FINESTYLER: TOWARDS DETAIL-PRESERVING VIDEO STYLIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-based video stylization transfers the style of a reference image to an input video while preserving its content and motion. Existing video diffusion methods often rely on overly compressed style representations and coarse structural conditioning, leading to washed-out micro-style details, reference content leakage, and temporal drift of fine structures. We present **FineStyler**, a single-stage inference framework for high-fidelity, temporally stable video stylization with a single style reference image. First, **Content-Anchored Style Modulation** aligns learned style embeddings to a natural-image feature domain, encouraging fine-grained texture encoding while suppressing instance-specific content leakage. Second, **High-Frequency Adapted LoRA** injects input-video high-frequency cues into DiT self-attention through a parameter-efficient adapter, preserving thin edges and local geometry beyond silhouette-level controls. Third, a style reference image curation pipeline synthesizes style-aligned yet content-mismatched references from paired raw and stylized videos using image stylization models, then selects high-quality references through automatic scoring to enable training that matches the inference interface. Extensive experiments demonstrate that FineStyler improves style faithfulness, content preservation, and temporal stability over prior diffusion-based baselines, producing sharper structures and higher perceptual quality in challenging content-mismatched reference settings.

1 INTRODUCTION

Diffusion models (Dhariwal & Nichol, 2021; Ho et al., 2020; Nichol & Dhariwal, 2021; Saharia et al., 2022; Sohl-Dickstein et al., 2015), particularly video systems built on Diffusion Transformer (Peebles & Xie, 2023) backbones, have advanced visual fidelity and dynamics. Yet video stylization remains challenging: given a reference image and an input video, it must faithfully transfer fine-grained style while preserving the source video's content and motion over time. Existing approaches struggle to meet these requirements, producing visually unstable results.

A key bottleneck is the style representation. Some methods (Liu et al., 2023; Ye et al., 2025) compress the reference into a single global embedding or a small set of tokens. Without explicit objectives for fine-grained style cues, these representations tend to collapse into coarse appearance hints, allowing the diffusion prior to wash out micro-style details and produce over-smoothed renderings with thickened contours and an overly glossy appearance. Weak style-content disentanglement may also allow reference semantics to enter the style pathway. StyleCrafter (Liu et al., 2023), for example, is prone to leaking reference content into the output videos (Fig. 5), indicating that its tokens retain instance-specific content and generalize poorly to content-mismatched references. AnyV2V (Ku et al., 2024) instead propagates an edited first frame, but our comparisons also show incomplete reference-style transfer and changes to source content (Fig. 5).

Another limitation is coarse structural conditioning. ControlNet (Zhang et al., 2023) is used in StyleMaster (Ye et al., 2025), constrains layout but often provides only silhouette-level guidance. Thin edges, small structures, and boundaries of the source video can still drift during denoising (Fig. 5). The input video's high-frequency components encode the missing sharp contours and detailed geometry, but are not explicitly used by these coarse controls for fine-grained content preservation.

We introduce **FineStyler** to transfer fine-grained reference style while preserving video content and temporal structure. First, Content-Anchored Style Modulation trains a query transformer with

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.

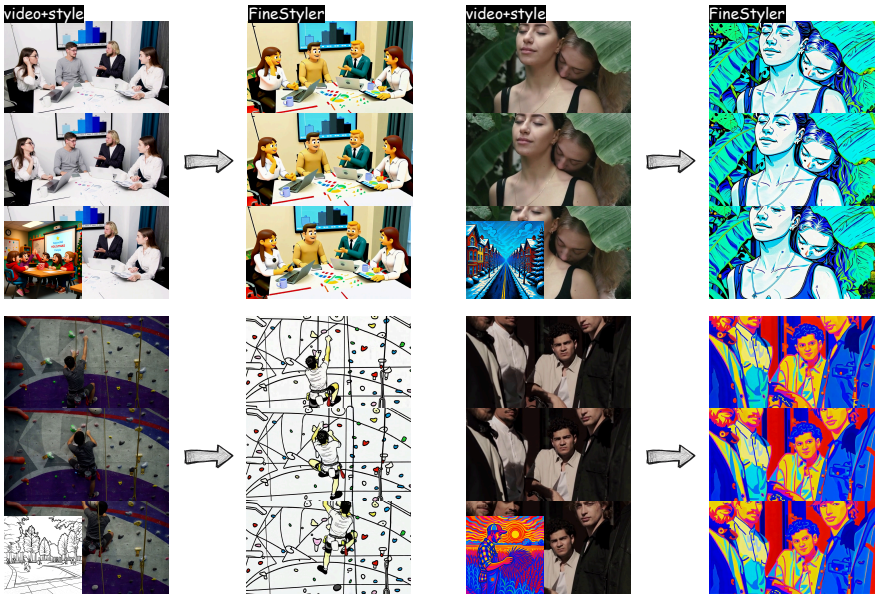

Figure 1: **Video stylization results generated by FineStyler.** FineStyler transfers reference style to the entire video while preserving content structure and temporal coherence, producing faithful style patterns and clean, stable results.

explicit reconstruction supervision from paired natural images. Given a stylized reference, the transformer extracts style tokens while reconstructing the corresponding natural-image latent. This anchors reference content in the reconstructed latent, while cross-content stylization supervision encourages fine-grained, content-invariant style tokens. Second, High-Frequency Adapted LoRA extracts high-frequency components from the input video latents and projects them alongside denoising features into a shared low-rank space. A learnable gate fuses the two feature streams before modulating the query, key, and value projections of DiT self-attention. This supplies fine structural cues throughout denoising to preserve thin edges and local geometry beyond outline supervision.

The existing video training data consist of paired raw and stylized videos, but lack the style-consistent yet content-mismatched reference images needed for adapter learning (Section 3.2). To fill this supervision gap, we select a representative style anchor frame from each stylized video and use image stylization models to transfer its style to images with different content. We then automatically score the generated candidates for visual quality, style consistency, and content leakage, retaining the highest-ranked references for each video. This curated reference set enables training that matches the inference interface and improves robustness to content-mismatched inputs.

To summarize, our main contributions are threefold:

- **Content-Anchored Style Modulation.** We introduce natural-image reconstruction supervision for the query transformer. Together with cross-content stylization, this objective encourages fine-grained style encoding while suppressing reference content leakage.
- **High-Frequency Adapted LoRA.** A parameter-efficient adapter injects video high-frequency cues into DiT self-attention, strengthening thin-edge and fine-structure preservation throughout denoising for higher content fidelity and perceptual quality.
- **Style Reference Image Curation Pipeline.** We synthesize style-aligned, content-mismatched references from paired raw/stylized videos and rank them by visual fidelity and style consistency, providing supervision that matches the inference interface.

2 Related Work

Video Diffusion Generation Diffusion models (Dhariwal & Nichol, 2021; Ho et al., 2020; Nichol & Dhariwal, 2021; Saharia et al., 2022; Sohl-Dickstein et al., 2015) have powered major breakthroughs

in the field of image and video generation. Latent Diffusion Models (Rombach et al., 2022) (LDMs) provide the core framework of performing diffusion in the Variational Autoencoder (Kingma & Welling, 2013) (VAE) latent space to reduce computation. A widely adopted instantiation of LDMs adopts a latent-space U-Net (Ronneberger et al., 2015) to enable high-quality, efficient text-to-image generation, which also bears many early video generation models (Guo et al., 2023; Blattmann et al., 2023; Wang et al., 2025b). More recently, the impressive quality demonstrated by large-scale systems has highlighted the promise of Diffusion Transformers (Peebles & Xie, 2023) (DiT) for video generation. Unlike convolutional U-Net backbones, DiT adopts a token-based formulation that enables unified modeling of spatial and temporal dependencies through attention mechanisms. Contemporary DiT-based video generation models (Wang et al., 2025a; Hong et al., 2022; Gao et al., 2025) therefore operate in a fully spatial-temporal manner, scaling model capacity with larger parameter counts, broader datasets, and increased computational resources.

Image Stylization With the rise of controllable image generation, the task of synthesizing a specific stylized image has also gained attention. IP-Adapter (Ye et al., 2023) trains a style reference image adapter for denoising U-Net (Ronneberger et al., 2015). InstantStyle (Wang et al., 2024) inserts the adapters into specific layers of the base model to only transfer the style of the reference image rather than content. StyleTokenizer (Li et al., 2024b) aligns style representations with the text embedding space and extracts style from the reference image. CSGO (Xing et al., 2024) designs different projectors for style and content while proposing IMAGStyle dataset. OminiControl (Tan et al., 2025) and EasyControl (Zhang et al., 2025) concatenate the conditioning signal and noisy latent to enable flexible multi-task controllable generation across diverse visual conditions.

Video Stylization Video stylization builds upon advances in image stylization while additionally addressing temporal consistency across frames. Training-free video stylization methods typically propagate appearance changes from a key frame across the sequence. AnyV2V (Ku et al., 2024) and TokenFlow (Geyer et al., 2023) leverage DDIM (Song et al., 2020) inversion and stylize the first frame to guide subsequent frames through feature propagation. More recently, FreeVis (Xu et al., 2025) integrates multiple stylized references into a pretrained image-to-video diffusion model with flow guidance to achieve temporally coherent and high-fidelity stylization. In contrast, supervised approaches learn explicit style representations from paired data. StyleCrafter (Liu et al., 2023) adopts Q-Former and injects CLIP-extracted style features into a denoiser while StyleMaster (Ye et al., 2025) extends it with more explicit global and local style extractors. Recently, DreamStyle (Li et al., 2026) proposes a unified framework for stylization with tasks of text, reference style image and first frame. TeleStyle (Zhang et al., 2026) relies on the stylized first frame as a temporal anchor and propagates its style cues across the video to maintain temporally consistent stylization.

3 METHOD

FineStyler comprises three key components. Section 3.1 introduces Content-Anchored Style Modulation for learning content-invariant and fine-grained style representations. Section 3.2 presents High-Frequency Adapted LoRA for preserving local structures and fine visual details, while Section 3.3 describes a style reference curation pipeline for constructing training reference images.

3.1 CONTENT-ANCHORED STYLE MODULATION

We train style modulation on paired samples sharing style but differing in content: $\{(I^{(a)}_{\text{org}}, I^{(a)}_{\text{sty}}), (I^{(b)}_{\text{org}}, I^{(b)}_{\text{sty}})\}$. The style branch extracts tokens from $I^{(a)}_{\text{sty}}$, while the generation branch receives $I^{(b)}_{\text{org}}$ and diffusion supervision from $I^{(b)}_{\text{sty}}$ (Fig. 2).

The style branch uses a query transformer to produce compact style tokens that condition image stylization in the generation branch. Both pairs are processed jointly, but only the style branch is updated. A reconstruction loss aligns its transformed latent with the corresponding natural image, suppressing content leakage and encouraging content-invariant style features.

Given a pretrained VAE encoder $\mathcal{E}(\cdot)$, we encode both original and stylized images into latents and apply the style encoder $\mathcal{T}_\varphi(\cdot)$ to obtain style tokens:

$$F^{(a)}_{\text{org}} = \mathcal{E}(I^{(a)}_{\text{org}}), \quad F^{(a)}_{\text{sty}} = \mathcal{E}(I^{(a)}_{\text{sty}}), \quad [Q; \hat{F}^{(a)}_{\text{sty}}] = \mathcal{T}_\varphi([Q_0; F^{(a)}_{\text{sty}}]), \quad S^{(a)} = Q. \tag{1}$$

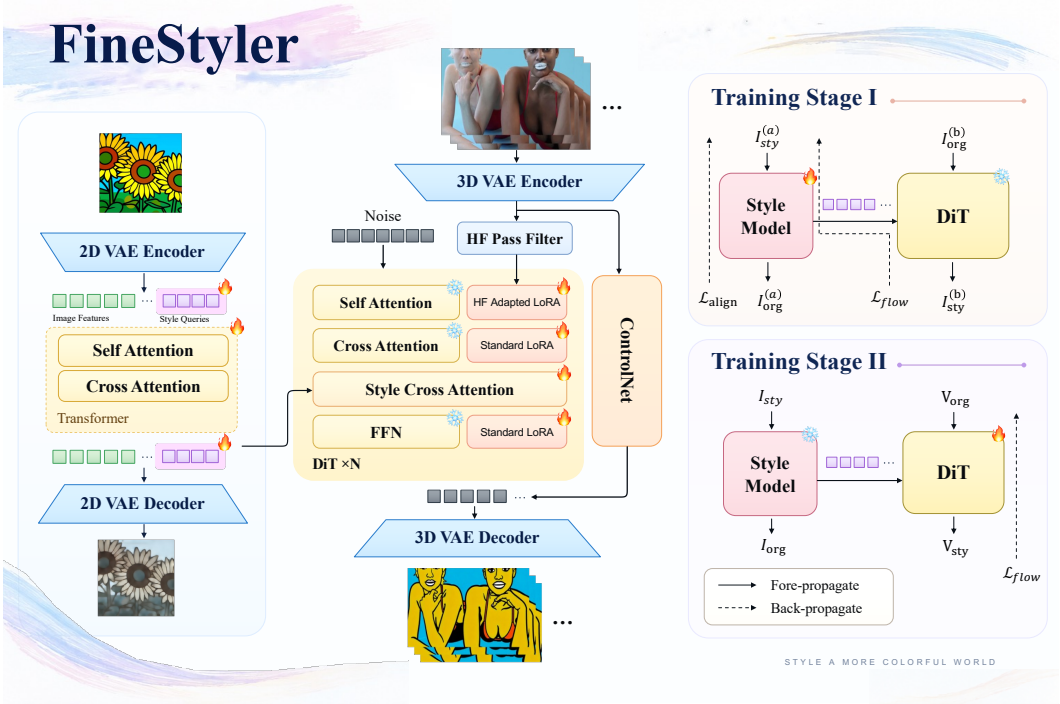

Figure 2: **Overview of FineStyler.** Left: reference-based video stylization with content preservation. Right: Stage I trains the style module; Stage II freezes it and optimizes the DiT LoRA adapters. Solid and dashed arrows denote forward and backward propagation, respectively.

Here, $\mathcal{T}_\varphi$ is the style extraction transformer with learnable queries Q_0. Its compact tokens S capture reusable style attributes, while the latent $\hat{F}^{(a)}_{\text{sty}}$ retains spatial structure for reconstruction.

To anchor the output to natural-image content and push style cues into the token pathway, the transformed latent is aligned with the original-domain latent:

$$\mathcal{L}_{align} = \|\hat{F}^{(a)}_{\text{sty}} - F^{(a)}_{\text{org}}\|_2^2. \tag{2}$$

Since $I^{(a)}_{\text{org}}$ and $I^{(a)}_{\text{sty}}$ share content, this constraint suppresses content leakage into style tokens and redundant style variation in the latent pathway, encouraging style information to concentrate in S for cross-content transfer.

For the generation branch, the extracted style tokens $S^{(a)}$ are injected into video diffusion transformer through style cross-attention, while supervision is provided by the second content stream. Let $z^{(b)}_{\text{org}}$ and $z^{(b)}_{\text{sty}}$ denote the latents of $I^{(b)}_{\text{org}}$ and $I^{(b)}_{\text{sty}}$, respectively. With ControlNet conditioning derived from the source-content stream, the flow-matching objective is defined as

$$\mathcal{L}_{flow} = \mathbb{E}_{t,\epsilon}\left[\left\| v_\theta\left(z_t, t; z^{(b)}_{\text{org}}, S^{(a)}, c_{text}\right) - \left(z^{(b)}_{\text{sty}} - \epsilon\right)\right\|_2^2\right]. \tag{3}$$

The overall objective for this stage is

$$\mathcal{L} = \mathcal{L}_{flow} + \lambda\mathcal{L}_{align}. \tag{4}$$

Here, $\mathcal{L}_{align}$ anchors content and forces style variation into the token pathway, while $\mathcal{L}_{flow}$ supervises its consistent use across different contents. Together, they stabilize optimization, prevent token collapse into weak global biases, and separate style modulation from content preservation.

3.2 High-Frequency Adapted LoRA

This stage uses curated references (Section 3.3) for video stylization training. ControlNet preserves coarse structure, but local textures and edges can drift because they depend on fine-scale feature

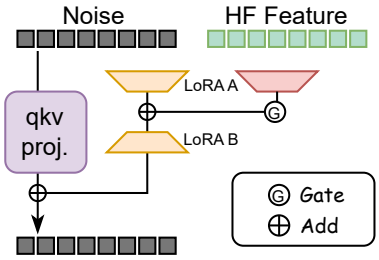

Figure 3: **High-Frequency Adapted LoRA.** LoRA-A branches project high-frequency and denoising features into a shared low-rank space. A learnable gate fuses them for shared LoRA-B updates to self-attention projections.

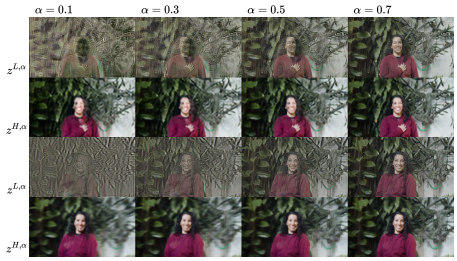

Figure 4: **Analysis of frequency fusion strategies.** For content guidance, we fuse low- and high-frequency components. In Eq. (6), α weights the low-frequency component in rows 1 and 3, and the high-frequency component in rows 2 and 4.

interactions. We separate reference latents into low- and high-frequency components and inject high-frequency cues into self-attention through a parameter-efficient adapter. This complements structural guidance and provides an adjustable trade-off between smooth appearance and crisp detail.

Given a source video latent x_{src}, we first convert it to frequency domain:

$$\mathcal{F} = FFT(x_{\text{src}}), \quad \mathcal{H}_r = 1 - \mathcal{L}_r, \tag{5}$$

where $\mathcal{L}_r$ denotes a centered low-pass mask controlled by ratio r, and $\mathcal{H}_r$ is the corresponding high-pass mask. A generalized attenuation form is

$$\mathcal{F}^{L,\alpha} = \alpha\mathcal{F} \odot \mathcal{L}_r + \mathcal{F} \odot \mathcal{H}_r, \quad \mathcal{F}^{H,\alpha} = \mathcal{F} \odot \mathcal{L}_r + \alpha\mathcal{F} \odot \mathcal{H}_r, \tag{6}$$

$$z^{L,\alpha} = IFFT(\mathcal{F}^{L,\alpha}), \quad z^{H,\alpha} = IFFT(\mathcal{F}^{H,\alpha}). \tag{7}$$

Here, α controls low- or high-frequency attenuation. As shown in Fig. 4, high-frequency emphasis preserves sharper structures and cleaner details across frames, whereas low-frequency emphasis smooths thin edges and weakens micro-textures. We therefore use a hard high-pass setting to construct $h = z^{H,\alpha}$ by removing the low-frequency cube.

We apply high-frequency adapted LoRA to self-attention $q/k/v$ projections, whose detailed framework is illustrated in Fig. 3. For query projection, we can formulate it as follows:

$$u_q = A_z^q z + g_q \odot (A_h^q h), \quad \Delta q = \frac{\alpha_{lora}}{r} B^q u_q, \quad q' = W_q z + \Delta q, \tag{8}$$

with the same form for k and v. These projections determine token matching and aggregation, controlling detail transport. The learnable gate g_q controls injection across layers and denoising steps, allowing conservative modulation at noisy steps and stronger detail emphasis later to avoid early over-sharpening. We freeze the style module and optimize LoRA parameters with flow-matching supervision from stylized targets under content guidance. Freezing prevents interference with the style interface while adaptation improves fine textures and temporal detail consistency.

3.3 Style Reference Image Curation Pipeline

Adapter training (Section 3.2) requires style-consistent, content-mismatched references, as typically provided by users. Many systems first stylize an input frame and propagate its style, adding workflow and computational cost. To enable end-to-end stylization despite the lack of references in the paired raw/stylized corpus, we use image stylization models to construct style-aligned, content-different references for each stylized video $\mathbf{x}^{sty}$.

We propose a three-step pipeline: (i) select a representative *style anchor frame* from $\mathbf{x}^{sty}$, (ii) use a high-quality image stylization model to generate diverse candidates from random content images under the same style cue, and (iii) apply lightweight automatic scoring to retain the best reference.

Style Anchor Frame Selection. For a stylized video with N frames $\{x_i\}_{i=1}^N$, we select a representative style frame via density estimation in a style-sensitive feature space. Concretely, for each frame

x_i we employ a powerful visual encoder (Oquab et al., 2023) to extract compact feature representations $F_i \in \mathbb{R}^{T\times C}$ and compute its Gram style descriptor. We then measure style representativeness by the average cosine similarity between Gram matrices using the Frobenius inner product:

$$G_i = \frac{1}{T} F_i^\top F_i, \qquad s_i \;=\; \frac{1}{N}\sum_{j=1}^{N} \frac{\langle G_i, G_j\rangle_F}{\|G_i\|_F \, \|G_j\|_F}, \tag{9}$$

where $\langle A, B\rangle_F$ denotes the Frobenius inner product. Finally, we select the frame with the highest s_i as the style anchor $x^\star$. This mode-seeking criterion selects the most typical style appearance, robust to occasional temporal style fluctuations.

Candidate Generation with Caption Guidance. We randomly sample a natural content image y from an image dataset (Wang et al., 2025c), and stylize it with strong image stylization models (Zhang et al., 2026; Wu et al., 2025) using the style anchor frame $x^\star$ and stylization instruction $\mathbf{t}^{sty}$ as the positive condition. To further prevent content leakage, we use the original video caption $\mathbf{t}^{cap}$ as a negative prompt in classifier-free guidance, encouraging the generated reference to be content-mismatched to the source video. This can be formally expressed as:

$$\hat{\epsilon}_\theta(z_t, t, \mathcal{C}_+, \mathcal{C}_-) = \epsilon_\theta(z_t, t, \mathcal{C}_-) + \omega_m(\epsilon_\theta(z_t, t, \mathcal{C}_+) - \epsilon_\theta(z_t, t, \mathcal{C}_-)), \tag{10}$$

where $\mathcal{C}_+ = \varphi(\mathcal{P}_+)$ and $\mathcal{C}_- = \varphi(\mathcal{P}_-)$ are the embeddings of the positive and negative prompts, respectively, with $\mathcal{P}_+ = (\mathbf{t}^{sty}, x^\star)$ and $\mathcal{P}_- = \mathbf{t}^{cap}$. We vary the guidance scale ω_m to obtain multiple stylization candidates of the content image y, forming a pool $\mathcal{S} = \{\tilde{s}_m\}_{m=1}^{M}$ for filtering.

Automatic Scoring and Filtering. Since the candidate pool may contain occasional failures (e.g., degraded aesthetics, weak stylization, or content leakage), we perform automatic filtering with off-the-shelf evaluators. Specifically, we score each candidate from four aspects: (i) *aesthetics* using human preference models (e.g., HPS (Wu et al., 2023), ImageReward (Xu et al., 2023)); (ii) *prompt-image alignment* using an instruction-following reward model (e.g., ORM (Guo et al., 2025)); (iii) *style consistency* with the style anchor $x^\star$ using CSD (Somepalli et al., 2024); and (iv) *content/leakage checks* by querying the existence and attributes of caption-related objects with VQA/LMMs (e.g., BLIP (Li et al., 2022), GIT (Wang et al., 2022), LLaVA (Li et al., 2024a)). We aggregate these scores with a weighted sum to select top-ranked candidates as style references $\mathbf{s}^{ref}$.

Finally, we augment each sample with curated references and obtain:

$$\mathcal{D} = \left\{(\mathbf{x}_i^{raw}, \mathbf{x}_i^{sty}, \mathbf{t}_i^{cap}, \mathbf{t}_i^{sty}, \mathbf{s}_i^{ref})\right\}_{i=1}^{|\mathcal{D}|}. \tag{11}$$

This pipeline produces content-agnostic yet style-consistent reference images, enabling end-to-end training while keeping deployment efficient and modular.

4 Experiments

Implementation Details. We build our method on top of Wan2.1-T2V-1.3B (Wan et al., 2025) as the frozen video diffusion base model, inheriting its pretrained VAE and DiT components. Unless otherwise stated, all architectural choices (latent resolution, noise schedule, and inference setup) follow the default configuration of Wan2.1. Training uses two stages with different data sources. Stage I is trained for style modulation on the OmniStyle (Wang et al., 2025c) image dataset, where we sample paired tuples of original images and its stylized counterparts with identical content under the same style label. Stage II is trained for high-frequency adapted LoRA on the Ditto (Bai et al., 2025) video dataset to adapt the video diffusion transformer to temporally coherent video stylization.

Evaluation Settings. We measure style adherence with CSD (Somepalli et al., 2024) and content preservation with source–output cosine similarity in DINOv2 (Oquab et al., 2023) feature space, which reflects semantic structure while tolerating minor photometric changes. We also report VBench (Huang et al., 2024) metrics: dynamic degree for motion dynamics, aesthetic quality for visual appeal, and subject/background consistency for temporal coherence.

Our benchmark randomly pairs 100 style references and 100 content videos into 200 test combinations. References span diverse artistic genres, visual patterns, and rendering characteristics. Videos cover single- and multi-person activities, facial motion, hand gestures, indoor/outdoor scenes, vehicles, animals, and everyday objects under varied backgrounds, viewpoints, and lighting conditions.

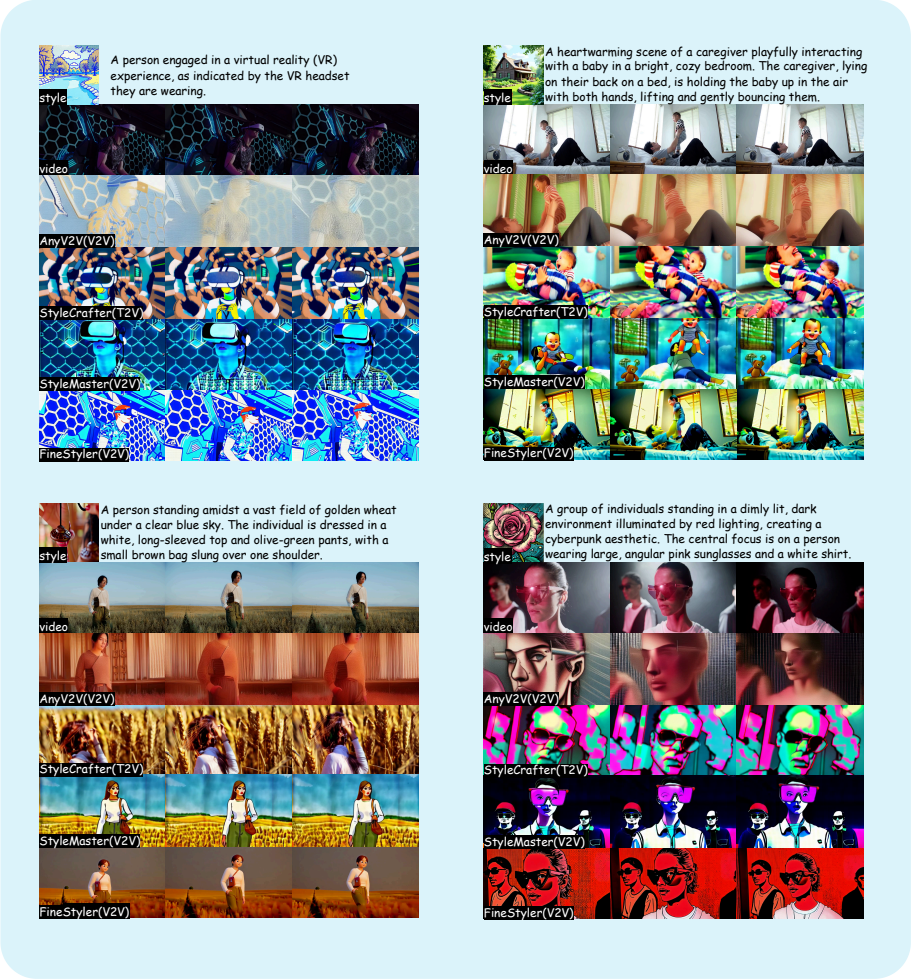

Figure 5: **Qualitative comparison of video stylization with other methods.** We compare FineStyler with StyleCrafter, StyleMaster, and AnyV2V. Our approach more faithfully transfers reference appearance while preserving source content. Here StyleCrafter uses its T2V setting.

4.1 Video Stylization Results

Quantitative Comparison. We compare with StyleCrafter, StyleMaster, and AnyV2V (Ku et al., 2024). StyleCrafter does not support an explicit content video, so we use its default style-conditioned setting with text and a style reference. FineStyler, StyleMaster, and AnyV2V are evaluated on V2V stylization with source videos and style references. AnyV2V first edits the source video's first frame and then propagates the edit through an image-to-video model with feature injection.

As shown in Table 1, FineStyler achieves higher CSD and DINO scores, indicating more faithful style transfer and stronger preservation of semantic structure and spatial layout. Compared with AnyV2V, CSD increases from 0.274 to 0.419 and DINO from 0.356 to 0.422. Dynamic degree and subject consistency also improve, further supporting the effectiveness of combining structural guidance with high-frequency injection for stable stylization. Our aesthetic quality score is slightly lower than that of StyleMaster, which we attribute to a mismatch between general visual appeal scoring and faithful style transfer under explicit content constraints.

Qualitative Comparison. In Fig. 5, StyleCrafter and StyleMaster show degraded fidelity and inconsistent reference-style patterns; StyleMaster also alters object boundaries and scene layouts. AnyV2V shows incomplete style transfer and content changes: the wheat-field example shifts to an indoor scene, while the sunglasses example varies in appearance across frames. FineStyler better captures reference appearance while preserving source subjects and scene structure across frames.

	StyleCrafter(T2V)	StyleMaster	AnyV2V	FineStyler
CSD score ↑	0.334	0.352	0.274	**0.419**
DINO score ↑	–	0.286	0.356	**0.422**
Dynamic Degree ↑	0.485	0.460	0.447	**0.780**
Aesthetic Quality ↑	0.554	**0.599**	0.528	0.535
Subject Consistency ↑	0.952	0.948	0.827	**0.953**
Background Consistency ↑	**0.958**	0.952	0.943	0.947

Table 1: **Quantitative comparison with state-of-the-art video stylization methods.** FineStyler leads in reported CSD and DINO scores. Dashes indicate unreported or inapplicable metrics.

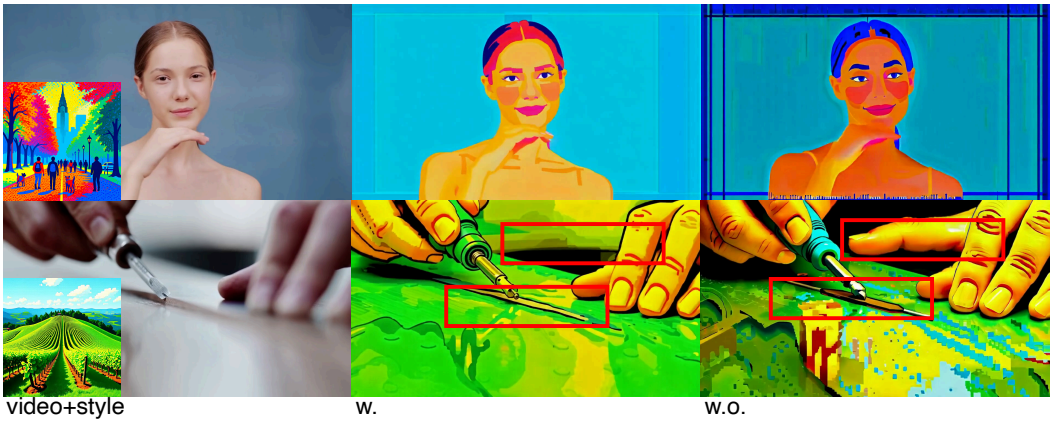

Figure 6: **Ablation on High-Frequency Adapted LoRA.** Compared with standard LoRA (*w.o.*), HF-LoRA (*w.*) produces cleaner renderings in row one and more accurate details in row two.

4.2 Ablation Studies

Style Modulation. We ablate Content-Anchored Style Modulation (CA-SM) in Table 2. Adding CA-SM improves both CSD and DINO scores while maintaining subject consistency, suggesting that the learned style representation better captures reference appearance without sacrificing source-content preservation. This is consistent with the qualitative results in Fig. 7, where CA-SM preserves finer line-level structures, more delicate textures, richer local shading patterns, and color transitions. Compared with the variant without CA-SM, the resulting stylization is less washed out or overly uniform, indicating that content anchoring helps retain fine-grained style cues while reducing content-dependent information in the style representation.

High-Frequency Adapted LoRA. Replacing standard LoRA with High-Frequency Adapted LoRA in self-attention substantially improves the DINO score (Table 2), demonstrating its effectiveness in preserving source-video structure during stylization. As shown in Fig. 6, High-Frequency Adapted LoRA better maintains thin object boundaries, local geometry, and subtle spatial details that are easily distorted by coarse structural conditioning. The resulting videos exhibit cleaner contours and more stable fine structures across frames, supporting our motivation of explicitly injecting high-frequency source cues into the denoising process. Together with CA-SM, which primarily improves

	CSD score ↑	DINO score ↑	Subject Consistency ↑
baseline	0.352	0.286	0.948
+ CA-SM	0.408	0.413	0.952
+ HF-LoRA	0.385	**0.438**	0.915
+ CA-SM + HF-LoRA	**0.419**	0.422	**0.953**

Table 2: **Quantitative ablation results.** CA-SM means Content-Anchored Style Modulation, and HF-LoRA refers to High-Frequency Adapted LoRA. StyleMaster is our baseline.



Figure 7: **Ablation on style modulation.** Content-Anchored Style Modulation (*w.*) preserves more line-level detail in the first row and stronger diffusion-like shading in the second than the variant without it (*w.o.*). We compare the two variants in a text-to-image setting using their Stage I weights.

fine-grained style representation, High-Frequency Adapted LoRA provides complementary structural guidance for achieving faithful stylization with stronger content preservation.

4.3 User Study

In a user study with 18 participants and 10 questions per criterion, we evaluated four complementary aspects of video stylization: style faithfulness, source structure preservation, dynamics quality, and overall visual quality. Participants consistently preferred FineStyler across these criteria, favoring its ability to reproduce reference textures, color palettes, and brushwork while preserving source structure and producing natural, temporally coherent motion (Table 3). Since StyleCrafter is a T2V method without source-video input, it was excluded from the source structure preservation criterion. Appendix F provides further details on the evaluation protocol.

Method	Style Faithfulness ↑	Source Structure Preservation ↑	Dynamics Quality ↑	Overall Visual Quality ↑
StyleMaster	6.11%	37.22%	12.78%	29.44%
StyleCrafter	15.56%	–	1.11%	1.11%
AnyV2V	19.44%	1.67%	2.78%	5.56%
FineStyler (Ours)	**58.89%**	**61.11%**	**83.33%**	**63.89%**

Table 3: **User study results.** Selection percentages (higher is better). A dash indicates that StyleCrafter (T2V) was excluded from the source-structure-preservation criterion.

5 Conclusion

We introduce FineStyler, a unified framework for high-fidelity and temporally stable video stylization from a single reference image. FineStyler addresses two key limitations of prior approaches: suboptimal style representations that overlook fine-grained appearance cues and coarse structural conditioning that struggles to preserve local details over time. **Content-Anchored Style Modulation** learns content-invariant style representations by aligning transformed features with a natural-image feature domain, enabling finer style transfer while suppressing reference-content leakage. **High-Frequency Adapted LoRA** explicitly injects high-frequency cues from the input video into DiT self-attention, strengthening the preservation of thin boundaries, local geometry, and fine spatial structures throughout denoising. In addition, our style reference curation pipeline constructs style-aligned yet content-mismatched references, providing supervision that better matches the inference setting. Extensive quantitative and qualitative experiments, together with user studies, demonstrate improved style faithfulness, content preservation, and temporal quality over prior baselines, while producing sharper structures and more faithful fine-grained stylization with stronger overall quality.

AI USE STATEMENT

Large language models were used to assist with manuscript revision, code development, and figure and table refinement. Generative models were also used to construct synthetic training data, as described in the methodology and appendices. All AI-assisted outputs, code, and synthetic data were reviewed and validated by the authors. The authors take full responsibility for the content, correctness, and integrity of this work.

REFERENCES

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, et al. Scaling instruction-based video editing with a high-quality synthetic dataset. *arXiv preprint arXiv:2510.15742*, 2025.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.

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

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Mengtian Li, Jinshu Chen, Songtao Zhao, Wanquan Feng, Pengqi Tu, and Qian He. Dreamstyle: A unified framework for video stylization. *arXiv preprint arXiv:2601.02785*, 2026.

Wen Li, Muyuan Fang, Cheng Zou, Biao Gong, Ruobing Zheng, Meng Wang, Jingdong Chen, and Ming Yang. Styletokenizer: Defining image style by a single instance for controlling diffusion models. In *European Conference on Computer Vision*, pp. 110–126. Springer, 2024b.

Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Yujiu Yang, and Ying Shan. Stylecrafter: Enhancing stylized text-to-video generation with style adapter. *arXiv preprint arXiv:2312.00330*, 2023.

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

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024.

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

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14940–14950, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 3(4):6, 2025a.

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.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025b.

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

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

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

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.

Jiacong Xu, Yiqun Mei, Ke Zhang, and Vishal M Patel. Freevis: Training-free video stylization with inconsistent references. *arXiv preprint arXiv:2510.01686*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 15903–15935, 2023.

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

Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2630–2640, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

Shiwen Zhang, Xiaoyan Yang, Bojia Zi, Haibin Huang, Chi Zhang, and Xuelong Li. Telestyle: Content-preserving style transfer in images and videos. *arXiv preprint arXiv:2601.20175*, 2026.

Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19513–19524, 2025.

A Base Model and Overall Structure

Base Model We build our method on top of *Wan2.1-T2V-1.3B* (Wan et al., 2025) as the video diffusion backbone. Specifically, we inherit its pretrained video VAE for encoding video frames into latents and the DiT denoiser for conditional generation. Each DiT block has Self-Attention, Cross-Attention and FFN module.

Conditioning Interface We integrate our High-Frequency Adapted LoRA (HF-LoRA) into every DiT layer by augmenting the self-attention blocks. Concretely, HF-LoRA is attached to the $q/k/v$ projection layers to modulate the corresponding linear mappings in a parameter-efficient manner. Unless otherwise specified, we use a LoRA scaling factor $\alpha_{\text{lora}} = 128$ and a low-rank dimension $r = 128$ throughout all layers. Since it is unclear a priori which DiT layers are most sensitive to high-frequency cues, we introduce a learnable gate to adaptively weight the HF-LoRA-A branch across layers, allowing the model to determine how much high-frequency information should be injected at each layer during training and inference. For structural conditioning, we initialize our ControlNet from the pretrained Gray Tile ControlNet released by StyleMaster (Ye et al., 2025), and use it as the default ControlNet backbone in all experiments.

B Comparison Methods

StyleCrafter StyleCrafter (Liu et al., 2023) proposes a generic *style control adapter* that augments a pretrained text-to-video diffusion model to enable reference-image-guided stylized generation. It encourages content–style disentanglement by removing explicit style descriptions from the text prompt and extracting style cues solely from the reference image, together with a scale-adaptive fusion module to balance text-driven content and image-driven style features. Note that StyleCrafter is primarily designed for *style-guided T2V generation* rather than video-to-video translation, and thus it does not take a content video as direct input in its default setting.

StyleMaster StyleMaster (Ye et al., 2025) is a DiT-based framework for reference-guided video stylization. It designs a *local–global* style extraction pipeline for both style features. The extracted style cues are injected through a decoupled style cross-attention, while a lightweight *motion adapter* is introduced to bridge the image-to-video gap and improve temporal quality. For video stylization (V2V) in particular, StyleMaster further proposes a *gray-tile ControlNet* as an accessible yet precise structural guidance, where grayscaling the tile condition mitigates color interference during style transfer. Because StyleMaster shares a similar overall design with FineStyler—both adopting an external style module plugged into a frozen diffusion backbone—we choose StyleMaster as the baseline for our ablation studies.

AnyV2V AnyV2V (Ku et al., 2024) is a tuning-free video editing framework that combines first-frame image editing with video generation. An image editor first modifies the source video's first frame, and an image-to-video diffusion model propagates the edit using features from the source video. It supports reference-based style transfer without task-specific video-model fine-tuning. AnyV2V serves as a V2V baseline in our quantitative and qualitative comparisons and user study.

C Style Reference Image Curation Pipeline

Goal and Setup. Our training in Stage II requires a *content-mismatched yet style-consistent* reference image for each stylized target video. However, the paired raw/stylized video corpus does not directly provide such references. We therefore curate additional style references in the image domain, leveraging the controllability of modern image stylization models.

Implementation Details for Reference Generation. In our implementation, we instantiate an image stylization pipeline built on a Qwen-based image editor with LoRA adapters (Zhang et al., 2026; Wu et al., 2025). The base model is `Qwen-Image-Edit-2509` (Wu et al., 2025), combined with a style-specific LoRA and a 4-step acceleration LoRA for efficient sampling. For each training instance, we: (i) extract a representative stylized frame as style anchor, (ii) randomly sample a natural content image from a large external content pool (Wang et al., 2025c), and (iii) generate a style reference under fixed inference settings. Unless otherwise noted, we use 4 denoising steps, a minimum edge length of 1024, and deterministic seeding by sample index. The positive instruction is

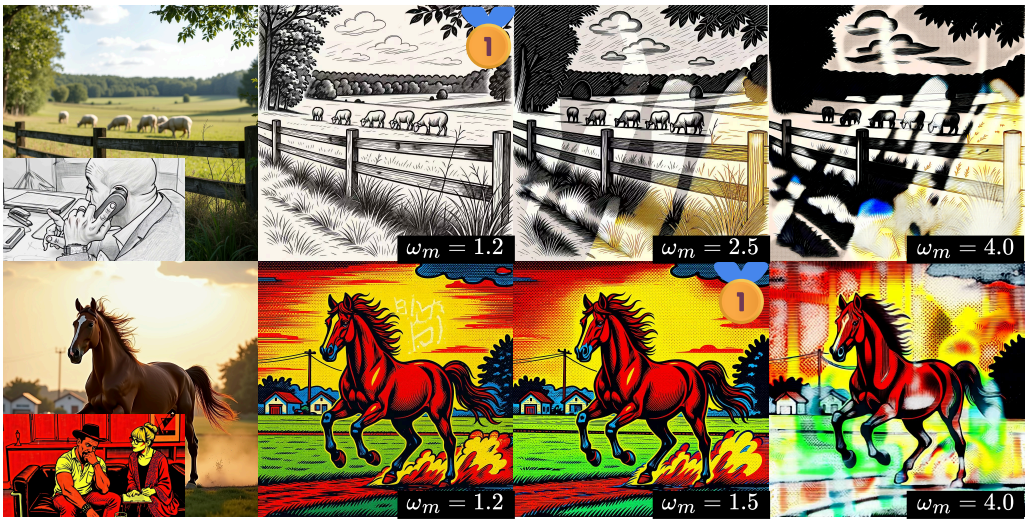

Figure 8: **Negative CFG scale sweep for reference generation.** We stylize the same content image y under the same style cue $(\mathbf{t}^{sty}, x^{\star})$ while varying the negative CFG scale ω_m (with $\mathbf{t}^{cap}$ as the negative prompt). Smaller ω_m may preserve caption-related semantics (content leakage), whereas larger ω_m can introduce artifacts or reduce stylization quality. Our pipeline generates a candidate pool via sweeping ω_m and selects the final reference through automatic scoring.

kept constant (preserve source structure while applying reference style), while the original caption is used as negative condition to suppress content leakage. To obtain robust candidates, we sweep the guidance scale as $\omega_m \in \{1.2, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$, producing a candidate pool for ranking.

Automatic Scoring and Final Metric Choice. For automatic filtering, we score each candidate reference using complementary signals from aesthetics, instruction faithfulness, caption consistency, and style similarity. Concretely, we combine HPS for aesthetic preference (Wu et al., 2023), ORM for instruction-following quality (Guo et al., 2025), GIT-based VQA signals for caption-level consistency (Wang et al., 2022), and CSD (Somepalli et al., 2024) for style consistency. The text condition is formed by composing the caption with the style instruction, and all candidates are evaluated on strictly aligned sample intersections to avoid comparison bias. As the dedicated style term, CSD uses a CLIP ViT-L/14-based encoder (Radford et al., 2021) to extract normalized style embeddings, and computes cosine similarity between the style anchor and each candidate. The final ranking is obtained by a weighted aggregation of these normalized scores, which favors references that preserve target rendering attributes (e.g., texture, color palette, and brushwork) while remaining content-mismatched. A qualitative visualization of generated references and the corresponding scoring trend across guidance settings is shown in Fig. 8, illustrating how the filtering stage retains style-faithful and content-disentangled candidates.

D Training Details

More Implementation Details on Stage I. In Stage I, we optimize only the style modulation branch while keeping the pretrained diffusion backbone fixed. In practice, style-related conditioning layers are updated, whereas the backbone denoising parameters remain frozen. This design preserves the pretrained content prior and concentrates adaptation capacity on style representation learning.

Each training item is built from two pairs that share the same style ID but use different content IDs, matching $\{(I^{(a)}_{\text{org}}, I^{(a)}_{\text{sty}}), (I^{(b)}_{\text{org}}, I^{(b)}_{\text{sty}})\}$, as shown in Fig. 9. The first pair is used to extract style features, while the second pair provides stylization supervision under different content. During optimization, style features from $I^{(a)}_{\text{sty}}$ condition the generation process for the second stream, where the model predicts the stylized target corresponding to $I^{(b)}_{\text{org}}$. We additionally monitor reconstruction

consistency between the transformed latent of the first stylized sample and its original counterpart to regularize content anchoring.

Overall, this stage learns a content-invariant style interface for subsequent stylization.

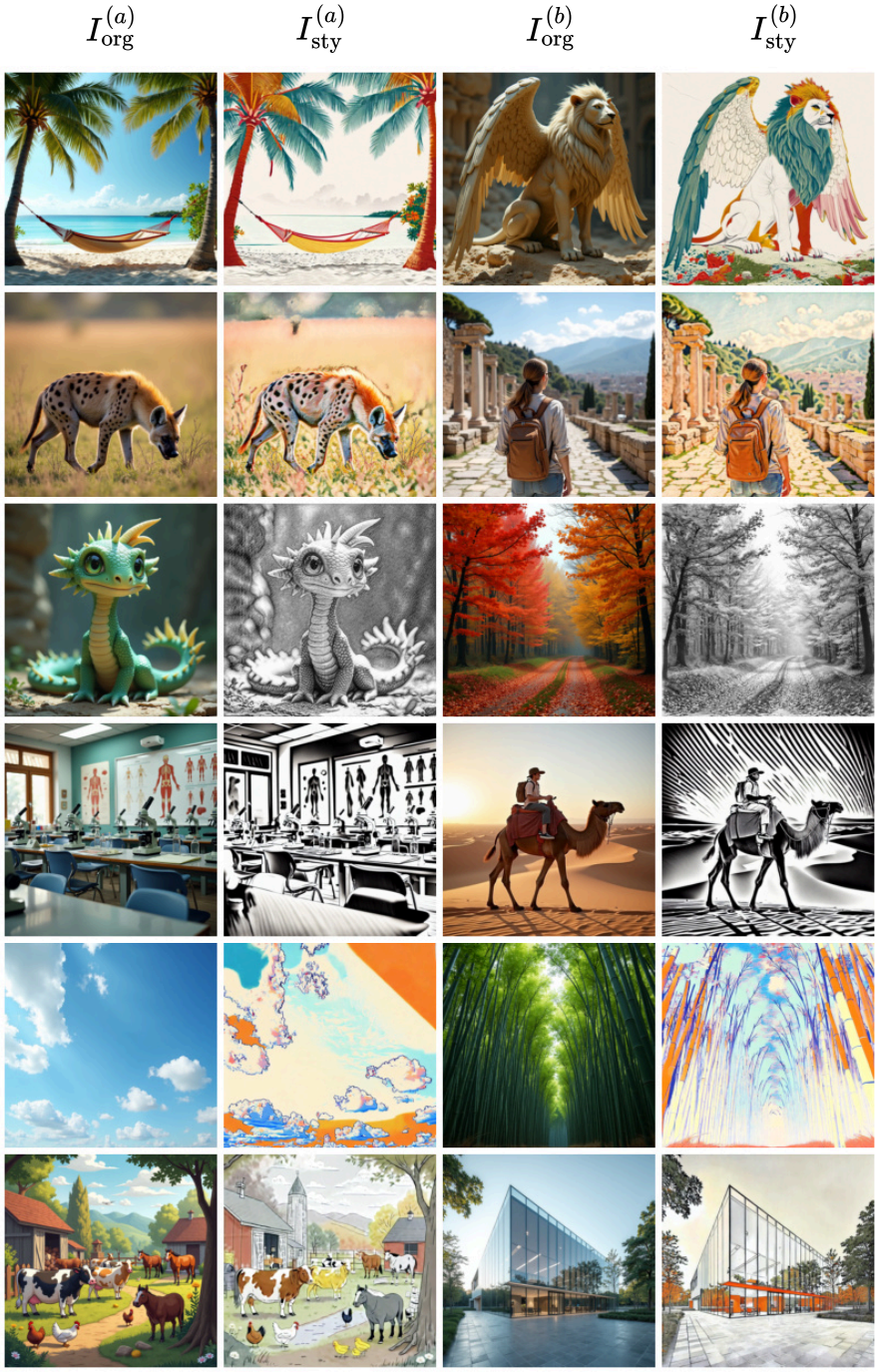

Figure 9: **Data pairs used in Training Stage I.** We construct the four-sample data pair every batch.

More Implementation Details on Stage II. In Stage II, we perform supervised training on paired raw and stylized videos while explicitly incorporating our curated style reference images.

In our implementation, mixed-precision training, cosine warmup scheduling, and gradient checkpointing are used to stabilize optimization and improve memory efficiency.

E Additional Qualitative Results

We provide additional qualitative comparisons on diverse videos and style references to complement the main paper. As shown in Fig. 10, our method consistently produces visually pleasing stylized videos with faithful texture transfer while preserving key content structures and temporal coherence. We also include challenging cases, where our approach better maintains local details and reduces temporal flickering and over-smoothing. These additional examples further demonstrate the robustness and generalization of our model across various styles and scenes.

F User Study Details

Participants and protocol. We collected responses from 18 participants to compare FineStyler, StyleMaster, StyleCrafter, and AnyV2V. The study has four sections of 10 questions each, evaluating *style faithfulness*, *source video structure preservation*, *dynamics quality*, and *overall visual quality*. For each question, participants selected the output best satisfying the given criterion.

Evaluation criteria. Style faithfulness concerns reference texture, palette, and brushwork; source video structure preservation concerns source subjects, structure, and layout. Dynamics quality concerns meaningful, natural motion and temporal coherence, including flicker and abrupt changes, without rewarding excessive motion or frozen outputs. Overall visual quality concerns the overall perceptual quality of the generated video. Since StyleCrafter is a T2V method and does not take a source video as input, it is excluded from the structure preservation section.

Aggregation. Each criterion has 180 responses (18 participants $\times$ 10 questions). We report selection percentages for each method; higher values indicate stronger preference (Table 3).

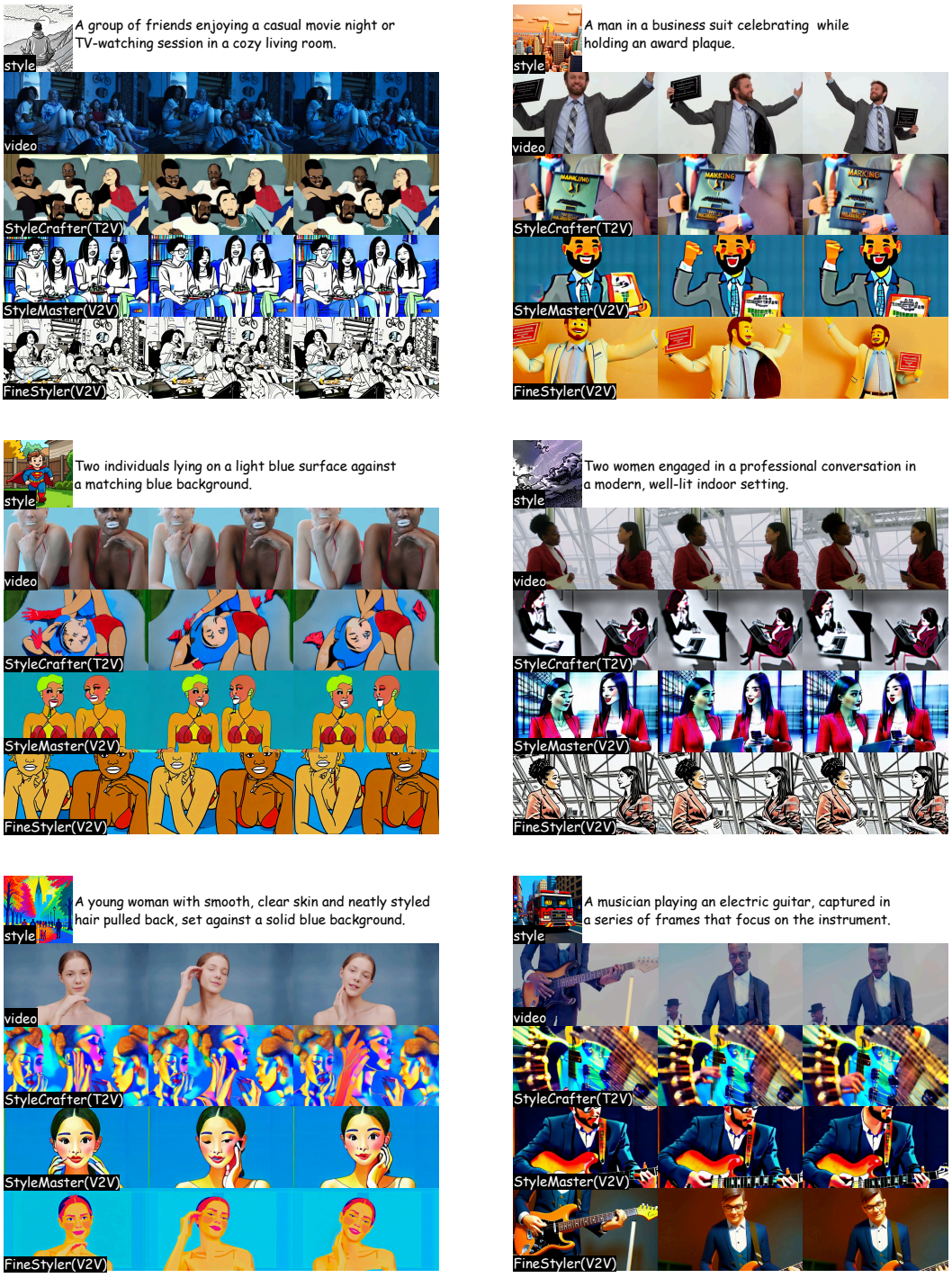

Figure 10: **Additional qualitative results.** We compare FineStyler with other SOTA methods.