GRANULARITY-ALIGNED SELECTIVE CONDITIONING FOR IMAGE STYLE TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite recent advances in reference-based image style transfer, existing methods still struggle to transfer global appearance and local stylistic details while preserving content structure. We introduce **Granularity-Aligned Selective Conditioning (GASC)**, a framework that aligns style granularity with the distinct conditioning interfaces in a pretrained Diffusion–Mamba denoiser. GASC assigns shared image-level style conditioning to state-transition dynamics and token-varying style carriers to input writes. To construct these conditions, per-head orthogonal residualization reduces content interference in reference features, while content-conditioned carrier retrieval captures local style information. A separate content stream anchors spatial structure throughout denoising. Together, these designs coordinate global and local style conditioning with content preservation. Extensive experiments demonstrate that GASC outperforms existing methods in overall stylization quality while maintaining competitive content preservation.

1 INTRODUCTION

Arbitrary reference-based image style transfer aims to reproduce the appearance of a reference image while preserving the recognizable structure of a content image (Gatys et al., 2015; 2016). A key difficulty is that reference style is expressed at different granularities. Some appearance cues are shared across the image, whereas textures and stylistic details may vary across regions. Effective reference conditioning should therefore preserve these distinct granularities while controlling how they influence generation.

Existing methods span diverse architectures and conditioning mechanisms. S2WAT (Zhang et al., 2024) models short- and long-range dependencies with hierarchical strip-window attention, StyleID (Chung et al., 2024) injects reference style through self-attention features in a pretrained diffusion model, OmniStyle (Wang et al., 2025) adopts an end-to-end Diffusion Transformer, and SaMam (Liu et al., 2025) uses a style-aware Mamba encoder–decoder. However, representative examples in Fig. 1 reveal recurring trade-offs, with some outputs only partially capturing the shared reference appearance while others alter recognizable content structure under stylization.

These visually different failures motivate examining a common conditioning issue, namely that existing designs do not explicitly match the granularity of style representations to the modes of action of their receiving interfaces. This compatibility matters because conditioning can preserve or discard the variation carried by a representation. Pooling a token-varying condition into a shared signal removes its position-dependent variation, whereas broadcasting an image-level condition along the token axis provides no token-specific reference information. We refer to this discrepancy between representation granularity and conditioning-interface action as *granularity mismatch*.

Selective state-space models(SSMs) expose functionally distinct conditioning interfaces. In Mamba, Δ modulates state-transition dynamics, whereas B controls how inputs are written into the state (Gu & Dao, 2023). This distinction makes Mamba a useful architecture for examining how token-shared and token-varying style conditions should be assigned. To study this allocation within iterative image generation, we build on a pretrained DiM denoiser, which uses Mamba blocks as its backbone (Teng et al., 2024). The pretrained denoiser provides a reusable image-synthesis prior, while diffusion allows reference conditions to guide synthesis across successive denoising steps.

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.

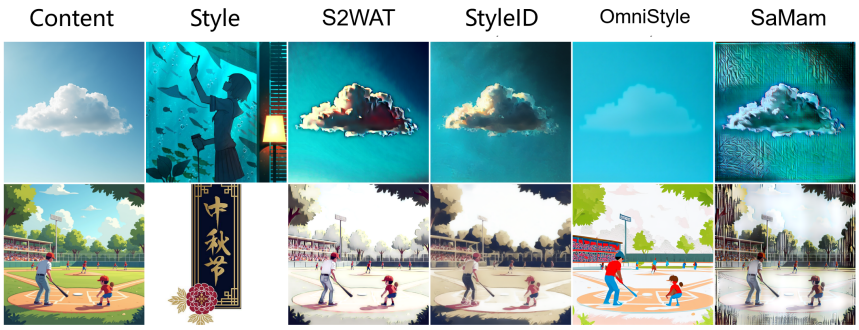

Figure 1: Representative examples motivating granularity-aware conditioning. The top row highlights differences in shared reference appearance transfer, while the bottom row highlights content-structure preservation under stylization.

We hypothesize that the granularity of a style condition should match the mode of action of the interface through which it enters the denoiser. We call this principle **granularity–interface alignment** (GIA). Guided by this principle, we introduce **Granularity-Aligned Selective Conditioning (GASC)**, a Diffusion–Mamba framework for reference-based image style transfer. GASC applies a token-shared style residual derived from an image-level descriptor to the Δ route and uses token-varying style carriers to condition input writes through B. To construct these conditions, Per-Head Orthogonal Residualization (HOD) suppresses content-dominant variation in reference features, while Content-Conditioned Group-Level Carrier Retrieval (C^2R) constructs content-adaptive carriers through retrieval and fusion. A separate pre-scan Content Stream injects spatially aligned content features into the hidden sequence to guide content structure.

Our contributions are summarized as follows:

- **Granularity mismatch in style conditioning.** We identify and formulate the mismatch between the granularity of style representations and the modes of action of their receiving conditioning interfaces, and propose granularity–interface alignment as a guiding principle.
- **Granularity-aligned Diffusion–Mamba framework.** GASC is the first Diffusion-Mamba framework for arbitrary reference-based image style transfer,, assigning image-level, token-shared style conditioning to Δ and token-varying style carriers to B.
- **Style-condition construction and content anchoring.** We introduce HOD to suppress content-related variation, C^2R to construct content-adaptive local carriers, and a pre-scan Content Stream to preserve recognizable content structure.
- **Superior empirical performance.** Extensive experiments show that GASC outperforms existing methods in overall stylization quality while maintaining competitive content preservation.

2 Related Work

2.1 Diffusion and SSMs

Diffusion models generate images through iterative denoising in pixel or latent space (Ho et al., 2020; Rombach et al., 2022). Common denoising backbones include convolutional U-Nets and, more recently, Transformer architectures. DiT (Peebles & Xie, 2023) models interactions among image tokens with self-attention and demonstrates scalable diffusion training. Selective state-space models offer another architecture for sequence modeling, with Mamba introducing input-dependent dynamics and linear complexity in sequence length (Gu & Dao, 2023). Building on this formulation, DiM integrates Mamba into a diffusion denoiser with multidirectional scans and local feature enhancement for high-resolution image synthesis (Teng et al., 2024), providing the backbone adopted in this work.

2.2 STYLE TRANSFER

Representative CNN-based methods transfer style by matching feature statistics, with AdaIN aligning channel-wise means and variances (Huang & Belongie, 2017) and WCT matching feature covariances through whitening and coloring transforms (Li et al., 2017). Transformer-based frameworks model content–style interactions through attention, as exemplified by StyTr^2 with separate content and style encoders and a Transformer decoder (Deng et al., 2022). Diffusion-based frameworks bring generative priors to reference-guided stylization (Zhang et al., 2023; Kwon & Ye, 2022; Wang et al., 2023; Deng et al., 2024). Within this family, ArtAdapter (Chen et al., 2024) and StyleShot (Gao et al., 2025) develop multi-level or style-aware reference representations, while DEADiff (Qi et al., 2024) and InstantStyle (Wang et al., 2024) emphasize disentanglement and selective conditioning. CSGO (Xing et al., 2024) explicitly decouples content and style through independent feature injection, while StyleID (Chung et al., 2024) transfers reference style through self-attention key–value manipulation.

Mamba-based frameworks explore content–style interaction through selective state-space processing. Mamba-ST (Botti et al., 2025) derives state-space parameters from separate content and style features, using style features for B and Δ and content features for C. SaMam (Liu et al., 2025) combines a Mamba encoder with a style-aware decoder, while StyMam (Hong et al., 2026) adopts a Mamba-based generator within a GAN framework.

3 METHOD

3.1 PRELIMINARIES

DiM (Teng et al., 2024) uses a Mamba-based denoiser for latent diffusion. At diffusion timestep t, the denoiser takes a noisy latent z_t and a timestep embedding to predict the added noise. We denote the spatial hidden sequence at block ℓ by $H^{(\ell)} \in \mathbb{R}^{N\times L\times D}$, where N, L, and D denote the batch size, number of spatial denoising tokens, and hidden width, respectively. For the feature projections below, vectors are treated as rows and projections act on the last dimension. Sequence operations are evaluated independently for each sample.

Within each block, let u_i be the input at scan position i formed from $H^{(\ell)}$. We write the selective state-space update (Gu et al., 2021; Gu & Dao, 2023) for a single scan channel, omitting batch, block, and channel indices. In this recurrence, states and input-write vectors are represented as columns:

$$\begin{aligned} \bar{A}_i &= \exp(\Delta_i A), & \bar{B}_i &= \mathrm{Disc}_B(A, \Delta_i, B_i), \\ s_i &= \bar{A}_i s_{i-1} + \bar{B}_i u_i, & y_i &= C_i s_i + d_{\mathrm{skip}} u_i. \end{aligned} \tag{1}$$

Here, s_i is the recurrent state, y_i is the scan output, and A is diagonal. For the selected channel, u_i, y_i, Δ_i, and d_{skip} are scalars; $B_i, \bar{B}_i$ are column vectors, and C_i is a row vector. The coefficients $\bar{A}_i$ and $\bar{B}_i$ are discretized state-transition and input-write terms, with Disc_B specifying input discretization. The input-dependent Δ_i, B_i, and C_i control the discretization scale, input write-in, and state readout, respectively.

3.2 OVERVIEW

Given a content image x^c and a reference image x^s, GASC constructs an image-level style descriptor g^s, token-varying style carriers $\widehat{S}^s$, and a spatially aligned content residual R_c (Fig. 2). HOD processes the reference features into a reference representation F^s, from which separate branches derive g^s and a reference carrier bank S^s_{raw}. C^{2}R then uses content anchors from x^c to adapt S^s_{raw} into content-adaptive carriers $\widehat{S}^s$. In parallel, a separate Content Stream encodes content structure into R_c.

The resulting conditions enter each DiM block through distinct conditioning paths. The descriptor g^s conditions Δ, the carriers $\widehat{S}^s$ condition B, and R_c is added before recurrent scanning, while C receives no dedicated external conditioning.

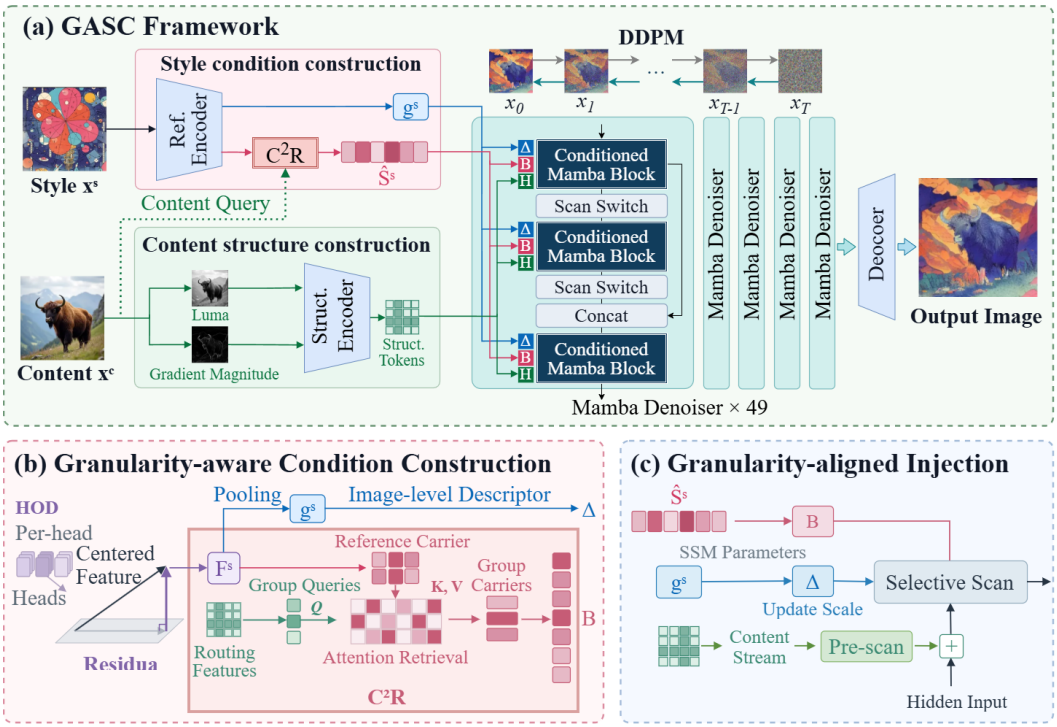

Figure 2: Overview of GASC. (a) Overall framework, with style-condition construction and a conditioned Mamba block expanded in (b) and (c), respectively. (b) Granularity-aware condition construction through HOD and C^{2}R. (c) Granularity-aligned injection into Δ, B, and the pre-scan content path. The final carrier-fusion step is omitted in (b) for clarity.

3.3 Granularity-Aware Style Condition Construction

3.3.1 Per-Head Orthogonal Residualization

HOD aims to reduce content-dominant variation in reference features before constructing the image-level descriptor and reference carrier bank. It removes variation along an offline-fitted principal subspace within each attention head. The association of the selected directions with reference content is examined in Sec. 4.5. For a reference token n, attention head h, and CLIP encoder layer j, let $x_{n,h}^{(j)}, \mu_h^{(j)} \in \mathbb{R}^{1\times d_h}$ be the per-head reference feature before the fixed CLIP head output projection and its fitted mean, respectively. The columns of $U_h^{(j)} \in \mathbb{R}^{d_h\times r_{j,h}}$ form an orthonormal basis for the fitted principal subspace, where d_h is the head width and $r_{j,h}$ is the number of fitted principal directions removed for layer–head pair (j,h). The means and leading principal directions are fitted offline by PCA on token-averaged reference features for each layer–head pair and kept fixed during training. HOD computes

$$\begin{aligned} e_{n,h}^{(j)} &= \left(x_{n,h}^{(j)} - \mu_h^{(j)}\right)\left(I - U_h^{(j)} U_h^{(j)\top}\right), \\ \widetilde{x}_{n,h}^{(j)} &= \left(e_{n,h}^{(j)} + \mu_h^{(j)}\right) W_{O,h}^{(j)}. \end{aligned} \tag{2}$$

Here, $I \in \mathbb{R}^{d_h\times d_h}$ is the identity matrix, $e_{n,h}^{(j)} \in \mathbb{R}^{1\times d_h}$ is the orthogonal residual, and $W_{O,h}^{(j)} \in \mathbb{R}^{d_h\times D_f}$ is the fixed CLIP head output projection. The fitted mean is added back after residualization and before the head projection. The projected contributions $\widetilde{x}_{n,h}^{(j)}$ are summed over heads within each selected layer and averaged across selected layers, yielding $F^s \in \mathbb{R}^{N\times(L_s+1)\times D_f}$. From F^s, which contains one CLS token and L_s patch tokens, separate branches construct the image-level descriptor and reference carrier bank:

$$\begin{aligned} g^s &= f_g\big(\mathrm{Pool}_{L_s+1}\left(\mathcal{N}_g^{\mathrm{enc}}(F^s)\right)\big) \in \mathbb{R}^{N\times D_g}, \\ S^s_{\mathrm{raw}} &= f_s\big(\mathcal{N}_B^{\mathrm{enc}}(F^s_{\mathrm{patch}})\big) \in \mathbb{R}^{N\times L_s\times D_s}. \end{aligned} \tag{3}$$

Here, D_f, D_g, D_s are the widths of reference features, descriptors, and carriers, respectively. F^s_{patch} excludes CLS, whereas Pool_{L_s+1} averages all tokens, including CLS. The operators $\mathcal{N}_g^{\text{enc}}, \mathcal{N}_B^{\text{enc}}$ apply token-wise normalization, and f_g, f_s are separate trainable heads.

3.3.2 Content-Conditioned Group-Level Carrier Retrieval

C^{2}R uses content anchors to organize reference carriers along the content-token axis, without requiring a predefined spatial correspondence during retrieval. The retrieval encoder E_{route} produces content anchors $A^c_{\text{route}} = E_{\text{route}}(x^c) \in \mathbb{R}^{N\times L\times D_a}$, where D_a is the anchor width. A GroupMLP maps each anchor to G group logits, and straight-through hard Gumbel–Softmax (Jang et al., 2016) assigns content tokens to groups at temperature τ:

$$P = \text{GumbelSoftmax}_{\text{ST}}(\text{GroupMLP}(A^c_{\text{route}}); \tau), \tag{4}$$

$$Q_{\text{grp}} = P^\top(A^c_{\text{route}} W_q), \qquad K^s = S^s_{\text{raw}} W_k, \qquad V^s = S^s_{\text{raw}} W_v,$$

$$M = \text{softmax}\left(\frac{Q_{\text{grp}}(K^s)^\top}{\sqrt{d_k}}\right), \tag{5}$$

$$R_{\text{grp}} = MV^s, \qquad S^{\text{route}} = PR_{\text{grp}}. \tag{6}$$

For the retrieval operations, shapes omit the batch dimension, and $(\cdot)^\top$ transposes the last two axes independently for each sample. The assignment matrix $P \in \mathbb{R}^{L\times G}$ groups content tokens, and $Q_{\text{grp}} \in \mathbb{R}^{G\times d_k}$ aggregates their projected anchors into group queries. With learned projections W_q, W_k, and W_v, these queries attend over the reference-token axis, producing group carriers $R_{\text{grp}} \in \mathbb{R}^{G\times D_s}$. Multiplication by P restores them to the content-token axis as $S^{\text{route}} \in \mathbb{R}^{L\times D_s}$. Thus, tokens within a group share a retrieved carrier, while different groups can select different reference style information.

At each content-token position p, a lightweight gate takes the content anchor $A^c_{\text{route},p}$ and retrieved carrier S^{route}_p, and predicts a scalar mixing weight $\lambda_p \in (0, 1)$ through a sigmoid. The final carrier is

$$\widehat{S}^s_p = \lambda_p S^{\text{route}}_p + (1 - \lambda_p)\overline{S}^s_{\text{raw},p}. \tag{7}$$

Here, $\overline{S}^s_{\text{raw}}$ is obtained by matching the reference carrier grid to the content-token grid, using bilinear interpolation only when their resolutions differ. The scalar gate is shared across carrier channels at each position, and stacking the fused carriers gives $\widehat{S}^s \in \mathbb{R}^{N\times L\times D_s}$. Together, group-level retrieval and carrier fusion form the C^{2}R carrier-construction pathway.

3.4 Granularity–Interface Alignment

The two style conditions represent reference appearance at different granularities. g^s is an image-level descriptor, whereas $\widehat{S}^s$ retains variation along the content-token axis. Accordingly, g^s provides token-shared conditioning to Δ, whereas $\widehat{S}^s$ provides token-varying conditioning to B. We match these granularities to the roles of the conditioning interfaces. A shared style residual on the Δ route provides a common reference-dependent modulation of state-transition dynamics while retaining the backbone's token-dependent discretization scales. Token-varying residuals on the B route retain token-wise reference variation when modulating input writes at content-token positions. Both routes use block-specific trainable scalar gates.

Image-level conditioning of Δ. Within block ℓ, the backbone produces a token-dependent feature $q_i^{\Delta,\text{base},(\ell)}$ at scan position i before the dt projection. Block-specific LayerNorm $\text{Norm}_g^{(\ell)}$ and projection $W_\Delta^{(\ell)}$ map g^s into the same dt-rank space. The operator Expand_L broadcasts the projected vector $d^{s,(\ell)}$ across L tokens to form the residual $D_\Delta^{s,(\ell)}$:

$$\begin{aligned}
d^{s,(\ell)} &= \text{Norm}_g^{(\ell)}(g^s)\big(W_\Delta^{(\ell)}\big)^\top, \\
D_\Delta^{s,(\ell)} &= \text{Expand}_L\Big(d^{s,(\ell)}\Big), \\
\widetilde{q}_i^{\Delta,(\ell)} &= q_i^{\Delta,\text{base},(\ell)} + \tanh\Big(\gamma_\Delta^{(\ell)}\Big) D^{s,(\ell)}_{\Delta,i}, \\
\widetilde{\Delta}_i^{(\ell)} &= \text{softplus}\Big(\widetilde{q}_i^{\Delta,(\ell)}\big(W_{dt}^{(\ell)}\big)^\top + b_{dt}^{(\ell)}\Big).
\end{aligned} \tag{8}$$

Here, $\gamma_{\Delta}^{(\ell)}$ controls the residual strength. $\widetilde{q}_i^{\Delta,(\ell)}$ is the style-conditioned pre-dt feature. The projection $W_{dt}^{(\ell)}$ maps it to the scan's inner channel width, with bias $b_{dt}^{(\ell)}$ broadcast across samples and tokens. The resulting discretization scale $\widetilde{\Delta}_i^{(\ell)}$ retains token dependence through the backbone feature despite the shared style residual.

Token-varying conditioning of B. The carriers $\widehat{S}^s$ are reordered using the same scan permutation as the hidden tokens. Let $\pi_\ell(i)$ map scan position i to its corresponding content-token index in block ℓ. Block-specific root mean square normalization $\mathrm{RMSNorm}_B^{(\ell)}$ and projection $W_B^{(\ell)}$ map each aligned carrier from width D_s to the SSM state dimension. The projected contribution is added to the backbone B parameter through a scalar gate:

$$\begin{aligned} B_i^{s,(\ell)} &= \mathrm{RMSNorm}_B^{(\ell)}\left(\widehat{S}^s_{\pi_\ell(i)}\right)\left(W_B^{(\ell)}\right)^\top, \\ B_i^{\mathrm{used},(\ell)} &= B_i^{\mathrm{base},(\ell)} + \tanh\left(\gamma_B^{(\ell)}\right) B_i^{s,(\ell)}. \end{aligned} \tag{9}$$

Here, $B_i^{s,(\ell)}$ is the projected style contribution, and $\gamma_B^{(\ell)}$ controls its residual strength. $B_i^{\mathrm{base},(\ell)}$ and $B_i^{\mathrm{used},(\ell)}$ denote the backbone and conditioned B parameters, respectively, before discretization in the selective scan. For each scan channel in block ℓ, Equation 1 uses the corresponding channel component of $\widetilde{\Delta}_i^{(\ell)}$ and the column vector $\left(B_i^{\mathrm{used},(\ell)}\right)^\top$ in place of Δ_i and B_i, respectively.

3.5 Pre-Scan Content Anchoring

Content Stream supplies structural guidance through the hidden sequence before SSM parameterization. It adds a spatially aligned residual $R_c \in \mathbb{R}^{N\times L\times D}$ to $H^{(\ell)}$ on the spatial token axis before the block normalization preceding the SSM. The same R_c is reused across all 49 blocks. Its encoder E_{stream} is parameterized independently of the retrieval encoder E_{route}.

The encoder E_{stream} includes structural preprocessing that extracts the luminance and normalized gradient magnitude of x^c. The content anchor $A^c_{\mathrm{stream}} = E_{\mathrm{stream}}(x^c) \in \mathbb{R}^{N\times L\times D}$ is normalized and projected before addition to the hidden sequence:

$$\begin{aligned} R_c &= \mathrm{LN}_c\left(A^c_{\mathrm{stream}}\right) W_c^\top, \\ \widetilde{H}^{(\ell)} &= H^{(\ell)} + R_c, \qquad \ell = 0,\ldots,48. \end{aligned} \tag{10}$$

The content-anchored sequence $\widetilde{H}^{(\ell)}$ is processed by the block's existing hidden-to-SSM transformations, from which the backbone base quantities for Δ, B, and C are computed. Thus, R_c influences these quantities through the shared hidden representation rather than through dedicated content-conditioning branches.

3.6 Training Objective

Training uses content–reference–target triplets $(x^c, x^s, x^{\mathrm{pt}})$ from OmniStyle-150K, with the dataset's paired stylized targets providing supervision (Sec. 4.1). The frozen VAE encoder E_{VAE} maps the target x^{pt} to the clean latent z_0. Given $\epsilon \sim \mathcal{N}(0, I_z)$ and diffusion timestep t,

$$z_0 = E_{\mathrm{VAE}}(x^{\mathrm{pt}}), \qquad z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon. \tag{11}$$

Here, I_z is the identity covariance in latent space, and $\bar{\alpha}_t$ denotes the cumulative signal retention factor.

We optimize the conditional denoiser with the standard noise-prediction objective:

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{x^c,x^s,x^{\mathrm{pt}},t,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, t \mid g^s, \widehat{S}^s, R_c\right)\right\|_2^2\right]. \tag{12}$$

Here, ϵ_θ denotes the conditional denoiser, with θ collecting all trainable denoiser and conditioning parameters. The CLIP(Radford et al., 2021) encoder, VAE, HOD statistics, and pretrained SSM transition parameter A remain fixed.

Table 1: Quantitative comparison with existing reference-based image style transfer methods. The best result for each metric is shown in bold and the second-best is underlined.

	Transformer-based		Diffusion-based					Mamba-based	Ours
Metric	StyTr^2	S2WAT	StyleID	DEADiff	OmniStyle	SSI	StyleFM	SaMam	GASC
FID ↓	23.848	20.028	22.250	25.971	<u>17.839</u>	22.851	23.180	19.370	**16.835**
LPIPS ↓	0.512	**0.411**	0.520	0.708	0.701	0.517	0.486	<u>0.475</u>	0.587
ArtFID ↓	37.570	<u>29.671</u>	35.340	46.066	32.045	36.182	35.931	30.046	**28.304**
CFSD ↓	0.415	<u>0.302</u>	0.587	0.643	0.521	0.625	0.522	0.503	**0.264**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

We conduct paired style transfer training on the OmniStyle-150K dataset (Wang et al., 2025). Each training sample consists of a content image x^c, a reference style image x^s, and a paired stylized target x^{pt}, which serves as the supervision target. Training is performed on four NVIDIA H20 GPUs. The default HOD rank ratio is 0.5, and C^{2}R uses $G = 12$ groups. At inference, we use DDPM (Ho et al., 2020) with 50 steps.

Evaluation protocol. We compare GASC with eight reference-based style transfer methods, including StyTr^2 (Deng et al., 2022), S2WAT (Zhang et al., 2024), StyleID (Chung et al., 2024), DEADiff (Qi et al., 2024), OmniStyle (Wang et al., 2025), SSI (Kulkarni, 2026), StyleFM (Ma et al., 2026), and SaMam (Liu et al., 2025), using their official implementations and recommended configurations. We sample 20 content images and 40 references from public datasets including COCO (Lin et al., 2014) and WikiArt (Tan et al., 2018), forming all $20 \times 40 = 800$ content–style pairs. All methods use the same evaluation pairs, preprocessing, and metric implementations.

Following StyleID (Chung et al., 2024), we report FID (Heusel et al., 2017), LPIPS (Zhang et al., 2018), ArtFID (Wright & Ommer, 2022), and content feature structural distance (CFSD) (Chung et al., 2024). FID measures distribution-level style similarity, while LPIPS and CFSD assess content preservation. ArtFID is defined as $\mathrm{ArtFID} = (1 + \mathrm{LPIPS})(1 + \mathrm{FID})$. Lower values are better for all four metrics.

4.2 QUANTITATIVE COMPARISON

Table 1 shows that GASC combines strong stylization with content structure preservation, achieving the lowest FID, ArtFID, and CFSD among the compared methods. While S2WAT yields greater perceptual similarity to the content input as measured by LPIPS, GASC better preserves content feature structure as measured by CFSD.

4.3 QUALITATIVE COMPARISON

Figure 3 compares how well each method transfers reference appearance while preserving content structure. In the first row, GASC incorporates the reference's red accents and light background tones while retaining the city's architectural layout. In the fourth row, it renders the bird with red cut-paper patterns while preserving its spread-wing pose and silhouette. For this example, DEADiff and OmniStyle produce orange-and-black and pink renderings, respectively, departing from the reference palette.

4.4 ANALYSIS OF GRANULARITY–INTERFACE ALIGNMENT

We analyze granularity–interface alignment through controlled condition allocation and route complementarity.

Condition allocation. We compare token-shared (S) and token-varying (V) conditioning in a 2×2 Δ/B design. S mean-pools and broadcasts conditioning, whereas V preserves token-wise variation.



Figure 3: Qualitative comparison. We compare GASC (Ours) with representative reference-based image style transfer methods across diverse content images and reference styles.

Table 2: Analysis of granularity–interface alignment. S and V denote token-shared and token-varying conditioning support, respectively, with assignments written in Δ/B order. S/S, V/V, V/S, and S/V form the 2×2 allocation comparison, while the single-route variants are independently retrained. Best result for each metric is shown in bold.

	Alternative Allocations			Single-Route Variants		Ours
Metric	S/S	V/V	V/S	B-only	Δ-only	**S/V**
FID↓	18.637	21.734	22.960	23.120	20.010	**16.835**
LPIPS↓	0.560	0.530	**0.502**	0.522	0.555	0.587
ArtFID↓	30.634	34.783	35.988	36.711	32.671	**28.304**
CFSD↓	0.259	0.325	0.306	0.258	**0.237**	0.264

For Δ, S pools normalized reference features, including CLS, before the descriptor head; V instead maps patch tokens without CLS and aligns the residuals with the denoising grid and scan order. The proposed S/V allocation uses shared Δ and varying B. All variants are independently trained under identical settings and evaluated on the same 800 content–style pairs.

Table 2 supports the proposed S/V allocation. At fixed $B = V$, shared Δ conditioning (S/V) outperforms V/V on FID and ArtFID; at fixed $\Delta = S$, token-varying B conditioning (S/V) similarly outperforms S/S. These controlled trends support shared conditioning for Δ and token-varying conditioning for B on FID and ArtFID, while other assignments achieve lower LPIPS or CFSD.

Route complementarity. We independently retrain B-only and Δ-only variants while keeping the remaining conditioning modules and training settings fixed. As shown in Table 2, joint $\Delta + B$ conditioning achieves lower FID and ArtFID than either single-route variant. Both single-route variants obtain lower LPIPS and CFSD than joint conditioning; among them, B-only yields the lower LPIPS and Δ-only the lower CFSD. This comparison supports complementary contributions from the two aligned routes to distribution-level stylization.

4.5 Analysis of Conditioning Components

We analyze style-condition construction and content anchoring within GASC (Table 3).

HOD. With C^{2}R disabled, enabling HOD lowers FID, ArtFID, and CFSD while leaving LPIPS nearly unchanged (Table 3(a)). Representation analysis further shows that the selected top PCs



Figure 4: Qualitative comparison of alternative conditioning allocations and single-route variants. S and V denote token-shared and token-varying support, respectively.

Table 3: Analysis of conditioning components. (a) Style-condition ablation, where "w/o" denotes without and GASC denotes the full configuration. (b) Independently trained content-conditioning variants with identical style pathways and training settings; the content paths are not capacity-matched.

(a) Style-condition ablation

Metric	w/o HOD, C^{2}R	w/o C^{2}R	GASC
FID ↓	19.180	17.655	**16.835**
LPIPS ↓	0.599	0.598	**0.587**
ArtFID ↓	32.268	29.811	**28.304**
CFSD ↓	0.281	**0.260**	0.264

(b) Content anchoring

Metric	Content-to-C	Content Stream
FID ↓	17.354	**16.835**
LPIPS ↓	0.603	**0.587**
ArtFID ↓	29.421	**28.304**
CFSD ↓	0.265	**0.264**

have higher descriptive content association than the remaining PCs (0.213 vs. 0.169), with positive differences in 83 of 96 layer–head pairs (Appendix Figure 5). This observation is consistent with HOD targeting content-dominant variation before style-condition construction; further analysis is provided in Appendix D.1.

C^{2}R. Removing the complete C^{2}R retrieval-and-fusion pathway increases FID, ArtFID, and LPIPS, with little change in CFSD (Table 3(a)). These results support content-conditioned carrier construction for forming the final token-varying style carriers.

Content anchoring. Content Stream improves all four metrics over Content-to-C, with lower FID, ArtFID, and LPIPS and a marginally lower CFSD (Table 3(b)). This favors pre-scan content anchoring, which conditions the hidden sequence before recurrent state updates rather than the readout C.

5 Conclusion

We addressed the problem of how style representations with different granularities should be assigned to conditioning interfaces in reference-based image style transfer. To this end, we introduced GASC and the principle of granularity–interface alignment, which couples representation granularity with how different interfaces act on the denoiser. GASC constructs complementary image-level and token-varying style conditions through HOD and C^{2}R, assigns them to the Δ and B pathways, and anchors content structure separately through a pre-scan Content Stream. Experiments demonstrate strong stylization quality while preserving content structure, and controlled assignment studies support the proposed allocation under the current model. Overall, our findings suggest that explicitly coordinating representation design with interface action provides a useful perspective for conditioning heterogeneous state-space denoisers.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting and polishing the manuscript. We have not used generative AI tools to fabricate or alter any experimental result. We reviewed all AI-assisted content and take full responsibility for the final text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

To support reproducibility, we document the GASC architecture and its key components, including style-condition construction, granularity–interface alignment, and pre-scan content anchoring. The main paper and appendix describe the model configurations, training and inference settings, and computing hardware. Our experiments use publicly accessible datasets, including MS COCO, WikiArt, and OmniStyle-150K, with data preprocessing procedures and evaluation protocols specified to facilitate replication of the reported results.

REFERENCES

Filippo Botti, Alex Ergasti, Leonardo Rossi, Tomaso Fontanini, Claudio Ferrari, Massimo Bertozzi, and Andrea Prati. Mamba-ST: State space model for efficient style transfer. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7797–7806. IEEE, 2025.

Dar-Yen Chen, Hamish Tennent, and Ching-Wen Hsu. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8619–8628. IEEE, 2024.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8795–8805. IEEE, 2024.

Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. StyTr2: Image style transfer with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11326–11336. IEEE, 2022.

Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. Z*: Zero-shot style transfer via attention reweighting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6934–6944. IEEE, 2024.

Junyao Gao, Yanan Sun, Yanchen Liu, Yinhao Tang, Yanhong Zeng, Ding Qi, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

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.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

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.

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

Zhou Hong, Rongsheng Hu, Yicheng Di, Xiaolong Xu, Ning Dong, Yihua Shao, Run Ling, Yun Wang, Juqin Wang, Zhanjie Zhang, et al. Stymam: A mamba-based generator for artistic style transfer. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10112–10116. IEEE, 2026.

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

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Amey Sunil Kulkarni. Scheduled style injection: Expanding the style-content pareto frontier in training-free diffusion-based style transfer. *arXiv preprint arXiv:2605.26538*, 2026.

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

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.

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*, pp. 740–755. Springer, 2014.

Hongda Liu, Longguang Wang, Ye Zhang, Ziru Yu, and Yulan Guo. Samam: Style-aware state space model for arbitrary image style transfer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28468–28478. IEEE, 2025.

Yingnan Ma, Zhenye Liu, Siying Liu, and Anup Basu. Stylefm: Frequency manipulation empowered by recursive attention on diffusion models for arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7865–7873, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205. IEEE, 2023.

Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8693–8702. Ieee, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10684–10695. ieee, 2022.

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.

Yao Teng, Yue Wu, Han Shi, Xuefei Ning, Guohao Dai, Yu Wang, Zhenguo Li, and Xihui Liu. Dim: Diffusion mamba for efficient high-resolution image synthesis. *arXiv preprint arXiv:2405.14224*, 2024.

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.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7847–7856. IEEE, 2025.

Zhizhong Wang, Lei Zhao, and Wei Xing. StyleDiffusion: Controllable disentangled style transfer via diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7677–7689. IEEE, 2023.

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

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.

Chiyu Zhang, Xiaogang Xu, Lei Wang, Zaiyan Dai, and Jun Yang. S2wat: Image style transfer via hierarchical vision transformer using strips window attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 7024–7032, 2024.

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

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10146–10156. IEEE, 2023.

A PRELIMINARIES

A.1 SELECTIVE STATE-SPACE UPDATE

A continuous state-space model maps a one-dimensional input signal through

$$\dot{s}(i) = As(i) + Bu(i), \qquad y(i) = Cs(i) + Du(i), \tag{13}$$

where $s(i)$ denotes the latent state, and $u(i)$ and $y(i)$ denote the input and output, respectively. For discrete inputs, the state-transition and input terms are discretized with step size Δ. Following the selective state-space formulation used by Mamba (Gu & Dao, 2023), the recurrent update can be written as

$$\begin{aligned} \bar{A}_i &= \exp(\Delta_i A), & \bar{B}_i &= \mathrm{Disc}_B(A, \Delta_i, B_i), \\ s_i &= \bar{A}_i s_{i-1} + \bar{B}_i u_i, & y_i &= C_i s_i + d_{\mathrm{skip}} u_i. \end{aligned} \tag{14}$$

Here, Disc_B denotes discretization of the input term under step size Δ_i. Mamba makes Δ_i, B_i, and C_i input-dependent, yielding selective state-space updates. For one scan direction and one scalar SSM channel, Δ_i controls the discretization of the state transition, B_i controls how the current input is written into the recurrent state, and C_i reads the recurrent state into the output.

A.2 DIM BACKBONE

GASC builds on Diffusion Mamba (DiM) (Teng et al., 2024), which adapts Mamba to two-dimensional diffusion features through spatial tokenization and alternating scan-based sequence modeling. At diffusion timestep t, the noisy latent $z_t \in \mathbb{R}^{4\times32\times32}$ is partitioned into 2×2 patches, yielding a 16×16 grid of 256 image tokens. DiM applies a lightweight 3×3 depth-wise convolution before the Mamba stack to enhance local spatial coherence and inserts learnable padding tokens at spatial boundaries before flattening the grid into a sequence. Successive Mamba blocks alternate forward and reverse row- and column-major scan orders to reduce directional bias, while long skip connections propagate shallow features to deeper stages.

After sequence modeling, the tokens are restored to the two-dimensional patch grid and passed through a lightweight depth-wise convolution and an output projection for noise prediction. We adopt the 49-block DiM configuration with hidden width 1536, SSM state dimension 16, and dt-rank 96.

GASC retains this DiM backbone and modifies how image-derived conditions enter its state-space blocks. The image-level style descriptor g^s provides token-shared conditioning to Δ, while the token-varying style carriers $\widehat{S}^s$ condition B after alignment with the denoising grid and the scan order of the corresponding block. A separate Content Stream adds the spatially aligned residual R_c to the hidden sequence before recurrent scanning, while C receives no dedicated external conditioning.

B IMPLEMENTATION AND REPRODUCIBILITY DETAILS

Architecture overview. The final GASC model uses a 49-block DiM denoiser with hidden width 1536. Table 4 summarizes its conditioning architecture.

HOD configuration. HOD extracts per-head features from zero-based transformer-block indices 18–23 of CLIP ViT-L/14, corresponding to its final six transformer blocks. Each block contains 16 attention heads with head width $d_h = 64$, giving 96 layer–head pairs in total. The PCA statistics are fitted offline using 500 style images randomly sampled with seed 42 from the OmniStyle-150K training style split. Each image is resized and center-cropped to 224×224 before CLIP normalization. For every layer–head pair (j, h), we average the features before the attention output projection over all tokens, including CLS, to obtain one 64-dimensional vector per image. We fit an independent PCA basis to the resulting 500×64 matrix. Let $k_{j,h}$ denote the minimum number of principal components whose cumulative explained variance reaches a fixed threshold of 90%. The residualized rank is

$$r_{j,h} = \max\left(1, \lfloor 0.5\, k_{j,h} \rfloor\right).$$

The fitted means and principal bases, together with the corresponding per-head slices of the CLIP output projection $W_{O,h}^{(j)}$, are fixed. Head contributions are summed after the fixed output projections,

Table 4: GASC conditioning architecture in the evaluated configuration. Shapes omit the batch dimension.

Component	Operation	Output	Use in the denoiser
HOD aggregation	Per-head residualization in CLIP ViT-L/14 blocks 18–23, followed by fixed output projection and aggregation	257×1024 sequence	Shared source for the two style branches
Image-level descriptor	Token-wise normalization, mean pooling, and descriptor head f_g	$g^s \in \mathbb{R}^{1536}$	Block-specific $1536 \to 96$ projection, broadcast on the Δ path
Reference carrier	Token-wise normalization, CLS removal, and carrier head f_s	$S^s_{\mathrm{raw}} \in \mathbb{R}^{256\times 1536}$	Carrier bank for C^{2}R
C^{2}R retrieval	256-D content anchors on a 16×16 grid; group assignment and carrier retrieval	$S^{\mathrm{route}} \in \mathbb{R}^{256\times 1536}$	Local carrier construction
C^{2}R fusion	Token-wise content-dependent interpolation between S^{route} and the length-matched raw carrier	$\widehat{S}^s \in \mathbb{R}^{256\times 1536}$	Block-specific $1536 \to 16$ projection on the B path
Content Stream	Independent structural encoder, shared LayerNorm, and $1536 \to 1536$ projection	$R_c \in \mathbb{R}^{256\times 1536}$	Same pre-scan residual reused in all 49 blocks

and the six selected layer outputs are averaged:

$$F^s = \frac{1}{6} \sum_{j=18}^{23} \left(\sum_{h=1}^{16} \widetilde{x}_h^{(j)} \right).$$

The aggregated sequence contains one CLS token and 256 patch tokens. In Equation 3, both $\mathcal{N}^{\mathrm{enc}}$ operators apply token-wise LayerNorm without learnable affine parameters. The global branch then averages all 257 tokens, while the carrier branch discards CLS. Each head f_g or f_s applies a second LayerNorm, with learnable affine parameters, followed by a $1024 \to 1536$ linear projection. The heads have separate parameters. In the evaluated configuration, $L_s = L = 256$.

C^{2}R configuration. The retrieval encoder E_{route} produces 256-dimensional content anchors on a 16×16 grid. We use $G = 12$ groups and query/key dimension $d_k = 64$. Group assignments use straight-through hard Gumbel–Softmax with temperature $\tau = 0.5$ during training and inference. Group queries sum the projected anchors assigned to each group without normalization by group size. Reference carriers are shuffled before group-level retrieval.

For carrier fusion, the content anchor and retrieved carrier are separately projected and concatenated to predict a scalar mixing weight at each content-token position p:

$$\lambda_p = \sigma\Big(\mathrm{Gate}\Big(A^c_{\mathrm{route},p}\big(W_{g,c}\big)^\top \,\|\, S^{\mathrm{route}}_p\big(W_{g,r}\big)^\top\Big)\Big).$$

Here, σ denotes the sigmoid, $\|$ denotes concatenation, and the projection matrices use the output-by-input convention. The gate uses only the content anchor and retrieved carrier, not the raw carrier. The final carrier $\widehat{S}^s$ is formed using Equation 7 before block-specific B-route normalization and projection. In the evaluated configuration, the reference and content-token grids both have size 16×16, so $\overline{S}^s_{\mathrm{raw}} = S^s_{\mathrm{raw}}$ without interpolation. This grid alignment does not imply semantic correspondence between reference and content positions.

Content Stream configuration. The Content Stream uses an independently parameterized instance of the structural encoder used for routing. Its four stages have hidden widths $(96, 192, 384, 384)$, and its structural input consists of image luma and Sobel-gradient magnitude. The output anchors are mapped once by the shared normalization and projection in Equation 10. The resulting pre-scan residual R_c is reused in all 49 blocks, as summarized in Table 4.

Carrier alignment with scan order. The carrier is first indexed on the content-token spatial grid. At each block, its ordering is advanced using the same scan-pattern transition as the hidden sequence, before the token-wise B projection. When padding or extra tokens are present, condition-sequence wrapping and unwrapping account for these positions so that the carrier follows the hidden sequence's ordering in the augmented layout. The full permutation acts on this augmented layout; π_ℓ in Equation 9 denotes the induced ordering of the L image tokens, excluding padding and extra positions.

Conditioning projections and gates. The projection widths in Table 4 correspond to a dt-rank of 96 and an SSM state dimension of 16. The B projection follows block-specific RMSNorm. Both routes initialize their scalar gate parameters to 0.03 and apply $\tanh$ before scaling the external residuals.

Training configuration. The CLIP reference encoder, Stable Diffusion VAE, HOD statistics, and pretrained SSM transition parameter A remain fixed throughout training. Reference images are resized and center-cropped to 224×224 before being processed by OpenAI CLIP ViT-L/14, while image latents are obtained using the Stable Diffusion VAE.

Training is performed for 480,000 optimization steps with a global batch size of 32. We optimize all trainable parameters with AdamW using a learning rate of 10^{-4}, weight decay 0.03, $(\beta_1, \beta_2) = (0.9, 0.999)$, and $\epsilon = 10^{-15}$. The learning rate is linearly warmed up over the first 5,000 steps and kept constant thereafter.

C Implementation of the 2×2 Condition-Allocation Study

This section specifies the conditions used in the 2×2 allocation study. Assignments are written in Δ/B order: S denotes a token-shared style residual, whereas V denotes a residual that retains variation along the denoising-token axis. Within each route, S and V use the same conditioning-map architecture, gating form, and injection location. Their parameters are learned independently in each experimental run.

Δ-route construction. Let

$$U^s = \mathcal{N}_g^{\mathrm{enc}}(F^s) \in \mathbb{R}^{N \times (L_s+1) \times D_f}$$

denote the normalized reference features after HOD. Omitting the batch index, U_0^s is the CLS token and $U_1^s, \ldots, U_{L_s}^s$ are the reference patch tokens. For block ℓ, define

$$\Phi_\Delta^{(\ell)}(u) = \mathrm{Norm}_g^{(\ell)}(f_g(u)) \left(W_\Delta^{(\ell)}\right)^\top.$$

The descriptor head f_g and the block-specific normalization and projection retain the order used in the main method. When applied to a sequence, $\Phi_\Delta^{(\ell)}$ acts on each token independently.

The S condition follows the main method: it averages the normalized reference features, including CLS, before applying the conditioning map. The resulting residual is shared by all L denoising tokens. The V condition instead applies the map to the normalized reference patch tokens without pooling:

$$D_{\Delta,i}^{S,(\ell)} = \Phi_\Delta^{(\ell)}\left(\frac{1}{L_s+1}\sum_{n=0}^{L_s} U_n^s\right), \qquad i = 1, \ldots, L, \tag{15}$$

$$D_{\Delta,i}^{V,(\ell)} = \left[\mathcal{R}_L\left(\Phi_\Delta^{(\ell)}(U_{1:L_s}^s)\right)\right]_{\pi_\ell(i)}, \qquad i = 1, \ldots, L. \tag{16}$$

The operator $\mathcal{R}_L$ reshapes the projected patch tokens to the reference patch grid, aligns them with the denoising-token grid by bilinear interpolation when the grid resolutions differ, and flattens the result in spatial order. No interpolation is applied when the grids match. The index $\pi_\ell(i)$ then maps scan position i to its spatial token index in block ℓ. This operation provides spatial-grid alignment; it does not perform content-conditioned retrieval.

Both residuals are added to the backbone's token-dependent pre-dt feature through the same gated Δ-route injection in Equation 8. Thus, a shared style residual does not make the resulting Δ token-independent. Normalization of F^s precedes pooling in S, whereas f_g, the block-specific normalization, and projection follow pooling. In V, these maps are applied token-wise before grid alignment.

B-route construction. Let

$$\widehat{S}^s \in \mathbb{R}^{N\times L\times D_s}$$

denote the final content-adaptive style carriers constructed by C^{2}R and carrier fusion. Both B-route variants use this carrier sequence. S averages it along the content-token axis before block-specific RMSNorm and projection, whereas V retains the individual carriers:

$$D_{B,i}^{S,(\ell)} = \mathrm{RMSNorm}_B^{(\ell)}\left(\frac{1}{L}\sum_{p=1}^{L}\widehat{S}_p^s\right)\left(W_B^{(\ell)}\right)^\top, \qquad i=1,\dots,L, \tag{17}$$

$$D_{B,i}^{V,(\ell)} = \mathrm{RMSNorm}_B^{(\ell)}\left(\widehat{S}^s_{\pi_\ell(i)}\right)\left(W_B^{(\ell)}\right)^\top, \qquad i=1,\dots,L. \tag{18}$$

The S residual is broadcast across scan positions. In V, each carrier follows the corresponding content token under the scan permutation. Both residuals use the same gated addition to the backbone B parameter as Equation 9; the backbone-generated B remains active in every variant.

Comparison scope. The four models combine the conditions above as S/S, V/V, V/S, and S/V in Δ/B order. In particular, S/V uses D_Δ^S and D_B^V and corresponds to the proposed GASC allocation. All variants are independently trained from the same pretrained DiM initialization under otherwise identical data and optimization settings, and are evaluated on the same 800 content–style pairs.

Pooling precedes the conditioning map in Δ-S but not in Δ-V. Because that map contains nonlinear operations, the shared residual need not equal the mean of the token-varying residuals. In addition, Δ-S includes CLS whereas Δ-V uses only patch tokens. The study therefore compares these specified conditioning configurations.

D Additional Analysis of HOD

D.1 HOD Representation Analysis

Table 3(a) reports the downstream effect of HOD within the full conditioning pipeline. We complement that system-level comparison with the representation analysis below.

Dataset and descriptive content-association score. The representation analysis uses 155 WikiArt images. WikiArt genre labels provide 15 genre categories, while manual annotation provides eight content classes: abstract, architecture, genre scene, landscape, other, portrait, religious/mythological, and still life. These labels are used only for post-hoc representation analysis and are not involved in fitting the HOD principal subspaces.

For each image and layer–head pair (j,h), we subtract the fitted PCA mean from each token feature, project onto principal component k, and average the projection scores over all tokens, including CLS, to obtain the image-level score $s_{j,h,k}$. Let V_{content} and V_{genre} denote the variances of the empirical content- and genre-category means, respectively, and let V_{total} denote the total score variance across images. We define the nonnegative remainder and descriptive content-association score as

$$V_{\mathrm{rem}} = \max(0, V_{\mathrm{total}} - V_{\mathrm{content}} - V_{\mathrm{genre}}), \quad \rho_c = \frac{V_{\mathrm{content}}}{V_{\mathrm{content}} + V_{\mathrm{genre}} + V_{\mathrm{rem}} + 10^{-12}}.$$

Category means are weighted equally when computing V_{content} and V_{genre}; all variances use the population convention, and 10^{-12} ensures numerical stability. Because genre and content categories may overlap in subject matter, their variance terms need not capture disjoint sources of variation. Thus, ρ_c remains descriptive, and V_{rem} is not an estimated interaction effect.

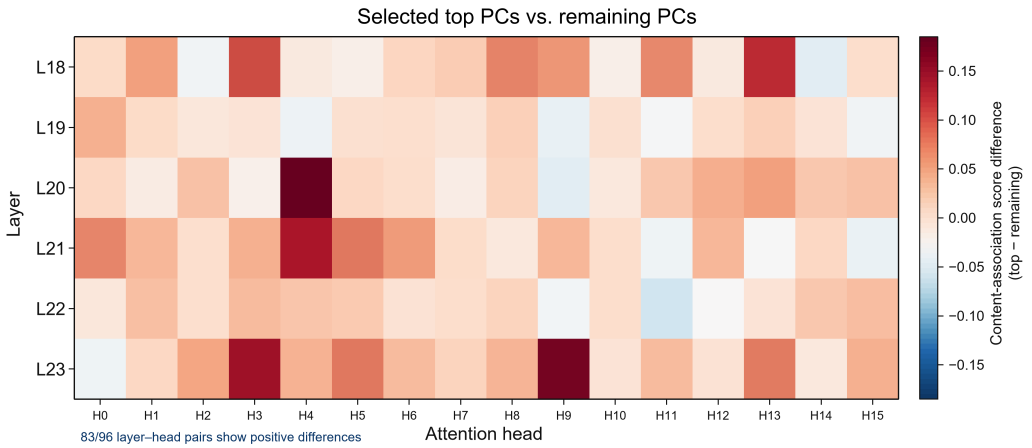

Figure 5: Difference in descriptive content-association score between selected top PCs and remaining PCs. Each cell represents a layer–head pair; positive values indicate higher content association for the selected top PCs.

Table 5: Fixed-checkpoint sensitivity to the HOD rank ratio.

HOD ratio ρ	ArtFID ↓	FID ↓	LPIPS ↓
`none`	30.555	18.400	**0.575**
0.25	29.562	17.663	0.584
0.50 (default)	**28.304**	**16.835**	0.587
0.75	29.646	17.787	0.578

Selected and remaining principal components. For each layer–head pair, the top-PC set consists of the first $r_{j,h}$ directions defined by the HOD rank rule above. The remaining-PC set contains directions $r_{j,h} + 1, \ldots, k_{j,h}$ within the retained PCA basis. We average the scores within each set and then equally across the 96 pairs; each heatmap cell shows the top-PC minus remaining-PC mean for one pair. The resulting means are 0.213 and 0.169, respectively, a difference of 0.044, and 83 of 96 pairs have a positive difference. Under this descriptive score, the directions selected by HOD exhibit stronger content-category association than the remaining PCs.

HOD Rank Sensitivity. Using a fixed checkpoint without retraining, we replace the default factor 0.5 in the rank rule with ρ; `none` disables HOD residualization. The default $\rho = 0.5$ yields the lowest FID and ArtFID, while `none` yields the lowest LPIPS (Table 5).

E More results

Finally, we provide additional qualitative results that complement the quantitative comparisons and ablation studies in the main paper. The examples cover additional content–style pairs and are intended to illustrate the visual behavior of the proposed model under the corresponding evaluation setting.



Figure 6: More style transfer results.

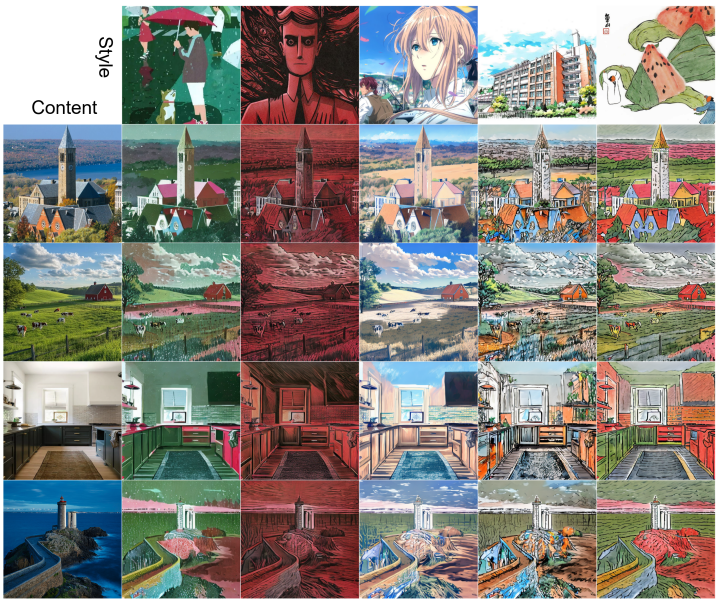

Figure 7: More style transfer results.