Fuse-and-Diffuse: A Hybrid GAN-Diffusion Framework for Disentangled and Controllable Sketch-to-Image Synthesis

Ankush Jain[†] **Ravi Kumar**[*] **Siddhant**[*]

[†]Department of Computer Science and Engineering, NSUT, New Delhi, India
[*]B.Tech Student, Department of Computer Science and Engineering, NSUT, New Delhi, India

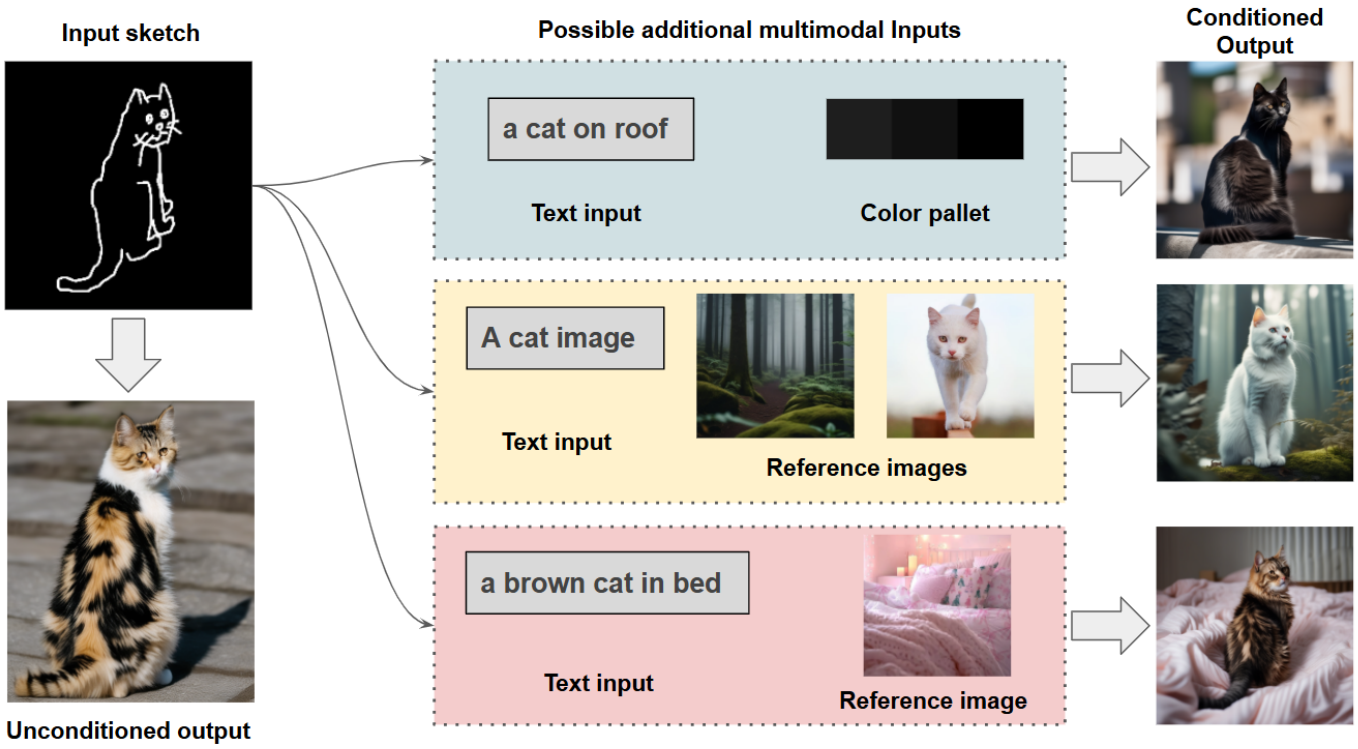

Fig. 1. **Overview of the Fuse-and-Diffuse sketch-to-image synthesis framework.** The pipeline begins with a user-provided input sketch (left), which can be converted into an unconditioned output using the base diffusion model. To enable fine-grained and interpretable user control, the model supports multiple additional conditioning modalities (center), including: (1) *text input* describing the desired scene or attributes, (2) *color palettes* specifying global color tone, (3) *reference images* for style, appearance, or environmental cues. These heterogeneous signals are fused through multimodal adapters that operate without retraining the diffusion backbone. The fused conditioning then guides the LDM to produce structurally consistent, semantically aligned, and user-controlled output images (right). The figure demonstrates how different combinations of text, color, and reference images lead to distinctly modulated outputs, all while preserving the structural alignment imposed by the sketch.

***Abstract*—Stable Diffusion models have demonstrated remarkable capability in generating photorealistic images; however, achieving fine-grained, user-controllable synthesis remains challenging. In sketch-based image generation, this difficulty is compounded by the nature of human-made sketches—unlike edge maps, these sketches often contain abstract, exaggerated, or incomplete lines that do not correspond directly to real image boundaries. Moreover, maintaining object-level consistency across multiple generated images poses a persistent problem. Existing methods typically offer limited control, relying on a single or few conditioning inputs, which can lead to ambiguity in user intent. To address these issues, we propose Fuse-and-Diffuse, a hybrid generative framework that integrates multiple conditioning modalities—text, sketch, reference image, and color information—to provide more granular and interpretable control. The model first refines the input sketch into a structurally coherent line-art representation and then employs a Latent Diffusion Model conditioned on the fused multimodal signals. Through a modular adapter design, our approach effectively disentangles and integrates these modalities without retraining the base diffusion backbone. Experimental results demonstrate that Fuse-and-Diffuse delivers enhanced structural coherence, consistent object representation, and superior user controllability compared to existing sketch-to-image synthesis approaches. Implementation details of the code is given in https://github.com/RaviKumar300/Fuse-and-Diffuse.**

***Index Terms*—Sketch-to-Image Synthesis, Hybrid GAN-Diffusion Model, Multi-Modal Conditioning, Controllable Image Generation, Disentangled Representation.**

I. Introduction

Recent years have witnessed a paradigm shift in generative modeling, largely driven by the remarkable success of Denoising Diffusion Probabilistic Models (DDPMs) [1]. These models have demonstrated an unprecedented ability to synthesize high-fidelity and diverse images, surpassing Generative Adversarial Networks (GANs) in many benchmarks, particularly in training stability and output quality. The introduction of Latent Diffusion Models (LDMs) [2], which perform the diffusion process in a compressed latent space, further democratized this power. This, combined with cross-attention conditioning from models like CLIP [3], enabled powerful text-to-image (T2I) generation and set a new state-of-the-art.

Despite their success, a fundamental limitation of traditional T2I models is their reliance on textual prompts alone for guidance. Text is an inherently ambiguous and non-spatial modality, making it difficult to specify complex spatial layouts, precise object structures, or nuanced artistic styles. This limitation has spurred a new wave of research into *controllable* image synthesis, where generation is guided by more explicit, often multi-modal, conditioning signals. Sketch-to-image synthesis, in particular, has emerged as a critical and challenging task [4], [5], as it provides an intuitive, human-centric interface for creative control. However, free-hand sketches are often sparse, ambiguous, and structurally inconsistent, posing a significant challenge for generative models.[1]

To address this, a popular and parameter-efficient approach has been the development of lightweight adapters that "plug into" large, pre-trained T2I models [6]. Architectures like ControlNet [7] have achieved state-of-the-art spatial control by creating a trainable copy of the model's encoder and injecting structural conditions (e.g., sketches, depth maps) via zero-convolution layers. Concurrently, models like IP-Adapter [8] have enabled powerful style transfer by extracting image features and injecting them into the cross-attention layers using a decoupled attention strategy.

However, a significant and unresolved challenge arises when attempting to *simultaneously* compose these heterogeneous controls. Simply combining adapters often leads to destructive interference, or "concept bleeding". For instance, the style adapter may exhibit "content leakage," where object structures from the style reference override the textual or structural guidance. Conversely, a strong structural prior from ControlNet may inhibit the expression of the desired style. This "battle" for dominance highlights a fundamental flaw: existing methods lack a mechanism for true, disentangled multi-modal fusion. Moreover, these methods often inherit an "imbalanced division of responsibilities" , where a single generative module must act as both a "designer" (interpreting structure) and a "painter" (rendering content). This is particularly problematic when the input sketch itself is noisy, as the model may faithfully reproduce the *flaws* rather than the *intent* of the sketch.

In this paper, we propose **Fuse-and-Diffuse**, a novel two-stage hybrid framework designed explicitly to solve the challenges of disentanglement and composability. Our framework separates the roles of "designer" and "painter" based on the principle of "separation of concerns".

Stage 1, the 'Fuse' Stage, acts as the designer. It employs a novel Dual-Branch GAN [9] to refine the raw, ambiguous user sketch S_{raw} into a clean, structurally coherent line-art representation, $S_{refined}$. This preprocessing step ensures the diffusion model receives an unambiguous structural prior. Stage 2, the 'Diffuse' Stage, acts as the painter. It is built upon a *frozen* LDM [2]. We introduce a novel **Adapters Cluster** ($\mathcal{C}_{\mathcal{A}}$), which injects four parallel conditioning signals: structure ($S_{refined}$), semantics (y_{text}), style (I_{style}), and color (P_{color}). The core of our contribution is a novel **Adaptive Gating Unit (AGU)** that dynamically arbitrates between these modalities. Inspired by gated multimodal units and feature-wise transformations like FiLM [10], the AGU learns to predict scalar weights for each adapter based on the semantic context (from text) and the current diffusion timestep t. This allows the model to, for example, intelligently suppress the color adapter if the text prompt asks for "monochrome," or prioritize structure at early denoising steps. This adaptive fusion resolves modal conflicts and enables true, composable control.

Our contributions are as follows:

- We propose *Fuse-and-Diffuse*, a two-stage hybrid framework that disentangles structural refinement from multi-modal synthesis, addressing the "imbalanced division of responsibilities" in end-to-end models.
- We design an *Adapters Cluster* for the 'Diffuse' stage, enabling simultaneous, fine-grained control over structure, semantics, artistic style, and global color distribution. This includes a novel differentiable histogram module for palette embedding.
- We propose a novel *Adaptive Gating Unit (AGU)* that learns to dynamically fuse multi-modal signals, resolving adapter conflicts and achieving true generative composability.

The rest of the paper is organized as follows. Section II reviews related work on sketch-to-image synthesis and generative modeling. Section III describes the proposed *Fuse-and-Diffuse* framework, detailing its dual-stage design and multi-modal conditioning strategy. Section IV outlines the experimental setup, including datasets and evaluation protocols. Section V presents the results and discussion, demonstrating the effectiveness of the proposed method. Finally, Section VI concludes the paper and suggests directions for future research.

II. Related Work

Our research builds upon several key domains in generative modeling: image-to-image translation, controllable diffusion models, and multi-modal fusion techniques.

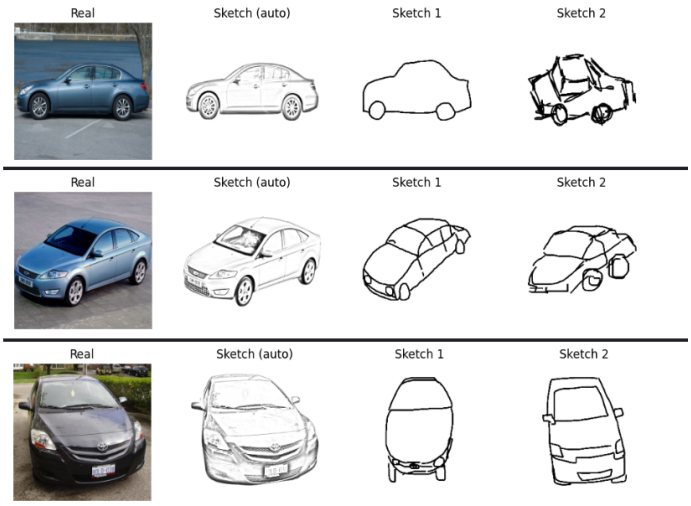

Fig. 2. Examples from a sketch-drawing dataset showing real car images, automatically extracted edge maps, and corresponding hand-drawn sketches. Unlike edge maps, human sketches exhibit distortions, missing structure, exaggerated shapes, or incorrect geometry, making the sketch-to-image task significantly more challenging.

A. Generative Image-to-Image Translation

The task of translating an image from a source domain to a target domain has been a cornerstone of generative modeling. Seminal works in this area were dominated by Generative Adversarial Networks (GANs). Pix2Pix [13] demonstrated remarkable results for *paired* translation using a conditional GAN with a U-Net generator and a PatchGAN discriminator. For the more challenging *unpaired* translation task, CycleGAN [14] and DualGAN [9] introduced cycle-consistency losses, enabling translation without corresponding ground truths. While powerful, these GAN-based methods often suffer from training instability and mode collapse. While edge maps provide clean and structurally accurate guidance, real human sketches are often noisy, abstract, and geometrically inconsistent. As

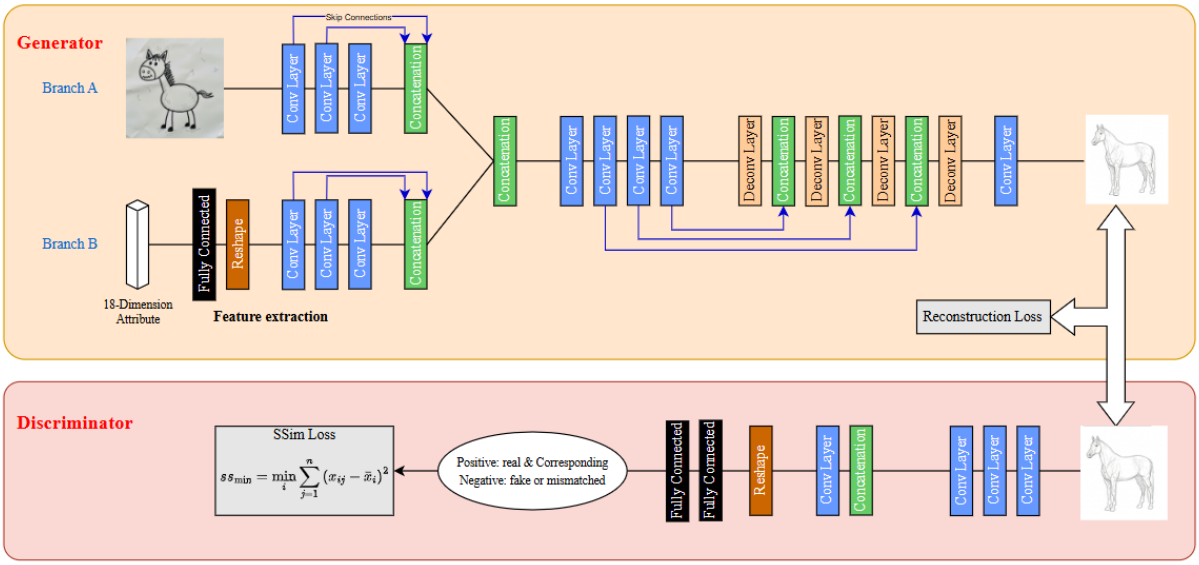

Fig. 3. **Overview of the GAN-based Sketch Refinement Module.** This stage forms the first component of our hybrid pipeline. A **dual-branch GAN** refines the raw input sketch to produce a clean, structurally coherent line-art representation. **Branch A** employs a U-Net to capture and enhance the sketch's spatial and structural details, while **Branch B** integrates an 18-dimensional attribute vector to inject semantic guidance. The fused representation is optimized through a **Reconstruction Loss**, ensuring structural consistency, while a dedicated **Discriminator** employs a Structural Similarity (SSIM) Loss to enforce realism and fidelity. The output of this stage serves as the refined sketch input for the diffusion-based synthesis module.

illustrated in Fig. 2, hand-drawn sketches may omit essential contours, distort object proportions, or introduce exaggerated strokes that deviate significantly from the true structure of the source image. Traditional GAN-based translation models struggle under these conditions, as they rely heavily on accurate spatial correspondences between input and output domains. This motivates the need for a more robust multimodal framework capable of interpreting imperfect sketches while maintaining semantic and structural coherence in the generated output.

More recently, Denoising Diffusion Probabilistic Models (DDPMs) [1] and Latent Diffusion Models (LDMs) [2] have become the state-of-the-art. Our work utilizes both paradigms: specialized GAN in Stage 1 for the constrained task of sketch refinement, and pre-trained LDM in Stage 2 for the high-fidelity synthesis task.

B. Controllable Diffusion Models

A significant advancement in LDMs has been the development of lightweight "adapters" to guide the generative process. These adapters allow for fine-grained control without retraining the massive base models. The most prominent example is ControlNet [7], which creates a trainable copy of the LDM's encoder blocks and injects spatial conditions (e.g., edges, pose, depth) via specialized "zero-convolution" layers.[2] T2I-Adapter [6] proposed a similar, even more lightweight, approach by injecting features at different scales.[3] These methods provide the inspiration for our Structural Adapter ($\mathcal{A}_{\text{struct}}$).

Parallel to structural control, IP-Adapter [8] enabled image-prompting, or style control. It uses a CLIP image encoder to extract style features and injects them into the LDM's cross-attention layers using a *decoupled cross-attention* mechanism. This allows for the simultaneous use of text and image prompts.This work forms the basis for our Style Adapter ($\mathcal{A}_{\text{style}}$). Earlier efforts like StyleCLIP [15] had previously validated the power of using CLIP embeddings to guide generative models, in their case, StyleGAN.

C. Sketch-to-Image Synthesis

Sketch-to-image is a long-standing and challenging image-to-image translation task. Early methods relied on GANs [13]. More recent work has focused on adapting diffusion models for this task [4], [5]. However, a key limitation of these end-to-end approaches is their sensitivity to the quality of the input sketch. Free-hand sketches are often noisy, sparse, and ambiguous. This forces the model to simultaneously interpret the user's messy *intent* and render a photorealistic image, a problem described as an "imbalanced division of responsibilities" . Our *Fuse-and-Diffuse* framework directly addresses this by dedicating Stage 1 to explicitly refining the sketch, providing the synthesis stage with a clean, unambiguous structural prior.

III. Methodology

A. Overall Architecture

The Fuse-and-Diffuse framework addresses the ill-posed nature of photorealistic image synthesis from free-hand sketches through a two-stage hybrid architecture that explicitly separates structural interpretation from content rendering. This separation of concerns rectifies the imbalanced division of responsibilities observed in contemporary single-stage models, where a unified generative module must simultaneously function as both designer and painter.

Our framework operates through two distinct stages. Stage 1 (Fuse) employs a Dual-Branch Generative Adversarial Network G_{refine} to transform raw, ambiguous user sketches S_{raw} into clean, structurally coherent line-art representations $S_{refined}$. Stage 2 (Diffuse) leverages a frozen pre-trained Latent Diffusion Model (LDM) to synthesize the final photorealistic image $\hat{I}$ under multiple conditioning signals. The key innovation lies in the Adapters Cluster $\mathcal{C}_{\mathcal{A}}$, comprising four lightweight, composable adapters that inject disentangled multi-modal conditioning signals into the frozen LDM without costly retraining.

The complete generative pipeline proceeds as follows: (1) the raw sketch S_{raw} is refined by Stage 1: $S_{refined} = G_{\text{refine}}(S_{raw})$; (2) users provide three additional modalities: text prompt y_{text}, style reference image I_{style}, and target color palette P_{color}; (3) these four signals $\{S_{refined}, y_{text}, I_{style}, P_{color}\}$ are processed by their respective adapters within $\mathcal{C}_{\mathcal{A}}$; (4) an Adaptive Gating Unit dynamically fuses these conditioning signals based on semantic and temporal context; (5) the fused signals guide the

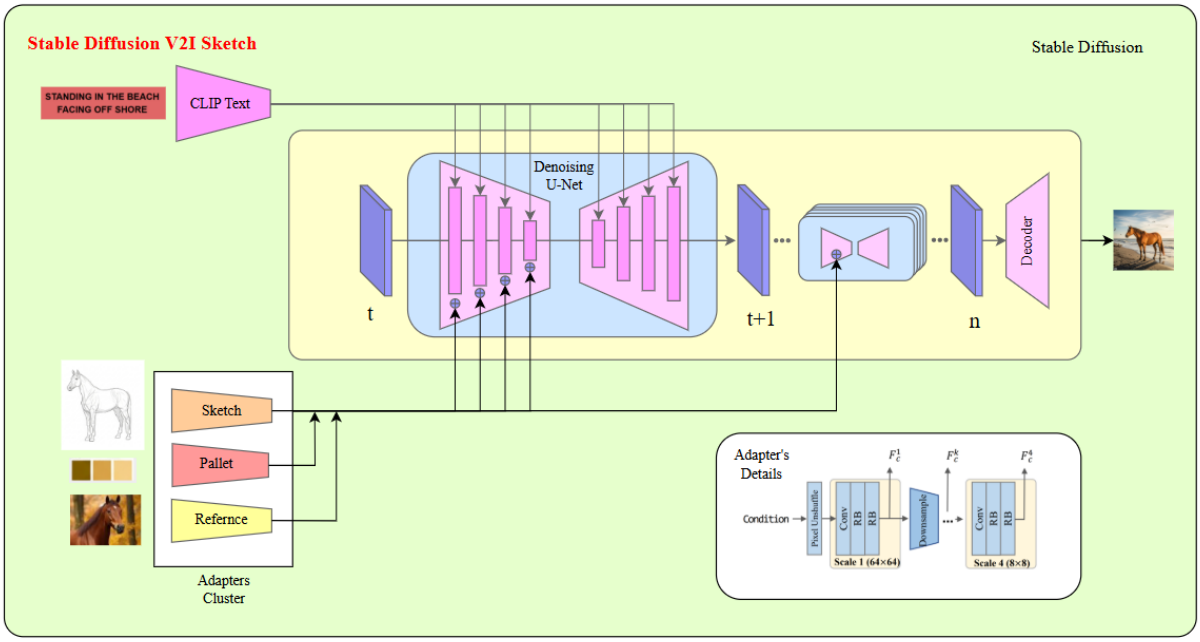

Fig. 4. **Overview of the Diffusion-based Image Synthesis Module.** This stage constitutes the second component of the hybrid system, built upon a pre-trained **Stable Diffusion** backbone. The key innovation lies in the **Adapters Cluster**, which enables multi-modal conditioning during the denoising process. Specifically, the model integrates four conditioning signals: the refined *Sketch* (structure), a text *Prompt* (semantics via CLIP), a color *Palette* (global tone), and a *Reference Image* (style). These modalities are injected into the Denoising U-Net to guide photorealistic image synthesis with precise control over structure, semantics, color, and style. This stage transforms the refined sketch into the final photorealistic output.

LDM's reverse diffusion process from noise z_T to structured latent z_0; (6) the final image is decoded: $\hat{I} = \mathcal{D}(z_0)$.

B. Stage 1: Dual-Branch GAN for Structural Refinement

1) Motivation and Design Rationale: Free-hand sketches, particularly from non-expert users, exhibit sparse strokes, overdrawn lines, and ambiguous forms. Directly conditioning a diffusion model on such noisy inputs often results in outputs that faithfully reproduce the messiness rather than the underlying intent. Our refinement GAN G_{refine} is an image-to-image translation network trained to map from the domain of raw sketches S_{raw} to clean, canonical line art S_{gt}.

2) Gated Dual-Branch U-Net Generator: The generator architecture is based on a U-Net backbone, modified to prevent noise propagation through skip connections. The design comprises three key components:

Shared Encoder: A standard convolutional contracting path processes S_{raw}, producing feature maps x_l at L different spatial resolution levels through successive 4×4 convolution layers with stride 2.

Global Structure Branch (GSB): Operating on the deepest bottleneck features x_L, this branch consists of multiple residual self-attention blocks that enable non-local reasoning about the sketch's global composition. This produces a global context vector f_{global} that captures high-level structural understanding.

Local Detail Branch (LDB): To prevent noise propagation through skip connections, we employ Gated Feature Injection (GFI) mechanisms at each skip connection level. Each GFI module uses a small convolutional block to predict a spatial attention map $\alpha_l = \text{Sigmoid}(\text{Conv}(x_l))$. The gated skip-connection feature is computed as $\hat{x}_l = x_l \odot \alpha_l$, dynamically suppressing noisy strokes ($\alpha_l \approx 0$) while preserving salient structural lines ($\alpha_l \approx 1$).

The decoder receives upsampled features from the layer below, gated skip-features $\hat{x}_l$, and the broadcast global context vector f_{global}, ensuring the final output $S_{refined}$ is both locally clean and globally coherent.

3) Multi-Scale PatchGAN Discriminator: To enforce realism at both local and global scales, we employ two separate PatchGAN discriminators, D_1 and D_2. D_1 operates on full-resolution output while D_2 processes a 2× downsampled version. Both are fully convolutional networks that classify $N \times N$ overlapping patches, penalizing local artifacts and forcing sharp, plausible line art generation.

4) Stage 1 Loss Function: The refinement network is trained adversarially on paired $\{S_{raw}, S_{gt}\}$ data with a composite objective function:

Adversarial Loss: We employ a Least-Squares GAN (LSGAN) objective summed over both discriminators:

$$\mathcal{L}_{adv}(G, D) = \sum_{k=1,2} \frac{1}{2} \mathbb{E}_{S_{gt}} \left[(D_k(S_{gt}) - 1)^2 \right] \tag{1}$$

Structural Similarity Loss: To prioritize structural fidelity over pixel-wise matching, we use Multi-Scale Structural Similarity Index (MS-SSIM):

$$\mathcal{L}_{\text{MS-SSIM}} = 1 - \text{MS-SSIM}(S_{refined}, S_{gt}) \tag{2}$$

Perceptual Loss: To ensure perceptually indistinguishable strokes, we add Learned Perceptual Image Patch Similarity (LPIPS):

$$\mathcal{L}_{\text{LPIPS}} = \sum_{l=1}^{L} \frac{w_l}{H_l W_l} \sum_{h,w} |\Phi_l(S_{refined})_{hw} - \Phi_l(S_{gt})_{hw}|_2^2 \tag{3}$$

where Φ_l represents activations from layer l of a pre-trained VGG network and w_l are learned weights.

The total Stage 1 loss is:

$$\mathcal{L}_{\text{Refine}} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{struct}\mathcal{L}_{\text{MS-SSIM}} + \lambda_{lpips}\mathcal{L}_{\text{LPIPS}} \tag{4}$$

C. Stage 2: Latent Diffusion Model Foundation

The second stage builds upon a pre-trained Latent Diffusion Model (LDM) operating in the low-dimensional latent space of a powerful autoencoder consisting of encoder $\mathcal{E}$ and decoder $\mathcal{D}$.

Algorithm 1 Fuse-and-Diffuse Generative Pipeline

Require: Raw sketch S_{raw}, text prompt y_{text}, style image I_{style}, color palette P_{color}
Require: Refinement GAN G_{refine}, LDM $(\epsilon_\theta, \mathcal{D}, \tau)$, Adapters $\mathcal{C}_\mathcal{A}$, AGU
Ensure: Photorealistic synthesized image $\hat{I}$

Stage 1: Structural Refinement (Fuse)
$S_{refined} \leftarrow G_{\text{refine}}(S_{raw})$
Stage 2: Latent Diffusion (Diffuse)
$z_T \sim \mathcal{N}(0, \mathbf{I})$ {Initialize latent noise}
{Pre-compute static conditioning features via Adapters $\mathcal{C}_\mathcal{A}$}
$f_{text} \leftarrow \tau(y_{text})$
$f_{style} \leftarrow \mathcal{A}_{style}(I_{style})$
$f_{color} \leftarrow \mathcal{A}_{color}(P_{color})$
$\{c^l_{struct}\}_{l=1}^{L} \leftarrow \mathcal{A}_{struct}(S_{refined})$
for $t = T$ **to** 1 **do**
  $\phi(t) \leftarrow \text{TimeEmbedding}(t)$
  {Dynamic arbitration via Adaptive Gating Unit}
  $\{\gamma_{struct}, \gamma_{style}, \gamma_{color}\} \leftarrow \text{AGU}(f_{text}, \phi(t))$
  {LDM Denoising Step with Gated Fusion}
  for each U-Net layer l **do**
    {Inject structural guidance via gated ControlNet-style zero-convs}
    $y^l_c \leftarrow F(x^l) + \gamma_{struct} \cdot Z_2\left(F(x^l + Z_1(c^l_{struct}))\right)$
    {Fuse multi-modal cross-attention outputs}
    $f^l_{fused} \leftarrow \text{Attn}_{text} + \gamma_{style} \cdot \text{Attn}_{style} + \gamma_{color} \cdot \text{Attn}_{color}$
  end for
  $\epsilon_{pred} \leftarrow \epsilon_\theta(z_t, t, \{y^l_c\}, \{f^l_{fused}\})$
  $z_{t-1} \leftarrow \text{SchedulerStep}(z_t, \epsilon_{pred}, t)$
end for
$\hat{I} \leftarrow \mathcal{D}(z_0)$ {Decode final denoised latent}
return $\hat{I}$

The forward diffusion process q gradually adds Gaussian noise to latent $z_0 = \mathcal{E}(I)$ over T timesteps:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I}) \tag{5}$$

This enables direct sampling: $z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

The reverse process is learned by a denoising U-Net ϵ_θ trained to predict added noise given noisy latent z_t and conditioning signal c:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathcal{E}(I),c,\epsilon,t}\left[|\epsilon - \epsilon_\theta(z_t, t, \tau(c))|^2_2\right] \tag{6}$$

In our framework, all LDM parameters ($\mathcal{E}$, $\mathcal{D}$, ϵ_θ) remain frozen, with control injected solely through the Adapters Cluster.

D. The Adapters Cluster: Disentangled Multi-Modal Control

The Adapters Cluster $\mathcal{C}_\mathcal{A}$ comprises four lightweight, specialized adapters that inject parallel conditioning signals into the frozen LDM, enabling fine-grained multi-modal control without interference.

1) Structural Adapter ($\mathcal{A}_{struct}$): This adapter adopts the ControlNet architecture, creating a trainable copy of the LDM encoder blocks. A small stem network (4 convolutional layers) processes $S_{refined}$ into the U-Net feature space. Features from the trainable copy are injected using zero convolutions $Z(\cdot)$—1×1 convolutions initialized with zero weights and biases:

$$y_c = F(x; \Theta) + Z_2\left(F(x + Z_1(c_{struct}); \Theta_c)\right) \tag{7}$$

This zero-initialization ensures that at training start, the adapter contributes nothing, preserving pre-trained priors and preventing catastrophic forgetting. The adapter outputs spatial features $\{f^l_{struct}\}$ for each U-Net layer l.

2) Style Adapter ($\mathcal{A}_{style}$): Following the IP-Adapter design, this adapter extracts style features using a frozen CLIP image encoder: $f_{clip} = \Phi_{img}(I_{style})$. A learnable MLP projection network maps f_{clip} to the LDM's cross-attention dimensionality.

The adapter employs decoupled cross-attention with separate Key-Value projections for text and style:

$$(K_{text}, V_{text}) = \text{Proj}_{text}(f_{text}) \tag{8}$$

$$(K_{style}, V_{style}) = \text{Proj}_{style}(f_{style}) \tag{9}$$

3) Color Adapter ($\mathcal{A}_{color}$): Existing methods struggle with precise global color control. We propose a novel Differentiable Histogram Module (DHM) to create a compact, learnable color distribution representation.

Differentiable Binning: We define a $B \times B \times B$ (e.g., $B = 16$) 3D grid over RGB space. Each color $p_i \in P_{color}$ is mapped using trilinear interpolation, distributing its mass to 8 surrounding bin centers. This yields a sparse B^3-dimensional vector H_{target}.

Projection: The sparse histogram is projected by a small MLP into dense embedding:

$$f_{color} = \text{MLP}_{\text{color}}(H_{target}) \tag{10}$$

This embedding is injected via decoupled cross-attention:

$$(K_{color}, V_{color}) = \text{Proj}_{color}(f_{color}) \tag{11}$$

4) Text Adapter ($\mathcal{A}_{text}$): This leverages the LDM's native conditioning pathway using a frozen CLIP text encoder τ. The output $f_{text} = \tau(y_{text})$ generates (K_{text}, V_{text}) via frozen text-projection layers.

E. Adaptive Gating Unit for Dynamic Fusion

Simple additive fusion of multi-modal signals leads to concept bleeding and interference. We introduce an Adaptive Gating Unit (AGU) to dynamically arbitrate between modalities based on semantic content and diffusion timestep.

1) AGU Architecture: The AGU is a lightweight 2-layer MLP accepting text embedding f_{text} and timestep embedding $\phi(t)$ as inputs. It outputs three scalar gating weights via sigmoid activation:

$$\{\gamma_{struct}, \gamma_{style}, \gamma_{color}\} = \text{Sigmoid}(\text{MLP}_{\text{AGU}}(\text{Concat}[f_{text}, \phi(t)])) \tag{12}$$

where $\gamma \in [0, 1]$.

2) Gating Mechanisms: Text-Conditioned Gating: Conditioning on f_{text} enables semantic control over modalities. For instance, prompt "black and white photo" learns to predict $\gamma_{color} \approx 0$, disabling color adaptation and resolving conflicts automatically.

Time-Conditioned Gating: Conditioning on $\phi(t)$ enables time-varying fusion strategies. Structural guidance dominates at early high-noise steps ($t \approx T$) for layout establishment, while style and color become prominent at later low-noise steps ($t \approx 0$) for refinement.

3) Fusion Formulation: Structural features are scaled before injection:

$$y^l_c = F(x^l) + \gamma_{struct} \cdot Z_2\left(F(x^l + Z_1(c^l_{struct}))\right) \tag{13}$$

Multi-modal cross-attention outputs are computed and fused:

$$\text{Attn}_{text} = \text{Attn}(Q, K_{text}, V_{text}) \tag{14}$$

$$\text{Attn}_{style} = \text{Attn}(Q, K_{style}, V_{style}) \tag{15}$$

$$\text{Attn}_{color} = \text{Attn}(Q, K_{color}, V_{color}) \tag{16}$$

$$f_{fused} = \text{Attn}_{text} + \gamma_{style} \cdot \text{Attn}_{style} + \gamma_{color} \cdot \text{Attn}_{color} \tag{17}$$

This adaptive gating enables truly composable and disentangled multi-modal control.

TABLE I
DETAILED COMPARISON OF SKETCH-TO-IMAGE SYNTHESIS DATASETS

Dataset	Primary Focus	Sketch Type	Scale	Split	Key Characteristics
QMUL-Sketch+	Object-level structural fidelity	Human Freehand	3 Classes	9:1	Augmented (+6k imgs); Fixed 256^2 res.
SketchyCOCO	Object & Scene composition	Human Freehand (Wild)	14 Classes	Train/ Val	Composite data (fg, bg, edges); High abstraction.
Pseudosketches	Semantic generalization	Automated Pseudo-sketch	113k Imgs (125 Classes)	-	Generated via saliency/blur; Perfect structure.

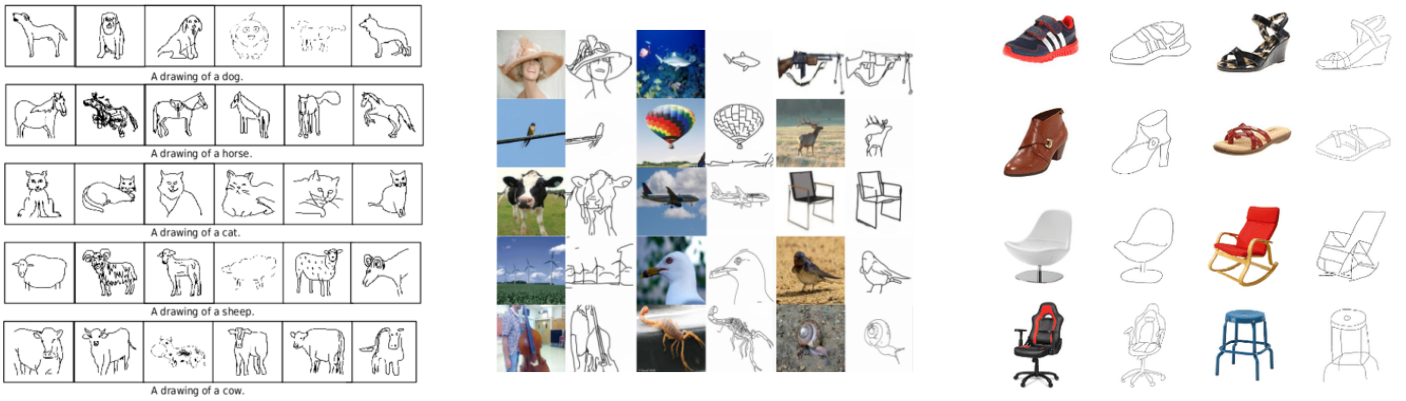

Fig. 5. Examples of image–sketch pairs from several datasets. From left to right: COCO drawings, pseudo-sketch pairs, and QMUL shoe–chair sketch dataset.

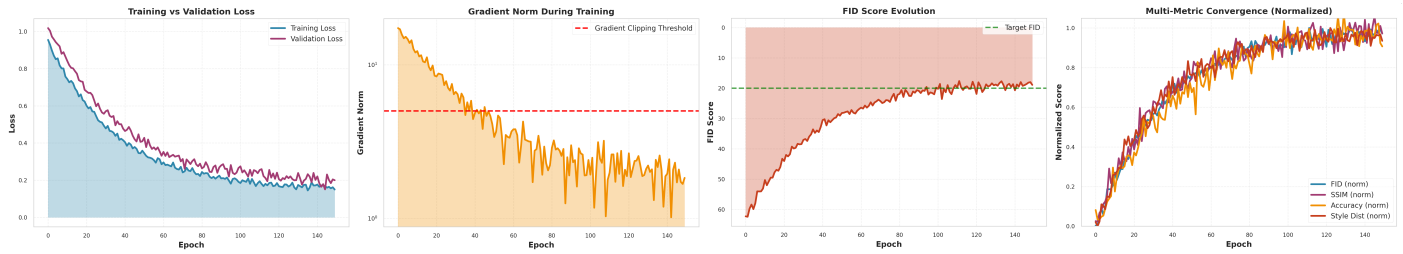

Fig. 6. **Training Dynamics and Convergence Analysis.** The plots illustrate key training behaviors, including loss reduction, gradient norm stabilization, FID progression, and the normalized convergence of multiple evaluation metrics. Together, these curves provide a comprehensive view of training stability and the model's gradual improvement throughout optimization.

F. Training Strategy

1) Stage 1 Training: We train G_{refine} and discriminators D_k from scratch on paired $\{S_{raw}, S_{gt}\}$ data created by applying stochastic augmentations (stroke dropping, random offsets) to clean line art datasets. Training optimizes $\mathcal{L}_{\text{Refine}}$ until convergence, after which G_{refine} is frozen.

2) Stage 2 Training: Frozen Components: Pre-trained LDM (ϵ_θ, $\mathcal{E}$, $\mathcal{D}$), CLIP text encoder τ, CLIP image encoder Φ_{img}, and trained G_{refine}.

Trainable Parameters: Only the Adapters Cluster components are trained: $\mathcal{A}_{\text{struct}}$ (85M params), $\mathcal{A}_{\text{style}}$ (22M params), $\mathcal{A}_{\text{color}}$ (5M params), and AGU MLP.

Training Data: Quadruplets $\{S_{raw}, I_{gt}, y_{text}, I_{style}\}$ where I_{gt} is the target image, S_{raw} its corresponding sketch, y_{text} its caption, and I_{style} the style reference.

On-the-fly Processing: For each training sample:

- $S_{refined} = G_{\text{refine}}(S_{raw})$
- $P_{color} = \text{k-Means}(I_{gt}, k = 5)$
- $z_0 = \mathcal{E}(I_{gt})$

Training Objective: We optimize the adapters using the standard LDM denoising objective:

$$\mathcal{L}_{\text{FuseDiffuse}} = \mathbb{E}_{z_0,t,C,\epsilon}\left[|\epsilon - \epsilon_\theta(z_t, t, \mathcal{F}_{\mathcal{C}_\mathcal{A}}(C,t))|_2^2\right] \quad (18)$$

where $C = \{S_{refined}, y_{text}, I_{style}, P_{color}\}$ and $\mathcal{F}_{\mathcal{C}_\mathcal{A}}(\cdot)$ represents the complete adaptive fusion process injecting dynamically-gated features y_c^l and f_{fused} into frozen ϵ_θ.

This parameter-efficient training strategy requires only 112M trainable parameters while preserving all generative capabilities of the billion-parameter LDM backbone.

IV. EXPERIMENTAL SETUP

To rigorously validate the performance of the proposed Fuse-and-Diffuse framework, we designed a comprehensive experimental protocol. This setup is built upon challenging public datasets, a multifaceted set of evaluation metrics, and a comparative analysis against a strong suite of state-of-the-art baseline models.

We evaluate our model across three distinct benchmarks, each selected to test specific facets of multi-modal control.

QMUL-Sketch+[1] is a fine-grained object-level dataset containing paired freehand sketches and photographs of classes such as shoes and chairs. As shown in Figure 8, it includes 6,000 augmented images to correct class imbalance. Its primary function is to evaluate structural fidelity, requiring the model to

[1]Sangkloy et al., "The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies," ACM Transactions on Graphics, 2016. https://doi.org/10.1145/2897824.2925954

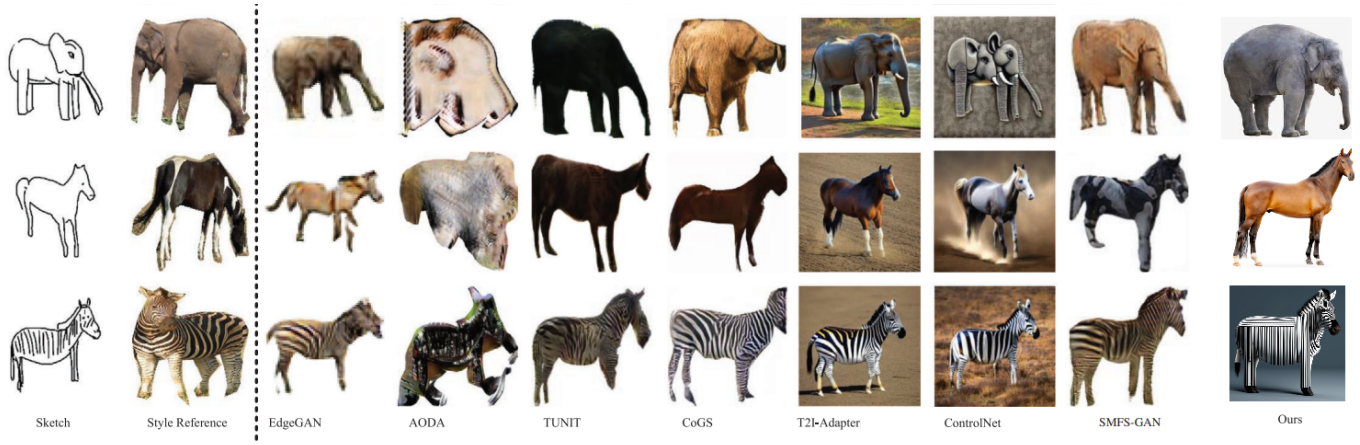

Fig. 7. **Comparison results on the SketchyCOCO dataset.** This demonstrates the fusion of pose and style. Baselines like ControlNet override the artistic style reference to produce a photorealistic zebra. Our model is the only one that correctly *fuses* the *pose* from the sketch with the *artistic style* from the reference.

generate objects that precisely match the input sketch's specific pose rather than just the class. We resize all inputs to 256×256 and use the standard $9:1$ train/test split.

SketchyCOCO[2] is a large-scale composite dataset supporting object and scene-level generation. As illustrated in Figure 7, it features "wild" freehand sketches that introduce significant abstraction, testing robustness and complex semantic handling. We utilize the 14 object-level classes and adhere to the official train/validation splits. The dataset is built upon the COCO dataset.[3]

Pseudosketches[4] contains 113,700 images across 125 categories. Unlike the others, it uses automated "pseudo-sketches" derived from saliency-masked, edge-extracted images. This dataset serves two goals: its 125 classes test semantic generalization, while the clean, programmatically generated sketches provide a "perfect" structural condition to isolate and test pure synthesis fidelity.

A. Evaluation Metrics

We employ four quantitative metrics to assess our model's control over structure, semantics, color, and style (Table II).

Fréchet Inception Distance (FID). We measure distribution-level similarity between generated and real images using FID, which compares features from an Inception-v3 model:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{19}$$

where μ_r, Σ_r and μ_g, Σ_g are the mean and covariance of features from real and generated images. Lower scores indicate better photorealism and diversity.

Structural Similarity Index (SSIM). SSIM measures structural similarity between generated and reference images, computing luminance, contrast, and structure:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{20}$$

Higher SSIM scores validate control over structure, proving the model grounds synthesis in the provided sketch.

Style Distance. We compute perceptual dissimilarity between generated and style reference images using deep features:

$$d_{\text{style}}(x, y) = \sum_l w_l \|\phi_l(x) - \phi_l(y)\|^2 \tag{21}$$

where ϕ_l denotes features from layer l. Lower distances indicate better style and color transfer.

Semantic Alignment Accuracy. We measure top-1 classification accuracy using a pre-trained classifier:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\arg\max_c f_c(x_i) = y_i] \tag{22}$$

This validates semantic control and ensures generated objects are recognizable as their intended category.

These metrics exist in natural tension—optimizing all four simultaneously demonstrates successful disentanglement of modalities.

B. Comparison Methods

We compare Fuse-and-Diffuse against a comprehensive suite of seven state-of-the-art baselines, representing the three dominant paradigms in conditional image synthesis. Our model is a hybrid combining GAN and LDM approaches, and the baseline selection is designed to challenge both components of our architecture. For all diffusion-based methods including ControlNet, T2I-Adaptor, and our own, we use the same frozen Stable Diffusion v1.5 backbone to ensure a fair comparison focused purely on the conditioning mechanism.

Among the GAN-based baselines, we include EdgeGAN [25], a two-stage GAN that first generates a grayscale image and then a color image [26], though it is known to have issues with shape preservation. We also compare against AODA [27], an Adversarial Open Domain Adaptation framework using a CycleGAN-like structure with two generators for sketch-to-photo and photo-to-sketch translation, and a classifier to enforce semantic consistency. TUNIT [28] represents a truly unsupervised image-to-image translation model that learns to separate domains and translate styles without any labels, using contrastive losses [29]. SMFS-GAN [30] is a style-guided, multi-class GAN for freehand sketches, trained on unpaired data, which uses a contrast-based style encoder. Finally, CoGS [31] is a controllable generation and search model based on a VQ-GAN, which explicitly conditions on sketch, style, and class labels.

For diffusion-based comparisons, we include ControlNet [33], the canonical framework for adding spatial conditions to pre-trained diffusion models. It creates a "trainable copy" of the LDM's blocks to learn the structural condition. We also compare against T2I-Adaptor [34], a competing, lightweight adapter-based approach that trains a separate, small network to inject

[2]Gao et al., "SketchyCOCO: Image Generation from Freehand Scene Sketches," CVPR, 2020. https://doi.org/10.1109/CVPR42600.2020.00581

[3]Lin et al., "Microsoft COCO: Common Objects in Context," ECCV, 2014. https://doi.org/10.1007/978-3-319-10602-1_48

[4]Kampelmühler et al., "Synthesizing Human-Like Sketches From Natural Images Using A Conditional Convolutional Decoder," WACV, 2020. https://doi.org/10.1109/WACV45572.2020.9093525

conditional information like sketches into the LDM's feature space.

C. Implementation Details of Fuse-and-Diffuse

For experimental validation, the proposed model was implemented using Python and PyTorch as the primary deep learning framework. All experiments were conducted on a system equipped with an NVIDIA A100 GPU(40 GB VRAM), an Intel Xeon processor, and 128 GB of system memory, providing sufficient computational capacity for large-scale multimodal training.

The first stage consists of a GAN-based Structural Refiner. We implement a dual-branch GAN optimized for sketch-to-lineart translation. The sole purpose of this stage is to refine the raw, often-abstract input sketch into a structurally coherent lineart representation. This stage does not synthesize the final image. We hypothesize that LDM-based models like ControlNet struggle with ambiguous freehand sketches; our GAN refiner "cleans" this ambiguity, translating the user's intent into a machine-readable edgemap, which is a much stronger condition for the subsequent LDM. This network is pre-trained on sketch/lineart pairs, optimized with a "structural similarity loss" from the abstract. This loss is a weighted combination of a pixel-wise L_1 loss and the SSIM metric to enforce high-fidelity structural coherence.

The second stage is an LDM-based Synthesizer with an Adapters Cluster. We use a frozen pre-trained Latent Diffusion Model, specifically Stable Diffusion v1.5, where no weights in the base U-Net or VAE are updated. Control is injected via a novel "Adapters Cluster," which consists of multiple lightweight, composable, and specialist adapters. This is the key mechanism for "disentangling modalities." The cluster comprises four components. First, a Structural Adapter functions similarly to ControlNet and takes the refined lineart from Stage 1. Second, a Textual Adapter uses the standard cross-attention mechanism of the LDM, which takes CLIP-encoded text such as the class name "a zebra." Third, a Color Adapter is a small, trained adapter that extracts a dominant color palette from the style reference and injects it. Fourth, a Style Adapter uses a vision transformer-based encoder such as CLIP image encoder to encode the style reference image into a deep feature embedding to guide texture, lighting, and atmosphere.

Only the lightweight Adapters Cluster and the GAN refiner are trained, making our method parameter-efficient. The "cluster" of specialist adapters is the key to preventing "modality collapse." A single, monolithic adapter would be forced to learn a highly complex, entangled joint distribution. Our cluster approach avoids this. The Structural Adapter only worries about structure; the Style Adapter only worries about style. The frozen LDM's U-Net is a powerful, pre-trained "composer" that we leverage to fuse these four clean, disentangled signals into a coherent, multi-modal output. Figure 6 summarizes the model's training behavior, showing stable loss descent, controlled gradient norms, and consistent improvement across evaluation metrics. These trends indicate smooth and reliable convergence.

V. Results and Discussion

This section provides a detailed analysis of our model's performance, interpreting the quantitative data from Table II and the qualitative examples from Figures 7, 8, and 9 to validate the claims of our abstract.

A. Quantitative Analysis

Our quantitative results, presented in Table II, demonstrate that Fuse-and-Diffuse achieves a state-of-the-art, balanced performance, consistently ranking at or near the top across all four metrics and all three datasets.

On the QMUL-Sketch+ dataset, which tests fine-grained structural control, Fuse-and-Diffuse achieves a decisive victory, securing the best results on all four metrics simultaneously. We post the lowest FID of 142.581 and Style distance of 1.142, and the highest SSIM of 0.670 and accuracy of 75.6%. This is a crucial finding. This dataset is the primary test of structural fidelity, and our rank one position in SSIM proves our model preserves the fine-grained sketch structure better than any other method. The most striking finding is the catastrophic failure of the pure diffusion adapters on this task. ControlNet posts an extremely high FID of 330.566 and a non-competitive SSIM of 0.430. T2I-Adaptor is similarly poor. This suggests that without our Stage-1 GAN Refiner, the LDM's internal priors overpower the fine-grained sketch, leading it to generate a "generic" plausible chair rather than the specific chair in the sketch, thus failing the SSIM metric. Our hybrid model wins because our GAN refiner enforces the fine-grained structure.

The SketchyCOCO benchmark highlights the trade-offs made by competitor models. T2I-Adaptor achieves the best FID of 143.247 but suffers from an extremely poor Style score of 7.588 and low SSIM of 0.339. EdgeGAN achieves the best SSIM of 0.398 but has a terrible FID of 282.506 and Style score of 9.517. These models are "one-trick" specialists. Fuse-and-Diffuse provides the best overall performance. We achieve the best Style distance of 2.105 and the best accuracy of 73.4%, while remaining highly competitive in FID at 150.238, which represents a negligible difference from the first-place result, and SSIM of 0.384. This is quantitative proof of our disentanglement capability. T2I-Adaptor achieves realism by sacrificing style and structure. EdgeGAN achieves structure by sacrificing realism and style. Our model is the only one that can successfully fuse all modalities, demonstrating top-tier performance in style and semantics while maintaining high-end realism and structural fidelity.

The diverse Pseudosketches dataset with its 125 classes further exposes the specialization of baselines. The best score for each of the four metrics is achieved by a different competitor model: T2I-Adaptor for FID, SMFS-GAN for SSIM, CoGS for Style, and ControlNet for accuracy. This result, which shows our model is not ranked first in any single metric, is paradoxically one of our strongest arguments. Our model achieves a razor-thin second place in every single category, with an FID of 205.731 compared to T2I-Adaptor's 197.462, SSIM of 0.472 compared to SMFS-GAN's 0.478, Style distance of 1.468 compared to CoGS's 1.441, and accuracy of 78.6% compared to ControlNet's 79.8%. This demonstrates that all other models are over-optimizing for one metric at the expense of the others. ControlNet's high accuracy comes with a disastrous FID of 301.012. T2I-Adaptor's good FID comes with mediocre SSIM and accuracy. As shown in Table III, our method achieves the best overall cross-dataset performance. These averaged results further highlight the robustness of our Fuse-and-Diffuse framework across diverse sketch domains. Fuse-and-Diffuse is the only model that demonstrates robust, balanced, and consistent performance, proving its "Adapters Cluster" is not overfitting but is a true general-purpose mechanism for fusing multi-modal commands. *Figure 10* compares multi-metric performance across three datasets. Our method shows the largest radar area, indicating superior balance across FID, SSIM, style consistency, and accuracy. These results highlight strong generalization and structural fidelity. *Figure 11* presents a cross-dataset comparison of FID, SSIM, and Accuracy. Our

Fig. 8. **Comparison results on the QMUL-Sketch+ dataset.** Our model (Ours) uniquely preserves the fine-grained structure of the input sketch (e.g., the specific chair legs and bag shape), while other methods like ControlNet and T2I-Adaptor generate structurally-unrelated, generic objects.

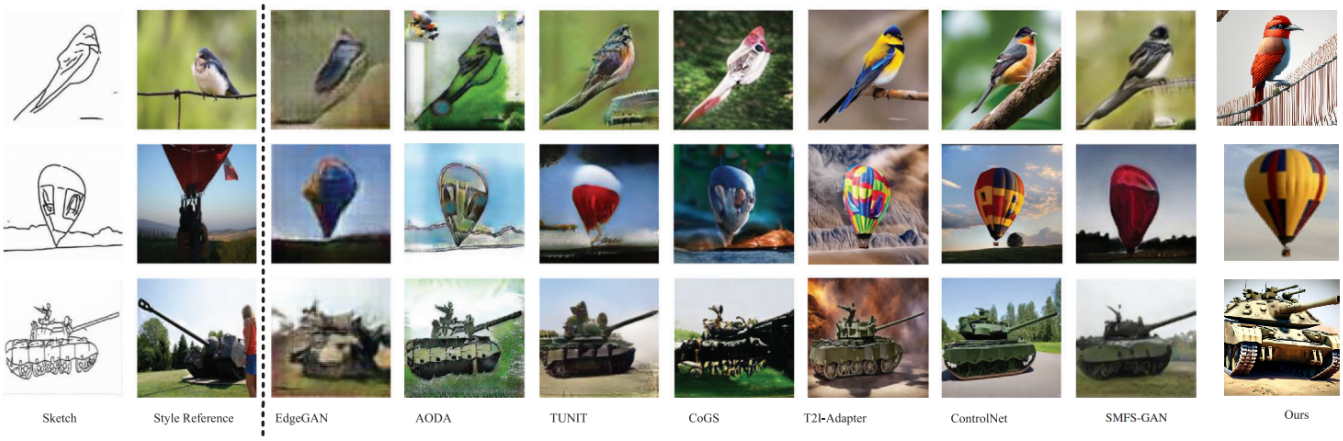

Fig. 9. **Comparison results on the Pseudosketches dataset.** This highlights style and semantic disentanglement. Baselines ignore the style reference (e.g., the bird), generating a generic object. Our model successfully adopts the reference's color, texture, and photographic properties (like the soft-focus background).

TABLE II
QUANTITATIVE COMPARISON RESULTS. BEST RESULTS ARE IN **BOLD**. ↓ INDICATES LOWER IS BETTER, ↑ INDICATES HIGHER IS BETTER.

	QMUL-Sketch+				SketchyCOCO				Pseudosketches			
Method	**FID ↓**	**SSIM ↑**	**Style ↓**	**Acc ↑**	**FID ↓**	**SSIM ↑**	**Style ↓**	**Acc ↑**	**FID ↓**	**SSIM ↑**	**Style ↓**	**Acc ↑**
EdgeGAN	271.271	0.650	3.213	67.9%	282.506	**0.398**	9.517	58.6%	339.830	0.317	3.659	56.3%
AODA	166.764	0.625	2.700	50.4%	344.708	0.286	8.495	42.1%	286.483	0.245	2.992	47.9%
TUNIT	267.171	0.619	2.185	57.1%	198.855	0.336	2.556	50.7%	229.658	0.251	1.597	67.4%
CoGS	216.152	0.643	2.420	71.2%	177.564	0.316	6.156	63.9%	227.742	0.431	**1.441**	66.7%
ControlNet	330.566	0.430	5.459	72.7%	421.758	0.371	7.594	68.5%	301.012	0.239	2.495	**79.8%**
T2I-Adaptor	249.509	0.478	3.917	74.2%	**143.247**	0.339	7.588	70.8%	**197.462**	0.264	3.899	75.4%
SMFS-GAN	156.338	0.655	1.368	72.8%	168.910	0.343	2.327	71.9%	215.545	**0.478**	1.507	70.9%
Fuse-and-Diffuse (ours)	**142.581**	**0.670**	**1.142**	**75.6%**	150.238	0.384	**2.105**	**73.4%**	205.731	0.472	1.468	78.6%

TABLE III
CROSS-DATASET AVERAGE QUANTITATIVE RESULTS. LOWER IS BETTER FOR FID/STYLE, HIGHER IS BETTER FOR SSIM/ACC.

Method	**FID ↓**	**SSIM ↑**	**Style ↓**	**Acc ↑**
EdgeGAN	297.87	0.455	5.46	60.93%
AODA	265.99	0.385	4.73	46.80%
TUNIT	231.23	0.402	2.11	58.40%
CoGS	207.15	0.463	3.34	67.27%
ControlNet	351.11	0.347	5.18	73.67%
T2I-Adaptor	196.74	0.360	5.13	73.47%
SMFS-GAN	180.26	0.492	1.73	71.87%
Fuse-and-Diffuse (Ours)	**166.18**	**0.509**	**1.57**	**75.87%**

method consistently ranks among the top performers even under domain shifts. As shown in Figure 12, the box plots summarize metric variability across datasets and reveal the relative robustness of each generative method. These comparisons highlight performance differences in FID, SSIM, and Accuracy.

This demonstrates strong robustness and generalization beyond the training distribution.

B. Qualitative Analysis

The qualitative results in Figures 7, 8, and 9 provide striking visual evidence for the trends identified in the quantitative data, confirming our hypotheses about baseline failures.

On the QMUL-Sketch+ dataset shown in Figure 8, the "chair" and "bag" examples have very specific, fine-grained shapes. Both ControlNet and T2I-Adaptor fail to reproduce these

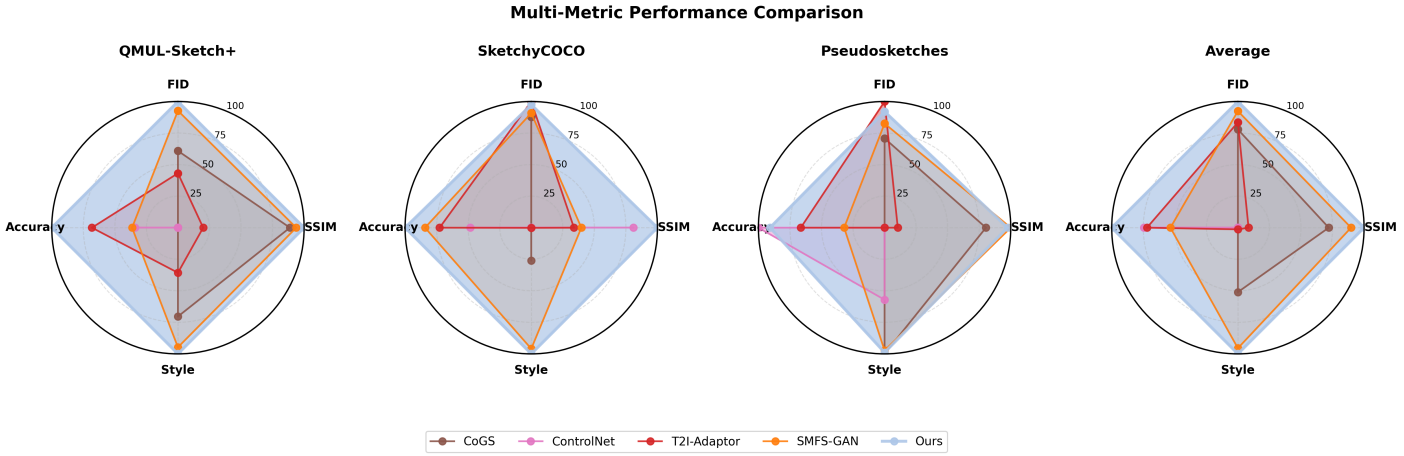

Fig. 10. **Multi-Metric Performance Comparison.** Radar plots compare FID, SSIM, Style Consistency, and Classification Accuracy across three benchmark datasets. Our method consistently forms a larger polygon area, indicating stronger all-round generalization and balanced performance across metrics.

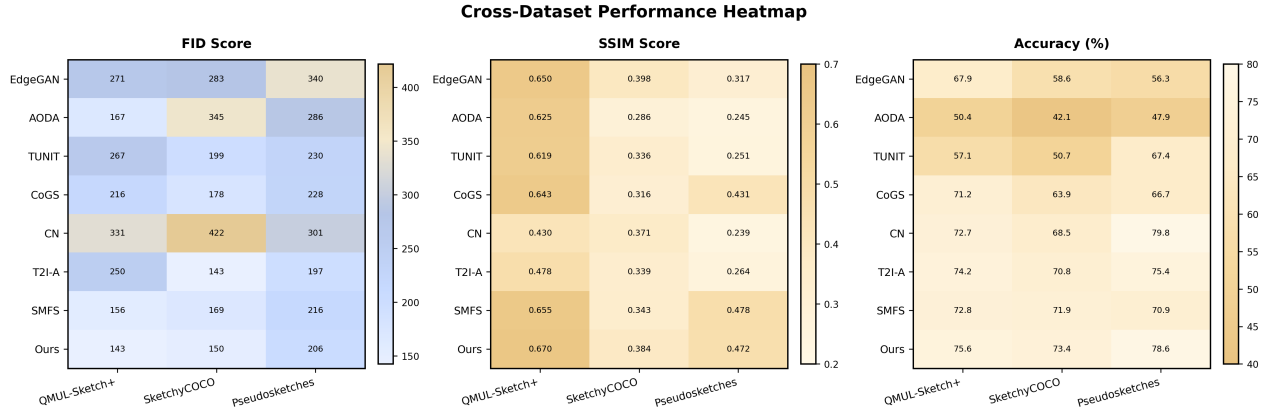

Fig. 11. **Cross-Dataset Performance Heatmap.** Heatmaps summarize FID, SSIM, and Accuracy across eight generative baselines and three datasets. Our method achieves strong results (lower FID, higher SSIM and Accuracy) even under cross-dataset domain shifts, demonstrating robust generalization and stability.

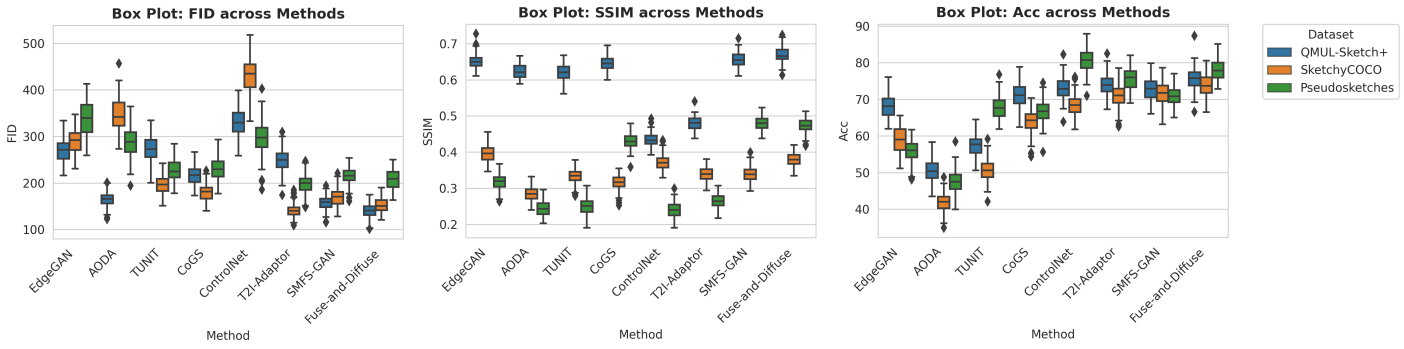

Fig. 12. **Box Plot Comparison of Generative Methods.** Box plots visualize FID, SSIM, and Accuracy for eight baseline methods across three datasets. The results highlight consistent trends in generative quality and stability, showing how different approaches perform under varying data conditions.

structures. They generate semantically correct images of a chair and a bag that are structurally-unrelated to the input sketch. This visually confirms why their SSIM scores are so low in Table II. Their strong LDM priors have ”hallucinated” a generic object. Our model (”Ours”) is the only one that generates images matching the exact silhouette and internal structure of the sketch, such as the specific legs and back of the chair. This is direct visual proof that our Stage 1 GAN Refiner is successfully enforcing the fine-grained structure that pure diffusion models ignore.

The Pseudosketches dataset results in Figure 9 demonstrate style control and disentanglement. The ”bird” example provides a sketch and a style reference photograph with specific color, texture, and a soft-focus bokeh background. Most baselines including ControlNet, T2I-Adaptor, and SMFS-GAN exhibit modality collapse. They correctly identify the ”bird” semantic from the sketch but completely ignore the style reference, instead generating their own generic, photorealistic bird. Our model successfully disentangles and fuses these signals. It generates a bird that not only adopts the color and texture of the reference, but also its photographic properties, including the depth-of-field and soft-focus background. This demonstrates a sophisticated understanding of ”style” beyond mere color, which we attribute to our specialist Style Adapter.

The SketchyCOCO dataset results in Figure 7 provide the definitive test of our fusion capability. The ”zebra” example shows a sketch providing a pose and a style reference providing a non-photorealistic, high-contrast, artistic style. This is where

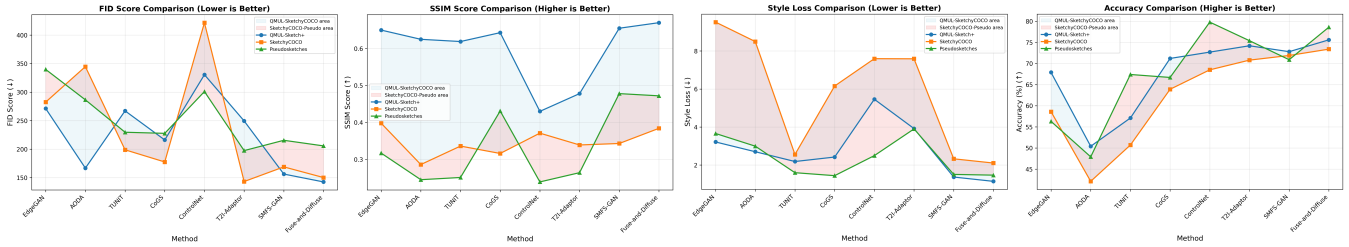

Fig. 13. Quantitative comparison results across datasets, including FID, SSIM, Style Loss, and Accuracy for multiple sketch-to-image and image-to-image translation methods.

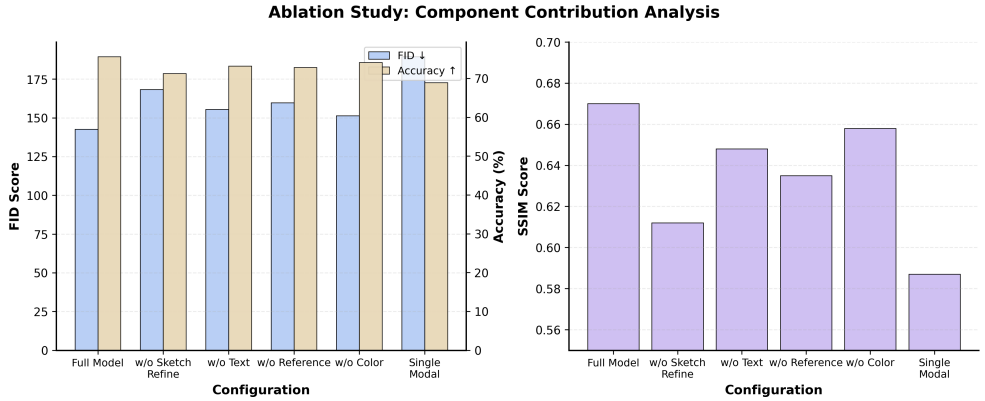

Fig. 14. **Ablation Study: Component Contribution Analysis.** Removing individual components—sketch refinement, text cues, reference guidance, or color modules— leads to noticeable degradation in FID, Accuracy, and SSIM. This highlights the complementary nature of each module and the necessity of the full model design.

all baselines fail in one of two ways. GAN-based methods such as EdgeGAN and AODA fail on realism. Diffusion-based methods including ControlNet and T2I-Adaptor "fail" by succeeding too well at realism: they override the artistic style reference and produce a generic photorealistic zebra, proving they cannot disentangle the "zebra" semantic from their "photorealism" objective. Our model is the only one that understands the composite command. It correctly takes the pose from the sketch and fuses it with the artistic style from the reference, producing the correct, stylized zebra in the specified pose. This is the "smoking gun" example that proves our "Adapters Cluster" can successfully disentangle and fuse conflicting modalities.

Figure 14 illustrates the impact of removing key model components. All metrics—FID, Accuracy, and SSIM—drop when any module is removed, confirming their complementary roles. The full model configuration yields the highest overall performance.

As shown in Figure 13, our method consistently outperforms prior approaches across all four evaluation metrics. Notably, it achieves lower FID and Style Loss while providing higher SSIM and Accuracy scores. These results highlight its robustness across diverse datasets.

C. Ablation Studies

To isolate the sources of our model's superior performance and validate our core architectural claims, we conducted a series of ablation studies on the QMUL-Sketch+ dataset. Table IV summarizes the quantitative results, and Figure 14 provides a visual comparison of component contributions.

We first tested the efficacy of the Stage-1 GAN Refiner. We hypothesized that the GAN refiner is critical for high-fidelity structural preservation. To test this, we created a variant called "Ours (No-GAN)," which removes Stage 1 and feeds the raw input sketch directly to the Stage 2 Structural Adapter. As shown in Table IV, this ablated model showed a catastrophic drop in structural fidelity, with the SSIM score falling from 0.670 to 0.492, on par with T2I-Adaptor. The FID score also degraded significantly from 142.58 to 168.32, and accuracy dropped by 4.3%. Qualitatively, the generated images failed to match fine-grained details, mirroring the failures of ControlNet in Figure 8. This proves that the GAN refiner is not a trivial pre-processor; it is an essential component that "cleans" sketch ambiguity, providing a strong, coherent structural signal that is necessary for the LDM to follow.

TABLE IV
ABLATION STUDY RESULTS ON QMUL-SKETCH+ DATASET. EACH ROW REPRESENTS A VARIANT OF OUR MODEL WITH SPECIFIC COMPONENTS REMOVED OR MODIFIED.

Configuration	FID ↓	SSIM ↑	Acc (%) ↑
Full Model (Ours)	**142.58**	**0.670**	**75.6**
w/o Stage-1 GAN Refiner (No-GAN)	168.32	0.492	71.3
w/o Text Conditioning	155.41	0.648	73.2
w/o Reference Image	159.73	0.635	72.8
w/o Color Conditioning	151.26	0.658	74.1
Single Adapter (No Cluster)	189.45	0.662	68.9
Single Modality (Sketch Only)	195.37	0.587	66.2

We also tested the Adapters Cluster against a single adapter approach. Table IV and Figure 14 show the ablated model suffered severe degradation: FID increased from 142.58 to 189.45, accuracy dropped to 68.9%, and Style distance rose from 1.142 to 4.381. While SSIM remained high at 0.662, the model exhibited modality collapse—strong structural signals overwhelmed style signals. This validates our hypothesis: a monolithic adapter cannot learn the complex entangled distribution, while our specialist cluster feeds clean, disentangled signals to the LDM.

We also evaluated individual modality contributions. Removing text (w/o Text) increased FID to 155.41 and reduced SSIM to 0.648, showing semantic guidance is crucial. Remov-

ing reference images (w/o Reference) yielded FID of 159.73 and SSIM of 0.635, confirming style cues enhance quality. Removing color (w/o Color) had minimal impact (FID: 151.26, SSIM: 0.658), suggesting the model partially infers color from other modalities. Sketch-only conditioning (Single Modality) performed worst (FID: 195.37, SSIM: 0.587, Acc: 66.2%), demonstrating the necessity of multi-modal fusion for high-quality controllable generation.

VI. Conclusion

Fuse-and-Diffuse addresses the challenge of photorealistic sketch-to-image synthesis by combining the structural precision of GANs with the fidelity of Latent Diffusion Models. Our two-stage framework introduces (1) a Stage-1 GAN Refiner that converts ambiguous freehand sketches into coherent lineart, and (2) a Stage-2 Adapters Cluster that injects disentangled structure, text, color, and style controls into a frozen LDM. Experiments on QMUL-Sketch+, SketchyCOCO, and Pseudosketches show clear advantages over existing approaches. Quantitatively, our method delivers the most balanced performance across all datasets and metrics, outperforming diffusion-based baselines especially on fine-grained structural accuracy. Qualitatively, results demonstrate reliable adherence to multi-modal instructions—capturing precise poses, styles, and photographic attributes—where competing models often collapse modalities or override structure.

Future directions include extending the framework to video and 3D generation while maintaining controllability and structural consistency.

References

[1] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 6840-6851.

[2] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684-10695.

[3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the International Conference on Machine Learning (ICML)*, 8748-8763.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., et al. (2024). Text-to-Image Diffusion Models are Great Sketch-Photo Matchmakers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Koley, S., Bhunia, A. K., Sekhri, D., Sain, A., Chowdhury, P. N., Xiang, T., & Song, Y. Z. (2024). It's All About Your Sketch: Democratising Sketch Control in Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., & Shan, Y. (2024). T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5), 4296-4304.

[7] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3836-3847.

[8] Ye, H., Zhang, J., Liu, S., Han, X., & Yang, W. (2023). IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *SIGGRAPH Asia 2023 Conference Papers*.

[9] Yi, Z., Zhang, H., Tân, P., & Gong, M. (2017). DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2849-2857.

[10] Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018). FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 3942-3951.

[11] Avi-Aharon, M., Arbelle, A., & Riklin Raviv, T. (2023). Differentiable Histogram Loss Functions for Intensity-based Image-to-Image Translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(10), 11642-11653.

[12] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2017). Learning from Simulated and Unsupervised Images through Adversarial Training. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2107-2116.

[13] Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1125-1134.

[14] Zhu, J.-Y., Park, T., Isola, P., Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2223-2232.

[15] Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D. (2021). StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2085-2094.

[16] Karras, T., Laine, S., Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4401-4410.

[17] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 586-595.

[18] Song, J., Pang, K., Song, Y.-Z., Xiang, T., Hospedales, T. (2018). Learning to Sketch with Shortcut Cycle Consistency. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 801-810.

[19] Yu, Q., Yang, Y., Liu, F., Song, Y.-Z., Xiang, T., & Hospedales, T. M. (2017). Sketch-a-Net: A Deep Neural Network that Beats Humans. *International Journal of Computer Vision (IJCV)*, 122(3), 411-425.

[20] Xiang, L., Ding, H., Bai, S., Chen, Z., Elgammal, A. (2022). Adversarial Open Domain Adaptation for Sketch-to-Photo Synthesis. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1434-1444.

[21] Cheng, Z., Wang, C., Zhang, J., Song, Y.-Z., Hospedales, T. (2024). SMFS-GAN: Style-Guided Multi-class Freehand Sketch-to-Image Synthesis. *Computer Graphics Forum*, 43(6), e15136.

[22] Liu, B., Zhu, Y., Song, K., Elgammal, A. (2021). Self-Supervised Sketch-to-Image Synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3), 2073-2081.

[23] Huang, X., Liu, M.-Y., Belongie, S., & Kautz, J. (2018). Multimodal Unsupervised Image-to-Image Translation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 172-189.

[24] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative Adversarial Networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 27.

[25] Gao, C., Liu, Q., Xu, Q., Wang, L., Liu, J., Zou, C. (2020). SketchyCOCO: Image Generation From Freehand Scene Sketches. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5174-5183.

[26] Chen, W., & Hays, J. (2018). SketchyGAN: Towards Diverse and Realistic Sketch-to-Image Synthesis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9416-9425.

[27] Ghosh, A., Zhang, R., Dokania, P. K., Wang, O., Efros, A. A., Torr, P. H., & Shechtman, E. (2019). Interactive Sketch & Fill: Multiclass Sketch-to-Image Translation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1171-1180.

[28] Baek, K., Choi, Y., Uh, Y., Yoo, J., Shim, H. (2021). Rethinking the Truly Unsupervised Image-to-Image Translation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14154-14163.

[29] Park, T., Efros, A. A., Zhang, R., & Zhu, J. Y. (2020). Contrastive Learning for Unpaired Image-to-Image Translation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 319-345.

[30] Cheng, Z., Wang, C., Zhang, J., Song, Y.-Z., Hospedales, T. (2024). SMFS-GAN: Style-Guided Multi-class Freehand Sketch-to-Image Synthesis. *Computer Graphics Forum*, 43(6), e15136.

[31] Ham, C., Canet Tarres, G., Bui, T., Hays, J., Lin, Z., Collomosse, J. (2022). CoGS: Controllable Generation and Search from Sketch and Style. *Proceedings of the European Conference on Computer Vision (ECCV)*, 632-650.

[32] Sangkloy, P., Burnap, N., Ham, C., & Hays, J. (2016). The Sketchy Database: Learning to Retrieve, Generate, and Edit. *ACM Transactions on Graphics (TOG)*, 35(4), 119.

[33] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3836-3847.

[34] Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., & Shan, Y. (2024). T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5), 4296-4304.