FREESTYLE: An Anchor-Free Mechanism for Training-Free Style-Aligned Image Generation

Minseok Oh* [1] Jihun Park* [1] Jongmin Gim [1] Minwoo Choi [1]
Kyoungmin Lee [1] Ferdinando Fioretto† [2] Sunghoon Im† [1]
[1]DGIST, Republic of Korea [2] University of Virginia, USA
{harrymark0, pjh2857, jongmin4422, subminu, kyoungmin, sunghoonim}@dgist.ac.kr
fioretto@virginia.edu

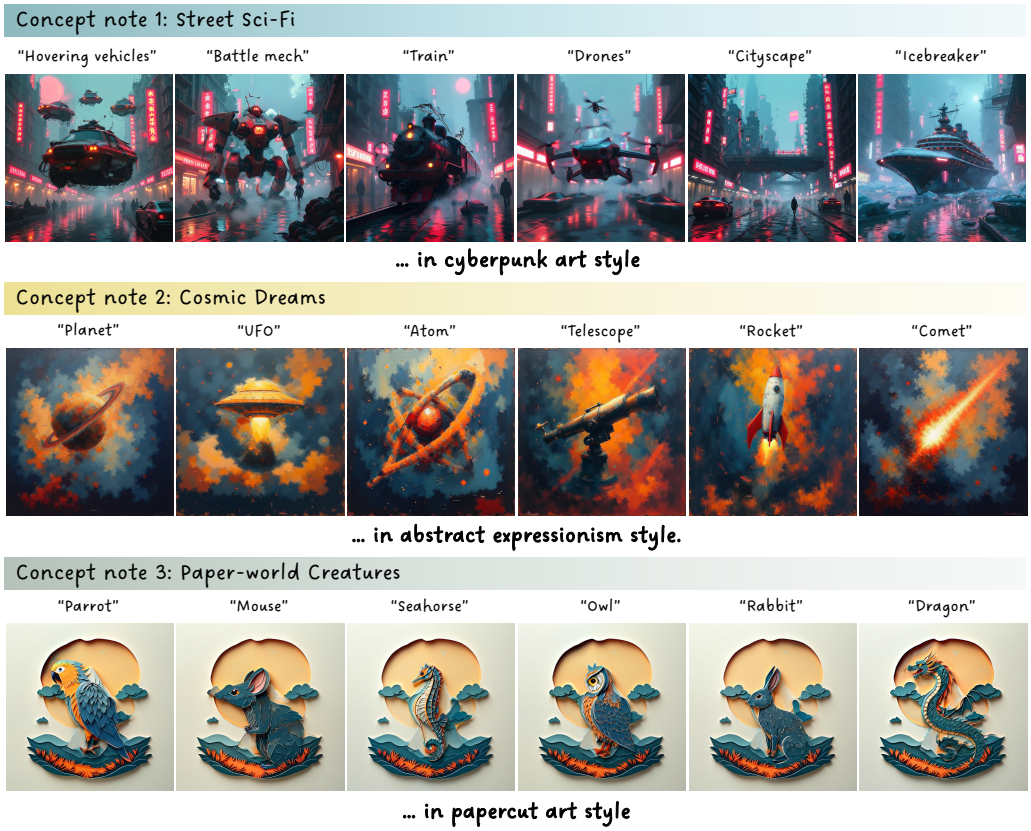

Figure 1. FREESTYLE generates images that maintain consistent style across varied objects given a shared style condition

Abstract

Text-to-Image (T2I) generation models have become central to creative workflows, where producing style-consistent image sets is crucial for applications such as visual identity design, illustration, and asset creation. While recent training-free methods have enabled efficient style-aligned synthesis, they commonly rely on ***anchor-dependent propagation****, where style features from a single reference image are shared across the batch. This dependency of-*

*Equal contribution.
†Co-corresponding authors.

*ten leads to inconsistent results: when the anchor fails to capture the intended reference style, its unintended artifacts are propagated to all samples, and excessive feature sharing further causes content leakage and semantic distortion. To overcome these limitations, we propose **FREESTYLE**, a training-free and anchor-free framework for robust style alignment within a scale-wise autoregressive generation paradigm. Our method integrates three core components—**Majority Voting (MV)**, which aggregates dominant style cues across the batch to form a representative style feature; **Majority Style Injection (MSI)**, which adaptively injects these aggregated features to enforce global style coherence; and **Set-Based Projection (SBP)**, which refines local object regions by projecting them onto a shared style manifold for context-aware adaptation. Extensive experiments demonstrate that, without any retraining or parameter updates, **FREESTYLE** achieves state-of-the-art performance in both style consistency and content preservation, while maintaining real-time inference efficiency.*

1. Introduction

Text-to-Image (T2I) generation models [7, 9, 11, 20, 21, 27, 29, 38] have emerged as a cornerstone of modern creative production. Owing to their accessibility and seamless integration into open conversational AI systems, they have rapidly expanded across a wide range of design and content creation domains. While these models excel at generating novel and diverse imagery, practical creative workflows increasingly demand *style-aligned* synthesis from user-specified references, as well as the ability to produce image sets that maintain consistent stylistic identity for applications such as logos, posters, and game assets. However, these settings introduce a fundamental challenge: *achieving cross-sample style coherence while simultaneously preserving faithful adherence to the textual semantics*.

To address this challenge, several works have introduced fine-tuning approaches incorporating lightweight adapters, such as LoRA [15], into large-scale T2I models [8, 32, 36, 44]. While effective at adapting styles from reference inputs, these methods require additional training costs and thus remain impractical for real-world applications. Consequently, recent studies [13, 22, 25, 26, 45] have shifted toward training-free paradigms that preserve the pre-trained model knowledge while enabling fast style-aligned generation.

Despite their effectiveness, diffusion-based approaches such as AlignedGen [45] and StyleAligned [13] inherently suffer from long inference times due to the iterative denoising process of diffusion models. In contrast, [25] employs a scale-wise autoregressive architecture [11, 39], which progressively refines features from coarse to fine scales, enabling significantly faster synthesis while maintaining competitive style consistency. This makes the autoregressive paradigm particularly appealing for training-free style alignment, where real-time or large-batch generation is essential.

Nevertheless, existing training-free approaches still suffer from the *anchor-dependency* problem: style cues extracted from a single (first) sample are shared with the rest of the batch. When this anchor fails to accurately reflect the intended target style, its unintended stylistic attributes are propagated to other images, yielding batch-wide inconsistency and artifacts (Fig. 2-second row). In addition, such global sharing can induce *content leakage*, where object semantics are inadvertently distorted by over-shared style features, as shown in Fig. 2-first row.

To address these limitations, we introduce *FREESTYLE*, a training-free framework that achieves fast, robust, and anchor-free style alignment within the scale-wise autoregressive generation paradigm. Our method identifies representative style information across the batch while preserving object integrity through three complementary mechanisms: (i) *Majority Voting (MV)*, which aggregates dominant style cues across samples to construct a representative style feature; (ii) *Majority Style Injection (MSI)*, which adaptively injects these aggregated features through spatially aware blending to enforce global coherence; and (iii) *Set-Based Projection (SBP)*, which refines object regions by projecting them onto a shared style manifold, ensuring localized style consistency without semantic distortion. Together, these components enable *FREESTYLE* to achieve state-of-the-art style-aligned generation performance while fully preserving the high-quality generation capability of the base model—without any additional training or modification to pretrained parameters.

In summary, our contributions are as follows:

- We present *FREESTYLE*, a *training-free, anchor-free* framework for style-aligned image generation within a scale-wise autoregressive paradigm, achieving state-of-the-art performance in both style consistency and object relevancy.
- We introduce *Majority Voting (MV)*, a batch-level mechanism that extracts representative style features by aggregating dominant cues across samples.
- We design two complementary modules—*Majority Style Injection (MSI)* and *Set-Based Projection (SBP)*—that collaboratively enforce global style coherence and local object-aware adaptation.

2. Related works

Text-to-image generation. Recent advances in text-to-image (T2I) generation have been driven by large-scale text–image datasets [2, 34] and rapid progress in generative modeling. Early GAN-based approaches [10, 18] enabled

Figure 2. Qualitative comparison between anchor dependent model StyleAR [42] (red box) and our method (green box). **Top row**: StyleAR enforces global style consistency, but the first sample acts as an anchor, causing content leakage from the anchor. **Bottom row**: When the anchor fails to realize the target style (e.g. watercolor painting), later samples inherit this failure. In contrast, our method maintains style consistency while minimizing cross-sample content leakage and expressing the target style across all samples.

conditional image synthesis but often suffered from training instability and limited text–image alignment. Subsequently, diffusion-based models and flow matching models [7, 14, 20, 21, 27, 29, 33] became the dominant paradigm, achieving superior visual and text fidelity through iterative denoising. These models have also been widely extended to tasks such as image editing, style transfer, and personalized generation [1, 12, 17, 19, 31, 40], but their slow inference remains a major bottleneck. To improve efficiency, scale-wise autoregressive approaches [11, 38, 39] have emerged, modeling discrete latent tokens via next-scale prediction to achieve faster synthesis while maintaining competitive image quality. Despite these advancements, both diffusion and scale-wise autoregressive models still struggle with style alignment and content fidelity, which are essential for practical creative workflows.

Style-Aligned image generation. Style-aligned image generation aims to produce image sets that share a unified visual style while preserving content fidelity and textual alignment. Early approaches addressed this through *fine-tuning*-based strategies [8, 23, 30, 32, 35, 36, 44], which extend personalized generation pipelines with lightweight adapters, such as LoRA [15], to encode reference style information. While effective, these methods require further optimization and model updates, which limit their practicality in real-time scenarios. To reduce training costs, recent *training-free* methods align style directly during inference while preserving the knowledge of the pretrained model. Diffusion-based frameworks, such as AlignedGen [45] and StyleAligned [13], achieve cross-sample consistency by sharing features or attention maps across denoising steps. However, their iterative process leads to high computational overhead and slow inference. To improve efficiency, [25] adopts a *scale-wise autoregressive* architecture [11, 39], enabling fast, progressive generation without fine-tuning. However, most training-free approaches remain *anchor-dependent*, sharing style features from a single reference image across the batch, which propagates style artifacts or causes content leakage when the anchor underrepresents the target style.

3. Preliminaries

Our approach builds upon the architecture of *Infinity* [11], a state-of-the-art text-to-image (T2I) generation model following the *next-scale prediction* paradigm [39]. The model consists of a pretrained text encoder E_T, an autoregressive transformer G for next-scale prediction, and a decoder D that reconstructs the final image from multi-scale feature representations. This progressive, resolution-aware refinement enables efficient, high-fidelity image synthesis while maintaining semantic coherence across scales.

Given a text prompt T, the text encoder produces contextual embeddings $E_T(T)$, which condition the autoregressive generation process. Starting from a start-of-sequence token $\langle \text{SOS} \rangle$ as the initial feature map F_0, the transformer G sequentially predicts residual feature maps of increasing spatial resolution across generation steps $s \in \{1, 2, \ldots, S\}$:

$$r_s = G(F_{s-1}, E_T(T)), \quad r_s \in \mathbb{R}^{C \times h_s \times w_s}, \tag{1}$$

where h_s and w_s denote the height and width of each step s. Each residual feature is then upsampled to a fixed spatial size $H \times W$ to form the high-resolution representation:

$$R_s = \text{up}_{H \times W}(r_s), \quad R_s \in \mathbb{R}^{C \times H \times W}. \tag{2}$$

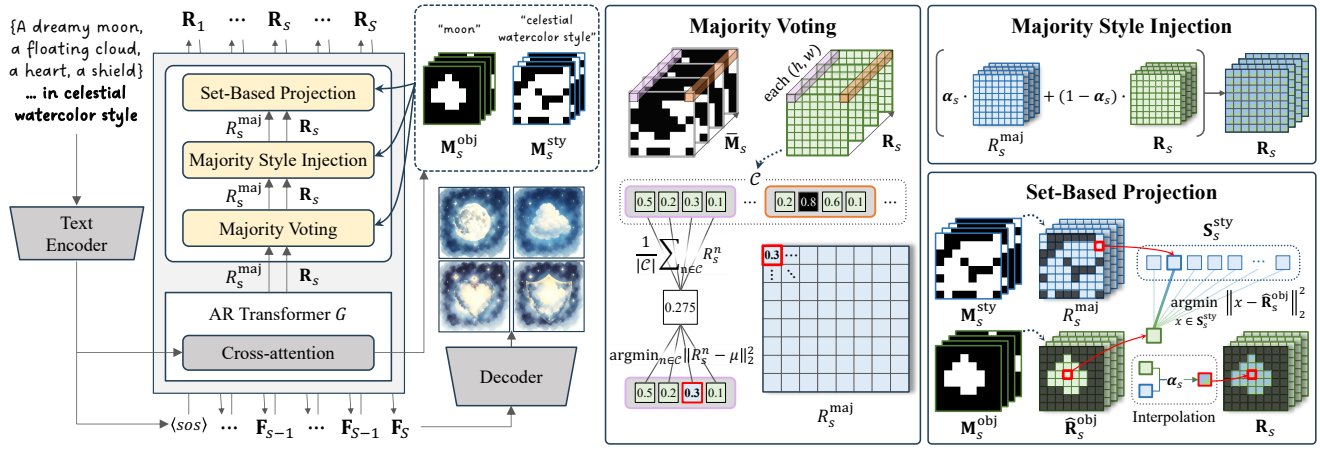

Figure 3. Overall pipeline. The ⟨SOS⟩ token and condition are encoded by the text encoder and fed into the transformer. Cross-attention maps used to derive the style and object masks, $\mathbf{M}_s^{\text{sty}}$ and $\mathbf{M}_s^{\text{obj}}$. *Majority Voting (MV)* then extracts a representative style residual feature R_s^{maj}, which is adaptively injected to unify the overall style through *Majority Style Injection (MSI)*, while *Set-Based Projection (SBP)* refines object regions to reflect style information from the style set.

The upsampled residuals are progressively accumulated to form the aggregated representation; after all scales are processed, the final aggregated feature map F_S encodes multi-scale semantic and structural information, which the decoder D transforms into the final image:

$$F_s = \sum_{i=1}^{s} R_i, \quad I = D(F_S). \tag{3}$$

4. Methods

4.1. Overall Pipeline

Given an N-samples batch of text prompts $\mathbf{T} = \{T^n\}_{n=1}^N$, our goal is to generate corresponding images $\mathbf{I} = \{I^n\}_{n=1}^N$ that maintain a unified visual style while depicting distinct objects. Each prompt T consists of distinct object prompts T^{obj} and a shared style prompt T^{sty}. All prompts are jointly processed as a batch, ensuring that style features can be shared across samples.

As illustrated in Fig. 3, given the encoded text prompts, the model generates latent representations for each sample while synchronously applying our batch-level style alignment modules. At each step, *Majority Voting (MV)* aggregates the most representative style patterns across the batch to form a style dominant residual feature R_s^{maj} (Sec. 4.2). This feature is then propagated through *Majority Style Injection (MSI)*, which enforces global style coherence by adaptively blending R_s^{maj} into each sample's latent features according to their style–object attention balance (Sec. 4.3). However, because MSI performs spatially uniform blending, it can overlook subtle, object-specific variations. To address this, *Set-Based Projection (SBP)* refines object features by projecting them onto a shared set of representative style exemplars, enriching under-styled or inconsistent regions with contextually compatible cues (Sec. 4.4).

4.2. Majority Voting (MV)

To mitigate the dependency on a single (anchor) sample, we extract the dominant style feature across the batch using a *Majority Voting (MV)* mechanism. The goal is to identify representative style cues that are consistently expressed among samples, rather than relying on the style of any specific sample. At each generation step s, the method operates on residual features $\mathbf{R}_s = \{R_s^n\}_{n=1}^N$ and utilizes attention masks defined for style and object tokens $(\mathbf{M}_s^{\text{sty}}, \mathbf{M}_s^{\text{obj}})$ to localize their respective influences. We first compute the cross-attention maps $\mathbf{H}_s^{\text{sty}}$ and $\mathbf{H}_s^{\text{obj}}$ for the style and object tokens, respectively. These maps are derived using the image query $\mathbf{Q}_s$ and the key embeddings $(\mathbf{K}^{\text{sty}}, \mathbf{K}^{\text{obj}})$, which are obtained from the corresponding token representations $\mathbf{T}^{\text{sty}}$ and $\mathbf{T}^{\text{obj}}$ as follows:

$$\begin{aligned}
\mathbf{H}_s^{\text{sty}} &= \text{Softmax}\left(\frac{\mathbf{Q}_s \mathbf{K}^{\text{sty}\top}}{\sqrt{d}}\right), \ \mathbf{H}_s^{\text{sty}} \in \mathbb{R}^{N\times H\times W}, \\
\mathbf{H}_s^{\text{obj}} &= \text{Softmax}\left(\frac{\mathbf{Q}_s \mathbf{K}^{\text{obj}\top}}{\sqrt{d}}\right), \ \mathbf{H}_s^{\text{obj}} \in \mathbb{R}^{N\times H\times W},
\end{aligned} \tag{4}$$

where d denotes the dimension of the feature. Then, binary masks $\mathbf{M}_s^{\text{sty}}$ and $\mathbf{M}_s^{\text{obj}}$ are obtained by thresholding the normalized attention maps:

$$\begin{aligned}
\mathbf{M}_s^{\text{sty}} &= \begin{cases} 1, & \text{if } \mathbf{H}_s^{\text{sty}} > \tau_s^{\text{sty}}, \\ 0, & \text{otherwise}, \end{cases}, \ \mathbf{M}_s^{\text{sty}} \in \mathbb{B}^{N\times H\times W}, \\
\mathbf{M}_s^{\text{obj}} &= \begin{cases} 1, & \text{if } \mathbf{H}_s^{\text{obj}} > \tau_s^{\text{obj}}, \\ 0, & \text{otherwise}, \end{cases}, \ \mathbf{M}_s^{\text{obj}} \in \mathbb{B}^{N\times H\times W},
\end{aligned} \tag{5}$$

Algorithm 1 Majority Voting

Input: Latent residual features $\mathbf{R}_s = \{R_s^1, \cdots, R_s^N\}$, style masks $\mathbf{M}_s^{\text{sty}}$, object masks $\mathbf{M}_s^{\text{obj}}$
Output: Representative style residual feature R_s^{maj}

$\bar{\mathbf{M}}_s \leftarrow \mathbf{M}_s^{\text{sty}} \cap (1 - \mathbf{M}_s^{\text{obj}})$ # Define style-centric region
for each spatial position (h, w) **do**
  $\mathcal{C} = \{\}$ # initialize index set.
  for each batch index n **do**
    if $\bar{\mathbf{M}}_s[n, h, w] = 1$ **then**
      $\mathcal{C} \leftarrow \mathcal{C} \cup \{n\}$ # Collect batch indices where style-centric region is active at (h, w).
    end if
  end for
  if $|\mathcal{C}| = 0$ **then**
    $\mathcal{C} \leftarrow \{1, \ldots, N\}$ # Fallback: use all samples if no style-centric region exists.
  end if
  $\mu \leftarrow \frac{1}{|\mathcal{C}|} \sum_{n \in \mathcal{C}} R_s^n[h, w]$
  $n^* \leftarrow \arg\min_{n \in \mathcal{C}} \|R_s^n[h, w] - \mu\|_2^2$
  $R_s^{\text{maj}}[h, w] \leftarrow R_s^{n^*}[h, w]$
end for
M_s^{obj}, $M_s^{\text{sty}} \leftarrow$ Union over all N samples of $\mathbf{M}_s^{\text{obj}}$, $\mathbf{M}_s^{\text{sty}}$
$R_s^{\text{obj}} \leftarrow R_s^{\text{maj}} \odot M_s^{\text{obj}}$, $R_s^{\text{sty}} \leftarrow R_s^{\text{maj}} \odot M_s^{\text{sty}}$
$R_s^{\text{ada}} \leftarrow \text{AdaIN}(R_s^{\text{obj}}, R_s^{\text{sty}})$ [16]
$R_s^{\text{maj}} \leftarrow R_s^{\text{ada}} \odot (M_s^{\text{obj}}) + R_s^{\text{maj}} \odot (1 - M_s^{\text{obj}})$
return R_s^{maj}

where τ_s^{sty} and τ_s^{obj} are determined leveraging Otsu's thresholding method [24]. Using these masks, the majority voting algorithm described in Alg. 1 compares latent features across all samples at each spatial position $(h, w) \in [1, H] \times [1, W]$ using the masks and selects the one that best represents the batch-wise majority style. This process yields a single representative style residual feature, R_s^{maj}.

4.3. Majority Style Injection (MSI)

Once the representative style residual feature R_s^{maj} is derived through Majority Voting (MV), we propagate it back to each sample to enforce global style consistency while preserving object fidelity. This process, termed *Majority Style Injection (MSI)*, adaptively blends the shared style feature into each sample's features. Specifically, at each generation step s, the residual features $\mathbf{R}_s$ are updated as follows:

$$\mathbf{R}_s \leftarrow (1 - \boldsymbol{\alpha}_s)\mathbf{R}_s + \boldsymbol{\alpha}_s R_s^{\text{maj}}, \quad \mathbf{R}_s \in \mathbb{R}^{N \times C \times H \times W} \tag{6}$$

Here, R_s^{maj} is broadcast across the batch dimension, while $\boldsymbol{\alpha}_s$ is broadcast along the channel dimension.

The adaptive blending weight $\boldsymbol{\alpha}_s \in \mathbb{R}^{N \times H \times W}$ is computed based on the relative strength of style and object attention maps as follows:

$$\boldsymbol{\alpha}_s = \begin{cases} \dfrac{\mathbf{H}_s^{\text{sty}}}{\mathbf{H}_s^{\text{sty}} + \mathbf{H}_s^{\text{obj}} + \epsilon}, & \mathbf{M}_s^{\text{obj}} = 1, \\ 1, & \mathbf{M}_s^{\text{obj}} = 0, \end{cases}, \tag{7}$$

where ϵ denotes a small constant for numerical stability. The adaptive weight set $\boldsymbol{\alpha}_s$ controls the balance between style and content: it preserves geometry in object-focused regions and strengthens style coherence elsewhere.

4.4. Set-Based Projection (SBP)

While Majority Style Injection (MSI) promotes global stylistic harmony across the batch, its spatially uniform blending restricts adaptation to object-specific variations, often failing to inject contextually appropriate styles for each object. To address this limitation, we introduce a *Set-Based Projection (SBP)* mechanism that adaptively refines object features by referencing a shared set of representative style exemplars.

Inspired by [5], which recast diffusion sampling as updates with projections onto constraint sets, we construct a *representative style set* $\mathbf{S}_s^{\text{sty}}$ that encapsulates style features from the style-dominant regions of R_s^{maj}:

$$\mathbf{S}_s^{\text{sty}} = \{R_s^{\text{maj}}[h, w] \mid M_s^{\text{sty}}[h, w] = 1\}, \tag{8}$$

where M_s^{sty} is the union over all N samples of $\mathbf{M}_s^{\text{sty}}$, and each element $R_s^{\text{maj}}[h, w]$ corresponds to a local feature vector of R_s^{maj} from spatial position $(h, w) \in [1, H] \times [1, W]$. This set serves as a compact and reusable memory of style exemplars shared across the batch.

For each sample, within the object regions $\mathbf{M}_s^{\text{obj}}$, we extract the local object residual features as follows:

$$\hat{\mathbf{R}}_s^{\text{obj}} = \mathbf{R}_s \odot \mathbf{M}_s^{\text{obj}}, \tag{9}$$

where $\odot$ denotes the Hadamard Product. We then project them toward the nearest representative style vector in $\mathbf{S}_s^{\text{sty}}$. To seamlessly blend the representative style into each local object feature, we identify the most similar style exemplar $\Pi_{\text{SBP}} \in \mathbb{R}^{N \times H \times W}$ in the feature space as:

$$\Pi_{\text{SBP}} = \underset{x \in \mathbf{S}_s^{\text{sty}}}{\text{argmin}} \left\| x - \hat{\mathbf{R}}_s^{\text{obj}}[h, w] \right\|_2^2, \quad \forall\, h \in \{1, \ldots, H\},\ w \in \{1, \ldots, W\}. \tag{10}$$

We then interpolate the projected feature map Π_{SBP} with the local object residual features $\hat{\mathbf{R}}_s^{\text{obj}}$ to obtain the final style-aligned residual feature $\mathbf{R}_s$:

$$\mathbf{R}_s \leftarrow (1 - \boldsymbol{\alpha}_s)\hat{\mathbf{R}}_s^{\text{obj}} + \boldsymbol{\alpha}_s \Pi_{\text{SBP}}, \tag{11}$$

where $\hat{\mathbf{R}}_s^{\text{obj}} \in \mathbb{R}^{N \times C \times H \times W}$, and $\boldsymbol{\alpha}_s$ (defined in Equ. (7)) is broadcast along the channel dimension. This projection-based refinement preserves object-specific characteristics while enforcing coherent and context-aware style integration across the batch.

5. Experiments

5.1. Implementation details

We build our framework upon the pretrained *Infinity 2B* model [11], which follows a scale-wise autoregressive generation paradigm. The model predicts residual feature maps over 12 steps and employs a quantized codebook with 2^{32} entries, producing latent feature maps of spatial resolution 64×64 with 32 channels. All components of the original architecture remain unchanged, and no additional training or parameter updates are performed during inference.

Our proposed modules, *Majority Voting (MV)*, *Majority Style Injection (MSI)*, and *Set-Based Projection (SBP)*, are applied throughout all generation steps. We conduct all experiments using a single NVIDIA A6000 GPU. Generating four 1024×1024 images in parallel requires approximately 7.92 seconds in total (1.98 seconds per image).

5.2. Evaluation Setup

Benchmark. Following the evaluation protocol of StyleAligned [13], we evaluate our method using the same benchmark consisting of 100 style–content prompt groups generated by ChatGPT. Each group includes one shared *style prompt* and four distinct *content prompts*, resulting in a total of 400 generated images per evaluation.

Evaluation Metrics. We employ four complementary metrics to comprehensively assess the model's performance: (1) **Object relevancy** (S_{obj}) measures the alignment between each generated image and its corresponding *content prompt*, computed as the cosine similarity between CLIP image and text embeddings [28]. (2) **DINO-based style consistency** (S_{DINO}) quantifies the stylistic uniformity across the generated images by calculating the average pairwise cosine similarity of DINO ViT-B/8 [4] embeddings within each batch. (3) **Whole-prompt relevancy** (S_{whole}) measures how well each image reflects the overall concept described by concatenating the content and style prompts, using CLIP image–text similarity. (4) **CLIP-based style consistency** (S_{CLIP}) measures the overall visual coherence within each generated batch by computing the average pairwise cosine similarity among CLIP image embeddings.

In addition, we report the **harmonic score** (S_{harmonic}) to provide an integrated evaluation that jointly considers content preservation, style coherence, and overall adherence to the intended style. It is computed as the harmonic mean of all four metrics—S_{obj}, S_{DINO}, S_{whole}, and S_{CLIP}:

$$S_{\text{harmonic}} = \frac{4}{\frac{1}{S_{\text{obj}}} + \frac{1}{S_{\text{DINO}}} + \frac{1}{S_{\text{whole}}} + \frac{1}{S_{\text{CLIP}}}}. \tag{12}$$

This unified metric provides a holistic assessment of text–image alignment and stylistic consistency, offering a fair and balanced comparison of overall generation quality across methods.

5.3. Comparison with state-of-the-art models

To demonstrate the effectiveness and superiority of our approach, we conduct comprehensive quantitative and qualitative comparisons against state-of-the-art style-aligned image generation models, including training-based methods (StyleDrop [36], IP-Adapter [44], B-Lora [8], CSGO [43], StyleAR [42]), training-free methods (StyleAligned [13], AlignedGen [45]), and baseline models (Infinity [11], SDXL [27], FLUX [20]).

Quantitative comparison. As shown in Tab. 1, our model achieves the highest overall performance across all evaluation metrics, demonstrating superior style coherence and text alignment while maintaining efficient inference. In particular, our method attains the best **style consistency** scores in both DINO-based (S_{DINO}) and CLIP-based (S_{CLIP}) measures, indicating that *FREESTYLE* effectively preserves stylistic uniformity across generated samples. Furthermore, our model also achieves competitive **object relevancy** (S_{obj}) and **whole-prompt relevancy** (S_{whole}) scores, reflecting faithful adherence to both content and style prompts. As a result, the overall **harmonic score** (S_{harmonic}), which jointly captures content preservation, style coherence, and global adherence, shows that *FREESTYLE* outperforms all training-based and training-free baselines.

Notably, while the training-based StyleAR [8] exhibits performance closest to ours, it requires additional fine-tuning with external datasets and suffers from significantly longer inference time (335.23 seconds per image), making it unsuitable for real-time or interactive generation scenarios. In contrast, our method operates fully *training-free* and achieves high-quality, style-consistent synthesis in just 1.98 seconds per image, comparable to the vanilla Infinity baseline, highlighting its effectiveness and practicality for real-world applications.

Qualitative comparison. Fig. 4 qualitatively compares images generated from prompt sets that share a common style prompt but contain distinct object prompts. We compare our method with the four best baselines by overall scores in Tab. 1. Overall, most baselines align the global style to some extent; yet, they often struggle to accurately realize the intended style or to generate all target objects. In the *psychedelic art style* examples, both StyleAR and IP-Adapter exhibit strong anchor dependency; trees from the reference image repeatedly appear across all generated samples, revealing poor prompt fidelity, and severe anchor dependency. In the *retro comic book style* case, StyleAR, StyleAligned, and AlignedGen fail to generate the intended *lightning bolt* object, indicating limited adherence to the object prompt. Furthermore, AlignedGen fails to reproduce the intended *papercut art style*, generating overly smooth

Table 1. Quantitative comparison with state-of-the-art style-aligned image generation models. We evaluate each method using four complementary metrics: object relevancy (S_{obj}, CLIP text-image similarity with content prompts), DINO-based style consistency (S_{DINO}, pairwise DINO embedding similarity), whole-prompt relevancy (S_{whole}, CLIP similarity with concatenated content–style prompts), and CLIP-based style consistency (S_{CLIP}, pairwise CLIP image similarity). We also report the harmonic score (S_{harmonic}), which provides an overall assessment of style coherence, content preservation, and global adherence. Inference time is measured per image. The symbols ↑ and ↓ indicate that higher or lower values are better, respectively.

Method	Train-Free	S_{harmonic} ↑	S_{obj} ↑	S_{DINO} ↑	S_{whole} ↑	S_{CLIP} ↑	Inference Time (s) ↓
Vanilla Infinity [11]	-	0.346	0.296	0.277	0.328	0.659	1.62
Vanilla SDXL [27]	-	0.370	0.288	0.341	0.341	0.676	9.51
Vanilla FLUX [20]	-	0.361	0.291	0.323	0.329	0.667	34.67
StyleDrop [36]	✗	0.365	0.267	0.399	0.304	0.704	544.06
IP-Adapter [44]	✗	0.405	0.278	0.529	0.324	0.772	<u>10.14</u>
B-Lora [8]	✗	0.323	0.259	0.271	0.308	0.625	633.20
CSGO [43]	✗	0.399	<u>0.282</u>	0.494	0.328	0.715	14.80
StyleAR [42]	✗	<u>0.414</u>	0.281	<u>0.559</u>	0.330	<u>0.772</u>	335.23
StyleAligned [13]	✓	0.409	0.281	0.530	<u>0.331</u>	0.762	11.25
AlignedGen [45]	✓	0.397	0.278	0.503	0.316	0.760	45.24
Ours	✓	**0.421**	**0.284**	**0.589**	**0.332**	**0.791**	**1.98**

Figure 4. Qualitative comparison with state-of-the-art style-aligned image generation models.

textures that deviate from the layered aesthetic, while also exhibiting uniform style propagation across all objects, a sign of anchor dependency. In contrast, our method produces images that exhibit consistent and faithful style alignment across samples while accurately preserving object identity, demonstrating strong agreement with the quantitative improvements in both style and object relevance metrics.

5.4. Ablation studies

Quantitative analysis. The quantitative results in Tab. 2 demonstrate the contribution of each proposed component across all evaluation metrics, including object relevancy (S_{obj}), DINO-based style consistency (S_{DINO}), whole-prompt relevancy (S_{whole}), CLIP-based style consistency (S_{CLIP}), and the overall harmonic score (S_{harmonic}). As

Table 2. Quantitative ablation study on *Majority Style Injection (MSI)* and *Set-Based Projection (SBP)*. The symbol ↑ indicates that higher values are better. **Bold** and underlined values denote the best and second-best results, respectively.

#	Component MSI	Component SBP	Quantitative Metrics S_{harmonic} ↑	S_{obj} ↑	S_{DINO} ↑	S_{whole} ↑	S_{CLIP} ↑
(a)			0.346	**0.296**	0.277	0.328	0.659
(b)	✓		0.414	0.285	0.535	**0.334**	0.771
(c)	✓	✓	**0.421**	0.284	**0.589**	0.332	**0.791**

Figure 5. Qualitative ablation studies on *Majority Style Injection (MSI)*, *Set-Based Projection (SBP)*. (a) - (c) correspond to the component in Tab. 2.

shown in Tab. 2-(b), *Majority Style Injection (MSI)* significantly enhances S_{DINO} and S_{CLIP}, indicating improved batch-level style consistency and visual coherence through adaptive blending of dominant style features. It also slightly improves S_{whole}, confirming better alignment with the joint style–content prompt. When incorporating *Set-Based Projection (SBP)* in Tab. 2-(c), further gains are observed in S_{DINO} and S_{harmonic}, reflecting enhanced contextual style adaptation within object regions while maintaining comparable S_{obj} and S_{whole} scores. Although the quantitative margin between MSI and SBP is moderate, their combined use achieves the highest overall performance across all metrics, validating that MSI enforces global style alignment, while SBP refines local, object-aware coherence. These results collectively confirm that the proposed components complement each other to achieve consistent style synthesis without compromising content fidelity.

Qualitative results. Fig. 5 presents a qualitative comparison corresponding to the quantitative ablation results. The samples are generated from the prompt set {A coral, A bell, A fish, A water wave, A jellyfish} in an *aquatic logo style*. In Fig. 5-(a), the baseline model produces inconsistent style characteristics across samples, showing noticeable variation in color tone and texture strength. Applying *Majority Style Injection (MSI)* in Fig. 5-(b) improves global style consistency and tonal harmony, consistent with the gains in S_{DINO} and S_{CLIP}. However, some undesired artifacts unrelated to object semantics emerge in certain samples (e.g., the coral), while style cues are insufficiently reflected within detailed object regions (e.g., the jellyfish), suggesting that uniform propagation alone cannot fully capture object-aware style correspondence. By incorporating *Set-Based Projection (SBP)* in Fig. 5-(c), the generated images exhibit coherent and contextually aligned style features across both object and background regions, maintaining semantic clarity and visual balance. This aligns with the improvement in S_{harmonic}, confirming that MSI enforces global consistency and adherence, while SBP complements it through localized, object-aware refinement.

Table 3. User study preference percentages.

Method	Prompt ↑	Style ↑
StyleAR [42]	6.50%	13.75%
StyleAligned [45]	18.75%	15.25%
IP-Adapter [44]	15.50%	6.50%
Ours	**59.25%**	**64.50%**

5.5. User study

To complement the quantitative evaluation, we conducted a user study, as summarized in Tab. 3. Forty participants evaluated two key aspects, prompt fidelity and style consistency, by comparing images generated by our model with those from StyleAR [42], StyleAligned [45], and IP-Adapter [44], the top-performing baselines in our quantitative benchmarks. The results show that our model consistently outperforms all competitors in both criteria, demonstrating strong human-perceived performance across diverse prompts.

6. Conclusion

In this paper, we introduce *FREESTYLE*, a training-free and anchor-free framework for style-aligned image generation within a scale-wise autoregressive paradigm. Our approach addresses the limitations of existing training-free methods, which rely on anchor-dependent propagation and often result in unintended style alignment and content leakage across samples. At the core of our framework lies a *majority voting (MV)* strategy that aggregates and extracts the dominant style representation across a batch. Building upon this, we propose two complementary mechanisms: *Majority Style Injection (MSI)*, which enforces global style coherence by adaptively modulating shared style cues, and *Set-Based Projection (SBP)*, which refines object regions through exemplar-based projection to achieve seamless content–style integration. Comprehensive experiments demonstrate that *FREESTYLE* surpasses both training-based and training-free baselines in terms of style consistency, object fidelity, and global adherence, while enabling real-time, high-quality style-aligned image generation without any fine-tuning or architectural modification.

Acknowledgements

This work was supported by the 2024 innovation base artificial intelligence data convergence project with the funding of the 2024 government (Ministry of Science and ICT) (S2201-24-1002) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02219277, AI Star Fellowship Support Project (DGIST)).

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[2] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 2

[3] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 2

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[5] Jacob K Christopher, Stephen Baek, and Nando Fioretto. Constrained synthesis with projected diffusion models. *Advances in Neural Information Processing Systems*, 37: 89307–89333, 2024. 5

[6] Quan Dao, Xiaoxiao He, Ligong Han, Ngan Hoai Nguyen, Amin Heyrani Nobar, Faez Ahmed, Han Zhang, Viet Anh Nguyen, and Dimitris Metaxas. Discrete noise inversion for next-scale autoregressive text-based image editing. *arXiv preprint arXiv:2509.01984*, 2025. 1

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2, 3

[8] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pages 181–198. Springer, 2024. 2, 3, 6, 7

[9] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7545–7556, 2023. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[11] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15733–15744, 2025. 2, 3, 6, 7

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[13] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 2, 3, 6, 7

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 5

[17] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Training-free content injection using h-space in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5151–5161, 2024. 3

[18] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10124–10134, 2023. 2

[19] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017, 2023. 3

[20] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3, 6, 7

[21] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2, 3

[22] Kyoungmin Lee, Jihun Park, Jongmin Gim, Wonhyeok Choi, Kyumin Hwang, Jaeyeul Kim, and Sunghoon Im. A training-free style-personalization via svd-based feature decomposition. *arXiv preprint arXiv:2507.04482*, 2025. 2

[23] Chang Liu, Viraj Shah, Aiyu Cui, and Svetlana Lazebnik. Unziplora: Separating content and style from a single image.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16776–16785, 2025. 3

[24] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975. 5

[25] Jihun Park, Jongmin Gim, Kyoungmin Lee, Minseok Oh, Minwoo Choi, Jaeyeul Kim, Woo Chool Park, and Sunghoon Im. A training-free style-aligned image generation with scale-wise autoregressive model. *arXiv preprint arXiv:2504.06144*, 2025. 2, 3

[26] Jihun Park, Kyoungmin Lee, Jongmin Gim, Hyeonseo Jo, Minseok Oh, Wonhyeok Choi, Kyumin Hwang, Jaeyeul Kim, Minwoo Choi, and Sunghoon Im. Infinite-story: A training-free consistent text-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8278–8286, 2026. 2

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3, 6, 7, 1

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[30] Aniket Roy, Shubhankar Borse, Shreya Kadambi, Debasmit Das, Shweta Mahajan, Risheek Garrepalli, Hyojin Park, Ankita Nayak, Rama Chellappa, Munawar Hayat, et al. Duolora: Cycle-consistent and rank-disentangled content-style personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15395–15404, 2025. 3

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[32] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2022. *URL https://github.com/cloneofsimo/lora*, 10:19, 2022. 2, 3

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[34] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 2

[35] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer, 2024. 3

[36] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 2, 3, 6, 7

[37] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024. 2

[38] Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 2024. 2, 3

[39] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024. 2, 3

[40] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 3

[41] Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024. 1

[42] Yi Wu, Lingting Zhu, Shengju Qian, Lei Liu, Wandi Qiao, Lequan Yu, and Bin Li. Stylear: Customizing multimodal autoregressive model for style-aligned text-to-image generation. *arXiv preprint arXiv:2505.19874*, 2025. 3, 6, 7, 8

[43] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024. 6, 7

[44] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3, 6, 7, 8

[45] Jiexuan Zhang, Yiheng Du, Qian Wang, Weiqi Li, Yu Gu, and Jian Zhang. Alignedgen: Aligning style across generated images. *arXiv preprint arXiv:2509.17088*, 2025. 2, 3, 6, 7, 8

FREESTYLE: An Anchor-Free Mechanism for Training-Free Style-Aligned Image Generation

Supplementary Material

A. Generalizability of our method

A.1. Incorporating ours into other scale-wise autoregressive models

To verify the generalization capability of our framework, we integrate it into Switti [41], another Text-to-Image model that adopts the scale-wise autoregressive generation paradigm. Our method is applied to each residual feature $\mathbf{R}_s$ of Switti without modifying any pretrained parameters or training procedures.

As shown in Fig. 6, the model equipped with our method produces image sets that exhibit more consistent style alignment across diverse object prompts while faithfully preserving the semantic content described by the text. Compared to the vanilla Switti, our integration enhances cross-sample style coherence, demonstrating that the proposed mechanisms are architecture-agnostic and can be seamlessly incorporated into other scale-wise autoregressive T2I models to improve stylistic consistency without retraining.

Figure 6. Qualitative results on Switti with *FREESTYLE*. The row (a) uses prompt: {A friendly robot, A woman walking a dog, Cherryblossom} in flat cartoon illustration style. The row (b) uses prompt: {Flowers, Golden gate bridge, Trees} in water color painting style.

A.2. Incorporating ours into diffusion-based model

While our method is developed within the scale-wise autoregressive paradigm, we further demonstrate its generality by adapting *FREESTYLE* to diffusion-based generation. Specifically, we apply our approach to SDXL [27], a representative diffusion T2I model. Unlike autoregressive models, where *FREESTYLE* operates on residual features at each scale, we inject style information at the noise residual of each denoising step without modifying model weights or architecture.

As shown in Fig. 7, applying *FREESTYLE* to SDXL enforces consistent style across images while preserving the semantic content defined by the prompts. This result highlights that our method is not restricted to scale-wise autoregressive models and can be seamlessly integrated into diffusion-based pipelines in a training-free, anchor-free manner.

Figure 7. Qualitative results on SDXL with *FREESTYLE*. The row (a) uses prompt: {A friendly robot, A woman walking a dog, Cherryblossom} in flat cartoon illustration style. The row (b) uses prompt: {Flowers, Golden gate bridge, Trees} in water color painting style.

B. Style alignment from user-provided reference

We further extend our framework to support user-specified style control by combining scale-wise autoregressive inversion with our method. Following [6], we first invert a reference image into its scale-wise feature trajectory. During generation, the recovered style representations are reintroduced at each scale as guidance signals, while our method enforces batch-wise style consistency without any additional training or parameter updates.

As shown in Fig. 8, our approach successfully transfers the reference style across diverse object prompts while preserving textual semantics. These results demonstrate that our framework not only yields style-consistent outputs in model-inherent styles but also faithfully adapts to externally

provided reference style images in a controlled, prompt-aligned manner.

Figure 8. Results of reference image-based style-aligned image generation.

C. Additional analysis

C.1. Impact of cross-attention maps

To further analyze the effect of coarse and noisy cross-attention maps, we provide additional results examining how the quality of cross-attention masks relates to the final generation outcomes. Specifically, we use SAM3 [3] to obtain object segmentation masks and measure the MIoU between these masks and the cross-attention masks across all backbone scales. Fig. 9 compares the cases with high and low MIoU samples and shows that the style alignment remains stable. This suggests that cross-attention maps provide sufficient localization of prompt-critical regions for generation, even when they are relatively noisy, and our method remains robust with these key areas.

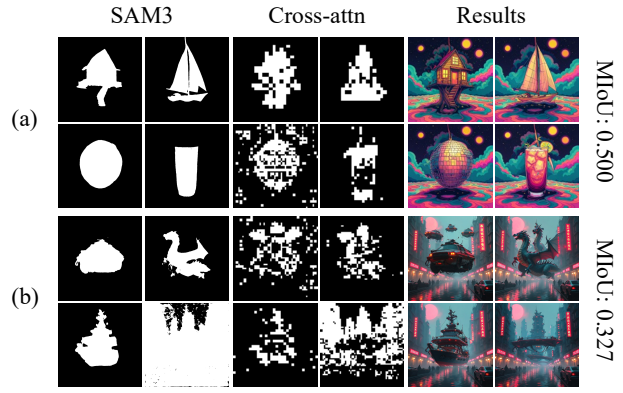

Figure 9. Visualization of the impact of cross-attention map accuracy.

C.2. Complex multi-object prompts

We provide additional qualitative results on complex multi-object prompts. These examples are intended to examine whether our method remains effective in more challenging settings involving multiple entities and attributes. While the models generally follow the prompts, some objects are occasionally omitted. As shown in Fig. 10, the baseline already fails to render the "fluffy cat," suggesting that such omissions mainly stem from limitations of the underlying backbone in the training-free regime. Nevertheless, our method yields more stylistically coherent results while largely maintaining the requested object composition.

Prompt : {a dreamy woman and a calm bulldog and a fluffy cat and a white swan, ... } in impressionist painting style.

Figure 10. Qualitative results with complex multi-object prompts.

C.3. Additional quantitative analysis

We provide additional quantitative comparisons with recent SOTA editing baselines using expanded evaluation metrics, including CSD [37] and $S_{\text{arithmetic}}$. Because image editing models operate in a reference-guided setting, a style reference image is required for evaluation. To this end, we first generate one image from the text prompt alone and then use it as the reference image for generating the remaining samples. Here, S_{CSD} measures style consistency and serves as an additional style-sensitive metric complementary to S_{DINO} and S_{CLIP}.

As shown in Tab. 4, our method achieves the best overall performance across most of the reported metrics, while remaining substantially faster than the other baselines. Qwen-Image-Edit achieves a high S_{DINO} score, but its relatively low S_{obj} and S_{whole} suggest less reliable preservation of the requested object composition and prompt semantics. In contrast, FLUX.2 obtains comparatively stronger object-related scores, but its lower style-related metrics suggest less faithful style alignment. Both S_{harmonic} and $S_{\text{arithmetic}}$ are included in the table for completeness.

C.4. AdaIN operation in Majority Voting (MV)

Majority Voting (MV) extracts a batch-wise representative residual style feature R_s^{maj} by selecting dominant residu-

Table 4. Quantitative results with additional models and metrics.

Method	$S_{\text{harmonic}} \uparrow$	$S_{\text{arithmetic}} \uparrow$	$S_{\text{obj}} \uparrow$	$S_{\text{CSD}} \uparrow$	$S_{\text{DINO}} \uparrow$	$S_{\text{whole}} \uparrow$	$S_{\text{CLIP}} \uparrow$	Time (s) $\downarrow$
StyleAR	<u>0.455</u>	<u>0.541</u>	<u>0.281</u>	<u>0.768</u>	0.559	0.330	0.772	335.23
StyleAligned	0.447	0.525	<u>0.281</u>	0.723	0.530	<u>0.331</u>	0.762	11.25
IP-Adapter	0.442	0.522	0.278	0.708	0.529	0.324	0.772	<u>10.14</u>
Qwen-Image-Edit	0.431	0.526	0.257	0.663	**0.625**	0.301	<u>0.784</u>	154.80
FLUX.2	0.445	0.529	0.271	0.702	0.561	0.327	0.782	146.93
Ours	**0.463**	**0.553**	**0.284**	**0.772**	<u>0.589</u>	**0.332**	**0.791**	**1.98**

als across samples. However, because this feature is taken from real generated samples, it may still retain object-specific signals that interfere with purely style-driven propagation. To further isolate stylistic information, during majority voting, we apply AdaIN to R_s^{maj}, normalizing object-associated regions using statistics from style-dominant regions, thereby suppressing content cues while preserving style characteristics.

Tab. 5 reports quantitative results with and without the AdaIN refinement. Applying AdaIN yields consistent improvements in style consistency—both DINO-based (S_{DINO}) and CLIP-based metrics (S_{CLIP})—while also providing slight gains in prompt-related metrics (S_{obj}, S_{whole}). These results confirm that AdaIN effectively removes residual content signals from R_s^{maj}, producing cleaner style representations that lead to more robust and faithful style alignment.

Table 5. Quantitative ablation study on AdaIN operation in Majority Voting (MV).

Method	$S_{\text{harmonic}} \uparrow$	$S_{\text{obj}} \uparrow$	$S_{\text{DINO}} \uparrow$	$S_{\text{whole}} \uparrow$	$S_{\text{CLIP}} \uparrow$
W/o AdaIN	0.086	0.283	0.584	0.331	0.787
Ours	**0.088**	**0.284**	**0.589**	**0.332**	**0.791**

D. Limitation and future work

Our anchor-free method is training-free and aggregates batch-wise style information from generated samples to enforce a representative, shared style across the batch. While this approach effectively enhances style consistency when the model has a well-formed prior for the target style, it also inherits limitations from the pretrained backbone. When the base model has limited exposure to the target style, individual samples may express it inconsistently, resulting in noisy batch statistics and making it difficult to extract a stable representative style. In such cases, the final output may fail to fully reflect the intended style, not because the alignment mechanism fails, but because the underlying model lacks a reliable style manifold to align toward. These limitations suggest future extensions that expand style priors or incorporate external style memory for more robust alignment under unfamiliar styles.

E. Additional qualitative results

We present additional qualitative samples in Fig. 11 and Fig. 12, demonstrating our method across diverse styles and prompt variations. The results show that our approach consistently preserves content semantics while enforcing coherent style characteristics across samples. Unlike anchor-based propagation, our method does not depend on any specific sample, thereby avoiding artifact transfer and maintaining stable performance even under large variations in object prompts and stylistic conditions. These findings further validate the robustness and generality of our framework in practical, style-aligned generation scenarios.

F. Details of User Study

We conducted a user study with 40 participants (ages 20–50) using the interface shown in Fig. 13. The study is divided into two evaluation tasks, each presented as a multi-choice selection from four options (Options 1–4). The four options correspond to our method and three top-performing baselines selected based on quantitative scores, and their order was randomized for every trial to avoid positional bias.

Part 1: Text Relevance. Each trial displayed a content prompt along with four candidate image sets. Participants were asked to select the option that best matched the semantic meaning of the prompt.

Part 2: Style Consistency. Each trial presented different objects paired with the same style prompt, again with four candidate image sets. Participants were instructed to choose the set that most consistently reflects a unified visual style across samples.

Responses for both tasks were collected as single categorical choices per trial, allowing separate aggregation for text relevance and style consistency.

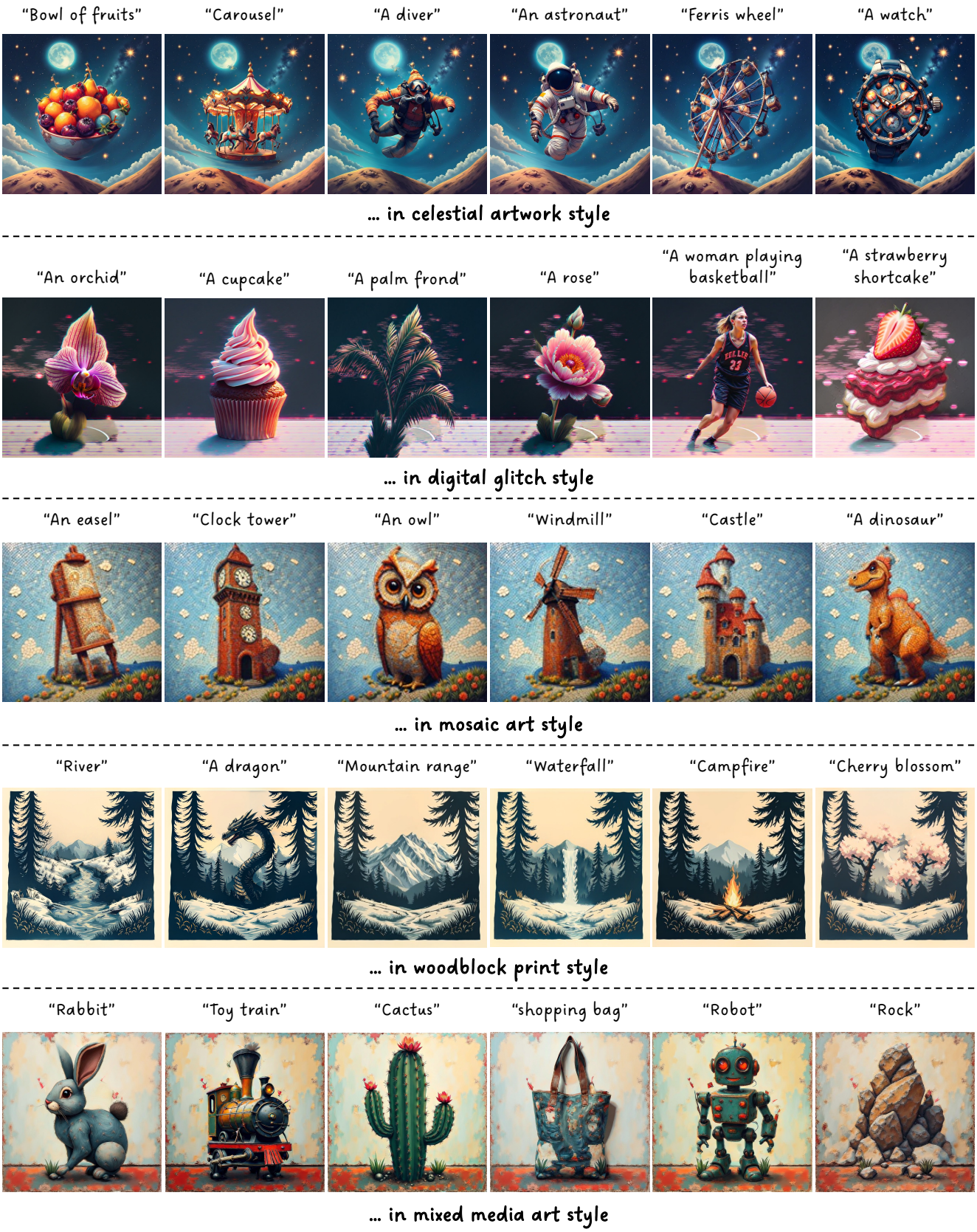

Figure 11. Additional qualitative results of *FREESTYLE* under diverse style-aligned image generation settings.

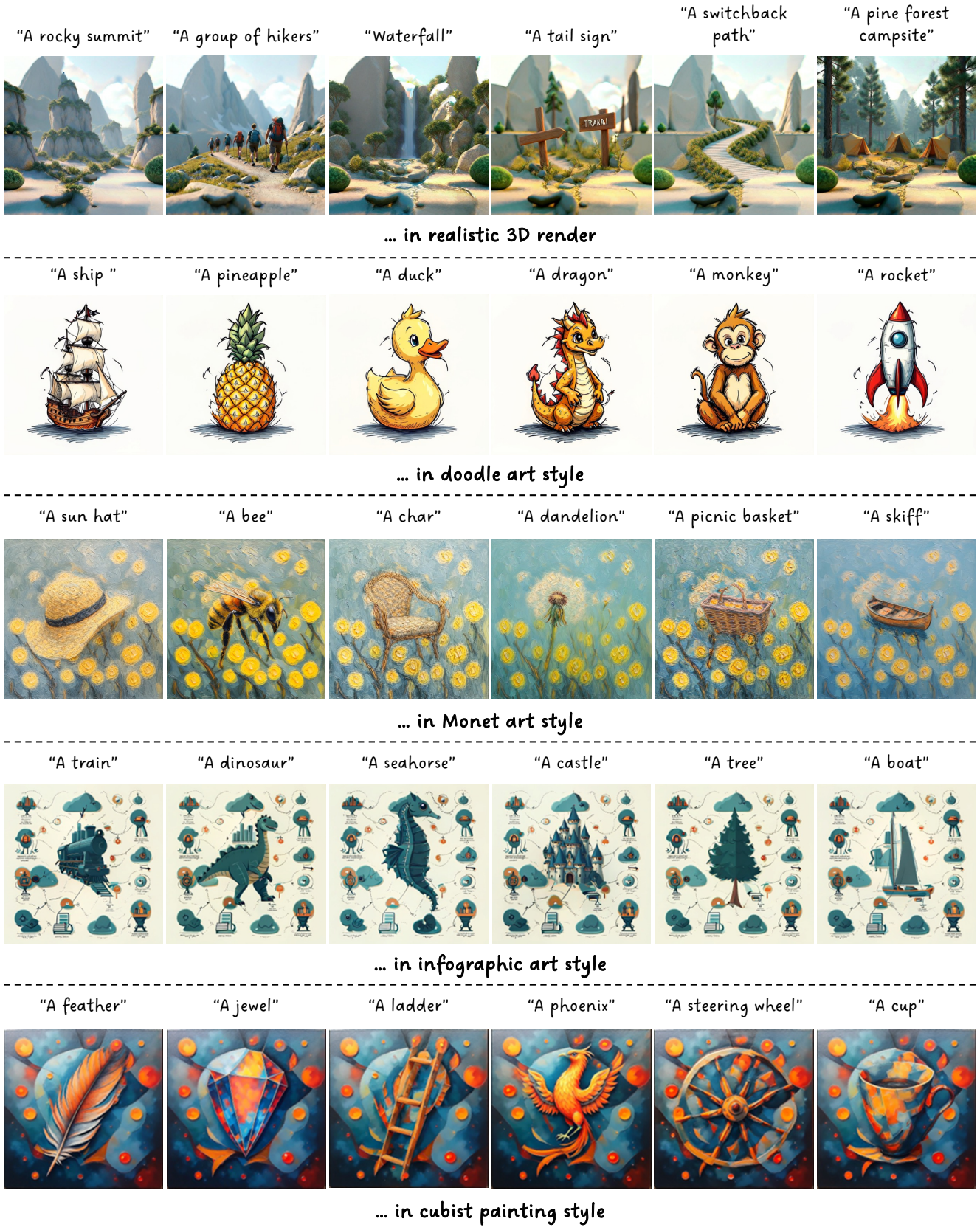

Figure 12. Additional qualitative results of *FREESTYLE* under diverse style-aligned image generation settings.

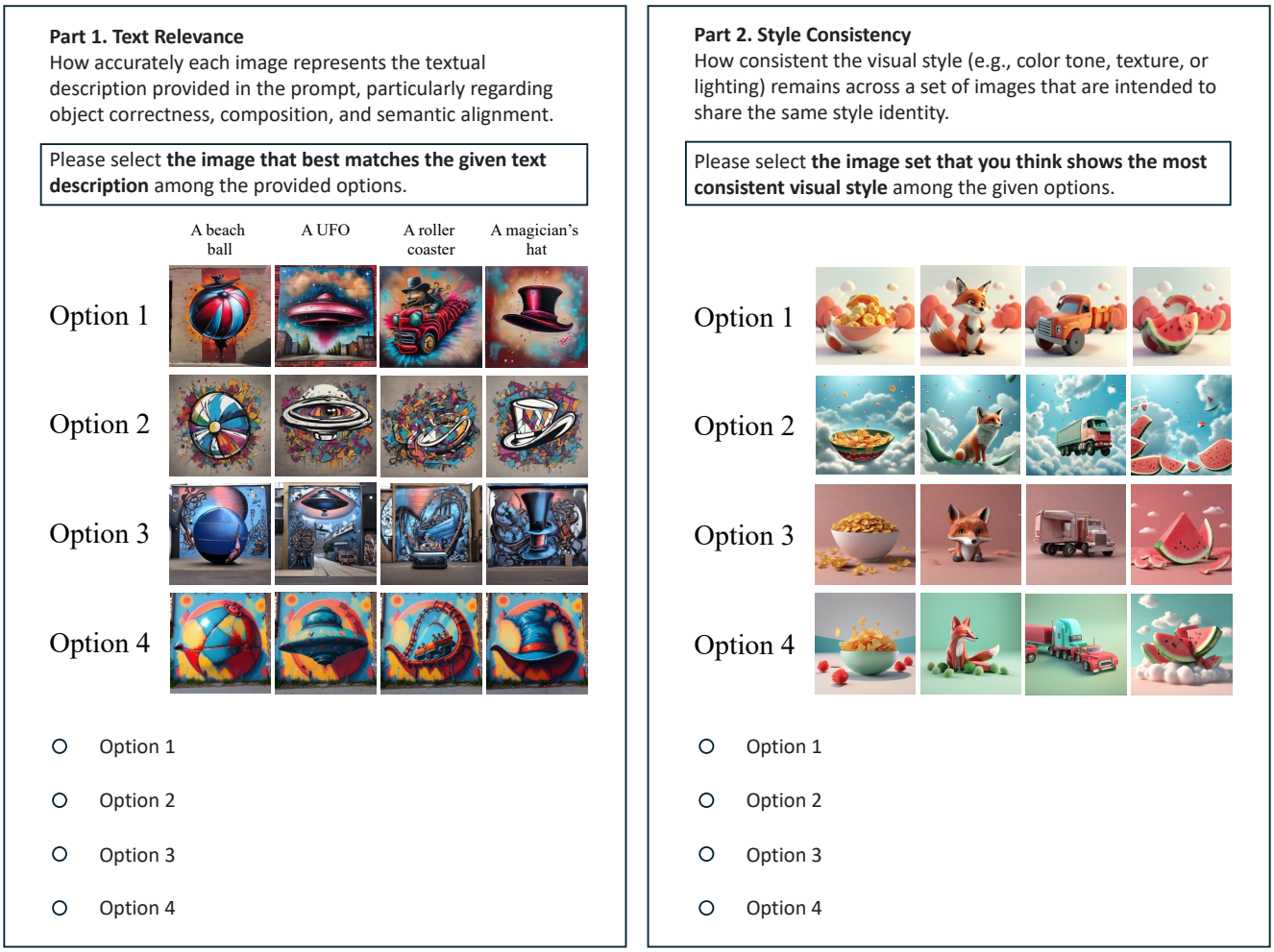

Figure 13. User study interface for text relevance (left) and style consistency (right). Among four randomized options, participants select the image that best matches the prompt or the set with the most consistent visual style.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[2] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 2

[3] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 2

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[5] Jacob K Christopher, Stephen Baek, and Nando Fioretto. Constrained synthesis with projected diffusion models. *Advances in Neural Information Processing Systems*, 37: 89307–89333, 2024. 5

[6] Quan Dao, Xiaoxiao He, Ligong Han, Ngan Hoai Nguyen, Amin Heyrani Nobar, Faez Ahmed, Han Zhang, Viet Anh Nguyen, and Dimitris Metaxas. Discrete noise inversion for next-scale autoregressive text-based image editing. *arXiv preprint arXiv:2509.01984*, 2025. 1

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2, 3

[8] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pages 181–198. Springer, 2024. 2, 3, 6, 7

[9] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7545–7556, 2023. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[11] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15733–15744, 2025. 2, 3, 6, 7

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[13] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 2, 3, 6, 7

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 5

[17] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Training-free content injection using h-space in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5151–5161, 2024. 3

[18] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10124–10134, 2023. 2

[19] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017, 2023. 3

[20] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3, 6, 7

[21] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2, 3

[22] Kyoungmin Lee, Jihun Park, Jongmin Gim, Wonhyeok Choi, Kyumin Hwang, Jaeyeul Kim, and Sunghoon Im. A training-free style-personalization via svd-based feature decomposition. *arXiv preprint arXiv:2507.04482*, 2025. 2

[23] Chang Liu, Viraj Shah, Aiyu Cui, and Svetlana Lazebnik. Unziplora: Separating content and style from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16776–16785, 2025. 3

[24] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975. 5

[25] Jihun Park, Jongmin Gim, Kyoungmin Lee, Minseok Oh, Minwoo Choi, Jaeyeul Kim, Woo Chool Park, and Sunghoon Im. A training-free style-aligned image generation with scale-wise autoregressive model. *arXiv preprint arXiv:2504.06144*, 2025. 2, 3

[26] Jihun Park, Kyoungmin Lee, Jongmin Gim, Hyeonseo Jo, Minseok Oh, Wonhyeok Choi, Kyumin Hwang, Jaeyeul

Kim, Minwoo Choi, and Sunghoon Im. Infinite-story: A training-free consistent text-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8278–8286, 2026. 2

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3, 6, 7, 1

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[30] Aniket Roy, Shubhankar Borse, Shreya Kadambi, Debasmit Das, Shweta Mahajan, Risheek Garrepalli, Hyojin Park, Ankita Nayak, Rama Chellappa, Munawar Hayat, et al. Duolora: Cycle-consistent and rank-disentangled content-style personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15395–15404, 2025. 3

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[32] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2022. *URL https://github.com/cloneofsimo/lora*, 10:19, 2022. 2, 3

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[34] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 2

[35] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer, 2024. 3

[36] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 2, 3, 6, 7

[37] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024. 2

[38] Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 2024. 2, 3

[39] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024. 2, 3

[40] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 3

[41] Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024. 1

[42] Yi Wu, Lingting Zhu, Shengju Qian, Lei Liu, Wandi Qiao, Lequan Yu, and Bin Li. Stylear: Customizing multimodal autoregressive model for style-aligned text-to-image generation. *arXiv preprint arXiv:2505.19874*, 2025. 3, 6, 7, 8

[43] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024. 6, 7

[44] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3, 6, 7, 8

[45] Jiexuan Zhang, Yiheng Du, Qian Wang, Weiqi Li, Yu Gu, and Jian Zhang. Alignedgen: Aligning style across generated images. *arXiv preprint arXiv:2509.17088*, 2025. 2, 3, 6, 7, 8