PIE-FRIDA: Personalized Interactive Emotion-Guided Collaborative Human-Robot Art Creation

Beverley-Claire A. Okogwu

CMU-RI-TR-24-18

May, 2024

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

Thesis Committee:
Dr. Jean Oh, *Chair*
Dr. Henny Admoni
Dr. Jim McCann
Peter Schaldenbrand

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

Abstract

The introduction of Generative AI has brought about many improvements in the artistic world. It allows many individuals to create artistic works via simple descriptive text prompts. This has, in particular, created an avenue for non-artistic individuals to express their thoughts through generated art.

Our work focuses on how emotion can be added as an additional modality to assist individuals and guide the system better to create more user-intended media in a collaborative space. In particular, we use audio as the primary input modality to best capture the emotion through the voice's raw tone, pitch, and pace. We hypothesize that users will best benefit from active emotional feedback during art generation in an interactive space instead of a simple generative pass or an interactive system without emotion analysis. Therefore, we propose using a personalized *Speech Emotion Recognition* system combined with a collaborative system, and the generation of desirable artistic media is obtained.

To address this, we consider (1) a personalized emotion calibration model, (2) an online emotion-guided *Interactive Detect and Respond* system from the finetuned model, and (3) the introduction of the personalized finetuned *Speech Emotion Recognition Detect and Respond* system in a collaborative artistic space. Our results support the advantage of introducing emotion in the generative space to foster a better collaborative experience.

Acknowledgments

First, I would like to thank my adviser, Dr. Jean Oh, for her guidance, patience, and flexibility during every aspect of the process and the overall opportunity to join the program. Without her support, I would be unable to pursue and complete this MSR thesis project.

I want to thank my committee members, Dr. Henny Admoni, Dr. Jim McCann, and Peter Schaldenbrand, for their many suggestions on building up and improving the project.

I also want to thank Zhixuan Liu, Lia Coleman, and Andrew Hundt for a collaborative experience with the *Multicultural Generative Media (M3C)* project. This experience not only introduced me to the HRI research space but also sparked the idea for my thesis.

Special thanks are also due to Rachel Burcin and John Dolan for inviting me to participate in the 2020 *Robotics Institute Summer Scholars (RISS)* program. It was this opportunity that fueled my developing passion for robotics-based projects.

Additionally, I extend my gratitude to the members of the *Bot Intelligent Group* for their constructive feedback on my practice talk. Their insights have immensely benefited my growth and confidence in presenting my work.

Finally, I would like to thank my parents and family members for their continuous love and support, which gave me the strength to move forward even when things were tough and I felt like giving up early in the process.

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

The intersection of artificial intelligence and creativity has witnessed unprecedented evolution over the past few years, heralding a new era in the artistic domain. However, the introduction of generative AI systems [14] has offered a platform where creativity is no longer bounded by technical proficiency. Users can now generate art through simple descriptive text prompts, thus bridging the gap between imagination and artistic realization.

Despite these advancements, the interaction between humans and these AI systems often remains surface in the art domain, primarily relying on the user to give specific instructions without a deeper understanding of the user's emotional intent [12]. In addition, for effective communication, humans must incorporate emotional feedback into their social and artistic interactions with other humans and intelligent systems [7, 36, 37].

Recognizing this gap, our work delves into an innovative approach incorporating emotion as a critical modality to enhance the collaborative process between the user and the robot. By leveraging audio inputs, we aim to capture and interpret emotional subtleties and use this as a way for the robot to form an interactive experience with the user that is often lost in text-based communication. This emotional insight promises to guide the generative process more effectively, creating more satisfying art for the user.

We hypothesize that incorporating emotional feedback actively during the art creation phase can significantly enrich the user experience, leading to outcomes that more accurately reflect the user's intentions. To achieve this, we propose a multifaceted approach involving a personalized emotion calibration model, an emotion-guided generative system finetuned to these calibrations, and deploying this enhanced system within a collaborative artistic space.

1.1 Motivation

The general foundation models are widely used and most common in today's Artificial intelligence space. They are built to be used by everyone, but they are not for everyone. A model that works for one person may not necessarily work for another. Improving the personalization of foundation models can significantly enhance user experience and effectiveness in applications virtual assistants [2]. Personalized models make the responses, suggestions, and interactions based on the individual user's preferences and behavior patterns, leading to more accurate, relevant, and engaging experiences. Conversely, a lack of personalization in foundation models could lead to less effective and engaging interactions, potentially causing user frustration due to generic or irrelevant responses, which we eventually want to avoid [2].

In our work, we aim to move up the levels of abstraction towards a more personalized model as shown in Figure 1.1. From this figure, the ultimate goal is to move away from the generic foundation models, and at each level, *personalize* some part of the models to cater more to our own needs.

In the *Multicultural Generative Media* [23, 24, 25] project, we leverage diffusion models, and the generated images produced are more diverse and cater to the user's needs. In Figure 1.2, we outline an example of how a diverse dataset can favor Nigerian culture images instead of the generic Stable Diffusion [30] model. Thus, we aim for more direct user-robot interaction by reaching a more personalized user experience and engaging with the user.

To achieve this, we utilize the FRIDA robot [33] to further expand the interaction space through art co-creation.

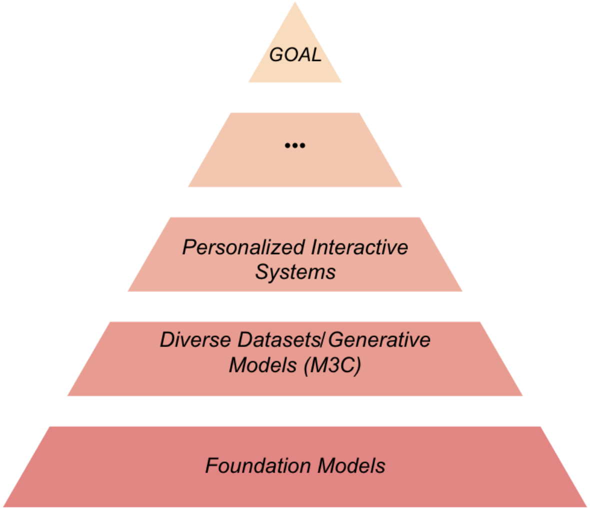

GOAL
•••
Personalized Interactive Systems
Diverse Datasets/Generative Models (M3C)
Foundation Models

Figure 1.1: High-Level Abstraction Diagram to achieve model personalization

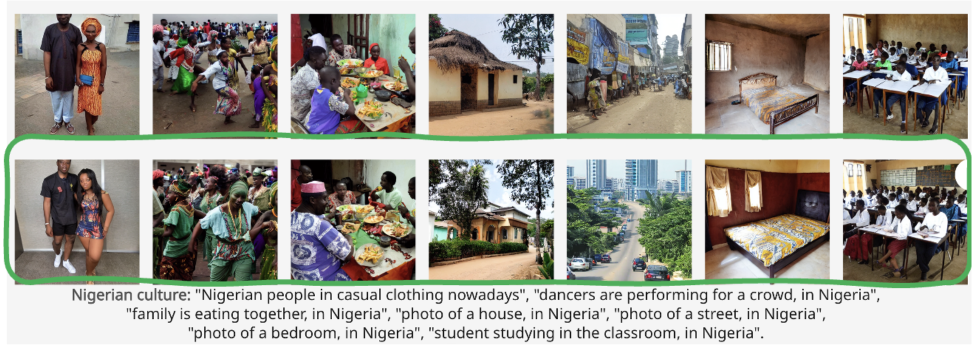

Nigerian culture: "Nigerian people in casual clothing nowadays", "dancers are performing for a crowd, in Nigeria", "family is eating together, in Nigeria", "photo of a house, in Nigeria", "photo of a street, in Nigeria", "photo of a bedroom, in Nigeria", "student studying in the classroom, in Nigeria".

Figure 1.2: Personalized Image Generation Tailored Towards a User from the Country, Nigeria

1.2 Thesis Organization

The following chapters are organized as follows:

- Chapter 2 reviews and summarizes existing work in this field;
- Chapter 3 describes the components leading to the emotion guidance;

- Chapter 4 compiles the set of studies and experiments that were carried out for testing the system's performance;
- Chapter 5 analyses the results obtained on the experiments, and;
- Chapter 6 provides final discussions and ongoing and future work.

Chapter 2

Related Work

2.1 Speech Emotion Recognition

Speech Emotion Recognition [10, 15, 35] is not a new concept and has been explored through various means. The overall framework adapted into the *Speech Emotion Recognition* models includes a series of classification tasks, with slight differences in the datasets, model architecture, and input streams.

This work focuses on integrating emotion into the generated art process, so choosing a good model is essential. We focus on the transformer-based WAV2VEC2 model[5] due to its ease of use and fine-tune, but other *Speech Emotion Recognition* models such as *Recurrent Neural Networks (RNNs)* [31], *Deep Neural Networks (DNNs)* [20], and *Domain-Adaptive Models* [13] can also perform similar tasks.

Although these *Speech Emotion Recognition* models are good at emotion recognition with an average user accuracy of about 79-80% [5, 20, 31], these models often fail to classify the user's audio correctly and need a good amount of attempts to do so. This can often be frustrating for the user and reduces the reliability of the system to the user. In our work, we address this through a ***Personalized*** *Speech Emotion Recognition* model that considers the user's own emotion meter.

2.2 Emotion in Image Generation

The exploration of emotion in image generation has garnered significant attention, seeking to enhance the expressiveness of generated imagery. Text-Guided Generative Adversarial Networks tailored for Image Emotion Transfer [41] were introduced, marking a significant stride toward bridging textual emotion descriptions with visual emotion conveyance. Similarly, StarGAN-EgVA [40], a model adept at Emotion-guided continuous affect synthesis, offers nuanced emotion manipulations within generated images. Xu et al. [39] further this discourse with their work on High-fidelity generalized emotional talking face generation, which innovatively incorporates multi-modal emotion space learning to enhance the emotional expressiveness of talking faces.

However, in our work, we are not focused on how the images are generated but on how using emotion as an additional input can help generate more satisfying output in a collaborative environment, improving the user's experience.

Furthermore, work has been done to create visuals from emotional input: Krcadnic et al. utilize Hoolovoo Visualizations [19] to create shades to show the strength of the emotional input from weak happiness to strong happiness. In addition, Multiconditional StyleGANs [18] has also been further modified to adapt art creations, taking in emotion and a prompt description of an artwork [9]. Robot Synesthesia[27]

These works, however, only use text input and are one-pass without feedback or check-in with the user. In addition, the Robot Synesthesia work also uses a generic *Speech Emotion Recognition* classifier to obtain the resulting visuals.

2.3 Interactive Feedback in Robotics

Over time, there have been many advancements with feedback in the *Human-Robot Interaction* space. Interactive feedback mechanisms represent a cornerstone in advancing robotics, facilitating more nuanced and effective human-robot interactions.

Early studies laid the groundwork by illustrating the critical role of immediate feedback in robotic systems for task efficiency and user satisfaction [34]. Building on this foundation, adaptive feedback algorithms were introduced that significantly enhanced robots' ability to understand and respond to human actions and intentions in real-time [6, 16, 21].

These contributions collectively underscore the transformative impact of interactive feedback on human-robot interactions. In our work, however, we make the user robot more personalized and tailored to the user.

2.4 CoFRIDA

"CoFRIDA: Self-Supervised FineTuning for Human-Robot Co-Painting" [33] is a framework that enhances human-robot collaboration in art, specifically in drawing and painting. Building upon prior work by the FRIDA system [32], which closed the simulation-to-reality gap and improved user interaction modalities during the initial stages of painting tasks. The CoFRIDA system shows improved alignment with user-provided text prompts. It can continue painting on canvases already started by humans without unnecessary overwrites, which maintains the collaborative artwork steps.

Currently, the CoFRIDA system primarily utilizes text-based input, with the only communication between the user and the robot being the FRIDA brush strokes and the display on the screen. In our work, we aim to close this gap by adapting audio input, incorporating emotion into the work, and having the system talk back to the user based on the emotion.

Chapter 3

Approach

To understand how integrating emotion makes for a better collaborative experience, we explore different stages that bring together user customization and robot interaction. This chapter describes the steps to ensure a seamless procedure in attaining a personalized user/robot experience with collaborative art in the *Human-Robot Interaction* space.

In finding ways to incorporate emotion into the generated media, we initially created a short video by taking in audio input, extracting the emotion from the audio, and stitching the generated frames together to create the video, which can be viewed from the following link. This work, however, used a generic emotion classifier and did not foster an interactive environment. Thus, we change the system to utilize a more *personalized Speech Emotion Recognition* system and be more collaborative using CoFRIDA[33].

Our novelty here lies in the *personalization*. We take the user's own voice and fine-tune the *Speech Emotion Recognition* model to introduce the bias to favor emotion recognition based on the user. Through *Personalized Interactive Emotion-Guided Collaborative Human-Robot Art Creation (PIE-FRIDA)* , we go up the model customization abstraction ladder by building from the ground up Figure 1.1. The following subsections discuss the sub-levels used in building up PIE-FRIDA.

3.1 Audio Data Collection

Current *Speech Emotion Recognition* models [15, 35, 38] have since been defined for generic use, where they are trained over large datasets, often in a language different from the intended user or used for a specific purpose not intended by the user. Standard datasets include the EMODB [4], a collection of German sentences, and the LSSED [11], an English dataset focusing on mental health datasets. In addition, these datasets are often collected by different people. Although this is standard procedure in *Speech Emotion Recognition* systems, this makes the process easy to misclassify the user intent.

A bigger problem includes the range of each individual's emotion meter. A person's happy emotional state can easily be perceived as sad or neutral. In the case of audio, people with different vocal intonations from the majority would find it challenging to benefit from generic *Speech Emotion Recognition* systems. To combat this, we collect each user's audio data and train to introduce bias in the *Speech Emotion Recognition* model, allowing the user to get more accurate emotion classifications. The data collection and fine-tuning process is summarized in Figure 3.1, Figure 3.2, and Section 3.2.

To collect the audio:

1. We create an online recording User Interface:
   (a) The users use their system's audio input device to record the provided sentences in 8 different emotions (*happiness, sadness, anger, disgust, neutral, calm, surprise, and fearful*) as defined by RAVDESS dataset[26]. The user will have 16 sentences: 2 sentences per emotion.

   (b) Different images and media are also shown along with each sentence to aid the required emotion to be evoked. For example, if we want the user to be surprised, we show them a surprising video favoring a more realistic "surprised" emotion. The goal here is to prime the user better to get their true audio for the particular emotion.

   (c) After all sentences are recorded, the user downloads a zip of the WAV recordings.
2. We then use the recordings obtained from the UI to train a new *Speech Emotion*

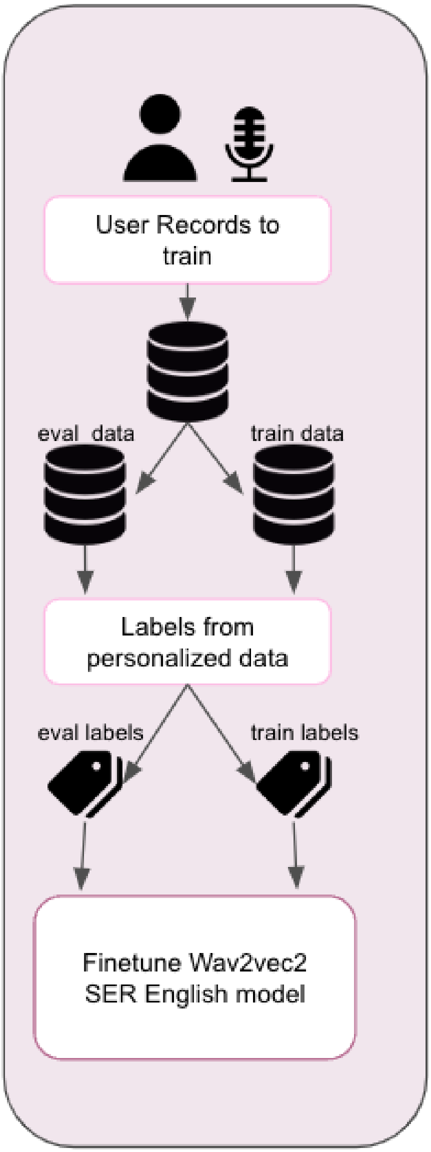

User Records to train
eval_data
train data
Labels from personalized data
eval labels
train labels
Finetune Wav2vec2 SER English model

Figure 3.1: An overview of the audio collection process.

Recognition system and fine-tune on an existing wav2vec2 *Speech Emotion Recognition* model [5, 29] to create a more *biased* and *personalized* *Speech Emotion Recognition* system.

3.2 Finetuning User Audio

The core of the fine-tuning process involves preparing and processing audio files to generate a suitable dataset for model training. We load the user's recorded audio files and perform data augmentation by adding noise, shifting time, and varying speed to enhance the model's robustness against real-world variations in input data.

The adaptation of the pre-trained Wav2Vec2 model is centered around customizing its configuration to align with the number of emotion categories. Each audio sample is processed to extract features using the Wav2Vec2 processor, followed by a forward pass through the model to obtain preliminary embeddings. These embeddings are then passed through additional classifier layers to predict the probability distribution over the emotion labels. Training involves adjusting the model's parameters by minimizing the self-contrastive loss between predicted probabilities and true labels, using gradient descent methods to improve the model in the training loop iteratively. The overall High-Level process is shown in Figure 3.2.

Following fine-tuning the Wav2Vec2 model on the individual users' audio, we would obtain the *Personalized Speech Emotion Recognition* model further needed for our system's emotion classification task.

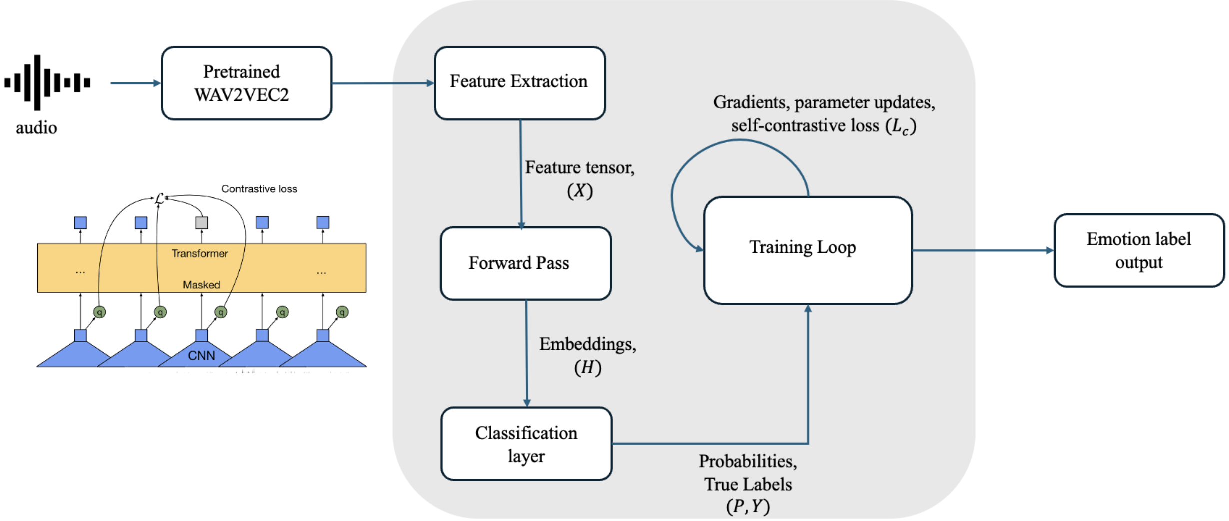


Figure 3.2: A High-Level overview of the Finetuning Process.

3.3 Detect and Respond

To simulate the interaction in an online environment, we first introduce the concept of *Detect and Respond* , which, by definition, detects the user's current emotional state while speaking a prompt and provides an image that reflects the detected emotion and the prompt. In Figure 3.3, we summarize the *Detect and Respond* system as follows:

1. The user speaks their prompt.
2. The audio obtained is:
   (a) Passed into an existing wav2vec2 *Speech Emotion Recognition* model [5] to extract the depicted emotion of the user, and is
   (b) transcribed to English text.
3. The transcribed text is then passed into the GPT-4 language model[28], which modifies the text prompt to give a more descriptive scene representation of the emotion called the *Emoprompt* .

–

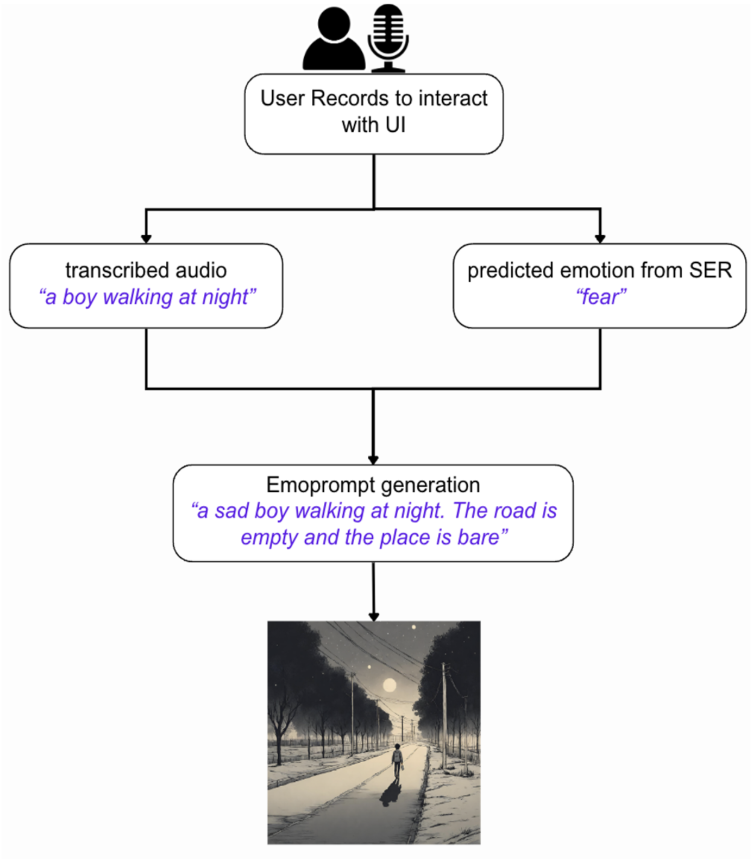

User Records to interact with UI
transcribed audio
"a boy walking at night"
predicted emotion from SER
"fear"
Emoprompt generation
"a sad boy walking at night. The road is empty and the place is bare"

Figure 3.3: An overview of the *Detect and Respond* process.

4. The *Emoprompt* is then used as input and is passed into an image generation model such as *Stable Diffusion* [30] to obtain the generated image.

The motivation for *Detect and Respond* involves finding a way to begin the analysis of whether or not it is possible to create an image that portrays the intended emotion of the user. This differs from existing work that uses image generation models to change elements of the image to evoke the emotion in facial synthesis [10, 39, 40, 41] or emotion editing[22]. Rather than being concerned with the *Generative Adversarial Network (GAN)* [14] architecture, and how it can be modified to accommodate emotional depiction, we are mainly concerned with gauging the user's prompt and emotional state to facilitate better interaction.

3.4 Interactive Detect and Respond

Although *Detect and Respond* gives the foundational approach in *Personalized Interactive Emotion-Guided Collaborative Human-Robot Art Creation* , it is a one-way system as it only provides the emotion and text input and produces a single image output. However, for a good collaborative system, we expect multiple rounds of user-system communication. With *Interactive Detect and Respond* , we extend the *Detect and Respond* system to engage the user to continuously obtain the best-desired image architecture. In Figure 3.4, we highlight the steps involved in the *Interactive Detect and Respond* process:

1. At the first iteration, the system operates just like *Detect and Respond* except for storing the perceived emotion.
2. for consecutive iterations, when a change in emotion is detected, we can assume that the user is currently dissatisfied with the image generated. The system then "talks back" to the user to know if the user wants a different image/emotional state depicted in the image. This step is important as it provides a way by which the user can provide feedback to the system and *vice versa* :
    (a) If the user does not want to have their current emotional state depicted in the image at the time of notification, then the system moves on to the next iteration.
    (b) But if the user wants a change, the system will derive three *Emoprompts* and subsequently show new image options for the user to choose from by either using the new emotion state, random choice of the new emotion from the system as a suggestion or the user manually selects the preferred emotion state.
3. This process continues until the desired image is obtained.

Now that we have established interaction via online simulation, we can deploy to the CoFRIDA system.

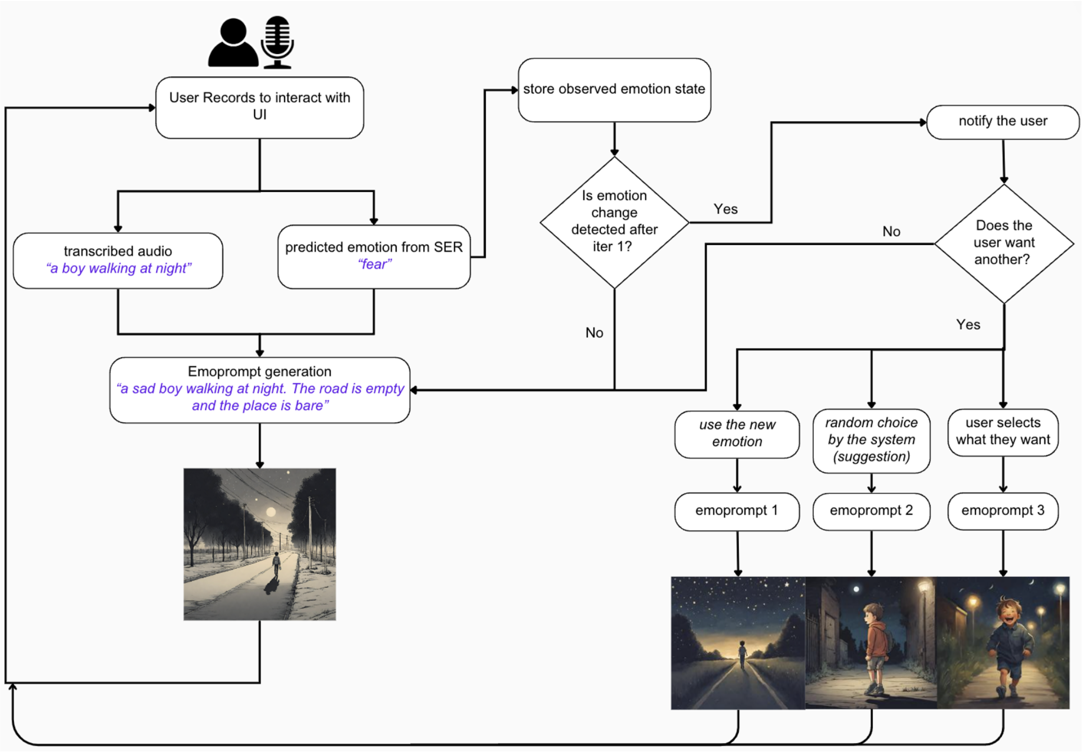

User Records to interact with UI
transcribed audio
“a boy walking at night”
predicted emotion from SER
“fear”
store observed emotion state
Is emotion change detected after iter 1?
Yes
No
notify the user
No
Does the user want another?
Yes
Emoprompt generation
“a sad boy walking at night. The road is empty and the place is bare”
use the new emotion
random choice by the system (suggestion)
user selects what they want
emoprompt 1
emoprompt 2
emoprompt 3

Figure 3.4: An overview of the *Interactive Detect and Respond* process.

3.5 Emotion-Guided Interaction with CoFRIDA (PIE-FRIDA)

In this section, we adapt the *Detect and Respond* and *Interactive Detect and Respond* , and systems in CoFRIDA [32, 33], which allows for co-painting between a robot and the user. We choose CoFRIDA because (1) It is already a turn-taking system that fosters *Human-Robot Interaction* to create art pieces, (2) CoFRIDA’s system currently does not support audio/ additional input to allow for diverse user space, and (3) Using CoFRIDA’s system, we can further utilize emotion in the turn-taking algorithm to determine whether at any turn the user’s emotional state changes for the final art piece (*Personalized Detect and Respond* , *Interactive Detect and Respond*).

In Figure 3.5, we abstractly show how *Personalized Interactive Emotion-Guided Collaborative Human-Robot Art Creation* fills the interactive gaps by adding emotion

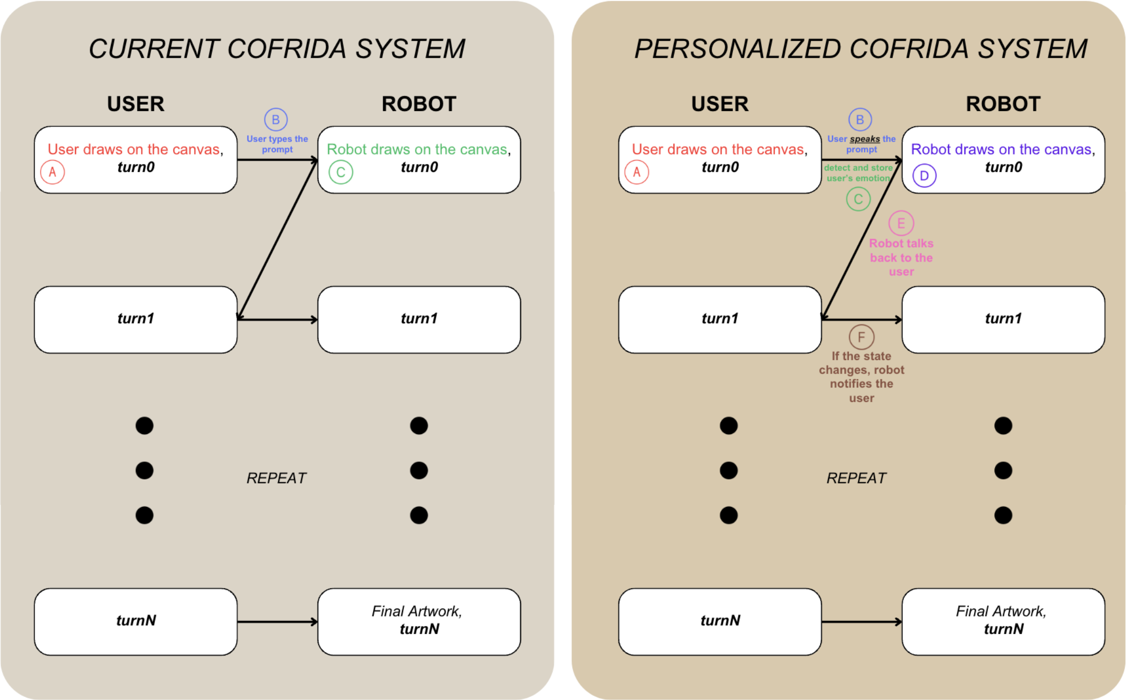

CURRENT COFRIDA SYSTEM
USER
ROBOT
User draws on the canvas,
A
turn0
B
User types the prompt
Robot draws on the canvas,
C
turn0
turn1
turn1
REPEAT
turnN
Final Artwork,
turnN
PERSONALIZED COFRIDA SYSTEM
USER
ROBOT
User draws on the canvas,
A
turn0
B
User speaks the prompt
detect and store user's emotion
C
Robot draws on the canvas,
D
turn0
E
Robot talks back to the user
turn1
turn1
F
If the state changes, robot notifies the user
REPEAT
turnN
Final Artwork,
turnN

Figure 3.5: An Abstract, *High-Level* overview of adding emotion and audio to CoFrida[33] to improve the co-Painting experience. In making the system more interactive, we introduce Ⓑ,Ⓒ, Ⓔ, and Ⓕ to represent the core changes as outlined in the *Interactive Detect and Respond* and *Personalized Detect and Respond* systems.

and audio feedback to the CoFRIDA system. As opposed to using *Stable Diffusion* as in the *Detect and Respond* systems, CoFRIDA utilizes *Instruct Pix2Pix* [3] as its primary image generation model to facilitate a seamless CoPainting. This difference is negligible in the overall project as we are not focused on how the images are generated but on how using emotion as an additional input can aid the generation of more satisfying output.

To accommodate FRIDA talking back to the user, we utilize Pylips [8] to give the robot a human-like voice. We chose a playful-sounding voice to get the most engagement from the user [1].

With these additions, we can create an informal dialogue between the user and CoFRIDA, which will help the user have much more control over the final artwork produced and improve the user's confidence in it. PIE-FRIDA can be further summarized in Figure 3.6.

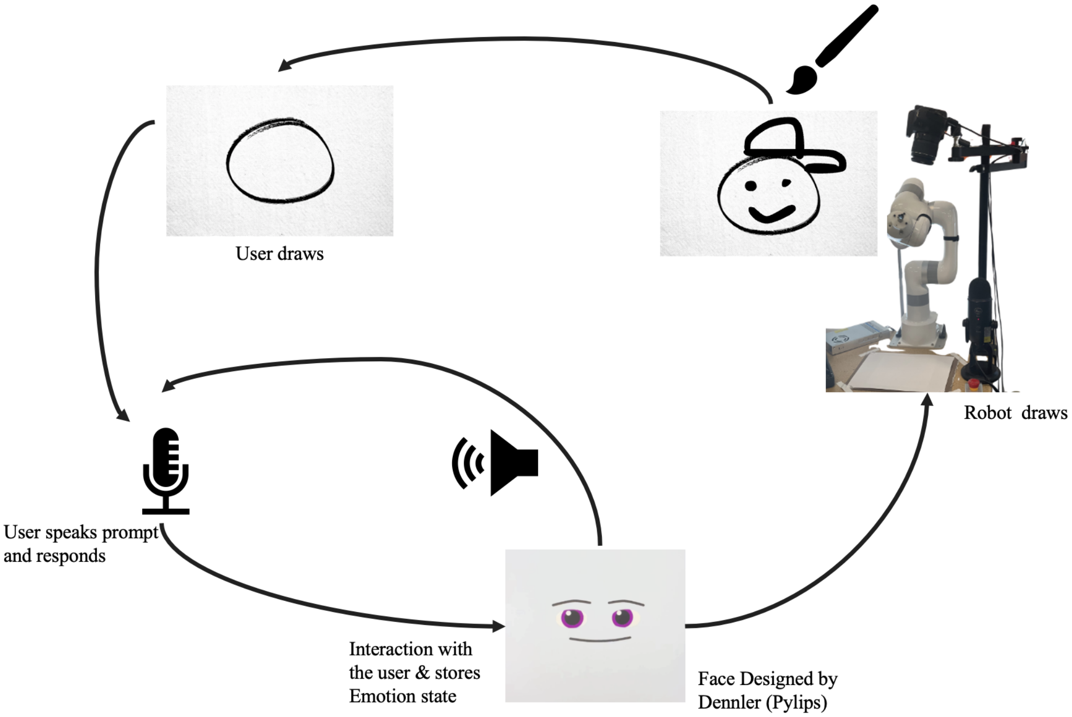

User draws
Robot draws
User speaks prompt and responds
Interaction with the user & stores Emotion state
Face Designed by Dennler (Pylips)

Figure 3.6: A High-Level Overview of PIE-FRIDA

Chapter 4

Experiments and Evaluation

In evaluating our procedure as defined in Chapter 3, we aim to assess the following:

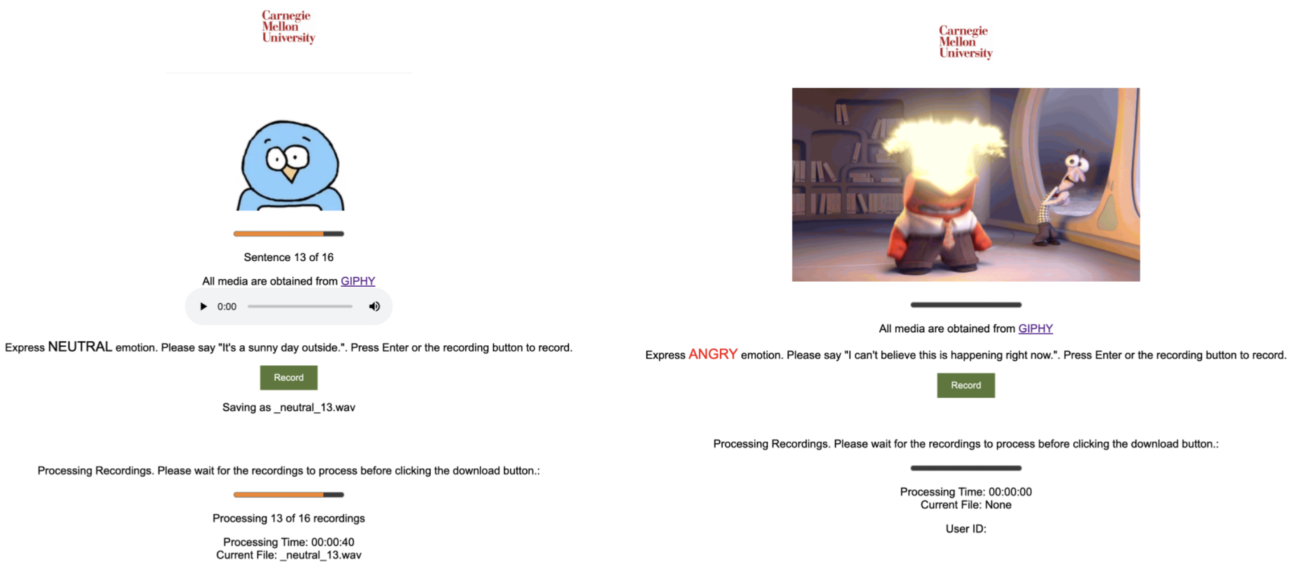
1. *Does integrating emotion in the art generation make for a better collaborative art experience?*
2. *Does it help make more satisfying art?*

To answer the above, we will primarily focus on user studies to gauge the user's interactions with the system.

4.1 Evaluation Procedures

4.1.1 Personalized Audio Dataset Collection

Building upon the section referenced as Section 3.1, our methodology for gathering user audio recordings to train/finetune is facilitated through a carefully designed user interface (UI). This interface, depicted in Figure 4.1, is the cornerstone of our data collection strategy, providing users with an intuitive and efficient means to submit their audio recordings. The UI is engineered to streamline the collection process while ensuring user comfort and compliance with privacy standards. This data collection consisted of **8** participants.



Sentence 13 of 16

All media are obtained from GIPHY

Express NEUTRAL emotion. Please say "It's a sunny day outside.". Press Enter or the recording button to record.

Saving as _neutral_13.wav

Processing Recordings. Please wait for the recordings to process before clicking the download button.:

Processing 13 of 16 recordings

Processing Time: 00:00:40
Current File: _neutral_13.wav

All media are obtained from GIPHY

Express ANGRY emotion. Please say "I can't believe this is happening right now.". Press Enter or the recording button to record.

Processing Recordings. Please wait for the recordings to process before clicking the download button.:

Processing Time: 00:00:00
Current File: None

User ID:

Figure 4.1: Two Pages of the Audio Collection User Interface

4.1.2 *Interactive Detect and Respond* with and without Personalization

The following experiment involves using *Detect and Respond* and the *Interactive Detect and Respond* to check and compare the image generation with and without personalization. We hope to (1) show a preference for user satisfaction using the personalized approach and (2) use the *Interactive Detect and Respond* to imitate a collaborative process.

We use the following UIs as shown in Figure 4.2 and Figure 4.3 to analyze this. Again, the user is unaware of which *Detect and Respond* system uses the personalization to account for a fair and unbiased study.

4.1.3 PIE-FRIDA vs CoFRIDA Interaction

Now that we have formulated an emotion guidance procedure (*Personalized Interactive Emotion-Guided Collaborative Human-Robot Art Creation*), we aim to assess the system with collaborative art using the CoFRIDA and PIE-FRIDA systems. In this study, we first test how well the current CoFrida system can depict emotion in the

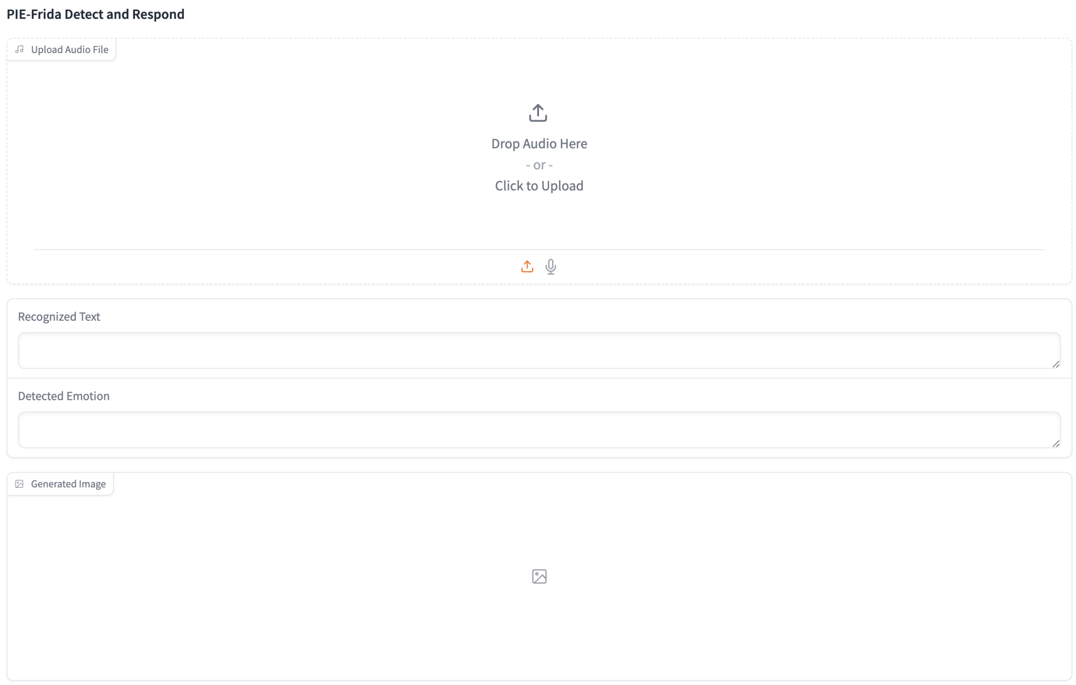
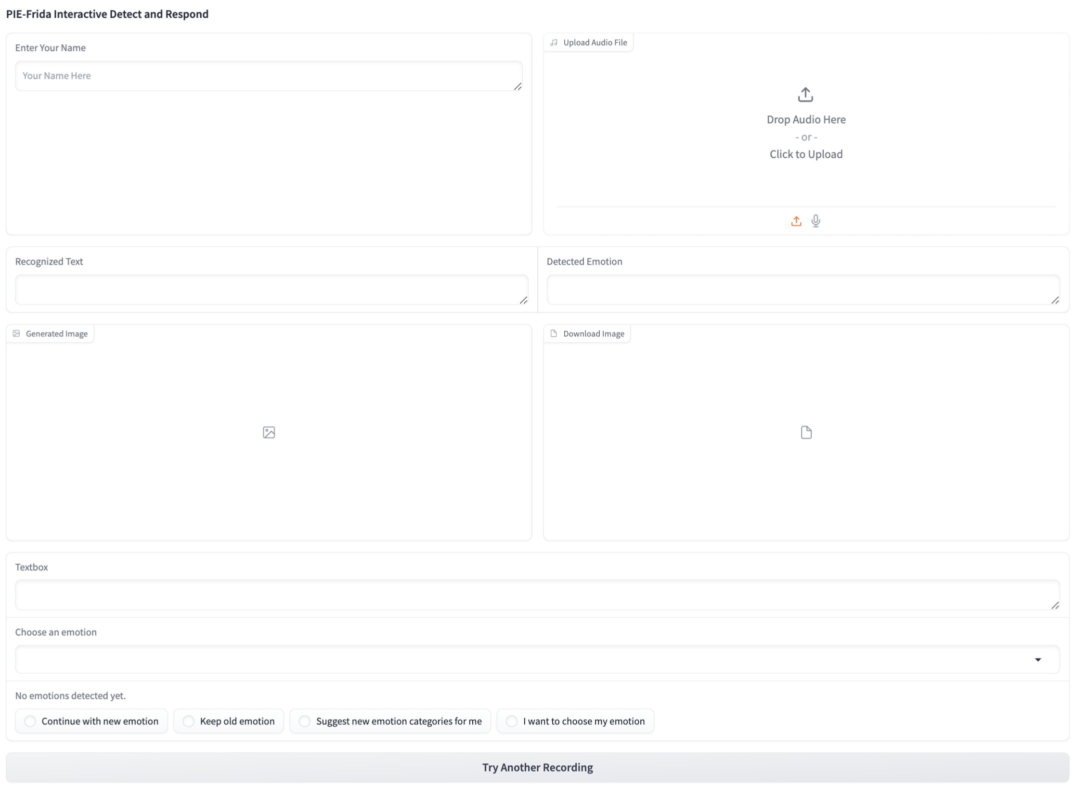


Figure 4.2: *Detect and Respond* Interface

simulation's "pen/brush-like" images.

In the simulation, we pass eight *Emoprompts* as inputs into the system. In this case, the controlled variable is the *number of strokes* needed to give the best image. We hypothesize that more strokes allow for more detail, but too many strokes could cluster the canvas and overwrite the details that represent the emotion in the image. Thus, we vary the stroke number when conducting this mini-experiment.

For the user interaction with the painting, we follow Figure 3.5, where the users take a specific number of turns to create a desired drawing using emotion as the input. With the addition of audio feedback and check-ins from the modified interactive coFRIDA system, we aim to show a boost in the overall mood of the user and satisfaction, as will be discussed in Section 4.1.4



No emotions detected yet.

Figure 4.3: *Detect and Respond* Interface

4.1.4 Assessing Overall User Confidence Level and Satisfaction

From the evaluation procedures in Section 4.1.2 and Section 4.1.3, the users fill out an exit survey once done interacting with each system. On the survey, they answer a series of Likert-Scale[17] type questions that will help gauge (1) how confident the users were with the system and (2) the degree to which the users are satisfied with the final artwork.

Chapter 5

Results

5.1 Depicting Emotion in CoFrida Simulation

We perform a mini experiment to see how well the current CoFRIDA system can depict emotions. From Figure 5.1, we see that the prompts can depict several emotional states respective to the scene. This makes it suitable to (1) Utilize the *Detect and Respond* and *Interactive Detect and Respond* systems to generate the emotion-based artwork simply in PIE-FRIDA. (2) Confirm if the emoprompts suitably depict emotion well.

5.2 Quantitative Results

To compare the generic to the personalized *Speech Emotion Recognition* models, we check the rank of each predicted value with the true value using the recorded audio given by the users as defined in Chapter 3. We do this as follows, also defined in Algorithm 1:

1. With **8** participants, we take **5** of the same audio sentence recordings from each participant and keep them as the test set.
2. For each n in rank 1,3 and 5:
   (a) Obtain the sorted probabilities of predictions, P and the true label, Y_t
   (b) If the predicted value Y is in the first n of the probabilities, then assign a

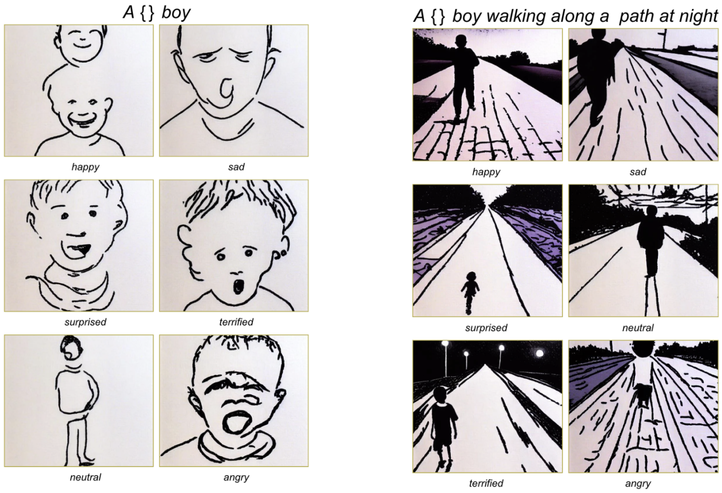


Figure 5.1: Results from evaluation emotion using the Frida Simulation

score of **1** or else **0**.

For example, if the personalized model predicts x and the generic has a prediction of z, and the sorted probabilities of true predictions are $[v, w, x, y, z, a]$:

- Rank 1 $[v]$ has the prediction of v, which does not match the personalized or generic models, so both models get a score of **0**.
- Rank 3 $[v, w, x]$ contains the personalized model's prediction, but not the generic model, so the personalized gets a score of **1** while the generic gets a score of **0**.
- Rank 5 $[v, w, x, y, z, a]$ contains the personalized and generic model's predictions, so they both get a score of **1**.

Tables Table 5.1, Table 5.3, and Table 5.5 show the total counts across the 8 participants for the ranks 1,3 and 5, and Table 5.2, Table 5.4, and Table 5.6 show their corresponding percentages. We see that the personalized performs slightly better for the small dataset, which shows promising results.

Algorithm 1 Audio Sentence Test Set Construction and Evaluation

```
Input: 8 participants, test set size of 5 audio sentences per participant
Output: Scores for ranks 1, 3, and 5
procedure PrepareTestData
    Select 5 audio sentence recordings from each of the 8 participants
    Keep these recordings as the test set
end procedure
procedure EvaluateRanks
    for each rank $n$ in $\{1, 3, 5\}$ do
        Obtain the sorted probabilities of predictions $P$
        Get the true label $Y_t$
        if predicted value $Y$ is within the top $n$ probabilities in $P$ then
            Assign score of 1
        else
            Assign score of 0
        end if
    end for
end procedure
```

Table 5.1: Overall Analysis Across All Users for Rank 1

ground truth	Baseline Score	Custom Score
disgust	3	2
fearful	0	1
happiness	1	0
neutral	1	2
sadness	1	2

Table 5.2: Overall Analysis Across All Users for Rank 1 (Percentage)

ground truth	Baseline Score (%)	Custom Score (%)
disgust	37.5	25.0
fearful	0.0	12.5
happiness	12.5	0.0
neutral	12.5	25.0
sadness	12.5	25.0

Table 5.3: Overall Analysis Across All Users for Rank 3

ground truth	Baseline Score	Custom Score
disgust	5	6
fearful	2	1
happiness	3	3
neutral	4	4
sadness	3	3

Table 5.4: Overall Analysis Across All Users for Rank 3 (Percentage)

ground truth	Baseline Score (%)	Custom Score (%)
disgust	25.0	30.0
fearful	10.0	5.0
happiness	15.0	15.0
neutral	20.0	20.0
sadness	15.0	15.0

Table 5.5: Overall Analysis Across All Users for Rank 5

ground truth	Baseline Score	Custom Score
disgust	4	6
fearful	3	5
happiness	4	4
neutral	6	6
sadness	5	8

Table 5.6: Overall Analysis Across All Users for Rank 5 (Percentage)

ground truth	Baseline Score (%)	Custom Score (%)
disgust	14.8	22.2
fearful	11.1	18.5
happiness	14.8	14.8
neutral	22.2	22.2
sadness	18.5	29.6

Chapter 6

Conclusions

We have explored the innovative intersection of emotional feedback and artificial intelligence in artistic co-creation, specifically through the interaction between users and robotic systems. We have further proposed *Personalized Interactive Emotion-Guided Collaborative Human-Robot Art Creation* : an approach that integrates personalized emotion calibration models and emotion-guided generative systems to enhance the collaborative artistic process. Furthermore, our results show (1) promising results for using a Personalized *Speech Emotion Recognition* model over a generic one and (2) the importance of constant feedback/interaction in human-robot co-creation. By actively incorporating user emotions, we show that art created by PIE-FRIDA can more accurately reflect the user's intent, thereby enriching the user experience.

Future work involves several steps, including:

1. ***Adapting Video Input:*** To allow for a more inclusive system, we would add video to use Computer Vision methods to analyze the user's emotional state and behavioral patterns that will help assess the user's confidentiality score.
2. ***Larger Study***: We hope to conduct a much larger study to help further analyze the *Personalized Speech Emotion Recognition* and the PIE-FRIDA's system. Although we have promising results from our smaller study, a more extensive analysis will help solidify the results.
3. ***Improving FRIDA's Interaction***: Currently, there are moments of silence

when FRIDA draws. We hope to improve this by having PIE-FRIDA explain what it does with each stroke. This will further engage the user throughout the interaction instead of some parts of it.

With the addition of future work, we hope to mitigate the limitations and foster a better interactive human-robot collaborative experience.

Bibliography

[1] Game-based learning and gamification to promote engagement and motivation in medical learning contexts. *Smart Learning Environments*, 2023. Accessed: 2023-05-02. 3.5

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. Opportunities and risks of foundation models. *ArXiv*, abs/2108.07258, 2021. 1.1

[3] 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. 3.5

[4] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005. 3.1

[5] Eduardo Henrique Calabres. wav2vec2-lg-xlsr-en-speech-emotion-recognition, 2024. Access date: 2024-04-07. 2.1, 2, 2a

[6] Emily Chen and Roberto Martinez. Deep learning for interactive robotic feedback systems. *Journal of Machine Learning Research*, 21:110–126, 2020. 2.3

[7] Sven-Ake Christianson, editor. *The Handbook of Emotion and Memory: Research and Theory.* Lawrence Erlbaum Associates, 1992. 1

[8] Nathaniel Dennler, Uksang Yoo, Stefanos Nikolaidis, and Maja Mataric. Pylips: A simple python package to develop animated screen-based text-to-speech interactions. Software available at URL, 2023. Accessed: 2023-05-02. 3.5

[9] Konstantin Dobler, Florian Hübscher, Jan Westphal, Alejandro Sierra-Múnera, Gerard de Melo, and Ralf Krestel. Art creation with multi-conditional stylegans. *arXiv preprint arXiv:2202.11777*, 2022. 2.2

[10] Sefik Emre Eskimez, You Zhang, and Zhiyao Duan. Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia*, 24:3480–3490, 2021. 2.1, 3.3

[11] Weiquan Fan, Xiangmin Xu, Xiaofen Xing, Weidong Chen, and Dongyan Huang. Lssed: a large-scale dataset and benchmark for speech emotion recognition. In

ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 641–645. IEEE, 2021. 3.1

[12] Kory Floyd. The role of emotion in computer-mediated communication: A review. *Communication Research*, 24(3):327–348, 1997. 1

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2.1

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1, 3.3

[15] Ashish B Ingale and DS Chaudhari. Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(1):235–238, 2012. 2.1, 3.1

[16] Alice Jones and David Kim. Adaptive feedback algorithms for human-robot interaction. *IEEE Transactions on Robotics*, 26(5):845–851, 2010. 2.3

[17] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4): 396–403, 2015. 4.1.4

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2.2

[19] Uros Krcadinac, Jelena Jovanovic, Vladan Devedzic, and Philippe Pasquier. Textual affect communication and evocation using abstract generative visuals. *IEEE Transactions on Human-Machine Systems*, 46(3):370–379, 2015. 2.2

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015. 2.1

[21] Michael Lee and Sung-Hyun Park. Affective feedback in human-robot interaction. In *Proceedings of the International Conference on Human-Robot Interaction*, pages 335–342. ACM, 2012. 2.3

[22] Qing Lin, Jingfeng Zhang, Yew Soon Ong, and Mengmi Zhang. Make me happier: Evoking emotions through image diffusion models. *arXiv preprint arXiv:2403.08255*, 2024. 3.3

[23] Zhixuan Liu, Youeun Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. Towards equitable representation in text-to-image synthesis models with the cross-cultural understanding benchmark (ccub) dataset. *arXiv preprint arXiv:2301.12073*, 2023. 1.1

[24] Zhixuan Liu, Youeun Shin, Beverley-Claire Okogwu, Youngsik Yun, Peter

Schaldenbrand, Jihie Kim, and Jean Oh. Culturally-aware stable diffusion: Supporting representation with culturally-aware text-to-image synthesis. 2023. 1.1

[25] Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. Scoft: Self-contrastive fine-tuning for equitable image generation. *arXiv preprint arXiv:2401.08053*, 2024. 1.1

[26] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 1a

[27] Vihaan Misra, Peter Schaldenbrand, and Jean Oh. Robot synesthesia: A sound and emotion guided ai painter. *arXiv preprint arXiv:2302.04850*, 2023. 2.2

[28] OpenAI. Introducing chatgpt-4, 2023. Accessed: 2023-05-02. 3

[29] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*, 2021. 2

[30] 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. 1.1, 4

[31] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986. 2.1

[32] Peter Schaldenbrand, James McCann, and Jean Oh. Frida: A collaborative robot painter with a differentiable, real2sim2real planning environment. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11712–11718. IEEE, 2023. 2.4, 3.5

[33] Peter Schaldenbrand, Gaurav Parmar, Jun-Yan Zhu, James McCann, and Jean Oh. Cofrida: Self-supervised fine-tuning for human-robot co-painting. *arXiv preprint arXiv:2402.13442*, 2024. (document), 1.1, 2.4, 3, 3.5, 3.5

[34] John Smith and Jane Doe. Interactive feedback in human-robot cooperation. *Journal of Robotics and Autonomous Systems*, 55(2):123–130, 2005. 2.3

[35] Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21:93–120, 2018. 2.1, 3.1

[36] Christian von Scheve. *Emotion and Social Structures: The Affective Foundations of Social Order*. Routledge, 2013. 1

[37] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective

computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52, 2022. 1

[38] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. A comprehensive review of speech emotion recognition systems. *IEEE access*, 9:47795–47814, 2021. 3.1

[39] Chao Xu, Junwei Zhu, Jiangning Zhang, Yue Han, Wenqing Chu, Ying Tai, Chengjie Wang, Zhifeng Xie, and Yong Liu. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6609–6619, 2023. 2.2, 3.3

[40] Li Yu, Dolzodmaa Davaasuren, Shivansh Rao, and Vikas Kumar. Stargan-egva: Emotion guided continuous affect synthesis. In *Proceedings of the 1st International Workshop on Human-centric Multimedia Analysis*, pages 53–61, 2020. 2.2, 3.3

[41] Siqi Zhu, Chunmei Qing, and Xiangmin Xu. Text-guided generative adversarial network for image emotion transfer. In *International Conference on Intelligent Computing*, pages 506–522. Springer, 2023. 2.2, 3.3