StyleRoute: Diffusion Style Transfer via Regional Routing and Conflict-aware Projection

Xunhao Lin[1,*] **Xinpeng Ding**[1,*] **Yu Sun**[1] **Fei Gao**[1,†]
Maoying Qiao[2] **Lihuo He**[1] **Nannan Wang**[1,†]

[1]Xidian University, Xi'an, China
[2]University of Technology Sydney, Sydney, Australia

Abstract

Diffusion-based style transfer faces two fundamental questions: which regions should prioritize structure preservation and which are more amenable to texture transfer (the *where* question), and how to achieve such region-aware stylization without corrupting content (the *how* question). Existing methods typically rely on a single mechanism to control local structure, local texture, and global style together, making them fragile in region-rich scenes: uniform stylization strength may distort salient objects, while conservative control leaves texture-dominant regions under-stylized. We propose StyleRoute, a region-aware diffusion stylization framework that explicitly separates *where* style should be transferred from *how* the transfer should be optimized. For the *where* question, StyleRoute infers a region-level routing variable from structure confidence, texture confidence, and route disagreement, deciding per region whether to protect structure or to inject texture. For the *how* question, StyleRoute coordinates the regional routing with diffusion optimization to guide global style evolution, suppress style-content conflicts, and stabilize coarse layout. *Together, StyleRoute transforms the implicit style-content trade-off into an explicit, interpretable, and region-aware transfer process.* Experiments on diverse content-style pairs show consistent improvements over prior diffusion-based and region-matching stylization methods. Project page: `https://lxihan.github.io/StyleRoute/`.

1 Introduction

Artistic style transfer aims to render a content image in the visual language of a reference style while preserving its semantic identity [1, 2, 3, 4]. In diffusion-based stylization, this goal becomes a *where-to-transfer* control problem rather than a simple content–style balancing task [5, 6, 7, 8]. Faces, animals, and building edges are sensitive to small geometric distortions, whereas foliage, fabric, and sky often require stronger appearance changes to appear stylized. The central question is therefore not only *how* to match content and style, but also *where structure should be protected and where texture should be transferred*, while keeping these local decisions compatible with a coherent global style.

Existing methods mostly improve the matching or injection mechanism, but they often leave these decisions coupled. Classical approaches align feature statistics [1, 2, 3]; attention-based models improve local correspondence [9, 4]; and diffusion methods manipulate latent features, self-attention, or semantic matches [5, 6, 7, 8]. Despite their differences, a single transfer strength is still expected to handle all regions uniformly, which leads to a clear failure: a strength that works for textured

*Equal contribution.
†Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: Motivation. A single implicit stylization strength is brittle: conservative transfer weakens painterly cues, while stronger transfer can damage fine structure.

backgrounds often distorts faces or object boundaries, while a conservative setting that protects structure leaves texture-rich areas looking under-stylized, as shown in Figure 1.

To address this challenge, we propose StyleRoute, a diffusion stylization framework that explicitly separates *where* to transfer style from *how* to optimize the transfer. For *where*, StyleRoute reformulates stylization as a region-level routing problem. It decomposes stylization into a structure carrier and a texture carrier, and uses a region-wise stability gate to adaptively balance these two carriers according to local structural sensitivity and texture tolerance. When the two carriers conflict, the gate conservatively prioritizes structure preservation, since structural distortions are typically harder to recover than insufficient stylization.

For *how*, StyleRoute couples the above regional routing mechanism with global style evolution in diffusion. Through directional style alignment, sparse texture refinement, conflict-aware gradient projection, and a low-frequency layout anchor, StyleRoute encourages coherent image-level stylization while preventing style optimization from overwriting locally protected content structure. *As a result, the implicit style-content trade-off is transformed into an explicit, interpretable, and region-aware transfer process.*

This design is both effective and easy to analyze. The routing weights and projected gradients are explicit, so we can examine exactly what each component contributes. Extensive experiments show that StyleRoute consistently outperforms existing methods, especially the closest region-matching baseline StyleGallery [8], achieving better structure preservation and stronger stylization at the same time. Our main contributions are three-fold:

- We formulate diffusion-based stylization as a regional routing problem, replacing a single uniform transfer strength with explicit structure and texture carriers.
- We introduce a stability-guided router that answers *where* to transfer, and a conflict-aware projection that answers *how* to transfer without hurting content structure.
- Experiments show that StyleRoute achieves a more favorable style–content trade-off than prior methods, improving across multiple evaluation metrics.

2 Related Work

2.1 Style Transfer and Diffusion-Based Stylization

Neural style transfer began with optimization-based matching of pretrained content features and style statistics [1, 10], but per-image optimization is slow and global statistics treat all regions uniformly. Feed-forward methods improved efficiency through perceptual losses and feature normalization [2, 3], while later methods used self-similarity, internal–external learning, contrastive objectives, and aesthetic pattern modeling for stronger local transfer [11, 12, 13]. Attention-based models such as AdaAttN [9] and StyTr^2 [4] further improved local correspondence with attention-weighted style statistics, yet they do not separate structure-sensitive regions from texture-tolerant ones. Diffusion methods expose disentangled style–content representations, inversion paths, h-space features, latent features, and attention maps for layout preservation, reference injection, semantic alignment, or training-free attention control [14, 15, 16, 5, 17, 6, 7, 18, 19]. Recent controllable stylization also studies structure–appearance separation [20]. Most methods still apply one transfer rule across the image; we instead route stylization at the region level.

The closest baselines are diffusion-based and semantic-aware stylization methods. Plug-and-play diffusion features [5] preserve spatial semantics through feature injection, but uses fixed injection strength. StyleID [6] exposes the content–style trade-off through query preservation and attention temperature scaling, yet its temperature remains global. InstantStyle [7] uses lightweight reference conditioning from image-prompt adapters [21], but does not decide which regions receive structure or texture transfer. SCSA [22] and Semantix [23] improve region-level alignment through semantic sparse attention or energy-guided sampling. StyleGallery [8] is closest to our setting because it combines region segmentation, clustered matching, and regional style losses in a training-free semantic pipeline. These methods show the value of matching and semantic control, but to our knowledge, among these semantic stylization methods, none provides an explicit variable that routes each region between structure-preserving and texture-seeking transfer.

2.2 Routing, Balancing, and Conflict-Aware Optimization

Uncertainty-based weighting balances competing objectives through data-dependent allocation rather than fixed scalar weights [24]. In mixture-of-experts models, DSelect-k [25] studies smooth routing, and Switch Transformers [26] show that sparse expert selection scales efficiently. In multi-task learning, CAGrad [27] modifies updates under incompatible gradient directions, and PCGrad [28] projects conflicting gradients to reduce negative transfer. These lines suggest that a single weighted sum can hide useful structure when objectives interact. We apply these ideas selectively. The stability gate deterministically routes stylization mass between two carriers using confidence signals, rather than learning a separate expert. Conflict-aware projection likewise targets only the global style-content conflict, leaving the local routing branch untouched. This keeps each component simple and scoped to its specific role.

3 Method

3.1 Overview

Given a content image x_c and a style image x_s, the goal is to synthesize an output that preserves the structure of x_c while adopting the appearance of x_s. Our key idea is to treat style transfer as a problem of *where to transfer*: different regions of the content image have different capacities for accepting stylistic changes, and forcing uniform stylization often distorts salient objects while leaving texture-rich areas under-stylized. To address this, we introduce a region planning stage before diffusion. It segments the content image into N regions $\mathcal{R} = \{1, \ldots, N\}$, each associated with a token-space mask m_r and area weight a_r, and retrieves from the style image a structure donor d_r^{str} and, when available, a texture donor d_r^{tex} for each region. This planning stage answers an initial question: *which regions should prioritize structure preservation, and which are more amenable to texture transfer*?

During diffusion, we optimize a latent z initialized from the content latent z_c. For the *where* decision, local structure routing pushes each region toward its structure donor, preventing over-stylization in salient areas. For the *how* decision, directional global style alignment guides overall appearance along the content-to-style residual, sparse texture refinement injects patterns only where needed, and conflict-aware projection removes the component of the style gradient that opposes content preservation. A low-frequency appearance anchor stabilizes coarse tone and layout throughout. Figure 2 summarizes the framework: regional routing decides *where* to transfer, while conflict-aware projection controls *how* to respect content constraints.

3.2 Where to Transfer: Region-Level Dual-Carrier Routing

3.2.1 Static Region Planning

Before denoising, StyleRoute builds a fixed region plan for the content image and the style candidate pool. The region masks are generated by automatic segmentation; DINOv2 supplies semantic region descriptors for representation and matching, but does not generate masks. The content and style images are decomposed into automatic regions, and each region is represented by a token-space mask, area, center, semantic prototype, and texture prototype, following the use of semantic region evidence in recent controllable stylization pipelines [8, 29]. The planner scores every content–style

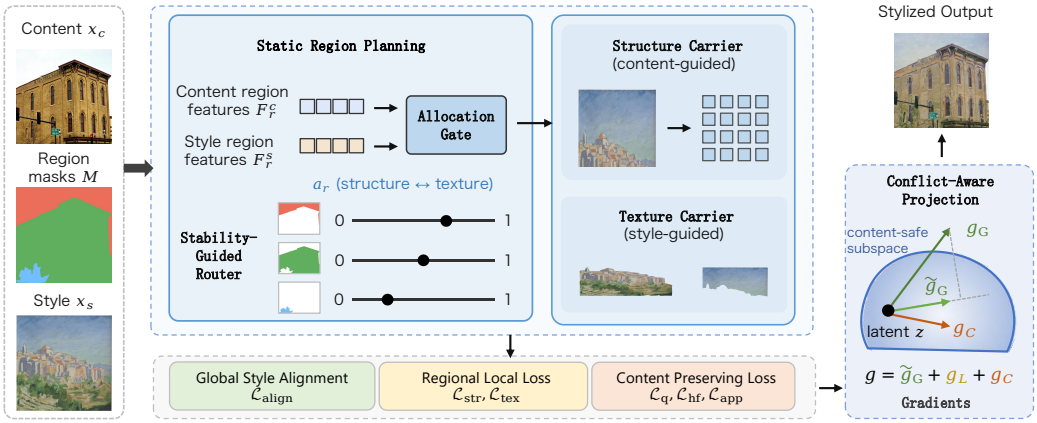


Figure 2: Method overview. The region-first stylization block answers *where* to transfer by estimating region-wise structure protection and texture eligibility. The global-control and diffusion-optimization blocks answer *how* to apply these regional decisions, using style alignment, conflict-aware projection, and appearance anchoring to preserve content structure.

region pair twice: a structure score favors semantic and spatial compatibility, while a texture score favors reliable local texture transfer. The best role-compatible semantic match becomes the structure donor d_r^{str}, and a texture donor d_r^{tex} is kept only when its texture score passes a routing threshold. The resulting masks, donors, confidence scores, and route disagreement values are fixed during diffusion and provide the inputs to the stability gate. Appendix A.1 gives the scoring definitions and implementation details.

3.2.2 Stability-Guided Dual-Carrier Routing

The main local decision is how much each region should rely on structure-preserving donor transfer rather than texture refinement. A uniform routing rule would treat a face and a patch of sky in the same way, although their tolerance to stylization is very different. Structural errors on salient objects are difficult to undo later, while an under-stylized sky can still absorb additional texture. We therefore define a region-specific stability gate g_r that favors structure when the evidence is ambiguous, using the same broad idea that data-dependent allocation can be preferable to a fixed scalar weight [24, 25, 26]:

$$g_r = \text{clip}_{[g_{\min}, g_{\max}]}\left(\eta_0 + \eta_s(\kappa_r^{(s)} - \kappa_r^{(t)}) + \eta_\delta \delta_r + b_r\right), \tag{1}$$

where η_0 is a base gate bias, $\kappa_r^{(s)}$ and $\kappa_r^{(t)}$ are structure and texture confidence, δ_r is route disagreement, b_r is a role prior for foreground-like regions, and $0 < g_{\min} < g_{\max} < 1$. The nonnegative coefficients η_s and η_δ encode two monotonic design constraints: stronger relative structure evidence and stronger route disagreement should not reduce structure protection. This conservative bias is useful because texture can still be added later, whereas geometric errors are hard to undo. The clipping interval $[g_{\min}, g_{\max}]$ prevents degenerate routing: every region keeps some structure protection, and no region is forced into pure structure transfer. The default gate uses $(\eta_0, \eta_s, \eta_\delta, g_{\min}, g_{\max}) = (0.48, 0.18, 0.22, 0.20, 0.85)$, with $b_r = 0.10$ for foreground-like regions and $b_r = -0.05$ otherwise.

The gate weights a mask-aware mixture of global style and region-specific donor responses. Let $o_{g,l}$ be the global style output at layer l, $o_{r,l}$ the donor-induced output for region r, and w_{bg} a background style-preservation weight. The structure target is

$$\hat{o}_l = \frac{w_{\text{bg}}\, o_{g,l} + \sum_{r\in\mathcal{R}} m_r(a_r g_r)\, o_{r,l}}{w_{\text{bg}} + \sum_{r\in\mathcal{R}} m_r(a_r g_r) + \varepsilon}, \qquad \mathcal{L}_{\text{str}} = \sum_l \|o_l(z) - \hat{o}_l\|_1. \tag{2}$$

Regions with large g_r are tied more strongly to structure donors, and the complementary weight $1 - g_r$ is used in the texture branch. We set $w_{\text{bg}} = 0.35$ for all experiments.

3.3 How to Optimize: Conflict-Aware Diffusion Stylization

3.3.1 Directional Global Style Alignment

Global alignment controls the image-level style direction, while gated sparse texture refinement controls local texture injection. Direct image-level style matching can pull the output toward a mismatched global style state even when each region receives a plausible donor, because locally reasonable assignments do not guarantee a coherent global style. We therefore measure style evolution relative to the content image and guide the diffusion trajectory along a controlled direction instead of forcing it to copy the style state exactly.

For each style-alignment layer $l \in \mathcal{S}_{\text{align}}$, a descriptor $\phi_l(\cdot)$ is constructed by spatially standardizing the attention output, applying mild smoothing, and pooling to a fixed grid. Let z_s be the VAE latent of the selected global style image, and let D denote the number of scalar entries in the flattened descriptor. The current and style residuals are

$$c_l = \phi_l(o_l(z)) - \phi_l(o_l(z_c)), \qquad s_l = \phi_l(o_l(z_s)) - \phi_l(o_l(z_c)). \tag{3}$$

Rather than forcing $c_l = s_l$, the current residual is decomposed into a component parallel to the style direction and an orthogonal residual:

$$u_l = \frac{s_l}{\|s_l\|_2 + \varepsilon}, \qquad p_l = \langle c_l, u_l \rangle, \qquad c_l^{\perp} = c_l - p_l u_l. \tag{4}$$

The style-alignment loss is

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{S}_{\text{align}}|} \sum_{l \in \mathcal{S}_{\text{align}}} \left[\frac{|p_l - \rho \|s_l\|_2|}{\|s_l\|_2 + \varepsilon} + \lambda_{\perp} \frac{\|c_l^{\perp}\|_1}{D(\|s_l\|_2 + \varepsilon)} \right]. \tag{5}$$

The parallel term moves the current image along the style direction to a target magnitude set by ρ, while the orthogonal term suppresses drift in unsupported directions. We use $\rho = 0.65$ and $\lambda_{\perp} = 0.35$, yielding a controlled global style trajectory instead of direct style copying.

3.3.2 Sparse Texture Refinement

Texture transfer is restricted in both space and time. Dense texture injection across all regions often damages semantic anchors. Adding brushstroke patterns to a face, for example, can harm identity even when the local texture match is reasonable. We therefore activate texture refinement only for regions with reliable texture evidence, and only for the top-K regions ranked by the texture budget score:

$$b_r^{\text{tex}} = (1 - g_r)\kappa_r^{(t)} \sqrt{a_r} \max(0, 1 - \eta_{\delta}^{\text{tex}} \delta_r), \qquad K = \min\big(K_{\max}, \lceil \sqrt{N_{\text{tex}}} \rceil\big), \tag{6}$$

where N_{tex} is the number of regions with texture donors. This score favors regions with strong texture evidence, sufficient support, and low route disagreement; we use $\eta_{\delta}^{\text{tex}} = 0.5$ and $K_{\max} = 3$ to keep the branch sparse even in region-rich images. For each active region, high-pass filtered query features from texture layers $\mathcal{S}_{\text{tex}}$ are compared inside the masked region through a grid descriptor ψ_l:

$$\mathcal{L}_{\text{tex}} = \frac{1}{|\mathcal{A}|} \sum_{r \in \mathcal{A}} (1 - g_r) \frac{1}{|\mathcal{S}_{\text{tex}}|} \sum_{l \in \mathcal{S}_{\text{tex}}} \big\| \psi_l(q_l(z), m_r) - \psi_l(q_l(d_r^{\text{tex}}), m_r) \big\|_1. \tag{7}$$

Let $\pi_t = (t+1)/T$ denote normalized diffusion progress. The texture schedule is

$$\gamma_t = \mathbf{1}\{\pi_t > \tau_{\text{tex}}\} \left[1 + \beta_{\text{tex}} \operatorname{clip}_{[0,1]} \left(\frac{\pi_t - \tau_{\text{tex}}}{1 - \tau_{\text{tex}}} \right) \right]. \tag{8}$$

We set $(\tau_{\text{tex}}, \beta_{\text{tex}}) = (0.55, 0.35)$, so γ_t keeps this branch inactive in early diffusion steps and raises its weight only after the global layout has largely stabilized.

3.3.3 Conflict-Aware Projection with Appearance Anchoring

Appearance Anchoring. The content branch preserves content features through query matching and high-frequency query matching. Let $q_l^c = q_l(z_c)$ be the content query feature at layer l, $\text{sg}(\cdot)$

stop gradients through the reference feature, and $\mathcal{H}_{\sigma_q}$ a token-grid high-pass filter. We use

$$\begin{aligned}\mathcal{L}_q &= \sum_l \left\| q_l(z) - \mathrm{sg}(q_l^c) \right\|_1, \\ \mathcal{L}_{\mathrm{hf}} &= \sum_l \left\| \mathcal{H}_{\sigma_q}(q_l(z)) - \mathrm{sg}\big(\mathcal{H}_{\sigma_q}(q_l^c)\big) \right\|_1.\end{aligned} \tag{9}$$

We also add a low-frequency appearance anchor to stabilize coarse tone:

$$\mathcal{L}_{\mathrm{app}} = \left\| \mathcal{G}_\sigma\big(\mathcal{V}(z)\big) - \mathcal{G}_\sigma(x_c) \right\|_1, \tag{10}$$

where $\mathcal{V}$ is the differentiable VAE decoder and $\mathcal{G}_\sigma$ is Gaussian blur. We use $\sigma = 5.0$; because the blur removes high-frequency details, this term constrains low-frequency appearance without suppressing fine stylization. In practice, it mainly prevents slow drift in overall tone and layout.

Conflict-Aware Projected Update. Based on the previously defined individual losses, the unprojected scalar combination is

$$\mathcal{L} = \alpha_t \mathcal{L}_{\mathrm{align}} + \mathcal{L}_{\mathrm{str}} + \lambda_{\mathrm{tex}} \gamma_t \mathcal{L}_{\mathrm{tex}} + \lambda_c \left(\mathcal{L}_q + \lambda_{\mathrm{hf}} \mathcal{L}_{\mathrm{hf}} \right) + \lambda_{\mathrm{app}} \mathcal{L}_{\mathrm{app}}, \tag{11}$$

where $\alpha_t = 0.10 + 0.45(1 - \pi_t)^2$ decays quadratically to emphasize global guidance early, and γ_t activates texture refinement after a fraction τ_{tex} of the diffusion trajectory. The loss weights are fixed to $\lambda_{\mathrm{tex}} = 0.12$, $\lambda_c = 0.25$, $\lambda_{\mathrm{hf}} = 0.10$, and $\lambda_{\mathrm{app}} = 0.08$. The actual update is not the plain gradient of this scalar combination, because the global style term can conflict with content preservation. These losses guide per-image latent optimization during diffusion inference only; all pretrained diffusion-model parameters remain frozen, without training or fine-tuning.

The main optimization conflict appears between the global style-alignment term and content preservation. The style-alignment loss can improve style coherence while pushing the latent away from content-consistent solutions. A fixed scalar weight does not resolve this conflict reliably because the conflict direction changes across the latent space and over diffusion steps, which is the setting addressed by gradient-conflict methods in multi-objective optimization [27, 28]. We therefore decompose the latent gradient into three components:

$$\begin{aligned} g_G &= \nabla_z(\alpha_t \mathcal{L}_{\mathrm{align}}), \\ g_L &= \nabla_z(\mathcal{L}_{\mathrm{str}} + \lambda_{\mathrm{tex}} \gamma_t \mathcal{L}_{\mathrm{tex}}), \\ g_C &= \nabla_z\big(\lambda_c(\mathcal{L}_q + \lambda_{\mathrm{hf}} \mathcal{L}_{\mathrm{hf}}) + \lambda_{\mathrm{app}} \mathcal{L}_{\mathrm{app}}\big). \end{aligned} \tag{12}$$

When the global and content gradients conflict ($\langle g_G, g_C \rangle < 0$), we project g_G to suppress the component that opposes g_C:

$$\begin{aligned} \tilde{g}_G &= g_G - \frac{\langle g_G, g_C \rangle}{\|g_C\|_2^2 + \varepsilon} g_C, \\ g &= \tilde{g}_G + g_L + g_C. \end{aligned} \tag{13}$$

Here, only the global branch is projected. The local branch is already moderated by the region gate, whereas the global style-alignment term is the main source of image-level drift. Under a conflict, the projection suppresses the component of g_G along $-g_C$ up to the numerical ε, so the global style update no longer directly cancels the content-preserving branch before the local gradient is added.

4 Experiments

4.1 Experimental Setup

Datasets and Settings. Following prior style transfer protocols [6, 4, 30], we randomly sample 30 content images from MS-COCO [31] and FFHQ [32] and 30 style images from WikiArt [33]. The sampled images are paired into a fixed held-out evaluation set, and all methods use the same content–style pairs. The content pool contains both region-rich scenes and structure-sensitive faces, so the evaluation tests object layout, hard boundaries, facial geometry, and texture-heavy backgrounds.



Figure 3: Qualitative comparison with baselines. StyleRoute preserves structure-sensitive geometry across diverse scenes while retaining visible style texture. Best viewed zoomed in.

Table 1: Comparison on the held-out random content–style evaluation set. Best and second-best results are bold and underlined.

Method	ArtFID↓	FID↓	LPIPS↓	CFSD↓	B-Gram ($\times 10^{-3}$)↓
StyTR^2 [4]	35.30	23.69	0.4299	0.0608	<u>2.268</u>
StyleID [6]	33.34	20.14	0.5776	0.1683	3.424
CoCoDiff [30]	33.99	20.27	0.5984	0.1882	3.436
InstantStyle [7]	43.70	26.50	0.5890	0.2726	3.669
AD [36]	30.21	<u>19.49</u>	0.4743	0.0759	2.492
CSGO [35]	33.37	20.85	0.5269	0.1643	3.044
StyleGallery [8]	<u>28.90</u>	19.51	0.4090	0.0602	2.314
StyleGallery+SAM [8]	28.97	19.56	<u>0.4087</u>	<u>0.0601</u>	2.314
StyleRoute (Ours)	**28.29**	**19.11**	**0.4070**	**0.0551**	**2.190**

Implementation Details. We compare StyleRoute with eight baselines: StyleGallery [8], StyleGallery with its released SAM option enabled [34], StyTR^2 [4], StyleID [6], CoCoDiff [30], InstantStyle [7], CSGO [35], and Attention Distillation (AD) [36]. StyleGallery is the direct semantic region-matching reference; StyleGallery+SAM isolates mask access by changing only the released mask option. The remaining baselines cover transformer feed-forward transfer, explicit content–style balancing, correspondence-aware diffusion, SDXL image-prompt/adapters, and attention manipulation. All baselines use public code and recommended settings, including default diffusion backbones. For StyleRoute, latent optimization uses Adam [37] under a DDIM scheduler [38] on a single NVIDIA A6000 GPU. We use one *fixed* scalar configuration for all content–style pairs; key parameter settings are summarized in Appendix A. The effect of changing the routing parameter is analyzed later in the gate-sensitivity study.

Metrics. We report ArtFID [39], FID [40], LPIPS [41], CFSD [6], and block-Hungarian Gram distance (B-Gram). The computation of B-Gram is detailed in Appendix A.3. These metrics cover style-domain quality, content and structural preservation, and style-statistic alignment; lower is better for all. Reporting all five is important because gains on one side of the trade-off can degrade the other, as discussed in Section 2.

4.2 Comparison with SOTAs

Qualitative Comparison. Figure 3 supports the quantitative trends. On animals, clocks, and other structure-sensitive regions, StyleRoute preserves geometry more reliably than StyleID and StyleGallery while avoiding the overly smooth appearance often produced by AD. On texture-dominant foliage, walls, and sky, it still introduces visible brushstroke and surface patterns. The gain is clearest on mixed scenes: regional routing estimates where texture can be absorbed and where structure should stay anchored, keeping object geometry stable while producing textured backgrounds.

Quantitative Comparison. Table 1 reports the quantitative comparison against existing style transfer methods. StyleRoute achieves the best performance across all five metrics, including Art-FID, FID, LPIPS, CFSD, and B-Gram. The baselines exhibit two recurring failure modes: conservative methods (e.g., StyleGallery, AD) preserve content structure but leave weak style alignment, while aggressive methods (e.g., StyTR^2, InstantStyle) push stylization at the cost of content distortion. By explicitly allocating style strength per region and resolving gradient conflicts globally, StyleRoute avoids both failure modes and attains a more favorable style–content operating point.

We further compare against two region-aware variants of StyleGallery [8]: the released no-SAM version and the SAM-augmented version. Adding SAM to StyleGallery yields nearly identical performance across all metrics, indicating that simply injecting segmentation masks without a principled routing mechanism does not effectively exploit region-level information. In contrast, StyleRoute incorporates regional routing and conflict-aware projection, achieving consistent improvements over both variants. This gap demonstrates that region awareness alone is insufficient; the key lies in *how* region information guides the allocation of structure preservation and style transfer strengths.

An additional paired evaluation against StyleGallery, including confidence intervals and computational costs, is reported in Appendix B.1–B.2.

User Study. We conducted a perceptual study on 30 content–style groups with 50 participants. StyleRoute received 62.3% preference votes, outperforming StyleGallery (17.7%), AD (11.0%), and StyleID (9.0%).

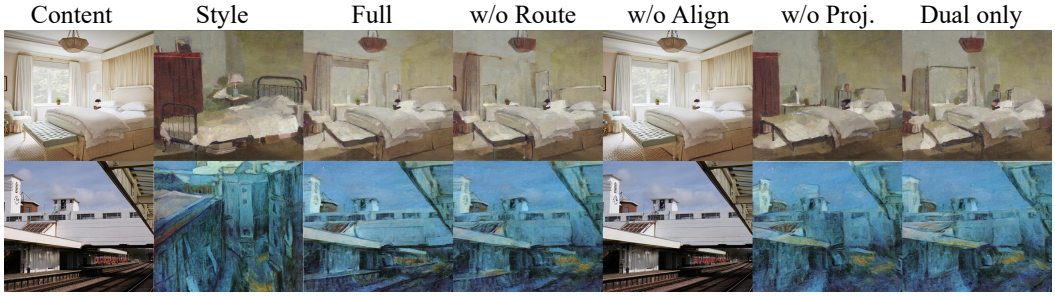


Figure 4: Qualitative ablation. Columns show content, style, full model, three removed-component variants, and dual-carrier-only transfer.

4.3 Model Analysis

Ablation Study. Figure 4 provides visual evidence for the component ablations. In the figure, *Content* and *Style* are the input content and reference style images, and *Full* is the complete StyleRoute model. *w/o Route* removes the stability-guided regional routing decision from the local branch; *w/o Align* removes the directional style-alignment loss $\mathcal{L}_{\text{align}}$; *w/o Proj.* keeps the losses but removes conflict-aware gradient projection; and *Dual only* keeps the dual-carrier local transfer branch without the global controls. The removed-component variants show that the ablations move along the style–content trade-off rather than only changing the apparent strength of stylization. This supports the interpretation that explicit regional decision points help separate structure preservation from texture transfer.

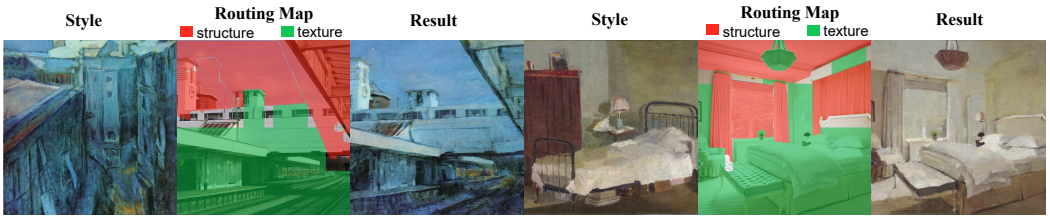


Figure 5: Routing maps from the stability gate: red denotes structure-protected regions and green denotes texture-eligible regions, revealing spatially selective transfer rather than a single global stylization strength.

Routing Maps. Figure 5 visualizes the gate-induced where-to-transfer allocation, showing that protected structure and texture-eligible regions are separated spatially rather than controlled by one global stylization strength.

Table 2: Ablation and gate diagnostics. Left: ablation study on routing and global controls; B-Gram is reported in 10^{-3} and lower is better. Right: gate-intercept diagnostic on a representative room case with a fixed cached region plan; increasing η_0 raises the structure gate and area-weighted structure mass $\sum_r a_r g_r$ while reducing the complementary texture budget $\sum_r b_r^{tex}$.

Ablation operating points

Config.	LPIPS↓	CFSD↓	B-Gram↓
StyleRoute	<u>0.4070</u>	**0.0551**	**2.19**
w/o routing	0.5809	0.0925	2.31
w/o style align	**0.3805**	<u>0.0847</u>	3.55
w/o projection	0.6829	0.1221	<u>2.30</u>
Dual-carrier	0.6240	0.1409	2.30

Gate-intercept diagnostic

η_0	mean g	FG	BG	Struct.	Tex.
0.36	0.389	0.465	0.313	0.354	1.312
0.42	0.449	0.525	0.373	0.411	1.185
0.48	0.509	0.585	0.433	0.467	1.058
0.54	0.569	0.645	0.493	0.524	0.931
0.60	0.629	0.705	0.553	0.581	0.804

Ablation Operating Points. The ablations reproduce the same two failure modes observed among the completed baselines. Removing style alignment collapses the optimization toward content reconstruction: content distances become small, but style-domain quality and style statistics degrade. Removing routing, projection, or the global controls moves in the opposite direction: style-domain metrics become more aggressive, but content and structural distances worsen. Mechanistically, style alignment keeps global style evolution active, routing decides which regions can absorb texture, and conflict-aware projection is designed to suppress the component of the global style gradient that opposes content preservation. The full model is therefore not designed to minimize one isolated metric; it is designed to mitigate both under-stylization and structure damage, which is the operating point supported by Table 2.

Budget-matched Uniform and Shuffled routing controls are reported in Appendix B.3 to isolate the effect of spatial allocation.

Gate Sensitivity. Table 2 isolates the intercept of the Stability Gate while keeping the cached region plan fixed. Increasing η_0 translates both foreground and background gates upward, which increases the area-weighted structure mass and reduces the complementary texture budget. The foreground–background gap remains nearly constant because it is controlled by the role prior rather than by η_0. This diagnostic shows that the gate coefficients act as interpretable routing controls with predictable structure–texture semantics, rather than as brittle hand-coded constants.

5 Conclusion

This paper presents StyleRoute for diffusion-based style transfer through regional routing and conflict-aware projection. Different regions need different amounts of structure protection and texture transfer, while global style updates must not cancel content preservation. StyleRoute implements this idea through stability-guided regional routing, directional style alignment, conflict-aware projection, and a low-frequency appearance anchor. Across the compared baselines under our evaluation protocol, StyleRoute improves the measured style-domain and content-preservation metrics

over the closest semantic matching baseline. This suggests that regional routing and projected global updates can improve the measured operating point, rather than merely shifting a fixed style–content weight. Despite such achievements, the current gate relies on hand-crafted confidence signals, and learned routing may adapt better to unfamiliar scene layouts. Besides, donor retrieval is still limited by the style candidate pool, so better retrieval could improve both structure and texture donors.

Performance depends on segmentation and donor quality, and the evaluated non-painting categories show mixed content–style trade-offs (Appendix B.4). The method incurs additional inference-time cost, and its compatibility with DiT/MMDiT backbones remains unvalidated.

Acknowledgements

This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project under Grant 2025ZD0123601, in part by the National Natural Science Foundation of China under Grants U22A2096, 62571395 and 62276203, in part by Scientific and Technological Innovation Teams in Shaanxi Province under Grant 2025RS-CXTD-011, in part by the Shaanxi Province Core Technology Research and Development Project under Grant 2024QY2-GJHX-11, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2026SYS-SYSZD-105, in part by the Key Research and Development Program of Shaanxi under Grant 2026JC-ZDGG-42, in part by the Establishment Fund of the State Key Laboratory of Wakefullness/Sleep and Cognition (Chongqing Institute for Brain and Intelligence), in part by the Fundamental Research Funds for the Central Universities under Grant QTZX26138.

References

[1] 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*, pages 2414–2423, 2016.

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

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

[4] 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*, pages 11326–11336, 2022.

[5] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1921–1930, 2023.

[6] 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*, pages 8795–8805, 2024.

[7] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024.

[8] Boyu He, Yunfan Ye, Chang Liu, Weishang Wu, Fang Liu, and Zhiping Cai. Stylegallery: Training-free and semantic-aware personalized style transfer from arbitrary image references. *arXiv preprint arXiv:2603.10354*, 2026.

[9] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6649–6658, 2021.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[11] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10051–10060, 2019.

[12] Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34:26561–26573, 2021.

[13] 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*, pages 22758–22767, 2023.

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

[15] 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*, pages 10146–10156, 2023.

[16] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Training-free style transfer emerges from h-space in diffusion models. *arXiv preprint arXiv:2303.15403*, 3(1):2, 2023.

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

[18] Bo Huang, Wenlun Xu, Qizhuo Han, Haodong Jing, and Ying Li. Attenst: A training-free attention-driven style transfer framework with pre-trained diffusion models. *arXiv preprint arXiv:2503.07307*, 2025.

[19] 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*, pages 18260–18269, 2025.

[20] Ruixiang Jiang and Chang Wen Chen. Diffartist: Towards structure and appearance controllable image stylization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9598–9607, 2025.

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

[22] Chunnan Shang, Zhizhong Wang, Hongwei Wang, and Xiangming Meng. Scsa: A plug-and-play semantic continuous-sparse attention for arbitrary semantic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13051–13060, 2025.

[23] Huiang He, Minghui Hu, Chuanxia Zheng, Chaoyue Wang, and Tat-Jen Cham. Semantix: An energy guided sampler for semantic style transfer. *arXiv preprint arXiv:2503.22344*, 2025.

[24] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[25] Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. *Advances in Neural Information Processing Systems*, 34:29335–29347, 2021.

[26] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[27] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021.

[28] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020.

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[30] Wenbo Nie, Zixiang Li, Renshuai Tao, Bin Wu, Yunchao Wei, and Yao Zhao. Cocodiff: Correspondence-consistent diffusion model for fine-grained style transfer. *arXiv preprint arXiv:2602.14464*, 2026.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[32] 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*, pages 4401–4410, 2019.

[33] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013.

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

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

[36] Yang Zhou, Xu Gao, Zichong Chen, and Hui Huang. Attention distillation: A unified approach to visual characteristics transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18270–18280, 2025.

[37] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

[40] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[41] 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*, pages 586–595, 2018.

A Implementation Details

A.1 Region Planning

The static region plan uses automatic masks. For each region, its semantic descriptor is the normalized masked mean of DINOv2 tokens, and its texture descriptor concatenates the masked mean and standard deviation of U-Net features. The area a_r and center μ_r are computed from the image-space mask. For content region r and style region j of candidate i, the compatibility terms are

$$\begin{aligned} D_{rij} &= \langle p_r^{\mathrm{sem}}, p_{ij}^{\mathrm{sem}} \rangle, & T_{rij} &= \langle p_r^{\mathrm{tex}}, p_{ij}^{\mathrm{tex}} \rangle, \\ P_{rij} &= \exp\!\left(-\frac{\|\mu_r - \mu_{ij}\|_2^2}{2\sigma_p^2} \right), & A_{rij} &= \exp\!\left(-\frac{|\log(a_r/(a_{ij}+\varepsilon))|}{\sigma_a} \right). \end{aligned} \tag{14}$$

After normalization over candidate regions, structure and texture compatibility scores are

$$\begin{aligned} s_{rij}^{\mathrm{str}} &= \omega_D D_{rij} + \omega_P P_{rij} + \omega_A A_{rij} + \omega_G \bar{s}_i^{\mathrm{cand}}, \\ s_{rij}^{\mathrm{tex}} &= \nu_T T_{rij} + \nu_D D_{rij} + \nu_A A_{rij} + \nu_G \bar{s}_i^{\mathrm{cand}}, \end{aligned} \tag{15}$$

where $\bar{s}_i^{\mathrm{cand}}$ is the normalized global candidate score. The structure donor is the highest-scoring role-compatible match. The texture donor is the highest-scoring candidate satisfying $s_{rij}^{\mathrm{tex}} \geq 0.14$; otherwise the texture branch is inactive for that region. The selected scores define structure/texture confidence. Route disagreement is the clipped complement of the semantic similarity between the selected donors (or 1 if no texture donor exists). All assignments are fixed before diffusion.

A.2 Key Parameters

Component	Setting used in the reported experiments
Diffusion optimization	Frozen U-Net; content-latent initialization; empty-prompt embedding; 150 DDIM steps; Adam learning rate 0.05; one NVIDIA A6000 GPU.
Routing gate	$(\eta_0, \eta_s, \eta_\delta) = (0.48, 0.18, 0.22)$; $[g_{\min}, g_{\max}] = [0.20, 0.85]$; foreground/background role priors $0.10/-0.05$.
Donor and structure	Texture donor threshold 0.14; tight/context reference blend 0.35; background support $w_{\mathrm{bg}} = 0.35$.
Global alignment	Middle and last self-attention layers; 8×8 descriptor; $(\rho, \lambda_\perp) = (0.65, 0.35)$.
Sparse texture	Top $K \leq 3$ regions; 4×4 descriptor; $\eta_\delta^{\mathrm{tex}} = 0.5$; $(\tau_{\mathrm{tex}}, \beta_{\mathrm{tex}}) = (0.55, 0.35)$.
Appearance / content	Query high-pass Gaussian $\sigma_q = 1.0$; appearance Gaussian $\sigma_{\mathrm{app}} = 5.0$.
Loss / projection	$(\lambda_{\mathrm{tex}}, \lambda_{\mathrm{c}}, \lambda_{\mathrm{hf}}, \lambda_{\mathrm{app}}) = (0.12, 0.25, 0.10, 0.08)$; $\varepsilon = 10^{-6}$.

Table 3: Implementation settings fixed across all evaluated content–style pairs.

A.3 B-Gram Evaluation Metric

For each VGG evaluation layer l, split output and style features into spatial blocks. With $F_l^b(x) \in \mathbb{R}^{C_l \times n_b}$, let $G_l^b(x) = F_l^b(x) F_l^b(x)^\top / (C_l n_b)$. A one-to-one Hungarian assignment $\pi_l^\star$ minimizes the sum of Frobenius distances between output and style block Gram matrices. The metric averages the matched distances across blocks and layers:

$$\text{B-Gram}(x, s) = \frac{1}{|\mathcal{L}_G|} \sum_{l \in \mathcal{L}_G} \frac{1}{|\mathcal{B}_l|} \sum_{b \in \mathcal{B}_l} \left\| G_l^b(x) - G_l^{\pi_l^\star(b)}(s) \right\|_F. \tag{16}$$

A.4 Perceptual Preference Study

For 30 fixed content–style groups, 50 participants compared anonymized, randomly ordered outputs from StyleRoute, StyleGallery, AD, and StyleID alongside the content and style references. They selected the result best balancing content preservation and style transfer; aggregate preferences are reported in Section 4.2.

B Additional Quantitative Evidence

B.1 Expanded Paired Evaluation

To assess the stability of the comparison with StyleGallery, we evaluate an additional 500 content–style pairs fixed before evaluation, using identical inputs for both methods and reporting results separately from Table 1. For the mean paired difference (StyleRoute minus StyleGallery), 95% CIs are the 2.5th and 97.5th percentiles of 10,000 bootstrap resamples of pair indices with replacement; they reflect variation across pairs, not repeated runs.

Metric ↓	Mean paired difference	95% CI
LPIPS	-0.0158	$[-0.0216, -0.0102]$
CFSD	-0.0091	$[-0.0109, -0.0074]$
B-Gram	-0.000108	$[-0.000124, -0.0000917]$

Table 4: Additional 500-pair comparison: negative differences favor StyleRoute.

All three intervals lie below zero; ArtFID improves slightly, while FID remains essentially unchanged.

B.2 Computational Cost

Under the same measurement protocol, StyleRoute increases per-pair runtime by approximately 8.7% and peak GPU memory by 71.8% relative to StyleGallery. StyleRoute uses 9.26 GB of peak GPU memory. This reflects additional inference-time computation and memory for explicit regional control.

B.3 Controlled Spatial Routing

We compare Full routing with Uniform and Shuffled controls. All variants use the same masks, donors, aggregate structure/texture budgets, initialization, losses, and schedules. Uniform removes spatially varying routing while preserving the total budgets; Shuffled reassigns the routing coefficients to different regions. This intervention isolates the spatial allocation of routing coefficients.

Regions with the highest structure-routing weights in the Full plan define the protected-region evaluation, and Full texture-active masks define the texture-region evaluation. Table 5 reports the percentage changes from the controlled evaluation.

Metric	Full vs. Uniform	Full vs. Shuffled
Protected-region SSIM ↑	$+1.54\%$	$+1.35\%$
Protected-region boundary discrepancy ↓	-2.39%	-2.16%
Protected-region pixel L1 ↓	-0.72%	-0.84%
Texture-region donor-Gram ↓	-0.37%	-0.24%
LPIPS-content ↓	-2.27%	-2.49%
CFSD ↓	-3.75%	-3.43%

Table 5: Reported routing-control changes (Full minus control). Positive SSIM changes and negative error changes favor Full.

The comparisons support the role of spatial allocation in protected-region structure preservation and texture-region consistency, beyond simply providing region masks or aggregate transfer budgets.

B.4 Style-Domain Breadth and Mask Perturbations

We evaluate five fixed mountain content images with four style references in each of five non-painting categories (20 pairs per category; 100 total). Both methods use the same 150-step setting, without output filtering or replacement. The limited content diversity constrains generalization beyond these evaluated scenes.

Category	LPIPS-content	CFSD	B-Gram
Anime/comic	-0.04679	-0.02112	$+3.04\times10^{-4}$
Graphic/poster	-0.02832	-0.03864	$+2.33\times10^{-4}$
Photographic grading	$+0.06029$	$+0.04449$	-3.88×10^{-4}
Pixel art	-0.02394	-0.10088	-2.41×10^{-4}
Sketch/engraving	$+0.04007$	$+0.01155$	-1.52×10^{-4}
All 100 pairs	$+0.000263$	-0.02092	-4.89×10^{-5}

Table 6: Non-painting evaluation: StyleRoute minus StyleGallery; negative values favor StyleRoute for all three metrics.

Results are mixed: photographic grading and sketch/engraving improve B-Gram but worsen LPIPS-content and CFSD, while anime/comic and graphic/poster show the opposite trade-off. Aggregate LPIPS is approximately tied. These findings do not establish uniformly better cross-domain performance.

For segmentation robustness, we test one-step erosion, one-step dilation, and deterministic two-pixel mask displacement while holding image pairs and optimization settings fixed. Across the reported LPIPS-content, CFSD, B-Gram, protected-region L1, and donor-Gram endpoints, the absolute relative changes are at most 0.32%. This supports stability under the tested mild geometric perturbations, but does not establish robustness to missing objects, severe segmentation errors, or incorrect donor assignments.

C Additional Qualitative Results

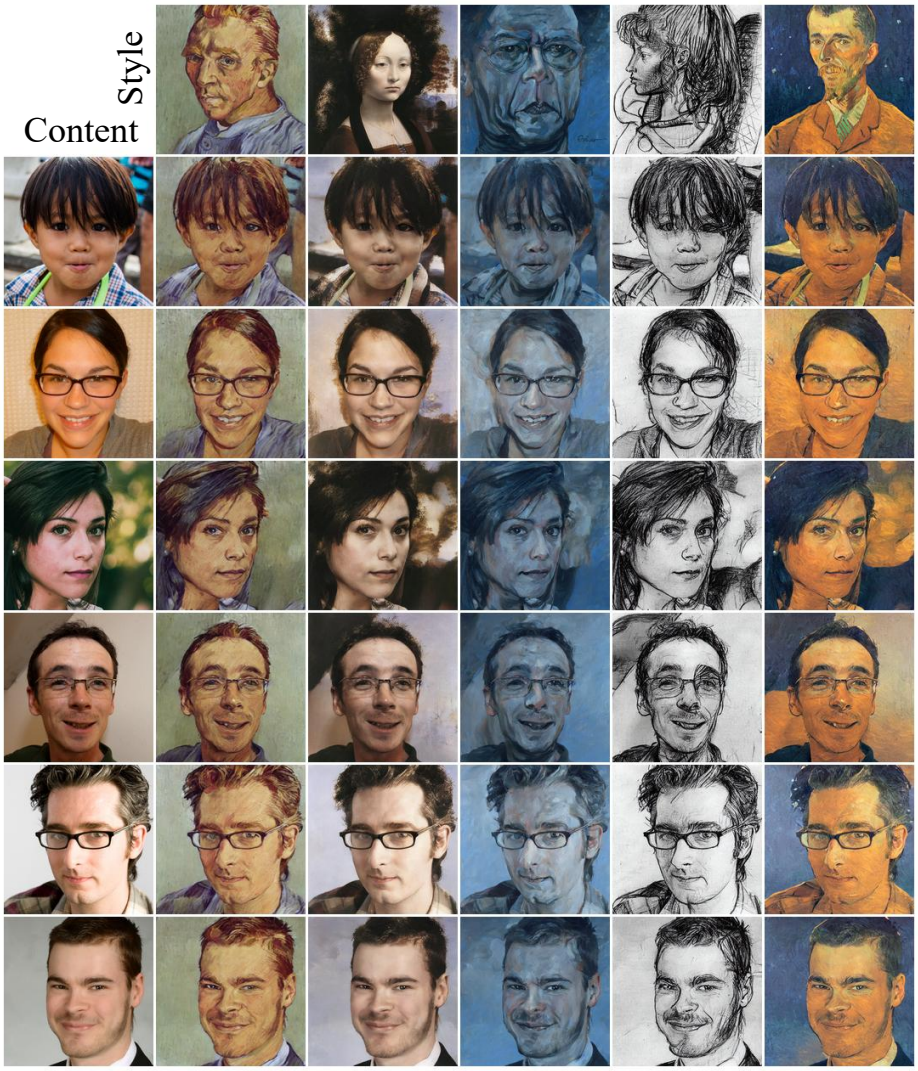


Figure 6: Additional style transfer results on portraits.



Figure 7: Additional style transfer results on buildings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the method scope as diffusion-based artistic style transfer with regional routing and conflict-aware projection. The claims are supported by the method description, quantitative comparison, ablations, and gate diagnostic in Sections 3 and 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.

- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion discusses hand-crafted routing signals, donor and segmentation quality, mixed cross-domain results, inference-time cost, and unvalidated DiT/MMDiT compatibility. Appendix B.4 reports the evaluated style-domain scope.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents algorithmic definitions and empirical evaluations, but does not claim new theoretical results requiring proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sections 3 and 4 and Appendix A describe the algorithm, latent optimization, evaluation protocol, and key implementation settings. Appendix B provides additional paired evaluations and controlled comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper uses public datasets and baseline implementations, but does not provide a public code link with complete reproduction scripts. The method and evaluation protocol are described in Sections 3 and 4 and Appendix A.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 and Appendix A report the datasets, baselines, optimizer, scheduler, and fixed hyperparameter settings. Appendix B describes the additional evaluation protocols.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix B.1 reports paired-bootstrap 95% confidence intervals for mean metric differences on 500 fixed content–style pairs, using 10,000 resamples. These intervals capture variability across evaluation pairs, not repeated-run or training variability.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4 and Appendix A.2 identify the GPU and optimization settings. Appendix B.2 reports 9.26 GB of peak GPU memory for StyleRoute and relative per-pair runtime and peak-memory increases over StyleGallery.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public image datasets and public baseline methods for artistic style transfer research. We are not aware of any aspect of the study that violates the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper focuses on the method and its empirical evaluation and does not include an explicit discussion of both positive and negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new dataset, pretrained image generator, or model weights. The evaluated framework uses existing models and standard benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper cites the datasets, models, and baseline methods used in the experiments, but does not explicitly document the licenses and terms of use for all assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new assets such as datasets, model weights, or code packages. The method is evaluated using existing models and standard experimental assets, which are cited in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Appendix A.4 summarizes the preference-study task, participant count, anonymized outputs, and randomized display order. Full participant instructions and compensation details are not provided.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: The paper does not report institutional review board approval or exemption, nor the information provided to participants concerning potential risks. No approval or exemption status is asserted here.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method does not use LLMs as an important, original, or non-standard component. Any use of language tools, if any, was limited to writing or formatting assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.