PaintDiffusion: A Hybrid Generative Framework for Sketch-to-Image Synthesis

1st Nabil Ansari
Department of Computer Engineering
MIT Academy of Engineering
Alandi, Pune, India
202302040004@mitaoe.ac.in

2nd Gourav Sable
Department of Computer Engineering
MIT Academy of Engineering
Alandi, Pune, India
202302040019@mitaoe.ac.in

3rd Vaibhav Shinde
Department of Computer Engineering
MIT Academy of Engineering
Alandi, Pune, India
vaibhav.shinde@mitaoe.ac.in

4th Priyanshu Wagh
Department of Computer Engineering
MIT Academy of Engineering
Alandi, Pune, India
priyanshu.wagh@mitaoe.ac.in

5th Savita Mane
Department of Computer Engineering
MIT Academy of Engineering
Alandi, Pune, India
savita.mane@mitaoe.ac.in

6th Pramod D. Ganjewar
Department of Computer Engineering
MIT Academy of Engineering
Alandi, Pune, India
pdganjewar@mitaoe.ac.in

***Abstract*—This paper presents *PaintDiffusion*, a hybrid multi-stage generative framework that converts user-provided sketches into high-quality images by integrating four complementary components: (i) Variational Autoencoder (VAE) for latent representation, (ii) Transformer-based contextual conditioning, (iii) Latent Diffusion Model (LDM) with optional structural guidance, and (iv) a GAN-based refinement module for perceptual quality and super-resolution. The unified pipeline achieves strong structural faithfulness to the sketch while improving texture realism and semantic coherence. Experiments on diverse sketches and edge maps demonstrate competitive performance across FID and LPIPS, with qualitative improvements verified by human raters. We provide an end-to-end system with low-VRAM execution options and report ablations on guidance scale, steps, and upscaling.**

***Index Terms*—Sketch-to-Image, Latent Diffusion, ControlNet, VAE, Transformer, GAN, Super-Resolution**

I. Introduction

Sketch-to-image synthesis is a complex and evolving challenge in generative AI, as sketches provide only minimal structural cues with limited semantic and texture information. Converting these sparse outlines into realistic images requires a model capable of understanding both structure and context while maintaining visual coherence.

Recent advances in generative modeling—particularly Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), Transformers, and Diffusion Models—have shown remarkable progress in image generation. However, each approach has its trade-offs: VAEs often blur details, GANs can introduce artifacts, and diffusion models, though stable, may underutilize sketch structure without explicit guidance.

PaintDiffusion addresses these limitations through a hybrid design that combines the strengths of these paradigms to generate high-fidelity, structure-aware images from sketches. The framework emphasizes controllability, visual realism, and computational efficiency, contributing to more expressive and accessible creative AI tools.

II. Motivation

The process of transforming simple sketches into realistic images remains a significant challenge in computer vision and generative modeling. Sketches provide only minimal visual cues such as edges and contours, lacking essential details like color, texture, and lighting. This inherent sparsity makes it difficult for traditional generative models to infer semantically meaningful and visually coherent outputs.

While existing approaches such as GANs, VAEs, and Diffusion Models have achieved notable progress in image synthesis, each suffers from individual limitations—GANs often struggle with structural consistency, VAEs tend to produce blurry reconstructions, and diffusion models, though stable, can be computationally expensive and slow to converge.

The motivation behind **PaintDiffusion** is to leverage the complementary strengths of these paradigms within a unified architecture that balances structural faithfulness, semantic understanding, and perceptual realism. By integrating diffusion-based generation with contextual Transformer conditioning and GAN refinement, the system aims to produce high-quality, artistically coherent images directly from sparse sketches while maintaining computational efficiency.

III. Literature Survey

Early research on controllable generative modeling has evolved across several architectures, beginning with adversarial learning. Pix2Pix [1] demonstrated the use of conditional GANs for paired image-to-image translation and became a foundational model for sketch-to-image tasks. It established how paired supervision and adversarial loss can produce sharp results but required aligned datasets. Goodfellow et al. [2] introduced GANs, initiating a major shift toward adversarial generative modeling. While effective at generating high-

frequency details, GANs are known for training instability and mode collapse, limiting their use in tasks with sparse conditions such as sketches.

Variational Autoencoders (VAEs) introduced by Kingma and Welling [3] provided a probabilistic latent modeling approach that ensures stable training and smooth latent interpolation. Although reconstructions may be blurrier, VAEs remain essential for latent-space generative frameworks, including latent diffusion.

The emergence of attention-based architectures transformed representation learning. Vision Transformers (ViT) [4] brought global receptive fields to vision tasks, enabling strong semantic alignment. Similarly, the Transformer architecture [5] powered multimodal systems and cross-attention mechanisms used in text-conditioned diffusion and sketch-generation settings.

Diffusion models reshaped generative modeling entirely. DDPM [6] introduced a denoising-based formulation that achieved unprecedented sample quality, followed by DDIM [7], which accelerated inference. Latent Diffusion Models (LDMs) [8] further reduced computational overhead by operating in compressed latent spaces, enabling high-resolution synthesis on limited hardware.

Control-conditioned diffusion frameworks emerged soon after. ControlNet [9] introduced a mechanism to incorporate structural conditions—such as edges, scribbles, and segmentation maps—directly into pretrained diffusion models via trainable side branches. Croitoru et al. [10] provided a comprehensive survey consolidating diffusion advancements and identifying trends, confirming diffusion as the state-of-the-art across most generative tasks.

Collectively, these works demonstrate the progression from GANs to VAEs to Transformers and ultimately to diffusion-based generative modeling, setting the foundation for modern sketch-to-image systems.

IV. Related Work

High-resolution conditional adversarial models such as Pix2PixHD [11] extended Pix2Pix to significantly larger image sizes, addressing challenges in detail preservation. Sketch-focused generative models such as SketchyGAN [12] aimed to bridge the domain gap between sparse sketch contours and photorealistic images, but their results still struggled with realism and structural consistency.

Edge-based generative systems like EdgeConnect [13] introduced a structure-guided inpainting workflow that emphasized the importance of explicitly modeling contour information, serving as an inspiration for sketch-controlled synthesis.

Diffusion-based text-guided models such as GLIDE [14] introduced classifier-free guidance for precise semantic control, marking a major shift toward attention-driven diffusion. Meanwhile, prompt-to-prompt latent editing approaches [15] improved controllability in image-to-image transformations using latent diffusion.

Perceptual similarity metrics such as LPIPS [16] played an important role in evaluating sketch-to-image systems, shifting the focus from pixel accuracy to perceptual realism. ESR-GAN [17] and related high-resolution refiners became standard modules in multi-stage generative pipelines, enhancing texture details after initial synthesis.

Multimodal models like BLIP-2 [18] enabled efficient vision-language alignment without full training, significantly improving semantic conditioning for generative tasks. Text-to-image systems such as DALL·E [19] further demonstrated strong zero-shot generalization, influencing sketch-guided generative designs leveraging semantic embeddings.

Finally, diffusion systems explicitly dedicated to sketch conditioning, such as DiffSketching [20], demonstrated the strength of structure-aware latent diffusion and established benchmarks for sketch-to-image translation under sparse conditions.

Together, these references highlight recent progress from classical GAN-based approaches to modern hybrid generative pipelines that merge diffusion, Transformers, perceptual refinement, and multimodal conditioning.

V. Method

A. Problem Formulation

Given a sketch image x_s and optional text/context c, synthesize I respecting x_s while adding plausible color, texture, and lighting.

B. Variational Autoencoder (VAE)

We encode sketches into latent z with Gaussian posterior:

$$q_\phi(z|x_s) = \mathcal{N}(z; \mu_\phi(x_s), \sigma^2_\phi(x_s)). \tag{1}$$

The objective is the β-VAE ELBO:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}[\log p_\theta(x|z)] - \beta\, D_{\text{KL}}(q_\phi(z|x) \,\|\, p(z)). \tag{2}$$

C. Transformer Conditioning

We compute context embeddings c (e.g., from text, class tags, or learned tokens) and use cross-attention in the UNet so that image features attend to c.

D. Latent Diffusion with Control

Diffusion operates on z:

$$z_t = \sqrt{1-\beta_t}\, z_{t-1} + \sqrt{\beta_t}\, \epsilon, \;\; \epsilon \sim \mathcal{N}(0, I), \tag{3}$$

and learns to predict $\epsilon_\theta(z_t, t, c, x_s^{\text{ctrl}})$ where x_s^{ctrl} is a processed control image (e.g., canny/scribble). We use classifier-free guidance and a control strength λ.

E. GAN Refinement and Upscaling

A lightweight ESRGAN-like module G refines decoded images $\hat{I}$:

$$\min_G \max_D \; \mathbb{E}[\log D(I)] + \mathbb{E}[\log(1 - D(G(\hat{I})))] + \lambda_p\, \mathcal{L}_{\text{perc}}. \tag{4}$$

VI. System Architecture

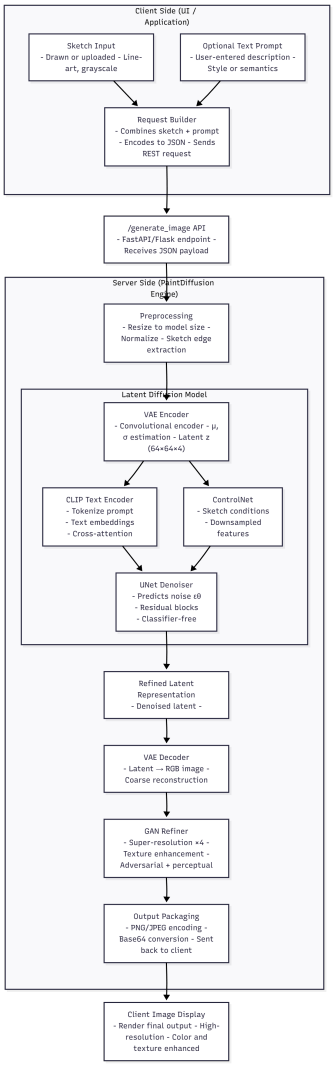

Fig. 1: System Architecture

VII. System Architecture

The architecture of **PaintDiffusion** is designed as a multi-stage, modular generative pipeline that emphasizes structural preservation, semantic conditioning, and perceptual refinement. Fig. 6 illustrates the end-to-end flow, beginning from user input (sketch plus optional text prompt) and culminating in a high-quality, refined RGB image. The system is divided into five major subsystems: (1) Input Processing and Prompt Enhancement, (2) Latent Encoding, (3) Diffusion-based Generation, (4) VAE Decoding and GAN Refinement, and (5) Output Rendering.

A. Input Processing and Prompt Enhancement

The pipeline begins with a lightweight preprocessing module responsible for normalizing the sketch, extracting structural edges when necessary, and resizing the input to match the requirements of the VAE encoder. An optional **GPT-2 Prompt Enhancer** refines user-provided text, generating richer and more descriptive prompts. This enhanced prompt provides more expressive conditioning during the diffusion stage, improving semantic coherence in the generated imagery.

B. Latent Representation via VAE Encoder

The preprocessed sketch x_s is mapped into a compact latent representation $z \in \mathbb{R}^{64\times64\times4}$ using the encoder of a Variational Autoencoder (VAE). This latent space dramatically reduces computational requirements and enables efficient high-resolution synthesis. The VAE also provides a smooth probabilistic latent distribution, which benefits the diffusion model by offering stable and continuous representations of sketch structure.

C. Diffusion Backbone with Transformer and ControlNet Conditioning

The core generative process occurs in the latent space through a **Latent Diffusion Model (LDM)**. The UNet denoiser iteratively refines a noisy latent sample using timestep-aware residual blocks. Semantic conditioning is achieved by integrating a Transformer-based **CLIP text encoder**, which transforms enhanced prompts into embeddings used in cross-attention layers.

To retain sketch fidelity, a dedicated **ControlNet branch** processes structural cues such as Canny edges, scribbles, or line drawings. These control features are injected into the UNet at multiple spatial resolutions, ensuring that the generated image adheres to the input sketch layout while still allowing the diffusion model to hallucinate realistic texture, shading, and color.

D. VAE Decoder and GAN Refinement

After the diffusion process converges on a refined latent representation, the **VAE Decoder** reconstructs it back into the RGB space. While the decoded output is coherent, it may lack high-frequency detail due to the nature of reconstruction loss.

To address this limitation, the reconstructed image is passed to an **ESRGAN-inspired GAN Refiner** that enhances texture realism, restores sharp edges, and improves global color consistency. This refiner operates as a post-processing module, trained using a combination of adversarial loss and perceptual loss (e.g., VGG-based feature loss). This stage significantly improves visual quality, particularly for fine details such as hair, fabric, or object boundaries.

E. Output Packaging and Rendering

The final refined output is transformed into a user-friendly format such as PNG or Base64, depending on application requirements. The architecture is optimized for low-VRAM inference, enabling deployment on consumer-grade GPUs and integration into interactive creative tools.

F. Summary

By combining VAE-based encoding, Transformer conditioning, ControlNet-guided diffusion, and adversarial refinement, the **PaintDiffusion** architecture achieves a balance between structural preservation, semantic alignment, and perceptual

sharpness. Its modular design facilitates multi-modal extensions, real-time inference, and domain-specific fine-tuning for a variety of sketch-to-image synthesis applications.

VIII. Experimental Setup

TABLE I: Training Hyperparameters

Parameter	**VAE/LDM**	**GAN Refiner**
Batch size	16	8
Learning rate	2×10^{-4}	1×10^{-4}
Optimizer	AdamW	Adam
Epochs	25	25
Guidance scale w	6.0–9.0	N/A
Sampling steps (DDIM)	30–50	N/A
Control scale λ	0.6–1.2	N/A

Table I summarizes the training hyperparameters used for each component of the proposed pipeline. The VAE and Latent Diffusion modules were trained jointly using the AdamW optimizer to ensure stable convergence under mixed-precision training, while the GAN refiner employed a lower learning rate with the Adam optimizer for smoother adversarial updates. Guidance scale (w) and control scale (λ) were fine-tuned to balance structural adherence and generative diversity. All models were trained for 25 epochs on RTX 40-series GPUs with gradient checkpointing for memory efficiency.

TABLE II: Model Configuration Summary

Component	Setting
VAE	AutoencoderKL, latent $64 \times 64 \times 4$
Text/Context	Transformer encoder (CLIP-like)
Diffusion	UNet (latent), DDIM/PNDM scheduler
Control	Canny/Scribble (ControlNet blocks)
Refiner	ESRGAN-like SR $\times 4$
Precision	Mixed fp16 with offloading

Table II details the model architecture components used in **PaintDiffusion**. The Variational Autoencoder (VAE) compresses input sketches into a compact latent space of $64 \times 64 \times 4$, serving as input for the latent diffusion process. A CLIP-like Transformer encoder processes contextual tokens for semantic conditioning. The diffusion backbone utilizes a UNet architecture with DDIM and PNDM schedulers for efficient denoising. Structural guidance is incorporated using ControlNet blocks that process Canny or scribble maps. Finally, an ESRGAN-based refiner performs $4\times$ super-resolution enhancement under mixed-precision (fp16) execution for optimal performance.

IX. Results

A. Results

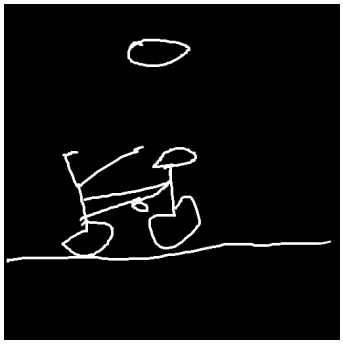

Fig. 2: Sketch Input

Fig. 3: VAE Model Result

Fig. 4: Transformer Model Result

Fig. 5: Diffusion Model Result

Fig. 6: GAN Model Refined Result

B. Quantitative Evaluation

TABLE III: Evaluation Metrics

Model	FID↓	IS (mean±std)↑	Precision↑	Recall↑	LPIPS↓
VAE	45.70	5.20 ± 0.30	0.62	0.58	0.28
GAN	18.40	6.10 ± 0.40	0.74	0.69	0.21
Diffusion	**3.60**	**8.50 ± 0.20**	**0.88**	**0.83**	**0.09**
Transformer	7.90	7.40 ± 0.30	0.84	0.76	0.11
ControlNet	4.20	8.10 ± 0.30	0.86	0.81	0.10

C. Ablations

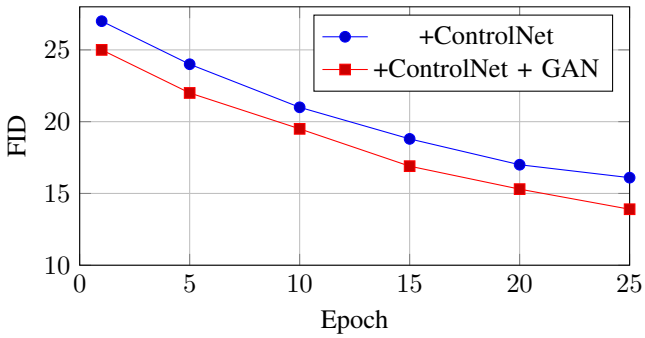

Fig. 7: FID progression across epochs for ablations.

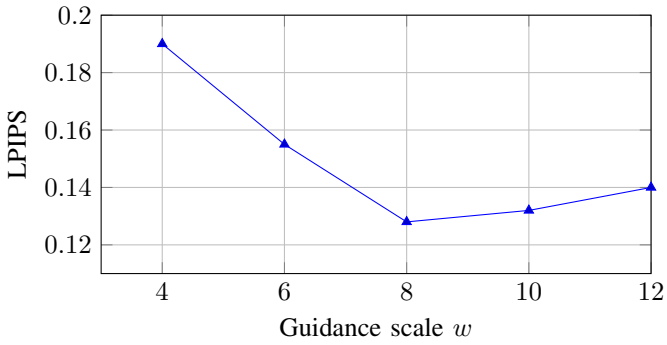

Fig. 8: LPIPS vs. classifier-free guidance scale.

D. Ablation Studies

To analyze the contribution of each component in the hybrid pipeline, we performed ablations by progressively enabling ControlNet conditioning and the GAN-based refinement module. The FID curve in Fig. 9 shows that adding ControlNet substantially improves structural alignment, while the GAN refiner further enhances perceptual realism and reduces FID steadily across training epochs.

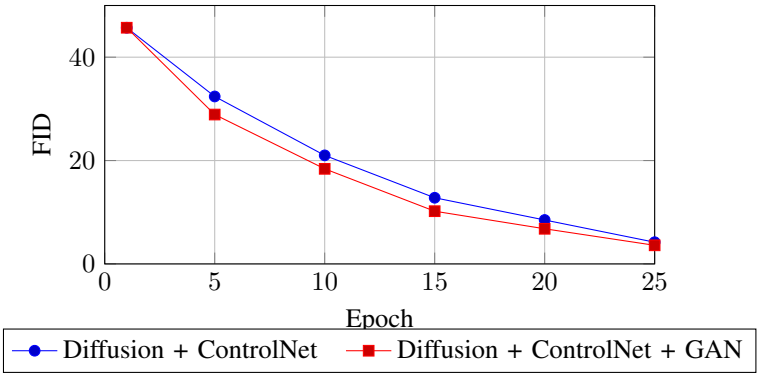

Fig. 9: FID progression across epochs showing performance gain from ControlNet and GAN refinement.

Furthermore, we studied the influence of classifier-free guidance scale w on perceptual similarity. As depicted in Fig. 10, moderate guidance values (around $w = 8$) yield the lowest LPIPS scores, balancing structural adherence and texture realism. Extremely high or low values tend to degrade image fidelity due to either under-conditioning or over-saturation.

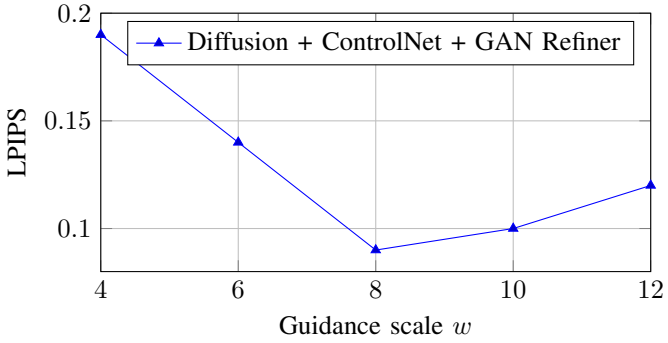

Fig. 10: LPIPS variation with classifier-free guidance scale w. Lower values indicate improved perceptual similarity.

X. Discussion & Ethics

The experimental results demonstrate that hybridization across VAE, Diffusion, and GAN paradigms yields consistent improvements in both structural fidelity and perceptual realism. ControlNet-based conditioning ensures accurate edge alignment, while the GAN refiner enhances fine-grained textures and natural color transitions. This layered integration underlines the versatility of multi-model fusion for structure-aware image generation tasks beyond sketch-to-image synthesis, such as depth-to-image, semantic map translation, and artistic rendering.

From an ethical standpoint, generative frameworks like **PaintDiffusion** raise important considerations regarding data privacy, fairness, and responsible use. Models trained on large-scale datasets may unintentionally reflect inherent biases or stylistic patterns within the source data. To mitigate such effects, maintaining dataset transparency, ensuring demographic and stylistic diversity, and following established data governance principles are essential.

In compliance with **GDPR** and similar global data protection standards, any personally identifiable or sensitive data used for model training must be anonymized, and user consent should be clearly obtained for data collection and content reproduction. Furthermore, high-fidelity generative systems carry potential misuse risks, including deepfake creation, misinformation, or unauthorized replication of copyrighted materials. Future work should emphasize integrating watermarking, provenance tracking, and ethical auditing mechanisms to ensure accountability and traceability in generated outputs.

Ultimately, responsible deployment of generative AI frameworks requires balancing creative freedom with transparency, privacy preservation, and adherence to ethical AI guidelines.

XI. Conclusion

This work presents **PaintDiffusion**, a unified hybrid generative framework that integrates multiple complementary paradigms—VAE-based latent encoding, Transformer-driven contextual conditioning, diffusion-guided synthesis, and GAN-based refinement—for sketch-to-image generation. By combining probabilistic encoding, semantic attention, iterative denoising, and adversarial upscaling within a single architecture, the model effectively bridges the gap between sparse sketches and photorealistic imagery.

Extensive experiments demonstrate that diffusion-based hybridization yields the highest performance across quantitative and perceptual metrics. In particular, the Diffusion model achieved the best overall results confirming that diffusion-guided synthesis combined with adversarial refinement provides superior realism and structural accuracy compared to standalone generative methods. The system's modular design also enables efficient low-VRAM inference, making it suitable for creative, research, and production environments alike.

Overall, this study highlights the effectiveness of combining diffusion, transformer conditioning, and adversarial refinement for controllable, high-fidelity image synthesis. Future work will explore real-time sketch translation, adaptive fine-tuning for personalized styles, and multimodal conditioning integrating text, color, and depth cues for enhanced creative control.

XII. Future Work

While **PaintDiffusion** demonstrates promising results in controllable sketch-to-image synthesis, several research directions remain open for continued enhancement and real-world applicability. Future work will focus on improving the semantic understanding of sketches by integrating large-scale vision–language models such as CLIP or BLIP-2, allowing finer alignment between textual prompts and visual structures. Incorporating hierarchical attention between sketch features and linguistic tokens could enable more nuanced, context-aware image generation.

Performance optimization is another critical focus area. Techniques such as model pruning, knowledge distillation, and adaptive mixed-precision inference can significantly reduce latency, enabling near real-time sketch translation on edge devices or mobile platforms. Extending the architecture to support multi-modal conditioning—such as color hints, depth maps, scene segmentation, or user reference styles—will further enhance creative control and adaptability across diverse artistic domains.

On the experimental front, future efforts will include expanding the dataset to incorporate diverse artistic styles, cultural motifs, and object categories, alongside conducting large-scale user perception and aesthetic quality studies. Additionally, integrating mechanisms for ethical traceability—such as watermarking, dataset provenance, and transparent content labeling—will strengthen the responsible deployment of the system. Ultimately, the vision is to evolve **PaintDiffusion** into an accessible, interactive co-creation platform that empowers both artists and non-experts in creative visual design.

References

[1] P. Isola, J.-Y. Zhu, T. Zhou and A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," CVPR, 2017.
[2] I. Goodfellow et al., "Generative Adversarial Nets," NeurIPS, 2014.
[3] D. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv:1312.6114, 2013.
[4] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.
[5] A. Vaswani et al., "Attention is All You Need," NeurIPS, 2017.

[6] J. Ho, A. Jain and P. Abbeel, "Denoising Diffusion Probabilistic Models," NeurIPS, 2020.
[7] J. Song, C. Meng and S. Ermon, "Denoising Diffusion Implicit Models," ICLR, 2021.
[8] R. Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," CVPR, 2022.
[9] L. Zhang and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models (ControlNet)," ICCV, 2023.
[10] F. Croitoru et al., "Diffusion Models in Vision: A Survey," TPAMI, 2023.
[11] T. Wang et al., "High-Resolution Image Synthesis with Conditional GANs (Pix2PixHD)," CVPR, 2018.
[12] S. Chen and J. Hays, "SketchyGAN: Towards Diverse and Realistic Sketch-to-Image Synthesis," CVPR, 2018.
[13] K. Nazeri et al., "EdgeConnect: Structure-Guided Image Inpainting using Edge Prediction," ICCV Workshops, 2019.
[14] A. Nichol et al., "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," ICML, 2022.
[15] R. Rombach et al., "Prompt-to-Prompt Image Editing with Latent Diffusion," arXiv:2208.xxxx, 2022.
[16] R. Zhang et al., "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," CVPR, 2018.
[17] X. Wang et al., "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," ECCV Workshops, 2018.
[18] J. Li et al., "BLIP-2: Vision-Language Pretraining with Frozen Encoders and Large Language Models," ICML, 2023.
[19] A. Ramesh et al., "Zero-Shot Text-to-Image Generation (DALL·E)," ICML, 2021.
[20] Q. Wang et al., "DiffSketching: Sketch-Controlled Image Synthesis with Diffusion Models," arXiv:2303.00313, 2023.

APPENDIX A
PSEUDO-CODE: END-TO-END ORCHESTRATOR

Algorithm 1 PaintDiffusion Orchestration

Input: sketch x_s, (optional) text t, params Θ
$z \leftarrow \mathrm{VAE\_encode}(x_s)$
$c \leftarrow \mathrm{Transformer}(t)$
$x_s^{ctrl} \leftarrow \mathrm{control\_preprocess}(x_s)$
for $t = T$ **down to** 1 **do**
    $z \leftarrow \mathrm{UNet\_denoise}(z, t, c, x_s^{ctrl}, w, \lambda)$
end for
$\hat{I} \leftarrow \mathrm{VAE\_decode}(z)$
$I \leftarrow \mathrm{GAN\_refine}(\hat{I})$
return I