Through Van Gogh's Eyes: Global Style Transfer with Diffusion Model

Jeongha Lee*[1,2], Yujin Kim*[3], Ghazanfar Ali[4], Suhyun Kim†[5], and Jae-In Hwang†[1]

[1]Korea Institute of Science and Technology, [2]University of Science and Technology, [3] Korea University, [4] Gachon University, [5] Kyung Hee University
wjdgk1029@gmail.com, lakeeye1220@gmail.com, ghazan@gachon.ac.kr, dr.suhyun.kim@gmail.com, hji@kist.re.kr

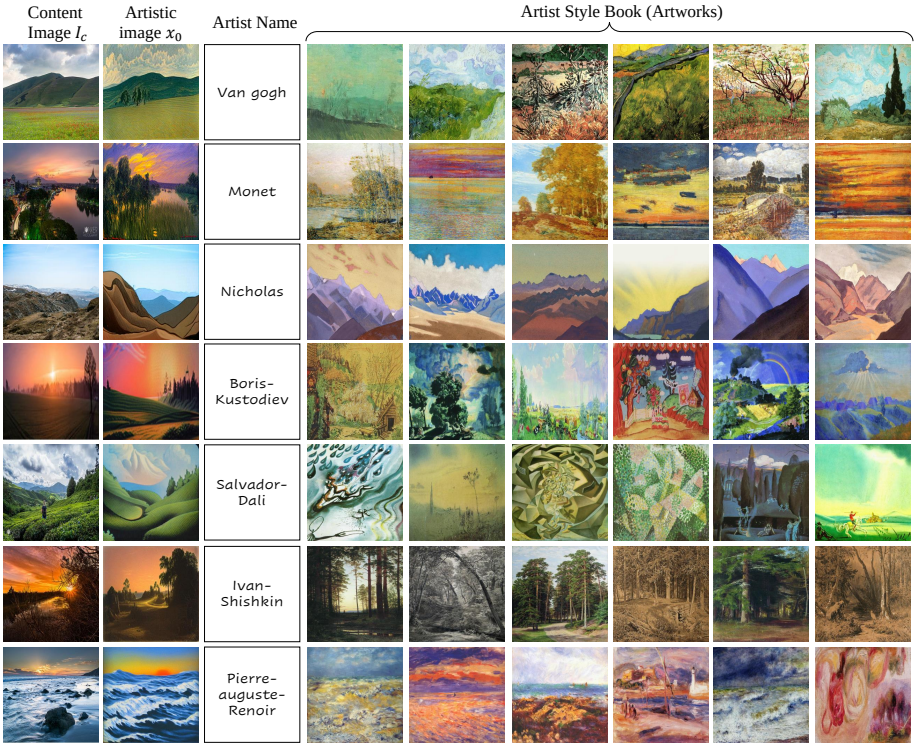

Fig. 1: Artist-level Global Style Transfer results. Given a content image I_c(first column), our framework synthesizes an artistic image x_0 (second column) that reflects the global style of the target artist. The Artist Style Book (remaining columns) displays the top-6 real artworks from the artist's corpus with the highest semantic similarity to x_0, measured by CLIP distance. This demonstrates that our generated results are influenced by a diverse range of the artist's genuine works, successfully capturing the global stylistic distribution rather than overfitting to a single iconic exemplar.

* Equal contribution, † Corresponding author

Abstract. Artistic image synthesis aims to recreate the expressive visual identity of a target artist, yet existing methods often fail to capture an artist's global style. Conventional style transfer methods transfer the style of one or a few reference artworks to a content image in a One-to-One manner, making them effective for artwork-level stylization but limited in representing the broader stylistic distribution of an artist. Text-to-image diffusion models conditioned on artist names, such as '$\sim$ `in Van Gogh style`', offer greater flexibility, but they often suffer from text-induced bias and reproduce patterns from only a few iconic works. To address these limitations, we introduce *Global Style Transfer* (GST), an artistic image synthesis paradigm, in a Many-to-One manner,that aggregates multiple artworks from a target artist and transfers their shared global style to a single content image. For GST, we propose *Global Style Guidance* (GSG), which learns a residual global style offset $\Delta\langle_t$ in the intermediate feature space, or h-space, of a diffusion model under a fixed prompt. By learning artist-level style semantics purely from visual statistics, GSG mitigates text-dependent artistic bias. We further propose *Content Alignment Guidance* (CAG), a training-free perceptual guidance mechanism that preserves the semantic structure of the content image while allowing artist-specific geometric deformation. Experiments on WikiArt demonstrate that GST achieves superior stylistic fidelity, content preservation, and output diversity compared to existing style transfer and diffusion-based artistic synthesis methods.

Keywords: Global Style Transfer · Diffusion Models · Artistic Image Synthesis

1 Introduction

If the great painters of the past were alive today, how would they perceive and depict our modern world on canvas? This question lies at the core of artistic image synthesis, which aims to recreate the expressive visual identity of a target artist. Existing approaches mainly fall into two paradigms: Style Transfer [6, 11, 16] and Text-to-Image (T2I) diffusion-based artistic image synthesis [21, 33]. Conventional style transfer methods extract style from a single or a few reference artworks and transfer it to a content image, effectively reproducing the characteristics of specific artworks, called the *One-to-One* approach. Unfortunately, such instance-level style transfer fails to capture the broader stylistic distribution that defines an artist's overall visual identity, shown in Fig. 2. On the other hand, T2I diffusion models conditioned on artist names (e.g., '$\sim$ `in the style of Van Gogh`') offer a more flexible synthesis pipeline but suffer from text-dependent artistic bias. T2I diffusion model often reproduces textures and compositions from a handful of iconic works, collapsing the stylistic diversity of an artist into a narrow, mode-biased distribution. Consequently, neither paradigm reliably captures the coherent visual identity that defines an artist's global style.

To overcome these limitations, we introduce *Global Style Transfer* (GST), a new paradigm that reformulates artistic image synthesis as a Many-to-One ap-

proach. Instead of relying on a single reference artwork or a text prompt, GST aggregates multiple artworks from a target artist and transfers their shared global style to a single content image, as shown in Fig. 1. This formulation naturally mitigates instance-level bias by capturing coherent style representations across the full breadth of an artist's artworks, and eliminates linguistic dependence by grounding stylistic guidance entirely in visual statistics. As a result, GST enables artist-level style transfer that faithfully reflects the coherent visual identity of the target artist while allowing diverse and expressive content re-rendering.

Concretely, we realize Global Style Transfer through two complementary components. First, we propose *Global Style Guidance* (GSG), which operates in the intermediate h-space of a diffusion U-Net bottleneck. GSG trains a lightweight Style Extraction Function (SEF) to predict a residual global style offset $\Delta\mathbf{h}_t$ from multiple artworks under a fixed, unified text condition (e.g., '`A painting`'), thereby learning purely visual style semantics free from text-induced variance. Second, we propose *Content Alignment Guidance* (CAG), a training-free mechanism that performs DDIM inversion [25] on the content image and applies perceptual guidance at each diffusion timestep. CAG preserves the abstract structure and semantic composition of the content while allowing style-driven geometric deformations, reflecting the expressive intent characteristic of a target artist.

Extensive experiments on WikiArt dataset [27] demonstrate that our framework achieves superior stylistic fidelity and content preservation compared to existing style transfer and T2I diffusion-based artistic synthesis methods, while producing diverse outputs that better reflect the global stylistic spectrum of the target artist. Our main contributions are as follows:

- We introduce ***Global Style Transfer***, a novel many-to-one artistic image synthesis paradigm that learns a unified global style distribution from multiple artworks of a target artist, overcoming the artwork (instance)-level and text-dependent artistic bias of prior methods.
- We propose ***Global Style Guidance*** (GSG), a guidance mechanism operating in h-space that learns artist-level global stylistic semantics only from visual statistics, mitigating linguistic dependence and artistic mode collapse.
- We propose ***Content Alignment Guidance*** (CAG), a training-free perceptual guidance that preserves the semantic integrity of the content while enabling artist-specific style-based deformation during the diffusion process.

2 Related Works

2.1 Style Transfer (1 content image, 1 style image)

Style Transfer aims to generate a stylized image by combining the semantic structure of a content image with the visual characteristics (e.g., texture, color, and brushstroke) of a style image, typically in a *One-to-One* setting (Content image: 1, Style image: 1) [8, 16]. Early Neural Style Transfer (NST) methods [6,15] achieved this by matching feature statistics between content and style

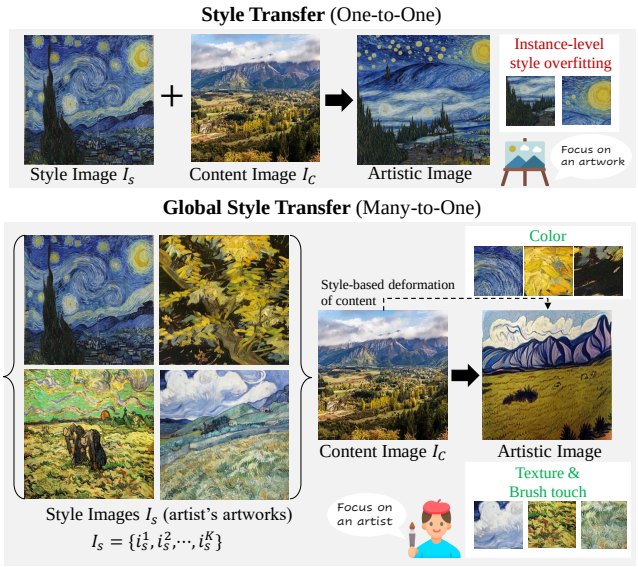

Fig. 2: Comparison between conventional style transfer and the proposed *Global Style Transfer*. (Top) Traditional style transfer applies the style of a single reference artwork to a content image, leading to instance-level overfitting. (Bottom) In contrast, *Global Style Transfer* leverages various artworks from the same artist to learn unified stylistic features, including color palettes, texture patterns, and brushwork, enabling richer style-based content deformation and more faithful artistic image synthesis.

images, while later methods improved stylization quality through normalization [11, 28], attention mechanisms [17, 20]. Recent diffusion-based approaches, such as StyleInjection [2], CSGO [30], InST [34] and Diff-NST [23] further improve style fidelity by injecting style features into the denoising process. However, because these methods rely on only one or few style images, they capture instance-level style rather than the broader stylistic distribution of an artist.

2.2 Style Personalization (× or 1 content image, Few style image)

Recent style personalization methods focus on adapting pretrained text-to-image diffusion models to generate images in a customized artistic style, typically using a small number of reference styles and either no content image or a single content image to be stylized (Content: × or 1, Style: few). These approaches fall into two categories. The first fine-tunes the diffusion backbone itself: DreamBooth [22], Custom Diffusion [12], and StyleDrop [24] fine-tune the model, cross-attention projections, or lightweight adapters, respectively, to bind the style to a unique identifier. The second embeds style into the specific module without modifying model weights: Textual Inversion [5] optimizes a word embedding for the target style, while StyleAligned [9] propagates stylistic consistency across generated

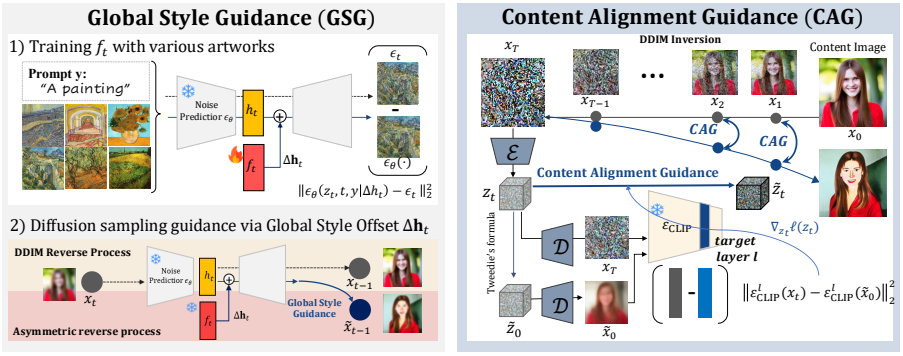

Fig. 3: Overall framework of *Global Style Transfer* (GST). (**Left**) *Global Style Guidance* (GSG) trains a lightweight Style Extraction Function f_t on multiple artworks under the fixed prompt '`A painting`', learning a residual global style offset $\Delta \mathbf{h}_t$ that modulates the U-Net bottleneck representation h_t in a text-independent manner. During reverse diffusion, $\Delta \mathbf{h}_t$ steers the denoising trajectory toward the artist's global style distribution. (**Right**) *Content Alignment Guidance* (CAG) first maps the content image I_c into a noisy latent via DDIM inversion, then applies CLIP-based perceptual guidance at each timestep to align the semantic structure of the generated image with the content while permitting style-driven deformation, yielding artist-level stylization.

images via shared attention at inference. However, these methods often memorize dominant visual patterns, limiting stylistic diversity.

In contrast to both methods, our *Global Style Transfer* learns an artist-level global style distribution from multiple artworks and applies it to a content image in *Many-to-One* manner, enabling faithful artistic synthesis without relying on single-instance style cues or text-dependent supervision.

3 Methods

3.1 Preliminary

Latent Diffusion Models. Latent Diffusion Models (LDM) [21] operate in a lower-dimensional latent space, enabling cost-efficient and scalable text-conditioned image synthesis. Given an image x_0, an autoencoder $\mathcal{E} : \mathbb{R}^k \to \mathbb{R}^d$ maps images into the latent vector $z_0 = \mathcal{E}(x_0)$. A forward diffusion process gradually adds Gaussian noise according to a variance schedule β_t, yielding:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \ \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$. Then, the denoising model ϵ_θ learns to predict the noise $\epsilon_\theta(z_t, t, c)$ given the noisy latent z_t at every time step t with text condition c encoded by the text encoder τ_ϕ during the sampling process as below:

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\tilde{z}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\,\epsilon_\theta^t(z_t) + \sigma_t \epsilon_t, \tag{2}$$

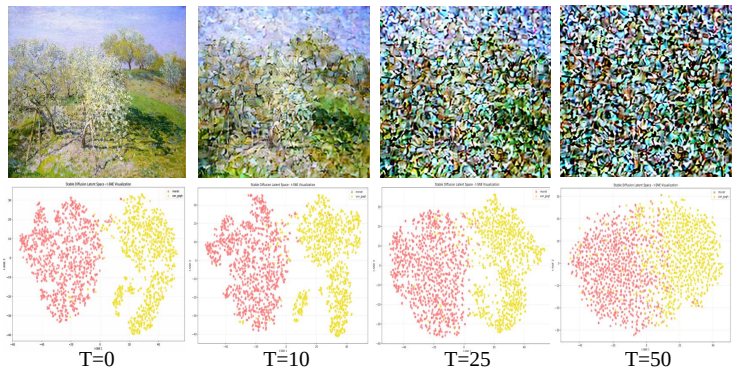

Fig. 4: Time-step robustness of *Global Style Guidance* in h-space. (**Top**) Artistic images generated at diffusion timesteps $t = \{0, 10, 25, 50\}$, showing increasing noise levels as t grows. (**Bottom**) t-SNE visualizations of 500 h_t feature vectors from artworks by Van Gogh (yellow) and Monet (red) at the corresponding t. Despite progressive noise corruption, clusters remain well-separated across all timesteps, validating h-space as a stable domain for artist-level style representation during the diffusion process.

where $\sigma_t = \eta\sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t-1}}$. $\epsilon_\theta^t(z_t)$ denotes the model-predicted noise, abbreviated from $\epsilon_\theta(\mathbf{z}_t, t, c)$, and $\tilde{\mathbf{z}}_0$ means the approximation of the clean latent $\mathbf{z}_0$ at time step t obtained via Tweedie's formula. The training objective of the LDM is defined as:

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{z\sim\mathcal{E}(x),\epsilon_t,t}\left[\|\epsilon_\theta(z_t, t, c) - \epsilon_t\|_2^2\right], \tag{3}$$

After training, the diffusion process starts from a noisy latent z_t and iteratively denoises it to z_0. The LDM then decodes z_0 into pixel space using the decoder $\mathcal{D} : \mathbb{R}^d \to \mathbb{R}^k$, yielding the clean image $x_0 = \mathcal{D}(z_0)$.

Asymmetric Reverse Process for Semantic Guidance. Asyrp [13] discovers that the pre-trained diffusion model inherently contains a semantic space, referred to as the h-space, corresponding to the activations of the U-Net bottleneck layer. To enable controllable generation within LDM, Asyrp reformulates the reverse process by modifying Eq. (2) into an asymmetric form as follows:

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\mathbf{P}_t(\epsilon_\theta^t(z_t|\Delta\mathbf{h}_t)) + \mathbf{D}_t(\epsilon_\theta^t(z_t)) + \sigma_t\epsilon_t, \tag{4}$$

where $\mathbf{P}_t$ denotes the predicted $\tilde{\mathbf{z}}_0$, and $\mathbf{D}_t$ represents the denoising direction toward $\mathbf{z}_t$. The asymmetric control updates only $\epsilon_\theta^t(\mathbf{z}_t)$ via $\epsilon_\theta^t(\mathbf{z}_t|\Delta\mathbf{h}_t)$, adding $\Delta\mathbf{h}_t$ as a residual to the U-Net bottleneck features $\mathbf{h}_t$, while preserving $\mathbf{D}_t$. This enables semantic manipulation in latent space without compromising reconstruction fidelity. Consequently, the h-space exhibits stable and transferable semantics across time steps, supporting consistent, and multi-attribute guidance for latent diffusion models.

3.2 Global Style Transfer

We introduce *Global Style Transfer*, a new paradigm that extends the conventional one-to-one style transfer setting to a many-to-one formulation, as illustrated in Fig. 2. Unlike traditional methods that rely on a single style instance to guide the appearance of a content image, our approach aggregates multiple artworks from a single artist to learn a unified and consistent representation of that artist's global style.

Given a content image I_c and a set of style exemplars $I_s = \{i_s^1, i_s^2, \cdot\cdot, i_s^K\}$ drawn from an artist's K artworks, *Global Style Guidance* (GSG) guides the diffusion model that encapsulates artist-specific semantic and style cues across various artworks. To preserve the semantic integrity of the content, we further introduce *Content Alignment Guidance* (CAG), which maintains the abstract structure and semantics of I_c while allowing flexible, style-based deformation. Through CAG, the content image is re-rendered with the distinctive visual characteristics of the target artworks without compromising its original composition.

Global Style Transfer offers two key advantages over text-conditioned artistic synthesis. (1) **Text-independent Guidance**: Rather than relying on textual prompts (e.g., '$\sim$ `in the style of Van Gogh`'), *Global Style Transfer* extracts visual semantics directly from the aggregated style artworks. (2) **Artistic bias mitigation in diffusion-based artistic image synthesis**: Vanilla Text-to-Image diffusion models conditioned on artist names tend to exhibit stylistic bias, over-representing textures and compositions from a few iconic works. By guiding a global style prior from multiple style exemplars, our framework regularizes the diffusion process, mitigating such artistic mode bias and yielding diverse yet faithful images that better reflect the full stylistic spectrum of the artist.

3.3 Global Style Guidance in h-Space

Finding Artist's Global Style. To guide the diffusion model toward the global style of a specific artist, we define the residual global style offset $\Delta\mathbf{h}_t$, which modulates the U–Net bottleneck representation h_t at each timestep t:

$$h_t = h_t + w \cdot \Delta\mathbf{h}_t, \tag{5}$$

where w controls the strength of stylistic modulation. The residual global style offset $\Delta\mathbf{h}_t$ captures the global stylistic semantics of the artist, enhancing the U-Net's intermediate representations with consistent style attributes. To estimate $\Delta\mathbf{h}_t$, we define a *Style Extraction Function* (SEF) f_t that transforms the given h_t into global style–conditioned feature:

$$\Delta\mathbf{h}_t = f_t(h_t; \theta). \tag{6}$$

f_t is a lightweight multilayer perceptron (MLP) with one hidden layer parameterized by θ and conditioned on timestep t. Specifically, we apply the *zero-initialization* strategy, which initializes the weights of f_t to zero [33] to avoid stochastic bias from random initialization. Zero-initialization ensures training

Algorithm 1 Global Style Guidance Algorithm

Require: Style exemplars (=artworks) I_s, Style Extraction Function (SEF) f_t, `Model`(Latent diffusion model) $\epsilon_\theta(\cdot, t)$, text encoder τ_ϕ, image decoder $\mathcal{D}$, Prompt y
Output: SEF f_t

$y =$'`A painting`' {All artworks are paired with this prompt}
for *epoch* $= 0$ **to** E **do**
  for $k = 1$ **to** K **do**
    /*Diffusion Forward Process*/
    for $t = 0$ **to** $T - 1$ **do**
      $z_0^k = \mathcal{E}(I_s^k)$ {instance style image sampling $i_s^k \in I_s$}
      $\epsilon_t \sim \mathcal{N}(0, I)$
      $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$
    end for
    /*Diffusion Denoising Process*/
    for $t = T - 1$ **to** 0 **do**
      $\mathcal{L}_{\text{SEF}} := \mathbb{E}_{z\sim\mathcal{E}(x),\epsilon_t,t}\left[\|\epsilon_\theta(z_t, t, \tau_\phi(y)|\Delta\mathbf{h}_t) - \epsilon_t\|_2^2\right]$
      //$\Delta\mathbf{h}_t = f_t(h_t; \theta)$ from Eq. (6)
      $f_t(h_t; \theta) \leftarrow f_t(h_t; \theta) - \eta\nabla_{f_t}\mathcal{L}_{\text{SEF}}$
      $\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\mathbf{P}_t(\epsilon_\theta^t(z_t|\Delta\mathbf{h}_t)) + \mathbf{D}_t(\epsilon_\theta^t(z_t)) + \sigma_t\epsilon_t$
    end for
  end for
end for
Return $f_t(h_t; \theta)$

Algorithm 2 Global Style Transfer Algorithm

Require: Content Image I_c, `Model`(Latent diffusion model) $\epsilon_\theta(\cdot, t)$, text encoder τ_ϕ, image decoder $\mathcal{D}$, prompt y, CLIP image encoder $\mathcal{E}_{\text{CLIP}}$
Output: Artistic image x_0

$\mathbf{z}_T \leftarrow$ DDIM Inversion with respect to the I_c
for $t = T - 1$ **to** 1 **do**
  $\epsilon_t \sim \mathcal{N}(0, I)$
  //obtain $\Delta\mathbf{h}_t$ from Algorithm 1
  $\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\mathbf{P}_t(\epsilon_\theta^t(z_t|\Delta\mathbf{h}_t)) + \mathbf{D}_t(\epsilon_\theta^t(z_t)) + \sigma_t\epsilon_t$
  //Approximate $\tilde{z}_0$ from Tweedie formula [4] and $\tilde{x}_0 \approx \mathcal{D}(\tilde{z}_0|z_t)$
  $\ell(z_t) = \|\mathcal{E}_{\text{CLIP}}^l(\tilde{x}_0) - \mathcal{E}_{\text{CLIP}}^l(x_t)\|_2$
  Content Alignment Guidance
  $\tilde{z}_t \leftarrow z_t - s\nabla_{z_t}\ell(z_t)$
  //Classifier Free Guidance
  $\tilde{\epsilon}_t \leftarrow \epsilon_\theta(\tilde{z}_t, t, \varnothing) + g(\epsilon_\theta(\tilde{z}_t, t, \tau_\phi(y)) - \epsilon_\theta(\tilde{z}_t, t, \varnothing))$
end for
Return $x_0 = \mathcal{D}(\tilde{\mathbf{z}}_0)$

process begins from an unbiased state and learns purely data-driven global style semantics. Then, we train the f_t with multiple artworks using the noise reconstruction loss same as Eq. (3):

$$\mathcal{L}_{\text{SEF}} := \mathbb{E}_{z\sim\mathcal{E}(x),\epsilon_t,t}\left[\|\epsilon_\theta(z_t, t, \tau_\phi(y)|\Delta\mathbf{h}_t) - \epsilon_t\|_2^2\right]. \tag{7}$$

Table 1: Comparison of implicit function learning between Asyrp and our framework.

Difference	**Asyrp**	**Global Style Transfer (Ours)**
Goal	Instance-conditional attribute manipulation (Image Editing)	Distribution-level global style training
Supervision signal	Text-driven attribute change	Only visual statistics across artworks (Text-independent)
Text conditioning on training f_t	Explicit source–target prompts (high variance), e.g., `'A girl'` (y_{source}) → `'A smiling girl'`(y_{ref})	Fixed prompt (zero variance) e.g., `'A painting'`
Training data	Single (or few) source images	Hundreds of artworks from a single artist
Loss function for learning f_t	CLIP directional editing loss + reconstruction loss: $\lambda_{\text{CLIP}} L_{\text{direction}}\left(P_t^{\text{edit}}, y_{\text{ref}}; P_t^{\text{source}}, y_{\text{source}}\right) + \lambda_{\text{recon}} \left\|x_t^{\text{edit}} - x_t^{\text{source}}\right\|_2$	Noise reconstruction loss over artist distribution: $\mathbb{E}_{z\sim\mathcal{E}(x),\,\epsilon_t,\,t}\left[\|\epsilon_\theta(z_t, t, \tau_\phi(y) \mid \Delta h_t) - \epsilon_t\|_2^2\right]$

Specifically, to ensure that training focuses purely on visual statistics rather than text conditioning, all artworks images I_s are paired with the fixed simple prompt y, '`A painting`'. By enforcing a unified prompt across all artworks, the variance of text conditioning effectively approaches zero, allowing f_t to rely purely on visual cues for capturing global style semantics. Consequently, our SEF learns global stylistic representations directly from the visual semantics of the artworks rather than from linguistic information, resulting in visually grounded and unbiased *Global Style Guidance*. After training, the diffusion process proceeds toward a unified global style distribution, enabling the generation of images that reflect the coherent stylistic identity of the target artist without additional textual conditioning. The overall algorithms of GSG is described in Algorithm 1.

Time-Step Robustness of Global Style Guidance. To motivate our design choice of performing *Global Style Guidance* (GSG) in the intermediate h-space, we analyze whether h-space provides a stable representation of global style across diffusion timesteps. As shown in Fig. 4, 500 artworks per artist form well-separated clusters in h-space even under Gaussian noise across different timesteps. This demonstrates that h_t remains robust to noise and timestep variation while preserving artist-level semantic global style, allowing GSG to maintain coherent stylization throughout the diffusion generative process.

Difference between Asyrp and our Global Style Transfer. Although our framework shares the same implicit function architecture as Asyrp [13], the two methods differ fundamentally in objective and supervision, as summarized in Tab. 1. Asyrp uses f_t as an instance-level editing operator for single-image manipulation, optimized with a CLIP directional loss between source–reference text prompts ('`A girl`' → '`A smiling girl`'). Consequently, the optimization is inherently text-dependent, aligning f_t with semantic attribute directions under highly varying prompt conditions. In contrast, we optimize f_t to model the global stylistic distribution of an artist from hundreds of artworks using only visual supervision. By fixing the text condition y to a constant prompt (e.g., '`A painting`'), we eliminate text-induced variance, allowing f_t to be learned purely from artwork visual statistics via noise reconstruction.

3.4 Content Alignment Guidance

Motivated by how artists reinterpret real-world objects, preserving abstract structure while expressing unique artistic style, we propose the training-free *Content Alignment Guidance* (CAG). In Neural Style Transfer (NST) [23, 35], style-based deformation of content can be desirable, as artistic styles often involve intentional geometric or structural alterations beyond color and texture changes. Inspired by NST, our CAG aims to preserve the recognizable structure of the original content while allowing controlled, style-driven deformations that reflect the artist's expressive intent.

We first perform DDIM inversion [25] to map the input content image into its noisy latent vector. Then, we measure perceptual similarity using the CLIP image encoder $\mathcal{E}_{\text{CLIP}}$ within a Stable Diffusion [21] between the generated image $x_t = \mathcal{D}(z_t)$ at every timestep t and approximated image $\tilde{x}_0 \approx \mathcal{D}(\tilde{z}_0|z_t)$ via Tweedie's formula [4]:

$$\ell(z_t) = \|\mathcal{E}^l_{\text{CLIP}}(\tilde{x}_0) - \mathcal{E}^l_{\text{CLIP}}(x_t)\|_2, \tag{8}$$

where $\mathcal{E}^l_{\text{CLIP}}$ extracts high-level semantic features from layer l of CLIP. Since the approximated image $\tilde{x}_0$ preserves the abstract structure of the content image, aligning its feature representation with that of the generated image x_t encourages structural consistency while allowing style-driven deformation. Following the SDE formulation [26], we incorporate this perceptual loss as the guidance by replacing the unconditional score $\nabla z_t \log p(z_t)$ with the conditional score $\nabla_{z_t} \log p(z_t|\ell)$, decomposed as:

$$\nabla_{z_t} \log p(z_t|\ell) = \nabla_{z_t} \log p(z_t) + s\, \nabla_{z_t} \log p(\ell|z_t), \tag{9}$$

where s is a parameter controlling the guidance strength. The second term is interpreted as the gradient of the perceptual loss with respect to the image latent $\nabla_{z_t}\ell(z_t)$. Formally, the guided image latent is expressed as:

$$\tilde{z}_t = z_t - s\nabla_{z_t}\ell(z_t), \tag{10}$$

where s controls the guidance strength. Consequently, CAG refines the latent trajectory in a *training-free* manner at each timestep to preserve content structure while enabling artist-specific compositional and geometric transformations. The overall algorithm and framework are illustrated in Algorithm 2 and Fig. 3.

4 Experiments

Datasets. We define the style reference images in the WikiArt dataset [27], which includes over 80,000 artworks created by more than 1,000 artists across 27 art movements. For content images, we utilize the VanGogh2Photo dataset [36] , which contains real-world photographic scenes paired with artistic counterparts, enabling consistent evaluation of content preservation in artistic synthesis. This dataset combination serves as a standard benchmark for the style transfer.

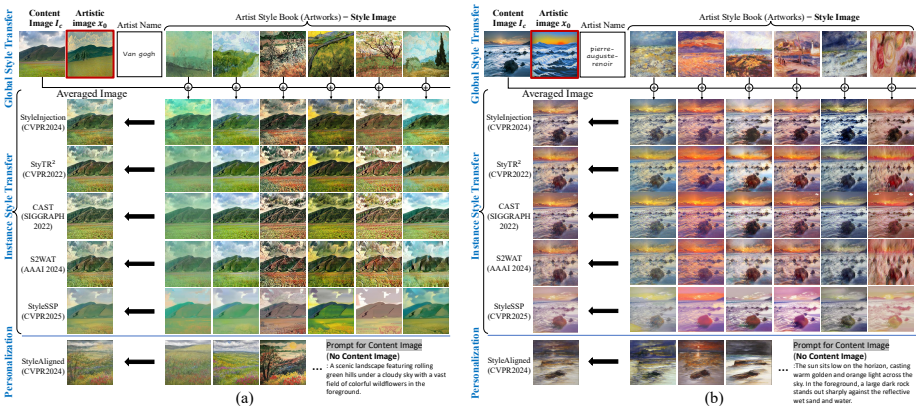

Fig. 5: Comparison of Global Style Transfer (GST) with representative style transfer baselines. We compare GST with instance-level style transfer and style personalization methods, using artworks from the same artist as style exemplars: (a) Van Gogh and (b) Pierre-Auguste Renoir. Since baseline methods utilize only a single style image I_s, we perform stylization separately using multiple artworks from the same artist and average the resulting outputs to approximate artist-level style. Baselines preserve only low-level appearance cues and produce inconsistent global styles, whereas GST learns the artist-level style and enables coherent global stylization.

Evaluation Metrics. We evaluate using five metrics: FID [10] and ArtFID [29] for global style fidelity, CFSD [2] for content preservation, and CLIP-Div [1] and 1-Precision [14,18] for stylistic diversity and memorization, respectively. Detailed definitions are provided in the Appendix.

4.1 Main Results

Sample Visualization. Fig. 1 shows the artistic images produced by the Global Style Transfer. This is achieved as the diffusion model performs *Global Style Guidance* (GSG) within the intermediate feature space (h-space), ensuring semantic stability across time steps and producing consistent style transformation throughout the generative process.

Comparison with Style Transfer Methods. Our method introduces a new paradigm, *Global Style Transfer* (GST), which fundamentally differs from conventional style transfer. Traditional style-transfer methods operate in a one-to-one setting, where a single content image I_c is transferred to an artistic image with only single style image I_s through feature injection or attention modulation. While effective for instance-level stylization, such approaches cannot capture an artist-level style distribution spanning hundreds or thousands of artworks. To verify this limitation, we compare our method with representative Style Transfer baselines, including StyleInjection [2], StyTR^2 [3], CAST [35], S2WAT [32], StyleSSP [31], as well as the shared-attention personalization method StyleAligned

Table 2: Quantitative comparison across artists. ArtFID measures style and content fidelity, CFSD measures content preservation, CLIP-Div and 1-Prec. assess stylistic diversity and memorization-avoidance. CLIP-Div is reported as *Ours/Real* for direct comparison with the real artwork corpus. **Bold**: best; underline: second best.

Artist	Method	FID (↓)	ArtFID (↓)	CFSD (↓)	CLIP-Div (↑)	1−Prec (↑)
Van Gogh	S.D	11.97	23.78	0.7428	0.178	0.257
	Textual Inversion	**7.67**	**13.68**	0.1674	0.220	0.550
	Custom Diffusion	9.49	14.71	**0.1208**	0.289	0.933
	LoRA based Fine-tuning	11.01	21.38	0.1886	0.261	0.913
	Ours	9.46	19.25	0.2896	**0.297**/0.335	**0.988**
Chagall	S.D	16.26	32.44	0.3117	0.224	0.844
	Textual Inversion	**11.72**	**21.15**	0.1683	0.238	0.877
	Custom Diffusion	18.06	26.70	**0.1108**	0.289	0.962
	LoRA based Fine-tuning	17.30	32.59	0.1674	0.266	0.928
	Ours	13.25	26.39	0.2626	**0.311**/0.293	**0.977**
Renoir	S.D	16.72	33.70	0.1584	0.225	0.987
	Textual Inversion	**12.14**	**21.99**	**0.1100**	0.238	0.895
	Custom Diffusion	15.15	22.77	0.1160	0.294	0.987
	LoRA based Fine-tuning	15.36	29.23	0.1749	0.279	0.988
	Ours	12.27	24.70	0.1566	**0.318**/0.313	**0.999**

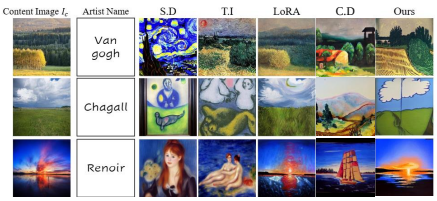

Fig. 6: Visual comparison of previous style personalization methods and vanilla Stable Diffusion. Qualitative results corresponding to Tab. 2 are shown, demonstrating the comparative performance of each method in terms of style alignment and visual coherence.

[9], grouped separately in Fig. 5. Since these baselines take a single style image, we run each on the artist's artworks individually and average the resulting outputs to approximate an artist-level style. As shown in Fig. 5, this strategy fails to recover coherent artist-level style. Existing methods primarily capture low-level appearance cues such as dominant colors or coarse brush textures from each exemplar, but fail to preserve consistent stylistic characteristics shared across the artist's full corpus. Even averaging multiple stylized outputs does not meaningfully represent the global style distribution, often producing blurred or inconsistent results. In contrast, our method directly learns the artist-level style manifold from multiple artworks, enabling coherent global stylization while avoiding overfitting to specific exemplars.

Comparison with Style Personalization Methods. We compare GST with representative personalization methods, including Textual Inversion (TI) [5], which optimizes a token embedding over the full artwork collection, DreamBooth [22] with LoRA-based fine-tuning, and Custom Diffusion [12], which fine-tunes only the cross-attention layers. We also provide a qualitative comparison with StyleAligned [9] in Fig. 5. For a fair comparison in the GST setting, we train each baseline on the full artwork collection of each artist rather than using the few-shot setting adopted in their original formulations.

As shown in Tab. 2, GST achieves consistently strong performance across all three artists, demonstrating effective modeling of artist-level global style. GST achieves competitive ArtFID, indicating strong overall style and content fidelity without overfitting to content structure. It also achieves the highest CLIP-Diversity across all artists, showing that GST captures a broader stylistic distribution rather than collapsing to a narrow set of visual patterns. Notably, the CLIP-Diversity values of GST are highly consistent with those computed from the real artwork collections (reported as Ours / Real), suggesting that GST

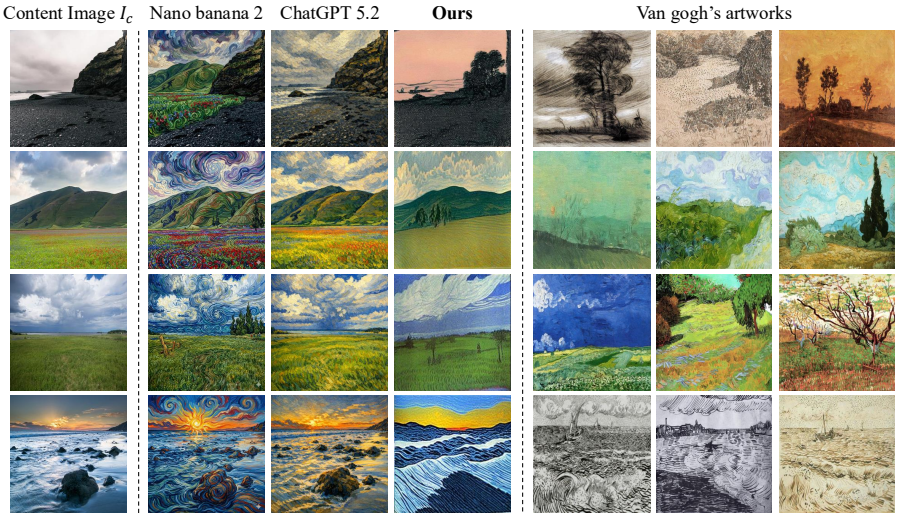

Fig. 7: Visual comparison with T2I models (Nano Banana 2, ChatGPT 5.2) prompted with '`Van Gogh style`' for a given content image I_c. The rightmost columns display the top three Van Gogh artworks with the lowest CLIP distance to our generated result.

successfully reproduces the diversity of the true artist-level style distribution. The 1-Precision results further show that GST best avoids memorization while preserving stylistic consistency. In contrast, prior personalization methods often rely on memorized exemplar patterns or overfit to dominant visual motifs, as also observed in Fig. 6. For example, Textual Inversion is limited by its fixed-dimensional token embedding, while fine-tuning methods often collapse toward iconic compositions. Consequently, our method scales effectively with increasing artwork diversity and better captures the full artist-level style distribution.

4.2 Analysis of Global Style Transfer

Analysis of Text-Independence in Global Style Transfer. We analyze whether the learned style representation depends on text prompts. Since the Style Extraction Function (SEF) is trained using a large collection of artworks, we use a fixed prompt, '`A painting`', during training to minimize textual influence and encourage the model to learn style primarily from visual cues. To evaluate prompt sensitivity, we replace the prompt with alternative generic prompts during training SEF and compute the corresponding style offset Δh.

Fig. 8: The sensitivity of prompt for training the Style Extraction Function f_t.

Table 3: Ablation study on Content Alignment Guidance (CAG). Removing CAG leads to a consistent increase in both FID and ArtFID, demonstrating its crucial role in preserving content fidelity under global style modulation

Method	FID	ArtFID
Ours w/o CAG	11.05	22.45
Ours	**9.46**	**19.25**

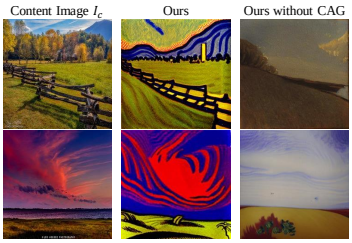

Fig. 9: Qualitative ablation on CAG. Without CAG, the images exhibit content degradation despite strong stylization.

As shown in Fig. 8, the generated images remain nearly identical across prompts, indicating that the learned representation is largely text-independent and primarily captures artist-level visual style.

Analysis of Stylistic Bias in T2I Models. We compare our framework with large-scale text-to-image baselines, Nano Banana 2 [7] and ChatGPT 5.2 [19], using the prompt '`Van Gogh style`'. As shown in Fig. 7, these baselines exhibit strong stylistic bias, repeatedly generating images dominated by iconic Van Gogh features such as swirling patterns and thick brushstrokes, often resembling Starry Night. In contrast, our Global Style Transfer learns a unified artist-level representation, mitigating such bias and producing diverse outputs that better reflect the true global style distribution.

Effect of the Content Alignment Guidance. To verify the effectiveness of CAG, we compare the full methods with a variant without CAG. As shown in Fig. 9 and Tab. 3, removing CAG leads to noticeable content degradation, reflected by a significant increase in ArtFID, which measures both style and content fidelity. These results demonstrate that CAG plays a crucial role in preserving content under strong global style modulation. Additional ablations on the embedding layer selection and training epochs are provided in Figs. 10 and 11.

Effect of Layer Level for Content Alignment Guidance. We examine how applying Content Alignment Guidance (CAG) at different layers of the CLIP image encoder affects the preservation of content structure during global style transfer. As shown in Fig. 10, applying CAG at lower or mid-level layers (1–9) produces unintended artifacts, such as house-like structures absent from the content image I_c, indicating poor preservation of global composition. This occurs because lower layers mainly encode low-level features such as edges, color and textures. In contrast, applying CAG at Layer 11 preserves the original scene structure and yields semantically consistent results, showing that higher-level semantic features provide more reliable content alignment.

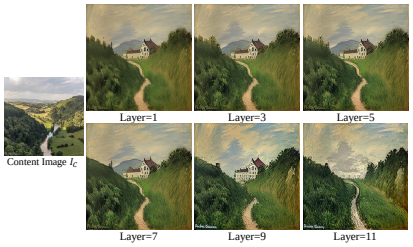

Fig. 10: Effect of layer level of CLIP encoder in Content Alignment Guidance for content preservation.

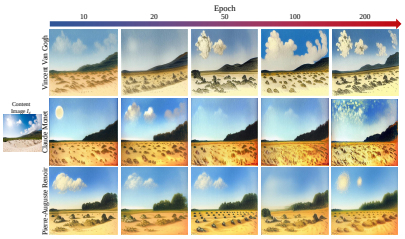

Fig. 11: Effect of training epochs on *Style Extraction Function*: higher epochs yield stronger artist-specific characteristics.

Effect of Training Epochs on Style Extraction Function. We investigate the effect of training epochs on the *Style Extraction Function* (SEF) f_t to evaluate how training duration influences global style learning. We train SEF with five epoch settings (10, 20, 50, 100, 200) using artworks from three representative artists: Van Gogh, Monet, and Renoir. As shown in Fig. 11, models trained for only 10 or 20 epochs fail to capture distinctive stylistic characteristics, producing visually similar results across artists. As training progresses, the model gradually learns artist-specific brushwork, composition, and color distributions. At 200 epochs, the generated images exhibit clear stylistic separation, indicating that sufficient training is essential for learning a stable global style representation.

5 Conclusion

We propose *Global Style Transfer*, a novel artistic image synthesis paradigm that represents the global artistic style of a target artist from multiple artworks. Through *Global Style Guidance* and *Content Alignment Guidance*, our method captures artist-level stylistic semantics while preserving the content structure and allowing style-driven geometric deformation. By moving beyond instance-level style transfer and text-dependent artistic synthesis, GST provides a more faithful and unbiased way to model an artist's coherent visual identity.

Acknowledgment

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (RS-2026-25516375), the Korea Institute of Science and Technology (KIST) Institutional Program (No. 26E0212). This work was supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2026-RS-2023-00258649,33%).

References

1. Alanov, A., Titov, V., Nakhodnov, M., Vetrov, D.: Styledomain: Efficient and lightweight parameterizations of stylegan for one-shot and few-shot domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2184–2194 (2023)
2. 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. pp. 8795–8805 (2024)
3. 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. pp. 11326–11336 (2022)
4. Efron, B.: Tweedie's formula and selection bias. Journal of the American Statistical Association **106**(496), 1602–1614 (2011)
5. 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. arXiv preprint arXiv:2208.01618 (2022)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
7. Google: Gemini image generation: Create images with gemini. `https://gemini.google/overview/image-generation/` (2026), accessed: 2026-01-22
8. Gu, S., Chen, C., Liao, J., Yuan, L.: Arbitrary style transfer with deep feature reshuffle. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8222–8231 (2018)
9. Hertz, A., Voynov, A., Fruchter, S., Cohen-Or, D.: Style aligned image generation via shared attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4775–4785 (2024)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
11. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
12. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2023)
13. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. arXiv preprint arXiv:2210.10960 (2022)
14. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in neural information processing systems **32** (2019)
15. Li, Y., Wang, N., Liu, J., Hou, X.: Demystifying neural style transfer. arXiv preprint arXiv:1701.01036 (2017)
16. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. Advances in neural information processing systems **30** (2017)
17. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6649–6658 (2021)

18. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: International conference on machine learning. pp. 7176–7185. PMLR (2020)
19. OpenAI: Chatgpt: Get answers. find inspiration. be productive. `https://openai.com/ko-KR/index/chatgpt/` (2026), accessed: 2026-01-22
20. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5880–5888 (2019)
21. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
22. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
23. Ruta, D., Tarrés, G.C., Gilbert, A., Shechtman, E., Kolkin, N., Collomosse, J.: Diff-nst: Diffusion interleaving for deformable neural style transfer. In: European Conference on Computer Vision. pp. 50–66. Springer (2024)
24. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: Text-to-image generation in any style. arXiv preprint arXiv:2306.00983 (2023)
25. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
26. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
27. Tan, W.R., Chan, C.S., Aguirre, H.E., Tanaka, K.: Improved artgan for conditional synthesis of natural image and artwork. IEEE Transactions on Image Processing **28**(1), 394–409 (2018)
28. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
29. Wright, M., Ommer, B.: Artfid: Quantitative evaluation of neural style transfer. In: DAGM German Conference on Pattern Recognition. pp. 560–576. Springer (2022)
30. 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. arXiv preprint arXiv:2408.16766 (2024)
31. Xu, R., Xi, W., Wang, X., Mao, Y., Cheng, Z.: Stylessp: Sampling startpoint enhancement for training-free diffusion-based method for style transfer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18260–18269 (2025)
32. Zhang, C., Xu, X., Wang, L., Dai, Z., Yang, J.: S2wat: Image style transfer via hierarchical vision transformer using strips window attention. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 7024–7032 (2024)
33. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
34. 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)

35. 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)
36. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)