arXiv:2601.06605v1 [cs.CV] 10 Jan 2026

Sissi: Zero-shot Style-guided Image Synthesis via Semantic–style Integration

Yingying Deng[1], Xiangyu He[2*], Fan Tang[3], Weiming Dong[2], Xucheng Yin[1]

[1]Department of Computer Science and Technology, University of Science and Technology Beijing, Beijing, 100083, China.
[2*]Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China.
[3]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, 100040, China.

*Corresponding author(s). E-mail(s): hexiangyu17@mails.ucas.edu.cn;
Contributing authors: yingying.deng@ustb.edu.cn; tfan.108@gmail.com; weiming.dong@ia.ac.cn; xuchengyin@ustb.edu.cn;

Abstract

Text-guided image generation has advanced rapidly with large-scale diffusion models, yet achieving precise stylization with visual exemplars remains difficult. Existing approaches often depend on task-specific retraining or expensive inversion procedures, which can compromise content integrity, reduce style fidelity, and lead to an unsatisfactory trade-off between semantic prompt adherence and style alignment. In this work, we introduce a training-free framework that reformulates style-guided synthesis as an in-context learning task. Guided by textual semantic prompts, our method concatenates a reference style image with a masked target image, leveraging a pretrained ReFlow-based inpainting model to seamlessly integrate semantic content with the desired style through multimodal attention fusion. We further analyze the imbalance and noise sensitivity inherent in multimodal attention fusion and propose a Dynamic Semantic-Style Integration (DSSI) mechanism that reweights attention between textual semantic and style visual tokens, effectively resolving guidance conflicts and enhancing output coherence. Experiments show that our approach achieves high-fidelity stylization with superior semantic-style balance and visual quality, offering a simple yet powerful alternative to complex, artifact-prone prior methods.

Keywords: Zero-shot, Style-guided, Image synthesis, Style transfer

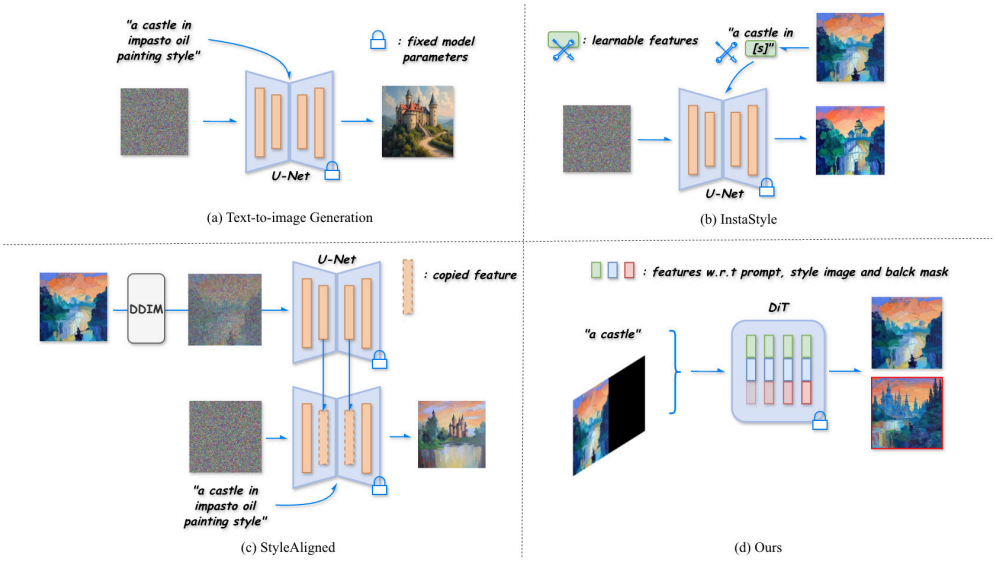


Fig. 1: Different pipelines for style-guided image synthesis: (a) style-related text-guided image synthesis; (b) training-based style-guided image synthesis (InstaStyle [1]); (c) training-free style-guided image synthesis (StyleAligned [2]); and (d) our method, which simply leverages in-context learning in a pre-trained ReFlow inpainting model enhanced by dynamic semantic-style integration for robust fusion.

1 Introduction

Text-guided image generation has gained increasing attention due to the rise of large-scale diffusion models, which allow users to specify desired styles through textual prompts. However, relying solely on text for style guidance presents several limitations. As shown in Fig. 1(a), the abstractness and ambiguity of textual style descriptions make it difficult to capture fine-grained visual details such as brush strokes, textures, subtle color variations, material gloss, or specific artistic effects (e.g., impasto in oil paintings). Moreover, a short prompt often cannot fully convey these details, potentially resulting in style transfer outputs that deviate from user expectations.

To overcome these limitations, recent research has incorporated style reference images to provide explicit visual examples of target styles. A set of methods [1, 3] directly learn the corresponding textual embedding from images, using the expressive power of the textual space to convey the stylistic characteristics of artistic images. Adapter-based fine-tuning for text-to-image diffusion models [4, 5] integrate style control through a learned adapter module, and encode the style image to style textual embeddings which is also concatenated to text embedding to guide the stylized results. However, the image-guided approaches mentioned above typically require task-specific training and exhibit limited generalization to unseen styles. Moreover, their tendency to entangle structural details from the reference style image often leads to undesired content distortions and overfitting (see Fig. 1(b)). Alternatively, training-free

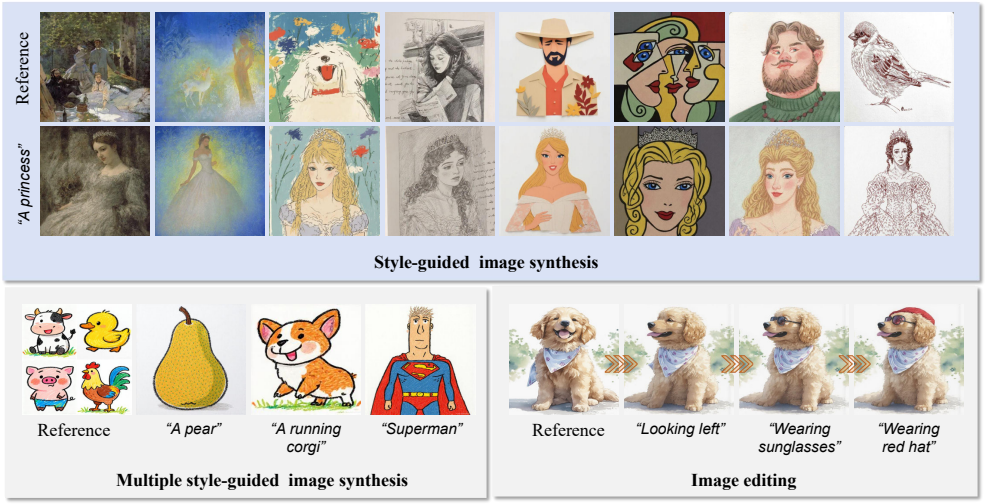


Fig. 2: Examples of style-guided image synthesis using our method **Sissi**. Given a style reference, our approach generates stylized images conditioned on a content prompt, achieving exceptional quality and fidelity. Furthermore, it seamlessly extends to multiple style-guided synthesis and image editing tasks, demonstrating versatility and superior performance.

inversion-based methods [2, 6, 7] leverage DDIM inversion [8] and latent feature fusion to inject visual style from a reference image into the text-conditioned latent space. While avoiding retraining, these approaches are computationally expensive during inference and sensitive to inaccurate attention estimation, often producing artifacts or unstable style alignment (see Fig. 1(c)). Collectively, existing approaches struggle to simultaneously achieve high visual quality, strong style–content consistency and robustness.

Recent progress in Rectified Flow (ReFlow) models provides a potential direction. Built upon Diffusion Transformer (DiT) backbones, ReFlow models substitute conventional U-Net convolutions with transformer-based sequence modeling, effectively capturing long-range dependencies across spatial and stylistic dimensions. This architectural design endows the model with exceptional in-context learning (ICL) capabilities [9], facilitating rapid adaptation to multimodal conditioning signals. However, existing ICL-based image generation tasks [10, 11] typically rely on explicit exemplar pairs for guidance, making it challenging to decouple appropriate style patterns from a single style exemplar and integrate them with a target semantic prompt in ambiguous scenarios.

In this paper, we propose a novel training-free style-guided text-to-image synthesis method that reformulates the problem as an in-context style learning task. Conditioned on textual semantic prompts, our method concatenates a reference style image with a masked target image to create a structured input, which is then fed into a pre-trained ReFlow-based inpainting model. This design enables the text prompt

to provide high-level semantic and content guidance, while the style image supplies concrete visual exemplars of textures, colors, and artistic characteristics, facilitating spatially intuitive integration between style and content regions. However, this multimodal fusion introduces core technical difficulties, including imbalance and noise sensitivity in attention mechanisms. Concatenated attention often results in dominance by stronger activations, such as texture-rich styles overwhelming semantics, and suppression of the weaker branch. Noise introduced by the zero-mask in the inpainting pipeline further exacerbates this issue, leading to amplified perturbation across sampling steps. To address these problems and ensure coherent integration of semantic and stylistic signals, we introduce a semantic-style integration mechanism that dynamically reweights attention between textual and visual tokens, adaptively resolving conflicts between the guidance sources. The entire framework operates in a training-free manner, thereby enabling effective generalization to unseen styles.

In summary, our key contributions are as follows:

- We propose a novel training-free framework for style-guided image synthesis that seamlessly integrates textual semantic guidance with visual style exemplars, reformulating the task as an in-context style learning problem, supported by analyses of imbalance and noise sensitivity in multimodal attention fusion.
- We introduce a dynamic semantic-style integration mechanism that adaptively balances cross-modal attention between text and style features, ensuring high fidelity to both semantic intent and stylistic characteristics.
- Extensive experiments demonstrate that our method achieves flexible generalization to unseen styles and textual prompts without requiring model retraining, and can be applied to multiple style-guided image synthesis and various image editing tasks (see Fig. 2).

2 Related Work

2.1 Personalized Image Generation

Diffusion models [12–14] have enabled remarkable progress in text-to-image generation by learning to reverse a noise-adding process and synthesizing high-quality images across diverse tasks [15–19]. Beyond generic generation, personalization has emerged as a key direction, aiming to generate images that preserve user-specific characteristics or follow user-defined visual styles [20–23].

Our work focuses on style-guided image synthesis, a sub-task of personalized generation that aims to produce images consistent with the style of a given reference image. Early approaches, such as Textual Inversion [24], learn a textual embedding from the input image to guide the generation process. DreamBooth [21] personalizes a diffusion model by fine-tuning it on a few images of a subject. The computational demands of DreamBooth's full fine-tuning spurred the development of parameter-efficient methods like Low-Rank Adaptation (LoRA) [25, 26]. To further improve inference efficiency, some works [4, 5, 27] introduce a lightweight conditioning framework that unlocks precise control over large, frozen text-to-image diffusion models.

Recently, ReFlow models [28–30] adopting diffusion transformer structures have emerged as a promising alternative to diffusion models. This new class of generative models achieves faster inference and enhanced visual fidelity by learning straighter probability paths, as demonstrated in several state-of-the-art text-to-image systems. In parallel, a growing line of research explores in-context personalization in the diffusion transformer model. In-Context LoRA [31] introduces a lightweight fine-tuning technique for diffusion transformers. By leveraging multiple images (rather than just text) as context, it efficiently adapts the model using task-specific LoRA modules while keeping the base model weights unchanged, achieving high-fidelity image generation and demonstrating excellent task-agnostic capabilities. Similarly, Diptych Prompting [32] entirely avoids model fine-tuning by reframing the generation task as a "diptych" inpainting process, enabling precise subject-driven generation in a zero-shot manner. IP-Prompter [33] focuses on training-free theme-specific generation, which involves dynamic visual prompts to guide the pre-trained model to recognize and replicate key visual themes from reference images. In-Context Brush [34] achieves zero-shot subject customizing by reformulating the task within the paradigm of in-context learning.

Despite these advances, effectively integrating semantic content from text with visual style from a reference image in large pretrained diffusion transformers remains challenging. In this work, we explore how content and style can be fused through in-context learning for style-guided text-to-image synthesis, enabling customized stylization with strong content preservation and faithful style alignment.

2.2 Style Transfer

Style transfer aims to render the stylistic patterns of a given style image onto a given content image, enabling non-professional users to generate artistic works. With the development of convolutional neural networks (CNNs), Gatys et al. [35] discover that perceptual features extracted from pre-trained VGG models can effectively represent both content and style information. Inspired by this work, a series of feed-forward style transfer methods [36–40] have been proposed to enable efficient inference. For instance, AdaIN [36] introduces adaptive instance normalization to merge content and style features, replacing second-order style statistics. WCT [38] employs whitening and coloring transformations to apply style patterns to content images. SANet [40] proposes a style-attention mechanism for local content-style fusion. CAST [41] adopts contrastive learning to improve style representation, moving beyond traditional Gram-based losses. Beyond CNNs, researchers have also explored alternative network architectures for style transfer. ArtFlow [42] utilizes a flow-based network to mitigate the content leak problem. StyTr^2 [43] designs a pure transformer-based network to better balance content structure and style patterns. More recently, SaMam [44] leverages a Mamba-based architecture to achieve a global receptive field while maintaining linear computational complexity.

The advent of diffusion and rectified flow models has further expanded the scope of style transfer. Methods such as InST [3] and VCT [45] employ inversion-based schemes to encode a style image into a compact style embedding, which guides the generation process. StyleShot [7], InstantStyle [46], DEADiff [47] and CSGO [48] fine-tune

the style-aware adapter module to achieve style transfer. To further reduce computational cost, several training-free approaches have been proposed. For example, Deng et al. [49, 50] rearrange attention maps in the latent space of pre-trained diffusion models, directly extracting and integrating style priors into content representations without inversion tuning. WSDT [51] introduces a training-free Wasserstein Style Distribution Transform method for precise and efficient stylized image generation. VSP [52] preserves content consistency by fusing key style features from the reference in a late self-attention layer. StyleAligned [2] proposes a shared attention mechanism to align style and content.

In contrast to existing approaches that depend on fine-tuning or unstable attention replacement, our method introduces a training-free paradigm that leverages style in-context learning for robust and superior content-style integration.

3 Preliminaries

Rectified Flow Models

The fundamental principle of ReFlow models is to learn a straight trajectory between two distributions, such as a noise distribution and a data distribution. This is achieved by training a neural ordinary differential equation (ODE) model that transforms samples from the source distribution to the target distribution along an approximately linear path. Specifically, Rectified Flow constructs a transport map T between two empirical distributions π_0 (e.g., noise) and π_1 (e.g., data). It is defined by the ODE:

$$d\mathbf{x}_t = \mathbf{v}_\theta(\mathbf{x}_t, t)\, dt, \quad \mathbf{x}_0 \sim \pi_0, \quad \mathbf{x}_1 \sim \pi_1, \tag{1}$$

where $\mathbf{v}_\theta(\mathbf{z}, t)$ denotes the velocity field parameterized by a neural network with parameters θ.

The training objective minimizes the expected squared difference between the predicted velocity and the straight-line direction:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0 \sim \pi_0, \mathbf{x}_1 \sim \pi_1} \left[\int_0^1 \|\mathbf{x}_1 - \mathbf{x}_0 - \mathbf{v}_\theta(\mathbf{x}_t, t)\|_2^2 \, dt \right], \tag{2}$$

where $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$.

Diffusion Transformer Structure

The Diffusion Transformer (DiT) is a novel transformer-based backbone designed for diffusion models, enabling scalable generation and editing across multiple modalities such as text and images [29]. DiT consists of a stack of L transformer blocks, each processing concatenated sequences of modality-specific tokens: text tokens $\mathbf{T} \in \mathbb{R}^{N_t \times d}$ (derived from pretrained CLIP [53] and T5 models [54]) and image latents $\mathbf{I} \in \mathbb{R}^{N_i \times d}$ (from the VAE), where N_t and N_i denote sequence lengths, and d is embedding dimensions. The forward pass can be formalized as:

$$\mathbf{O}_t, \mathbf{O}_i = \mathrm{DiT}(\mathbf{T}, \mathbf{I}; \theta), \tag{3}$$

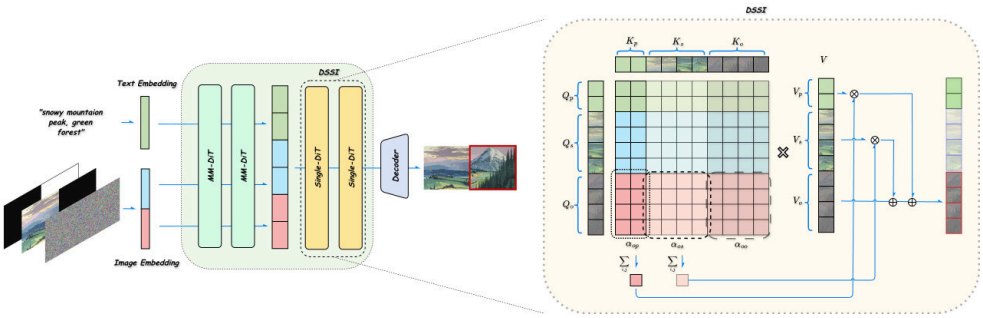


Fig. 3: Overview of our framework's pipeline. Given a style image I_s, we concatenate it with a target mask and input the combined image to the generation network, leveraging its in-context style adaptation capability. The DSSI module adaptively balances attention between textual and visual style cues, yielding outputs that incorporate target style patterns while preserving content semantics.

where $\mathbf{O}_t$ and $\mathbf{O}_i$ are the updated text and image representations, θ denotes the model parameters.

The core module in the DiT-based model is the attention mechanism [55], which has become a fundamental component in neural networks for information aggregation:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}. \tag{4}$$

The attention mechanism serves as the key driver behind these deep token interactions, which are crucial for achieving coherence in generated images.

4 Method

4.1 In-context Style Enhancement

In contrast to the UNet-based text-to-image architecture, which incorporates textual prompts via cross-attention, Transformer-based models integrate both text and image embeddings as a unified token sequence. These multimodal tokens are processed jointly through self-attention within DiT blocks, enabling deeper fusion between semantic and visual representations.

This unified multimodal representation suggests that the model can flexibly attend to multiple types of conditioning information in the same latent space. Building on this insight, we extend the notion of a "prompt" beyond text and posit that a style image can serve an equivalent role, acting as a visual conditioning signal that guides the model toward desired stylistic attributes. Instead of relying solely on text to describe abstract visual styles, directly encoding the style image provides explicit visual cues, enabling more accurate and transferable style control.

Given the target style image $\mathbf{I}_s \in \mathbb{R}^{h\times w\times 3}$ and text prompt $\mathbf{c}$, we aim to generate the output image $\mathbf{I}_o \in \mathbb{R}^{h\times w\times 3}$. To leverage the consistency of the DiT-based model,

we concatenate $\mathbf{I}_s$ and $\mathbf{I}_o$ as $[\mathbf{I}_s; \mathbf{I}_o]$ and input them as a single image. Since $\mathbf{I}_o$ is unknown, we initialize the input for $\mathbf{I}_o$ to zero and provide a mask $[0; 1]$ to the pre-trained inpainting model. As the generation process occurs in the latent space, we denote the input embedding as $\mathbf{x} = [\mathbf{x}_p; \mathbf{x}_s; \mathbf{x}_o]$, where $\mathbf{x}_p$, $\mathbf{x}_s$, and $\mathbf{x}_o$ represent the text tokens, style image tokens, and output tokens, respectively.

Through the learnable projection matrices $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$, we obtain the query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) features in the attention module:

$$\begin{aligned} \mathbf{Q} &= [\mathbf{Q}_p; \mathbf{Q}_s; \mathbf{Q}_o] = \mathbf{x}\mathbf{W}_q, \\ \mathbf{K} &= [\mathbf{K}_p; \mathbf{K}_s; \mathbf{K}_o] = \mathbf{x}\mathbf{W}_k, \\ \mathbf{V} &= [\mathbf{V}_p; \mathbf{V}_s; \mathbf{V}_o] = \mathbf{x}\mathbf{W}_v. \end{aligned} \tag{5}$$

The attention output can be expressed as:

$$\begin{aligned} \mathbf{H} &= [\mathbf{H}_p; \mathbf{H}_s; \mathbf{H}_o] = \mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \\ &= \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \\ &= \mathrm{Softmax}\left(\frac{1}{\sqrt{d}}\begin{bmatrix} \mathbf{Q}_p\mathbf{K}_p^\top & \mathbf{Q}_p\mathbf{K}_s^\top & \mathbf{Q}_p\mathbf{K}_o^\top \\ \mathbf{Q}_s\mathbf{K}_p^\top & \mathbf{Q}_s\mathbf{K}_s^\top & \mathbf{Q}_s\mathbf{K}_o^\top \\ \mathbf{Q}_o\mathbf{K}_p^\top & \mathbf{Q}_o\mathbf{K}_s^\top & \mathbf{Q}_o\mathbf{K}_o^\top \end{bmatrix}\right)\begin{bmatrix} \mathbf{V}_p \\ \mathbf{V}_s \\ \mathbf{V}_o \end{bmatrix}. \end{aligned} \tag{6}$$

For the output corresponding to the query tokens $\mathbf{Q}_o$, the attention can be expressed as:

$$\begin{aligned} \mathbf{H}_o &= \mathrm{Softmax}\left(\frac{1}{\sqrt{d}}\begin{bmatrix} \mathbf{Q}_o\mathbf{K}_p^\top & \mathbf{Q}_o\mathbf{K}_s^\top & \mathbf{Q}_o\mathbf{K}_o^\top \end{bmatrix}\right)\begin{bmatrix} \mathbf{V}_p \\ \mathbf{V}_s \\ \mathbf{V}_o \end{bmatrix} \\ &= \boldsymbol{\alpha}_p\mathbf{V}_p + \boldsymbol{\alpha}_s\mathbf{V}_s + \boldsymbol{\alpha}_o\mathbf{V}_o, \end{aligned} \tag{7}$$

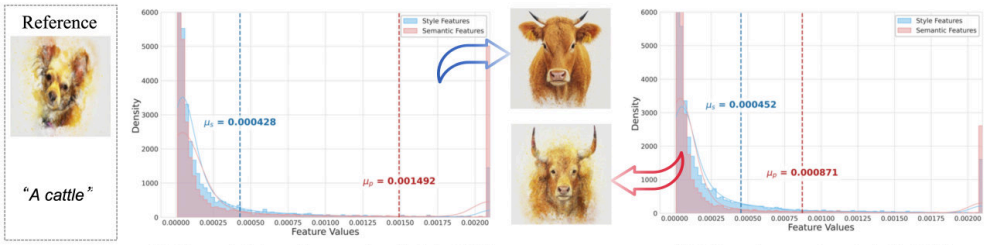
where $\boldsymbol{\alpha}_i = \frac{\exp\left(\frac{\mathbf{Q}_o\mathbf{K}_i^\top}{\sqrt{d}}\right)}{\sum_j \exp\left(\frac{\mathbf{Q}_o\mathbf{K}_j^\top}{\sqrt{d}}\right)}$, with the exponential and softmax operations applied element-wise over the concatenated attention logits for each row of $\mathbf{Q}_o$.

4.2 Problems in the Vanilla Setting

In the vanilla formulation, the attention weights $\boldsymbol{\alpha}_p, \boldsymbol{\alpha}_s, \boldsymbol{\alpha}_o$ are derived directly from the concatenated softmax without adaptive reweighting. This fixed aggregation assumes balanced contributions from the prompt ($\boldsymbol{\alpha}_p\mathbf{V}_p$), style ($\boldsymbol{\alpha}_s\mathbf{V}_s$), and output ($\boldsymbol{\alpha}_o\mathbf{V}_o$) branches, but in practice, it leads to the following issues:

Inherent Imbalance in Semantic-Style Contributions

In the vanilla formulation, the concatenated softmax over the prompt, style, and output branches amplifies differences in activation magnitudes due to the "winner-takes-all" behavior of softmax. This normalization exacerbates imbalance when the



(a) Inherent imbalance in semantic-style (w/o. DSSI) (b) Balance in semantic-style (with DSSI)

Fig. 4: An illustration of semantic-style imbalance. Subfigure (a) shows a failure case from the baseline model, where the prompt dominates the output ($\mu_p \gg \mu_s$). In contrast, our approach achieves a more balanced generation where the style (μ_s) and prompt (μ_p) contributions are comparable.

raw activations from one branch (e.g., style) are systematically larger than another's (e.g., prompt).

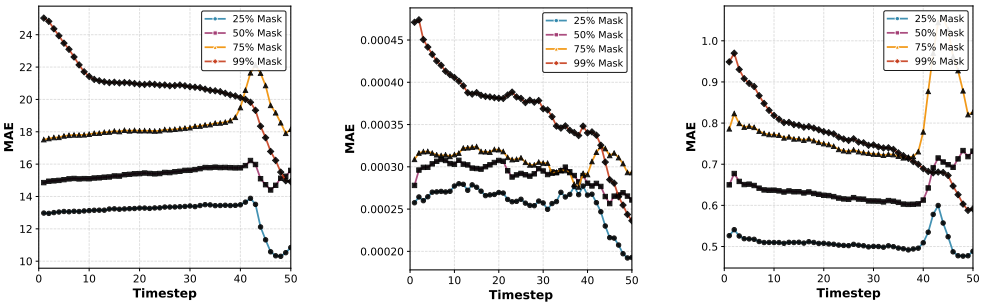
To derive this, recall that the dot products $\vec{\mathbf{q}} \cdot \vec{\mathbf{k}}^T$ for vectors $\vec{\mathbf{q}}, \vec{\mathbf{k}} \in \mathbb{R}^d$ with unit norms have expected variance of 1 after scaling by $1/\sqrt{d}$ (assuming i.i.d. components with variance 1). The softmax then computes probabilities as $\text{Softmax}(\mathbf{Z})_{i,j} = \frac{\exp(\mathbf{Z}_{i,j})}{\sum_k \exp(\mathbf{Z}_{i,k})}$, where $\mathbf{Z} = \frac{1}{\sqrt{d}}[\mathbf{Q}_o\mathbf{K}_p^\top \quad \mathbf{Q}_o\mathbf{K}_s^\top \quad \mathbf{Q}_o\mathbf{K}_o^\top]$ (concatenated across branches). For activations with unit variance, softmax exhibits a strong winner-takes-all effect: if one branch's mean activation value μ_s (for style branch $\vec{\mathbf{q}_o}\mathbf{K}_s$) exceeds another's μ_p (for prompt branch $\vec{\mathbf{q}_o}\mathbf{K}_p$) by a large margin (e.g., $\mu_s - \mu_p > 1$), the probability mass concentrates on the higher branch, as $\exp(\mu_s) \gg \exp(\mu_p)$. Quantitatively, let the concatenated activations have branch means μ_p, μ_s, μ_o, we have the following statement.

Proposition 1 *Assume that for each branch $b \in \{p, s, o\}$, the entries of $\mathbf{Z}$ satisfy $\mathbf{Z}_{i,j} = \mu_b + \xi_{i,j}$, where μ_b is the branch-specific mean and $\xi_{i,j}$ are i.i.d. zero-mean random variables with variance $\sigma^2 \approx 1$. Then the ℓ_1 norm of each row of the style-branch attention weights satisfies:*

$$\|\vec{\alpha}_s\|_1 \approx \frac{N_s \exp(\mu_s)}{N_p \exp(\mu_p) + N_s \exp(\mu_s) + N_o \exp(\mu_o)} = \frac{1}{1 + \frac{N_p}{N_s}\exp(\mu_p - \mu_s) + \frac{N_o}{N_s}\exp(\mu_o - \mu_s)},$$

up to an approximation error of order $\mathcal{O}(\sigma^3)$.

The implications are immediate: when $\mu_s \gg \mu_p$, the attention becomes dominated by the style branch, driving $\|\vec{\alpha}_s\|_1 \to 1$ and $\|\vec{\alpha}_p\|_1 \to 0$, which produces over-stylization and semantic distortion. Conversely, as shown in Fig. 4, when $\mu_p \gg \mu_s$, the style branch is effectively suppressed, giving weak or washed-out stylization. Although the $1/\sqrt{d}$ scaling ensures variance stability as the dimension grows, it does not prevent inappropriate cross-branch imbalances: differences arising from input statistics (e.g., texture-heavy style encodings vs. semantic prompt encodings) can push the system into regimes where one branch almost fully dominates.



(a) MAE of Softmax logits $\tilde{\mathbf{Z}}$ v.s. $\mathbf{Z}$.

(b) MAE of attention weights $\tilde{\boldsymbol{\alpha}}$ v.s. $\boldsymbol{\alpha}$.

(c) MAE of output tensor $\tilde{\mathbf{h}}$ v.s. $\mathbf{h}$.

Fig. 5: An illustration depicting how noise introduced by masking the input propagates through logits, attention weights, and the output tensor. The Mean Absolute Error (MAE) quantifies the differences between results obtained with masking and the reference results. As the masked area increases, these differences become more pronounced.

Sensitivity to Noise

In the previous subsection, we demonstrated that the vanilla attention configuration becomes unsatisfactory when the model strongly favors either the content branch ($\mu_p \gg \mu_s$) or the style branch ($\mu_s \gg \mu_p$). A less obvious, yet equally significant issue arises from perturbations introduced by the input mask in the inpainting setting. To investigate this, we conduct a zero-mask reconstruction experiment: within the FLUX-Fill pipeline, we concatenate a reference image with the same image masked at different percentages (25%, 50%, 75% and 99%) and attempt to reconstruct the original. As shown in Fig. 5, applying a zero-mask for reconstruction perturbs the activations in the attention operations relative to those produced by a clean input. These perturbations are further amplified by the softmax operation, ultimately causing the output to diverge from the original image. In our style-guide image synthesis experiments, we may also regard image synthesis guided by a prompt as the reconstruction of an oracle image. Therefore, reducing the noise introduced by the zero-mask is essential for achieving higher-quality generation.

Formally, let $\tilde{\mathbf{Z}} = \mathbf{Z} + \boldsymbol{\epsilon}$, perturbed by $\|\boldsymbol{\epsilon}\|_\infty \leq \delta$, and $\tilde{\boldsymbol{\alpha}}_* = \mathrm{Softmax}(\tilde{\mathbf{Z}}_*)$. The following result characterizes how logit perturbations propagate through the softmax.

Proposition 2 *Let $s \in \mathbb{R}^n$ be an arbitrary logit vector and let $\alpha = \mathrm{Softmax}(s)$ denote the corresponding probability distribution, i.e., $\alpha_j = \frac{\exp(s_j)}{\sum_{k=1}^{n}\exp(s_k)}, \quad j = 1, \ldots, n$. Let $\tilde{s} \in \mathbb{R}^n$ be a perturbed logit vector satisfying $\|\tilde{s} - s\|_\infty \leq \delta$ for some $\delta \geq 0$, and define $\tilde{\alpha} = \mathrm{Softmax}(\tilde{s})$. Then the total variation distance between the two distributions is bounded by $\mathrm{Total_Variance}(\tilde{\alpha}, \alpha) := \frac{1}{2}\|\tilde{\alpha} - \alpha\|_1 \leq 1 - e^{-2\delta}$. Equivalently, $\|\tilde{\alpha} - \alpha\|_1 \leq 2(1 - e^{-2\delta})$. For small perturbations δ, the bound behaves asymptotically as $\|\tilde{\alpha} - \alpha\|_1 \leq 4\delta + O(\delta^2)$.*

Using this bound, the perturbation of the output row $\tilde{\mathbf{h}}_o$ compared to the ideal case $\mathbf{h}_o$ satisfies:

$$\begin{aligned}\|\tilde{\mathbf{h}}_o - \mathbf{h}_o\|_2 &= \|(\vec{\tilde{\alpha}}_p - \vec{\alpha}_p)\mathbf{V}_p + (\vec{\tilde{\alpha}}_s - \vec{\alpha}_s)\mathbf{V}_s + (\vec{\tilde{\alpha}}_o - \vec{\alpha}_o)\mathbf{V}_o\|_2 \\ &\leq \sum_* \|\vec{\tilde{\alpha}}_* - \vec{\alpha}_*\|_1 \, \|\mathbf{V}_*\|_2 \\ &\leq \max_* \left(\|\vec{\tilde{\alpha}}_* - \vec{\alpha}_*\|_1\right) \, (\|\mathbf{V}_p\|_2 + \|\mathbf{V}_s\|_2 + \|\mathbf{V}_o\|_2) \\ &= 2(1 - e^{-2\delta}) \, (\|\mathbf{V}_p\|_2 + \|\mathbf{V}_s\|_2 + \|\mathbf{V}_o\|_2),\end{aligned} \tag{8}$$

where α_* denotes either α_p, α_s or α_o. The factor $2(1 - e^{-2\delta})$ is intrinsic, reflecting the stochastic noise inherent to the FLUX-Fill pipeline. Therefore, we focus on the simplest strategy: reducing $\max_* \left(\|\vec{\tilde{\alpha}}_* - \vec{\alpha}_*\|_1\right)$, which in turn leads to a tighter upper bound on the output perturbation.

4.3 Dynamic Semantic-Style Integration

In the above formulation, the attention output $\mathbf{h}_o$ integrates three types of contextual information: semantic guidance from textual prompts $(\mathbf{Q}_o, \mathbf{K}_p, \mathbf{V}_p)$, stylistic cues from the reference image $(\mathbf{Q}_o, \mathbf{K}_s, \mathbf{V}_s)$, and output $(\mathbf{Q}_o, \mathbf{K}_o, \mathbf{V}_o)$. Ideally, these sources should complement each other, text guiding the global semantic structure, while the style image provides local texture and color cues. However, in practice, we observe that the contributions of semantic and style signals are often unbalanced. Specifically, when the style image exhibits strong visual features, its attention response tends to dominate the update of $\mathbf{h}_o$, leading to over-stylization, content distortion, or unintended introduction of style structures. Conversely, when the text prompt exerts the primary influence, the model may overemphasize textual semantics, resulting in weak style transfer and diminished visual fidelity. This imbalance indicates that a fixed attention formulation (as in Eq. (7)) cannot dynamically adapt to varying semantic-style relationships during inference.

To address this, we introduce a Dynamic Semantic-Style Integration (DSSI) mechanism, which adaptively adjusts the attention balance between textual and visual style cues, enabling the model to maintain semantic consistency while preserving fine-grained stylistic fidelity. This approach draws inspiration from gated fusion mechanisms in multimodal vision-language models, where adaptive weights prevent one modality from overwhelming others, ensuring coherent integration without training.

We reformulate DSSI as an optimization problem over the attention outputs. Define the semantic-guided reference output as:

$$\mathbf{s} = \kappa \cdot \boldsymbol{\alpha}_p \mathbf{V}_p + \boldsymbol{\alpha}_o \mathbf{V}_o, \tag{9}$$

emphasizing text prompts and output, ignoring style, and the style-guided reference output as:

$$\mathbf{t} = \kappa \cdot \boldsymbol{\alpha}_s \mathbf{V}_s + \boldsymbol{\alpha}_o \mathbf{V}_o, \tag{10}$$

emphasizing the style reference and output, ignoring text prompts, where κ serves as a hyper-parameter to adjust the strength of the prior of prompt and style. The goal is to find a fused output:

$$\mathbf{h}_o^{\text{DSSI}} = (1-\lambda)\mathbf{s} + \lambda\mathbf{t}, \tag{11}$$

parameterized by $\lambda \in [0,1]$, that balances proximity to $\mathbf{s}$ (for semantic fidelity) and $\mathbf{t}$ (for stylistic fidelity). This can be formalized as the following unconstrained optimization problem:

$$\min_{\lambda\in[0,1]} \mathcal{L}(\lambda) = \|\mathbf{h}(\lambda) - \mathbf{s}\|_F^2 + \gamma\|\mathbf{h}(\lambda) - \mathbf{t}\|_F^2, \tag{12}$$

where $\mathbf{h}(\lambda) = (1-\lambda)\mathbf{s} + \lambda\mathbf{t}$, $\|\cdot\|_F^2$ denotes the squared Frobenius norm, and $\gamma > 0$ is a trade-off parameter controlling the relative emphasis on stylistic versus semantic fidelity. The Frobenius norm is particularly apt here as it respects the matrix structure of attention outputs (rows as tokens, columns as dimensions), providing a holistic measure of discrepancy while being computationally efficient and differentiable.

Since $\mathbf{h}(\lambda)$ is linear in λ, $\mathcal{L}(\lambda)$ is a convex quadratic function, admitting a closed-form solution. To confirm convexity, note the second derivative: $\frac{d^2}{d\lambda^2}\mathcal{L}(\lambda) = 2D + 2\gamma D = 2D(1+\gamma) > 0$ (for $D > 0$), ensuring a unique global minimum and stable optimization during inference. Let $D = \|\mathbf{t} - \mathbf{s}\|_F^2$ (assuming $D > 0$; otherwise, $\mathbf{s} = \mathbf{t}$ and any λ suffices). Substituting $\mathbf{h}(\lambda)$ yields:

$$\begin{aligned} \|\mathbf{h}(\lambda) - \mathbf{s}\|_F^2 &= \|\lambda(\mathbf{t} - \mathbf{s})\|_F^2 = \lambda^2 D, \\ \|\mathbf{h}(\lambda) - \mathbf{t}\|_F^2 &= \|(1-\lambda)(\mathbf{s} - \mathbf{t})\|_F^2 = (1-\lambda)^2 D. \end{aligned} \tag{13}$$

Thus,

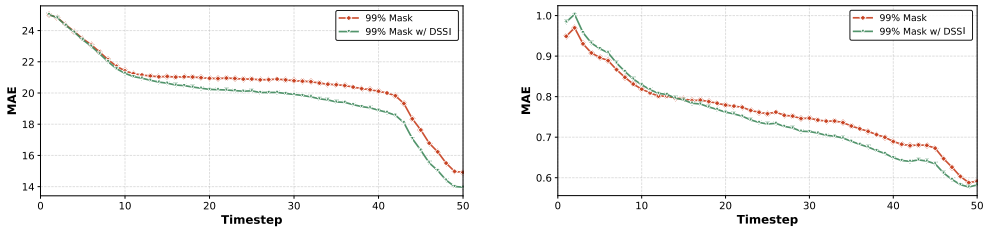
$$\mathcal{L}(\lambda) = D\left[\lambda^2 + \gamma(1-\lambda)^2\right]. \tag{14}$$

Minimizing with respect to λ (ignoring the constant D):

$$\begin{aligned} &\frac{d}{d\lambda}\left[\lambda^2 + \gamma(1-\lambda)^2\right] = 2\lambda - 2\gamma(1-\lambda) = 0 \\ &\implies \lambda + \gamma\lambda = \gamma \\ &\implies \lambda^* = \frac{\gamma}{1+\gamma}. \end{aligned} \tag{15}$$

Since $\gamma \geq 0$, $\lambda^* \in [0,1]$, satisfying the bounds without additional constraints. Alternatively, this can be interpreted probabilistically: treating $\mathbf{s}$ and $\mathbf{t}$ as "priors" from semantic and style distributions, λ^* acts as a posterior weight, akin to Bayesian model averaging, where γ reflects the relative "evidence" from each modality.

The trade-off parameter γ is intuitively selected as the ratio of style integration strength to semantic integration strength, i.e., $\gamma = \frac{\lambda_p}{\lambda_s}$. Here, λ_p and λ_s measure the



(a) MAE of Softmax logits $\tilde{\mathbf{Z}}$ v.s. $\mathbf{Z}$.

(b) MAE of output tensor $\tilde{\mathbf{h}}$ v.s. $\mathbf{h}$.

Fig. 6: An illustration showing the effect of DSSI in reducing the noise introduced by the zero-mask, with the reference image serving as the baseline for MAE computation.

overall softmax alignment between the target queries and the respective keys:

$$\begin{aligned} \lambda_p &= \log \sum_{i=1}^{N_o} \sum_{j=1}^{N_p} \left[\text{Softmax}\left(\frac{\mathbf{Q}_o \mathbf{K}_p^\top}{\sqrt{d}} \right) \right]_{i,j}, \\ \lambda_s &= \log \sum_{i=1}^{N_o} \sum_{j=1}^{N_s} \left[\text{Softmax}\left(\frac{\mathbf{Q}_o \mathbf{K}_s^\top}{\sqrt{d}} \right) \right]_{i,j}, \end{aligned} \tag{16}$$

where N_o, N_p, and N_s are the number of target, prompt, and style tokens, respectively. This choice ensures that if the style cues are strongly aligned ($\lambda_s \gg \lambda_p$), γ is small, prioritizing semantic preservation ($\lambda^* \approx 0$). Conversely, if semantic cues dominate ($\lambda_p \gg \lambda_s$), γ is small, favoring stylistic fidelity ($\lambda^* \approx 1$).

Substituting $\gamma = \frac{\lambda_p}{\lambda_s}$ into the closed-form solution derives our DSSI weighting:

$$\lambda = \frac{\gamma}{1+\gamma} = \frac{\frac{\lambda_p}{\lambda_s}}{1+\frac{\lambda_p}{\lambda_s}} = \frac{\lambda_p}{\lambda_s + \lambda_p}. \tag{17}$$

The reweighted output attention is then:

$$\mathbf{h}_o^{\text{DSSI}} = \kappa \cdot \big((1-\lambda) \cdot \boldsymbol{\alpha}_p \mathbf{V}_p + \lambda \cdot \boldsymbol{\alpha}_s \mathbf{V}_s \big) + \boldsymbol{\alpha}_o \mathbf{V}_o. \tag{18}$$

This mechanism allows the model to adaptively modulate the contribution of textual and visual cues based on the current attention distribution, effectively mitigating conflicts between semantic and stylistic features to make a balance. When the textual prompt provides strong semantic cues, λ increases to enhance stylistic fidelity. Conversely, when the style reference exhibits rich texture patterns, the model preserves more structural content by reducing λ.

4.4 Noise Sensitivity Analysis

Having shown that the proposed method encourages more balanced attention between the prompt and style branches, we now analyze how the scaling factor λ contributes to

robustness under noise. Let the clean DSSI output be $\mathbf{H}_o = (1-\lambda)\,\boldsymbol{\alpha}_p\mathbf{V}_p + \lambda\,\boldsymbol{\alpha}_s\mathbf{V}_s + \boldsymbol{\alpha}_o\mathbf{V}_o$, where we set $\kappa = 1$ for simplicity, $\lambda = \frac{\lambda_p}{\lambda_p+\lambda_s}$ is the dynamic scaling coefficient, and $\boldsymbol{\alpha}_p, \boldsymbol{\alpha}_s, \boldsymbol{\alpha}_o$ are the attention weights produced by the concatenated softmax. Under the noise introduced by our training-free modification for the pre-trained FLUX-Fill model, the output becomes $\tilde{\mathbf{H}}_o = (1-\tilde{\lambda})\,\tilde{\boldsymbol{\alpha}}_p\mathbf{V}_p + \tilde{\lambda}\,\tilde{\boldsymbol{\alpha}}_s\mathbf{V}_s + \tilde{\boldsymbol{\alpha}}_o\mathbf{V}_o$, with $\tilde{\lambda} = \frac{\tilde{\lambda}_p}{\tilde{\lambda}_p+\tilde{\lambda}_s}$ and perturbed attention weights $\tilde{\boldsymbol{\alpha}}_*$. Subtracting the clean output yields:

$$\begin{aligned}\tilde{\mathbf{H}}_o - \mathbf{H}_o &= (1-\tilde{\lambda})\tilde{\boldsymbol{\alpha}}_p\mathbf{V}_p - (1-\lambda)\boldsymbol{\alpha}_p\mathbf{V}_p + \tilde{\lambda}\,\tilde{\boldsymbol{\alpha}}_s\mathbf{V}_s - \lambda\boldsymbol{\alpha}_s\mathbf{V}_s + (\tilde{\boldsymbol{\alpha}}_o - \boldsymbol{\alpha}_o)\mathbf{V}_o \\ &= (1-\lambda)(\tilde{\boldsymbol{\alpha}}_p - \boldsymbol{\alpha}_p)\mathbf{V}_p - (\tilde{\lambda}-\lambda)\tilde{\boldsymbol{\alpha}}_p\mathbf{V}_p + \lambda(\tilde{\boldsymbol{\alpha}}_s - \boldsymbol{\alpha}_s)\mathbf{V}_s \\ &\quad + (\tilde{\lambda}-\lambda)\tilde{\boldsymbol{\alpha}}_s\mathbf{V}_s + (\tilde{\boldsymbol{\alpha}}_o - \boldsymbol{\alpha}_o)\mathbf{V}_o.\end{aligned} \tag{19}$$

Applying the triangle inequality row-wise and using the same norm relations as in Eq. (8), we obtain:

$$\begin{aligned}\|\tilde{\mathbf{h}}_o - \mathbf{h}_o\|_2 \le &\underbrace{(1-\lambda)\|\vec{\tilde{\alpha}}_p - \vec{\alpha}_p\|_1\|\mathbf{V}_p\|_2 + \lambda\|\vec{\tilde{\alpha}}_s - \vec{\alpha}_s\|_1\|\mathbf{V}_s\|_2}_{\text{term I}} \\ &+ \underbrace{|\tilde{\lambda}-\lambda|\|\vec{\tilde{\alpha}}_p\|_2\|\mathbf{V}_p\|_2 + |\tilde{\lambda}-\lambda|\|\vec{\tilde{\alpha}}_s\|_2\|\mathbf{V}_s\|_2}_{\text{term II}} \\ &+ \underbrace{\|\vec{\tilde{\alpha}}_o - \vec{\alpha}_o\|_2\|\mathbf{V}_o\|_2}_{\text{term III}}.\end{aligned} \tag{20}$$

Term III matches the vanilla setting exactly. Term I is always smaller than its vanilla counterpart because $\lambda \in [0,1]$. We now show that Term II is negligible.

Proposition 3 *Let* $\lambda_p = \log\sum_{i,j}[\mathrm{Softmax}(\mathbf{Q}_o\mathbf{K}_p^\top/\sqrt{d})]_{i,j}$ *and* $\lambda_s = \log\sum_{i,j}[\mathrm{Softmax}(\mathbf{Q}_o\mathbf{K}_s^\top/\sqrt{d})]_{i,j}$ *denote the clean alignment strengths for the prompt and style branches, respectively, and* $\lambda = \frac{\lambda_p}{\lambda_p+\lambda_s}$. *Under noisy alignments* $\tilde{\lambda}_p = \lambda_p + \Delta_p$ *and* $\tilde{\lambda}_s = \lambda_s + \Delta_s$ *with noise bounds* $|\Delta_p| \le \epsilon_p$ *and* $|\Delta_s| \le \epsilon_s$, *the perturbed scaling factor* $\tilde{\lambda} = \frac{\tilde{\lambda}_p}{\tilde{\lambda}_p+\tilde{\lambda}_s}$ *satisfies* $|\tilde{\lambda}-\lambda| \le \frac{\max(\epsilon_p,\epsilon_s)}{\lambda_p+\lambda_s}$.

Following the noise model in Section 4.2, $\tilde{\mathbf{Z}} = \mathbf{Z} + \boldsymbol{\epsilon}$, $\|\boldsymbol{\epsilon}\|_\infty \le \delta$, we further obtain $\max(\epsilon_p, \epsilon_s) \le 2\delta$ (see Appendix). Since $\|\vec{\tilde{\alpha}}_*\|_1 \le 1$, Term II satisfies:

$$|\tilde{\lambda}-\lambda|\,\|\vec{\tilde{\alpha}}_*\|_2 \le |\tilde{\lambda}-\lambda|\|\vec{\tilde{\alpha}}_*\|_1 \le \frac{2\delta}{\lambda_p+\lambda_s}.$$

In both the imbalance regime (where one branch dominates, e.g., $\lambda_p \gg \lambda_s$) and the balanced regime (where prompt and style have comparable strength, e.g., $\lambda_p \approx \lambda_s$), we always have $\lambda_p + \lambda_s \gg \delta$. Thus, Term II is vanishingly small and can be ignored. Therefore, Eq. (20) clearly shows a lower bound than Eq. (8); hence, the proposed

method is less sensitive to the noise introduced by the mask for stylized image generation than the vanilla setting. To further illustrate this, we present a comparison of the effect of DSSI on noise reduction during masked image reconstruction in Fig. 6. The results demonstrate that DSSI progressively reduces noise in the logits and diminishes its impact on the outputs.

5 Experiment

5.1 Evaluation

Our method uses Flux-Fill [56] as the base generation model. All experiments are conducted on a single NVIDIA Tesla P40 GPU. Unless otherwise specified, we use the default hyperparameters provided by the original Flux-Fill repository and perform inference with 30 denoising steps. As demonstrated in Fig. 7, our approach exhibits remarkable generalization capability across a wide variety of style references and content scenarios. In addition, we comprehensively compare our approach against several state-of-the-art style-guided image synthesis methods, including WSDT [51], VSP [52], StyleShot [7], StyleAligned [2], DEADiff [47], InstaStyle [1], InstantStyle [46] and CSGO [48]. The generation process for a 512×512 image takes approximately 15 seconds. This inference speed is on par with training-free baselines (e.g., WSDT, VSP, StyleAligned) and significantly faster than methods requiring day-level training (e.g., StyleShot, DEADiff, InstaStyle, InstantStyle, CSGO).

5.1.1 Qualitative Evaluation

As shown in Fig. 8, we present representative qualitative comparisons with SOTA baselines on a diverse set of challenging content prompts and reference style images. Existing methods exhibit clear limitations in achieving robust style consistency and content fidelity simultaneously. StyleAligned, DEADiff, InstantStyle, InstaStyle and CSGO frequently fail to faithfully reproduce the target style, resulting in outputs that either marginally reflect the reference (weak stylization) or introduce inconsistent and patchy style patterns. StyleShot, VSP and WSDT demonstrate improved style expressiveness in relatively simple cases. However, when confronted with complex cases, these methods struggle to effectively adapt the style elements to the content layout, such as the "A smiling girl" and "A dolphin" cases. As can be observed, our approach achieves superior visual fidelity in both style expressiveness and content preservation, producing stylized images that closely adhere to the reference style while maintaining the semantic structure of the input content. Importantly, beyond mere low-level attributes such as color distribution and local texture, our approach effectively learns and transfers higher-order stylistic features, including global composition principles, rhythm of brushstrokes, spatial organization of visual elements, and overall artistic intent, as illustrated in the fourth and sixth rows of Fig. 8.

5.1.2 Quantitative Evaluation

This section presents a comprehensive quantitative comparison between the proposed method and existing state-of-the-art approaches to validate its effectiveness.

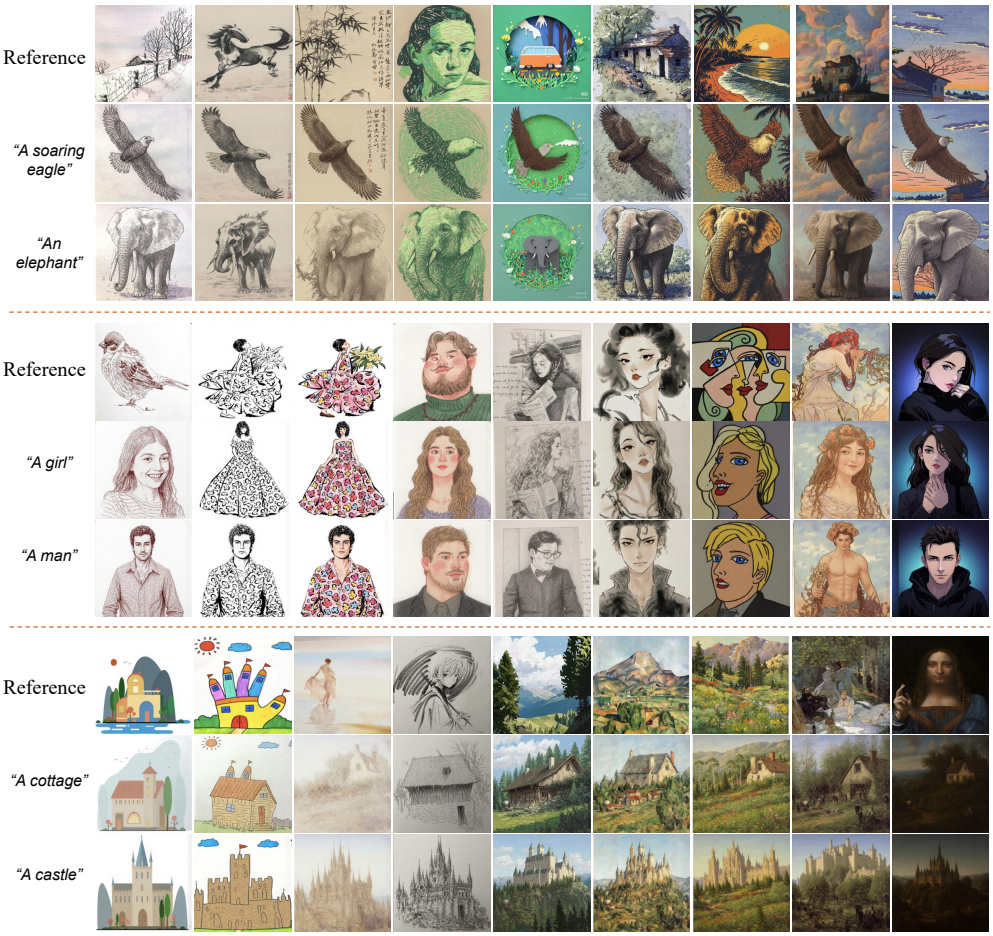


Fig. 7: Demonstration of our method's generation results across diverse style references and content scenarios, showcasing robust stylization while preserving semantic fidelity.

Evaluation Metrics. Following InstaStyle [1], we use the object classes from CIFAR-100 [57] and their corresponding class names as text prompts to generate content images. For style references, we adopt the original 60 artistic style images curated by InstaStyle and further enrich the dataset with 40 additional diverse style images (including abstract, sketch painting, cartoon, etc.), resulting in a total of 100 style references. These additional images are provided in the supplementary material. Performance is measured from two complementary perspectives using a pretrained CLIP [53]. **Content Alignment**: Cosine similarity between the CLIP image embedding of the generated result and the CLIP text embedding of the input content prompt (higher is better). **Style Alignment**: Cosine similarity between the CLIP image

Fig. 8: Comparison with state-of-the-art methods.

embedding of the generated result and the CLIP image embedding of the reference style image (higher is better).

As summarized in Table 1(a), our method achieves the highest style consistency score while maintaining competitive content alignment. This result confirms our approach's unique capability to strike an optimal balance, producing images that exhibit both strong fidelity to the content prompt and faithful adherence to the reference style.

User Study. To further evaluate subjective quality, we conducted a user preference study involving 30 participants, evenly split between technical (50%) and non-technical (50%) backgrounds. Each participant was asked to rate the generated results on three criteria (including content preservation, style alignment, and overall visual quality) on a scale of 1-5 (1 = poorest, 5 = best). The evaluation included 30 randomly selected cases generated from each compared method. The averaged scores are summarized in Table 1(b). Our approach consistently obtains the highest ratings across all three evaluated dimensions. These subjective preferences confirm the marked perceptual superiority of our method over existing state-of-the-art approaches. A key observation is the discrepancy between objective metrics and user study results. This arises because objective metrics provide a compartmentalized evaluation by design, while users naturally synthesize a unified impression from all aspects of the output.

Table 1: (a) Quantitative comparison of style transfer performance in terms of content and style alignment. (b) User study results on three perceptual criteria: content alignment, style alignment, and overall visual quality. The most favorable outcomes are highlighted in **bold**.

(a)

Method	Content ↑	Style ↑
Ours	0.287	**0.750**
WSDT	**0.297**	0.699
VSP	0.292	0.729
StyleShot	0.269	0.718
StyleAligned	0.281	0.700
DEADiff	0.286	0.687
InstaStyle	0.284	0.683
InstantStyle	0.280	0.708
CSGO	0.277	0.698

(b)

Method	Content ↑	Style ↑	Overall↑
Ours	**4.09**	**4.31**	**4.17**
WSDT	3.74	4.14	4.10
VSP	3.66	4.20	4.03
StyleShot	3.67	3.85	3.83
StyleAligned	3.64	3.71	3.72
DEADiff	3.08	2.93	2.96
InstaStyle	3.54	3.27	3.60
InstantStyle	3.55	3.34	3.56
CSGO	3.01	2.79	2.85

5.2 Ablation Study

5.2.1 Influence of κ.

To evaluate the role of the hyper-parameter κ in our Dynamic Semantic-Style Integration (DSSI) framework, we conduct an ablation study by varying its value and assessing its impact on the generated outputs. As described in Eq. 21, κ acts as a global scaling factor controlling the relative strengths of the prompt and style branches, effectively modulating the relative influence of semantic and stylistic cues during the fusion process. By adjusting κ, we can change the degree of stylization and content preservation. We systematically vary κ across a range of values (e.g., from 1.0 to 3.0) and generate images using diverse style references and content prompts. The results are visualized in Fig. 9 and quantitatively summarized in Table 2, where we report metrics such

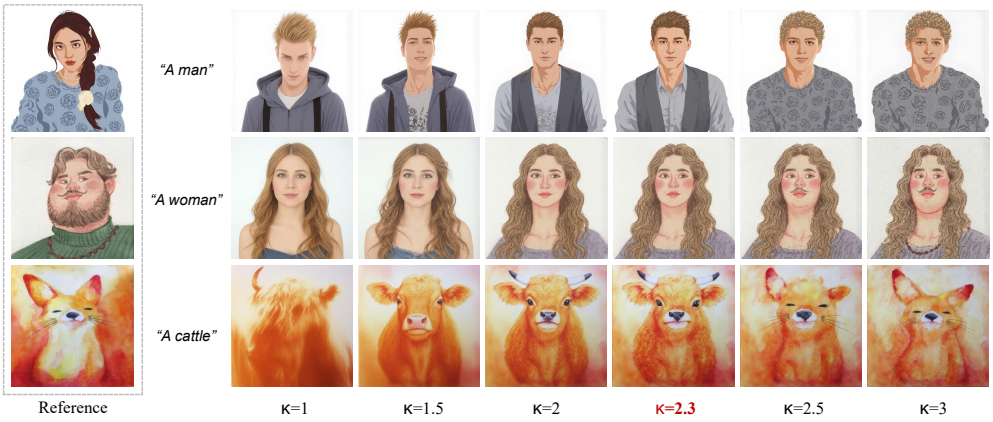


Fig. 9: Effects of varying κ on style-guided image synthesis. Smaller values of κ result in diminished style patterns, whereas larger values allow the semantics of the style image to influence the generated content. At $\kappa = 2.3$, the outputs achieve an optimal balance between stylization and semantic preservation.

as style alignment and content alignment. As κ increases, the style patterns become progressively more pronounced, enhancing the visual transfer of textures, colors, and artistic elements from the reference image. For small κ (e.g., $\kappa \leq 1.5$), the outputs exhibit subdued stylization, often resulting in images that lack distinctive artistic flair. Conversely, larger κ (e.g., $\kappa \geq 2.5$) amplifies style dominance, potentially introducing artifacts or semantic distortions, as the reference image's features overpower the content guidance. Empirically, we find that $\kappa = 2.3$ achieves an optimal balance, yielding high-quality results where stylization is vivid yet harmonious with the semantic content.

Table 2: Quantitative results demonstrating the influence of κ on generation performance across metrics, including content alignment and style alignment. The optimal parameter is marked with (*).

Metric	κ=1	κ=1.5	κ=2	κ=2.3(*)	κ=2.5	κ=3
Content ↑	0.276	0.279	0.289	0.287	0.277	0.271
Style ↑	0.632	0.646	0.700	0.750	0.756	0.760

5.2.2 Influence of DSSI Module.

To assess the impact of our proposed Dynamic Semantic-Style Integration (DSSI) module on the overall performance of the framework, we conduct an ablation study by simplifying and removing the module and comparing the results across three configurations. This analysis aims to isolate the contributions of DSSI's adaptive balancing mechanism, highlighting its role in achieving robust semantic-style fusion without

Fig. 10: Effects of three settings (full DSSI, simplified FSSI, and without DSSI) on style-guided image synthesis.

training. First, we simplify the DSSI module to a fixed-weight variant (Fixed-weight Semantic-Style Integration, FSSI), where the dynamic parameter λ is replaced with a static weight of 0.5 between the prompt and style branches:

$$\mathbf{h}_o^{\text{FSSI}} = \kappa \cdot \left(0.5 \cdot \alpha_p \mathbf{V}_p + 0.5 \cdot \alpha_s \mathbf{V}_s\right) + \boldsymbol{\alpha}_o \mathbf{V}_o. \tag{21}$$

Second, we completely remove the DSSI module, reverting to the vanilla attention formulation: $\mathbf{h}_o = \boldsymbol{\alpha}_p \mathbf{V}_p + \boldsymbol{\alpha}_s \mathbf{V}_s + \boldsymbol{\alpha}_o \mathbf{V}_o$.

We evaluate these three settings, including full DSSI, simplified FSSI, and without DSSI, using a diverse set of style references and content prompts, generating images under consistent hyperparameter (fixed $\kappa = 2.3$). Qualitative results are visualized in Fig. 10, which showcases side-by-side comparisons. Quantitative metrics, including style fidelity and semantic consistency, are reported in Table 3.

The results reveal that the full DSSI module significantly outperforms the ablated variants. Using simplified FSSI or without DSSI, outputs often exhibit noticeable imbalances: strong prompt alignments lead to weak stylization, while dominant style cues cause content distortion. In contrast, the adaptive λ in full DSSI dynamically adjusts to varying λ_p and λ_s, achieving the best metrics and visually coherent results without artifacts.

Table 3: Quantitative ablation results for the DSSI module across metrics including content alignment and style alignment. The optimal setting is marked with (*).

Metric	Full DSSI(*)	Simplified FSSI	Without DSSI
Content ↑	0.287	0.258	0.264
Style ↑	0.750	0.740	0.717

5.3 Application

5.3.1 Multiple Style-guided Image Synthesis

A key advantage of our approach lies in its inherent scalability to multiple style references, enabling the extraction of more robust and intrinsic stylistic representations

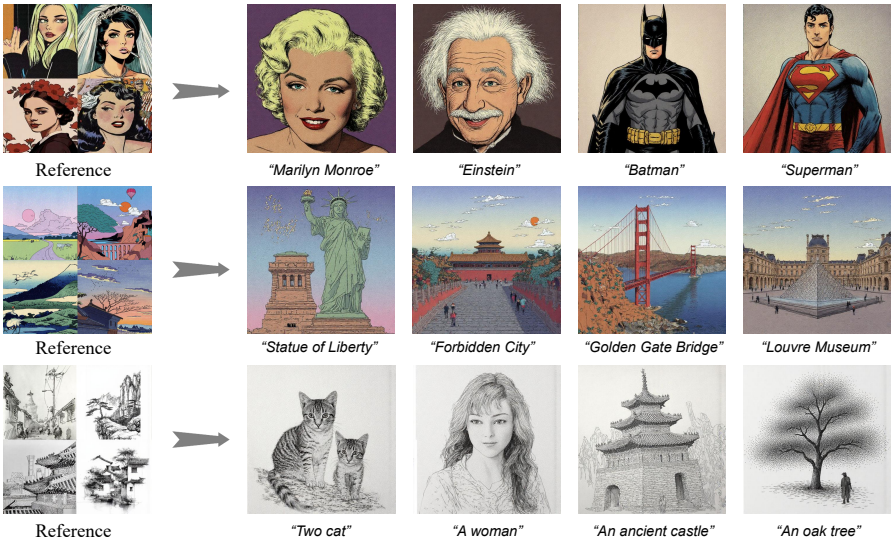


Fig. 11: Multiple style-guide image synthesis.

that transcend the limitations of single-image references. As depicted in Fig. 11, our method seamlessly accommodates four style images by simply concatenating them. When provided with multiple style images that share a coherent artistic theme (such as comic and sketch style), our method can distill intrinsic style representation. The resulting stylized outputs exhibit significantly enhanced style consistency and richness, manifesting as more authentic and professionally coherent artistic renderings.

5.3.2 Image Editing

Beyond single-reference style transfer, our method naturally extends to a wide range of real-world image editing tasks, demonstrating remarkable versatility and generalization across both photorealistic and highly artistic inputs. As illustrated in Fig. 12, our approach seamlessly supports diverse editing operations, including adding objects, non-rigid editing, object replacement and attribute editing (e.g., color, material and style). By changing the textual prompt, regardless of whether the input image is a natural photograph or an already heavily stylized artwork, our method adheres to the user's textual editing instructions while preserving the content and appearance of non-edited regions.

6 Limitation and Discussion

As shown in Fig. 13, the limitations of Sissi can be summarized in two main aspects. (a) Due to the inpainting paradigm, the generated results may inadvertently complete or extend elements from the style reference image at the image junctions, though this does not significantly impact the overall stylized output. (b) When the input is a pure texture image, the model struggles to reasonably arrange and integrate the textures

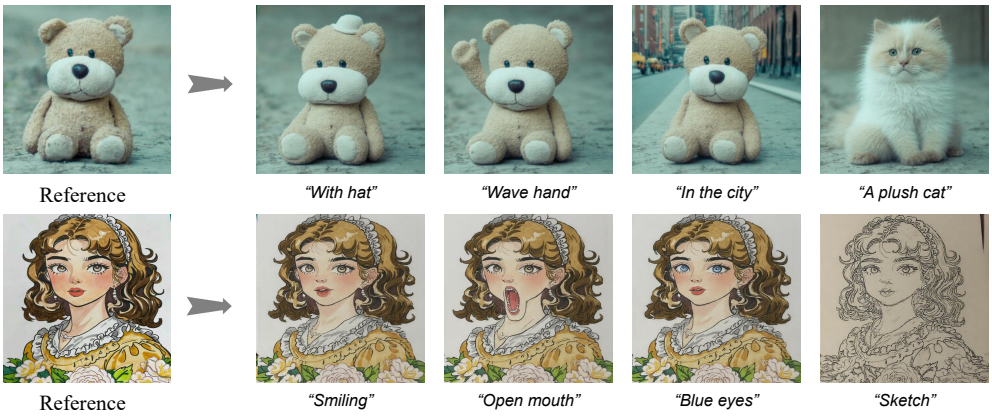


Fig. 12: Image editing results.

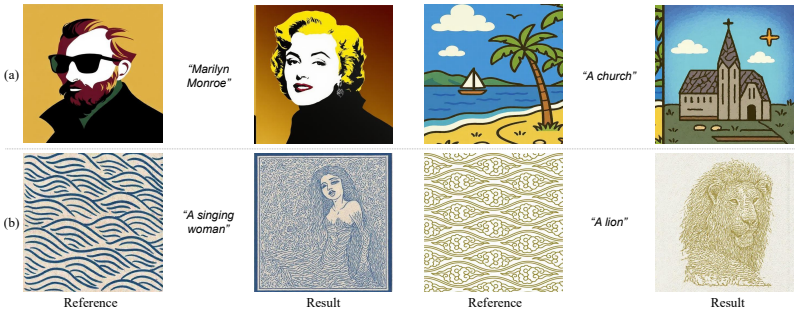


Fig. 13: Bad cases. (a) Inpainting Paradigm Constraint. (b) Texture Input Sensitivity.

within the generated content, thereby compromising style preservation performance. To address these limitations, future research could explore advanced masking strategies or multi-resolution fusion techniques to minimize junction artifacts in inpainting-based methods. Additionally, incorporating texture-specific priors or hierarchical attention mechanisms could improve handling of pure texture inputs, enhancing style fidelity and arrangement in diverse scenarios.

7 Conclusion

In this paper, we introduce Sissi, a training-free style–image-guided synthesis framework that leverages the in-context capabilities of the ReFlow-based inpainting model to achieve robust and coherent semantic–style fusion. Through the proposed Dynamic Semantic–Style Integration module, our method adaptively balances textual semantic prompts with visual style references, enabling high-quality stylized generation with strong fidelity to both content and artistic attributes. Furthermore, Sissi extends naturally to a wide range of downstream applications, including multiple style-guided

image synthesis, and fine-grained image editing, demonstrating its versatility and effectiveness across diverse generative tasks.

References

[1] Cui, X., Li, Z., Li, P., Huang, H., Liu, X., He, Z.: INSTASTYLE: inversion noise of a stylized image is secretly a style adviser. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) European Conference on Computer Vision (ECCV), vol. 15109, pp. 455–472 (2024)

[2] Hertz, A., Voynov, A., Fruchter, S., Cohen-Or, D.: Style aligned image generation via shared attention. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4775–4785 (2024)

[3] Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10146–10156 (2023)

[4] Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Thirty-Eighth AAAI Conference on Artificial Intelligence, pp. 4296–4304 (2024)

[5] Chen, D., Tennent, H., Hsu, C.: Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8619–8628 (2024)

[6] Chung, J., Hyun, S., Heo, J.-P.: 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 (CVPR), pp. 8795–8805 (2024)

[7] Gao, J., Liu, Y., Sun, Y., Tang, Y., Zeng, Y., Chen, K., Zhao, C.: Styleshot: A snapshot on any style. CoRR **abs/2407.01414** (2024)

[8] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: 9th International Conference on Learning Representations (2021)

[9] Huang, L., Wang, W., Wu, Z.-F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-context lora for diffusion transformers. arXiv preprint arxiv:2410.23775 (2024)

[10] Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6830–6839 (2023)

[11] Najdenkoska, I., Sinha, A., Dubey, A., Mahajan, D., Ramanathan, V., Radenovic, F.: Context diffusion: In-context aware image generation. In: European Conference on Computer Vision, pp. 375–391 (2024). Springer

[12] Wu, C.H., Torre, F.: A latent space of stochastic diffusion models for zero-shot image editing and guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7378–7387 (2023)

[13] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. CoRR **abs/2204.06125** (2022)

[14] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Lopes, R.G., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)

[15] Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L., Liu, Z., Qiao, Y., Kot, A.C., Wen, B.: Sinsr: Diffusion-based image super-resolution in a single step. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 25796–25805 (2024)

[16] Wang, J., Yue, Z., Zhou, S., Chan, K.C.K., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision **132**(12), 5929–5949 (2024)

[17] Deng, Y., He, X., Mei, C., Wang, P., Tang, F.: Fireflow: Fast inversion of rectified flow for image semantic editing. CoRR **abs/2412.07517** (2024)

[18] Zhou, Z., Deng, Y., He, X., Dong, W., Tang, F.: Multi-turn consistent image editing. CoRR **abs/2505.04320** (2025)

[19] Ge, S., Park, T., Zhu, J., Huang, J.: Expressive image generation and editing with rich text. International Journal of Computer Vision **133**(7), 4604–4622 (2025)

[20] Deng, Y., He, X., Tang, F., Dong, W.: Z-magic: Zero-shot multiple attributes guided image creator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18390–18400 (2025)

[21] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22500–22510 (2023)

[22] Dong, Z., Wei, P., Lin, L.: Dreamartist: Controllable one-shot text-to-image generation via positive-negative adapter. International Journal of Computer Vision **133**(10), 7037–7053 (2025)

[23] Jin, J., Shen, Y., Zhao, X., Fu, Z., Yang, J.: Unicanvas: Affordance-aware unified real image editing via customized text-to-image generation. International Journal of Computer Vision **133**(6), 3456–3480 (2025)

[24] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, ??? (2023)

[25] Smith, J.S., Hsu, Y., Zhang, L., Hua, T., Kira, Z., Shen, Y., Jin, H.: Continual diffusion: Continual customization of text-to-image diffusion with c-lora. Trans. Mach. Learn. Res. **2024** (2024)

[26] Soboleva, V., Alanov, A., Kuznetsov, A., Sobolev, K.: T-lora: Single image diffusion model customization without overfitting. CoRR **abs/2507.05964** (2025)

[27] Wang, Z., Wang, X., Xie, L., Qi, Z., Shan, Y., Wang, W., Luo, P.: Styleadapter: A unified stylized image generation model. International Journal of Computer Vision **133**(4), 1894–1911 (2025)

[28] Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space (2025). https://arxiv.org/abs/2506.15742

[29] Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning, ICML (2024)

[30] Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023 (2023)

[31] Huang, L., Wang, W., Wu, Z., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-context lora for diffusion transformers. CoRR **abs/2410.23775** (2024)

[32] Shin, C., Choi, J., Kim, H., Yoon, S.: Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7986–7996 (2025)

[33] Zhang, Y., Luo, M., Dong, W., Yang, X., Huang, H., Ma, C., Deussen, O., Lee, T., Xu, C.: Bringing characters to new stories: Training-free theme-specific image generation via dynamic visual prompting. CoRR **abs/2501.15641** (2025)

[34] Xu, Y., Tang, F., Wu, Y., Gao, L., Deussen, O., Yan, H., Li, J., Cao, J., Lee, T.: In-context brush: Zero-shot customized subject insertion with context-aware latent space manipulation. CoRR **abs/2505.20271** (2025)

[35] Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423 (2016)

[36] Huang, X., Serge, B.: Arbitrary style transfer in real-time with adaptive instance normalization. In: IEEE International Conference on Computer Vision (ICCV), pp. 1501–1510 (2017)

[37] Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: European Conference on Computer Vision (ECCV), pp. 702–716 (2016)

[38] Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.-H.: Universal style transfer via feature transforms. In: Advances Neural Information Processing Systems (NeurIPS), pp. 386–396 (2017)

[39] Li, Y., Wang, N., Liu, J., Hou, X.: Demystifying neural style transfer. In: International Joint Conference on Artificial Intelligence (IJCAI) (2017)

[40] Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5880–5888 (2019)

[41] Zhang, Y., Tang, F., Dong, W., Huang, H., Ma, C., Lee, T.-Y., Xu, C.: Domain enhanced arbitrary image style transfer via contrastive learning. In: ACM SIGGRAPH 2022 Conference Proceedings, pp. 1–8 (2022)

[42] An, J., Huang, S., Song, Y., Dou, D., Liu, W., Luo, J.: ArtFlow: Unbiased image style transfer via reversible neural flows. In: IEEE/CVF Conferences on Computer Vision and Pattern Recognition (CVPR), pp. 862–871 (2021)

[43] Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: Stytr2: Image style transfer with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11326–11336 (2022)

[44] Liu, H., Wang, L., Zhang, Y., Yu, Z., Guo, Y.: Samam: Style-aware state space model for arbitrary image style transfer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 28468–28478 (2025)

[45] Cheng, B., Liu, Z., Peng, Y., Lin, Y.: General image-to-image translation with one-shot image guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22736–22746 (2023)

[46] Wang, H., Spinelli, M., Wang, Q., Bai, X., Qin, Z., Chen, A.: Instantstyle: Free lunch towards style-preserving in text-to-image generation. CoRR **abs/2404.02733** (2024)

[47] Qi, T., Fang, S., Wu, Y., Xie, H., Liu, J., Chen, L., He, Q., Zhang, Y.: Deadiff: An efficient stylization diffusion model with disentangled representations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pp. 8693–8702. IEEE, ??? (2024)

[48] Xing, P., Wang, H., Sun, Y., Wang, Q., Bai, X., Ai, H., Huang, R., Li, Z.: CSGO: content-style composition in text-to-image generation. CoRR **abs/2408.16766** (2024)

[49] Deng, Y., He, X., Tang, F., Dong, W.: Z*: Zero-shot style transfer via attention reweighting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6934–6944 (2024)

[50] Deng, Y., He, X., Tang, F., Dong, W.: Z-STAR+: A zero-shot style transfer method via adjusting style distribution. CoRR **abs/2411.19231** (2024) https://doi.org/10.48550/ARXIV.2411.19231 2411.19231

[51] Yu, X., Gu, X., Shi, Z., Sun, J.: Wasserstein style distribution analysis and transform for stylized image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17496–17505 (2025)

[52] Jeong, J., Kim, J., Choi, Y., Lee, G., Uh, Y.: Visual style prompting with swapping self-attention. CoRR **abs/2402.12974** (2024)

[53] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of International Conference on Machine Learning, ICML, vol. 139, pp. 8748–8763 (2021)

[54] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)

[55] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations (2015)

[56] Black Forest Labs: FLUX (2023). https://github.com/black-forest-labs/flux

[57] Krizhevsky, A.: Learning multiple layers of features from tiny images. (2009). https://api.semanticscholar.org/CorpusID:18268744