DiffArtist: Towards Structure and Appearance Controllable Image Stylization

Ruixiang Jiang
rui-x.jiang@connect.polyu.hk
The Hong Kong Polytechnic University
Hong Kong, China

Chang Wen Chen
chen.changwen@polyu.edu.hk
The Hong Kong Polytechnic University
Hong Kong, China

Abstract

Artistic styles are defined by both their structural and appearance elements. Existing neural stylization techniques primarily focus on transferring appearance-level features such as color and texture, often neglecting the equally crucial aspect of structural stylization. To address this gap, we introduce **DiffArtist**, the first 2D stylization method to offer fine-grained, simultaneous control over both structure and appearance style strength. This dual controllability is achieved by representing structure and appearance generation as separate diffusion processes, necessitating no further tuning or additional adapters. To properly evaluate this new capability of dual stylization, we further propose a Multimodal LLM-based stylization evaluator that aligns significantly better with human preferences than existing metrics. Extensive analysis shows that DiffArtist achieves superior style fidelity and dual-controllability compared to state-of-the-art methods. Its text-driven, training-free design and unprecedented dual controllability make it a powerful and interactive tool for various creative applications. Project homepage: https://diffusionartist.github.io.

CCS Concepts

• **Computing methodologies** → *Appearance and texture representations*; **Image manipulation**; • **Applied computing** → **Fine arts**.

Keywords

Generative art; Text-driven stylization; Structure and appearance; Stylization evaluation; Multimodal LLM applications

ACM Reference Format:
Ruixiang Jiang and Chang Wen Chen. 2025. DiffArtist: Towards Structure and Appearance Controllable Image Stylization. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3746027.3755010

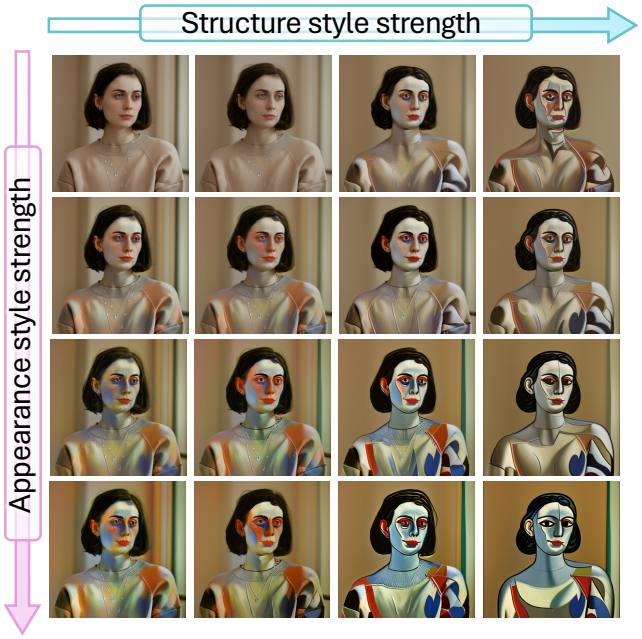

Figure 1: DiffArtist enables disentangled and fine-grained control of style strength from two orthogonal perspectives: structure and appearance. The style prompt is *"The Dream, by Picasso."*

1 Introduction

The essence of an artistic style lies not only in its appearance—color and texture—but also its structure—geometry and composition [17, 18, 30]. For example, the fragmentation of figures in Picasso's Cubist works and the undulating sky in Van Gogh's "Starry Night", each contributing distinctly to their artistic expression. Existing neural stylization approaches [10, 11, 24, 25, 37, 56] predominantly focus on manipulating appearance-level attributes. The structural elements in the source image, however, are often viewed as part of "content" and are explicitly preserved [55, 59, 61, 62]. This fundamental limitation prevents them from capturing the true essence of an art style, severely restricting their expressive potential and customizability.

The root of this limitation lies in the inherent complexity of structural stylization. Unlike appearance-style transfer, structural stylization requires a delicate balance between three competing objectives: (1) aligning with the target style, (2) harmonizing with the source image's composition, and (3) preserving the core semantic integrity of the content. These objectives operate at a high-level semantic plane, exposing a critical gap in current methods. While descriptors like AdaIN [25] and Gram loss [15] may suffice for appearance-style modeling, the lack of adequate structure representation and structure-style evaluators presents significant obstacles

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755010

in the development of structural stylization techniques. This challenge is amplified in recent multimodal generation scenarios, where a style prompt offers no explicit visual template [10, 22, 24, 53].

The advent of Diffusion Models (DMs) offers a powerful new paradigm for achieving this dual controllability, as their generative sampling process enables far greater structural and appearance diversity than prior methods. This reframes stylization as a conditional generation task, guided by a source image and a style prompt (image or text). However, this generative power comes with a critical, unaddressed challenge: the diffusion process inherently **entangles** the generation of structure and appearance. We identify this as a fundamental **Structure-Appearance (S-A) Tradeoff**: intensifying structural changes inadvertently corrupts appearance style, while strengthening appearance washes out structural transformations. This tradeoff directly explains the core failures of existing diffusion-based methods, which are either prone to severe content degradation [8, 44, 55] or suffer from weak, constrained stylization [58, 62]. Achieving dual controllability in the stylization thus remains an open question.

To solve this, we introduce **DiffArtist**, the first framework to our knowledge that offers explicit, disentangled control over both structure and appearance in 2D stylization. At its core, DiffArtist explicitly disentangles the structural and appearance generation as separated diffusion processes, with shared semantic information. This design directly overcomes the fundamental S-A tradeoff and functions as a zero-shot, plug-and-play module for any pretrained U-Net-based DM, requiring no costly fine-tuning or external adapters [63, 66]. As evidenced in this paper, this design provides true disentanglement—a key advantage over ControlNet-based methods [42, 54] where adjusting one style factor adversely impacts the other. As demonstrated in Fig. 1, this unprecedented level of control allows DiffArtist to achieve strong, semantically coherent stylization, unleashing the full creative potential of dual-style customization.

Evaluating this novel capability of dual control requires understanding on image semantics, where existing evaluation metrics obsolete. This exposes a critical need for a new evaluation paradigm. To address this, we introduce our second major contribution: **a Multimodal LLM (MLLM)-based evaluator** designed for dual stylization. We argue that any such evaluator must satisfy three key criteria: (1) operate at a **high-level semantic plane** to assess structure, (2) possess **contextual awareness** to maintain semantic integrity, and (3) perform robust **cross-modal association** between text prompts and visual forms. By leveraging the zero-shot reasoning of MLLMs, our proposed metric meets these criteria. We empirically show that it aligns significantly better with human artistic judgment than existing stylization metrics [34, 48, 57], establishing a more reliable and human-centric standard for future stylization research.

We summarize our contributions as follows:

(1) We identify the S-A tradeoff in diffusion models as the key challenge for disentangled dual controllability.
(2) We propose DiffArtist, the first 2D stylization method that enables the dual controllability of structure and appearance.
(3) We present a novel MLLM-based evaluator for evaluating structure and appearance in artistic stylization, which aligns better with human perception.
(4) Extensive experiments demonstrate that DiffArtist achieves superior stylization fidelity, control editability, and disentanglement than existing approaches.

2 Related Works

The Stylization of Structure and/or Appearance. Structure and appearance collectively define the style of a visual representation. Existing neural stylization methods [10, 11, 24, 25, 37, 56] have predominantly focused on appearance stylization, generally achieved via an encoder-decoder architecture. Only a few papers [33, 71] focus on transferring structural style components between 2D images. This is usually accomplished by calculating a correspondence between content and style image and performing a non-rigid deformation. However, these methods typically operate on images in specific domains like portraits, requiring an in-domain stylized reference. Moreover, they do not enable dual controllability. Very recently, dual controllability has been explored in Ctrl-X [42], but its framework is designed for localized, image edits rather than the global, harmonious style transformations. Meanwhile, the concept of disentangled control is explored in 3D synthesis, where explicit representations of shape (e.g., meshes [45] and radiance fields [53]) and appearance (e.g., texture mapping [5]) make separation natural. The success in 3D, however, is not directly transferable to the 2D domain due to the lack of explicit geometric priors in single images. In this paper, we explore the first dual controllable 2D image stylization method.

Text-Driven Image Stylization. Text-driven image stylization aims to stylize a source image according to style prompts. Early methods achieve it by optimizing certain image representation [13, 32, 37, 45, 53] with a multimodal alignment objective, typically implemented as the CLIP loss [48]. Recently, it was discovered that text-to-image (T2I) DMs could also be adapted for similar optimization schemes [20, 29, 31, 47]. These optimization-based methods are costly and slow, motivating recent exploration of the feed-forward paradigm. Instruct-Pix2Pix [4] tunes the diffusion model along with a language model for generalized editing tasks. Diffstyler [24] learn a content and style-specific denoiser for disentanglement. FreeStyle [19] modulates the U-Net feature for training-free stylization. Concurrent with this exploration is the stylized image generation [7, 14, 22, 54, 55]. While related, they focus on a different setting where the style is extracted from an image and the content is a prompt. In this work, we focus on the structure and appearance control in text-driven stylization scenarios.

Quantitative Evaluation of Style Transfer. Quantitatively evaluating style transfer is a long-standing problem. Initial approaches repurposed low-level metrics, including Gram Loss [16], LPIPS [67], and FID [23]. However, recent literature found that these metrics are fundamentally incapable of capturing the holistic and semantic qualities of human artistic perception [3, 6, 27, 50, 60, 64]. In response to these shortcomings, art-specific evaluators like Art-FID [60] and ArtScore [6] were developed to better quantify the abstract concept of "artness". Nevertheless, they cannot handle open-vocabulary text-driven stylization and lack the mechanism to

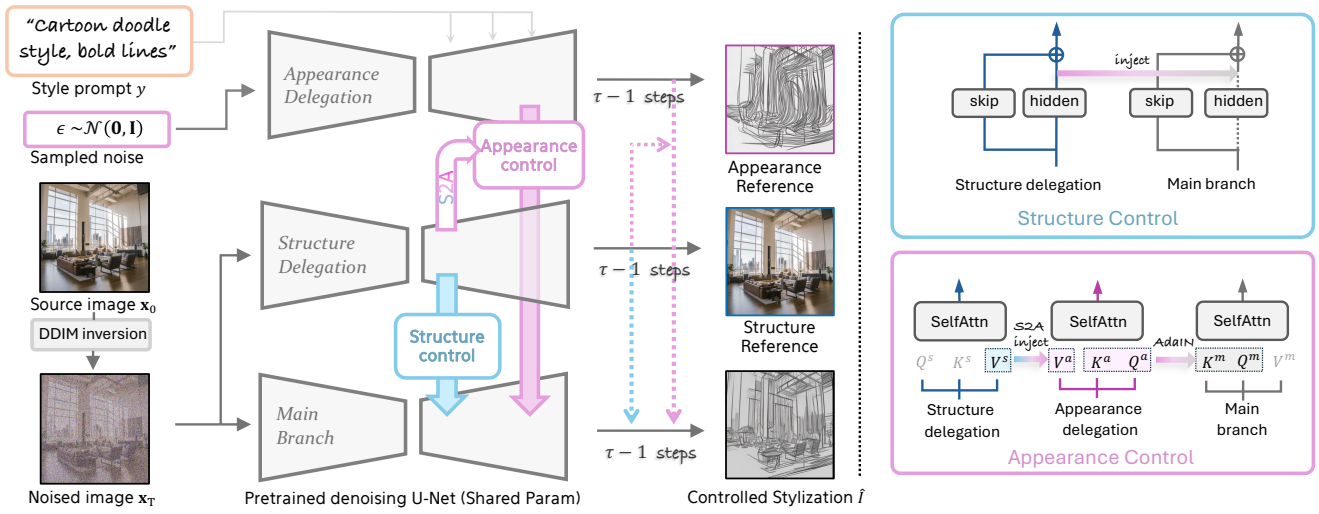

Figure 2: Overview of DiffArtist.Our method disentangles stylization by processing structure and appearance through two independent diffusion trajectories (delegations). At each denoising step, the main stylization branch is conditioned on semantic-level features from the structure and appearance delegations. All three branches share the same pretrained U-Net parameters, and perform full denoising of τ steps. The entire framework operates without requiring any fine-tuning or adapters.

evaluate structure and appearance separately. This paper propose a semantic-level MLLM-based evaluator to assess the structure and appearance fidelity, which aligns better with human perception.

3 Methodology

3.1 Objective: Disentangled Dual Controllability

Given a source image I, a text-based style prompt y, and a pre-trained diffusion model $\mathcal{G}(\cdot)$, our primary objective is to generate a stylized image $\hat{I}$ that preserves the semantic content of I while harmoniously embodying the style described by y. The core innovation we pursue is **disentangled dual control**, meaning that decomposing the style prompt y into two orthogonal components—structure and appearance—and controlling their strength independently. Our definition of structure and appearance in a 2D image is mainly based on fine art [17]. A formal definition of them is challenging as it relates to visual semiotics [18, 52], extending beyond the scope of this paper. Generally speaking, structure corresponds to the shapes, like contours and curvatures, while appearance corresponds to local patterns, like strokes and color palettes. We also aim to develop an evaluator $\mathcal{E}$, which can evaluate the fidelity of structure and appearance style in a way aligned with human perception.

The remaining parts of this section are organized as follows. In Sec. 3.2, we review the basics of inversion-based image manipulation. In Sec. 3.3, 3.4, we explain the motivation and design of control at a high level, and details are described in Sec. 3.5. Sec. 3.6 outlines the proposed MLLM-based structure and appearance evaluators.

3.2 Preliminary: DDIM Inversion

To stylize a source image $\mathbf{x}_0 := I$ using DMs, inversion-based methods first approximate the noise latents $\mathbf{x}_{1:T}$ of I, achieved via techniques such as DDIM inversion [51]. Stylization is then performed through re-generation with altered conditions (usually specified as a style prompt y). Specifically, one may start with an intermediate step $\mathbf{x}_\tau$ (i.e., control point), where $\tau \in [1, T]$ for iterative DDIM sampling. Each denoising step is formulated as follows:

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}_t, t; y)}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(\mathbf{x}_t, t; y), \tag{1}$$

where ϵ_θ is the denoiser, t is the timestep, and $\alpha_{1:T}$ is a predefined noise schedule. The assumption of this paradigm is that with a proper τ, the resulting stylized image $\hat{I} := \hat{\mathbf{x}}_0$ harmoniously integrates the structure and appearance of style in prompt y with the source image I.

3.3 Structure and Appearance in Noise Space

Prevailing neural stylization methods are built on a paradigm that separates an image into "content" and "style" [12, 25, 36, 39, 49, 58, 68, 69]. In this view, the style usually refers to the feature maps extracted from certain layers of a neural network. To advance stylization towards both structure and appearance controllability, we adopt different modeling that decomposes an image as its structure and appearance components: $\mathbf{x}_0 = \mathcal{G}_0(\mathbf{z}_0^s, \mathbf{z}_0^a)$, where $\mathcal{G}_0 = (\cdot, \cdot)$ is a composition function, $\mathbf{z}_0^s$ and $\mathbf{z}_0^a$ are the latent structure and appearance factorization, respectively. This is a "static", image-level perspective. In the diffusion process, the distribution of image x_0 is tied with the intermediate distributions in $\mathbf{x}_{1:T}$, where the denoiser ϵ_θ learns the transition $q_\theta(x_{t-1}|x_t, y)$ via ϵ-prediction. Therefore, we posit similar factorization of predicted noise residual, which is a "dynamic" decomposition across the full frequency bands:

$$\epsilon_\theta(\mathbf{x}_t, t; y) = \mathcal{G}_t(\kappa_t, \psi_t), \quad t \in [0, T] \tag{2}$$

where κ_t and ψ_t denote the structure and appearance representation at diffusion time-step t (detailed later), respectively. $\mathcal{G}_t(\cdot,\cdot)$ is a conceptual noise-space composition function at time t.

3.4 Structure and Appearance as Delegate Diffusion Process

We argue that the fundamental obstacle to achieving dual control in diffusion-based stylization is the inherent entanglement of structure and appearance. Our analysis, detailed in Appendix D, pinpoints the source of this problem: the reliance on a single latent trajectory $\mathbf{x}_\tau \rightarrow \mathbf{x}_0$. This monolithic generation process forces structural and appearance attributes to compete for influence at every denoising step, creating the S-A tradeoff that fundamentally limits controllability.

To break this bottleneck, we propose a novel mechanism that stylizes an image with **separate** diffusion trajectories, as illustrated in Fig. 2. Specifically, we leverage two supplementary diffusion processes with shared information, called *delegate branches*. We initialize the structure and main branch from the inverted noise $\mathbf{x}_T$, while appearance delegation starts from a Gaussian $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. These delegations enable controlling the stylization over the entire diffusion process. The controlled main branch can be denoted as:

$$\epsilon_\theta^m(\mathbf{x}_t, t; y, \kappa_t^s, \psi_t^a) = \mathcal{G}_t\left(\kappa_t^s \circ \kappa_t^m, \psi_t^a \diamond \psi_t^m\right), \quad t \in [0, T] \tag{3}$$

where the superscripts s, a, and m denote the factorization extracted from the structure, appearance delegation, and main branch, respectively. The $\circ$ and $\diamond$ are two non-commutative control operators.

3.5 Structure and Appearance Representations in Denoising U-Net

Having established the control mechanism in Eq. 3, we now formulate the κ and ψ in a U-Net-based denoiser for disentangling structure and appearance control.

Pyramidal Structure Representation κ. To effectively control structural stylization, we require a representation that captures image semantics at multiple levels of abstraction. We identify the hidden features in the ResBlock of denoising U-Net as the ideal substrate for this purpose, which robustly encode appearance-invariant image semantics across varying t (see Fig. 3). Formally, we denote the hidden feature of a ResBlock as $h_i(\mathbf{x}_t)$, where $i \in \{1, 2, \ldots, N_{res}\}$ indexes the ResBlocks up to N_{res}, with increasing spatial resolution. Stacking such feature from all layers forms a pyramidal structure representation of $\mathbf{x}_0$ at t:

$$\begin{aligned} \kappa_t^s &:= \{h_i(\mathbf{x_t})\}_{i \in S_{res}}, \\ &\text{where } h_i \text{ extracted from } \epsilon_\theta^s(\mathbf{x}_t, t; \varnothing), \\ &\text{and } S_{res} \subseteq \{1, 2, \ldots, N_{res}\}. \end{aligned} \tag{4}$$

Our representation is distinct as it captures multi-scale semantics and provides continuous guidance across the full denoising trajectory ($t \in [0, T]$). This fundamentally differs from methods relying on solitary control points (e.g., $\mathbf{x}_\tau$) or single-scale conditions (e.g., ControlNet, IP-Adapter). Such approaches are constrained to a fixed resolution and/or SNR, which architecturally limits their ability to generate complex structural styles. As evidenced in Sec. 5, this constraint often leads to undesirable semantic trade-offs. With this pyramidal representation, we implement the structure control operator $\circ$ as *injection* (i.e., $a \circ b = a$).

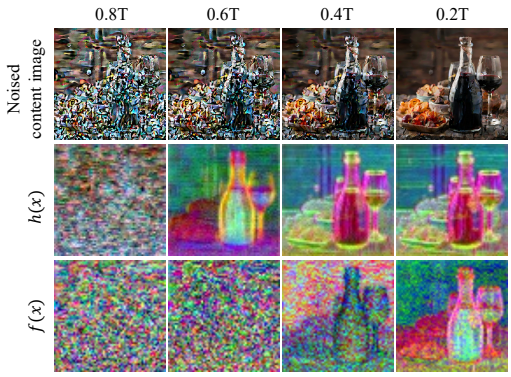

Figure 3: ResBlock feature map visualization. We apply t-sne to visualize the feature map of different feature maps in U-Net decoder. The hidden features $h(\mathbf{x})$ better preserves the semantics than the ResNet feature $f(\mathbf{x})$ throughout all T.

Semantic-aware Appearance Representation ψ. We represent the appearance of the target style as self-attention maps extracted from all layers of ϵ_θ^s. For the style to be applied harmoniously, its generation must be guided by the image's semantics. However, until now, we denoise appearance delegation from Gaussian noise and hence has no information-sharing with the source image. To compensate for this, we propose **Structure-to-Appearance injection (S2A)** that propagates the high-level semantics into appearance generation. Specifically, we inject the self-attention value V from early layers of ϵ_θ^s to ϵ_θ^a. Let N_{attn} denote the total number of attention blocks within the U-Net decoder, $S_{s2a} \subseteq \{1, 2, \ldots, N_{attn}\}$ be the selected blocks for S2A injection. The appearance representation at t is:

$$\begin{aligned} \psi_t^a := \{A_i^a\}_{i=1}^{N_{attn}}, \quad &\text{where } A_i^a \text{ is extracted from } \epsilon_\theta^a\left(\mathbf{x}_t, t; y, \{V_j^s\}_{j \in S_{s2a}}\right), \\ &\text{with } \{V_j^s\}_{j \in S_{s2a}} \text{ extracted from } \epsilon_\theta^s(\mathbf{x}_t, t; \varnothing). \end{aligned} \tag{5}$$

Inspired by StyleAligned [22] we design the appearance-style control operator $\diamond$ as the AdaIN [25],

$$a \diamond b = \sigma(b)\left(\frac{a - \mu(a)}{\sigma(a)}\right) + \mu(b). \tag{6}$$

We visualize the feature interactions in the right part of Fig. 2. Adjusting the control layer in S_{res} and style strength of ϵ_t^a enables disentangled control for structure and appearance, respectively.

3.6 Structure and Appearance Evaluation via MLLMs

Recent research demonstrates the powerful semantic-level multimodal understanding of MLLMs [35, 38, 40, 65, 70]. We leverage state-of-the-art MLLMs as a zero-shot evaluator to assess two key axes of our method: structure preservation and appearance fidelity.

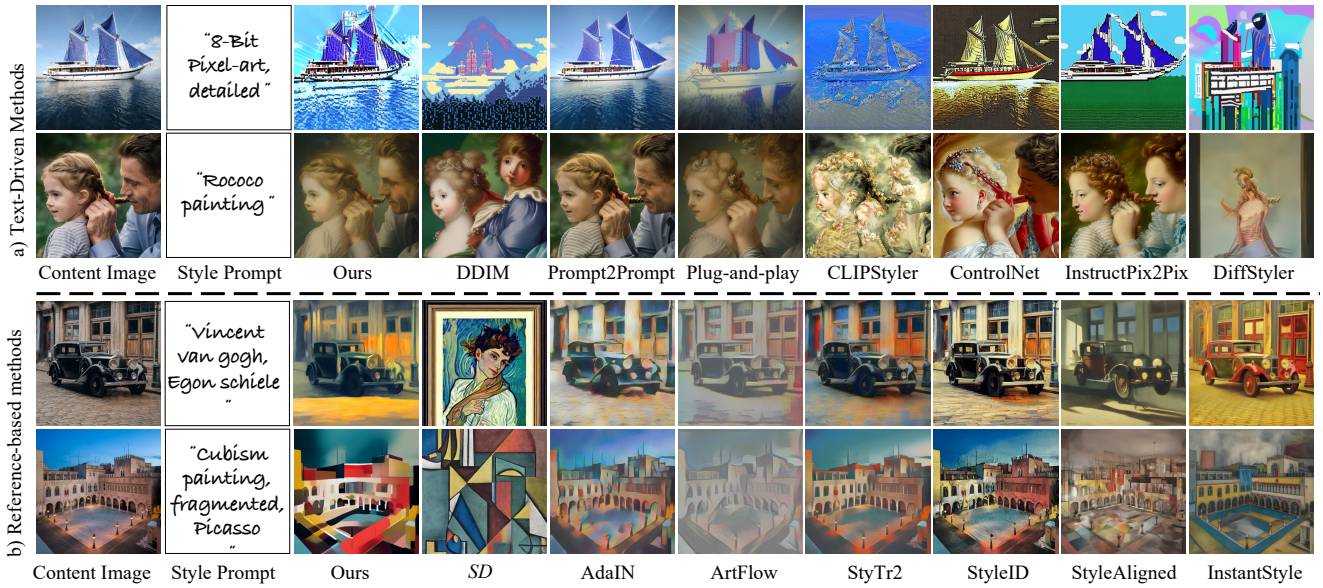

Figure 4: Qualitative comparison with existing methods. We compare our work with representative text-driven image manipulation method in (a), and image-based stylization methods in (b). Stylized images generated by DiffArtist produce high-fidelity structural and appearance-level style with semantic integrity. We suggest readers for more visualizations in Appendix A.

Crucially, our goal is to measure fidelity of stylized images, not subjective qualities like visual appeal.[1].

Given the inherent subjectivity and the difficulty of assigning absolute scores, we design a relative evaluation framework. Specifically, we query MLLM with the tuple $(\hat{I}, I, y, y_i)$, where $\hat{I} = \{\hat{I}_1, \ldots, \hat{I}_k\}$ is the stylization result generated by k different models, $I := \mathbf{x}_0$ is the source image, and y_i is the instruction dedicated for structure or appearance fidelity evaluation. The MLLM is tasked with ranking the outputs of k different methods for each criteria. We show in Sec. 5.2 that this design achieves superior alignment with human perception compared with existing metrics.

4 Experiments

4.1 Experiment Setup

Implementation Details. Our experiments are built upon the publicly available Stable Diffusion 2.1 model[2]. We perform DDIM sampling with $T = 50$ steps. During the inversion, we record the intermediate noise predictions to overwrite the input of ϵ_θ^s during denoising. Further implementation details are available in Appendix B. Experiments were conducted with a single RTX 4090-D GPU, with an approximate runtime of 2 seconds for inversion and 8 seconds for the final stylization.

Default Parameters. In main experiment, we default S_{res} to be the first four ResNet layers (`[1, 2, 3, 4]`), and S_{s2a} as the first two attention layer features (`[1, 2]`). The classifier-free guidance (CFG) scale is set as 7.5. These default control parameters correspond to moderate structural and appearance variations, used to set a fair comparison with existing works to avoid per-image parameter tuning. However, it should be noted that users can adjust these parameters for customization.

Compared Methods. We compare our method against existing text-driven stylization and manipulation methods: DDIM Inversion [51], CLIPStyler (optimization-based) [37], DiffStyler [24], Plug-and-Play (PnP), Prompt2Prompt (with null text inversion) (P2P) [21, 46], ControlNet [66], and InstructPix2Pix [4]. Additionally, we consider a baseline named *SD*, which generates images with Stable Diffusion according to y.

We also indirectly compare our method with reference-based stylization methods, including AdaIN [25], ArtFlow [1], StyTr2 [11], StyleID [8], StyleAligned [22] (with ControlNet), and InstantStyle [54] (with ControlNet). Images generated by *SD* are used as reference.

Conventional Metrics: **LPIPS** [67] measures the content preservation by calculating the feature distance between the source and stylized image. For style fidelity, we leverage **CLIP Score** [48] and **Pick Score** [34], both of which quantify the alignment between the stylized image $\hat{I}$ and prompt y. We also include a **human study** crowd-sourced from $n_1 = 200$ users and report the average preference rate for our method.

MLLM-based Metrics: We prompt the MLLM to rank the fidelity of k stylized images from best (rank 1) to worst (rank k). We normalize the integer ranking and average it over the whole evaluation set. Therefore, a score closer to 1 indicates a stronger fidelity. We use `Gemini-v2.0-flash` for its strong multimodal capability. The full prompt templates can be found in Appendix C.

4.2 Comparisons

Qualitative Comparisons. We first provide a comprehensive comparison against previous methods, visualized in Fig. 4-(a, b). **(a)**:

[1]Note that visual-appeal is also not equivalent to *aesthetic quality* [17, 28]
[2]https://huggingface.co/stabilityai/stable-diffusion-2-1

Table 1: Quantitative comparison against existing methods. We show conventional (in gray font) and MLLM-based metrics for representative methods. For each metric, indicates the best score, indicates the second best score , and indicates the third best score (best viewed in color). Win rate means the percentage that our method wins in pair-wise comparison.

Metric	Ours	DDIM	SD	PnP	P2P	InstructP2P	ControlNet	InstantStyle	DiffStyler	CLIPStyler
Inference time (sec)	10.5	9.7	3.9	55.3	29.1	9.2	7.8	7.8	18.2	24.2
Training & adapter free	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
LPIPS ↓	0.52	0.57	0.76	0.67	0.47	0.42	0.65	0.59	0.71	0.46
CLIP Score [48] ↑	25.91	25.25	27.46	24.89	23.48	21.94	24.93	22.85	25.79	27.14
PickScore [34] ↑	20.51	20.58	20.68	20.34	20.50	20.06	20.46	19.97	19.24	20.13
Structure (MLLM) ↑	0.61	0.22	0.29	0.52	0.65	0.60	0.58	0.56	0.35	0.51
Appearance (MLLM) ↑	0.67	0.46	0.31	0.60	0.47	0.59	0.55	0.67	0.30	0.59
Avg. (MLLM) ↑	0.64	0.34	0.30	0.56	0.56	0.60	0.57	0.62	0.33	0.55
Structure Win (Human) ↑	-	78.2%	62.4%	64.7%	57.3%	62.2%	71.2%	59.8%	81.3%	73.0%
Appearance Win (Human) ↑	-	74.2%	86.4%	62.0%	73.7%	68.7%	75.0%	60.1%	85.3%	76.3%

Compared with text-driven methods, DiffArtist is the best at following the style prompt while maintaining semantic integrity. Our method enables harmonious structural variations, such as pixelation, without compromising intricate details like facial identity and hair. By contrast, the compared methods may produce misaligned styles (*e.g.,* CLIPStyler, Plug-and-Play) or introduce undesired modifications that violate semantics (*e.g.,* DiffStyler, ControlNet). **(b)**: When broadly compared with reference-based methods, DiffArtist still stands out for its high stylization fidelity from two perspectives. To fully demonstrate the superiority, we **highly suggest readers for additional visualizations** in Appendix A.

Quantitative Comparison. For our quantitative evaluation, we first sample 50 art styles from WikiART, with both abstract (e.g., "Cubism") and realistic styles (e.g., "High Renaissance"), which are further diversified by GPT-4o in terms of description. This diversification sets a broad spectrum of styles to align with real-world user inputs. The content comprises 50 images from MSCOCO [43] and 50 photorealistic images generated by another model [26]. For each of the 100 content images, we randomly draw 10 style prompts from all possible styles, resulting in a total of 1,000 unique combinations for comparison. Tab. 1 presents the results.

For conventional metrics, DiffArtist achieves an LPIPS of 0.52, a CLIP Score of 25.91, and a PickScore of 20.51, outperforming most of the compared methods. However, these metrics do not measure stylization quality in structure and appearance. As a simple counterexample, the baseline *SD* has the highest CLIP Score and PickScore, whereas it is not even performing stylization. We include these scores solely for reference.

When evaluated with MLLM-based metrics, DiffArtist attains the highest average score of 0.64. Specifically, our method achieves the second-highest structure score of 0.61, demonstrating DiffArtist's effectiveness in generating structural styles. While the editing-focused P2P method scores higher, it does so by sacrificing stylistic strength, evidenced in qualitative comparisons. Besides, our method achieves the best appearance fidelity score, confirming its superior ability to render the appearance details from the text prompt. In human evaluations, our method is preferred by an average of 67.8% of users in pairwise comparison, further validating its superiority.

5 Analysis and Discussion

This section provide in-depth analysis on the proposed system. In Sec. 5.1, we analyze the controllability of DiffArtist in detail. Sec. 5.2 validates the effectiveness of MLLMs as style evaluators. Sec. 5.3 provides ablations on delegations and the S2A injection. We conclude with a discussion in Sec. 5.4.

5.1 DiffArtist has Strong Controllability

This subsection validates the unprecedented dual controllability of DiffArtist. To achieve this, we conduct a fine-grained comparative analysis against representative methods, categorizing them by their core control mechanism: (a) **Semantic pyramid:** include DiffArtist, DiffArtist implemented with Plug-and-Play structure representation, $f(\mathbf{x})$ (Ours + PnP), and Ctrl-X; (b) **Pixel-level map:** include InstantStyle [10], which is based on ControlNet [66]; (c) **Noise inversion**, which corresponds to the DDIM baseline. For structure control, we define five levels from weakest to strongest. For group (a) we use the following control layers: (∅, `[1]`, `[1-4]`, `[1-6]`, `[1-8]`); for group (b), we evenly sample their respective control strength parameters; and we use $\tau = [0, 5, 10, 15, 20]$ for group (c). Appearance strength is controlled by sampling CFG weights in `[2.5,5,7.5,10]` for all groups except for Ctrl-X, which is achieved by adjusting its appearance schedule parameter.

Qualitative Comparison on Control. As visualized in Fig. 5, DiffArtist demonstrates superior controllability, with harmonious, consistent, and disentangled interpolations across a sequence of control levels. It correctly captures the essential geometric principles of Cubism, whereas competing methods merely apply a superficial texture (e.g., Ours+PnP) or fail to produce meaningful structural variations (e.g., InstantStyle). This nuanced control is further evident when adjusting style strength; DiffArtist provides an artistically meaningful interpretation by producing bolder strokes according to style prompt, while other methods resort to simplistic and often undesirable increases in color saturation. Most critically, our approach preserves semantic integrity. It avoids the catastrophic failures of pixel-map methods like InstantStyle, which can render the face unrecognizable, and also prevents the facial structure corruption seen in inversion-based methods that inherently entangle form with appearance. The extended visualizations in Appendix A

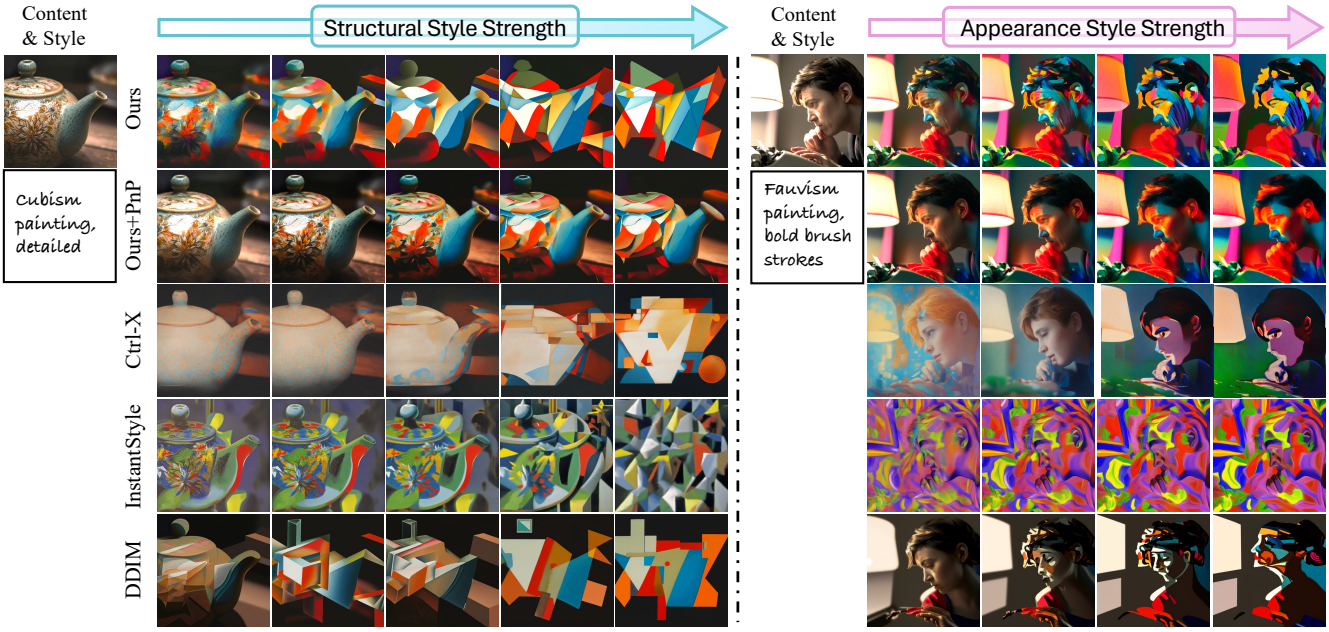

Figure 5: DiffArtist offers disentangled and stronger controllability. (Left): DiffArtist enables smooth and artistically meaningful structural stylization at varying degree, without violating appearance style strength. (Right): DiffArtist allows fine-grained control of appearance-related style strength while preserving structural and semantic integrity.

with diverse style and content further confirm the superiority of DiffArtist.

Quantitative Comparisons on Control. We provide quantitative experiments to substantiate our visual observations, evaluating control based on two properties: **fidelity** and **editability**. We measure the fidelity as structure and appearance MLLM scores at different control levels. The results are reported in Tab. 2. DiffArtist outperforms others significantly and consistently, demonstrating **superior control fidelity**.

The editability defines the quality of the manipulation itself, which we assess via three criteria: **range, monotonicity**, and **disentanglement**. An ideal control of stylization should cover a wide range of stylistic effects (a large spread in MLLM scores), exhibit predictable monotonicity (Spearman's $\rho \approx 1$ for the target attribute), and maintain disentanglement from other attributes. We measure the disentanglement using Kendall's W on the off-target scores, where a stable, unaffected score sequence yields $W \approx 0$. For instance, when controlling structural strength, W for the appearance score is desired to be near 0. As shown in Tab. 2, the control of DiffArtist is the most editable: it covers the broadest range of effects, demonstrates the strongest monotonicity, and achieves the best disentanglement, reaffirming the superior control visualized in Fig. 5.

5.2 MLLMs are Human-Aligned Stylization Evaluators

We evaluate how each stylization metric aligns with human feedback by calculating the statistical correlation with the rankings

Table 2: Fidelity of structure and appearance control via cross-method comparison. The values correspond to the structure or appearance score (MLLM, ↑). Note that the magnitudes of scores are only comparable within each column. The best result for each column is in bold.

Method	— Structure →					— Appearance →			
	lv.1	lv.2	lv.3	lv.4	lv.5	lv.1	lv.2	lv.3	lv.4
Ours	**0.62**	**0.65**	**0.74**	**0.63**	**0.66**	**0.70**	**0.74**	**0.80**	**0.78**
Ours+PnP	0.43	0.38	0.28	0.42	0.34	0.21	0.26	0.27	0.32
Ctrl-X	0.49	0.46	0.36	0.46	0.47	0.41	0.37	0.33	0.42
InstantStyle	0.42	0.42	0.33	0.43	0.35	0.34	0.34	0.35	0.60
DDIM	0.49	0.52	0.52	0.51	0.55	0.59	0.53	0.50	0.52

from human feedback. To achieve this, we first construct a comparison set of 800 stylized images[3], and compare how human and MLLM preferences correlate.

To gather human feedback, we consider two groups of users. For the non-expert group, we recruited a large-scale group of $n_1 = 200$ participants through a crowdsourcing platform. Each participant performed a series of randomly sampled ranking tasks. To ensure the integrity of the collected data, we implemented attention checks and consistency filters to remove unreliable responses. We also recruited an expert group of $n_2 = 12$ participants with knowledge of fine art.

We measured the alignment between each metric's rankings and the human-derived preferences using Spearman's rank correlation (ρ). The averaged (of all content-style pairs) ρ for both

[3]The images evaluated here do not overlap with the main experiment in Tab. 1

Table 3: Control editability and disentanglement via inter-method comparison. Higher ρ indicates stronger monotonicity while lower W means ranks are indistinguishable. When controlling from one perspective, a high ρ is desired for editability, and a low W for the other aspect is expected for disentangled control. The controls in DiffArtist are the most editable and disentangled.

	Sequential Structure-Control Only						Sequential Appearance-Control Only				
	Ours	Ours+PnP	Ctrl-X	InstantStyle	DDIM		Ours	Ours+PnP	Ctrl-X	InstantStyle	DDIM
$\rho(S)\uparrow$	**0.82**	0.54	0.32	0.39	0.70	$W(S)\downarrow$	0.37	**0.32**	0.36	0.45	0.69
$W(A)\downarrow$	**0.32**	0.44	0.45	0.34	0.72	$\rho(A)\uparrow$	**0.71**	0.42	0.35	0.26	0.68

Table 4: Metrics correlation with human feedback. We report correlation ρ and combined significance p. The MLLM scores show stronger alignment with both expert and non-expert users.

		Corr. (Non-expert)		Corr. (Expert)	
	Metrics	$\rho\uparrow$	p-value $\downarrow$	$\rho\uparrow$	p-value $\downarrow$
S	SSIM [57]	0.29	0.12	0.25	0.14
	MLLM (GPT-4o)	0.44	0.004	0.34	0.20
	MLLM (Gemini 2.0)	**0.42**	0.02	**0.45**	0.03
A	CLIP Score	0.05	0.73	0.01	0.75
	Pick Score [34]	0.27	0.11	0.25	0.13
	MLLM (GPT-4o)	0.25	0.05	0.22	0.06
	MLLM (Gemini 2.0)	**0.48**	0.04	**0.41**	0.02

Table 5: Ablation on delegation branches. The proposed two delegations are complementary to each other, and the full methods achieves the highest fidelity.

Metric \ Method	**full**	**w/o structure**	**w/o appearance**
LPIPS $\downarrow$	0.51	0.76	**0.42**
CLIP Score $\uparrow$	25.91	**27.69**	21.75
Pick Score [34] $\uparrow$	20.55	**20.57**	20.41
Structure (MLLM) $\uparrow$	**0.72**	0.37	0.33
Appearance (MLLM) $\uparrow$	**0.62**	0.59	0.22

groups is reported in Tab. 4. As the table shows, the MLLM-based metrics are **better aligned with human perception**, validating its effectiveness as an evaluation metric for style transfer.

5.3 Ablations

The delegations enable dual controllability. DiffArtist's controllability stems from delegating structure and appearance generation to separate processes. To test the necessity of each, we created two ablated variants for comparison where the structure or appearance delegation is removed. Tab. 5 presents the result of this ablation. The full methods achieves the best results, demonstrating the synergistic effect of delegations for dual controllability.

S2A injection promotes semantic-aligned spatial distribution of style strength in the style delegation process, thereby avoiding artifacts in the final stylization result. We visualize the denoised style image (from appearance delegation) and the final stylization result in Fig. 6. Without S2A injection, the appearance delegation fails to align with content semantics, generating an appearance reference image with undesired patterns and an uneven texture distribution. These flaws manifest directly in the final output as distracting visual artifacts. In contrast, the full model leverages S2A to produce a coherent style representation, resulting in a clean and high-quality final image.

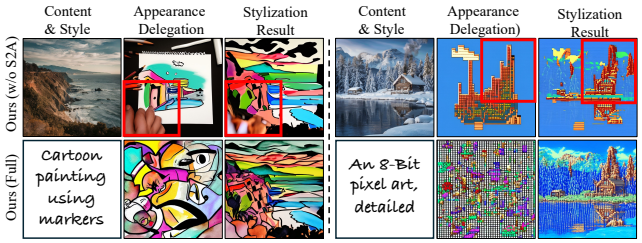

Figure 6: Ablation on S2A injection. S2A injection propagates the high-level semantic to the appearance generation. It avoids spatial misalignment of appearance-style strength.

5.4 Limitations & Future work

While DiffArtist marks a significant step towards disentangled structure and appearance control, we identify several limitations that open exciting avenues for future research. For instance, the structure control in DiffArtist is at a global level, and it cannot control the structure for each object separately. Many art styles exhibit mixed structure variation, like Surrealism and Collage art. Developing dense structure evaluators with 2D feedback signals is a promising direction [41], which may be further utilized as a reward model for reinforcement learning [2, 9].

6 Conclusion

We present the first exploration of structure- and appearance-controllable image stylization. Our contributions include DiffArtist, a styler that fully disentangles structure and appearance during the diffusion process, and a human-aligned evaluator to assess structural and appearance fidelity at the semantic level. Our extensive analysis proves that semantically-rich representations are essential for both structure and appearance style. We demonstrated that our design allows for high style fidelity and controllability, similar to that of a human artist. We believe the objective established in this paper—to stylize in both structure and appearance—offer a roadmap for the next generation of generative art tools to produce artistically meaningful paintings.

7 Acknowledgment

This research was supported by the Hong Kong Research Grants Council (GRF-15229423).

References

[1] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. 2021. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 862–871.

[2] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2023. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301* (2023).

[3] Yihang Bo, Jinhui Yu, and Kang Zhang. 2018. Computational aesthetics and applications. *Visual computing for industry, biomedicine, and art* 1 (2018), 1–19.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

[5] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. 2023. Text2tex: Text-driven texture synthesis via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 18558–18568.

[6] Junyu Chen, Jie An, Hanjia Lyu, Christopher Kanan, and Jiebo Luo. 2024. Learning to Evaluate the Artness of AI-generated Images. *IEEE Transactions on Multimedia* (2024).

[7] Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. 2023. Controlstyle: Text-driven stylized image generation using diffusion priors. In *Proceedings of the 31st ACM International Conference on Multimedia*. 7540–7548.

[8] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. 2024. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8795–8805.

[9] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. 2023. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400* (2023).

[10] Xing Cui, Zekun Li, Pei Pei Li, Huaibo Huang, and Zhaofeng He. 2023. InstaStyle: Inversion Noise of a Stylized Image is Secretly a Style Adviser. *arXiv preprint arXiv:2311.15040* (2023).

[11] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. 2022. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11326–11336.

[12] Guanchen Ding, Lingbo Liu, Zhenzhong Chen, and Changwen Chen. 2024. Domain-agnostic crowd counting via uncertainty-guided style diversity augmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 1642–1651.

[13] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.

[14] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. 2024. Styleshot: A snapshot on any style. *arXiv preprint arXiv:2407.01414* (2024).

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

[16] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.

[17] Ernst Hans Gombrich and EH Gombrich. 1995. *The story of art*. Vol. 12. Phaidon London.

[18] Nelson Goodman. 1976. *Languages of art: An approach to a theory of symbols*. Hackett.

[19] Feihong He, Gang Li, Mengyuan Zhang, Leilei Yan, Lingyu Si, Fanzhang Li, and Li Shen. 2024. Freestyle: Free lunch for text-guided style transfer using diffusion models. *arXiv preprint arXiv:2401.15636* (2024).

[20] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. 2023. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2328–2337.

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

[22] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2024. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4775–4785.

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[24] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Weiming Dong, and Changsheng Xu. 2024. Diffstyler: Controllable dual diffusion for text-driven image stylization. *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[25] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*. 1501–1510.

[26] Ideogram. 2024. Ideogram: Explore. https://ideogram.ai/t/explore Accessed: 2024-03-20.

[27] Eleftherios Ioannou and Steve Maddock. 2024. Evaluation in Neural Style Transfer: A Review. In *Computer Graphics Forum*. Wiley Online Library, e15165.

[28] Ruixiang Jiang and Changwen Chen. 2025. Multimodal LLMs Can Reason about Aesthetics in Zero-Shot. *arXiv preprint arXiv:2501.09012* (2025).

[29] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2023. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14371–14382.

[30] Hyunyoung Jung, Seonghyeon Nam, Nikolaos Sarafianos, Sungjoo Yoo, Alexander Sorkine-Hornung, and Rakesh Ranjan. 2024. Geometry transfer for stylizing radiance fields. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8565–8575.

[31] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

[32] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2426–2435.

[33] Sunnie SY Kim, Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. 2020. Deformable style transfer. In *European Conference on Computer Vision*. Springer, 246–261.

[34] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2023), 36652–36663.

[35] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[36] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. 2019. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4422–4431.

[37] Gihyun Kwon and Jong Chul Ye. 2022. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18062–18071.

[38] Yanshu Li, Hongyang He, Yi Cao, Qisen Cheng, Xiang Fu, and Ruixiang Tang. 2025. M2iv: Towards efficient and fine-grained multimodal in-context learning in large vision-language models. *arXiv preprint arXiv:2504.04633* (2025).

[39] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. 2017. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036* (2017).

[40] Yanshu Li, Tian Yun, Jianjiang Yang, Pinyuan Feng, Jinfa Huang, and Ruixiang Tang. 2025. TACO: Enhancing Multimodal In-context Learning via Task Mapping-Guided Sequence Configuration. *arXiv preprint arXiv:2505.17098* (2025).

[41] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. 2024. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19401–19411.

[42] Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. 2024. Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance. *Advances in Neural Information Processing Systems* 37 (2024), 128911–128939.

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 740–755.

[44] Yueming Lyu, Yue Jiang, Bo Peng, and Jing Dong. 2023. InfoStyler: Disentanglement information bottleneck for artistic style transfer. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 4 (2023), 2070–2082.

[45] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. 2022. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13492–13502.

[46] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6038–6047.

[47] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[49] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. 2018. A style-aware content loss for real-time hd style transfer. In *proceedings of the European conference on computer vision (ECCV)*. 698–714.

[50] Chaehan So. 2023. Measuring aesthetic preferences of neural style transfer: More precision with the two-alternative-forced-choice task. *International Journal of Human–Computer Interaction* 39, 4 (2023), 755–775.

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[52] Theo Van Leeuwen and Carey Jewitt. 2000. *The handbook of visual analysis*. Sage.

[53] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2023. Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization and Computer Graphics* (2023).

[54] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. 2024. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733* (2024).

[55] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. 2024. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788* (2024).

[56] Xinhao Wang, Wenjing Wang, Shuai Yang, and Jiaying Liu. 2022. CLAST: Contrastive learning for arbitrary style transfer. *IEEE Transactions on Image Processing* 31 (2022), 6761–6772.

[57] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[58] Zhizhong Wang, Lei Zhao, and Wei Xing. 2023. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7677–7689.

[59] Linfeng Wen, Chengying Gao, and Changqing Zou. 2023. CAP-VSTNet: Content affinity preserved versatile style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18300–18309.

[60] Matthias Wright and Björn Ommer. 2022. Artfid: Quantitative evaluation of neural style transfer. In *DAGM German Conference on Pattern Recognition*. Springer, 560–576.

[61] Jinchao Yang, Fei Guo, Shuo Chen, Jun Li, and Jian Yang. 2022. Industrial style transfer with large-scale geometric warping and content preservation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7834–7843.

[62] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. 2023. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22873–22882.

[63] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).

[64] Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L Rosin. 2023. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22388–22397.

[65] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601* (2024).

[66] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[68] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2023. A unified arbitrary style transfer framework via adaptive contrastive learning. *ACM Transactions on Graphics* 42, 5 (2023), 1–16.

[69] Yexun Zhang, Ya Zhang, and Wenbin Cai. 2020. A unified framework for generalizable style transfer: Style and content separation. *IEEE Transactions on Image Processing* 29 (2020), 4085–4098.

[70] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923* (2023).

[71] Yang Zhou, Zichong Chen, and Hui Huang. 2024. Deformable one-shot face stylization via dino semantic guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7787–7796.

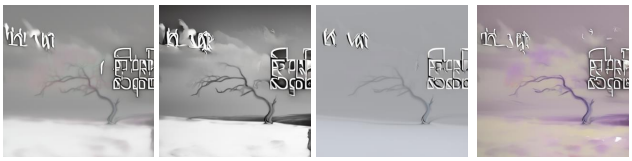

Figure 9: Example of failure cases.

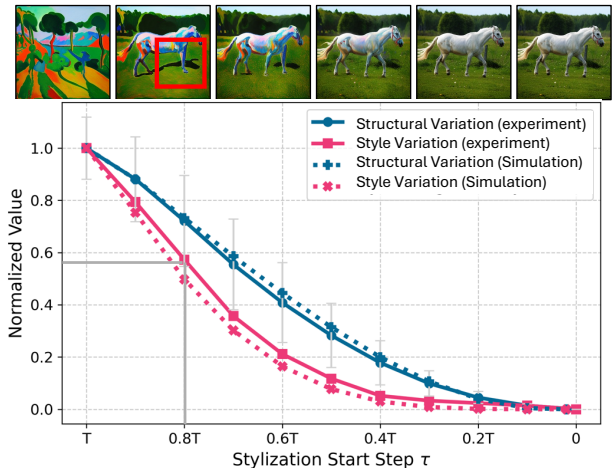

Figure 7: Tradeoff between Structure and Appearance Style Control. We present the experimental (solid lines) and simulation (dotted lines) trends of structural and appearance variation in the diffusion process. Due to quadratic growth, highly noisy steps are required to achieve strong appearance styles, which are associated with significant structural variation, which can violate semantics. Top: Example stylization results starting from different steps, using the prompt: *"Fauvism painting"*. When the appearance strength is high ($t = 0.8T$), the structure (legs of the horse) is incorrectly modified.

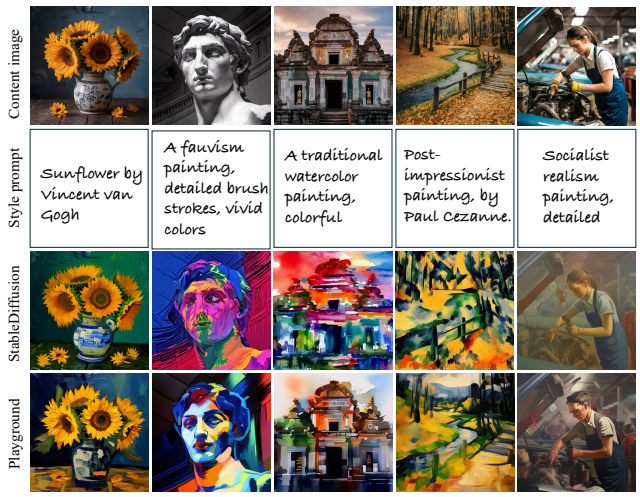

Figure 8: DiffArtist implemented with different diffusion architecture. We implement DiffArtist on the playground-v2 diffusion model. Similar stylization results could be achieved, demonstrating the generalizability of proposed method.

Appendix A Additional Qualitative Results

Visualization Additional appearance stylization is in Fig. 10, the source image is in Fig. 11. A grid of different images with different styles is in Fig. 12. Additional structure control is in Fig. 13 and Fig. 14.

Additional qualitative comparisons on stylization. We show an extended comparison with the previous *reference-based* method in Fig. 15. More qualitative comparisons with existing *text-driven* image stylization and manipulation methods can be found in Fig. 16 and Fig. 17.

Additional comparisons on fine-grained control. We provide additional comparisons with other control methods in Fig. 18 and Fig. 19. These results demonstrate the advantage of DiffArtist in providing disentangled structural and appearance-level style control. In particular, the Ctrl-X, as an image editing method, produces less visually pleasing results when applied to image stylization. This is because they have a different definition of appearance and structure for editing real photos.

Appendix B On the Structure and Appearance Entanglement in Diffusion Process

B.1 Theoretical analysis

We now explore how the factorization of structure and appearance factorization, defined in Eq. 2, interact and evolve throughout the denoising trajectory $p_\theta(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{T-1}, |\mathbf{x}_T)$. Specifically, suppose the content image I_c is inverted into $\mathbf{x}_{1:T}$, inversion-based stylization starts from an intermediate step $\mathbf{x}_\tau, \tau < T$ for DDIM denoising process. By rearranging Eqn. 1, we obtain:

$$\begin{aligned} \mathbf{x}_{t-1} &= \mathcal{A}_t \mathbf{x}_t + \mathcal{B}_t \epsilon_\theta(\mathbf{x}_t, t; y), \quad \mathcal{A}_t := \frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}} \\ \mathcal{B}_t &:= \sqrt{1-\alpha_{t-1}} - \frac{\sqrt{\alpha_{t-1}(1-\alpha_{t-1})}}{\sqrt{\alpha_t}} \end{aligned} \tag{7}$$

Based on above formulation, the full stylization process could be expressed as:

$$\begin{aligned} \hat{\mathbf{x}}_0 = \prod_{j=1}^{T} \mathcal{A}_j \cdot \mathbf{x}_T + \underbrace{\sum_{k=\tau+1}^{T} \left[\mathcal{B}_k \prod_{j=\tau+1}^{k-1} \mathcal{A}_j \right] \epsilon'(\mathbf{x}_{k-1}, k)}_{\text{preserve original structure and appearance}} \\ + \underbrace{\sum_{k=1}^{\tau} \left[\mathcal{B}_k \prod_{j=1}^{k-1} \mathcal{A}_j \right] \epsilon_\theta(\mathbf{x}_{k-1}, k; y)}_{\text{generate new structure and appearance}}, \end{aligned} \tag{8}$$

where ϵ' denotes an ideal denoiser that perfectly models the transition distribution $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$. In Eq. 8, the conceptual denoising term from $\mathbf{x}_T$ to $\mathbf{x}_\tau$ preserves the structure and appearance in I_c. The stylization trajectory from $\mathbf{x}_\tau$ to $\mathbf{x}_t$ is what is actually computed, which introduces the desired appearance based on the prompt y but

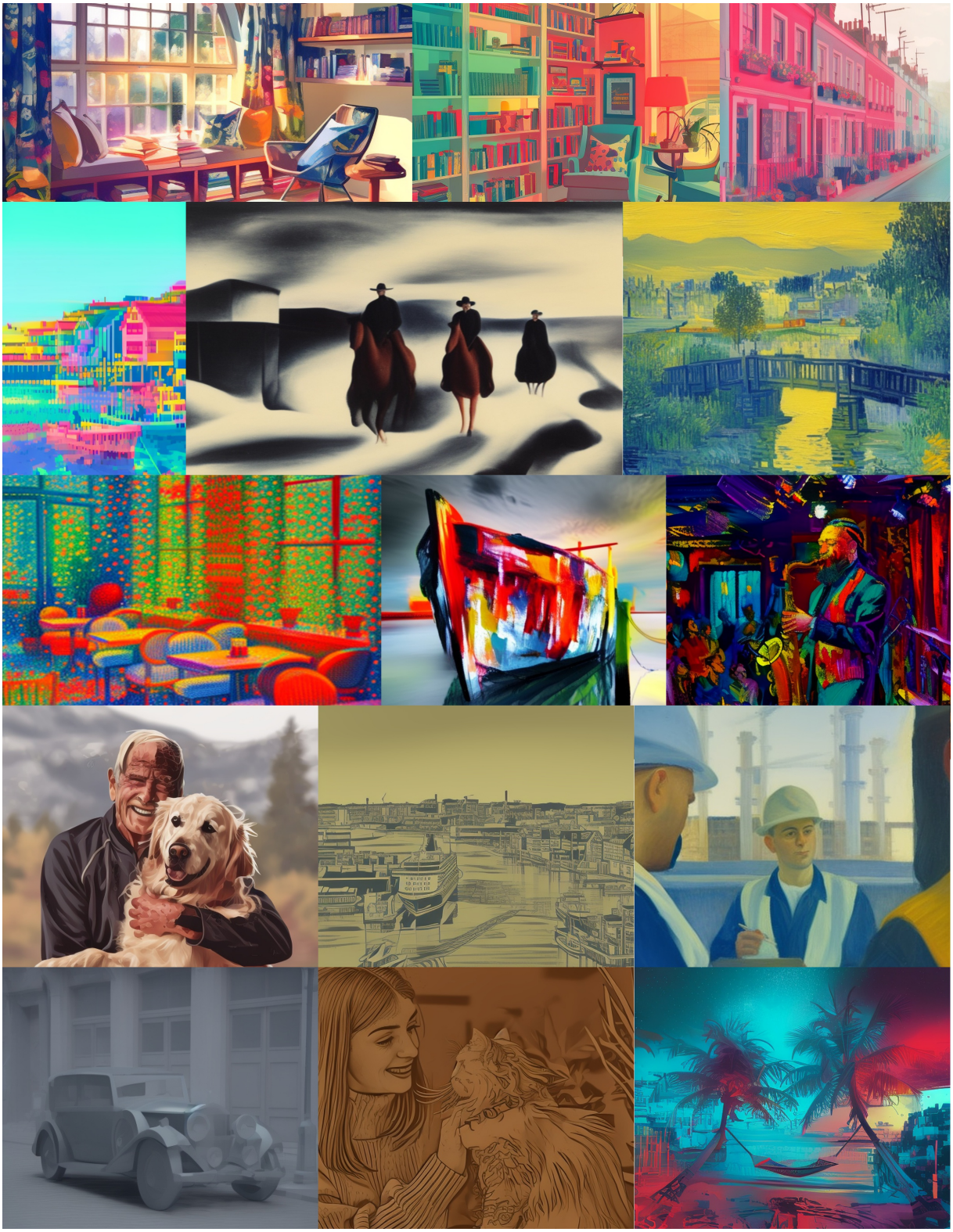

Figure 10: Addition results for DiffArtist (with default control parameters). The image semantics are preserved with strong and high-fidelity styles harmoniously integrated.

Figure 11: The corresponding source images of Figure 10.

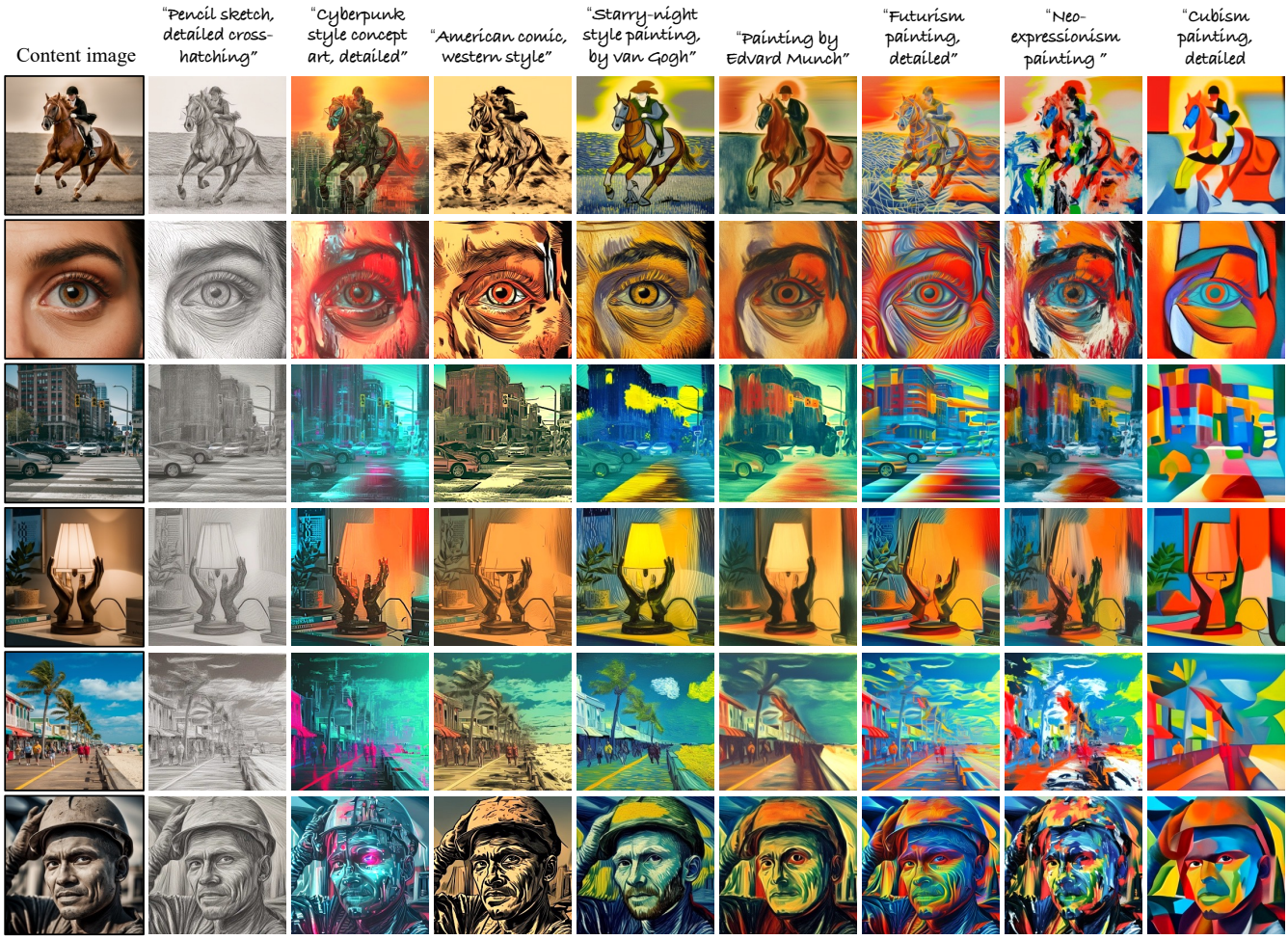

Figure 12: A grid experiment with different content and styles.

may also lead to *uncontrolled* structure generation. In this paradigm, preserving original appearance and generating new structure are undesirable side effects. Intuitively, a larger τ leads to a shorter trajectory of $\mathbf{x}_T \rightarrow \mathbf{x}_\tau$, resulting in a stronger appearance in $\hat{\mathbf{x}}_0$ with weaker structure preservation (uncontrolled structure stylization), while a lower τ sacrifices stylization strength for stronger structure preservation. In other words, one can not arbitrarily control the strength of appearance and structure without affecting the other.

By combining Eqn. 2 and Eqn. 8, we can quantitatively assess the strength of content preservation and stylization in the diffusion process under a particular noise schedule $\alpha_{1:T}$. Specifically, we further assume the SNR of structure and appearance has a linear relationship with that of the $\mathbf{x}_t$ for t:

$$\text{SNR}(\mathbf{z}_t^c) \propto \frac{\alpha_t}{1-\alpha_t}; \text{SNR}(\mathbf{z}_t^s) \propto \frac{\alpha_t}{1-\alpha_t} \tag{9}$$

B.2 Simulation

To derive the theoretical trends of structure and appearance strength during the diffusion process, we introduce an additional assumption. Specifically, we assume that the relative significance of each unweighted denoising step $\epsilon(\mathbf{x}_t, t; y)$ on the final stylized image remains consistent across different timesteps for both structure and appearance. In other words, we assume the SNR of structure and appearance has a linear relationship with that of the $\mathbf{x}_t$ at time t as characterized by the noise schedule $\alpha_{1:T}$:

$$\text{SNR}(\mathbf{z}_t^c) \propto \frac{\alpha_t}{1-\alpha_t}; \text{SNR}(\mathbf{z}_t^s) \propto \frac{\alpha_t}{1-\alpha_t}, \tag{10}$$

It is important to note that we do not assume that the relative proportions of structure and appearance are identical at each denoising step.

With the above assumption, the effect of varying τ on the structure and appearance of the final stylized image could be derived in closed form. In practice, we use the following code to calculate iteratively:

```
def cum_score(low_t, hi_t, alphas):
    res = 0
    for k in range(low_t + 1, hi_t):
        for j in range(low_t + 1, k - 1):
            res += A_t(j, alphas) * B_t(k, alphas)
    return res
```

Figure 13: Additional result on structure control - 1.

```
struct_scores = []
appear_scores = []

for tau in range(0, 50):
    crt_struct = cum_score(50 - tau, 50, sampled_alphas)
    crt_appear = cum_score(0, tau, sampled_alphas)
    struct.append(crt_struct)
    appear_scores.append(crt_appear)
```

B.3 Empirical Result

Due to the inherent inaccuracy of DDIM inversion, the estimation of $\mathbf{x}_\tau$ may be imperfect, resulting in unintended modifications in the final sampled image even if no style prompt y is used. To address this issue, we adopt an alternative strategy by randomly sampling 500 Gaussian noise as the $\mathbf{x}_T$ of content, which are paired with 500 content prompts. We treat the images denoised using content prompts for $\tau = T$ steps as the content image, which simulates a perfect inversion technique. To stylize an image, we first denoise the $\mathbf{x}_T$ with the content prompt for τ steps, which is subsequently denoised with the style prompt for $T - \tau$ steps. The LPIPS between the stylized image and the content image is used as the empirical structural strength. In contrast, the CLIP Deception score (correct classification rate among a set of styles) is used as the empirical appearance strength. The following 10 style prompts are used:

- "watercolor style"
- "fauvism style"
- "pencil sketch style"
- "pointillism style"
- "art deco style"
- "impressionism style"
- "surrealism style"
- "pop art style"
- "cubism style"
- "abstract expressionism style"

The results for both simulation and empirical results are in Fig. 7. The result shows a good fit, and it turns out that the structure modification appears to be linear, with the stylization strength being quadratic with respect to τ. Moreover, this result further evidenced the issue of S-A entanglement in the diffusion process.

Appendix C Details on MLLM-based metrics

C.1 Implementation Details

The stylized images, style prompt, and the instruction prompts are fed to MLLM for inference. We compose stylized images as a grid image with numbers at the top-left corner. The full prompt template for structure and appearance score is available in Tab. 6.

C.2 Correlation with Human Preference

Human Question Collection We distributed the questionnaire on a crowd-sourcing platform, where each participant was required to complete up to 20 randomly sampled ranking tasks. An example of the user interface is provided in Fig. 20. A total of 200 participants took part in this study. To ensure the validity of the responses,

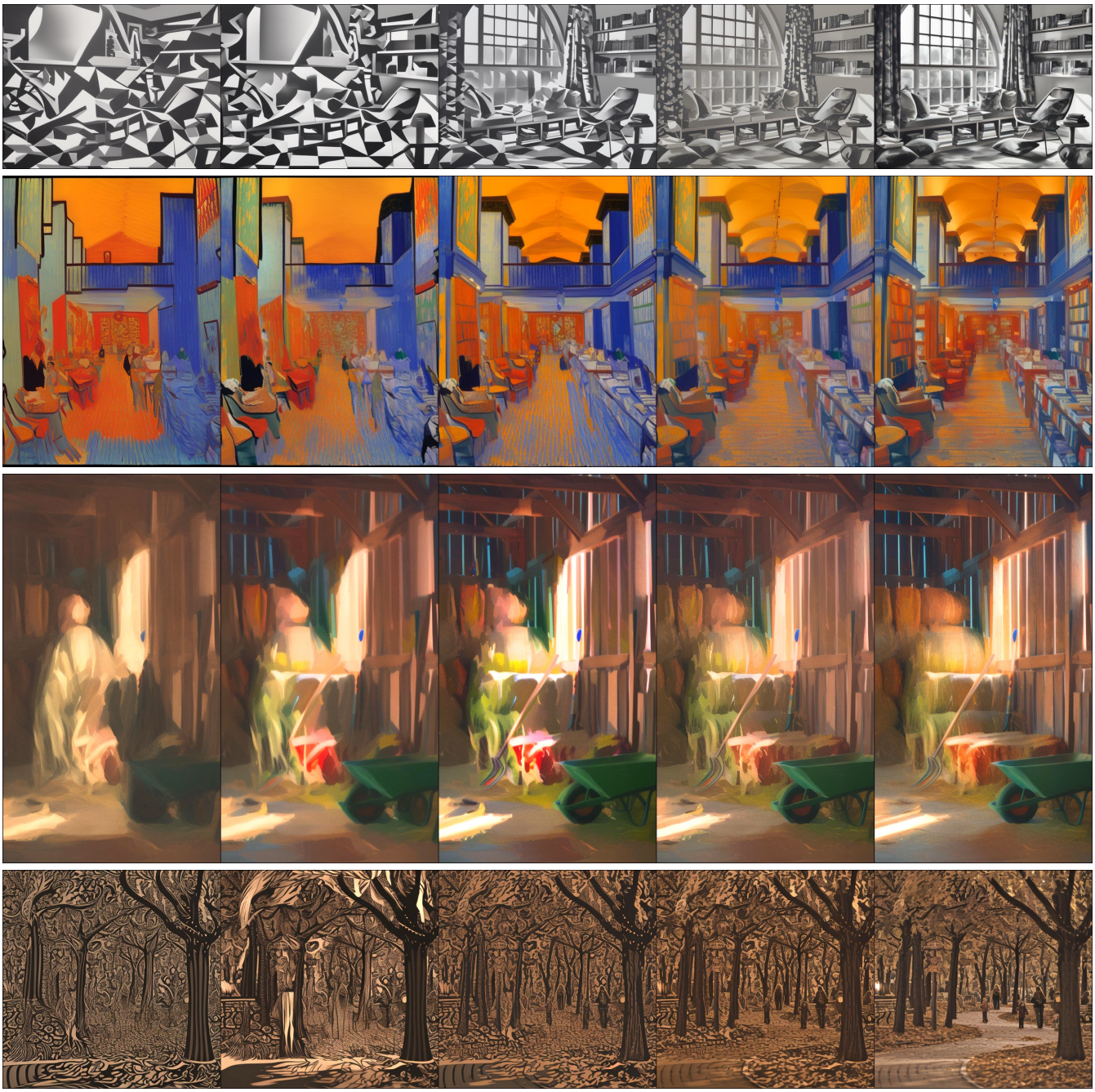

Figure 14: Additional result on structure control - 2.

we included attention-check questions. If a participant answers an attention-check question incorrectly, all of their responses will be marked invalid. Responses that are made with less than 20 seconds are also removed.

Appendix D Additional Discussion and Analysis

D.1 Generalizability of DiffArtist

To demonstrate the generalizability of DiffArtist across different U-Net-based diffusion architectures, we implement our method on Playground v2[4], which utilizes the SDXL architecture, distinct from Stable Diffusion 2.1. Several results are provided in Fig. 8. These

[4] https://huggingface.co/playgroundai/playground-v2-1024px-aesthetic

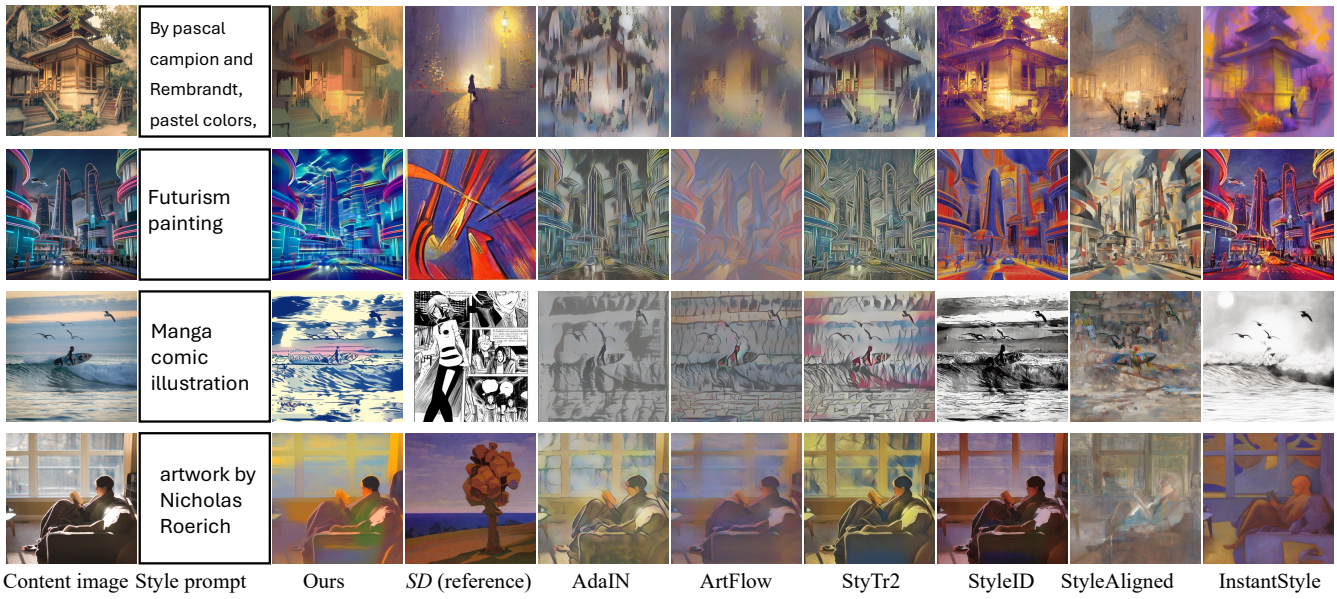

Figure 15: Extended comparison with reference-based style transfer methods.

results validate that DiffArtist serves as a versatile control method applicable to various U-Net-based diffusion models, regardless of their underlying architectural differences.

D.2 Additional Ablations on S2A Injection

In this section, we provide additional ablations to study the effect of proposed S2A design.

Ablation on S2A layers S_{S2A}. We ablate the number of injection layer used in the S2A injection (i.e., S_{S2A}). As illustrated in Fig. 21, the S2A layers influence the frequency bands of style delegation. Incorporating only early layers (e.g., `[1, 2]`) focuses on generating low-frequency style features such as tones and small objects, while adding more layers facilitates the creation of high-frequency style details like stroke shapes. Empirically, we set the S2A injection layers to `[1, 2]` by default, as using additional layers typically results in blurriness in the stylized outputs.

D.3 Failure Case

In our experiment, we identify a rare (<1%) and special failure case in the proposed methods. Specifically, for certain content image, its stylization result will consistently contains black and white chessboard-pattern artifacts. We provide one example in Fig. 9.

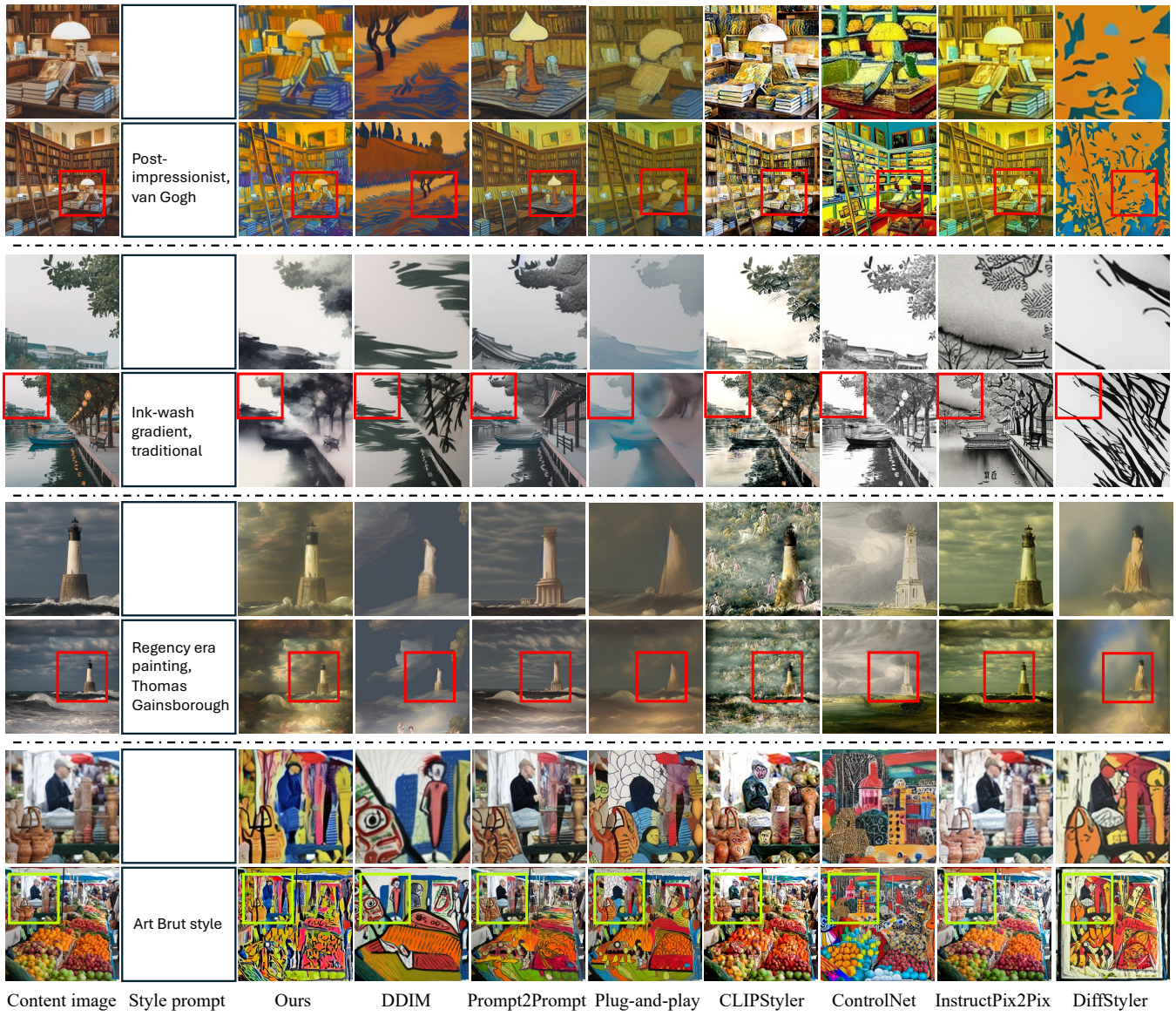

Figure 16: Extended comparison with existing text-driven stylization and manipulation methods.

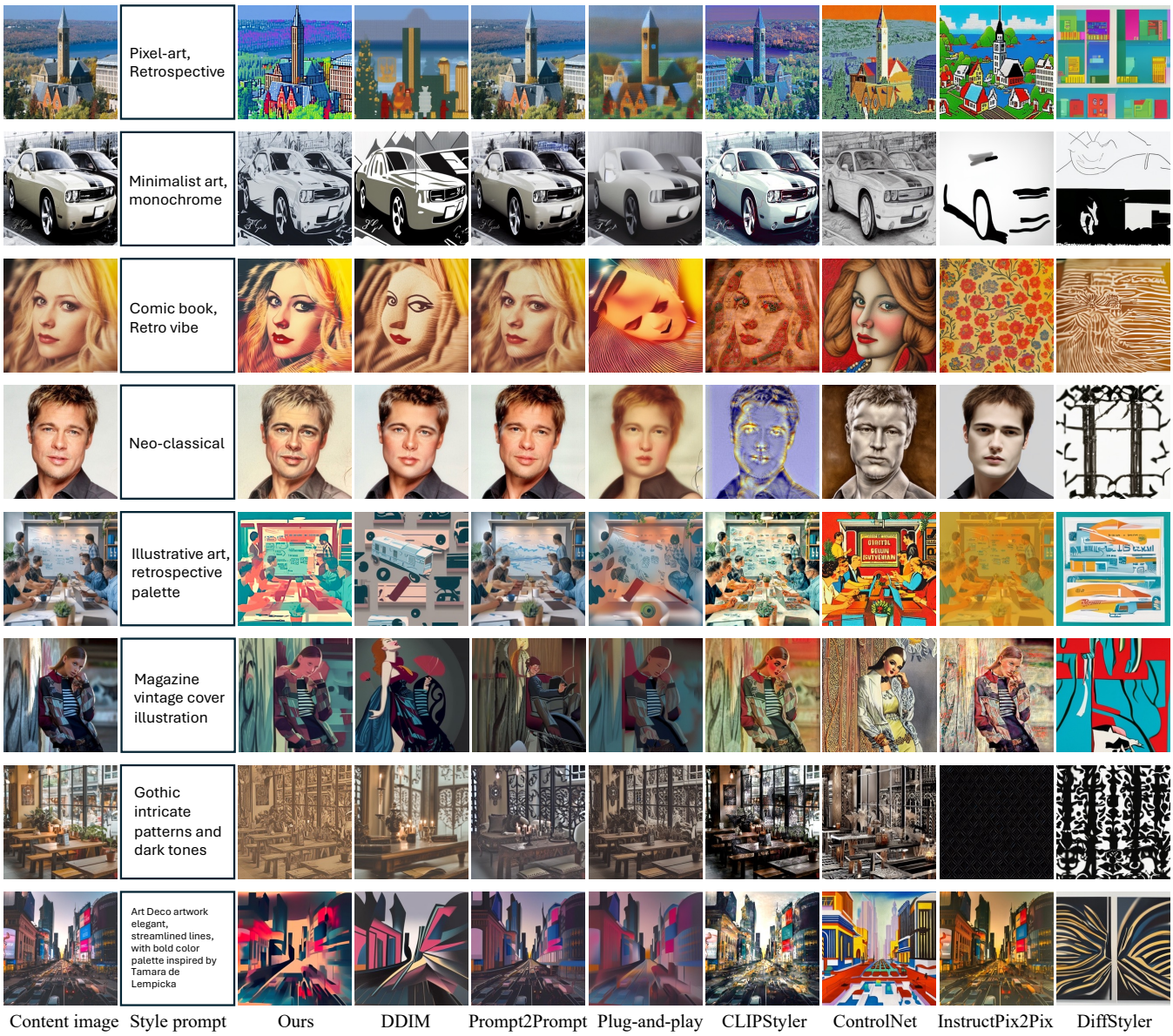

Figure 17: Extended comparison with existing text-driven stylization and manipulation methods.

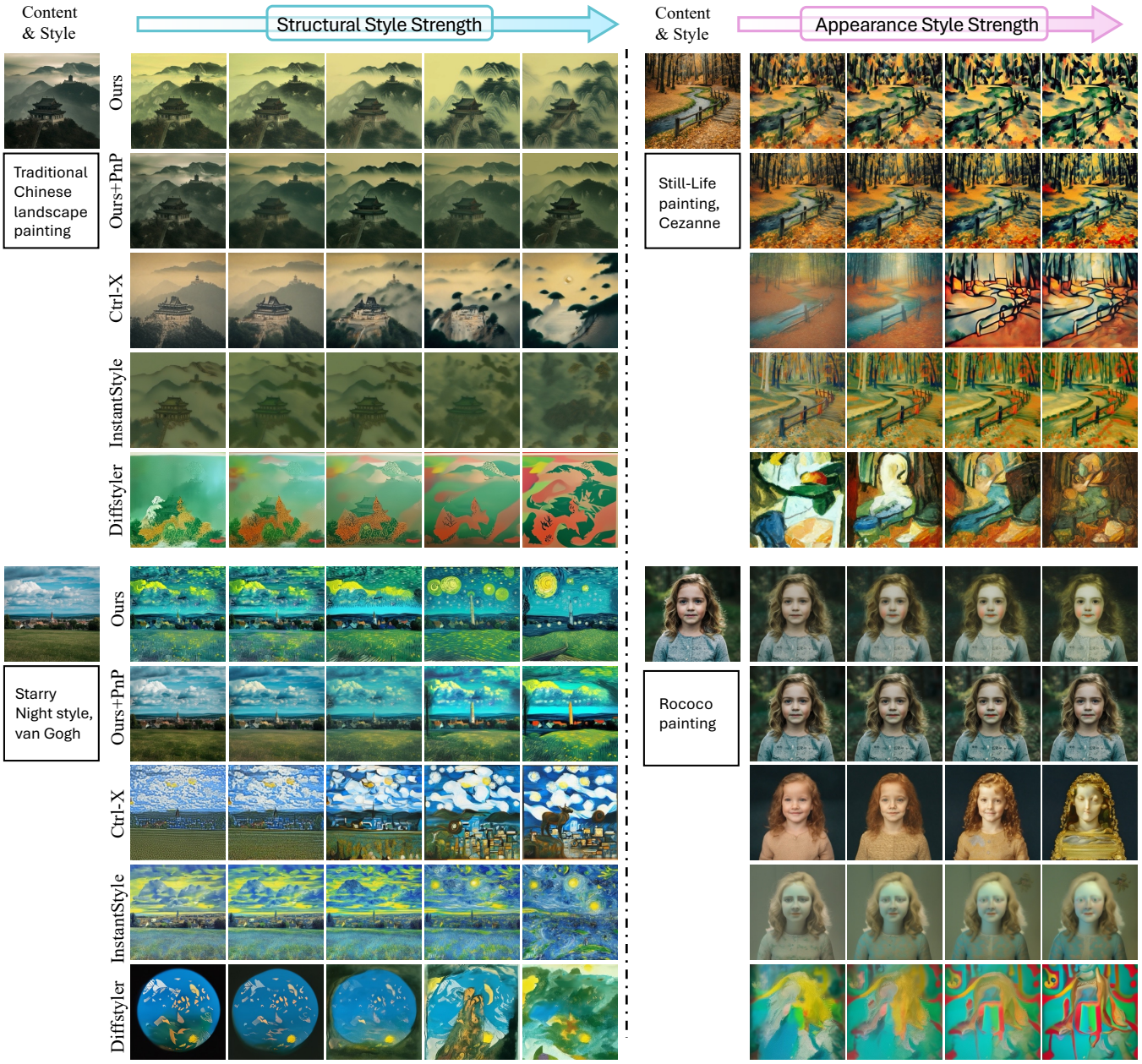

Figure 18: Extended comparison on fine-grained structural and appearance-based style control

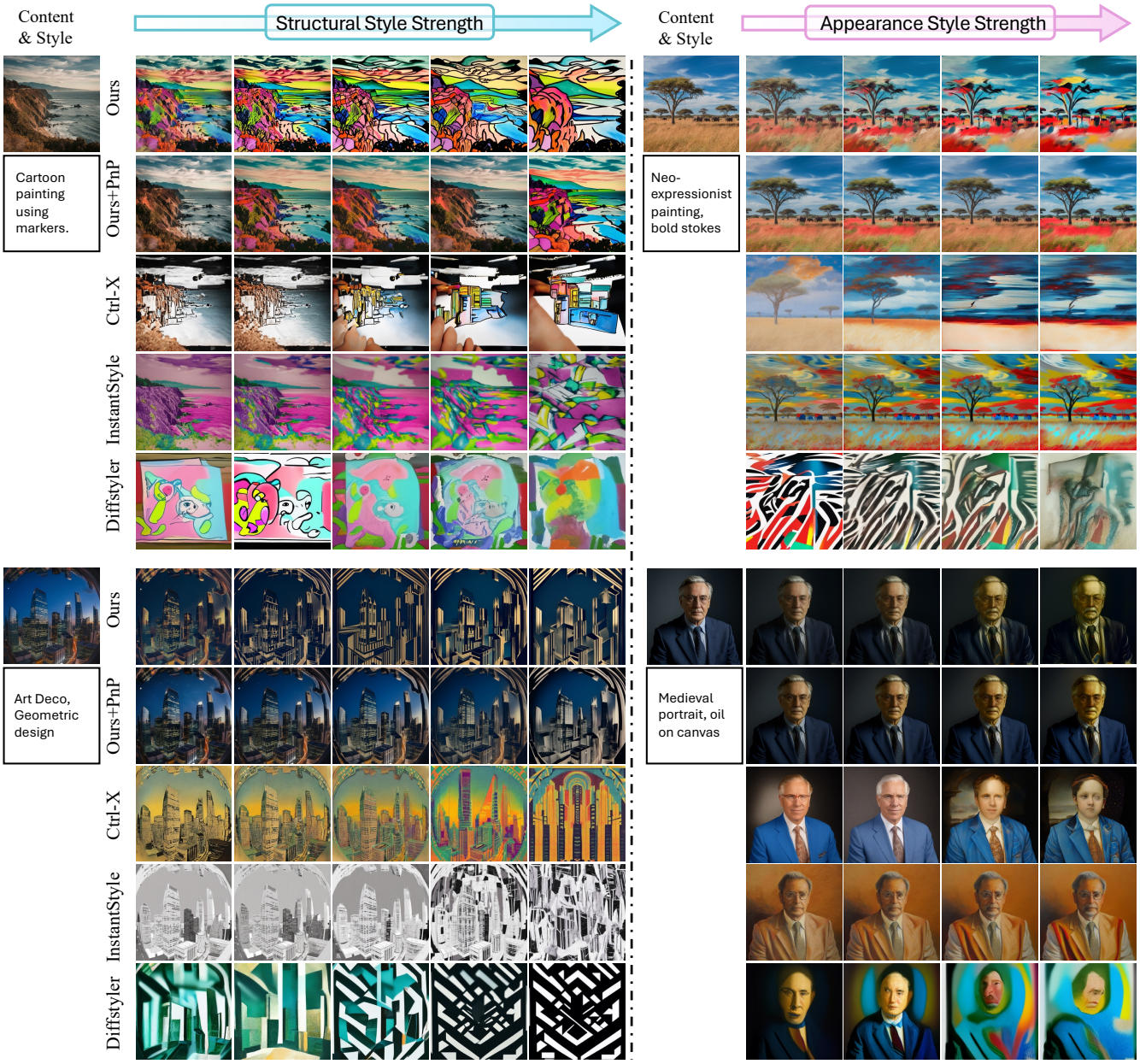

Figure 19: Extended comparison on fine-grained structural and appearance-based style control

Table 6: Prompt templates for MLLM-based metrics. [IMG], [STYLE] and [NUM_METHOD] is the placeholder for combined image, style and number of methods, respectively.

Structure Score	Appearance Score
"[IMAGE] A content (source) image (top left) and [NUM_METHOD] stylized images in the style of [STYLE] are placed as a grid. " "The stylized images are indexed from left to right, and from top to bottom. " "Compare, analyze and discriminately rank the fidelity to which the structure described in the style of [STYLE] is transferred to the source image." "You should focus on the fidelity of structure-related style component only, such as the lines, shapes, geometry, layout, and perspective. You should not consider the style related to appearance (e.g., texture, color, stroke, and pattern). You should also consider how the structure of [STYLE] is integrated with the source image." "Stylized image that has (1) limited style strength, (2) structure that is mis-aligned with the style, or (3) significant artifacts and distortions of the semantic integrity (e.g., the original object and scene become unrecognizable) unless the distortion is explicitly intended by the style of [STYLE], and (4) un-harmonious integration with the source image should be considered of in lower rank. In other words, if a stylized image is not an artistically meaningful painting of the source image in target style, then it should be rated lower. Images that harmoniously integrate the structure of [STYLE] with the source image should be rated higher." "Rank the [NUM_METHOD] images in ascending order from 1 to [NUM_METHOD], where the highest rank of [NUM_METHOD] means the best structural fidelity. No images shall have the same ranking. " "As an expert in art, return your thinking in short (what structure is desired, and how the ranking is decided for each image in short), and ranks for each image id in a Python Dict, ['thinking':str, 'rank':List[[NUM_METHOD]]]. Do not include any other string in your response."	"[IMAGE] A content (source) image (top left) and [NUM_METHOD] stylized images in the style of [STYLE] are placed as a grid. " "The stylized images are indexed from left to right, and from top to bottom. " "Compare, analyze and discriminately rank the fidelity to which the appearance described in the style of [STYLE] is transferred and to the source image." "You should focus on the fidelity of appearance-related art style component only, such as the texture, color, stroke, and pattern. Note that it does not simply means color pallette and saturation. You should not consider the style related to structure (e.g., lines, shapes, geometry, layout, and perspective), unless the original scene become unrecognizable. You should also consider how the appearance of [STYLE] is integrated with the source image." "Stylized image that has (1) limited style strength, (2) visual appearance that is mis-aligned with the style, (3) significant artifacts and distortions of the semantic integrity (e.g., the original object and scene become unrecognizable) unless the distortion is explicitly intended by the style of [STYLE] and (4) un-harmonious integration with the source image should be considered of in lower rank. In other words, if a stylized image is not an artistically meaningful painting of the source image in target style, then it should be rated lower. Images that harmoniously integrate texture, color, stroke, and pattern should be rated higher." "Rank the [NUM_METHOD] images in ascending order from 1 to [NUM_METHOD], where the highest rank of [NUM_METHOD] means the best appearance fidelity. No images shall have the same ranking." "As an expert in art, return your thinking (what appearance is desired, and how the ranking is decided for each image in short), and ranks for each image id in a Python Dict, ['thinking':str, 'rank':List[[NUM_METHOD]]]. Do not include any other string in your response."

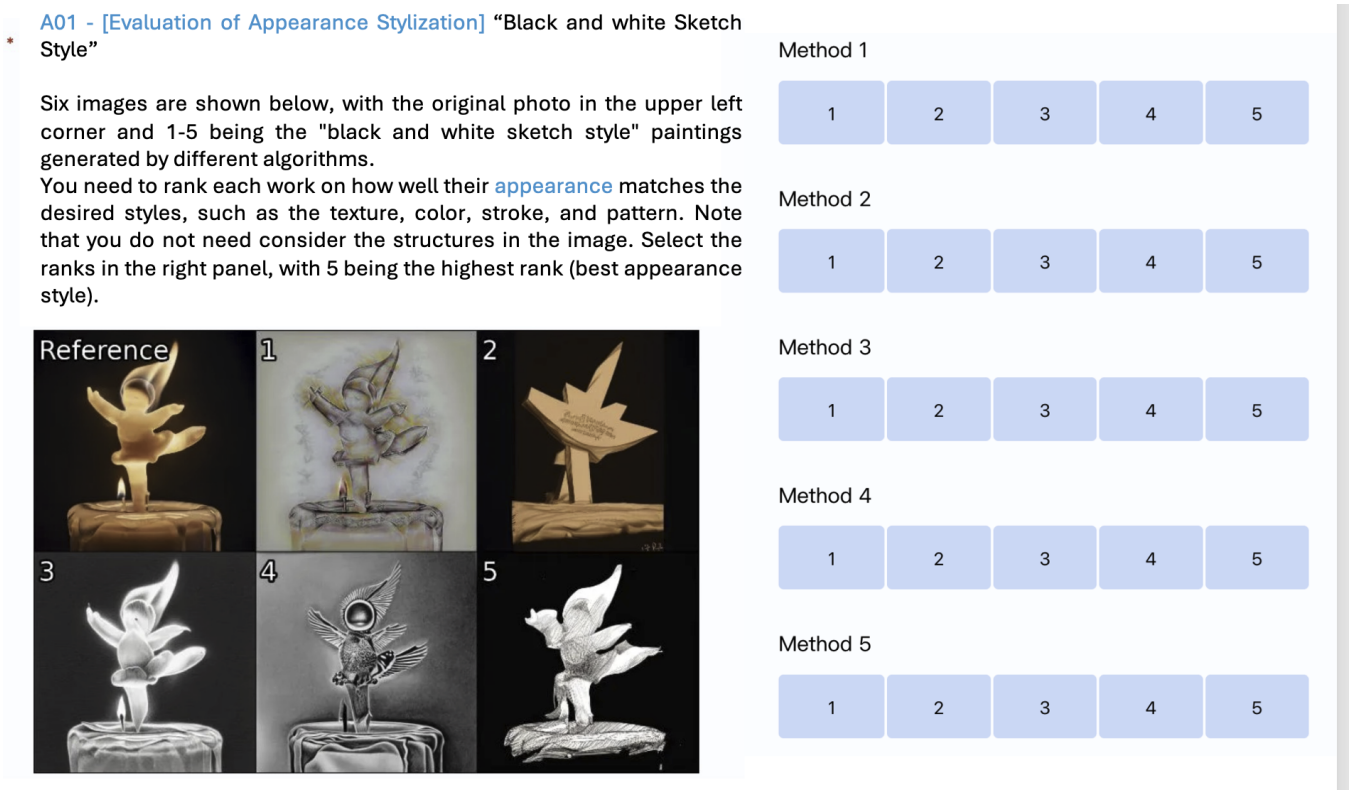

Figure 20: Example user interface in collecting human preference. The system will prevent user from selecting the same ranking.

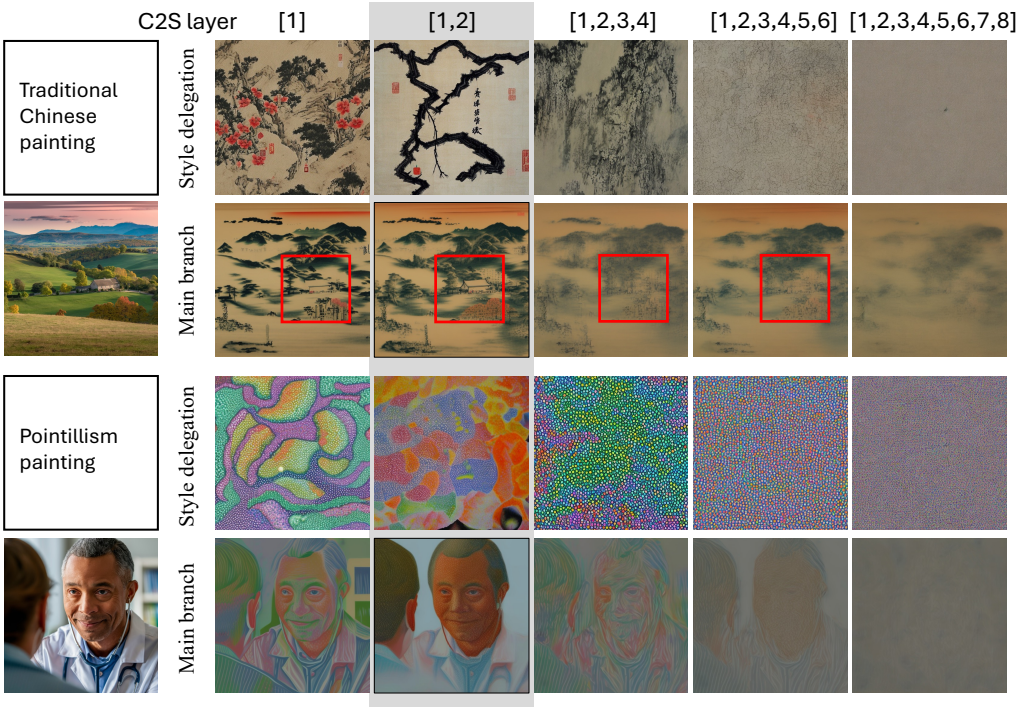

Figure 21: Ablation Study on S2A Layers S_{S2A}. Increasing the number of S2A layers compels the appearance delegation to generate higher-frequency style features (stokes, points) while diminishing low-frequency tonal components (large color fields). Empirically, our default configuration $[1, 2]$ strikes an optimal balance between enhancing style detail and preserving essential content structure.