position x_k, y_k and orientation α_k) from the input arrangement using SAM [22]. These parameters are used to render the current canvas via DiffVG. The goal concept is encoded through CLIP, enabling Multi-Scale Score Distillation Sampling that generates gradients to update shape parameters (ΔP). Simultaneously, our Content-Aware Collision Resolution Module utilizes GPT to identify semantic regions in the arrangement and employs geometric adjustments with SAT (Separating Axis Theorem) to ensure shapes remain physically valid while respecting semantic relationships.

leveraging pretrained diffusion priors. In DreamFusion [19], a 3D scene rendered from a random viewpoint produces an image x , which is perturbed to yield $\tilde{x}_t = \sqrt{\bar{\alpha}_t} x + \sqrt{1 - \bar{\alpha}_t} \epsilon$. The SDS loss is derived as:

$$\nabla_{\phi} L_{SDS} = \mathbb{E}_{t, \epsilon} \left[w(t) \left(\hat{\epsilon}_{\theta}(\tilde{x}_t; y, t) - \epsilon \right) \frac{\partial x}{\partial \phi} \right],$$

where $w(t)$ is a noise-dependent weighting factor. VectorFusion [10] extended this methodology to text-to-SVG generation, computing the SDS loss in latent space to guide vector graphics optimization.

Our approach adapts SDS for arrangement generation with explicit geometric primitives.

3.0.2 Minimal Shape Parameterization

In our framework, each object in the layout is modeled by a fixed geometric template that can represent both simple primitives (e.g., circles, rectangles, or triangles) and complex object masks extracted from segmentation, approximated as polygonal contours. Regardless of the underlying complexity, every element is parameterized by a minimal set of variables: its two-dimensional position of its calculated centroid (x_i, y_i) and its orientation α_i . This concise representation confines the optimization to a low-dimensional space, mitigating excessive deformations and providing a way to prevent overlaps. Moreover, by directly representing elements as precise vector shapes or polygons—rather than relying on coarse bounding box approximations—we enable accurate geometry-based collision detection, e.g., using SAT [6].

3.0.3 Multi-Scale Rendering and Semantic Guidance

The objects can be rendered onto a photo of a scene in new positions and orientations. This can be performed with the actual cutout photos of the objects (as in Fig. 3.2) or with the contours of the images using DiffVG [13].

To robustly capture both the overall structure and fine details of the layout, we render the current scene at multiple resolutions. Lower-resolution renders emphasize the coarse spatial arrangement and global silhouettes of the elements, while higher-resolution renders capture precise orientations and subtle edge alignments. Formally, let $x^{(\sigma)}$ denote the rendered image at scale σ . Each $x^{(\sigma)}$ is normalized and mapped into the latent space of a pretrained diffusion model (e.g., Stable Diffusion [24]), where the noise predictor estimates the noise component given the conditioning text y and timestep t .

The SDS loss computed at scale σ is used to compute the final multi-scale SDS loss, which is obtained by averaging the SDS losses across K scales:

$$\mathcal{L}_{\text{MSDS}} = \frac{1}{K} \sum_{j=1}^K \mathcal{L}_{\text{SDS}}^{(\sigma_j)}.$$

This multiscale approach not only mitigates artifacts that may occur at a single resolution but also ensures that both coarse and fine semantic features guide the

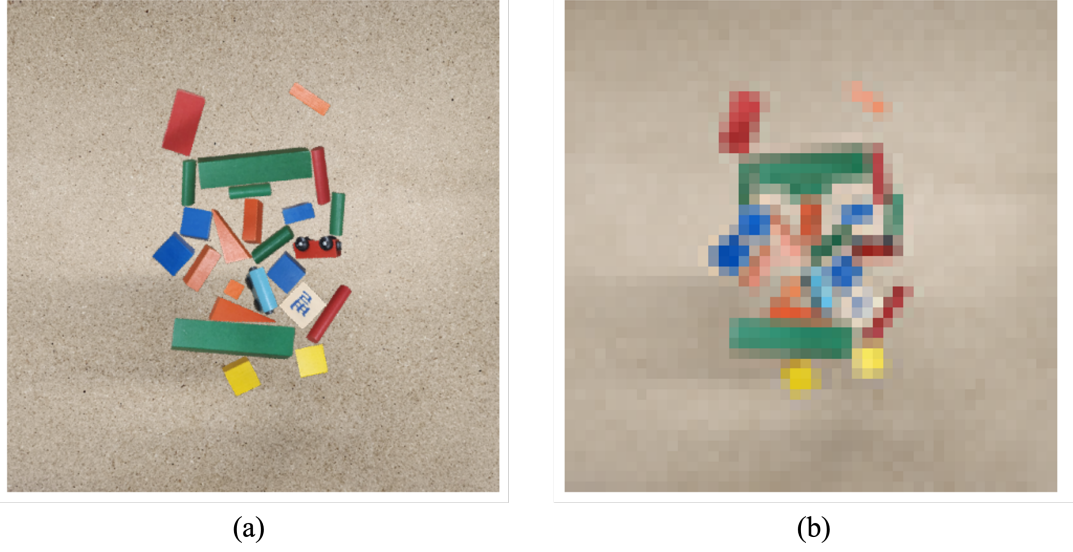


Figure 3.2: **Multi-Scale Rendering.** (a) High-resolution render capturing fine details and precise edge alignments; (b) Low-resolution render emphasizing global layout and overall shape placement; This provides for a robust semantic guidance and avoids overfitting to pixel-level noise. We use aggregated SDS loss computed by averaging losses across scales, ensuring that both coarse and fine semantic features guide the optimization.

optimization process effectively.

3.0.4 Content-Aware Collision Resolution

To generate collision-free layouts that are both geometrically precise and semantically coherent, we propose a unified collision resolution module that integrates geometric collision detection with cross-modal semantic guidance. Our method consists of two complementary components: a geometric stage based on the SAT for detecting and resolving overlaps, and a semantic stage that modulates the collision resolution using cross-modal information via a dynamic semantic graph.

For each object, represented as a shape S_i , parameterized by its position $\mathbf{p}_i = (x_i, y_i)$ and orientation α_i , we first compute a convex approximation (e.g., the convex hull) of its footprint. When two shapes S_i and S_j overlap, our SAT-based routine computes a minimal translation vector (MTV) \mathbf{m}_{ij} by considering candidate separating axes derived from the edges of the convex hulls. Specifically, for each candidate axis

\mathbf{a}_k , we project the vertices of both shapes and identify the axis \mathbf{a}_{ij}^* corresponding to the smallest overlap magnitude d_{ij} . The MTV is then given by:

$$\mathbf{m}_{ij} = d_{ij} \cdot \frac{\mathbf{a}_{ij}^*}{\|\mathbf{a}_{ij}^*\|},$$

which represents the smallest displacement required to separate the overlapping shapes along the optimal direction \mathbf{a}_{ij}^* .

To incorporate semantic context, we first extract a visual embedding for each shape S_i using a vision-language model (e.g., CLIP[20]) and obtain textual embeddings for a set of semantic concepts c using a large language model (e.g., GPT-4 Turbo). Let ϕ and ψ denote learnable projection functions that align the visual and textual embeddings, respectively. We then compute a semantic attention weight for each concept as

$$a_{c,i} = \frac{\exp\left(\phi(\text{CLIP}(S_i)) \cdot \psi(\text{GPT}(c))\right)}{\sum_{c'} \exp\left(\phi(\text{CLIP}(S_i)) \cdot \psi(\text{GPT}(c'))\right)}.$$

These attention weights $a_{c,i}$ quantify the relevance of concept c to shape S_i , creating a semantic-geometric bridge that guides collision resolution. In our implementation, these weights are managed within a dynamic semantic graph that captures contextual relationships among shapes, allowing for intelligent distribution of displacement vectors across multiple colliding objects while preserving semantic coherence.

The final displacement update for shape S_i is computed by blending the physical correction with the semantically preferred adjustment:

$$\Delta \mathbf{p}_i = \eta \frac{\sum_{j \in C_i} a_{c,i} \mathbf{m}_{ij}}{|C_i|}$$

where η is a scaling factor. In cases where a shape lacks a clear semantic assignment, a default scaling of the physical collision vector is employed. Both the physical and semantic components are integrated iteratively within an end-to-end optimization framework that updates the positions and orientations of all shapes concurrently.

By combining precise SAT-based geometric collision resolution with dynamic cross-modal semantic guidance, our approach produces natural, context-aware layouts. For example, when resolving collisions involving a sword (see 3.3, the content-aware

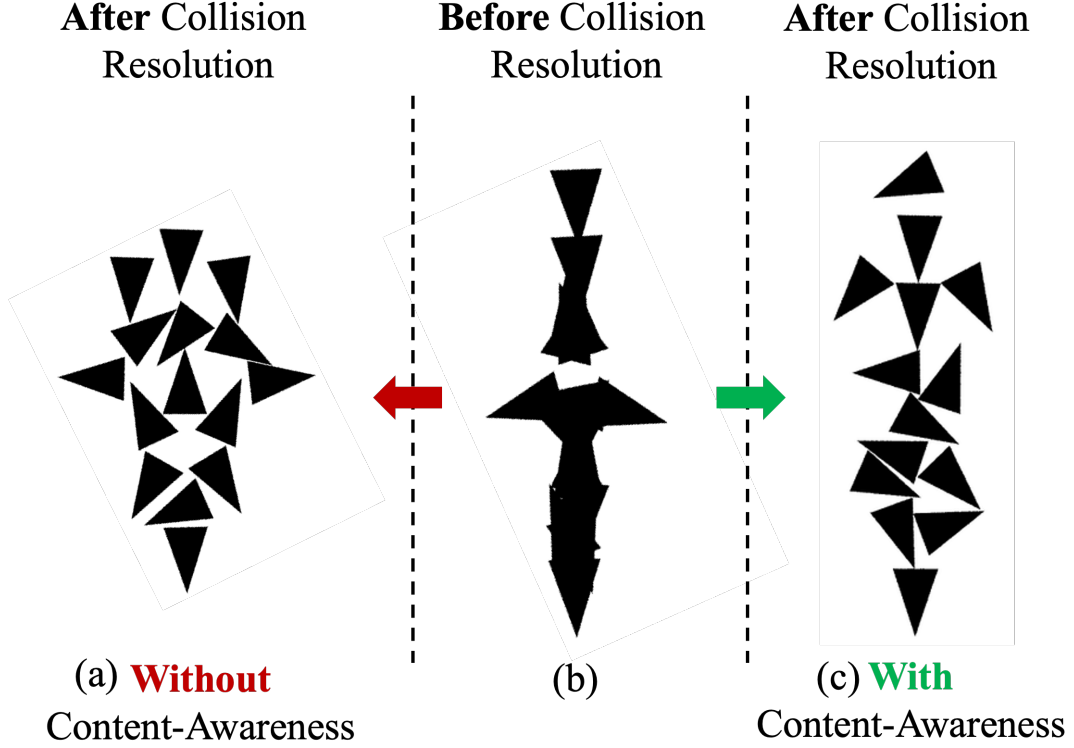


Figure 3.3: **Content-Aware Collision Resolution Process for the goal concept “Sword”** (a) Plain collision resolution relies solely on geometric penetration, often producing unnatural displacements. (b) The initial render state before collision detection. (c) Our content-aware method integrates extracted semantic concepts and dynamic attention weights to preserve object orientations and contextual relationships (e.g., maintaining a sword’s natural alignment).

mechanism preserves the blade’s intended alignment rather than displacing it arbitrarily.

3.0.5 Overall Optimization Process and Weighting

Our optimization process refines shape parameters $P = \{(x_i, y_i, \alpha_i)\}_{i=1}^N$ through a composite loss function balancing semantic fidelity with geometric coherence:

$$\mathcal{L}_{\text{total}}(P) = \lambda_{\text{SDS}} \cdot \mathcal{L}_{\text{MSDS}}(P) + \lambda_{\text{center}} \cdot \mathcal{L}_{\text{center}}(P),$$

where $\mathcal{L}_{\text{MSDS}}$ is the multi-scale SDS loss guiding semantic alignment with the input text prompt. We incorporate a centering regularization term:

$$\mathcal{L}_{\text{center}} = \frac{1}{N} \sum_{i=1}^N \|(x_i, y_i) - \mathbf{c}\|^2,$$

with \mathbf{c} denoting canvas center and empirically determined weights $\lambda_{\text{SDS}} = 1.0$ and $\lambda_{\text{center}} = 0.05$.

We implement a two-phase strategy addressing the trade-off between semantic exploration and collision avoidance. Initially, we optimize using only the composite loss function without geometric constraints, allowing shapes to explore semantically optimal regions while preventing local minima traps. This exploratory phase establishes a semantic context that informs our subsequent collision resolution about meaningful spatial relationships.

In the second phase, we activate content-aware collision resolution after each gradient update:

$$P_{t+1} = \mathcal{C}(P_t - \alpha \nabla_P \mathcal{L}_{\text{total}}(P_t)),$$

where $\mathcal{C}(\cdot)$ represents the collision resolution operator that applies minimal displacements based on both SAT-computed geometric corrections and semantically weighted adjustments. This progression from unrestricted exploration to constrained refinement produces layouts that satisfy both semantic relevance and geometric validity.

Chapter 4

Results

We rigorously evaluate ShapeShift through a series of quantitative metrics, baseline comparisons, and user studies. Our experiments assess both the geometric integrity and semantic fidelity of the generated layouts, addressing the central question: can our framework generate coherent and meaningful arrangements while maintaining collision-free configurations?

4.0.1 Adding Physical Constraints to Image Generators

While pixel-based methods can generate visually appealing images that align with the prompt semantically, they inherently operate in a continuous latent space that lacks the explicit geometric representation necessary for maintaining shape integrity. In Table 4.1, we show qualitative results of using state-of-the-art text-to-image models to rearrange objects based on a text prompt. We compare our ShapeShift method with text-guided image-to-image models Stable Diffusion Img2Img [24] and Instruct-Pix2Pix [1] along with an image and text conditioned video generation model, SORA [18] (displaying only the last frame from the generated video). To ensure a fair comparison, we augmented baseline prompts with various instructional suffixes (e.g., “maintain all original object shapes,” “keep objects separate and distinct,” “ensure no overlapping objects”) in an attempt to explicitly guide these models toward preserving primitive integrity, but observed consistently similar results.

The baseline pixel-generating models in Table 4.1 produce images that often



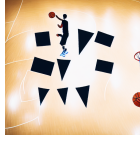












Reference	Text Prompt	Stable Diffusion	InstructPix2Pix	SORA	Ours
	“Arrange these shapes into a person playing basketball”				
	“Create a teapot using only these objects”				
	“Form a giraffe using these 10 triangles”				

Table 4.1: **Qualitative Comparison with Baselines.** We evaluate how different generative approaches respond to the task of arranging specific objects and basic shapes according to semantic prompts. While Stable Diffusion Img2Img [24] introduces extraneous elements and distortions, InstructPix2Pix [1] sacrifices discrete shape boundaries for photorealism, and SORA [18] exhibits good quality objects but inconsistent spatial organization despite its temporal capabilities. ShapeShift (ours) preserves both the integrity of each input primitive and their semantic arrangement, demonstrating the advantage of explicit geometric parameterization over pixel-space transformations when physical constraints must be maintained.

contain semantic features of the prompt but lack any physical grounding as they often remove or distort the objects in the input image. These limitations are not shortcomings of the models themselves but rather a consequence of their design paradigm—transforming pixels rather than manipulating parameterized objects. While highly effective for general image generation, these approaches cannot maintain geometric integrity. The explicit geometric control, combined with semantic guidance from large vision-language models in ShapeShift, enables applications that would be inherently challenging for pixel-based methods.

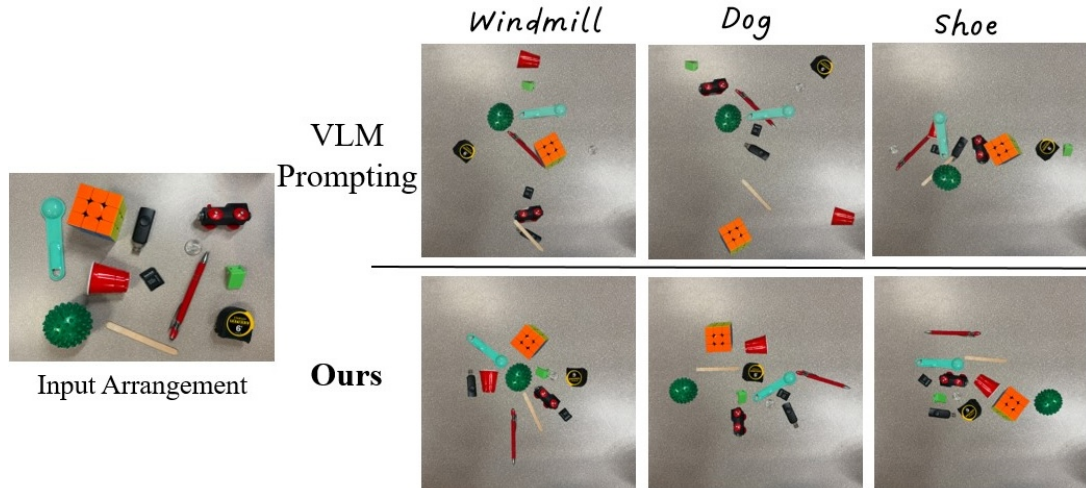


Figure 4.1: **Comparison with VLM Planning.** The figure shows arrangements of everyday objects into three semantic concepts (Windmill, Dog, Shoe). The top row displays arrangements following the VLM’s coordinate and orientation specifications, while the bottom row shows our method’s results for the same prompts.

4.0.2 Improving Semantic-Guidance in Vision-Language Model Planning

Vision-Language Models (VLMs) have successfully been used to plan new arrangements of table top items [5, 16], rearrange furniture [3, 29], and add strokes to complete partial drawings [30]. We evaluated ChatGPT-4 as a representative zero-shot planner by providing it with (1) an image of the initial shape arrangement, (2) the target goal concept, and (3) an explicit instruction to output coordinates and orientations for each shape. To ensure fairness, we provided a coordinate grid overlay with labeled shapes and instructions to preserve shape integrity while avoiding overlaps. For collision cases, we iteratively requested refined positions based on grid feedback.

The results reveal both the capabilities and limitations of current language-vision models for arrangement tasks. ChatGPT-4 demonstrates reasonable spatial reasoning for basic semantic concepts, often producing recognizable high-level configurations. However, it struggles with physical constraints consistent with [3], producing significant shape overlaps and imprecise semantic alignment. These limitations stem from its one-shot planning approach without geometric feedback, inability to optimize iteratively, and difficulty reasoning about precise spatial relationships among multiple

Object Count	Method	Overlap % (\downarrow)	# Coll. (\downarrow)	CLIP Score (\uparrow)
10	SDS only	28.60	5.94	0.2328
	SDS + SAT	0.68	0.33	0.1421
	Ours	0.70	0.38	0.2285
15	SDS only	35.25	12.60	0.2364
	SDS + SAT	0.24	0.59	0.1222
	Ours	0.27	0.65	0.2265
20	SDS only	36.83	19.51	0.2425
	SDS + SAT	1.49	1.71	0.1648
	Ours	1.12	1.55	0.2288
25	SDS only	41.15	22.13	0.2611
	SDS + SAT	1.52	1.92	0.1985
	Ours	1.78	2.10	0.2502

Table 4.2: **Ablation Study.** Comparison of three variants of our pipeline across different shape counts. We report the average overlap area percentage (Overlap %), number of collisions (# Coll.), and CLIP Score.

objects simultaneously consistent with [30].

4.0.3 Ablation Study

To understand the effect of each component of ShapeShift, we performed an ablation study. Table 4.2 presents a comparison of three variants: (1) using only Score Distillation Sampling (SDS), (2) SDS with SAT for basic collision detection and resolution (SDS+SAT), and (3) our full approach, ShapeShift, which incorporates content-aware collision resolution. We evaluated these variants on arrangements with different object counts using three metrics: average overlap area percentage (lower is better), number of collisions (lower is better), and CLIP score measuring semantic alignment with the prompt (higher is better).

The ablation results reveal several key insights. First, using SDS alone produces semantically rich layouts but suffers from significant overlaps that increase with the number of objects. Adding SAT-based collision resolution substantially reduces overlaps but at the expense of semantic alignment, as shapes are displaced based purely on geometric considerations without context. Our full approach, which incorporates

content-aware collision resolution, achieves the best balance—maintaining strong semantic alignment while drastically reducing overlaps compared to SDS-only and yielding more natural, contextually appropriate arrangements than SDS+SAT.

4.0.4 Human Rater Evaluation

Image semantics can be subjective and difficult to evaluate automatically. CLIPScore gives a sense of how well the arranged shapes fit the given text description. However, to measure semantic comprehensibility beyond these computational metrics, we conducted a controlled identification study with human raters on Amazon Mechanical Turk. Participants were presented with an image of the arranged objects (48 unique images for each method, see 4.2) and then asked to select what concept they see in the image, given four different options. Only one option was correct, creating a forced-choice task with 25% random baseline accuracy. We evaluated both complete arrangements and versions with 30% of objects randomly removed. Examples can be seen in Figure 4.2. Our results presented in Table 4.3 were based on 576 ratings (144 for each of the six methods) from 161 unique participants.

Our participants found the arrangements in SDS and ShapeShift results easiest to identify. While our approach had a larger accuracy, there was no statistical significant difference (p-value= 0.15). However, the object removal test exposed a crucial distinction: while removing shapes from the image made no statistical impact on recognizability for SDS and SDS+SAT, for our approach, removing shapes *significantly* hurt the ability to recognize content. This result implies possibly overlapping shapes in the configurations generated by the baselines, which is consistent with the overlap measures. By contrast, this indicates that each object in our ShapeShift arrangement is important for recognizability, showing evidence that the Content-Aware Collision Avoidance is accurately understanding how each shape contributes to the overall semantics of the arrangement.

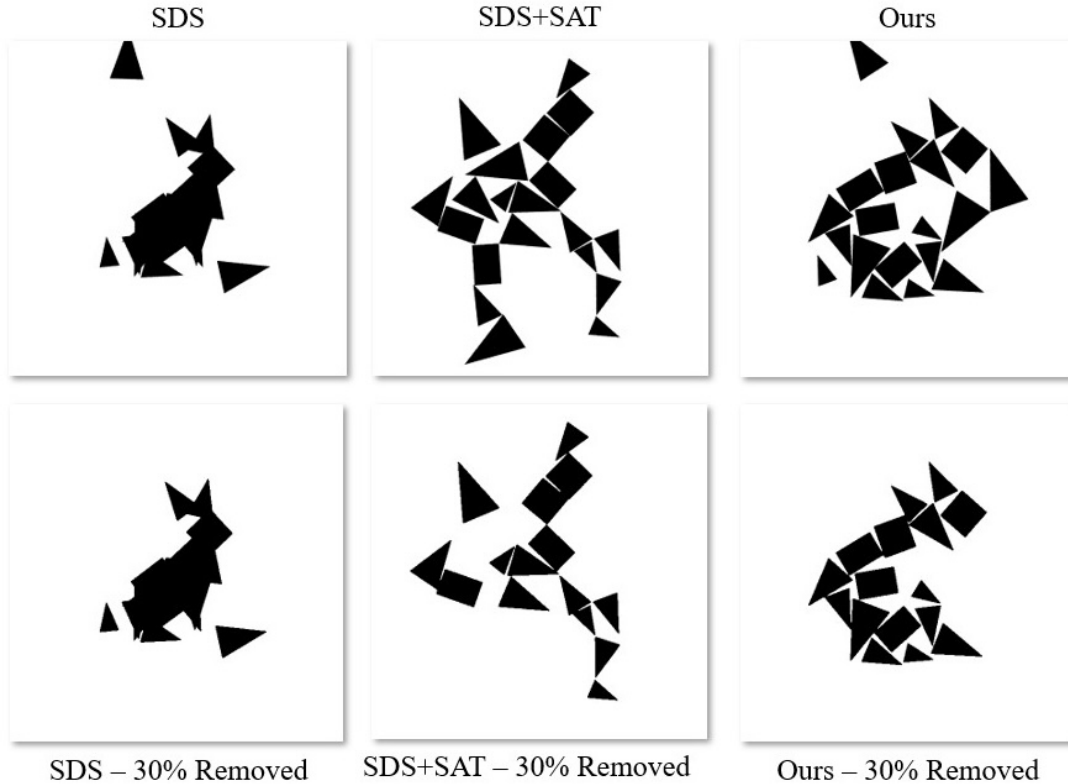


Figure 4.2: **Human-Rater Study.** Example arrangements of the “rabbit” concept generated using three methods (columns) and shown with complete or partial (30% removed) object sets (rows).

	CLIP Score	Accuracy	Accuracy w/ Blocks Removed	Effect of Block Removal
SDS	0.2579	46.9 ± 5.1	51.0 ± 5.1	$+4.2 \ p = 0.566$
SDS+SAT	0.2349	38.5 ± 5.0	31.3 ± 4.8	$-7.3 \ p = 0.292$
Ours	0.2526	57.3 ± 5.1	39.5 ± 5.0	$-17.7^* \ p = 0.014$

Table 4.3: **Human-Rater Results.** We report CLIP Scores for survey images (example in Figure 4.2) alongside participant accuracy in identifying target concepts, along with standard error. Comparing accuracy before and after random 30% image removal reveals that the only significant difference occurred with our proposed method.

Chapter 5

Conclusion

We presented **ShapeShift**, a framework that synthesizes semantically meaningful, collision-free arrangements of shape primitives from text prompts. Our approach bridges semantic guidance and geometric control through: (1) minimal shape parameterization (position and orientation), (2) multi-scale rendering for hierarchical relationships, and (3) content-aware collision resolution preserving semantic intent. Evaluations demonstrate ShapeShift outperforms pixel-based approaches on physical validity while maintaining semantic coherence, with user studies confirming both technical soundness and perceptual clarity.

Chapter 6

Discussion and Limitations

ShapeShift demonstrates the efficacy of combining diffusion-based semantic guidance with explicit geometric parameterization but faces several challenges. Arrangement quality deteriorates with increasing shape count—both overlap percentage and collision count rise from 0.70% with 10 shapes to 1.78% with 25 shapes (Table 4.2). This stems from canvas space constraints and the challenge of satisfying multiple constraints simultaneously. Our SAT-based collision detection, while effective for convex shapes, requires decomposition for concave geometries and scales quadratically with primitive count, introducing computational overhead for complex scenes.

Our experiments reveal a fundamental trade-off between semantic alignment and geometric integrity. Both automatic metrics and human evaluation (Tables 4.2 and 4.3) show that naive collision resolution (SDS + SAT) significantly degrades semantic quality compared to unconstrained optimization (SDS only). Our Content-Aware Collision Resolution successfully mitigates this tension, maintaining semantic clarity while ensuring physical plausibility. The shape removal experiment further confirms the semantic efficiency of our approach—human evaluators had a much more difficult time identifying the semantics compared to our ablation methods.

While effective for 2D arrangements, ShapeShift is currently limited to top-down representations with position and orientation parameters. Future work could extend this framework to three-dimensional space for robotic assembly tasks, building upon approaches like Dream2Real [12] to bridge high-level instructions with precise spatial control.

Chapter 7

Future Work

While ShapeShift demonstrates significant advances in text-guided shape arrangement synthesis with geometric constraints, several promising research directions remain unexplored. We outline two major extensions that could substantially expand the impact and applicability of this work.

7.0.1 Robotic Implementation of 2D Arrangements

A natural extension of our work is the physical realization of ShapeShift arrangements using robotic manipulation. This would bridge the gap between virtual design and physical assembly, enabling tangible interaction with semantically meaningful compositions. Implementing such a system presents several key challenges:

- **Perception-Action Alignment:** Translating the precise position and orientation parameters from our differentiable framework to robotic actuation requires addressing uncertainties in perception, grasping, and placement. Robust object recognition and localization would be essential for accurately reproducing ShapeShift’s virtual arrangements in the physical world.
- **Sequential Assembly Planning:** With the final position and orientation parameters already determined by ShapeShift, the robotics challenge becomes one of optimal execution sequence. Standard motion planning algorithms can determine collision-free paths to place each object in its designated position and orientation. The key challenge is determining an efficient placement order

that minimizes the risk of disturbing already-positioned objects.

- **Physical Constraints:** Real-world manipulation introduces additional considerations such as friction, stability, and grasp limitations that are not modeled in our current framework. Extending our collision resolution mechanism to account for these physical factors while preserving semantic intent would enhance the practicality of robotic implementations.

Following recent approaches in language-guided robotic manipulation [12], we envision a system where ShapeShift provides the high-level arrangement planning while low-level robotic controllers handle the physical execution. This integration would enable applications in educational robotics, assistive technologies, and interactive design where meaningful physical arrangements can be created from simple textual descriptions.

7.0.2 Extension to 3D Shape Assemblies

Extending ShapeShift to three-dimensional space would significantly expand its expressive capabilities and practical applications. While conceptually similar to our 2D approach, 3D arrangement introduces several fundamental challenges:

- **Expanded Parameter Space:** In addition to 2D position (x, y) and planar rotation α , 3D arrangements require z-coordinates and full 3D rotational parameters (quaternions or Euler angles), substantially increasing the optimization complexity. This expanded parameter space would necessitate more sophisticated regularization techniques to ensure convergence while maintaining semantic alignment.
- **Physics-Based Constraints:** Beyond simple non-overlap constraints, 3D assemblies must account for gravity, balance, and structural stability. Integrating differentiable physics simulation within our optimization framework would allow for arrangements that are both semantically meaningful and physically stable without external support.
- **Viewpoint Ambiguity:** 3D arrangements can be viewed from multiple angles, complicating the semantic alignment with text prompts. Multi-view Score Distillation Sampling would be necessary to ensure the arrangement is recognizable

7. Future Work

from different perspectives, requiring extensions to our rendering pipeline and loss functions.

- **Hierarchical Assembly:** Complex 3D structures often involve nested or stacked components. Our content-aware collision resolution would need to be extended to handle hierarchical relationships between objects, potentially incorporating dependency graphs to represent structural requirements in the final assembly.

The extension to 3D would enable applications in architectural design, educational tools for spatial reasoning, and automated construction of physical models from textual descriptions. By leveraging advances in 3D diffusion models and differentiable rendering, ShapeShift could become a powerful tool for bridging natural language understanding with physical 3D design.

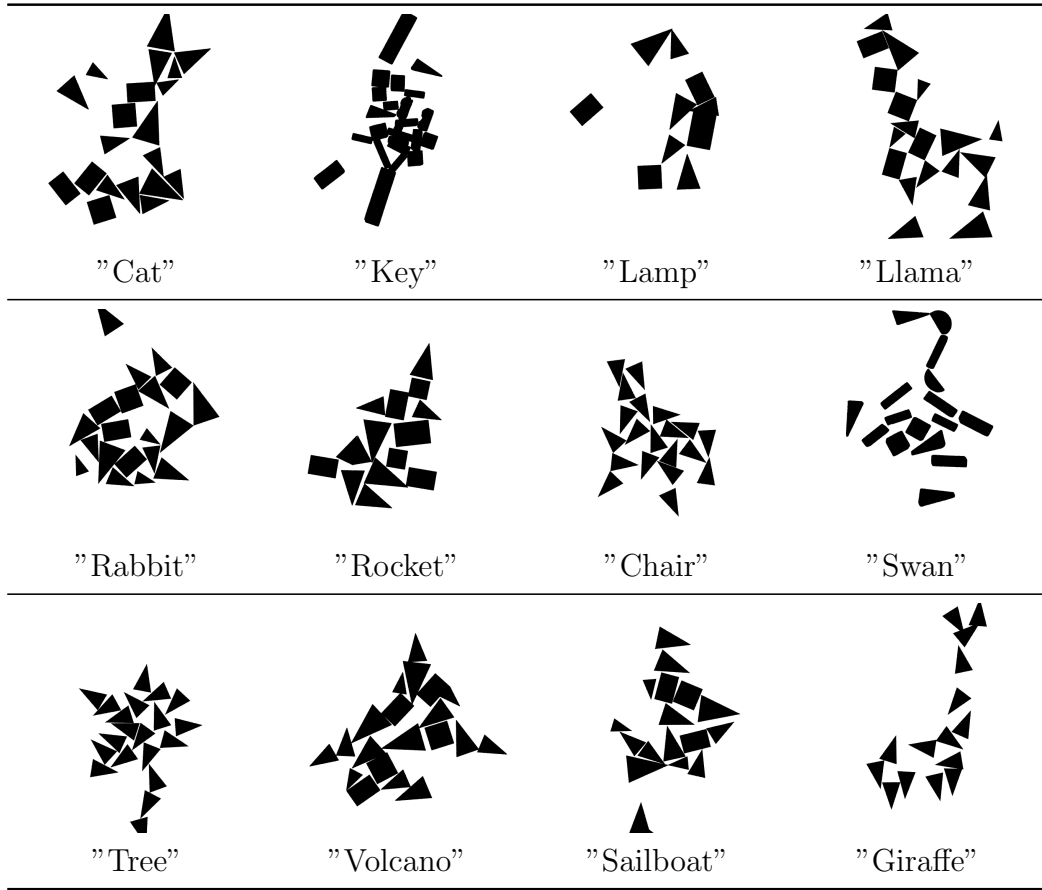
These future directions would not only advance the technical capabilities of our system but also expand its practical impact across domains such as education, design, robotics, and human-computer interaction. By maintaining our core approach of combining semantic guidance with explicit geometric parameterization, these extensions would preserve the interpretability and physical validity that distinguish ShapeShift from purely pixel-based generative methods.

Appendix A

Appendix

This appendix presents additional qualitative results from the ShapeShift framework across a diverse set of semantic concepts and shape configurations. Table [A.1](#) showcases the versatility of our approach in generating recognizable arrangements from varying numbers of primitive shapes. These examples were selected to demonstrate the range of semantic concepts that can be effectively communicated through minimal shape arrangements while adhering to our strict non-overlapping constraints.

Table A.1: Additional arrangement examples generated by ShapeShift. Each arrangement uses the exact same set of shapes from the input image, rearranged to match the specified concept. The numbers in filenames indicate the experimental batch identifier.



A.1 Effect of Shape Count on Semantic Alignment

While the main paper examines the relationship between shape count and collision metrics, this appendix further investigates how the number of available shapes affects semantic alignment with the target concept. Intuitively, one might expect that more shapes allow for greater expressivity and thus better semantic alignment. However, as shown in Figure A.1, this relationship is non-linear and exhibits interesting characteristics.

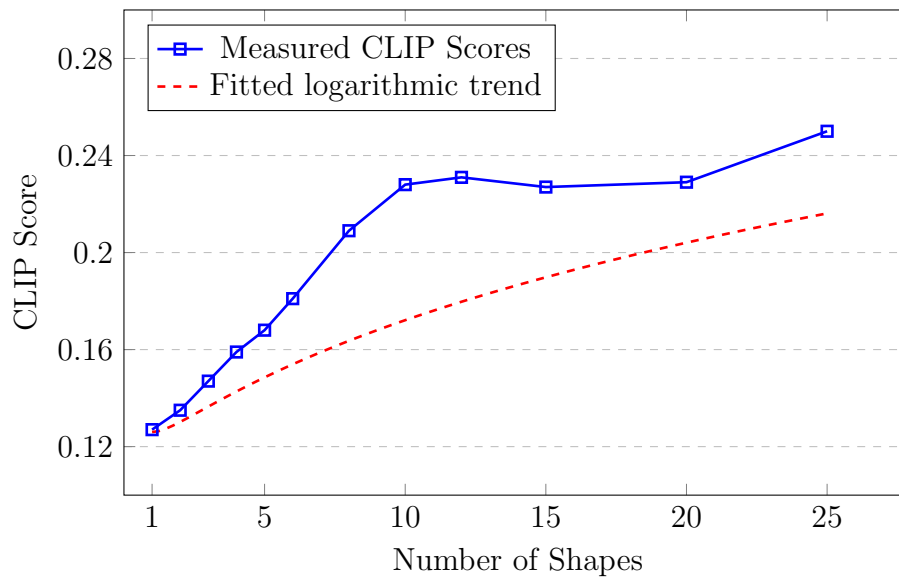


Figure A.1: Relationship between number of shapes and semantic alignment (measured by CLIP score). With fewer than 6 shapes, arrangements struggle to capture sufficient detail for strong semantic alignment. Between 10-20 shapes, semantic clarity plateaus, suggesting an optimal range for balancing expressivity and arrangement simplicity. Beyond 20 shapes, we observe further improvements as more complex concepts become representable.

Table A.2 provides a more detailed analysis of this relationship, including both quantitative metrics and qualitative observations across different shape count ranges.

These findings suggest that while more shapes generally improve semantic alignment as measured by CLIP score, there exists a “sweet spot” of approximately 8-15 shapes that balances expressivity with arrangement simplicity. This insight could

Table A.2: Detailed analysis of shape count effect on arrangement quality and semantic alignment

Shape Count	CLIP Score	Overlap %	Qualitative Observations
1-3	0.127-0.147	0.00-0.05	Extremely limited expressivity; only very simple concepts recognizable (e.g., basic letters)
4-6	0.159-0.181	0.08-0.25	Recognizable but highly abstracted; limited to concepts with distinctive silhouettes
7-10	0.195-0.228	0.38-0.70	Good balance of abstraction and recognizability; suitable for many common concepts
11-15	0.230-0.227	0.27-0.41	Strong expressivity with diminishing returns; collision resolution becomes more challenging
16-20	0.228-0.229	1.12-1.33	Marginal improvements in semantic clarity; arrangement complexity increases significantly
21-25	0.235-0.250	1.78-2.01	Highest semantic alignment but with notable increase in arrangement optimization difficulty

inform future work in both computational design and educational applications, where finding the minimal number of elements needed to express a concept clearly is often desirable.

Notably, the logarithmic trend visible in Figure A.1 indicates diminishing returns as shape count increases, with the most dramatic improvements occurring in the transition from very few shapes (1-6) to a moderate number (7-15). This pattern aligns with human cognitive principles of perception, where recognizability often follows similar non-linear relationships with visual complexity.

Bibliography

- [1] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. URL <https://arxiv.org/abs/2211.09800>. (document), 4.0.1, 4.1
- [2] 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. (document), 1, 1.2
- [3] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36:18225–18250, 2023. 4.0.2
- [4] Kevin Frans, L. B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders, 2021. URL <https://arxiv.org/abs/2106.14843>. 2
- [5] Andrew Goldberg, Kavish Kondap, Tianshuang Qiu, Zehan Ma, Letian Fu, Justin Kerr, Huang Huang, Kaiyuan Chen, Kuan Fang, and Ken Goldberg. Blox-net: Generative design-for-robot-assembly using vlm supervision, physics simulation, and a robot with reset. *arXiv preprint arXiv:2409.17126*, 2024. 1, 2, 4.0.2
- [6] S. Gottschalk, M. C. Lin, and D. Manocha. *OBBTree: A Hierarchical Structure for Rapid Interference Detection*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. ISBN 9798400708978. URL <https://doi.org/10.1145/3596711.3596791>. 2, 3.0.2
- [7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL <https://arxiv.org/abs/2006.11239>. 2
- [8] Juncheng Hu, Ximing Xing, Jing Zhang, and Qian Yu. Vectorpainter: Advanced stylized vector graphics synthesis using stroke-style priors, 2024. URL <https://arxiv.org/abs/2405.02962>. 2
- [9] Shir Iluz, Yael Vinker, Amir Hertz, Daniel Berio, Daniel Cohen-Or, and Ariel

- Shamir. Word-as-image for semantic typography, 2023. URL <https://arxiv.org/abs/2303.01818>. 2
- [10] Ajay Jain, Amber Xie, and Pieter Abbeel. Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models, 2022. URL <https://arxiv.org/abs/2211.11319>. 1, 2, 3.0.1
- [11] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. Dall-e-bot: Introducing web-scale diffusion models to robotics. *IEEE Robotics and Automation Letters*, 8(7):3956–3963, 2023. 2
- [12] Ivan Kapelyukh, Yifei Ren, Ignacio Alzugaray, and Edward Johns. Dream2real: Zero-shot 3d object rearrangement with vision-language models, 2024. URL <https://arxiv.org/abs/2312.04533>. 1, 2, 6, 7.0.1
- [13] Tzu-Mao Li, Michal Lukáč, Michaël Gharbi, and Jonathan Ragan-Kelley. Differentiable vector graphics rasterization for editing and learning. *ACM Trans. Graph.*, 39(6), November 2020. ISSN 0730-0301. doi: 10.1145/3414685.3417871. URL <https://doi.org/10.1145/3414685.3417871>. 2, 3, 3.0.3
- [14] Patrick Lindemann. The gilbert-johnson-keerthi distance algorithm. 2009. URL <https://api.semanticscholar.org/CorpusID:17068679>. 2
- [15] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. URL <https://arxiv.org/abs/2303.05499>. 3
- [16] Weiyu Liu, Yilun Du, Tucker Hermans, Sonia Chernova, and Chris Paxton. Structdiffusion: Language-guided creation of physically-valid structures using unseen objects. *arXiv preprint arXiv:2211.04604*, 2022. 1, 2, 4.0.2
- [17] Weiyu Liu, Chris Paxton, Tucker Hermans, and Dieter Fox. Structformer: Learning spatial structure for language-guided semantic rearrangement of novel objects. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6322–6329. IEEE, 2022. 1, 2
- [18] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models, 2024. URL <https://arxiv.org/abs/2402.17177>. (document), 4.0.1, 4.1
- [19] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022. URL <https://arxiv.org/abs/2209.14988>. 1, 3, 3.0.1
- [20] 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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021. [2](#), [3.0.4](#)
- [21] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. [2](#)
- [22] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL <https://arxiv.org/abs/2408.00714>. ([document](#)), [3](#), [3.1](#)
- [23] Scott Reed, Zeynep Akata, Xincheng Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. [2](#)
- [24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. ([document](#)), [1](#), [2](#), [2](#), [3.0.1](#), [3.0.3](#), [4.0.1](#), [4.1](#)
- [25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with enhanced text understanding. pages 3647–3660, 2022. [1](#)
- [26] Peter Schaldenbrand, Zhixuan Liu, and Jean Oh. Styleclipdraw: Coupling content and style in text-to-drawing translation, 2022. URL <https://arxiv.org/abs/2202.12362>. [2](#)
- [27] 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. *Advances in neural information processing systems*, 35:25278–25294, 2022. [2](#)
- [28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL <https://arxiv.org/abs/2010.02502>. [2](#)
- [29] Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. Layoutvlm: Differentiable optimization of 3d layout via vision-language models. *arXiv preprint arXiv:2412.02193*, 2024. [2](#), [4.0.2](#)

- [30] Yael Vinker, Tamar Rott Shaham, Kristine Zheng, Alex Zhao, Judith E Fan, and Antonio Torralba. Sketchagent: Language-driven sequential sketch generation. *arXiv preprint arXiv:2411.17673*, 2024. [4.0.2](#)
- [31] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. [2](#)
- [32] 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*, 2(3):5, 2022. [2](#)
- [33] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. [2](#)