2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

CoFRIDA: Self-Supervised Fine-Tuning for Human-Robot Co-Painting

Peter Schaldenbrand[1], Gaurav Parmar[1], Jun-Yan Zhu[1], James McCann[1], and Jean Oh[1]

Abstract— **Prior robot painting and drawing work, such as FRIDA, has focused on decreasing the sim-to-real gap and expanding input modalities for users, but the interaction with these systems generally exists only in the input stages. To support interactive, human-robot collaborative painting, we introduce the Collaborative FRIDA (CoFRIDA) robot painting framework, which can *co-paint* by modifying and engaging with content already painted by a human collaborator. To improve text-image alignment–FRIDA's major weakness–our system uses pre-trained text-to-image models; however, pre-trained models in the context of real-world co-painting do not perform well because they (1) do not understand the constraints and abilities of the robot and (2) cannot perform co-painting without making unrealistic edits to the canvas and overwriting content. We propose a self-supervised fine-tuning procedure that can tackle both issues, allowing the use of pre-trained state-of-the-art text-image alignment models with robots to enable co-painting in the physical world. Our open-source approach, CoFRIDA, creates paintings and drawings that match the input text prompt more clearly than FRIDA, both from a blank canvas and one with human created work. More generally, our fine-tuning procedure successfully encodes the robot's constraints and abilities into a foundation model, showcasing promising results as an effective method for reducing sim-to-real gaps.** `https://pschaldenbrand.github.io/cofrida/`

I. INTRODUCTION

While recent breakthroughs in text-to-image synthesis technologies have ignited a boom in digital content generation, using them to produce art with robots is still in its infancy due to a significant gap between simulated and real-world environments. FRIDA [1] is a robotic framework that can take user inputs, such as language descriptions or input images, to paint on a physical canvas using a paintbrush and acrylic paint. While FRIDA aims at giving users control over content generation, the users are allowed to add their input only with an initial input prompt or image, after which they are excluded from the creative process. While it is still debatable whether such an autonomous creation is desired by humans practicing art [2], there is strong evidence of the potential value of a co-creative agent [3], [4], [5], [6], [7], [8], [9] specifically in the domain of art therapy [10], [11], [12], [13]. The benefits can be further increased when paired with a physical embodiment of such an agent and drawing in the real world [14], [15]. To invite users into the creative process and bring the benefits of both co-creation and robotic embodiment, we build on FRIDA to propose a Collaborative Framework and Robotics Initiative for Developing Arts (CoFRIDA), as illustrated in Fig. 1.

One of the biggest challenges in human-robot co-creation is enabling the robot to create new content that engages with the existing content that the human drew, hereby referred to as *co-painting*. While there exist related image editing problems, such as in-painting, co-painting is a new class of problems with unique challenges as it is undesirable in co-painting to make radical changes to the image that would overwrite the human's previous work. In in-painting, the area for editing an image is coarsely specified by the user and the model is expected to drastically change the content within that local region. By contrast, with co-painting, the edit is expected to preserve and engage with the full canvas rather than re-imagining a local region. Whereas in-painting is a localized edit by definition, co-painting is a continuous, iterative completion, e.g., adding detail to an existing human-drawn rough sketch.

Besides the challenges of co-painting, robotic image cre-

Fig. 1. **Co-Painting with CoFRIDA.** We showcase how CoFRIDA collaboratively paints with artists. The process begins with the artist sketching a table. Building on that foundation, CoFRIDA adds to the canvas, guided by the artist's initial prompt: "A bulky robot arm on a table." The artist then iterates on the painting with additional strokes to add detail to the robot arm, and provides a new text prompt, "A robot arm with a hand." CoFRIDA responds by completing the painting to match this new description.

[1]The Robotics Institute, Carnegie Mellon University
{pschalde, jmccann, gparmar, junyanz, hyaejino}@andrew.cmu.edu

Fig. 2. **Co-Painting.** We introduce Co-Painting as a task in which a robot must add content to a painting that engages with the current content without destroying the existing work. We demonstrate that existing models (Instruct-Pix2Pix, bottom row) often cannot successfully add content without making unreasonably large edits to the canvas, overwriting any prior work, while CoFRIDA (top row) adds content that harmonizes with the existing work.

ation is difficult due to real-world constraints, such as existing canvas state, limited abilities of the robot, tools and materials available to the robot, and stochasticity in robot performance. These robotic constraints vastly limit the content that is capable of being created, as illustrated in the left side of Fig. 8. With a large paintbrush, fine-details are not achievable, and with a single marker, multi-color images are not possible. Multiple works address these constraints to decrease the Sim2Real gap, but only paint from image inputs [16], [17], [18], [19]. Even fewer existing works use cameras to enable co-creation of images [14].

FRIDA uses the data from a real robot to be able to simulate high-fidelity brush strokes using the idea known as Real2Sim2Real. To paint from a language input, FRIDA uses CLIP [20] to align language and image which tends to generate noisy output. To improve the quality of paintings for CoFRIDA, we use powerful image generators pre-trained using gigantic text-image paired data, e.g., StableDiffusion [21] or Instruct-Pix2Pix [22]. Because such pre-trained image generators do not know the capabilities of the robot, there is both a large difference in pixel value and semantic meaning between the image generator output and FRIDA's simulated plan. The former difference is a traditional Sim2Real gap, whereas the latter is a concept we introduce as the *Semantic Sim2Real Gap*.

To reduce the Semantic Sim2Real Gap, we propose a self-supervised fine-tuning for CoFRIDA. CoFRIDA adapts a pre-trained image generator to both generate content within the abilities of the robot and perform co-painting to enable human-robot collaborative drawing from language guidance, e.g., in this paper, we use Instruct-Pix2Pix [22] as our base text-image model. To adapt a pre-trained model for co-painting and encode robotic constraints, first we create the self-supervised fine-tuning dataset by using FRIDA to generate full drawings or paintings of images from a text-image dataset. Strokes from the full paintings are removed selectively to form partial paintings. We fine-tune Instruct-Pix2Pix by retraining it with a low learning rate to predict the full painting from the partial painting and text prompt.

CoFRIDA can successfully use an existing canvas state to generate future actions towards a language goal without completely overwriting the existing work as shown in Fig. 2. Based on a survey on Amazon Mechanical Turk (MTurk) of 24 participants, CoFRIDA's completed drawings from partial sketches were found to be substantially more similar to the language goal when compared to those by the baselines.

Our main contributions are summarized as: 1) we introduce co-painting, a new class of image editing that is required for human-robot collaborative creation; 2) we propose Collaborative FRIDA (CoFRIDA) to support human-robot co-painting to produce real-world arts, e.g., paintings and drawings on canvas; 3) we propose a generalizable method for reducing the Sim2Real gap using self-supervised fine-tuning, enabling generic pre-trained models to be used with physical robots; 4) CoFRIDA is open-source[1] and available on XArm, Franka Emika, and Rethink Sawyer robot platforms.

II. RELATED WORK

A. Computer-Based Image Co-Creation

Computer-based image co-creation generally involves turn taking between a human and a computer in applying brush stroke primitives towards one of a discrete set of goals, as in `sketch-rnn` [23] and Drawing Apprentice [4], or even towards natural language goals [6], [5]. Computer-based studies have shown creativity augmentation benefits of co-creation [6], [5], [4] since computer agents can add serendipity and reformulate user's original intentions leading to unexpected by enjoyable outputs [7]. However, Computer-based painting models do not transfer well out-of-the-box into the real world due to the Sim2Real gap [16], [1], [24].

B. Robotic Image Co-Creation

There exists many real-world methods for robot painting and drawing [25], [19], [17], [26], however, few systems have incorporated perception into their systems to enable co-painting. Cobbie [14] is a co-drawing system that boosted ideation for novice drawers, however, it is limited to drawing on blank areas of the paper rather than engaging with the user drawn content. [13] created a robot arm that can draw from speech inputs that are limited to simple objects found in the Quick, Draw! dataset [27]. The FRIDA [1] system is the only system directly capable of making physical paintings that engage with existing content conditioned on natural language goals. While FRIDA plans based on current canvas state, it uses CLIP and gradient descent for planning which produces paintings that are very noisy and only loosely resemble the input text.

III. METHOD

Our approach, CoFRIDA shown in Fig. 3, is made up of three primary components: (1) The Co-Painting Module, which produces images illustrating how the robot should add content to an existing canvas given a text description, (2) FRIDA [1], a robotic painting system for planning actions from given images, and (3) a self-supervised method for creating training data using FRIDA to fine-tune pre-trained models in the Co-Painting Module.

[1] `https://github.com/cmubig/Frida`

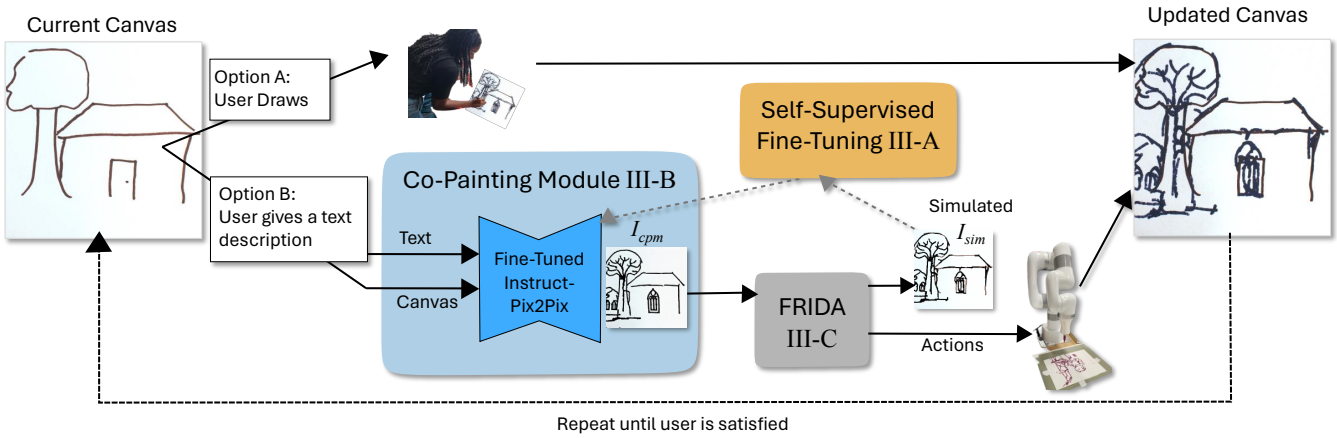

Fig. 3. **Method Overview.** Offline, we fine-tune a pre-trained Instruct-Pix2Pix model on our self-supervised data. Online, the user can either draw or give the robot a text description. The Co-Painting Module takes as input the current canvas and text description to generate a pixel prediction of how the robot should finish the painting using the fine-tuned Instruct-Pix2Pix model. FRIDA predicts actions for the robot to create this pixel image and produces a simulation. This process is repeated until the user is satisfied.

A. *Self-Supervised Data Creation*

While there exist some supervised data of human-created co-paintings [28], [29], they are only on the order of tens of examples and were not made using the same materials available to our robot. To support co-painting tasks, we propose a self-supervised method for generating training data to train a Co-Painting Module. We simulate paintings of images from the art subset of the LAION image-text dataset [30] using FRIDA with image-guidance loss (difference of CLIP embeddings of images). To create partial paintings, strokes are removed selectively to support a variety of co-painting tasks: remove all strokes, a random subset of strokes, strokes corresponding to a salient region (defined with CLIP as in [31]) of the image, and strokes from a semantic region (using Segment Anything [32]). Illustrative examples are shown in Fig. 4.

Some source images cannot be accurately represented with the robot's abilities. We filter out such images by removing instances that have a CLIPScore between the simulated full paintings and the text less than 0.5.

We use this self-curated data to fine-tune a base text-to-image generation model to be able to 1) continue to create content on an existing canvas and 2) generate images that the target robot is capable of painting.

B. *Co-Painting Module*

The goal of the Co-Painting Module (Fig.3) is to generate an image of how the robot should complete the painting given a photograph of the current canvas and a user given text description. The Co-Painting Module uses Instruct-Pix2Pix [22] as a pre-trained model as it enables conditioning the output on an input canvas. The pre-trained Instruct-Pix2Pix, however, has two shortcomings to be used for co-painting: (1) the generated images do not reflect actual robotic constraints, and (2) the existing canvas can sometimes be overwritten completely as shown in Fig. 2. To overcome these limitations, we fine-tune Instruct-Pix2Pix using the dataset of partial and full drawings with their captions described in Sec. III-A.

Fine-tuning is performed using FRIDA's simulated canvases because (1) it would be infeasible to generate a large-scale dataset with the physical robot, and (2) the Co-Painting Module output is eventually used with the FRIDA simulation.

C. *FRIDA*

In this work, we use the FRIDA [1] robotic painting system for both planning actions from given images and creating self-supervision data. FRIDA uses Real2Sim2Real methodology to enable it to paint with a variety of media such as acrylic paint or markers. Because FRIDA plans using a perceptual loss, it is capable of making, for example, drawings using markers from color photographs.

FRIDA plans actions to guide the robot to make the current canvas look like the output of the Co-Painting Module. Using FRIDA's internal simulation, we visualize a prediction of what the actions will look like, and the robot can execute the actions to update the real canvas, Fig. 3.

IV. EXPERIMENTS

A. *Baselines*

We compare CoFRIDA to the original FRIDA's [1] CLIP-guided text-to-painting method. We investigate the effects of our fine-tuning procedure on Instruct-Pix2Pix in the Co-Painting Module by comparing our method (CoFRIDA) with pre-trained Instruct-Pix2pix (CoFRIDA w/o fine-tuning).

B. *Different Painting Settings*

The FRIDA painting system can paint with various brushes and can have different color constraints. We test CoFRIDA using three different painting settings (1) acrylic painting using one brush and 12 colors which can differ from painting to painting, (2) acrylic painting with a fixed 4-color palette, and (3) a black Sharpie marker. Examples of these three settings are shown in Fig. 4, 9, and 8 . The robot can only be used in one of these settings at a time. However, users can paint using any media of choice, leading to mixed media paintings in Fig. 7.

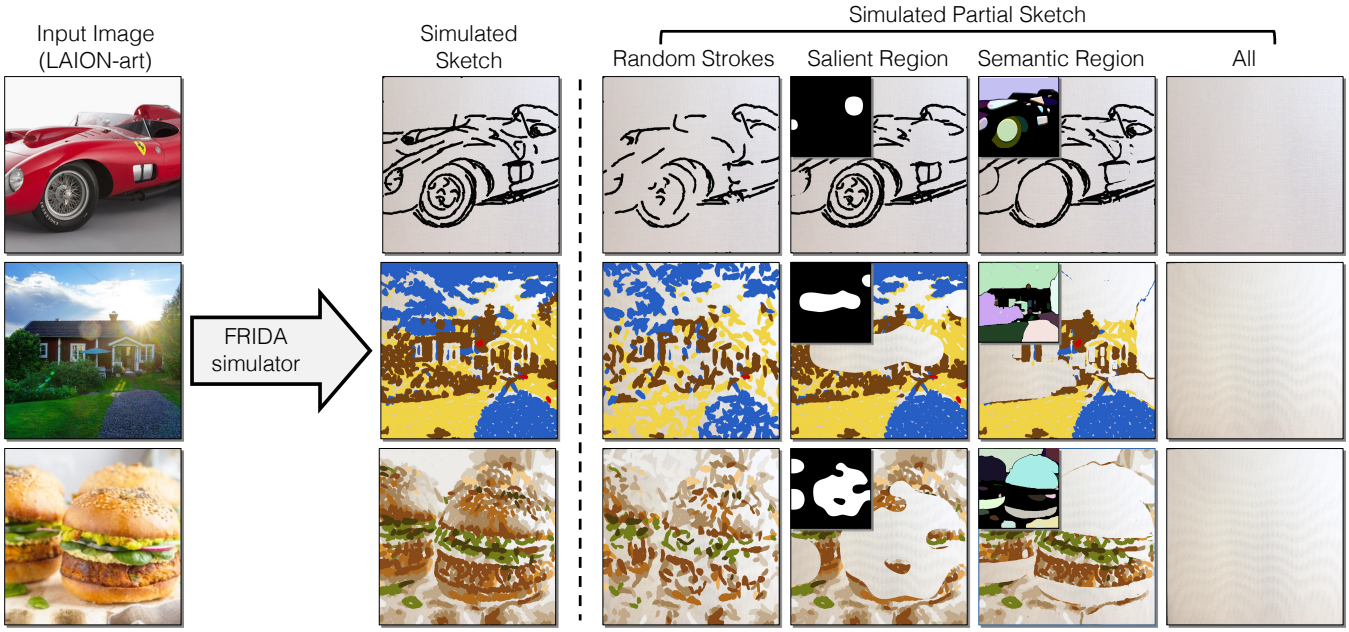

Fig. 4. **Self-Supervised Dataset Creation.** We describe the process of generating the self-supervised training data pairs for fine-tuning the Co-Painting Module. We start with the input images from the LAION-art dataset and convert them into simulated sketch outputs with the FRIDA simulator. Next, we create partial sketches in four different ways: removing random strokes, removing the salient region, removing a semantic region, and removing all strokes.

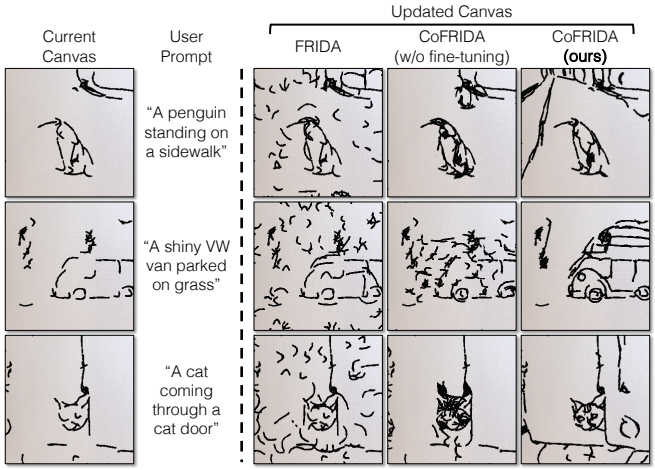

Fig. 5. **Qualitative Comparison.** We show a comparison between three methods of performing text-based canvas updates: FRIDA, CoFRIDA without fine-tuning, and CoFRIDA with fine-tuning (ours). FRIDA uses a CLIP based optimization and generates outputs that are noisy. CoFRIDA without fine-tuning, is not aware of the constraints of the robot and generates an output that is difficult for the robot to execute and often does not satisfy the text prompt specified by the user. In contrast, CoFRIDA outputs an updated canvas that reflects the user prompt without being noisy.

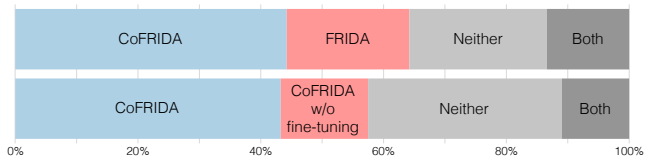

Fig. 6. **User Preference Study.** Results from two MTurk Surveys. Presented with a text description, participants chose which of two drawings (CoFRIDA versus either FRIDA or CoFRIDA without fine-tuning) was more similar to the text, neither, or both. See Fig. 5 for examples.

TABLE I

CLIPSCORES AND BLIPSCORES COMPUTED ON ROBOT SIMULATED DRAWINGS (SEE FIG. 5). SIM-TO-REAL GAP MEASUREMENTS, Δ_{pix} AND Δ_{sem}, MEASURE THE DIFFERENCE BETWEEN THE CO-PAINTING MODULE OUTPUT AND THE SIMULATED DRAWING OF THAT IMAGE.

	CLIPScore ↑	BLIPScore ↑	Δ_{pix} ↓	Δ_{sem} ↓
FRIDA	**0.741**	**0.192**	—	—
CoFRIDA w/o fine-tuning	0.595	0.162	0.195	0.241
CoFRIDA	0.624	0.178	**0.052**	**0.035**

C. Evaluation

Text-Image Alignment - Two automatic methods of comparing image and text are CLIPScore [33] and BLIP-Score [34], which measure the similarity between images and text with a pre-trained image-text encoders. Because FRIDA directly optimizes the CLIPScore to create images from text, this method is unfairly advantaged when using CLIPScore. We use MTurk to achieve large-scale fair evaluation of text-image alignment.

Semantic Sim2Real Gap - It is important that between the output of the Co-Painting Module (I_{cpm}) and the FRIDA simulation (I_{sim}) there is little loss in semantic meaning. A naive approach at measuring this loss is the mean-squared-error between the images' pixels (Eq. 1). However, this is sensitive to low-level variation in details such as color or tone differences which are tolerable as long as the high-level content in the images is the same. To measure the high-level difference, we propose to use the cosine distance between CLIP image embeddings, Δ_{sem}, Eq. 2, referred to as the Semantic Sim2Real Gap.

$$\Delta_{pix} = ||I_{cpm} - I_{sim}||_2^2 \quad (1)$$

$$\Delta_{sem} = \cos(CLIP(I_{cpm}), CLIP(I_{sim})) \quad (2)$$

A proper Sim2Real gap measurement would compare the output of the Co-Painting Module to the real drawing, however, it is infeasible to generate a robust number of real-world samples. Because the Sim2Real gap between the FRIDA simulation and the real drawing is the same across all tested methods, we can fairly use the FRIDA simulations in lieu of the real drawings for comparing the Sim2Real gaps of Co-Painting Module variations.

V. RESULTS

A. Co-Painting

To test the ability of CoFRIDA to work with an existing canvas state, we focus on Sharpie marker drawings where no erasing is possible, forcing the model to have to adapt to and use the existing markings on the page. To create the partial drawing, we generate an image with Stable Diffusion using prompts from the PartiPrompts [35] dataset, then simulate the drawing with just 35 strokes as depicted in Fig. 5. We generated 40 images from different prompts per method. CLIPScore [33], BLIPScore, and Sim2Real gap measures are reported in Table I. Since FRIDA maximizes CLIPScore, it was expected and confirmed to have the highest CLIPScore. BLIP is also expected to correlate with CLIP, leading FRIDA to have an artificially high BLIPScore.

To properly assess the image-text similarity of the drawings from partial sketches, we conducted an MTurk survey summarized in Fig 6. 24 unique participants were shown a language description then two images (one from ours and the other one of the two baselines, in random order). Participants were instructed to choose which image fits the given caption better, or to select neither or both. Each image pair was evaluated by 4 unique participants leading to 160 comparisons per baseline. While many participants found neither image fit the text description (an indicator of the challenging nature of co-painting), CoFRIDA was generally indicated as having clearer content over FRIDA and CoFRIDA without our fine-tuning.

In terms of the proposed Semantic Sim2Real Gap, CoFRIDA outperforms the baselines indicating that our fine-tuning guided Instruct-Pix2Pix to produce images that were less likely to change meaning when painted by FRIDA.

B. Multiple Turns

A co-painting system must be capable of accommodating multiple iterations of human-robot interaction in which the robot adds content but does not completely overwrite the human's prior work. We simulate this by having the robot create sequences of modifications to a simulated painting with different text prompts in Fig. 2. The baseline methods tend to either avoid making changes or make huge changes to the canvas, whereas CoFRIDA makes updates that are more reasonable for the robot to achieve and integrate naturally with prior work.

C. Text Conditioned Paintings

FRIDA's text-to-painting method relies on feedback through CLIP which results in noisy, unclear imagery. We compare CoFRIDA which uses a pre-trained generative model to FRIDA in Fig. 9. CoFRIDA's paintings are far more clear and capture the caption better than FRIDA in various painting settings.

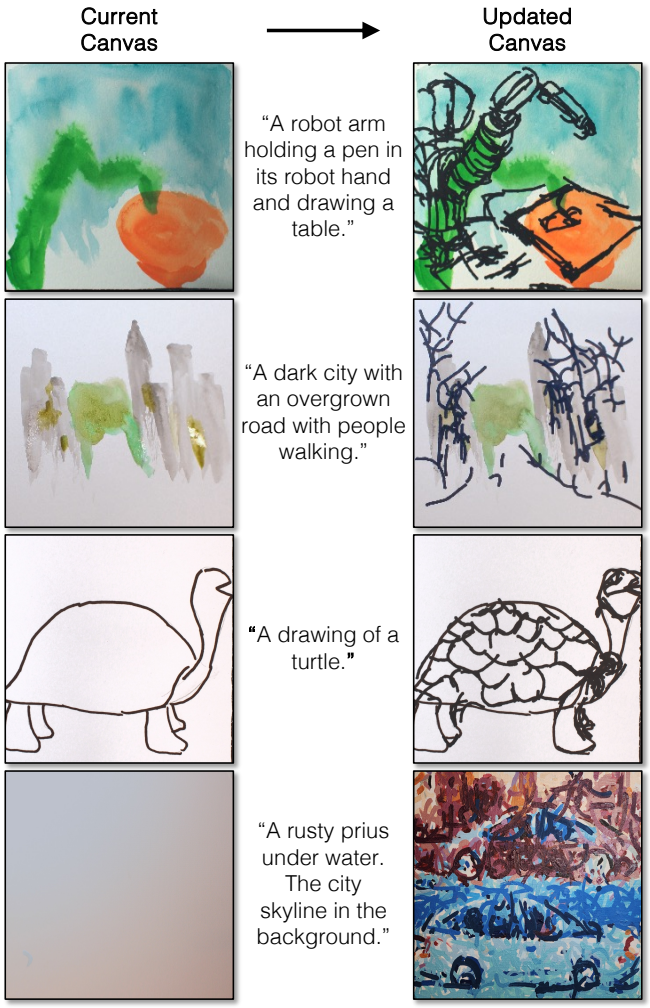

Fig. 7. **Mixed-Media Paintings.** CoFRIDA can use markers and paintbrushes to co-paint with a human. Despite being fine-tuned with a single medium, CoFRIDA can still perform co-painting when a user uses different media such as watercolors.

D. Real Paintings

We used FRIDA's simulation to make large scale data creation and evaluation feasible. Fig. 7 displays multiple real-world examples of CoFRIDA's drawings and paintings. CoFRIDA is able to successfully use content on canvases that is out of distribution from its fine-tuning training data as with the watercolor and marker examples in Fig. 7.

VI. DISCUSSION

Limitations and Ethical Considerations CoFRIDA stands out as a successful collaborative painting system, but is limited to discrete turn-taking interactions. While our self-supervised training data creation method (Fig. 4) was informed by real co-painting data, a more end-to-end approach where the system learns how to form the partial paintings could result in even better results.

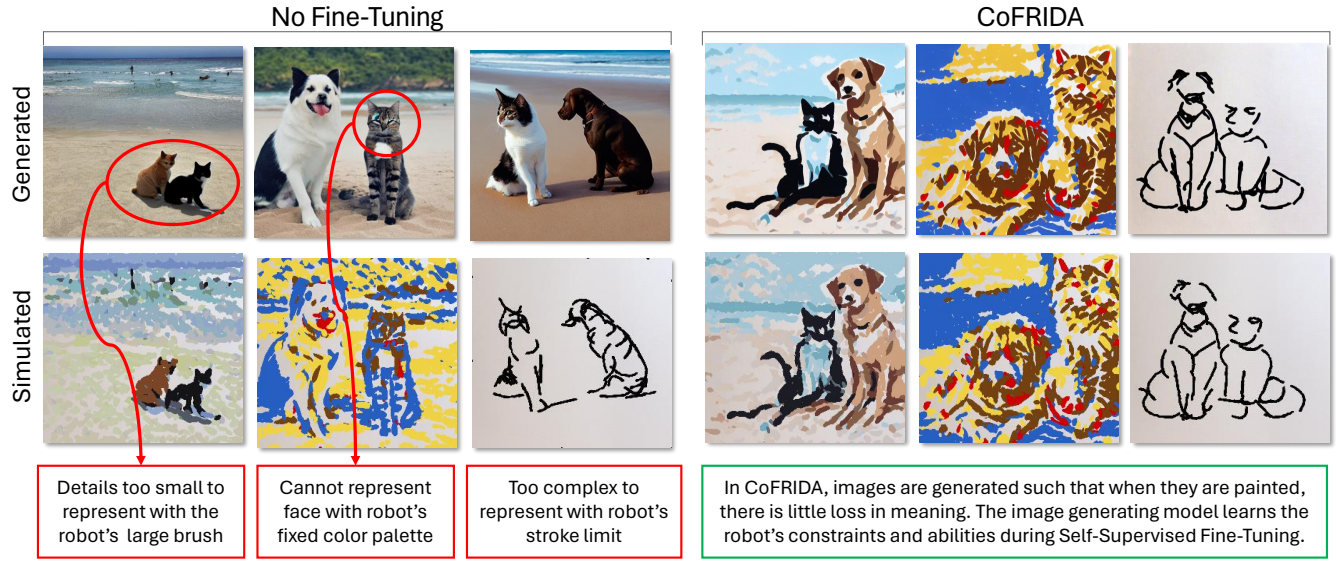

Fig. 8. **Learning Robotic Constraints.** We compare images generated by a pre-trained Stable Diffusion model (left) to those generated by our proposed CoFRIDA module (right) with the prompt "A dog and a cat sitting next to each other on the beach" in three different painting settings (Sec.IV-B). The top row shows the images generated by each of the models and the bottom row shows the corresponding FRIDA simulation.

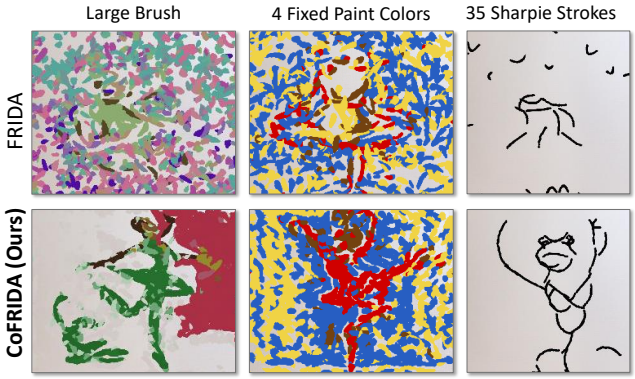

Fig. 9. Comparing CoFRIDA's fine-tuned pre-trained image generator versus FRIDA's CLIP-guided method for generating paintings from the text "A sad, frog ballerina doing an arabesque" in three painting settings.

In light of recent discoveries of harmful content in the LAION dataset, we have inspected the subset of data used to fine-tune the models used in this paper and have not found harmful content. We have changed the code to use the COCO dataset [36] and have not seen a degradation in the quality of results. CoFRIDA is subject to the biases of Stable Diffusion [21] and its training data [37], and so we recommend the usage of CoFRIDA with caution and solely for research purposes.

Learning Robotic Abilities Our self-supervised fine-tuning procedure guided the pre-trained model to generate images that, at a pixel-level, appeared similar to what FRIDA can paint, but is it learning the actual robot constraints or just a low-level style transfer? We computed the Sim2Real gap measurements between the LAION images and their FRIDA simulations (as seen in Fig. 4) along with the CLIPScore of the simulation and text prompt. We found that Δ_{pix} had a small, insignificant Pearson correlation (-0.08, 0.08 p-value) with the CLIPScore of the painting whereas Δ_{sem} had a significant, negative correlation (-0.48, $2.4e-31$ p-value). Because CoFRIDA greatly decreases the Δsem, this indicates that CoFRIDA's fine-tuning technique is not solely changing the low-level appearance (akin to style-transfer) over the output of its base model. It appears that CoFRIDA is learning the robot's abilities, as seen in Fig. 8 where CoFRIDA's Co-Painting Module produces images with (1) very prominent and clear content, when the robot's brush is large (2) select and limited colors, when the robot paints with fixed palettes or markers, and (3) sparse, concise drawings when the number of strokes is limited.

VII. CONCLUSIONS

An end-to-end approach, like FRIDA, that optimizes the brush strokes towards the text goal tends to produce noisy looking paintings that only loosely resemble the text because it operates in a low-level space without a global context. Additionally, it is hard to incorporate interactivity beyond an initial input. We present Collaborative FRIDA (CoFRIDA), a hierarchical approach for interactive human-robot co-painting where semantic planning via pre-trained models happens in a high-level, pixel space before being transferred to a low-level brush stroke planner. Pre-trained models do not immediately provide the requirements for co-painting, as they do not know the capabilities of the robot. Whereas the Real2Sim2Real methodology improves low-level action-space planning in FRIDA, the proposed self-supervised fine-tuning procedure provides a method for adapting powerful pre-trained models for high-level robotic planning. CoFRIDA uses this hierarchical approach for reducing the Sim2Real gap, achieving enhanced performance over the baselines.

VIII. ACKNOWLEDGMENTS

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

REFERENCES

[1] P. Schaldenbrand, J. McCann, and J. Oh, "Frida: A collaborative robot painter with a differentiable, real2sim2real planning environment," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 712–11 718.

[2] H. H. Jiang, L. Brown, J. Cheng, M. Khan, A. Gupta, D. Workman, A. Hanna, J. Flowers, and T. Gebru, "Ai art and its impact on artists," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 363–374. [Online]. Available: https://doi.org/10.1145/3600211.3604681

[3] C. Bateman, "Creating for creatives: A humanistic approach to designing ai tools targeted at professional animators," Ph.D. dissertation, Harvard University, 2021.

[4] N. Davis, C.-P. Hsiao, K. Yashraj Singh, L. Li, and B. Magerko, "Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 2016, pp. 196–207.

[5] F. Ibarrola, T. Lawton, and K. Grace, "A collaborative, interactive and context-aware drawing agent for co-creative design," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[6] T. Lawton, F. J. Ibarrola, D. Ventura, and K. Grace, "Drawing with reframer: Emergence and control in co-creative ai," in *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023, pp. 264–277.

[7] T. Lawton, K. Grace, and F. J. Ibarrola, "When is a tool a tool? user perceptions of system agency in human–ai co-creative drawing," in *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 2023, pp. 1978–1996.

[8] C. Jansen and E. Sklar, "Exploring co-creative drawing workflows," *Frontiers in Robotics and AI*, vol. 8, p. 577770, 2021.

[9] S. Lee and W. Ju, "Adversarial robots as creative collaborators," *arXiv preprint arXiv:2402.03691*, 2024.

[10] M. D. Cooney and M. L. R. Menezes, "Design for an art therapy robot: An explorative review of the theoretical foundations for engaging in emotional and creative painting with a robot," *Multimodal Technologies and Interaction*, vol. 2, no. 3, p. 52, 2018.

[11] M. Cooney and P. Berck, "Designing a robot which paints with a human: visual metaphors to convey contingency and artistry," in *ICRA-X Robots Art Program at IEEE International Conference on Robotics and Automation (ICRA), Montreal QC, Canada*, 2019, p. 2.

[12] M. Cooney, "Robot art, in the eye of the beholder?: Personalized metaphors facilitate communication of emotions and creativity," *Frontiers in Robotics and AI*, vol. 8, p. 668986, 2021.

[13] S. Z. Shaik, V. Srinivasan, Y. Peng, M. Lee, and N. Davis, "Co-creative robotic arm for differently-abled kids: Speech, sketch inputs and external feedbacks for multiple drawings," in *Proceedings of the Future Technologies Conference (FTC) 2020, Volume 3*. Springer, 2021, pp. 998–1007.

[14] Y. Lin, J. Guo, Y. Chen, C. Yao, and F. Ying, "It is your turn: Collaborative ideation with a co-creative robot through sketch," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–14.

[15] D. Herath, J. McFarlane, E. A. Jochum, J. B. Grant, and P. Tresset, "Arts+ health: New approaches to arts and robots in health care," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 1–7.

[16] P. Schaldenbrand and J. Oh, "Content masked loss: Human-like brush stroke planning in a reinforcement learning painting agent," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 505–512.

[17] T. Lindemeier, "e-david: Non-photorealistic rendering using a robot and visual feedback," Ph.D. dissertation, University of Konstanz, 2018.

[18] S. Wang, J. Chen, X. Deng, S. Hutchinson, and F. Dellaert, "Robot calligraphy using pseudospectral optimal control in conjunction with a novel dynamic brush model," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 6696–6703.

[19] Rob Carter and Nick Carter, "Dark factory portraits," http://www.robandnick.com/dark-factory-portraits, 2017.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *CoRR*, vol. abs/2103.00020, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[22] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402.

[23] D. Ha and D. Eck, "A neural representation of sketch drawings," *arXiv preprint arXiv:1704.03477*, 2017.

[24] A. Bidgoli, M. L. De Guevara, C. Hsiung, J. Oh, and E. Kang, "Artistic style in robotic painting; a machine learning approach to learning brushstroke from human artists," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 412–418.

[25] G. Lee, M. Kim, M. Lee, and B.-T. Zhang, "From scratch to sketch: Deep decoupled hierarchical reinforcement learning for robotic sketching agent," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 5553–5559.

[26] M. C. Sola and V. Guljajeva, "Dream painter: Exploring creative possibilities of ai-aided speech-to-image synthesis in the interactive art context," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 5, no. 4, pp. 1–11, 2022.

[27] J. Jongejan, H. Rowley, T. Kawashima, J. Kim, and N. Fox-Gieg, "The quick, draw!-ai experiment," *Mount View, CA, accessed Feb*, vol. 17, no. 2018, p. 4, 2016.

[28] D. Parikh and C. L. Zitnick, "Exploring crowd co-creation scenarios for sketches," *arXiv preprint arXiv:2005.07328*, 2020.

[29] e. a. Juliet Shen, "Co-drawings," 2016. [Online]. Available: https://www.codrawseattle.com/

[30] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[31] Y. Vinker, E. Pajouheshgar, J. Y. Bo, R. C. Bachmann, A. H. Bermano, D. Cohen-Or, A. Zamir, and A. Shamir, "Clipasso: Semantically-aware object sketching," *arXiv preprint arXiv:2202.05822*, 2022.

[32] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[33] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[34] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[35] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, vol. 2, no. 3, p. 5, 2022.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[37] A. Birhane, V. U. Prabhu, and E. Kahembwe, "Multimodal datasets: misogyny, pornography, and malignant stereotypes," *arXiv preprint arXiv:2110.01963*, 2021.