STYLIZATION AS GEODESIC NAVIGATION ON DIFFUSION MANIFOLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Facial stylization asks for a delicate transformation: the image should look unmistakably artistic, yet the person should remain unmistakably the same. Existing training-free and instruction-driven methods often fail to expose such control, either drifting away from identity or producing only weak stylization. We introduce **Stylus**, a training-free framework that treats facial style transfer as navigation on the curved latent manifold of a pretrained diffusion model. Rather than interpolating in flat Euclidean space, Stylus refines the interpolated content–style point by lowering a ratio-weighted discrete geodesic energy under the pullback metric, made tractable by a rank-one power-iteration approximation of the Jacobian. This principle drives two complementary controls: content–style latent interpolation for smooth semantic fusion, and self-attention query optimization for spatially adaptive style injection. The resulting dual-control mechanism provides continuous, fine-grained control over stylization strength while preserving identity-critical facial structure, all without additional training. Experiments show that Stylus achieves the strongest identity and content preservation at competitive stylization strength.

1 INTRODUCTION

Image stylization, which aims to transform images into artistic styles while preserving their original content characteristics Gatys et al. (2016); Kwon & Ye (2022a); Deng et al. (2022); Zhang et al. (2023b); Hong et al. (2023); Zhang et al. (2022); An et al. (2021); Yang et al. (2023), has attracted significant research attention in recent years. While traditional style transfer methods have shown success with general images, facial image stylization presents unique challenges that require special consideration. While artistic style transfer is a well-established field, the specific challenge of stylizing facial images while maintaining identity has received surprisingly limited research attention in recent years Khowaja et al. (2024); Yi et al. (2020); Shi et al. (2019); Chen et al. (2020). The key challenge of facial image stylization lies in striking a delicate balance - the stylization must maintain recognizable facial features that preserve the subject's identity while simultaneously achieving compelling artistic effects (Cohen et al., 2025). This requires more precise control over the stylization process compared to general image stylization, making it a particularly demanding problem in the field.

Recent advances in diffusion models have demonstrated remarkable capabilities in various image generation and manipulation tasks Ho et al. (2020); Nichol & Dhariwal (2021); Kim et al. (2022); Wang et al. (2023a); Kong et al. (2023), with promising applications in style transfer Jeong et al. (2024); Wang et al. (2023b); Chung et al. (2024); Kwon & Ye (2022b). More recently, large-scale DiT-based editing models Wu et al. (2025); Labs et al. (2025) have pushed the frontier of instruction-driven image manipulation and can be repurposed for stylization via natural-language prompts. Despite their impressive generative quality, several critical limitations persist. First, adapting conventional diffusion models to new artistic styles typically demands extensive computational resources and training time Zhang et al. (2023b). Second, existing diffusion-based stylization methods often lack precise control mechanisms over the stylization process, making it challenging to achieve desired artistic effects Chung et al. (2024). Third, they frequently struggle to maintain an optimal equilibrium between faithful style transfer and the preservation of essential identity features, which is particularly crucial for facial stylization. Finally, DiT-based editing models inherit a distinct set of weaknesses Wu et al. (2025); Labs et al. (2025): they rely almost exclusively on textual prompts to specify the target appearance, which is inherently ambiguous for describing visual style; they provide no fine-grained knob to modulate stylization strength; and they tend to either over-edit, erasing the subject's identity, or under-edit, failing to faithfully transfer the reference style.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We trace these failures to a shared oversight: prior methods implicitly treat the diffusion latent space as *flat Euclidean space*, where linear interpolation and uniform attention substitution are assumed to produce semantically valid results. In reality, diffusion models learn a *curved manifold* on which valid images reside Kwon & Ye (2022b); Park et al. (2023), and straight-line traversal of this manifold routinely crosses low-probability regions, yielding ghosting artifacts, identity drift, or weak stylization. Our key insight is to instead use the pullback geometry of this manifold to refine ratio-selected intermediate points toward locally low-energy geodesic points Shao et al. (2018), exposing an interpolation parameter that continuously modulates stylization strength.

Guided by this principle, we propose **Stylus**, a training-free framework that performs geodesic-guided point refinement at two complementary stages. First, *linear style interpolation* refines a ratio-selected content–style latent by minimizing a ratio-weighted discrete energy. Second, *dynamic style injection control* applies the same refinement to a ratio-selected point between content and style queries in the attention space, yielding spatially adaptive injection while recovering fine-grained details that latent-space interpolation alone cannot capture.

The main contributions of our work can be summarized as follows:

- A novel training-free method for facial stylization that can be generalized across various stable diffusion architectures, achieving state-of-the-art performance while providing fine-grained control over the stylization process.
- A dual-control framework combining geodesic-based style interpolation and adaptive style injection, enabling precise content-style fusion through optimal trajectories in latent space while dynamically optimizing the balance between stylization and identity preservation.
- Extensive empirical evaluation and rigorous experimental analysis demonstrating significant performance improvements across a comprehensive suite of quantitative metrics and systematic qualitative assessments.

2 Related Work

2.1 Image Style Transfer

Image style transfer has been an active research area in computer vision. Early approaches like Gatys et al. Gatys et al. (2016) pioneered neural style transfer by using CNN features to separate and recombine content and style. Since then, numerous methods have been proposed to improve style transfer quality and efficiency Johnson et al. (2016); Huang & Belongie (2017). Recent transformer-based methods like StyTr2 Deng et al. (2022) leverage self-attention mechanisms for better feature transformation. CLIPStyler Kwon & Ye (2022a) enables text-guided style transfer by utilizing CLIP's multimodal representations. With the advent of diffusion models Nichol & Dhariwal (2021), several works have explored their potential for style transfer. Wang et al. Wang et al. (2023b) proposed StyleDiffusion for controllable disentangled style transfer. Zhang et al. Zhang et al. (2023b) developed an inversion-based approach for more precise style control. Domain-aware methods like Zhang et al. (2022) incorporate contrastive learning to better capture domain-specific style features. Recent works have also focused on improving artistic quality while maintaining content fidelity Liu et al. (2025); Zhang et al. (2025); Wang et al. (2025); Chittersu et al. (2025). AesPA-Net Hong et al. (2023) introduces aesthetic pattern awareness for more visually pleasing results. ArtFlow An et al. (2021) employs reversible neural flows to achieve unbiased style transfer. Yang et al. Yang et al. (2023) proposed a zero-shot contrastive loss for text-guided diffusion style transfer, demonstrating the potential of diffusion models in this domain. More recently, DiT-based image editing models, such as Qwen-Image-Edit Wu et al. (2025) and FLUX.2-dev Labs et al. (2025), have shown strong editing capabilities, and can be adapted to stylization through textual prompts. However, their reliance on language prompts makes visual style specification ambiguous, and they provide limited fine-grained control over stylization strength and identity preservation.

2.2 Facial Image Stylization

Recognizing the unique challenges of facial stylization, researchers have developed specialized approaches for face images Khowaja et al. (2024). Early attempts adapted general style transfer methods with face-specific constraints Selim et al. (2016). More recent works have incorporated facial landmark detection Zhang et al. (2017) and identity preservation mechanisms Chung et al.

(2024). StyleGAN-based approaches Karras et al. (2019; 2020b) and unsupervised image-to-image translation methods have shown promise in facial stylization tasks like selfie-to-anime conversion. However, these methods typically require task-specific training to achieve satisfactory results. While many existing approaches excel at single-style transfer, they often struggle with multi-style fusion and customized style combinations. Methods like DreamBooth Ruiz et al. (2022) and Textual Inversion Gal et al. (2022) face similar challenges when attempting to combine multiple artistic styles like cartoon, anime, and arcane aesthetics. A key limitation across these approaches is their difficulty in preserving facial identity features during style transfer. The misalignment between source and target domain facial features frequently results in artifacts that compromise the quality of the stylized output.

Our work advances the field with a training-free dual-control framework that enables fine-grained control over both initial style fusion and ongoing style injection in diffusion models, achieving precise stylization while preserving facial identity.

3 METHODOLOGY

3.1 THEORETICAL FOUNDATION

Let the frozen U-Net be decomposed as $\epsilon_\theta(x, t, c) = g_u(g_d(x, t, c), t, c)$. Throughout this section, $g(x) := g_d(x, t, c) \in \mathbb{R}^p$ denotes the vectorized down-path/bottleneck hidden state at fixed timestep t and conditioning c, while $x \in \mathcal{X} = \mathbb{R}^n$ is the vectorized diffusion latent. In our setting $p > n$; g is not the final noise prediction. Assuming that J_g has full column rank in the local region visited by refinement, g is an immersion and $G = J_g^\top J_g$ defines a pullback Riemannian metric on that region. This gives a local interpretation of distances after the hidden-state map; it is not a global claim about the complete diffusion trajectory.

Definition 3.1. Let $\gamma : [0, 1] \to \mathcal{X}$ be an absolutely continuous (AC) curve. The arc length of γ with respect to the pullback Riemannian metric g is

$$L_{(\mathcal{X},\mathrm{g})}(\gamma) \;=\; \int_0^1 \big\|\dot\gamma(t)\big\|_{g,\gamma(t)}\, dt \;=\; \int_0^1 \sqrt{\dot\gamma(t)^\top G\big(\gamma(t)\big)\,\dot\gamma(t)}\; dt.$$

Let $\eta(t) = g(\gamma(t)) \in \mathcal{M}$. The arc length of η with respect to the metric induced by the ambient Euclidean inner product is

$$L_{(\mathcal{M},\|\cdot\|_2)}(\eta) \;=\; \int_0^1 \big\|\dot\eta(t)\big\|_2\, dt \;=\; \int_0^1 \Big\| \tfrac{d}{dt}\big(g\circ\gamma\big)(t) \Big\|_2\, dt.$$

Geodesic distances. For $x_1, x_2 \in \mathcal{X}$, the geodesic distance induced by g is

$$d_{\mathrm{g}}(x_1, x_2) \;:=\; \inf\Big\{\, L_{(\mathcal{X},\mathrm{g})}(\gamma) \;:\; \gamma \in AC([0,1],\mathcal{X}), \gamma(0) = x_1,\; \gamma(1) = x_2 \Big\}.$$

For $y_1, y_2 \in \mathcal{M}$, the geodesic distance induced by the Euclidean metric restricted to $\mathcal{M}$ is

$$d_{\mathcal{M}}(y_1, y_2) \;:=\; \inf\Big\{\, L_{(\mathcal{M},\|\cdot\|_2)}(\eta) \;:\; \eta \in AC([0,1],\mathcal{M}), \eta(0) = y_1,\; \eta(1) = y_2 \Big\}.$$

The following proposition exhibits the local isometry between $(\mathcal{X}, \mathrm{g})$ and $(\mathcal{M}, \|\cdot\|_2)$. The proof can be found in Appendix B.

Proposition 3.2 (Local isometry of $(\mathcal{X}, \mathrm{g})$ into $(\mathcal{M}, \|\cdot\|_2)$)**.** *Let $g : \mathcal{X} \to \mathbb{R}^p$ be continuously differentiable and assume $J_g(x)$ has full column rank at $x \in \mathcal{X}$. Consequently, there exists a neighborhood $U \ni x$ such that $g|_U : U \to \mathcal{M}_U$ is a Riemannian isometry onto its image (lengths of curves are preserved), i.e.,*

$$L_{(\mathcal{X},\mathrm{g})}(\gamma) = L_{(\mathcal{M}_U,\|\cdot\|_2)}(g\circ\gamma) \quad \textit{for all absolutely continuous } \gamma : [0,1] \to U.$$

Proof. By definition of the pullback metric, $\mathrm{g}_x(u, v) = \langle J_g(x)u,\; J_g(x)v\rangle_2$, so $J_g(x)$ preserves inner products on tangent spaces. Since full column rank is an open condition, J_g has constant rank n on a sufficiently small neighborhood U of x. By the constant rank theorem, after possibly shrinking

U, the restriction $g|_U$ is a C^1 embedding and hence a diffeomorphism onto its image $\mathcal{M}_U$. For any absolutely continuous $\gamma : [0,1] \to U$,

$$\big\|\dot\gamma(t)\big\|_{g,\gamma(t)} = \big\|J_g(\gamma(t))\dot\gamma(t)\|_2 = \big\|\tfrac{d}{dt}\big(g\circ\gamma\big)(t)\big\|_2,$$

hence lengths are preserved after integration. This is precisely the definition of a local Riemannian isometry onto $\mathcal{M}_U$ endowed with the Euclidean induced metric. □

Within U, optimizing arc length or taking Riemannian gradients in $\mathcal{X}$ under the pullback metric is thus identical to doing so on $\mathcal{M}$. Practically, the chord length $\|g(x_{k+1}) - g(x_k)\|_2$ serves as a surrogate for the short geodesic arc length between nearby points; this is a local, per-step approximation, valid for small steps and moderate curvature. The following proposition gives the corresponding distance equality on U. Since it requires both points to lie in one such neighborhood, we use it only locally, around the current iterate x_m, not as a global content-to-style guarantee.

Proposition 3.3 (Local distance equality between (U, g) and $\mathcal{M}_U$)**.** *Endow U with the pullback Riemannian metric $\mathrm{g}_x(u,v) := \langle J_g(x)u, J_g(x)v\rangle$ and $\mathcal{M}_U$ with the Euclidean induced metric. Then for all $x_1, x_2 \in U$,*

$$d_{\mathrm{g},U}(x_1,x_2) \;=\; d_{\mathcal{M}_U}\big(g(x_1), g(x_2)\big),$$

where both distances are defined using curves restricted to U and $\mathcal{M}_U$, respectively.

Proof. For any absolutely continuous curve $\gamma : [0,1] \to U$ we have, by definition of the pullback metric,

$$\big\|\dot\gamma(t)\big\|_{g,\gamma(t)} = \sqrt{\dot\gamma(t)^\top J_g(\gamma(t))^\top J_g(\gamma(t))\,\dot\gamma(t)} = \big\|\tfrac{d}{dt}\big(g\circ\gamma\big)(t)\big\|_2.$$

Since length is preserved pointwise and hence after integration, we have $L_{(U,\mathrm{g})}(\gamma) = L_{(\mathcal{M}_U,\|\cdot\|_2)}(g\circ\gamma)$. Taking the infimum over all such γ in U with $\gamma(0) = x_1, \gamma(1) = x_2$ yields

$$d_{\mathrm{g},U}(x_1,x_2) \;\geq\; d_{\mathcal{M}_U}(g(x_1), g(x_2)),$$

because any competitor in U maps to a competitor in $\mathcal{M}_U$ with the same length.

Conversely, since $g|_U$ is a diffeomorphism, every absolutely continuous $\eta : [0,1] \to \mathcal{M}_U$ with $\eta(0) = g(x_1)$, $\eta(1) = g(x_2)$ lifts uniquely to $\gamma = (g|_U)^{-1} \circ \eta$ in U, and the same length identity shows $L_{(\mathcal{X},\mathrm{g})}(\gamma) = L_{(\mathcal{M}_U,\|\cdot\|_2)}(\eta)$. Taking infimum over such η gives the reverse inequality and hence equality of distances. □

The above results explain why our interpolation objective is defined after the latent points are passed through g. A small displacement in the raw latent space may have very different effects after the generator mapping, and the pullback metric measures such effects by the change in $g(x)$.

3.2 Geodesic-Guided Style Interpolation

In the context of style transfer, we aim to find an optimal path between content and style representations. Let us consider two points $x_c, x_s \in \mathcal{X}$ representing the content and style images in the coordinate space $\mathcal{X}$. These points can be connected by a continuous curve $\gamma : [0,1] \to \mathcal{X}$ with $\gamma(0) = x_c$ and $\gamma(1) = x_s$. Following Section B, when this curve is mapped through the hidden-state map g, it creates a trajectory $\gamma_g : [0,1] \to \mathcal{M}$ on the Riemannian manifold $\mathcal{M}$, where $\gamma_g = g \circ \gamma$. Each point along this manifold curve represents a unique blend of content and style characteristics, enabling controlled style transfer.

We start with a simple linear interpolation $x_m = \delta x_c + (1-\delta)x_s$, where $\delta \in [0,1]$ controls the mixing ratio between content and style (larger δ keeps the initial point closer to the content). While this linear interpolation provides an initial point in the coordinate space $\mathcal{X}$, our goal is to refine it into a more geometrically meaningful point that better preserves both content and style characteristics when mapped onto the manifold $\mathcal{M}$ through the hidden-state map g.

The curve γ follows a direction from the content point x_c to the style point x_s, and this directional path is preserved when mapped to the corresponding curve on the Riemannian manifold $\mathcal{M}$ through the hidden-state map g. By performing forward finite difference, we can obtain the approximate

velocity of the curve γ at point x_m, which under our initialization lies at parameter $1-\delta$ along the path:

$$v_m = (g(x_s) - g(x_m))/\delta. \tag{1}$$

To quantify the optimality of this point, we use the discrete two-chord version of the energy function in Equation 12 (the derivation is given in Appendix B):

$$E \approx \alpha \left(\delta||g(x_m) - g(x_c)||^2 + (1-\delta)||g(x_s) - g(x_m)||^2\right), \tag{2}$$

where α is a constant scale factor. By Section 3.1, the chord lengths in $\mathcal{M}$ locally match geodesic lengths in $\mathcal{X}$, which motivates Equation 2. The ratio-dependent weights make the unconstrained mapped-space target $\delta g(x_c)+(1-\delta)g(x_s)$, consistent with the initialization; if this target lies outside $g(\mathcal{X})$, the full objective instead seeks a local projection onto $g(\mathcal{X})$.

To refine x_m with fixed endpoints x_c and x_s, we minimize the discrete energy function by gradient descent on x_m. The gradient on the point x_m is:

$$\nabla_{x_m} E = -2\alpha J_g^\top(x_m)\left(\delta g(x_c) + (1-\delta)g(x_s) - g(x_m)\right). \tag{3}$$

The gradient in Equation 3 forms a finite-difference residual in $\mathcal{M}$ and pulls it back to $\mathcal{X}$ through $J_g^\top$, which removes its component normal to the manifold (Figure 1). This update is another direct consequence of the pullback formulation in Section 3.1: the finite-difference residual is first formed in the mapped space $\mathcal{M}$, and the transpose Jacobian $J_g^\top$ pulls this descent signal back to the coordinate space $\mathcal{X}$. Through this projection mechanism, we can perform gradient descent directly on x_m in the coordinate space $\mathcal{X}$ to find the geodesic path while respecting the manifold's geometric structure.

3.3 Efficient Jacobian Approximation

The gradient in Equation (3) shows how the pullback geometry enters the actual update: the descent signal is computed in the mapped space and then pulled back through $J_g^\top$. Thus, the practical computation only requires an efficient approximation of this Jacobian-induced direction. However, computing the Jacobian of a generative model is computationally expensive, particularly for diffusion-based models. To address this challenge, we approximate the singular value decomposition (SVD) of the Jacobian using power iteration Haas et al. (2024) over the matrix $J_g^\top J_g$, whose eigenvectors correspond to the right singular vectors of J_g.

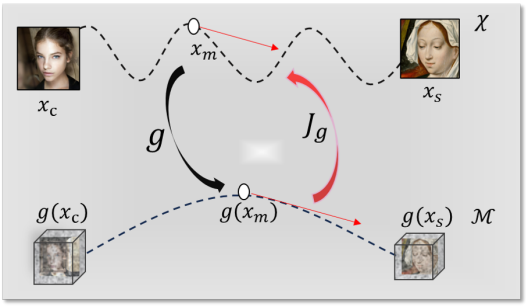


Figure 1: Geometric interpretation of the gradient computation: the hidden-state map g maps points from coordinate space $\mathcal{X}$ to the manifold $\mathcal{M}$, while the Jacobian J_g projects the manifold gradient onto the tangent space of $\mathcal{X}$.

In our implementation, $g(x_m) = g_d(x_m, t, c)$ is the fixed-timestep down-path/bottleneck hidden-state map defined in Section 3.1. Its Jacobian and SVD are:

$$J_g(x_m) = \frac{\partial g_d(x_m, t, c)}{\partial x_m} = U\Sigma V^\top = \sum_i \sigma_i u_i v_i^\top, \tag{4}$$

where v_1 is the leading right singular vector. We begin with a random vector v; the product $J_g^\top J_g v$ can be computed as:

$$J_g^\top J_g v = \frac{\partial}{\partial x_m}\left\langle g_d(x_m, t, c),\, \mathrm{sg}(J_g v)\right\rangle \tag{5}$$

and

$$J_g v = \frac{\partial}{\partial a} g_d(x_m + av, t, c)\Big|_{a=0}, \tag{6}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Iterating Equation 5 with normalization converges to v_1 at rate $O\big((\sigma_2^2/\sigma_1^2)^K\big)$ after K iterations.

We then replace the exact pullback direction $J_g^\top r_m = \sum_i \sigma_i (u_i^\top r_m)\, v_i$, with $r_m := \delta g(x_c) + (1-\delta)g(x_s) - g(x_m)$, by its leading term:

$$\eta_m = -2\alpha\, v\, (J_g v)^\top \big(\delta g(x_c) + (1-\delta)g(x_s) - g(x_m)\big), \tag{7}$$

where $(J_g v)^\top r_m$ is a scalar. This rank-one surrogate drops the subdominant directions $v_{i>1}$; a top-k comparison (Appendix D) shows that adding them changes the results only marginally at a substantially higher cost. We update $x_m \leftarrow x_m - \lambda \eta_m$ with step size λ and stop once the energy change drops below 10^{-4} (within 5–10 iterations in practice).

3.4 Style Injection Control

Interpolating latent codes as in Section 3.2 can degrade content details. To better preserve content, prior work Chung et al. (2024) injects style via self-attention (SA) layers, using query features from the content image and key/value features from the style image. However, since query features largely determine spatial structure, using only the content query weakens the applied style. We propose to find an optimal interpolation between the content query (Q_c) and style query (Q_s). A naive linear interpolation, $Q_m = (1-\beta)Q_c + \beta Q_s$, often produces ghosting artifacts. Instead, we find a geodesic path in the query feature space by adapting the energy minimization from Section 3.2. We initialize Q_m with linear interpolation and then minimize the ratio-weighted energy

$$E_\beta = (1-\beta)\|f(Q_m) - f(Q_c)\|^2 + \beta\|f(Q_s) - f(Q_m)\|^2.$$

Its gradient is:

$$\nabla_{Q_m} E_\beta = -2J_f^\top \big((1-\beta)f(Q_c) + \beta f(Q_s) - f(Q_m)\big), \tag{8}$$

where f is the attention module. The key and value features are held fixed. As in the latent stage, the unconstrained mapped-space target is $(1-\beta)f(Q_c) + \beta f(Q_s)$, and the truncated refinement is a small correction of the linear query interpolant. This query-space refinement follows the same pullback-energy principle as Section 3.1, with the hidden-state map g replaced by the attention module f. In other words, the interpolation between Q_c and Q_s is measured after the queries are transformed by f, rather than by their raw Euclidean distance in query space.

We approximate the Jacobian J_f using power iteration as before. The vector-Jacobian products are computed via:

$$J_f^\top J_f v = \frac{\partial}{\partial Q_m}(f(Q_m), J_f v) \tag{9}$$

and

$$J_f v = \frac{\partial}{\partial a} f(Q_m + av)|_{a=0}, \tag{10}$$

The update uses the same rank-one surrogate as Equation 7. By performing gradient descent on Q_m, we obtain modified query features that preserve content structure while faithfully representing the desired style.

4 Experiment

4.1 Experiment Setting

Implementation: We use the official pretrained Stable Diffusion models (SD1.4, SD1.5, SD2.0, SD2.1), and the sample step is set to 50, the guidance scale is set to 3.0. The parameter δ is set to 0.9, α is set to 5, β is set to 0.25 in our method. All experiments are conducted on a single NVIDIA H800 GPU. The max iteration for geodesic optimization is set to 10, with early stopping at an energy change below 10^{-4}. The implementation details can refer to the Appendix H and Appendix H.3.

Evaluation: We quantitatively evaluate our method along three complementary axes. Style alignment is measured with ArtFID Wright & Ommer (2022); content preservation is assessed using LPIPS Zhang et al. (2018) and Content Loss (CS) Li et al. (2017); and identity preservation is assessed by identity loss (ID), computed as one minus the dot product between normalized ArcFace Deng et al. (2019) embeddings of the content and stylized images. For a comprehensive comparison, we benchmark Stylus against representative baselines drawn from three categories: (i) *DiT-based editing models*, including Qwen-Image-Edit Wu et al. (2025) and FLUX.2-dev Labs et al. (2025); (ii) *diffusion-based style transfer methods*, including StyleSSP Xu et al. (2025), StyleID Chung et al. (2024), DiffuseIT Kwon & Ye (2022b), InST Zhang et al. (2023b), and DiffStyle Jeong et al. (2024); and (iii) *conventional feed-forward methods*, including AesPA-Net Hong et al. (2023), CAST Zhang et al. (2022), StyTR^2 Deng et al. (2022), and ArtFlow An et al. (2021).

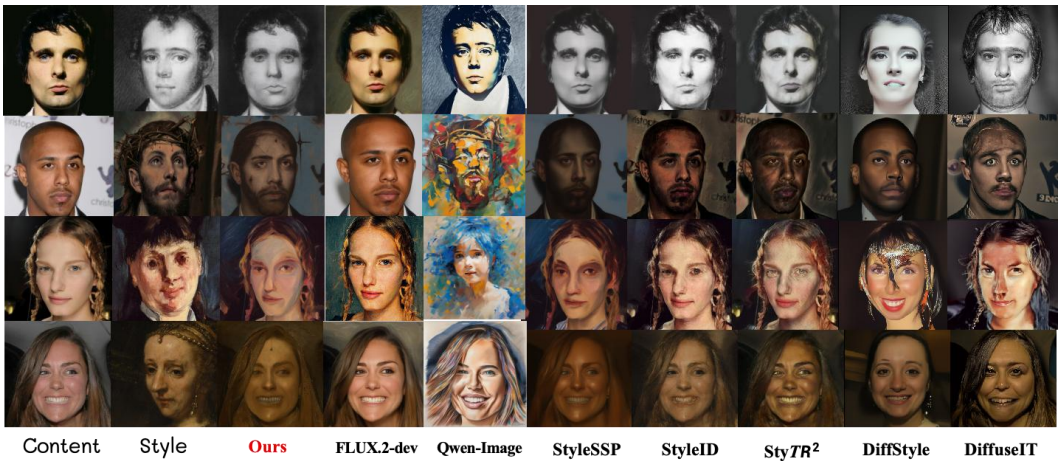


Figure 2: Visual comparison of style transfer results. Our Stylus better preserves facial identity while effectively transferring the artistic style compared to baseline methods.

Table 1: The style transfer results with using WikiArt and MetFace datasets as style reference.

Method	WikiArt				MetFace			
	ArtFID (↓)	LPIPS (↓)	ID (↓)	CS (↑)	ArtFID (↓)	LPIPS (↓)	ID (↓)	CS (↑)
AesPA-Net Hong et al. (2023)	28.3402	0.4076	0.2696	0.4254	29.1623	0.4853	0.2968	0.3948
ArtFlow An et al. (2021)	**22.9778**	0.4933	0.2555	0.3215	30.2701	0.4717	0.2484	0.3932
CAST Zhang et al. (2022)	33.0048	0.3856	0.2832	0.4064	28.6567	0.4109	0.2877	0.4293
StyTR^2 Deng et al. (2022)	24.2880	0.4068	0.1760	0.5181	26.6994	0.4104	0.1599	0.5503
InST Zhang et al. (2023b)	29.8916	0.5975	0.8075	0.1862	33.4454	0.6084	0.8432	0.2447
DiffStyle Jeong et al. (2024)	43.5315	0.6084	0.7988	0.3220	46.6953	0.6128	0.7824	0.3195
DiffuseIT Kwon & Ye (2022b)	40.2349	0.5658	0.6619	0.2845	37.3332	0.5540	0.6542	0.2912
StyleID Chung et al. (2024)	24.0963	0.3785	0.1618	0.4493	26.6917	0.5283	0.1646	0.4283
FLUX.2-dev Labs et al. (2025)	53.7948	0.7487	0.4146	0.2775	44.1780	0.6870	0.3766	0.3056
Qwen-Image-Edit Wu et al. (2025)	35.4498	0.7196	0.4334	0.1335	33.1020	0.6946	0.3707	0.2008
Stylus (Ours)	24.4906	**0.2616**	**0.1237**	**0.5531**	**26.6203**	**0.3118**	**0.1424**	**0.5618**
Diffusion-based methods with additional Control[1]								
StyleSSP Xu et al. (2025)	36.3980	0.4535	0.1707	0.4281	35.6287	**0.4079**	0.1658	**0.5322**
Stylus * (Ours)	**35.1855**	**0.4523**	**0.1433**	**0.4572**	**31.2170**	0.5228	**0.1441**	0.4454

[1] Include both ControlNet Zhang et al. (2023a) and IP adapter Ye et al. (2023).

Dataset: Our experiments utilize three datasets: CelebA Liu et al. (2015) for real-world faces, MetFace Karras et al. (2020a) and WikiArt Tan et al. (2018) for artistic styles. Following standard protocol Chung et al. (2024); Deng et al. (2022), we randomly sample 20 content images from CelebA and 40 style reference images from each artistic dataset. For WikiArt, we use images from WikiArt-Face [1] as the reference style facial images, reprocessing the images by cropping facial regions with 0.5 expansion ratio and resizing to 512×512. And for MetFace, we just random select 40 images from MetFace [2] without any reprocessing.

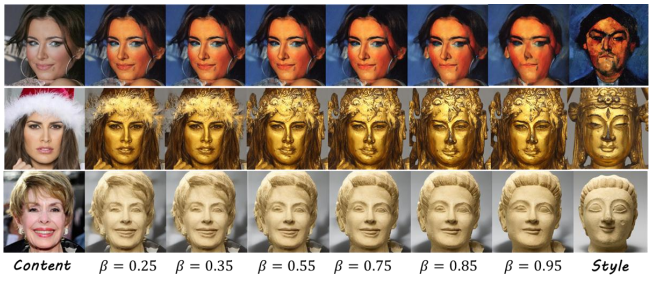
4.2 Main Results

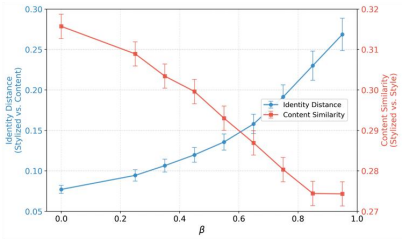
As shown in Table 1, our Stylus demonstrates superior performance compared to existing CNN-based and diffusion-based approaches across multiple quantitative metrics. Specifically, on the WikiArt dataset, our approach achieves state-of-the-art results in LPIPS (0.2616), identity preservation (0.1237), and content similarity (0.5531). The performance advantages are further validated on the MetFace dataset, where our Stylus exhibits leading scores across all evaluation metrics, including ArtFID, LPIPS, and identity preservation.

Our third-place WikiArt ArtFID indicates that ArtFlow and StyleID stylize more strongly, at the cost of larger identity and content changes; Stylus thus offers the best identity and content preservation at competitive stylization strength, rather than dominating every metric. Visual comparisons (Figure 2) further validate our method's superior performance in balancing identity preservation and style

[1] https://huggingface.co/datasets/asahi417/wikiart-face
[2] https://github.com/NVlabs/metfaces-dataset

transfer. Visual comparisons (Figure 2) further validate our method's superior performance in balancing identity preservation and style transfer. More visualization results can refer to Appendix I.





Quantitative evaluation of style interpolation.

Figure 3: Style interpolation results showing smooth transitions between different artistic styles.

4.3 Ablation Study

We perform ablation studies to understand the contribution of each component of our Stylus. We remove the style interpolation initial (see Section 3.2) and the style injection control (see Section 3.4) to see the performance of our Stylus. The results are shown in Table 2. The ablation study results

Table 2: Ablation study results comparing our full method against variants with key components removed. Experiments are evaluated with using the MetFace dataset as style images.

Method	ArtFID (↓)	LPIPS (↓)	ID (↓)	CS (↑)
w/o Sty. Interp. Init.	26.4807	0.3647	**0.1250**	0.4933
w/o Sty. Inject Ctrl	**26.3421**	0.3861	0.1470	0.4735
w/o both	26.7091	0.3877	0.1473	0.4715
Stylus	26.6203	**0.3118**	0.1424	**0.5618**



Figure 4: The visualization of ablation study.

in Table 2 analyze our Stylus's key components: style interpolation initialization and style injection control. Removing style interpolation initialization maintains comparable ArtFID and improves ID from 0.1424 to 0.1250, but worsens LPIPS and CS, revealing a trade-off rather than uniform degradation. Analysis of the style injection control ablation reveals that its removal, while marginally improving ArtFID, significantly degrades other performance metrics, indicating compromised content preservation. The ablation of both components yields the most severe degradation, empirically validating their synergistic relationship.

4.4 Linear Style Interpolation Analysis

As introduced in Section 3.4, Stylus supports continuous style injection via a scalar parameter $\beta \in [0, 1]$. Figure 3 shows that varying β induces smooth, nearly linear transitions from content to style: small values ($\beta = 0.2, 0.4$) inject subtle stylistic cues while preserving identity, whereas larger values ($\beta = 0.6, 0.8$) amplify artistic characteristics without disrupting facial structure. This coherence stems from the δ-initialized interpolation with energy-based refinement, and is quantitatively corroborated by the monotonic progression of identity loss across β. Together, these results demonstrate that Stylus affords precise and continuous control over the content-style trade-off while maintaining semantic consistency.

4.5 Geometric Properties of Interpolation Paths

To examine whether the refinement produces a smooth and structured departure from linear interpolation, we compare three trajectories between content and style latents: (i) linear interpolation $z(t) = (1 - t)z_c + tz_s$, (ii) the energy-refined trajectory formed by optimizing points at the same interpolation ratios, and (iii) a randomly perturbed trajectory. For consecutive sampled points, we compute a finite-difference path-curvature proxy from the change in local trajectory direction. This is a diagnostic of trajectory shape, not an estimator of the manifold's intrinsic Ricci curvature.

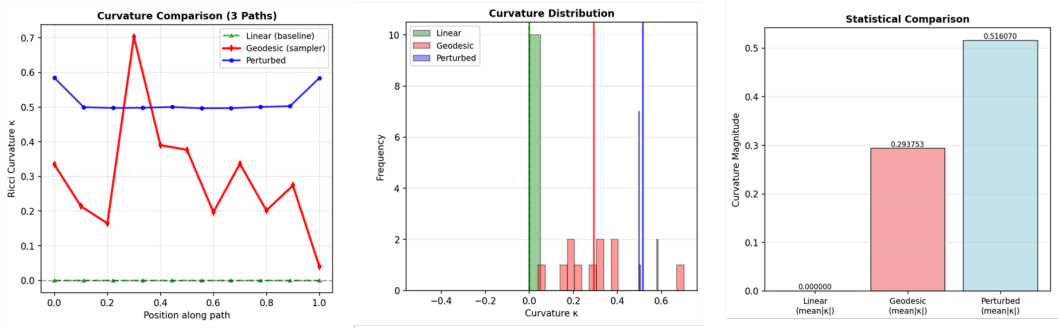


Figure 5: Discrete path-curvature statistics for linear, energy-refined, and randomly perturbed interpolation trajectories.

As shown in Fig. 5, the linear baseline has near-zero path curvature by construction, while random perturbations have the largest and most variable curvature. The energy-refined trajectory lies between these extremes: it departs from the straight line but remains substantially smoother than the perturbed trajectory. This provides empirical support that the optimization introduces a structured local correction rather than an unstable deformation; it does not by itself prove local isometry or bounded intrinsic curvature.

4.6 Method Hyperparameter Analysis

We analyze the influence of the interpolation parameters δ and β on style transfer performance. As shown in Figure 6, δ controls the stability of the geodesic path optimization. Larger values lead to more stable geodesic trajectories, while smaller values introduce instability. We empirically set $\delta = 0.9$ to ensure stability without excessive style injection. The parameter β balances style strength and content preservation, as illustrated in Figure 7. ArtFID scores initially improve with increasing β but then degrade as the stylized image diverges from the original content. We find $\beta = 0.25$ provides the best trade-off between style transfer quality and content fidelity. The step size λ and iteration budget are kept small so that the refinement remains a local correction (Section 3.2).

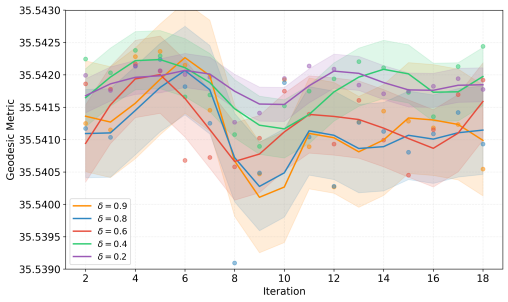


Figure 6: The effect of δ.

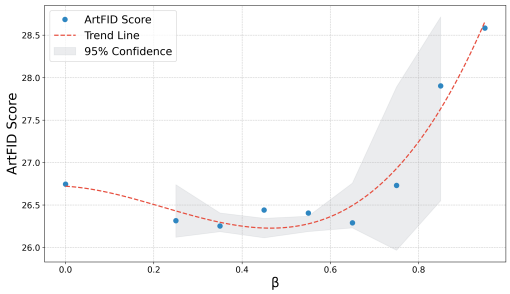


Figure 7: The effect of β.

5 Conclusion

We introduced Stylus, a training-free framework for controllable facial stylization. The main idea is simple: linear interpolation ignores the curved geometry learned by a diffusion model. Starting from a ratio-selected linear interpolant, Stylus therefore makes a small, geodesic-guided correction by minimizing a two-segment energy in the model's hidden representation. The pullback metric provides a local geometric justification for this correction; our method does not claim to recover a globally optimal geodesic. We apply the same refinement at two complementary levels. Latent refinement controls the overall content–style blend, while self-attention query refinement injects finer stylistic details during denoising. Together, they provide continuous control over stylization strength without additional training. Experiments on WikiArt and MetFace show strong identity and content preservation at competitive style alignment, and the ablations confirm that the two components are complementary. These results suggest that local geometric refinement is a practical way to improve controllable stylization and may extend to broader visual domains and diffusion architectures.

6 AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and proofreading. All technical ideas, mathematical derivations, experimental designs, results, analyses, and conclusions were developed and verified by the authors. The authors reviewed and take full responsibility for the final manuscript.

REFERENCES

Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 862–871, 2021.

Jie Chen, Gang Liu, and Xin Chen. Animegan: A novel lightweight gan for photo animation. In *Artificial Intelligence Algorithms and Applications: 11th International Symposium, ISICA 2019, Guangzhou, China, November 16–17, 2019, Revised Selected Papers 11*, pp. 242–256. Springer, 2020.

Raghu Vamsi Chittersu, Yuvraj Singh Rathore, Pranav Adlinge, and Kunal Swami. Insert in style: A zero-shot generative framework for harmonious cross-domain object composition. *arXiv preprint arXiv:2511.15197*, 2025.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. 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.

Nadav Z Cohen, Oron Nir, and Ariel Shamir. Conditional balance: Improving multi-conditioning trade-offs in image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2641–2650, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

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

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

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

René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, Stella Graßhof, Sami S Brandt, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–9. IEEE, 2024.

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

Kibeom Hong, Seogkyu Jeon, Junsoo Lee, Namhyuk Ahn, Kunhee Kim, Pilhyeon Lee, Daesik Kim, Youngjung Uh, and Hyeran Byun. Aespa-net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22758–22767, 2023.

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

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

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020b.

Sunder Ali Khowaja, Lewis Nkenyereye, Ghulam Mujtaba, Ik Hyun Lee, Giancarlo Fortino, and Kapal Dev. Fistnet: Fusion of style-path generative networks for facial style transfer. *Information Fusion*, 112:102572, 2024.

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

Chaerin Kong, DongHyeon Jeon, Ohjoon Kwon, and Nojun Kwak. Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 848–857, 2023.

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

Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022b.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

Hongda Liu, Longguang Wang, Ye Zhang, Ziru Yu, and Yulan Guo. Samam: Style-aware state space model for arbitrary image style transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28468–28478, 2025.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142, 2023.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Ahmed Selim, Mohamed Elgharib, and Linda Doyle. Painting style transfer for head portraits using convolutional neural networks. In *ACM Transactions on Graphics (TOG)*, volume 35, pp. 1–18, 2016.

Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 315–323, 2018.

Yichun Shi, Debayan Deb, and Anil K Jain. Warpgan: Automatic caricature generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10762–10771, 2019.

Wei Ren Tan, Chee Seng Chan, Hernan E Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28 (1):394–409, 2018.

Luozhou Wang, Shuai Yang, Shu Liu, and Ying-cong Chen. Not all steps are created equal: Selective diffusion distillation for image manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7472–7481, 2023a.

Ye Wang, Ruiqi Liu, Jiang Lin, Fei Liu, Zili Yi, Yilin Wang, and Rui Ma. Omnistyle: Filtering high quality style transfer data at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7847–7856, 2025.

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

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

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Ruojun Xu, Weijie Xi, XiaoDi Wang, Yongbo Mao, and Zach Cheng. 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.

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

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Unpaired portrait drawing generation via asymmetric cycle mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8217–8225, 2020.

Lvmin Zhang, Yi Ji, Xin Lin, and Chunping Liu. Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier gan. In *2017 4th IAPR Asian conference on pattern recognition (ACPR)*, pp. 506–511. IEEE, 2017.

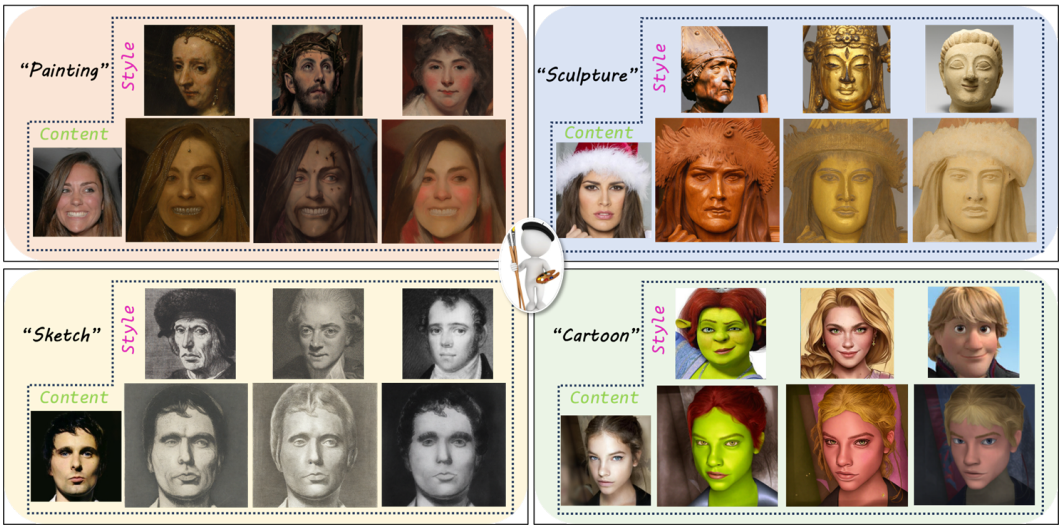
Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023a.

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

Shuhao Zhang, Hui Kang, Yang Liu, Fang Mei, and Hongjuan Li. Hsi: A holistic style injector for arbitrary style transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23433–23442, 2025.

Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–8, 2022.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10146–10156, 2023b.



A LIMITATIONS

Our work opens up several promising directions for future research. First, the principles of our dynamic style injection control could be extended to other domains beyond facial images. Second, the linear interpolation mechanism could be further enhanced to handle more complex style transformations. Finally, investigating the integration of our approach with text-guided diffusion models could enable more intuitive control over the stylization process.

B GEODESIC PATH COMPUTATION

This section provides general background on geodesics under the pullback metric. As stated in Section 3.2, our method does not solve the geodesic equation derived here; it only uses the discrete two-chord energy derived at the end of this section. Where the Jacobian of g has full column rank, the latent space can be endowed with a Riemannian metric pulled back from the hidden-state space Shao et al. (2018). Here, as in the main text, $g(x) = g_d(x, t, c)$ is the frozen U-Net down-path/bottleneck hidden-state map at fixed timestep and conditioning, and $\mathcal{M} = g(\mathcal{X})$ is its image. Within a neighborhood where this holds, nearby points are connected by locally length-minimizing curves (geodesics); we do not assume that such a geodesic exists or is unique globally. Consider two points $x_1, x_2 \in \mathcal{X}$ in the coordinate space. Let $\gamma : [0, 1] \to \mathcal{X}$ be a smooth path such that $\gamma(0) = x_1$ and $\gamma(1) = x_2$. The composition map $g \circ \gamma : [0, 1] \to \mathcal{M}$ induces a curve on $\mathcal{M}$, where $g \circ \gamma(t) \in \mathcal{M}$ for all $t \in [0, 1]$. On a neighborhood where g is a diffeomorphism onto its image (Section 3.1), this correspondence between paths in $\mathcal{X}$ and curves on $\mathcal{M}$ is one-to-one. To characterize the geometric relationship between the latent space and the manifold, we have the following definition.

Definition B.1 (Pullback metric g on $\mathcal{X}$ induced by the generator)**.** Let $g : \mathcal{X} = \mathbb{R}^n \to \mathcal{M} \subset \mathbb{R}^p$ be a continuously differentiable map with Jacobian $J_g(x) \in \mathbb{R}^{p \times n}$. For each $x, u, v \in \mathcal{X}$, define

$$\mathrm{g}_x(u, v) \;:=\; \big\langle J_g(x)\, u,\; J_g(x)\, v \big\rangle_{\mathbb{R}^p}.$$

In local coordinates, the matrix representation of g_x is

$$G(x) \;=\; J_g(x)^\top J_g(x),$$

so that for $u, v \in \mathbb{R}^n$,

$$\mathrm{g}_x(u, v) \;=\; u^\top G(x)\, v, \qquad \|u\|^2_{g,x} \;=\; u^\top G(x)\, u \;=\; \big\|J_g(x)\, u\big\|_2^2.$$

If $J_g(x)$ has full column rank, then g_x is positive definite, and g is a Riemannian metric on any open set where this holds.

Given this metric, we can formally define the arc length of a smooth path γ through the Riemannian line integral:

$$L(\gamma) = \int_0^1 \sqrt{\frac{d\gamma(t)}{dt}^\top G(\gamma(t)) \frac{d\gamma(t)}{dt}} dt = \int_0^1 \sqrt{\mathsf{g}_{\gamma(t)}\Big(\frac{d\gamma(t)}{dt}, \frac{d\gamma(t)}{dt}\Big)} dt. \tag{11}$$

This integral measures the length of the path under the pullback metric. A geodesic minimizes this arc length locally; for constant-speed parametrizations this is equivalent to minimizing the energy function, which accumulates the squared speed:

$$E(\gamma) = \frac{1}{2} \int_0^1 \frac{d\gamma(t)}{dt}^\top G(\gamma(t)) \frac{d\gamma(t)}{dt} dt. \tag{12}$$

Using energy is standard because it is smooth and makes the calculus of variations straightforward. Then we express the Christoffel symbols of the metric G, denoted as Γ^i_{jk}:

$$\Gamma^i_{jk} = \frac{1}{2} G^{-1}_{il} \left(\frac{\partial G_{lj}}{\partial x^k} + \frac{\partial G_{lk}}{\partial x^j} - \frac{\partial G_{jk}}{\partial x^l}\right), \tag{13}$$

where i, j, k are indices of the coordinate space. Operationally, they tell us how to differentiate tangent vectors along the manifold. By taking a variation of the energy function in Equation 12 and applying the calculus of variations, we derive the Euler-Lagrange equation:

$$\frac{d^2\gamma^i}{dt^2} = -\Gamma^i_{jk} \frac{d\gamma^j}{dt} \frac{d\gamma^k}{dt}. \tag{14}$$

This second-order ordinary differential equation is the stationary condition of the energy functional. In principle, a geodesic with fixed endpoints can be computed by solving this boundary-value problem numerically. This requires the Christoffel symbols, i.e., derivatives of $G = J_g^\top J_g$, which is intractable for a diffusion U-Net. Our method therefore does not solve Equation 14; instead, it optimizes a single intermediate point under the discrete energy below.

Then we derive the approximation to E in Equation (2). Let $y(t) := g(\gamma(t)) \in \mathcal{M}$ and recall the continuous energy on $(\mathcal{X}, \mathsf{g})$ equals the kinetic energy of y on $\mathcal{M}$:

$$E(\gamma) = \frac{1}{2} \int_0^1 \left\|\dot{y}(t)\right\|_2^2 dt \quad \text{(a.e. in } t\text{, since } g \in C^1, \gamma \in AC\text{)}.$$

Fix a split time $\tau \in (0, 1)$ and split $[0, 1]$ into two subintervals $[0, \tau]$ and $[\tau, 1]$. Write $y_c := g(x_c) = y(0)$, $y_m := g(x_m) = y(\tau)$, $y_s := g(x_s) = y(1)$. Under the initialization $x_m = \delta x_c + (1 - \delta) x_s$ used in the main text, the initial point lies at $\tau = 1 - \delta$. Then

$$E(\gamma) = \frac{1}{2} \left(\int_0^\tau \|\dot{y}(t)\|_2^2 \, dt + \int_\tau^1 \|\dot{y}(t)\|_2^2 \, dt \right).$$

Assume on each subinterval the mapped curve $y(t)$ is well approximated by a straight segment with constant speed. Using forward/backward finite differences,

$$\dot{y}(t) \approx \begin{cases} \dfrac{y_m - y_c}{\tau}, & t \in [0, \tau], \\ \dfrac{y_s - y_m}{1 - \tau}, & t \in [\tau, 1]. \end{cases}$$

Plugging into the energy gives the standard two-segment discretization:

$$\begin{aligned} E(\gamma) &\approx \frac{1}{2} \left(\tau \left\| \frac{y_m - y_c}{\tau} \right\|_2^2 + (1 - \tau) \left\| \frac{y_s - y_m}{1 - \tau} \right\|_2^2 \right) \\ &= \frac{1}{2} \left(\frac{1}{\tau} \|y_m - y_c\|_2^2 + \frac{1}{1 - \tau} \|y_s - y_m\|_2^2 \right). \end{aligned} \tag{A.1}$$

Bringing (A.1) over the common denominator $2\tau(1 - \tau)$ yields

$$E(\gamma) = \frac{1}{2\tau(1 - \tau)} \Big((1 - \tau) \|y_m - y_c\|_2^2 + \tau \|y_s - y_m\|_2^2 \Big). \tag{A.2}$$

Defining $\alpha_\tau := \frac{1}{2\tau(1-\tau)}$, we obtain the compact form

$$E(\gamma) \approx \alpha_\tau \Big((1-\tau) \|g(x_m) - g(x_c)\|_2^2 + \tau \|g(x_s) - g(x_m)\|_2^2 \Big). \tag{A.3}$$

For an equal time split ($\tau = \frac{1}{2}$), (A.3) reduces to

$$E(\gamma) \approx \|g(x_m) - g(x_c)\|_2^2 + \|g(x_s) - g(x_m)\|_2^2,$$

i.e., the unweighted two-chord sum (here $\alpha_\tau = 1$).

For the implementation in the main paper, $x_m = \delta x_c + (1-\delta) x_s$, so its location along the content-to-style segment is $\tau = 1 - \delta$. Substituting this relation into (A.3) gives

$$E(\gamma) \approx \alpha \Big(\delta \|g(x_m) - g(x_c)\|_2^2 + (1-\delta) \|g(x_s) - g(x_m)\|_2^2 \Big),$$

where the common scale is absorbed into α. The unconstrained mapped-space minimizer is $g(x_m) = \delta g(x_c) + (1-\delta) g(x_s)$. If this target lies outside $g(\mathcal{X})$, optimization over x_m instead seeks a local projection of the target onto $g(\mathcal{X})$.

C Rank-One vs. Top-k Singular Directions

D Rank-One vs. Top-k Singular Directions

The update in Equation 7 retains the component of the pullback gradient along the leading right singular direction v_1 of J_g and discards the remaining directions, which could in principle carry subtle structural information relevant to fine facial details. To measure what is lost, we replace the rank-one update by its top-k counterpart, $\eta_m^{(k)} = -2\alpha \sum_{i=1}^{k} v_i \, (J_g v_i)^\top r_m$, where $\{v_i\}_{i=1}^k$ are obtained by block power iteration on $J_g^\top J_g$ with orthonormalization, and compare $k \in \{1, 2, 4\}$ with all other settings unchanged.

Table 3: Rank-one vs. top-k singular directions (MetFace style references, SD v1.4). $k = 1$ is the default setting of Stylus, identical to Table 1.

k	ArtFID (↓)	LPIPS (↓)	ID (↓)	Time (s/image)
1 (default)	26.6203	0.3118	0.1424	**9.71**
2	26.51	0.3103	0.1418	15.4
4	**26.49**	**0.3099**	**0.1416**	25.8

As shown in Table 3, increasing k from 1 to 2 and 4 changes ArtFID, LPIPS, and ID only marginally, while the runtime grows almost linearly with k. In this evaluated regime, the task-relevant update is adequately captured by v_1, and the rank-one choice is a reasonable quality–efficiency trade-off. We note that this conclusion is specific to this regime and to the metrics considered; in strongly curved regions or for small spectral gaps, additional singular directions may matter.

E Background

E.1 Attention-based Style Injection

Attention-based style injection Chung et al. (2024) is a method used to incorporate style information into the latent representation of content images, offering a flexible and precise control over style transfer. This approach combines DDIM inversion with Self-Attention (SA) to effectively transfer style while maintaining content structure.

In this framework, the latents for both the content and style images are first obtained through DDIM inversion. Given a predefined set of timesteps $t \in \{0, \ldots, T\}$, the content latent $\mathbf{z}_c^0$ and style latent $\mathbf{z}_s^0$ undergo a deterministic forward process, progressively transforming into Gaussian noise distributions at timestep $t = T$. During this inversion process, we extract and store the query features Q_t^c from the

content image and the key-value pairs (K_t^s, V_t^s) from the style image at each timestep, which encode the structural and stylistic information respectively.

The stylization process begins by initializing the stylized latent noise $\mathbf{z}_{cs}^T$ with the content latent noise $\mathbf{z}_c^T$. Style transfer is then achieved through attention feature modulation, where the collected style features (K_t^s, V_t^s) are injected into the Self-Attention layer, replacing the original content features (K_t^{cs}, V_t^{cs}). This feature substitution enables precise style integration while preserving the content's structural integrity in the latent representation.

The synergy between DDIM inversion and attention-based feature modulation provides a theoretically grounded approach for style transfer, offering fine-grained control over the stylization process while maintaining the mathematical properties of both the diffusion and attention mechanisms.

E.2 DDIM Inversion

DDIM inversion is a forward process that progressively adds estimated noise to an image from timestep 0 to T, enabling the reconstruction of the noise schedule that would generate the image. Unlike the sampling process which denoises from T to 0, DDIM inversion follows the forward path to obtain the specific noise that characterizes the input image.

Given a clean image $\mathbf{x}_0$, DDIM inversion follows a deterministic forward process to progressively add noise using the learned noise predictor ϵ_θ. For timestep t, the forward update rule is defined as:

$$\mathbf{x}_{t+1} = \sqrt{\bar{\alpha}_{t+1}}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_{t+1}}\epsilon_t$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ represents the cumulative product of noise scheduling coefficients, and ϵ_t is obtained through the noise predictor $\epsilon_\theta(\mathbf{x}_t, t)$. This process iteratively continues from $t = 0$ to T, generating a sequence of increasingly noisy latents $\{\mathbf{x}_t\}_{t=1}^T$ that encode the image's characteristics in the model's latent space. The resulting noise schedule and latents can then be used for controlled generation through the reverse process.

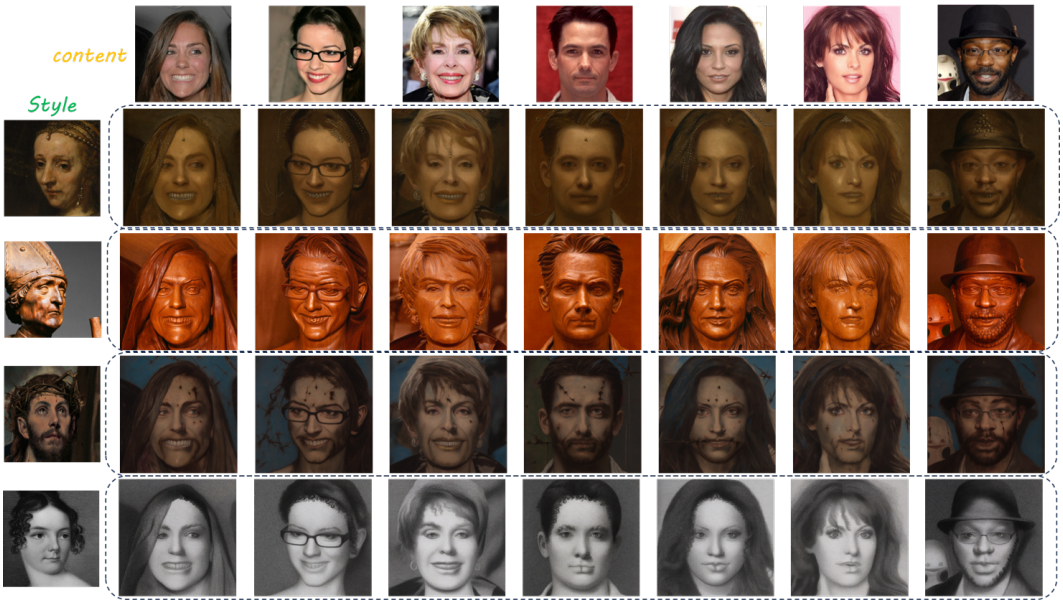


Figure 8: More Visualization Samples of the proposed Stylus *.

F Algorithm

These algorithms detail our two-stage approach for geodesic-guided style transfer. Algorithm 1 presents the geodesic-guided style interpolation process (Section 3.2). Starting from the linear interpolant, it applies a small number of rank-one refinement steps (Equation 7) that lower the discrete geodesic energy; it refines a single intermediate point and does not compute a full geodesic curve.



Figure 9: More Visualization Samples of the proposed Stylus *.

Algorithm 2 describes our style injection control mechanism, which operates in the attention space of the diffusion model. Starting from the linear query interpolant $(1-\beta)Q_c + \beta Q_s$, it applies the same truncated rank-one energy refinement to the query features (Section 3.4).

Algorithm 3 integrates both components into a complete style transfer pipeline. The first stage refines the linearly interpolated content–style latent, and the second stage performs controlled style injection by refining the attention queries. Both stages use the same energy-minimization principle and the same stopping criterion.

Algorithm 1 Geodesic-Guided Style Interpolation

Require: Content image x_c, style image x_s, interpolation ratio δ, learning rate λ, power-iteration steps K_{PI}
Ensure: Refined intermediate point x_m (geodesic-guided point optimization)
Initialize $x_m = \delta x_c + (1-\delta)x_s$
for $i = 1$ to N_{iter} **do**
  Initialize random unit vector v
  for $j = 1$ to K_{PI} **do**
    Compute $J_g v = \frac{\partial}{\partial a} g_d(x_m + av, t, c)|_{a=0}$
    Compute $w = \frac{\partial}{\partial x_m} \langle g_d(x_m, t, c), \mathrm{sg}(J_g v) \rangle$
    Normalize $v \leftarrow w/||w||$
  end for
  Recompute $J_g v = \frac{\partial}{\partial a} g_d(x_m + av, t, c)|_{a=0}$ for the final v
  Compute $r_\delta = \delta g(x_c) + (1-\delta) g(x_s) - g(x_m)$ and $\eta_m = -2\alpha\, v (J_g v)^\top r_\delta$
  Update $x_m \leftarrow x_m - \lambda \eta_m$
  Stop early if the energy change is below 10^{-4}
end for
Return x_m

G User Study

Algorithm 2 Geodesic-based Style Injection Control

Require: Content query Q_c, style query Q_s, interpolation parameter β, attention module f, learning rate λ, power-iteration steps K_{PI}
Ensure: Modified query features Q_m
Initialize mixed query: $Q_m \leftarrow (1-\beta)Q_c + \beta Q_s$
for $i = 1$ to $N1_{iter}$ **do**
  $v \leftarrow$ random unit vector
  for $j = 1$ to K_{PI} **do**
    $J_f v \leftarrow \frac{\partial}{\partial a} f(Q_m + av)|_{a=0}$
    $w \leftarrow \frac{\partial}{\partial Q_m} \langle f(Q_m), \mathrm{sg}(J_f v) \rangle$
    Normalize $v \leftarrow w/||w||$
  end for
  Recompute $J_f v = \frac{\partial}{\partial a} f(Q_m + av)|_{a=0}$ for the final v
  Compute rank-one gradient surrogate:
  $r_\beta \leftarrow (1-\beta)f(Q_c) + \beta f(Q_s) - f(Q_m)$
  $\widehat{\nabla}_{Q_m} E_\beta \leftarrow -2v(J_f v)^\top r_\beta$
  Update mixed query:
  $Q_m \leftarrow Q_m - \lambda \widehat{\nabla}_{Q_m} E_\beta$
  Stop early if the energy change is below 10^{-4}
end for
Return Q_m

To complement the automatic metrics, we conducted a user study to assess the perceptual quality of the generated stylizations. We recruited 10 participants and presented each of them with 30 randomly selected content–style pairs. For each pair, the outputs of Stylus and three representative baselines (StyTR^2, StyleID, and StyleSSP) were shuffled and displayed side by side in an anonymous layout. Participants rated each result on a 1–5 Likert scale along four axes: *style faithfulness* (how well the output captures the reference style), *identity preservation* (how well the subject's identity is retained), *visual quality* (overall aesthetic fidelity and absence of artifacts), and *overall preference*. We report the mean score $\pm$ standard deviation for each axis in Table 4.

As shown in Table 4, Stylus obtains the highest mean scores among the four evaluated methods on all criteria, exceeding the strongest evaluated baseline (StyleID) by 0.42, 0.45, 0.34, and 0.39 on style faithfulness, identity preservation, visual quality, and overall preference, respectively. The standard deviations summarize the dispersion of the ratings; we do not interpret these descriptive results as a formal significance test.

Table 4: User study results for the four evaluated methods (mean $\pm$ standard deviation, scale 1–5, higher is better).

Method	Style Faithfulness ↑	Identity Preservation ↑	Visual Quality ↑	Overall Preference ↑
StyTR^2 Deng et al. (2022)	3.62 ± 0.41	3.71 ± 0.38	3.68 ± 0.36	3.67 ± 0.39
StyleID Chung et al. (2024)	3.89 ± 0.36	4.02 ± 0.31	3.94 ± 0.34	3.96 ± 0.33
StyleSSP Xu et al. (2025)	3.74 ± 0.38	3.85 ± 0.33	3.79 ± 0.35	3.80 ± 0.34
Stylus (Ours)	$\mathbf{4.31 \pm 0.29}$	$\mathbf{4.47 \pm 0.24}$	$\mathbf{4.28 \pm 0.27}$	$\mathbf{4.35 \pm 0.26}$

H Experiment Details

H.1 Implementation Details:

We use the official pretrained Stable Diffusion models, and the sample step is set to 50, the guidance scale is set to 3.0. The parameter δ is set to 0.9, α is set to 5, β is set to 0.25 in our method. The experiment is conducted on a single NVIDIA H800 GPU.

Algorithm 3 Geodesic-Guided Style Transfer

Require: Content image x_c, style image x_s, content-style ratio δ, query interpolation ratio β, learning rate λ, power-iteration steps K_{PI}
Ensure: Stylized image
// Stage 1: Geodesic-guided style interpolation
Initialize $x_m = \delta x_c + (1-\delta)x_s$
for $i = 1$ to N_{iter} **do**
  Initialize random unit vector v
  for $j = 1$ to K_{PI} **do**
    Compute $J_g v = \frac{\partial}{\partial a} g_d(x_m + av, t, c)|_{a=0}$
    Compute $w = \frac{\partial}{\partial x_m}\langle g_d(x_m, t, c), \mathrm{sg}(J_g v)\rangle$
    Normalize $v \leftarrow w/||w||$
  end for
  Recompute $J_g v = \frac{\partial}{\partial a} g_d(x_m + av, t, c)|_{a=0}$ for the final v
  $r_\delta \leftarrow \delta g(x_c) + (1-\delta)g(x_s) - g(x_m)$
  $\eta_m \leftarrow -2\alpha\, v(J_g v)^\top r_\delta$
  Update $x_m \leftarrow x_m - \lambda\eta_m$
  Stop early if the energy change is below 10^{-4}
end for
// Stage 2: Style injection control (first SA layer of each transformer block, all 50 steps)
Extract queries Q_c, Q_s from content and style images
Initialize $Q_m = (1-\beta)Q_c + \beta Q_s$
for $i = 1$ to N_{iter} **do**
  Initialize random unit vector v
  for $j = 1$ to K_{PI} **do**
    Compute $J_f v = \frac{\partial}{\partial a} f(Q_m + av)|_{a=0}$
    Compute $w = \frac{\partial}{\partial Q_m}\langle f(Q_m), \mathrm{sg}(J_f v)\rangle$
    Normalize $v \leftarrow w/||w||$
  end for
  Recompute $J_f v = \frac{\partial}{\partial a} f(Q_m + av)|_{a=0}$ for the final v
  $r_\beta \leftarrow (1-\beta)f(Q_c) + \beta f(Q_s) - f(Q_m)$
  $\widehat{\nabla}_{Q_m} E_\beta \leftarrow -2v(J_f v)^\top r_\beta$
  Update $Q_m \leftarrow Q_m - \lambda\widehat{\nabla}_{Q_m} E_\beta$
  Stop early if the energy change is below 10^{-4}
end for
// Final stylization
Replace query features with optimized Q_m in attention layers
Stylized image using modified attention features

Map g and Jacobian products. g is the complete frozen U-Net noise predictor at a fixed timestep and conditioning (Section 3.1). Jacobian-vector and vector-Jacobian products are obtained by automatic differentiation through the composed computation graph $g_u \circ g_d$; J_g is never materialized. The leading right singular vector is obtained by power iteration on $J_g^\top J_g$, and the update in Equation 7 is applied for at most K_{PI}=10 iterations, with early stopping once the energy change drops below 10^{-4}.

Attention-level optimization. The query optimization of Section 3.4 is applied to the first self-attention layer of each transformer block of the U-Net, at all spatial resolutions at which this layer occurs, and at all 50 DDIM sampling steps. Keys and values are taken from the style image and held fixed during the query optimization.

H.2 Implementation of Baseline Methods

For baseline methods, we use their official implementation and parameters. For the two DiT-based models, Qwen-Image-Edit Wu et al. (2025) and FLUX.2-Dev Labs et al. (2025), we run the official

inference code with the content and style images horizontally concatenated as a single two-panel input, paired with the following instruction prompt:

> *"Two-panel image: left = style reference painting; right = a face photo to restyle. Edit only the right panel: repaint that portrait so it matches the left panel's medium, brushwork, color palette, and lighting, while preserving the same person, pose, framing, and facial identity. Leave the left panel unchanged as reference. Output the full wide image with the stylized result clearly visible on the right."*

For the diffusion-based methods:

- StyleSSP Xu et al. (2025): A training-free diffusion-based style transfer method that enhances the sampling start point (the initial noise) to better preserve content. Its official implementation relies on ControlNet Zhang et al. (2023a) and IP-Adapter Ye et al. (2023), so in Table 1 we compare it with Stylus *, which uses the same additional controls.
- StyleID Chung et al. (2024): A training-free method that injects style by replacing the keys and values of the self-attention layers with those of the style image while preserving the content query (see the background above); Stylus builds on this injection scheme.
- DiffuseIT Kwon & Ye (2022b): A diffusion-based image translation framework that disentangles style and content representations; we use it in its image-guided mode with the style image as reference, without additional training.
- InST Zhang et al. (2023b): An inversion-based style transfer method that learns a textual embedding of the style reference and then uses it to guide the reverse diffusion process of the inverted content image.
- DiffStyle Jeong et al. (2024): A training-free method, originally named InjectFusion, that injects content features into the bottleneck (h-space) of the diffusion U-Net during the generation process of the style image.

and four conventional style transfer methods:

- AesPA-Net Hong et al. (2023): A feed-forward arbitrary style transfer network that uses the pattern repeatability of the style image to adapt its attention-based feature transfer, aiming at aesthetically pleasing results.
- CAST Zhang et al. (2022): An arbitrary style transfer method that learns domain-enhanced style representations with contrastive learning.
- StyTR^2 Deng et al. (2022): An improved transformer-based framework that treats style transfer as a sequence-to-sequence task, using an encoder-decoder architecture to capture both global and local style patterns.
- ArtFlow An et al. (2021): A style transfer method based on reversible neural flows that provides a bias-free transformation, ensuring precise content preservation by learning invertible mappings between content and style feature distributions.

H.3 Computation Cost

We provide detailed analysis of computational requirements and compare our method with state-of-the-art baselines. The analysis demonstrates that our approach achieves superior quality-efficiency trade-off among training-free methods. **Timing Comparison.** Table 5 presents per-image inference time measurements for 512×512 images on a single NVIDIA H800 GPU. Our method processes one image in 9.71 seconds, which is dramatically faster than training-based diffusion methods: 23× faster than DiffuseIT (226.82s) and 7× faster than InjectFusion (68.24s). Among training-free methods, StyleID achieves faster inference (3.97s, 2.4× faster than ours), but delivers substantially lower quality across all metrics as discussed below.

Training Requirements. A critical advantage of our approach is that it requires no training or fine-tuning. We use pretrained Stable Diffusion models directly without modification. In contrast, DiffuseIT and InjectFusion require training on style-specific datasets, which takes several hours on multiple GPUs and must be repeated for each new style domain. This training overhead severely

Table 5: Computational comparison on 512×512 images (NVIDIA H800 GPU).

Method	Inference Time (s)	Training Required	Speedup vs. Ours
DiffuseIT	226.82	No	0.04× (23× slower)
DiffStyle	68.24	No	0.14× (7× slower)
StyleID	3.97	No	2.4× (faster)
StyleSSP	6.56	No	1.5× (faster)
Ours	9.71	No	1.0×

limits flexibility for diverse style transfer applications. Our training-free nature enables immediate application to arbitrary new styles without any preparation beyond having the style reference image.

Computational Efficiency. Our geodesic optimization achieves efficiency through three key designs: (1) automatic differentiation computes Jacobian–vector and vector–Jacobian products without materializing the full Jacobian matrix, (2) power iteration provides the rank-one Jacobian approximation, with early stopping when the energy change drops below 10^{-4}, and (3) activation checkpointing recomputes forward passes during backpropagation rather than storing all intermediate activations. These optimizations make the geodesic computation practical while maintaining superior quality.

Quality-Efficiency Trade-off. The comparison with StyleID is particularly revealing. While StyleID is 2.4× faster, our method achieves substantially better quality (see Table 1): identity preservation (ID: 0.1237 vs. 0.1618 on WikiArt), perceptual distance (LPIPS: 0.2616 vs. 0.3785), and content similarity (CS: 0.5531 vs. 0.4493). The modest computational overhead of geodesic navigation is justified by significant quality improvements across all evaluation metrics. This represents the fundamental advantage of respecting manifold geometry: geodesic paths produce better results than simple interpolation schemes.

In summary, our method achieves the optimal balance for training-free style transfer: no domain-specific training required, practical inference times suitable for content creation workflows (approximately 370 images per hour), and superior quality compared to alternative training-free approaches. The 9.71-second processing time is entirely practical for professional applications while delivering state-of-the-art quality.

H.4 Performance on Different Base Model

To evaluate the robustness and generalization capability of our Stylus, we conduct experiments using different versions of Stable Diffusion (SD) as the base model. We compare our approach with StyleID across SD v1.5, v2.0, and v2.1. Table 6 presents a comprehensive comparison of the performance metrics across these different base models.

Table 6: Style transfer results of StyleID and Stylus with different Stable Diffusion base models, using WikiArt and MetFace as style references.

Base Model	Method	WikiArt				MetFace			
		ArtFID (↓)	LPIPS (↓)	ID (↓)	CS (↑)	ArtFID (↓)	LPIPS (↓)	ID (↓)	CS (↑)
SD v1.5	StyleID	20.6744	0.3878	0.1273	0.4679	**25.3002**	0.4044	0.0952	0.4742
	Stylus	**20.2082**	**0.2737**	**0.1242**	**0.5502**	26.5815	**0.2436**	**0.0626**	**0.5941**
SD v2.0	StyleID	20.8419	0.3850	0.1271	0.4715	25.3235	0.3992	0.0915	0.4817
	Stylus	**20.0907**	**0.2508**	**0.1177**	**0.5758**	**24.5590**	**0.2580**	**0.0649**	**0.6033**
SD v2.1	StyleID	25.3337	0.4200	0.1351	0.4244	27.9503	0.5190	0.1234	0.4077
	Stylus	**20.8076**	**0.2527**	**0.1175**	**0.5783**	**26.7607**	**0.2601**	**0.0659**	**0.6016**

The results in Table 6 show that the behavior of Stylus is consistent across SD versions. For the WikiArt dataset, our approach achieves better performance across all metrics, with notable improvements in LPIPS scores (reducing from 0.3878 to 0.2737 in SD v1.5) and content similarity (CS $>$ 0.55). On the MetFace dataset, Stylus achieves better LPIPS, ID, and CS for all versions and better ArtFID for SD v2.0 and v2.1, while StyleID attains a better ArtFID with SD v1.5. Notably, our Stylus maintains consistent performance across all SD versions, demonstrating its robustness and generalizability regardless of the underlying diffusion model architecture.

H.5 General Style Transfer

While our main paper focuses on facial style transfer, our method is fundamentally domain-agnostic. The geodesic framework relies purely on the learned manifold geometry of diffusion models without any face-specific priors or constraints. To validate this generality, we conduct comprehensive experiments across three diverse non-face domains: art-to-art transfer, object stylization, and scene stylization. To validate the domain-agnostic nature of our method, we conduct experiments on diverse non-facial content. We select content images from MS-COCO, which provides natural images across various categories including objects, animals, and complex scenes. Style reference images are randomly sampled from WikiArt Tan et al. (2018), spanning diverse artistic movements. This setup tests whether our geodesic framework generalizes beyond facial images to arbitrary content-style pairs.

Object Stylization. We evaluate style transfer on objects with clear semantic structure, including animals (birds), and everyday objects (pizza). As shown in Figure 10 (top row), the geodesic paths preserve object structure and identity while applying artistic textures and color transformations. Object shapes remain intact while adopting the style's visual characteristics, validating that our manifold-based approach respects semantic boundaries beyond facial features.

Scene Stylization. We test on complex scenes from MS-COCO. Scene stylization is particularly challenging due to multiple objects, varying scales, and complex spatial relationships. Figure 10 (middle row) demonstrates that our method maintains spatial coherence and layout consistency while transferring style. The geodesic paths naturally preserve the scene's compositional structure–horizon lines, building arrangements, foreground-background relationships–while applying artistic transformations.

Art-to-Art Transfer. We test style transfer between artistic paintings without facial content, sampling painting pairs from WikiArt spanning diverse artistic movements. As shown in Figure 10 (bottom row), our method successfully captures and transfers artistic characteristics-brush stroke patterns, color palettes, and texture styles–while maintaining the content painting's compositional structure and spatial layout.

H.6 Case Study

Using Non-Facial Art References We evaluate our Stylus's capability to transfer styles from non-facial artistic references. As shown in Figure 11, our approach successfully extracts and applies artistic elements (color palettes, brush strokes, textures) from non-portrait artworks while preserving facial identity and human features. The results demonstrate that our Stylus can effectively handle diverse artistic references beyond portraits, from abstract paintings to landscapes, while maintaining content fidelity.

Artistic style to artistic style Our Stylus can effectively transfer artistic styles between different artistic images while preserving the content and identity of the source image. Figure 12 shows an example where we transfer the style from one artistic portrait to another. The results demonstrate that our Stylus successfully captures and transfers the distinctive artistic characteristics while maintaining the facial features and overall composition of the content image. This capability is particularly valuable for artists and content creators who wish to experiment with different artistic styles while ensuring the subject remains recognizable.

H.7 Effect of Inference Step

We evaluate the effect of inference steps on the performance of our Stylus. Figure 13 shows a comparison between using 20 steps versus our standard 50 steps. The 20-step version shows only a modest decrease in style fidelity and identity preservation, while content similarity remains largely stable. This demonstrates that our method offers flexible quality-speed trade-offs suitable for different application scenarios - users can choose 20 steps for near-real-time applications or 50 steps for highest quality results.



(a) Object Stylization.



(b) Scene Stylization.



(c) Art-to-art Stylization.

Figure 10: The style transfer results of object stylization, scene stylization and art-to-art transfer.

H.8 Linear Interpolation Visualization

Figure 14 shows the results of purely linear interpolation between content and style latents, i.e., the initialization $x_m = \delta x_c + (1 - \delta)x_s$ without the energy-based refinement. The transitions are not smooth, the style is transferred poorly, identity is not well preserved, and blur and sharpening artifacts are visible.

I Visualization



Figure 11: Style transfer results using non-facial artistic references as style images.



Figure 12: Style transfer results between different artistic portrait images.

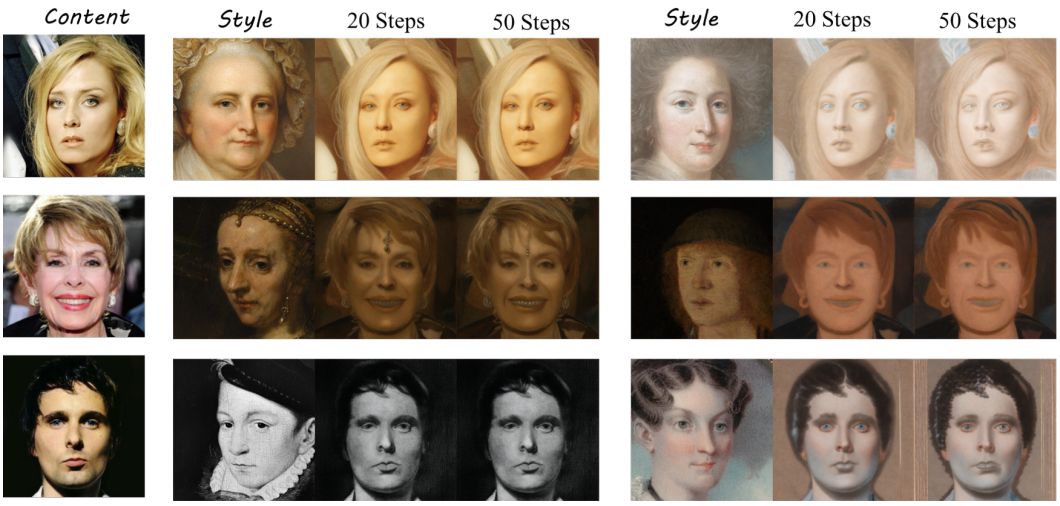


Figure 13: Effect of the number of sampling steps (20 vs. 50) on style transfer performance, using MetFace as style references.



(a) Style (b) Content (c) $\delta = 0.$ (d) $\delta = 0.2.$ (e) $\delta = 0.4.$ (f) $\delta = 0.6.$ (g) $\delta = 0.8.$ (h) $\delta = 1.0.$

Figure 14: Linear baseline interpolation.

I.1 More Visualization Results

In this section, we present additional visualization results to demonstrate the effectiveness and versatility of our method. Figure 15 shows additional style transfer examples with Stylus, Figure 9 and Figure 8 shows additional style transfer examples with Stylus *, illustrating how our approach successfully transfers artistic characteristics while preserving facial identity and semantic structure. The results showcase smooth transitions between content and style domains, with consistent quality across diverse artistic styles and facial features. These visualizations complement the main experimental results by providing a more comprehensive view of our method's capabilities in handling various artistic references and content images.



Figure 15: More visualization results showing style transfer examples with different content and style image pairs.

I.2 More Comparison Results

Owing to space constraints in the main paper, we split the qualitative comparisons across two figures. Figure 16 compares our method against recent DiT-based editing models, while Figure 17 reports comparisons against diffusion-based and conventional style transfer baselines. Together, these results further demonstrate the superiority of our approach in balancing faithful style transfer with identity preservation.

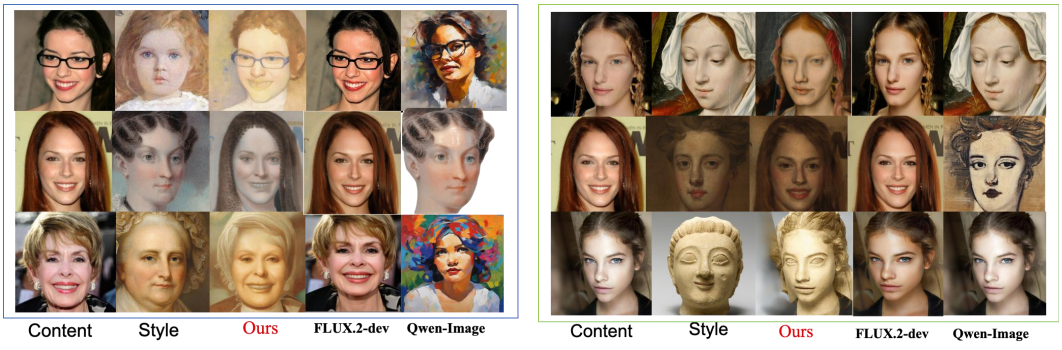


Figure 16: Qualitative comparison of style transfer results against DiT-based editing models.

I.3 Zoom-in Visualization

To provide a detailed view of how our method handles fine-grained style transfer, we present zoom-in visualizations that highlight specific regions of interest. Figure 18 shows magnified portions of our stylization results, focusing on facial features, textures, and artistic elements. These detailed views demonstrate how our method successfully preserves important facial characteristics while incorporating artistic style elements at different scales.

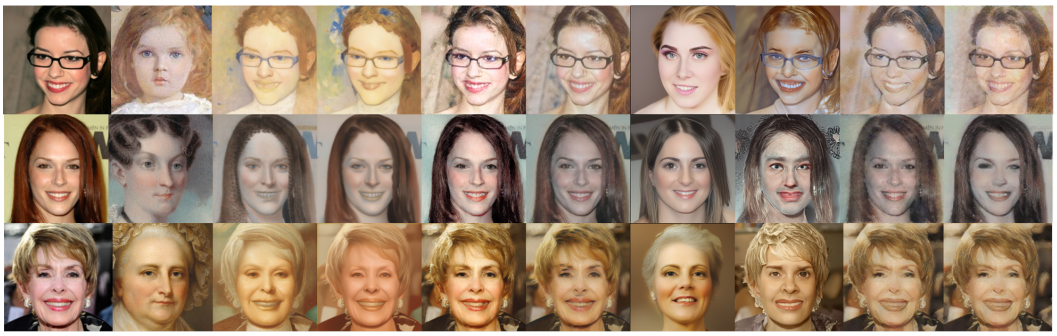


Figure 17: Qualitative comparison of style transfer results against diffusion-based and conventional baselines.

I.4 More Visualization of Artistic to Artistic Style Transfer

We provide additional visualization results of artistic-to-artistic style transfer to demonstrate our method's capability in handling diverse artistic styles. As shown in Figure 19, our method successfully transfers styles between different artistic portraits while preserving the key characteristics of both content and style images.

I.5 More Visualization of Non-face Artistic to Realistic Style Transfer

We provide additional visualization results of style transfer from non-facial artistic references to realistic face images. As shown in Figure 20, our method successfully extracts and applies artistic elements from diverse non-portrait artworks while preserving facial identity and human features.

I.6 Style Transfer Results Using Different Stable Diffusion Model Versions

To evaluate the generalization capability and robustness of our proposed method, we conduct comprehensive experiments using multiple versions of Stable Diffusion models, specifically SD v1.5, v2.0, and v2.1. Figure 22 and Figure 21 present a comparative analysis of style transfer results across these model variants. Our empirical findings demonstrate that the method exhibits consistent performance characteristics across different model architectures, maintaining robust facial identity preservation, semantic content fidelity, and effective artistic style element extraction and application. The consistent performance across model versions validates the architecture-agnostic nature of our approach, showing superior results compared to baseline methods, particularly StyleID. This characteristic is particularly valuable as it ensures reliable style transfer capabilities regardless of the underlying diffusion model implementation, contributing to the method's practical applicability and deployment flexibility.

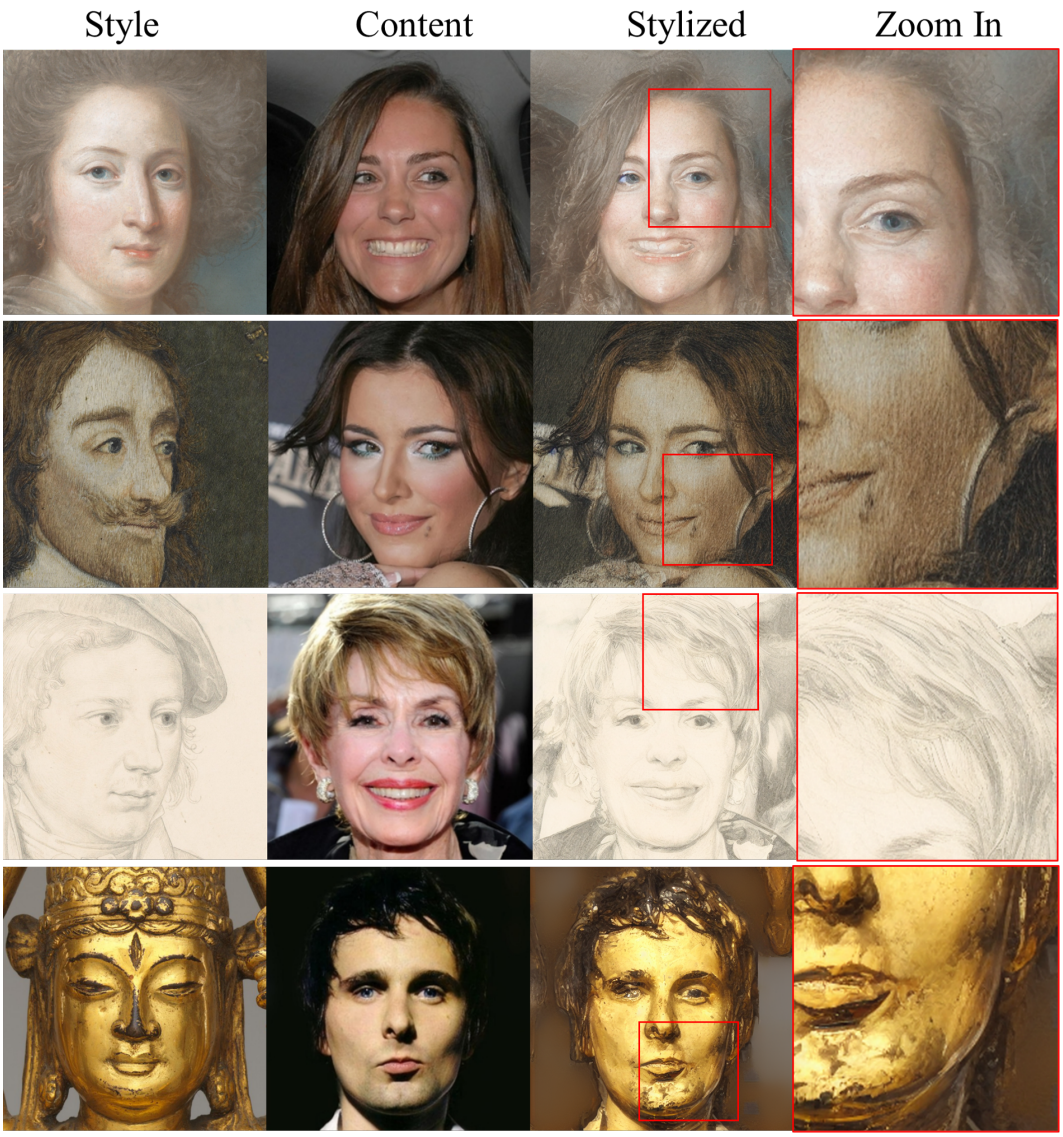


Figure 18: Zoom-in visualization comparing original and stylized image regions. The detailed views highlight our method's ability to preserve facial features while incorporating artistic elements.



Figure 19: Additional artistic-to-artistic style transfer results. Each row shows a different example with content image, style reference, and our stylization result.



Figure 20: Additional style transfer results using non-facial artistic references. Each row shows a different example with content image, non-facial style reference, and our stylization result.

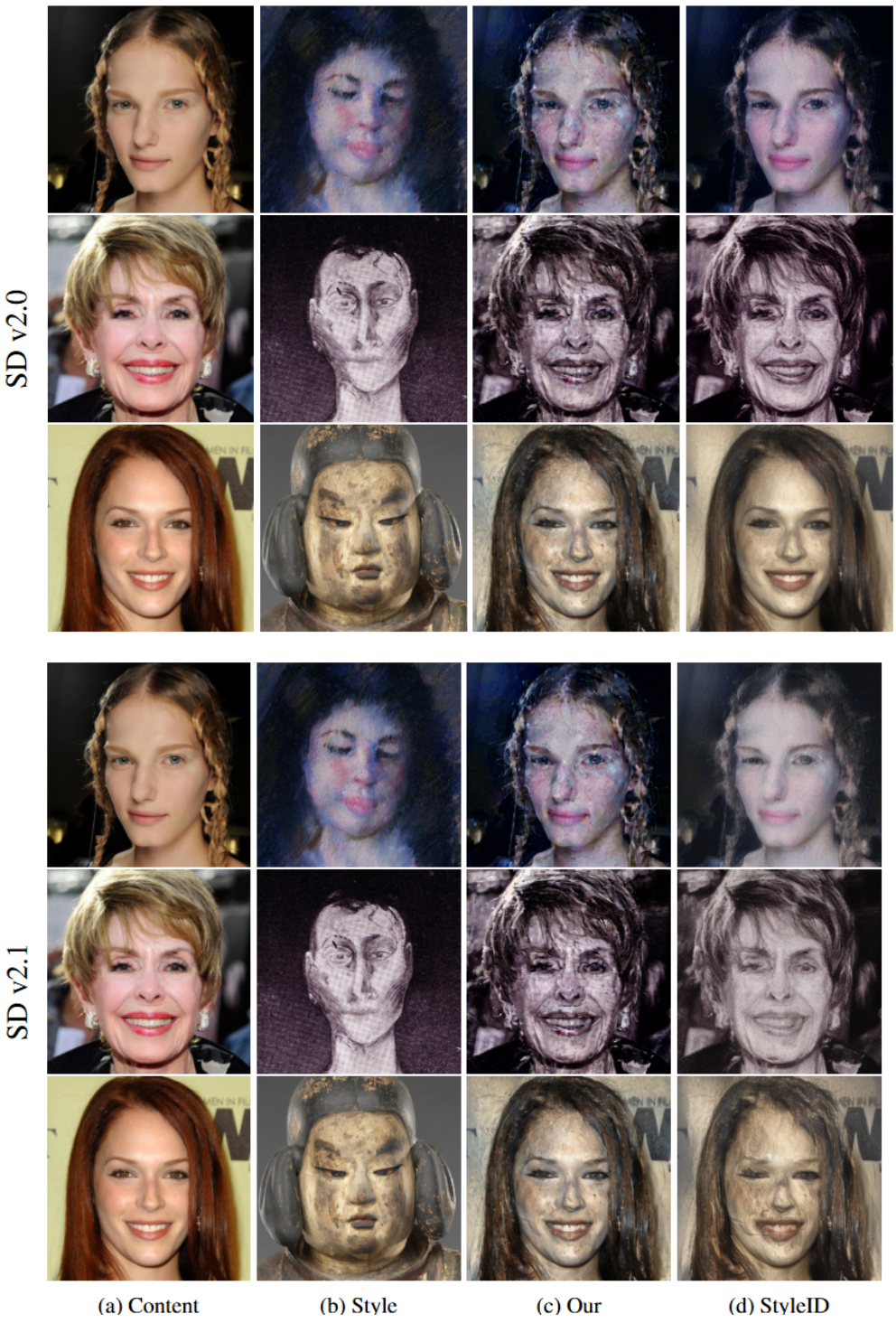


Figure 21: Style transfer results using different versions of Stable Diffusion models.

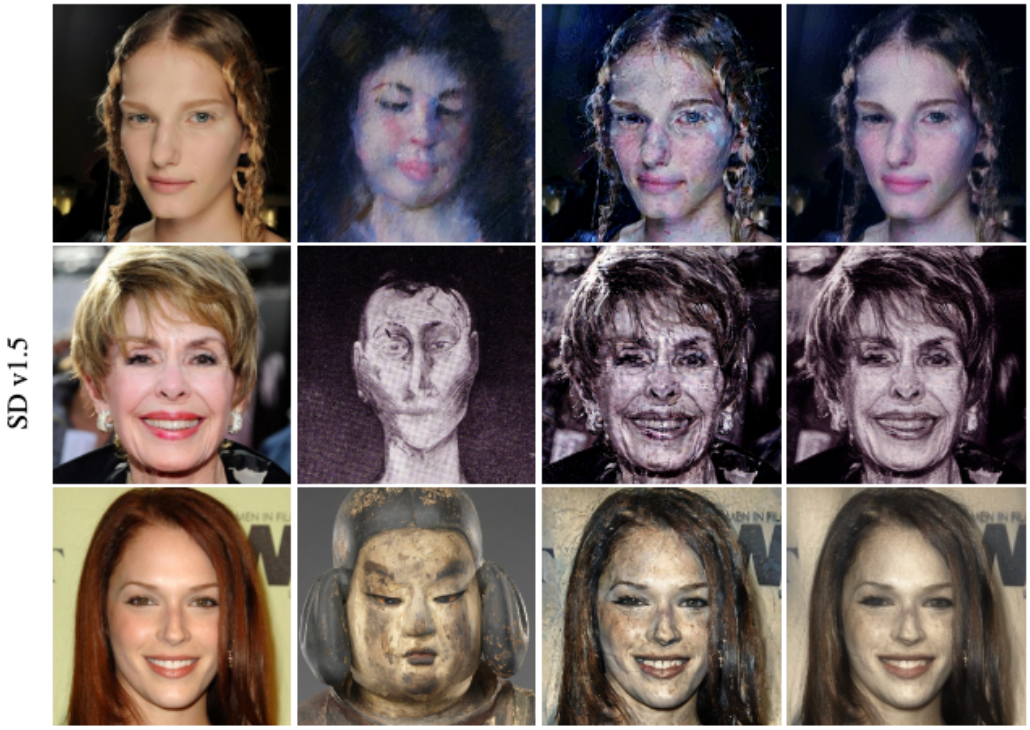


Figure 22: Style transfer results using different versions of Stable Diffusion models.