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

Anonymous ECCV 2026 Submission

Paper ID #10389

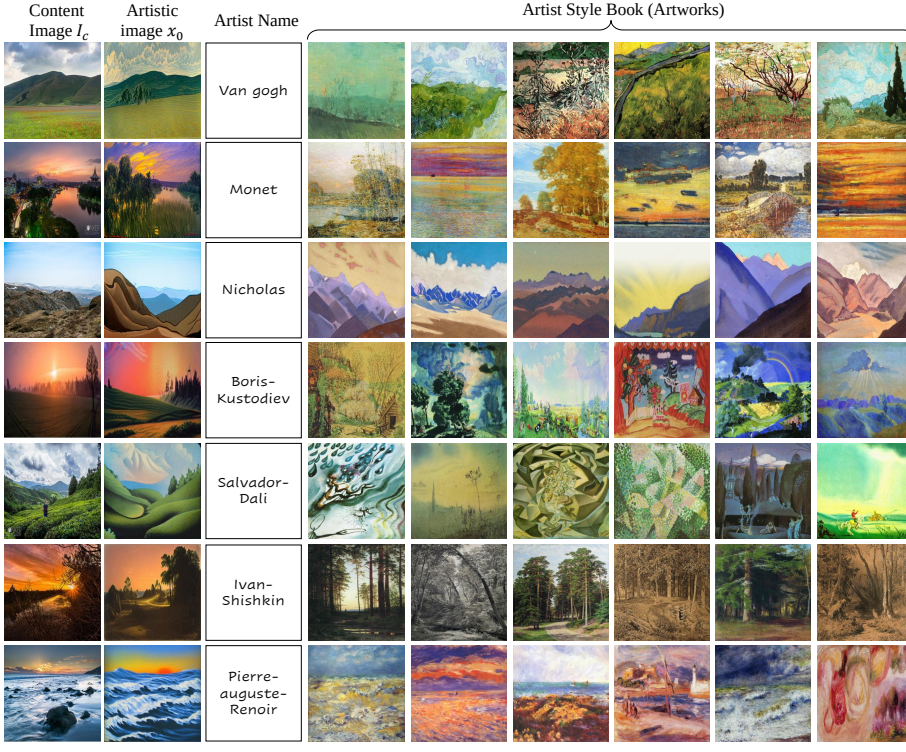

Fig. 1: Artist-level Global Style Transfer results. For each content image I_c (first column), our framework synthesizes an artistic image x_0 (second column) that reflects the global style of the target artist. To visualize the source of stylistic influence, the Artist Style Book section displays the top six real artworks from the artist's corpus that exhibit the highest semantic similarity (i.e., the lowest CLIP distance) to the generated image x_0. This explicitly demonstrates that our generated results are actively influenced by a diverse range of the artist's genuine works, successfully capturing their comprehensive stylistic distribution rather than overfitting to a single iconic exemplar.

Abstract. Diffusion models have achieved strong performance in artistic image synthesis, yet they still suffer from stylistic bias: when instructing the diffusion model to create 'Van Gogh-style' images, we observed that the model tends to repeatedly generate textures and compositions

characteristic of a narrow subset of iconic works. This bias limits the model's ability to fully represent an artist's stylistic diversity. To address this, we introduce *Global Style Transfer* (GST), a text-independent framework that learns an artist's unified style distribution by training global visual statistics from hundreds of artworks. GST guides the diffusion process in the intermediate feature space (**h**-space), enabling generation that reflects an artist's global stylistic spectrum rather than relying on prompt-specific cues or memorized exemplars. In addition, we propose *Content Alignment Guidance* (CAG), a training-free guidance that aligns the semantic structure of a given content image while permitting flexible, style-based deformation. CAG preserves content identity without constraining artistic variation, allowing structural reinterpretations that naturally arise in artistic expression. Experiments on WikiArt benchmarks demonstrate that our method produces images with improved style consistency, reduced prompt-induced bias, and greater fidelity to global artistic semantics compared to the vanilla diffusion model. Our findings establish GST as a new direction for bias-robust artistic image generation.

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? While we can only infer their unique perspectives through their surviving artworks, techniques like style transfer attempt to imitate their aesthetics in the digital realm. However, existing style transfer models do not truly transfer a "painter's style." Because they extract stylistic features from a single or a few reference paintings [**?**], they merely apply the specific traits of those individual artworks to a content image described in Fig. 2. This achieves the style transfer of a specific artwork, but fails to capture the comprehensive style of the artist. Furthermore, while recent large-scale text-to-image diffusion models [3, 5] can stylize images based on text prompts (e.g., "in the style of Van Gogh") [**?**], they suffer from a similar limitation due to training data imbalance. These models are heavily biased toward a few overrepresented iconic artworks seen during their training [**?**], simply mimicking their superficial characteristics rather than reflecting the artist's full aesthetic spectrum as shown in Fig. 5. To resolve these limitations, we propose *Global Style Transfer* (GST), which learns from a large-scale collection of an artist's artworks to transfer their unbiased, global style to a content image. GST goes beyond superficial imitation; it aims to capture the true global style of a painter by learning and reproducing the aesthetic tendencies and compositional characteristics consistently exhibited across their entire body of work.

To achieve *Global Style Transfer* (GST), we propose a novel text-independent framework comprising two key components: *Global Style Guidance* (GSG) and *Content Alignment Guidance* (CAG). By employing a single, fixed prompt (e.g.,

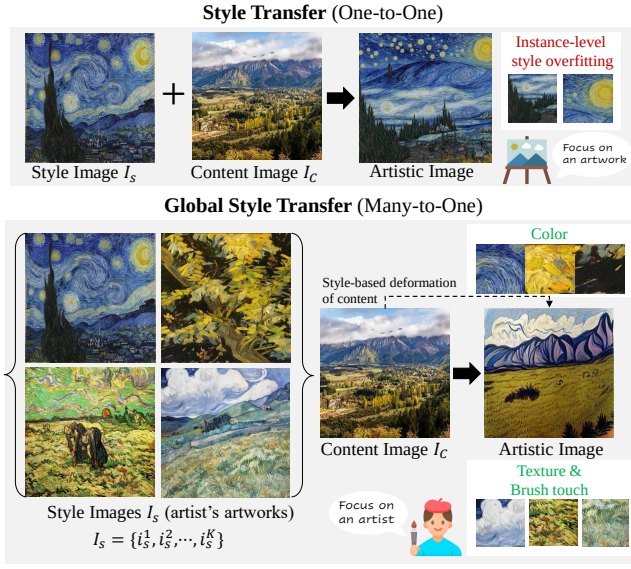

Fig. 2: Comparison between conventional style transfer and the proposed *Global Style Transfer*. (Top) Traditional style transfer applies the style of a single reference artwork to a content image, leading to instance-level overfitting. (Bottom) In contrast, *Global Style Transfer* leverages various 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.

"A painting") for all reference artworks, our framework removes the variance and linguistic bias typically induced by text conditioning.

First, *Global Style Guidance* (GSG) learns the artist's overall stylistic characteristics from hundreds of artworks and injects them into the diffusion model's intermediate latent space (the U-Net bottleneck, or h-space). Because h-space inherently captures stable semantic representations, GSG enables semantically consistent style modulation throughout the diffusion process without degrading the image quality.

Second, *Content Alignment Guidance* (CAG) preserves the structural and semantic integrity of the content image. Addressing the limitations of strict structural freezing, CAG uses a pre-trained CLIP image encoder to measure perceptual similarity, allowing for controlled, style-driven geometric deformations. This mechanism ensures the generated images remain faithful to the original content's identity while naturally reflecting the artist's structural reinterpretations. We validate our approach on the large-scale WikiArt dataset [11]. Compared to vanilla diffusion models and text-based personalization methods [**?**, **?**], our framework demonstrates superior stylistic fidelity, diversity, and content preservation. Our main contributions are summarized as follows:

- We propose the first ***Global Style Transfer*** framework for artist-level diffusion-based synthesis.
- We introduce ***Global Style Guidance***, which captures global stylistic semantics in the intermediate feature space (h-space) with various artworks.
- We develop ***Content Alignment Guidance*** to preserve the structure of the content image while enabling flexible, style-driven deformation.
- We conduct extensive experiments demonstrating that our method achieves bias-robust and semantically consistent artistic image generation.

2 Related Works

2.1 Traditional and Neural Style Transfer

Classical and CNN-based neural style transfer (NST) methods established the foundation of stylization, with subsequent feed-forward and Transformer-based approaches significantly improving semantic consistency and generalization [**?**, **?**, **?**, **?**, **?**]. However, these approaches predominantly rely on a *single-exemplar* definition of style, often causing texture-driven structural drift and failing to capture an artist's *global* stylistic behavior [**?**, **?**, **?**, **?**]. In contrast, our *Content Alignment Guidance* stabilizes structure while explicitly targeting aggregated, corpus-level stylistic tendencies.

2.2 Style Transfer with Text-conditioned Models

Diffusion models [**?**, **?**, **?**, **?**] enable stylistic prompting, but text-conditioned generation suffers from linguistic and dataset bias [**?**], often collapsing toward over-represented iconic works rather than reflecting full stylistic diversity. Personalization techniques DreamBooth [7], Textual Inversion [**?**], Custom Diffusion [**?**], and Perfusion [**?**] improve concept binding but remain prompt-dependent and prone to style–content coupling. Inversion methods [**?**, **?**, **?**] support editing but do not capture large-scale style distributions. Reference-guided diffusion [**?**, **?**, **?**, **?**] and multi-style diffusion [**?**, **?**, **?**, **?**] aggregate multiple exemplars but operate on limited reference sets and do not learn an artist's *global stylistic manifold*. ControlNet and adapters [**?**, **?**, **?**] preserve geometry but do not address prompt bias. We instead remove textual conditioning entirely and derive style solely from large visual corpora, enabling consistent, bias-free modeling of an artist's full stylistic space.

3 Methods

3.1 Preliminary

Latent Diffusion Models. Latent Diffusion Models (LDM) [6] 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

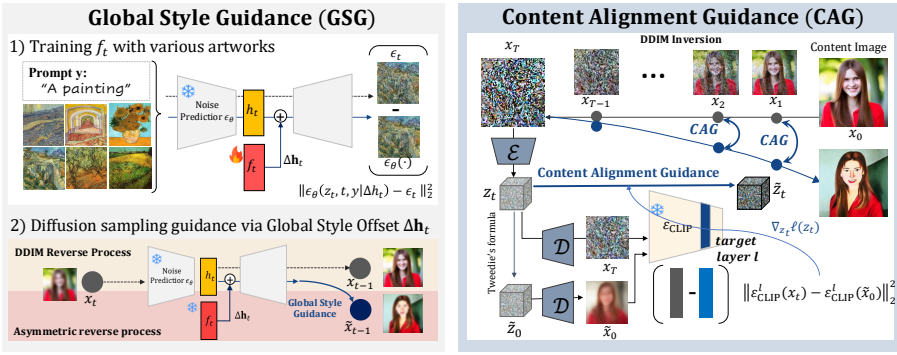

Fig. 3: Overall framework of *Global Style Transfer* (GST). Starting from a content image I_c, we apply DDIM inversion and two guidance terms, *Global Style Guidance* (GSG) and *Content Alignment Guidance* (CAG), during reverse diffusion to generate artist-specific outputs. First, we train the *Style Extraction Function* f_t with hundreds of artworks using the text-independent prompt "`A painting`". Then our *Global Style Offset* $\Delta\mathbf{h}_t$ guides the diffusion model in the $h-$space. CAG uses CLIP-based perceptual gradients to align the semantic structure of the content image while allowing style-based deformation, enabling coherent and artist-faithful stylization.

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:

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

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 [4] discovers that the frozen pre-trained diffusion model inherently contains a semantic

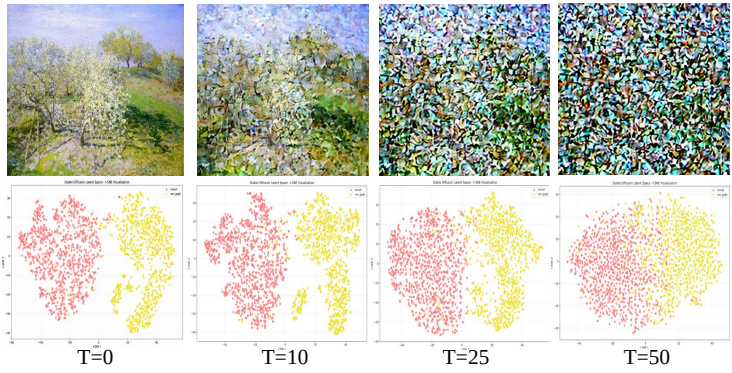

Fig. 4: (Top) Artistic images generated at different diffusion time steps $t = \{0, 10, 25, 50\}$. (Bottom) t-SNE visualization of 500 feature vectors sampled from the h-space for artworks by Van Gogh (yellow) and Monet (red) across the same time steps. Despite increasing noise levels, the clusters remain clearly separable, indicating that the proposed *Global Style Guidance* preserves each artist's semantic manifold without blending or degradation.

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 $I_s = \{i_s^1, i_s^2, .., i_s^K\}$ drawn from an artist's K artworks, *Global Style Guidance* (GSG) guides the diffusion model that encapsulates artist-specific semantic and structural cues across various artworks. To preserve the semantic integrity of the content, we further introduce *Content Alignment Guidance* (CAG), which maintains the abstract

Algorithm 1 Global Style Guidance Algorithm

Require: Style images(=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 $\mathbf{y}$
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**
      $z_0^k = \mathcal{D}(I_s^k)$ {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 $f_t(h_t;\theta)$

structure and semantics of I_c while allowing flexible, style-driven transformations. Through CAG, the content image is re-rendered with the distinctive visual characteristics of the target artwork without compromising its original form or 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, achieving prompt-agnostic style consistency. (2) **Bias mitigation in diffusion-based artistic image synthesis**: Conventional 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 reconstructions 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 diffusion step:

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

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 $\mathbf{y}$, 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 [2] and $\tilde{x}_0 = \mathcal{D}(\tilde{z}_0)$
  $\ell(z_t) = \|\mathcal{E}^l_{\text{CLIP}}(\tilde{x}_0) - \mathcal{E}^l_{\text{CLIP}}(x_t)\|_2$
  Content Alignment Guidance
  $\tilde{z}_t \leftarrow z_t - s\nabla_{z_t}\ell(z_t)$
  //Classifier Free Guidance
  $\tilde{\epsilon}_t \leftarrow \epsilon_\theta(\tilde{z}_t, t, \varnothing) + g(\epsilon_\theta(\tilde{z}_t, t, \tau_\phi(\mathbf{y})) - \epsilon_\theta(\tilde{z}_t, t, \varnothing))$
end for
Return $x_0 = \mathcal{D}(\tilde{\mathbf{z}}_0)$

where w controls the strength of stylistic modulation. The residual $\Delta\mathbf{h}_t$ 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 [13] to avoid stochastic bias from random initialization. Zero-initialization ensures training 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):

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

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 "`A painting`." By enforcing a unified prompt across all artworks, the variance of text conditioning effectively approaches zero, allowing the 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.

Table 1: Comparison of implicit function learning between Asyrp and our Global Style Transfer 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	Visual statistics across artworks
Text conditioning on training f_t	Explicit source–target prompts (high variance), e.g., `'A girl'` → `'A smiling girl'`	Fixed prompt (`'A painting'`, zero variance)
Training data	Single (or few) source images	Hundreds of artworks from a single artist
Loss function for learning f_t	Attribute-driven LPIPS editing loss: $\xi_t = \text{LPIPS}(x, P_T) - \text{LPIPS}(x, P_1)$	Noise reconstruction loss over the artist distribution: $\mathcal{L}_{\text{SEF}} = \mathbb{E}_{z_t \sim \mathcal{E}(x), \epsilon_t, t}\left[\|\epsilon_\theta(z_t, t, \tau_\theta(y) \mid \Delta h_t) - \epsilon_t\|_2^2\right]$

Time-Step Robustness of Global Style Guidance. We perform *Global Style Guidance* (GSG) within the intermediate feature space (h-space) to achieve semantically stable transformations throughout the diffusion process. As shown in Fig. 4, we visualize 500 artworks per artist perturbed with Gaussian noise across diffusion timesteps. The intermediate representations h_t form well-separated, consistent clusters in h-space, demonstrating strong robustness to both noise and timestep variation. This stability indicates that h-space preserves high-level semantic attributes of an artist's global style, independent of the diffusion step. Consequently, performing GSG in this space ensures coherent style modulation while maintaining structural and semantic integrity across the entire generative trajectory.

Difference between Image Editing with Asyrp and our Global Style Transfer. Asyrp focuses on "text-driven, instance-level image editing", where h-space directions are learned to modify attributes of a single source image under prompt supervision. The implicit functions f_t are trained to satisfy a specific source-target prompt change (e.g., "`a girl`" → "`a smiling girl`") while preserving the source structure via DDIM inversion and early-stage editing. In contrast, we instead learn a 'distribution-level global style from hundreds of artworks by fixing the text prompt to a constant ("`A painting`"), which effectively removes text-conditioning variance and forces the implicit function f_t to rely purely on shared visual statistics across the artist's dataset. Consequently, our $\Delta h_t = f_t(h_t)$ captures artist-consistent global style semantics rather than a text-aligned attribute axis, enabling *prompt-agnostic, many-to-one global style guidance*. (Tab. 1)

3.4 Content Alignment Guidance

Motivated by how artists reinterpret real-world objects, preserving abstract structure while expressing unique artistic style, we propose the *Content Alignment Guidance* (CAG). In Neural Style Transfer (NST) [8, 12, 14], 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 [1, 9], a process used in diffusion models to deterministically map an input content image into its corresponding noisy latent. We then extend diffusion guidance by drawing inspiration from Neural

Style Transfer, where minimizing feature-map distances between content and synthesized images induces a style-based deformation of the content. Building upon Stable Diffusion [6], we employ the CLIP image encoder $\mathcal{E}_{\text{CLIP}}(\cdot)$ to measure perceptual similarity between the approximated image $\tilde{x}_0 = \mathcal{D}(\tilde{z}_0)$ from [2] and the sampled image $\tilde{x}_t$:

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

where $\mathcal{E}^l_{\text{CLIP}}$ extracts features from a designated target layer l of CLIP to capture high-level semantics of content. In this process, we project the latent variable into pixel space at each time step by $\hat{x}_t = \mathcal{D}(z_t)$ to compute the gradient. At every time step, we guide the feature map of the generated image $\tilde{x}_t$ to follow that of the approximated $\tilde{x}_0$, since intermediate reconstructions preserve the abstract, recognizable structure of the content image but fail to capture fine-grained details. Enforcing this alignment effectively enhances style-based deformation of content, maintaining structural coherence while adapting the global style. To guide the image latent z_t, from the SDE formulation [10], we replace the unconditional score $\nabla_{z_t} \log p(z_t)$ with its conditional form $\nabla_{z_t} \log p(z_t|\ell)$, 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 is a scaling factor controlling the gradient influence. At each time step, CAG thus refines the latent trajectory to align the generated image with the content feature distribution without external supervision or training. Consequently, CAG enables the generated image to remain faithful to the semantic identity of the content while capturing the distinctive compositional and geometric characteristics found in the artist's works. The overall algorithms of *Global Style Transfer* with CAG are described in Algorithm 2, and an overview of our framework is described in Fig. 3.

4 Experiments

Datasets. We define the style reference images in the WikiArt dataset [11], 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 [15] , 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 artistic image synthesis and style transfer tasks.

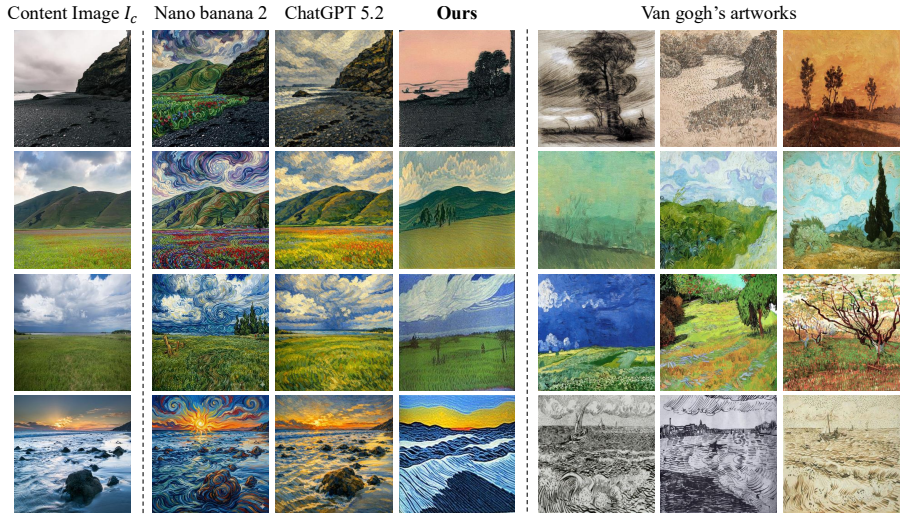

Fig. 5: Compare our Global Style Transfer with baselines (Nano Banana 2, ChatGPT 5.2) prompted with '`Van Gogh style`' for a given content image I_c. The rightmost columns display the top three real Van Gogh artworks with the lowest CLIP distance to our generated result.

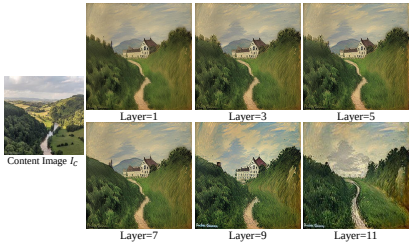

Fig. 6: Effect of different CLIP encoder layers in Content Alignment Guidance on content preservation.

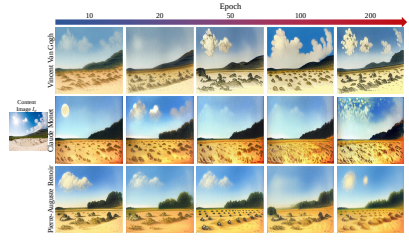

Fig. 7: Effect of training epochs on *Global Style Guidance*: higher epochs yield clearer stylistic separation and stronger artist-specific characteristics.

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.

Analysis of Stylistic Bias We compared our framework with large-scale text-to-image baselines (Nano Banana 2 and ChatGPT 5.2) [3, 5] conditioned on the

prompt "Van Gogh style" As shown in Fig. 5, the baselines exhibit severe stylistic bias, collapsing into formulaic repetitions of overrepresented iconic features (e.g., swirls, thick brushstrokes) and resulting in severe over-stylization In contrast, our Global Style Transfer learns a unified global representation from a comprehensive corpus through the MLP, effectively buffering against this bias and yielding diverse synthesis that reflects the true global aesthetic without collapsing into a formulaic representation

Quantitative Results. We report quantitative results in Tab. 2, where a lower ArtFID indicates higher stylistic fidelity. Our model achieves superior performance for most painters, including Chagall, Saryan, Dali, and Van Gogh, demonstrating its ability to generalize across diverse stylistic domains without overfitting to specific exemplars. Each artist was evaluated using between 500 and 1400 real artworks, ensuring sufficient data diversity to reflect the full range of individual stylistic variations. Notably, painters such as Van Gogh and Dali, whose works exhibit high stylistic variance and contain several iconic yet visually distinct masterpieces, benefit from our bias-free formulation, yielding significantly lower ArtFID scores. In contrast, for artists like Renoir and Kustodiev, whose paintings share highly consistent brushwork and color composition, the vanilla Stable Diffusion model sometimes achieves comparable or slightly better results. This is because even when the baseline model overfits to certain reference images, the resulting synthesis still aligns closely with the artist's overall aesthetic, given the homogeneity of their corpus. Nevertheless, our method consistently maintains stylistic robustness across heterogeneous artistic domains.

Table 2: Comparison of the ArtFID between ours and vanilla Stable Diffusion. **Bold** entries are the best results.

Model	ArtFID (↓)						
	Chagall	Saryan	Renoir	Dali	Kustodiev	Vangogh	**Avg.**
Ours	**34.13**	**24.99**	33.25	**25.73**	28.46	**29.75**	**29.38**
SD	37.64	29.03	**31.84**	35.29	**28.30**	30.33	32.07

4.2 Analysis of Global Style Transfer

Effect of Training Epochs on Global Style Guidance. We investigated the impact of training epochs on the MLP-based *Global Style Guidance* (GSG) module to assess how the duration of training influences the strength and stability of the global style representation. The GSG was trained with five epoch configurations(10, 20, 50, 100, 200) using artworks from three representative artists, namely Van Gogh, Monet, and Renoir. As shown in Fig. 7, models trained with fewer epochs such as 10 and 20 failed to reproduce the structural composition, brushwork and stylistic characteristics of the artists, resulting in visually similar results among painters. However, as training epochs increased, the model gradually captured more distinctive stylistic features and achieved greater cross-artist separation. At 200 epochs, the generated images exhibited distinct color palettes and compositional traits characteristic of each painter, confirming that

Table 3: Quantitative comparison of different style transfer methods.

Artist	Method	FID (↓)	ArtFID (↓)	CFSD (↓)
Van Gogh	S.D	11.97	23.78	0.7428
	Textual Inversion	7.67	13.68	0.1674
	LoRA based Fine-tuning	11.01	21.38	0.1886
	Ours	9.46	19.25	0.2896
Chagall	S.D	16.26	32.44	0.3117
	Textual Inversion	11.72	21.15	0.1683
	LoRA based Fine-tuning	17.30	32.59	0.1674
	Ours	13.25	26.39	0.2626
Renoir	S.D	16.72	33.70	0.1584
	Textual Inversion	12.14	21.99	0.1100
	LoRA based Fine-tuning	15.36	29.23	0.1749
	Ours	12.27	24.70	0.1566

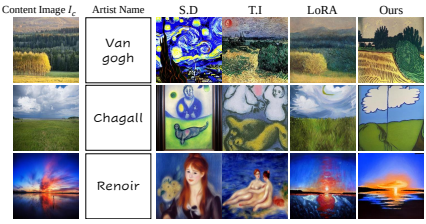

Fig. 8: Visual comparison of previous style-transfer techniques and vanilla Stable Diffusion.

prolonged training allows GSG to converge toward a stable and semantically meaningful global style embedding.

Effect of layer index 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 style transfer. As shown in Fig. 6, applying CAG at lower or mid-level layers (Layers 1, 3, 5, 7, and 9) consistently produces unintended artifacts, such as house-like shapes appearing in the center that are absent from the original content image. This suggests that low-level features primarily capture edges, textures, and local color gradients, emphasizing fine details while failing to maintain global composition and semantic coherence. In contrast, applying CAG at Layer 11 achieves stable composition and semantically faithful results, maintaining the original scene structure without introducing artificial elements. Since higher layers in CLIP encode semantic correspondence between vision and language, guiding at this level provides more effective control for content alignment, preventing low-level overfitting and ensuring stable semantic preservation during style transfer.

Comparison between Style Transfer Methods. Our method introduces a new paradigm for global style transfer, which differs fundamentally from traditional one-to-one style transfer that maps a single style image to a single content image. As a result, direct comparison is inherently limited. To enable a fair evaluation, we adapt representative style-transfer baselines to operate in a global-style setting using multiple style exemplars.

We first consider Textual Inversion (TI) [**?**], which encodes style by optimizing a token embedding from a small set of images. We extend this method by optimizing the embedding using the full set of an artist's artworks, allowing it to capture broader stylistic characteristics. This experiment evaluates whether TI can scale to represent an artist's global style. Second, inspired by DreamBooth [7], we evaluate a LoRA-based fine-tuning strategy that adapts the diffusion model using multiple artworks from the same artist. While DreamBooth typically binds a unique identifier to a subject using a few images, we fine-tune

the model on a larger artwork collection to assess its ability to learn global stylistic patterns. We compare both adapted baselines with our proposed framework.

Tab. 3 reports the style- and content-fidelity metrics, and Fig. 8 shows qualitative results. Our approach achieves consistently strong performance across FID, ArtFID, and CFSD, demonstrating its effectiveness in capturing global artistic styles. Although a lower CFSD typically indicates higher content fidelity, our framework intentionally allows style-driven deformation of the content to reflect the expressive characteristics of the original artworks. Therefore, excessive structural preservation may not correspond to better style-transfer performance.

TI encodes stylistic information into a single fixed-dimensional token embedding, which limits its ability to represent the large stylistic diversity across artists with hundreds or thousands of artworks (e.g., Van Gogh: 1,889; Renoir: 1,400; Chagall: 765). While TI achieves lower FID and ArtFID, this mainly results from stronger content preservation rather than meaningful stylistic transformation. As the number of artworks increases, this capacity limitation becomes more evident and TI performance plateaus. In contrast, our method continues to improve with more artworks, indicating better scalability for capturing global artistic styles.

Ablation Study - The effect of the CAG. To verify the effectiveness of CAG, we conduct an ablation study comparing our full model with a variant where CAG is removed. As shown in Fig. 9 and Tab. 4, disabling CAG leads to a noticeable loss of content information, which is reflected by a notable decrease in ArtFID, a metric that accounts for content similarity. This shows that CAG is essential for preserving content under strong global style modulation. Additional ablations on the CAG scale and embedding layer choice are reported in Figs. 6 and 7.

Table 4: Comparison of the metric values.

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

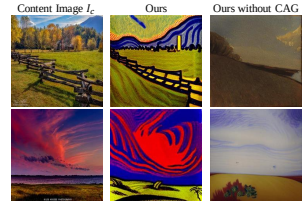

Fig. 9: The effect of CAG.

5 Conclusion

We propose *Global Style Transfer*, which learns to represent the global artistic style of an artist from multiple artworks via *Global Style Guidance* and *Content Alignment Guidance*. Our approach enables unbiased artistic image synthesis, effectively capturing global stylistic semantics while preserving the structural integrity of the content.

References

1. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
2. Efron, B.: Tweedie's formula and selection bias. Journal of the American Statistical Association **106**(496), 1602–1614 (2011)
3. Google: Gemini image generation: Create images with gemini. `https://gemini.google/overview/image-generation/` (2026), accessed: 2026-01-22
4. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. arXiv preprint arXiv:2210.10960 (2022)
5. OpenAI: Chatgpt: Get answers. find inspiration. be productive. `https://openai.com/ko-KR/index/chatgpt/` (2026), accessed: 2026-01-22
6. 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)
7. 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)
8. 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)
9. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
10. 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)
11. 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)
12. Wang, Z., Zhao, L., Chen, H., Qiu, L., Mo, Q., Lin, S., Xing, W., Lu, D.: Diversified arbitrary style transfer via deep feature perturbation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7789–7798 (2020)
13. 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)
14. 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)
15. 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)