Robot Synesthesia: A Sound and Semantics Guided AI Painter

Vihaan Misra,[1,2] **Peter Schaldenbrand,**[2] **Jean Oh** [2]

[1] Netaji Subhas University of Technology, New Delhi, India
[2] The Robotics Institute, Carnegie Mellon University, Pittsburgh, USA
vihaan.ee19@nsut.ac.in, pschalde@andrew.cmu.edu, hyaejino@andrew.cmu.edu

Abstract

If a picture is worth a thousand words, sound may cost a million. Additionally, it is incredibly challenging to accurately use words to describe the nuances and complexities of sound. Recent robotic painting and other image synthesis methods have achieved progress in generating visuals from language inputs, but the translation of sound into images is vastly unexplored. Audio data has the potential to expand the accessibility and controllability for the user and provide a means to convey complex emotions and the dynamic aspects of the real world. Here, we propose to extend the recent robotic painting framework, FRIDA, by incorporating an additional generalized sound-semantics step that encodes sound into the image-text embedding space and aids in the manipulation of the painting planning process for controlling the robot painting. We illustrate how our approach may be used in conjunction with existing modalities to create paintings that adhere to the guiding semantics while also enhancing the user's control capabilities. While sound-guidance has been used in image manipulation, few existing work uses sound inputs to create general image content. In this paper, we share our preliminary results in a qualitative form.

Introduction

One of the first approaches to make mechanical drawings can be traced back several decades to (McCorduck 1991). Since then, roboticists have studied a variety of interesting approaches for robot-based drawings. Recent developments in generative models like Generative Adversarial Networks(GAN), Vision Transformers(ViT) and Diffusion models have further spurred this research direction with their ever-improving quality of image synthesis and understanding of semantics.

Existing style transfer and image-to-image translation techniques treat the process as mapping pixels (McCorduck 1991; Zhu et al. 2017) or a continuous pixel space optimization (Gatys, Ecker, and Bethge 2016). However, painting is an expression of ideas and emotions of an artist through a visual language. It is a dynamic approach that begins with a vague notion of an artist and evolves dynamically during the course of painting, producing a picture that satisfies the artist's semantic and high-level objectives (Hertzmann 2022). This procedure is organically in contrast to how deep learning methods have been used to create artwork. To create an approach that incorporates the dynamics of the process of painting an image, (Schaldenbrand, McCann, and Oh 2022) proposed a Framework and Robotics Initiative for Developing Arts (FRIDA). FRIDA aims to give robots the ability to paint by taking in inputs in the form of images, texts, and sketches to produce an artistic visualisation that is consistent with the painter's intent. They accomplish this by emulating the painting as a planning problem, with the canvas as the state space and the brush strokes as the available actions. This gives the robot the ability to combine content creation with action planning and adhere to a continuous content optimization strategy.

Figure 1: Sound-guided image manipulation. The figure shows the source images (top row), their paintings drawn by FRIDA (middle row) and the paintings drawn by FRIDA with the proposed sound semantics (bottom row)

However, using sketch and text inputs have numerous limitations when it comes to applying sound semantics into an image. It can be challenging to fully express the complexity and nuances of sounds using something discrete such as

text. For instance, laughter is a complex sound denoted by unique loudness and rhythmic properties. While people can hear the differences in laughter easily, it can be challenging to describe them.

To incorporate sound as an input modality, we propose a form of *robot synesthesia*–a perceptual phenomenon in which a person may perceive visuals when listening to sounds. We extend the FRIDA framework by introducing an additional intuitive and powerful image generation planning method driven by sound semantics. As highlighted in (Lee et al. 2021), we leverage the approach of encoding images into a latent representation using a $CLIP$ (Radford et al. 2021) and $CLIP_{Audio}$ for audio (Lee et al. 2021). The difference of the latent spaces of the planned painting and the input audio forms a loss function, which is optimized using gradient descent to create painting plans for paintings that fit audio inputs.

Since the generation process relies solely on guidance, any input image—whether original or taken from a well-researched dataset—can employ the sound guidance to produce new and creative paintings as opposed to existing sound-guided image manipulation work which is constrained in its content creation abilities (Lee et al. 2021). Our experimental results demonstrate how the use of sound semantics along with the existing modalities facilitates a larger range of user-control for producing natural paintings. In contrast to text and sketch-based painting synthesis approaches, it enables the user to give a wide range of intricate information.

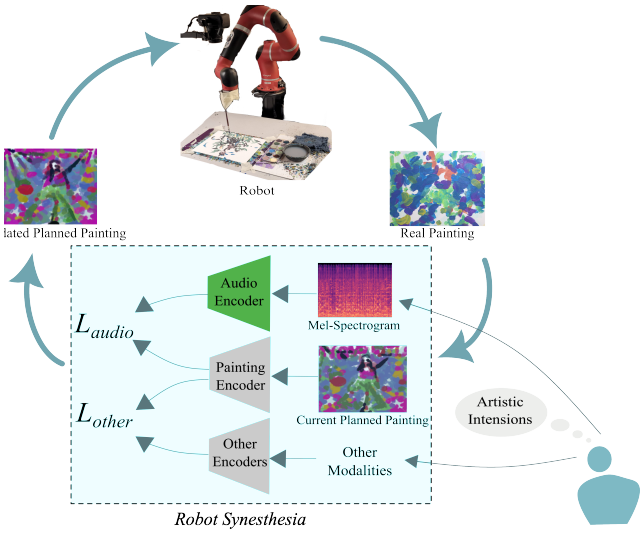

Figure 2: Following FRIDA (Schaldenbrand, McCann, and Oh 2022) human users specify their artistic intentions via text, styles, and sketches. We add audio as an input in this paper. These intentions influence the planning in simulation which is carried out via a Rethink Sawyer robot. The painting process is monitored via camera perception and the plan can be modified during the painting process.

Related Work

Robot Painting

Most prior work on robot painting and stroke-based rendering formalize their tasks as rendering a given image using given primitives, materials, and tools (Huang, Heng, and Zhou 2019; Schaldenbrand and Oh 2021; Singh et al. 2021; Hertzmann 1998). Painting, as a form of art, involves creatively expressing a message, and this aspect extends it as a high-level task than simply printing an image with challenging constraints. Proper robot painting involves achieving an artists high-level goals. To this end FRIDA (Schaldenbrand, McCann, and Oh 2022) introduced multiple inputs to the system in addition to a reference image to allow further expression of the human user of the system.

Sound-Image Encoding

Encoding sounds and images into the same latent space in a differentiable manner allows for comparison of the two modalities that can be exploited such that one of the modalities can be altered to match the other. (Lee et al. 2021) train a sound encoder, $CLIP_{Audio}$, to encode sound and images into the same latent space. $CLIP_{audio}$ was trained on the VGG-Sound Dataset (Chen et al. 2020) which contains over 200k clips for 309 different classes. The dataset contains 200,000 10-second audio clips from each class captured from YouTube videos, with no more than two clips per video. The dataset's sound categories can be broadly divided into: people, animals, music, sports, nature, vehicles, homes, and tools, among others.

Sound-Guided Image Manipulation

There is a growing body of research on cross-modal generative models that use audio samples for guidance based tasks. The previous works in this field have mostly focused on music rather than sound semantics and follow a cross-modal learning strategy for style transfer from music to image(Lee et al. 2020). There also have been works that map music embeddings to visual embedding-space using Style-GAN models (Jeong, Doh, and Kwon 2021).

In recent studies, there has been more interest in the semantics of the sound for navigational direction in the latent space of Style-GAN models. (Lee et al. 2021) demonstrate that the latent space of a pretrained image synthesis model, StyleGAN2, can be manipulated such that the original generated image fits a given audio sample better. They use layerwise masking to keep compact content information within style latent code. This edited latent code is then passed into the StyleGAN2 generator to obtain the modified image. This follows the StyleCLIP (Patashnik et al. 2021) methodology for sound-guided image manipulation. This is done through sound guidance in which the generated image and input sound are encoded into the same latent space. The comparison of these latent spaces forms a loss function that can be back propagated. Stochastic gradient descent alters the latent input to produce images that are more aligned with the input sound.

While there has been success in using this methodology for image manipulation, previous work fails to create gen-

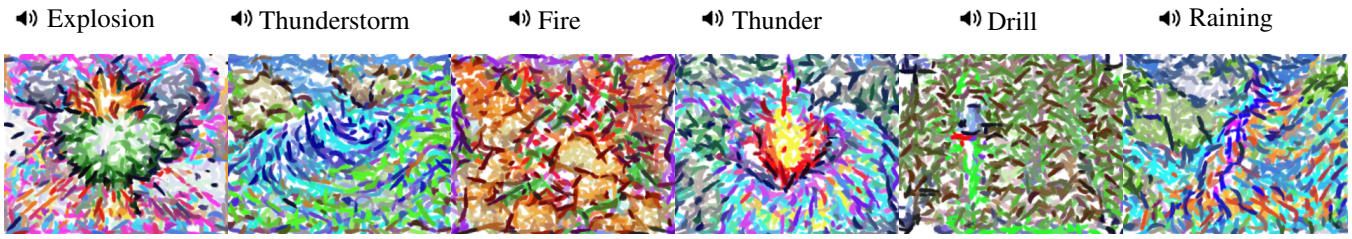

Figure 3: Paintings generated using sounds described in labels above each painting.

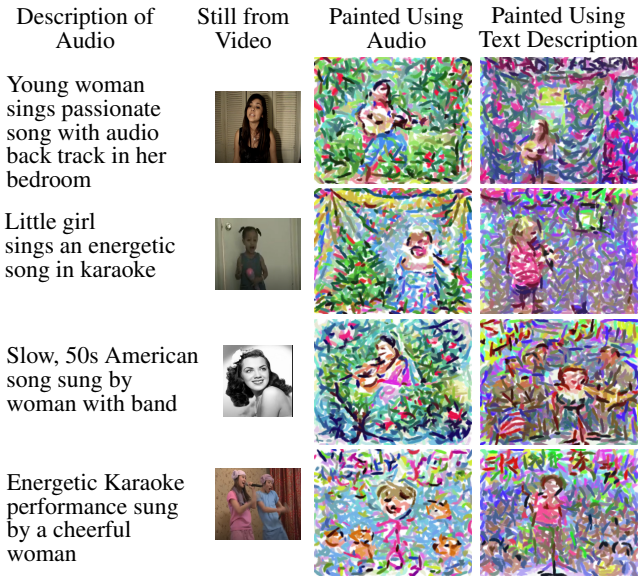

Figure 4: Painting various examples from the VGG-Sound (Chen et al. 2020) categorized as "female singing" from just audio versus using a description of the audio. A description of the sound used and a still from the video that it was captured from are displayed left of each painting.

eralizable output images fitting the sound descriptions because they utilize pre-trained image synthesis models that can only generate images within their training distribution such as churches, faces, or artwork. Therefore, these prior works use sound only to manipulate a particular image that the synthesis model is capable of producing. In this work, our approach creates in a more general content domain and can generate images purely from audio inputs.

Approach

Our approach adds audio input to the FRIDA robotic painting system (Schaldenbrand, McCann, and Oh 2022). The brush strokes are the actions which are parameterized by shape (length, bend, and thickness, location, orientation, and color. Given a set of strokes and canvas, FRIDA's simulation environment can differentiably render these into a simulated painting represented as an RGB image. We encode the simulated painting, p, and input audio, a, into the same latent space using $CLIP$ (Radford et al. 2021) and $CLIP_{audio}$ (Lee et al. 2021). The encodings are compared using cosine distance to form a loss function. In practice, the painting is augmented using various perspective warps and cropping, as is customary in CLIP-guided image synthesis, for robust loss backpropagation.

$$l_{\text{audio}}\,(p, a) = \cos(CLIP(p), CLIP_{Audio}(a)) \tag{1}$$

$$\ddot{p} = \min_{p} \left[l_{audio}(p, a) + \sum_{i=1}^{4} (w_i l_i) \right] \tag{2}$$

The FRIDA paper introduces 4 other loss functions, l_i, that connect the paintings to modalities text, images, sketches, and styles. An objective function can be formulated to find the painting plan, Equation 2: Given the weights for each loss function, w_i, find the painting, p, that minimizes the weighted sum of losses.

Our full painting pipeline is displayed in Figure 2. A human user expresses their intentions via a combination of modalities. To create a painting we follow the FRIDA framework and randomly initialize a set of brush strokes then render a simulated painting on to the current canvas. After that, we initialize n strokes and sample evenly over the brush stroke parameters to begin the painting process. The plan is then optimized to use the objectives passed by the user. These objectives can include, image-text similarity, style, simple replication, semantic replication and the proposed image-audio similarity objective. These objectives enable the system to achieve the user-specified goals for the painting being drawn by the robot, and does away with the need of having trained generators for synthesizing the paintings, making it more generalized.

Results

Painting with Sound

We introduce sound-semantics to the FRIDA framework and demonstrate how it can be used to produce a variety of contemporary paintings, offering the user greater freedom and control. Using just audio as guidance in Figure 3, our approach is capable of generating paintings that capture various natural sounds.

We conducted a survey to quantify the results of Figure 3 and measure the correlation between the input audio and the generated painting. In the survey, participants listened to one of the six audio samples used to generate the images in Figure 3, then were shown the six paintings and asked to "select the image below that looks like it was created from the

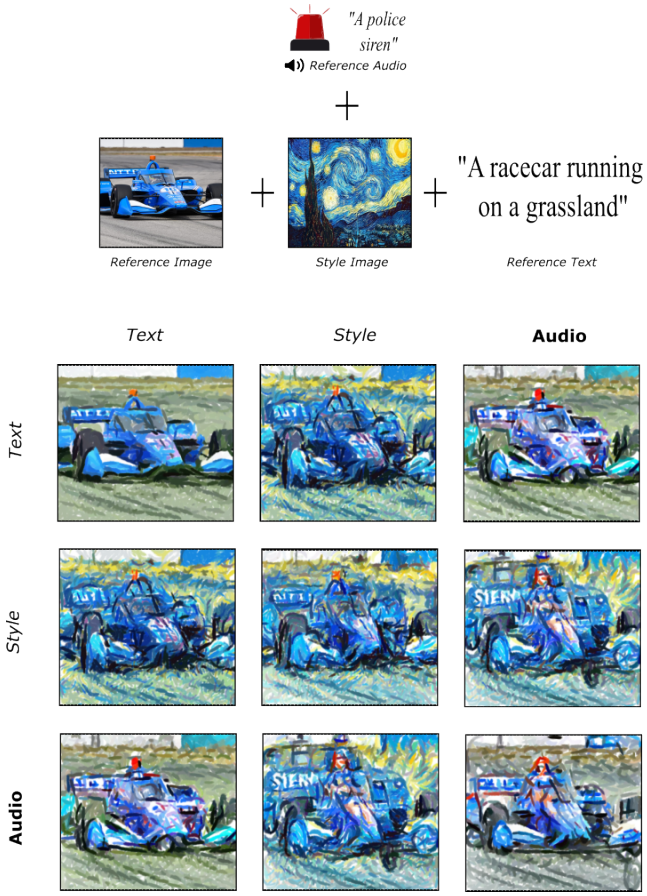

Figure 5: An image matrix demonstrating the results with different modality combinations of text, style and audio. The reference image, text, audio and the style image signify the style image, input text, audio and the base image used for creating the paintings

above sound". A random selection would result in 16.7% accuracy as there were six options, however, participants selected the correct painting 43.3% of the time. We conducted the survey through Amazon Mechanical Turk recruiting 28 participants. Each audio was evaluated by five different participants. The confusion matrix can be seen in Table 1. Many of the sounds were similar and this can be reflected with the frequent confusion of thunder, thunderstorm, and raining sounds.

Sound is nuanced and challenging to describe accurately with language. We paint multiple examples from the held-out set of the VGG Sound (Chen et al. 2020) dataset that were all categorized as "female singing", Figure 4. We attempt to describe the sound used and capture stills from the YouTube videos that these sounds were scraped from. Despite all the sounds having the same category and some similarities to sound, instrumentation, and atmosphere, the generated paintings are vastly dissimilar to each other. The paintings generated from the sound of the video were also vastly different from those generated from just the text description of the sound, thereby demonstrating the importance of sound as an input modality since it cannot be represented well with other modalities.

	D	E	F	R	T	TS
Drill	3	1	0	0	0	1
Explosion	0	2	1	0	2	0
Fire	0	0	2	1	2	0
Raining	0	1	0	1	1	2
Thunder	2	0	0	2	0	1
Thunderstorm	0	0	0	0	0	5

Table 1: The confusion matrix from our survey where participants listened to audio then decided which of the six paintings in Figure 3 was generated using that audio. Rows display true labels and columns show the predicted labels.

Painting with Sound and Reference Images

In Figure 5, we demonstrate how the different modalities in FRIDA affect the painting synthesis in various combinations with each other when painting from a reference image. It illustrates the creative scope of using audio-based guidance and the way it can be used to make unique paintings without prior training.

Figure 1 further emphasizes that this technique enables a wide variety of unique painting manipulations that improves the existing framework and makes it more robust. We contrast the image editing capabilities of sound and text in Figure 6 in order to emphasise the distinctions between the proposed sound mode and the current text modalities. The figure shows that sound-based guidance tends to lead the painting generation to follow a more distinct, and arguably more meaningful path, than the text-based guidance. These results indicate that the audio-based semantic assistance allows the user to make creative adjustments and synthesize unique artworks.

Painting with Audio and Other Modalities

We paint with audio inputs paired with FRIDA's existing modalities in Figure 8. Loss function weights are adjusted to allow the appearance of both modalities to become prominent. The results are abstract, but represent both modalities strongly.

Painting from Music

The VGG-Sound dataset used to train our sound encoder contains some music in various forms. Music is a modality that is extremely challenging to represent in any other form such as language or images accurately. We experiment with the generalization of the image encoder in Figure 7. The results are abstract but make some resemblance to content and general atmosphere of the given songs.

Discussion

This work forms as ground work towards creative human-robot interaction research. Increasing the input space of the robotic painting system should increase the control of the user. This may allow the human interacting with FRIDA to

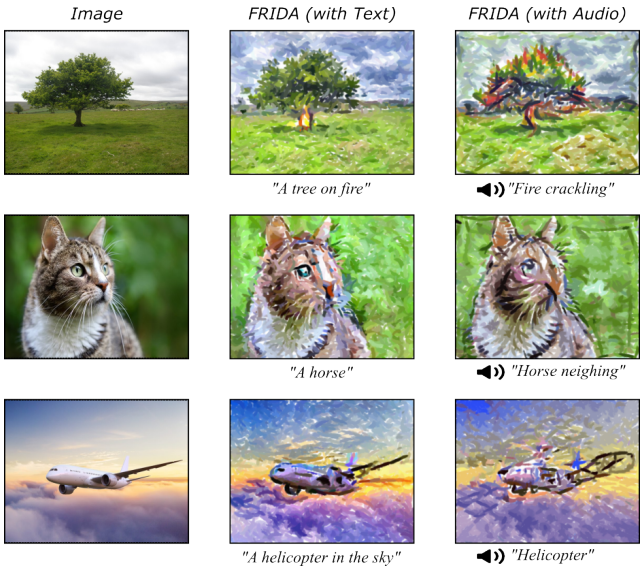

Figure 6: Comparison of modified paintings with text-guided and sound-guided semantic manipulation. The figure shows the source images (first column), their paintings drawn by FRIDA with text-guidance (middle column) and the paintings drawn by FRIDA with audio-guidance (last column)

Figure 7: Painting using various Pop songs as input. Genres of the songs from left to right are Disco, Traditional American Folk, and Rap/Hip-Hop

Figure 8: Combining audio as an input modality with other modalities that FRIDA can handle.

feel more ownership over the artwork created such that the human and robot are collaborators rather than creative automation. Sound can act as an accessible input modality for people with visual impairments who like to paint. In future work, we hope to engage with different communities of people who face physical barriers to performing painting to get FRIDA to enable more people to express themselves through the visual art of painting.

Conclusions

We present a novel connection between an existing robotic painting system (Schaldenbrand, McCann, and Oh 2022), and an image-audio encoder (Lee et al. 2021). Our approach enables image synthesis from purely sound inputs whereas previous work is limited to sound-guided image manipulation. Because our approach does not rely on a pretrained image synthesis model, it is not constrained by any training dataset and can produce images only constrained by the image-audio encoders. This allows for painting from musical inputs as well as various natural sounds.

References

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725. IEEE.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image Style Transfer Using Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2414–2423.

Hertzmann, A. 1998. Painterly rendering with curved brush strokes of multiple sizes. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 453–460.

Hertzmann, A. 2022. Toward Modeling Creative Processes for Algorithmic Painting. *arXiv preprint arXiv:2205.01605*.

Huang, Z.; Heng, W.; and Zhou, S. 2019. Learning to paint with model-based deep reinforcement learning. In *Proceed-*

ings of the IEEE/CVF International Conference on Computer Vision, 8709–8718.

Jeong, D.; Doh, S.; and Kwon, T. 2021. TräumerAI: Dreaming Music with StyleGAN.

Lee, C.-C.; Lin, W.-Y.; Shih, Y.-T.; Kuo, P.-Y. P.; and Su, L. 2020. Crossing You in Style: Cross-Modal Style Transfer from Music to Visual Arts. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 3219–3227. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Lee, S. H.; Roh, W.; Byeon, W.; Yoon, S. H.; Kim, C. Y.; Kim, J.; and Kim, S. 2021. Sound-Guided Semantic Image Manipulation.

McCorduck, P. 1991. *Aaron's code: meta-art, artificial intelligence, and the work of Harold Cohen*. Macmillan.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision.

Schaldenbrand, P.; McCann, J.; and Oh, J. 2022. FRIDA: A Collaborative Robot Painter with a Differentiable, Real2Sim2Real Planning Environment.

Schaldenbrand, P.; and Oh, J. 2021. Content masked loss: Human-like brush stroke planning in a reinforcement learning painting agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 505–512.

Singh, J.; Smith, C.; Echevarria, J.; and Zheng, L. 2021. Intelli-Paint: Towards Developing Human-like Painting Agents. *arXiv preprint arXiv:2112.08930*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.