FACT: FREQUENCY AND ATTENTION CONTROL FOR TRAINING-FREE VIDEO STYLIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-guided video stylization transfers the appearance of a reference image to a source video while preserving source structure and temporal consistency. Recent training-free approaches leverage pretrained diffusion models to avoid costly training, yet they can lose fine-grained source details during style transfer and produce inconsistent appearance across frames. We analyze the frequency components of intermediate denoising latents and find that low-frequency components evolve gradually while high-frequency components change more rapidly across steps. Motivated by this finding, we propose FACT, a training-free framework that combines latent frequency modulation with attention control. At each denoising step, frequency modulation attenuates the low-frequency component of the current latent while retaining its high-frequency component, preserving fine spatial details throughout stylization. Logarithmic Length Attention Scaling (LAS) adaptively adjusts content-to-style attention strength across U-Net feature scales, balancing style injection with source-structure preservation. Confidence-Weighted Temporal Attention (CTA) guides cross-frame attention along tracked point correspondences, weighting interactions by tracking confidence and temporal distance to improve frame-to-frame consistency. Experiments show that FACT outperforms existing methods in reference-style transfer, source-structure preservation, and temporal consistency.

1 INTRODUCTION

Reference-guided video stylization applies the appearance of a reference image to a source video while preserving its scene structure and motion. Pretrained diffusion models offer a natural backbone for this task, as their denoising process can be steered to incorporate a reference appearance without task-specific training (Rombach et al., 2022; Chung et al., 2024). A successful result requires two properties: recognizable content and fine details within each frame, and coherent appearance across frames.

Existing methods pursue this goal through learned conditioning or inference-time attention manipulation (Liu et al., 2024; Chung et al., 2024; Shi et al., 2024; Song et al., 2026; Xu et al., 2026). These approaches improve style control and temporal coherence, yet the transferred appearance can still lose fine-grained source details or fluctuate across frames. Figure 1 illustrates such cases on three consecutive video frames.

To understand how spatial details evolve during stylization, we analyze the frequency components of intermediate denoising latents. Figure 2 shows these components at selected denoising steps. We find that low-frequency components evolve gradually, while high-frequency components change more rapidly. The high-frequency views capture contours and fine texture; the low-frequency views retain broad color and coarse scene structure.

Motivated by this observation, we propose FACT, a training-free framework for reference-guided video stylization that combines latent frequency modulation with attention control. At each denoising step, frequency modulation attenuates the low-frequency component of the current latent while retaining its high-frequency component, preserving fine spatial details throughout stylization. Logarithmic Length Attention Scaling (LAS) adaptively adjusts content-to-style attention strength across U-Net feature scales, balancing style injection with source-structure preservation. Confidence-Weighted Temporal Attention (CTA) guides cross-frame attention along tracked point

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.

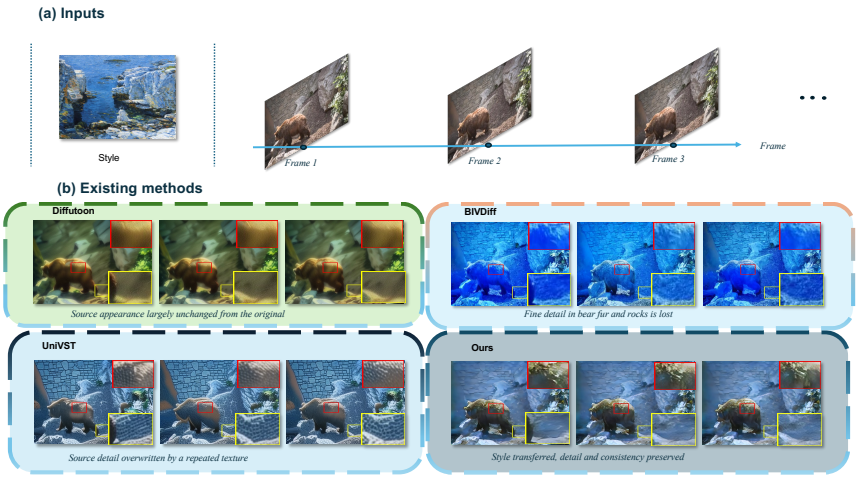

Figure 1: Reference-guided stylization over three consecutive frames. Enlarged regions compare source-detail preservation; the frame sequences show appearance consistency across Diffutoon, BIVDiff, UniVST, and FACT.

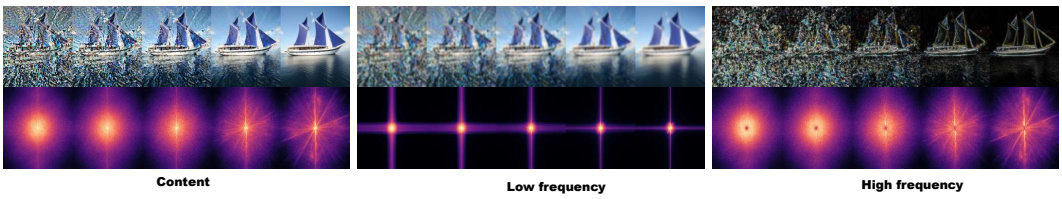

Figure 2: Frequency decomposition during denoising. From left to right, the groups show evolving content, low-frequency components, and high-frequency components. Within each group, the top row shows spatial images, the bottom row their frequency-domain views, and steps progress from left to right.

correspondences (Karaev et al., 2024), weighting interactions by tracking confidence and temporal distance to improve frame-to-frame consistency.

Experiments show that FACT outperforms existing methods in reference-style transfer, source-structure preservation, and temporal consistency. Ablations confirm the contribution of each module.

Our contributions are:

(1) We propose FACT, a training-free framework that combines latent frequency modulation with attention control for reference-guided video stylization. Frequency modulation attenuates low-frequency latent components at each denoising step, preserving fine spatial details throughout stylization.

(2) We propose Logarithmic Length Attention Scaling (LAS), which adaptively adjusts content-to-style attention across feature scales, balancing style transfer with source-structure preservation.

(3) We introduce Confidence-Weighted Temporal Attention (CTA), which guides cross-frame attention along tracked correspondences weighted by confidence and temporal distance, improving appearance consistency across frames.

2 Related Work

2.1 Image Style Transfer

Reference-guided image style transfer aims to reproduce a style image's appearance while retaining the content image's structure. Feature-statistics methods match channel-wise moments or covariance matrices between content and style representations, as in AdaIN and WCT (Huang & Belongie, 2017; Li et al., 2017). EFDM aligns full empirical feature distributions (Zhang et al., 2022a), while CAST introduces contrastive learning to improve cross-domain transfer (Zhang et al., 2022b). SANet, AdaAttN, and StyTr2 take a different route and use attention mechanisms to relate content and style features (Park & Lee, 2019; Liu et al., 2021; Deng et al., 2022).

Diffusion-based methods leverage pretrained generative models for stylization. StyleID retains content queries and injects style-derived keys and values into self-attention within a frozen diffusion model (Chung et al., 2024); Z* reweights attention to balance style and content (Deng et al., 2024). InstantStyle conditions text-to-image generation on a reference appearance (Wang et al., 2024). StyleSSP attenuates low-frequency components of a DDIM-inverted content latent to construct the denoising starting point (Xu et al., 2025). Our frequency modulation differs in that it acts on the current latent at every denoising step, not only at initialization. When applied to video frames independently, these image methods do not link corresponding points across time.

2.2 Video Style Transfer

Video style transfer adds an extra requirement: the stylized appearance must stay coherent as content moves between frames. Early approaches used optical flow to connect stylized frames (Ruder et al., 2016) or learned to stabilize the outputs of image-level models (Lai et al., 2018). Recent diffusion-based methods tackle both style conditioning and temporal coherence. StyleCrafter learns a reference-augmented adapter for style-conditioned video generation (Liu et al., 2024). Diffutoon combines structure guidance with a consistency module for toon shading (Duan et al., 2024).

BIVDiff bridges image and video diffusion to produce temporally coherent edits without additional training (Shi et al., 2024). UniVST combines latent- and attention-level transfer with optical-flow-based temporal smoothing (Song et al., 2026), while FreeViS uses stylized references, high-frequency compensation, and motion cues for training-free video stylization (Xu et al., 2026). These methods each improve certain aspects of style control or temporal coherence, but balancing source-detail preservation with consistent cross-frame appearance remains difficult. FACT addresses both through scale-adaptive style attention, latent frequency modulation at each denoising step, and attention guidance along tracked cross-frame correspondences, all without additional training.

3 Method

To address the detail loss and temporal inconsistency in existing training-free stylization methods, we propose FACT, which operates on a frozen pretrained diffusion model during inference. The overall pipeline is shown in Figure 3. FACT contains three components: latent frequency modulation (Section 3.3) to preserve spatial detail, Logarithmic Length Attention Scaling (Section 3.2) to balance style injection with source-structure preservation, and Confidence-Weighted Temporal Attention (Section 3.4) to improve cross-frame consistency.

3.1 Preliminary

Diffusion models progressively add Gaussian noise to data and learn to reverse this corruption through denoising (Ho et al., 2020). We use pretrained Stable Diffusion (SD) v1.5, which implements latent diffusion (Rombach et al., 2022). Its VAE encoder maps an input image to a clean latent z_0, and its U-Net predicts the noise in a latent z_t at timestep t. The Gaussian forward marginal is

$$q(z_t \mid z_0) = \mathcal{N}\big(z_t; \sqrt{\bar{\alpha}_t}\, z_0, (1 - \bar{\alpha}_t) I\big)\,, \tag{1}$$

where $\bar{\alpha}_t$ is the cumulative signal-retention coefficient and $\bar{\alpha}_0 = 1$. Thus, a larger t corresponds to a noisier latent, while the pretrained denoiser predicts the noise needed for the reverse process.

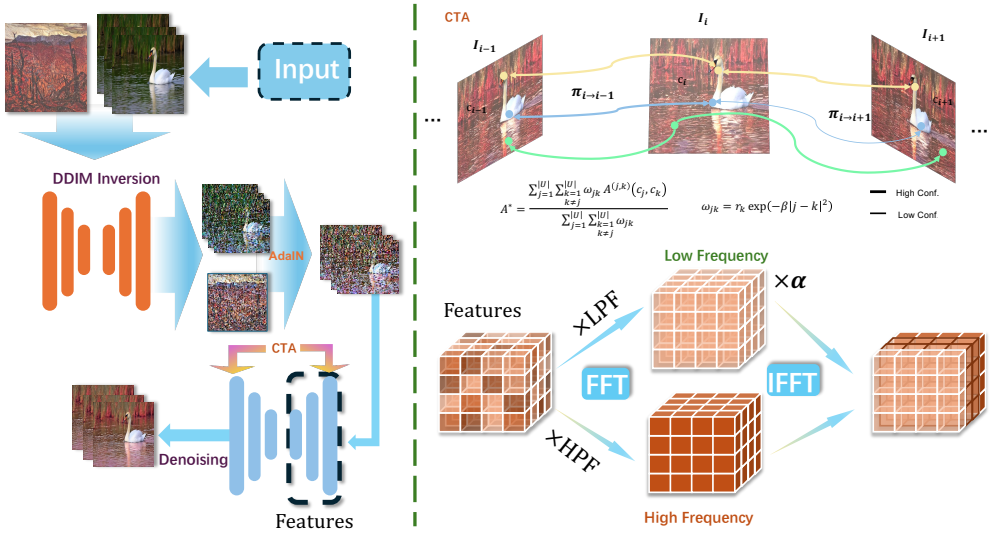

Figure 3: FACT overview. Separate DDIM inversions and latent AdaIN initialize denoising; frequency modulation, LAS, and CTA control spatial detail, content-to-style attention, and tracked temporal attention.

To obtain a denoising starting point from a real image, we use DDIM inversion (Song et al., 2021). DDIM admits a deterministic denoising trajectory; inversion approximately traverses it in the opposite direction. Starting with $\hat{z}_0 = z_0$, the ascending-time update is

$$\hat{z}_{t+1} = \sqrt{\frac{\bar{\alpha}_{t+1}}{\bar{\alpha}_t}}\,\hat{z}_t + \sqrt{\bar{\alpha}_{t+1}}\left(\sqrt{1/\bar{\alpha}_{t+1} - 1} - \sqrt{1/\bar{\alpha}_t - 1}\right)\epsilon_\theta(\hat{z}_t, t), \tag{2}$$

where ϵ_θ is the frozen SD noise predictor and t indexes consecutive points on the chosen DDIM schedule. Applying this inversion to each content frame and to the reference style image yields the terminal latents at T used to initialize stylization, as illustrated in Figure 3.

3.2 Style Injection and Logarithmic Attention Scaling

We apply DDIM inversion to each content frame and to the reference style image separately. Following StyleID (Chung et al., 2024), we then match each content latent's channel-wise mean and standard deviation to those of the style latent via AdaIN (Huang & Belongie, 2017). The resulting latent combines the content frame's spatial layout with the style statistics and serves as the starting point for DDIM denoising.

During denoising, we pair content-preserving queries with style-derived keys and values in self-attention to transfer local appearance. Content and reference-style images can have dissimilar local features, so pairing their queries and keys may reduce the spread of similarity logits. This can yield a diffuse attention map after softmax (Chung et al., 2024). StyleID addresses this with a fixed logit-scaling coefficient $\lambda = 1.5$:

$$\mathrm{Attn}_\lambda\left(\tilde{Q}_t^{cs}, K_t^s, V_t^s\right) = \mathrm{softmax}\left(\frac{\lambda \tilde{Q}_t^{cs}(K_t^s)^\top}{\sqrt{d}}\right)V_t^s, \qquad \lambda = 1.5. \tag{3}$$

Here $\tilde{Q}_t^{cs}$ is the content-preserving query, K_t^s and V_t^s are the style keys and values, and d is the per-head feature dimension.

Attention blocks at different U-Net scales contain different numbers of spatial positions. Chiang & Cholak (2022) proposed scaling attention logits by the logarithm of sequence length, which in

row-wise matrix notation is

$$\mathrm{Attn}_{\log n}(Q,K,V) = \mathrm{softmax}\left(\frac{(\log n)QK^\top}{\sqrt{d}}\right)V. \tag{4}$$

The factor $\log n$ grows with sequence length and can make the softmax map overly concentrated when n is large.

We propose Logarithmic Length Attention Scaling (LAS) for content-to-style attention. Let n be the number of style-key positions at the current U-Net scale and m_{ref} the spatial sequence length of the input image. We set

$$\lambda = \log_{m_{\mathrm{ref}}} n \tag{5}$$

and multiply the content-to-style logits by λ before softmax:

$$A_t^{cs} = \mathrm{softmax}\left(\frac{\lambda \tilde{Q}_t^{cs}(K_t^s)^\top}{\sqrt{d}}\right), \qquad Y_t^{cs} = A_t^{cs}V_t^s. \tag{6}$$

Thus, LAS adjusts the attention map according to the current feature scale during denoising. In the example in Figure 7, LAS produces a clearer stylized result that retains recognizable content while capturing the reference appearance. A derivation of this scaling factor from an entropy-invariance argument is given in Appendix B.

3.3 Latent Frequency Modulation

To keep content details recognizable after stylization, we modulate the current latent at each DDIM denoising step in the framework of Figure 3. For each frame and latent channel, let $f \in \mathbb{R}^{M\times N}$ denote its spatial latent slice. Its two-dimensional discrete Fourier transform, computed by FFT, is

$$F(u,v) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\exp\Big[-\mathrm{i}2\pi\left(\frac{um}{M}+\frac{vn}{N}\right)\Big], \tag{7}$$

where $f(m,n)$ is the latent value at spatial location (m,n) and $F(u,v)$ its Fourier coefficient at frequency (u,v). We apply a high-pass filter (HPF) and a Gaussian low-pass filter (LPF) to obtain

$$F^{\mathrm{H}} = H_{\mathrm{HP}} \odot F, \qquad F^{\mathrm{L}} = H_{\mathrm{LP}} \odot F, \tag{8}$$

where H_{LP} is the low-pass frequency mask and $H_{\mathrm{HP}} = 1 - H_{\mathrm{LP}}$ is its complementary high-pass mask.

We scale the low-frequency component by $0 < \alpha < 1$, retain the high-frequency component, and apply the inverse Fourier transform to obtain the modulated latent slice:

$$f^{\mathrm{FM}} = \mathcal{F}^{-1}\big(F^{\mathrm{H}} + \alpha F^{\mathrm{L}}\big). \tag{9}$$

This reduces the relative contribution of low-frequency variation while retaining the high-frequency component for denoising. As shown in Figure 4, frequency modulation preserves more recognizable details in the stylized content in this example.

3.4 Confidence-Weighted Temporal Attention

To preserve consistency across stylized frames, we identify image patches corresponding to the same region in nearby frames. We use the point-tracking framework CoTracker (Karaev et al., 2024), which follows points from one frame to their positions in subsequent frames to establish spatiotemporal correspondences. Given a point c_i in frame I_i, CoTracker finds its corresponding point $c_j = \pi_{i\to j}(c_i)$ in frame I_j (Figure 3). Each trajectory yields corresponding point pairs (c_j, c_k) across frames j and k. Here, $A^{(j,k)}$ is the post-softmax attention map between these frames, and $A^{(j,k)}(c_j, c_k)$ is its value at the tracked point pair. We combine these map values, giving greater influence to pairs with higher tracking confidence and shorter temporal distance:

$$\omega_{jk} = r_k \exp\left(-\beta|j-k|^2\right), \qquad j,k \in U, \quad j \neq k, \tag{10}$$

Here, $r_k \in [0,1]$ is the tracking confidence at c_k, and $\beta > 0$ controls temporal decay. The set U contains all frames where the trajectory is visible.

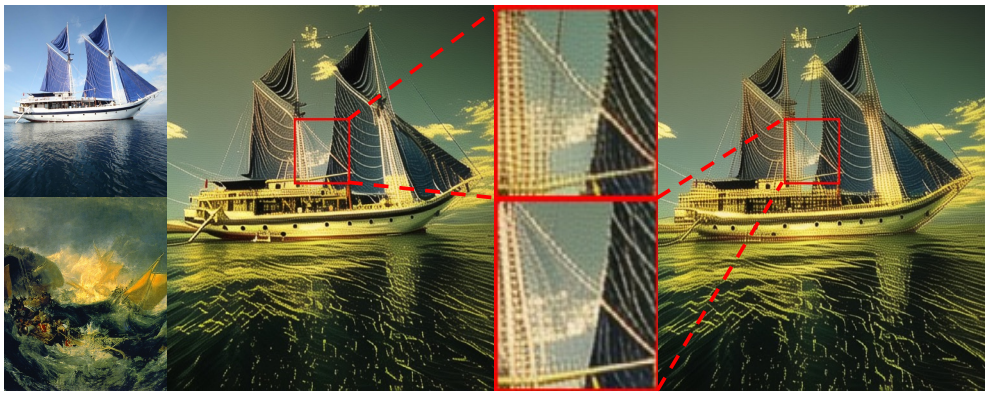

Figure 4: Effect of latent frequency modulation on a stylized frame. The content frame and reference style image are shown on the left, followed by results without and with frequency modulation. The enlarged views compare details in the same sail region.

Their weighted mean gives a consensus attention-map value:

$$A^* = \frac{\sum_{j \in U} \sum_{\substack{k \in U \\ k \neq j}} \omega_{jk} A^{(j,k)}(c_j, c_k)}{\sum_{j \in U} \sum_{\substack{k \in U \\ k \neq j}} \omega_{jk}}, \tag{11}$$

We set the corresponding entries of the guided attention maps to this value:

$$\hat{A}^{(i,k)}(c_i, c_k) = A^*, \qquad k \in U \setminus \{i\}. \tag{12}$$

The supplied foreground mask guides CoTracker point sampling only; stylization covers the entire frame. Applied to both foreground and background tracks, this trajectory-guided variant of temporal attention guidance (Samuel et al., 2025) replaces only the selected entries before value aggregation.

4 Experiments

4.1 Experimental Setup

Backbone and inference settings. We use pretrained Stable Diffusion v1.5 (Rombach et al., 2022) as the frozen diffusion backbone. Its VAE encoder and decoder map frames to and from the latent space; all model parameters remain frozen throughout. Latent frequency modulation, LAS, and CTA operate only during inference. CoTracker also remains frozen.

We invert each source-video frame and the reference style image separately with 50 DDIM inversion steps. AdaIN combines the inverted content and style latents to initialize stylization, followed by 50 DDIM sampling steps. All experiments are conducted on one NVIDIA Tesla V100 GPU with 32 GB memory. For frequency modulation, the Gaussian low-pass filter uses the normalized spatial frequency parameter $d_s = 0.3$, and the low-frequency scaling coefficient is $\alpha = 0.8$. CTA retains 80% of the available tracked points across the video. For LAS, $m_{\text{ref}} = 512 \times 512$ is the spatial sequence length of an input frame.

Datasets. We construct the evaluation set from 30 DAVIS (Pont-Tuset et al., 2017) videos and 30 ViDSOD-100 (Lin et al., 2024) videos. We retain the first 20 consecutive frames of each video. The reference set contains 40 paintings selected from the WikiArt website (WikiArt, n.d.), covering diverse palettes, textures, and painting styles. Each of the 60 videos is paired with every style image, yielding $60 \times 40 = 2{,}400$ video–style pairs and 48,000 output frames per method. Videos

are selected before stylization if at least 20 consecutive frames are available. The complete video and style-image identifiers will be released in the supplementary material. Frames and style images are resized to 512×512.

Baselines. We compare FACT with six methods: the image stylization methods EFDM (Zhang et al., 2022a) and CAST (Zhang et al., 2022b), the video stylization methods Diffutoon (Duan et al., 2024) and StyleCrafter (Liu et al., 2024), the video editing framework BIVDiff (Shi et al., 2024), and the training-free video stylization method UniVST (Song et al., 2026). Because these methods were designed for different tasks, we adapt their inputs for reference-guided video stylization. Diffutoon is text-guided and does not directly accept a reference style image; we therefore use GPT-5 (OpenAI, 2025) to generate a style description from that image. StyleCrafter originally generates videos from a content text prompt and a reference style image, but does not directly edit a source video. We invert the source video to obtain initial noise for its subsequent generation. For BIVDiff, we use InstantStyle (Wang et al., 2024) as its image stylization component. EFDM and CAST retain their original settings and are applied frame by frame.

UniVST receives the source video and reference style image with its mask module disabled. FACT uses a supplied foreground mask only for CoTracker query-point sampling; it does not mask the stylized frames. We evaluate the output frames from every method.

Following UniVST (Song et al., 2026), we evaluate overall transfer quality, source preservation, and video quality. LPIPS (Zhang et al., 2018) measures perceptual deviation from the source video, whereas FID (Heusel et al., 2017) measures the distributional discrepancy between stylized and reference-style images. ArtFID (Wright & Ommer, 2022) combines the two quantities as

$$\mathrm{ArtFID} = (1 + \mathrm{LPIPS})(1 + \mathrm{FID}). \tag{13}$$

SSIM (Wang et al., 2004) and CLIP-I measure structural and semantic similarity, respectively, between each output frame and its source. CLIP-I and CLIP-F are based on CLIP (Radford et al., 2021); CLIP-F measures feature similarity between adjacent output frames, while VBench (Huang et al., 2024) provides a complementary assessment of overall visual and temporal quality. Lower ArtFID, FID, and LPIPS are better; higher SSIM, CLIP-I, CLIP-F, and VBench are better.

4.2 Qualitative Comparison

Figure 5 compares FACT with six baselines on three video–style pairs. EFDM and CAST introduce reference colors and textures, but some source contours become less distinct in the vehicle and animal sequences. Diffutoon, StyleCrafter, and BIVDiff produce more pronounced appearance changes; in several frames, these changes also alter subject geometry or scene layout. UniVST follows the source structure more closely, although the reference painting's appearance is less apparent in the bird and monkey sequences.

FACT applies the reference palette and texture throughout each frame while keeping the main subjects recognizable. In the bird sequence, the blue–gray painterly treatment retains the birds' thin legs and body contours. In the vehicle sequence, the car silhouette and road layout remain visible as the car moves through the smoke. In the monkey sequence, the pose and head–torso structure remain recognizable under the transferred texture. The appearance of each scene is also visually coherent across the three displayed frames.

4.3 Quantitative Comparison

Table 1 compares video stylization over 2,400 video–style pairs. FACT attains the highest CLIP-F score of 0.986, exceeding UniVST by 0.005, and the highest VBench score of 75.43, exceeding UniVST by 1.18 points. Because CLIP-F compares adjacent output frames, its result provides direct evidence of greater feature-level frame consistency in this evaluation. VBench provides a complementary measure of overall visual and temporal quality.

The temporal results are accompanied by the lowest ArtFID (31.96), 5.7% below UniVST (33.89), and the highest SSIM (0.989) and CLIP-I (0.994). Together, these results show that FACT retains source structure and semantic content while achieving the strongest adjacent-frame consistency on the evaluated video–style pairs.

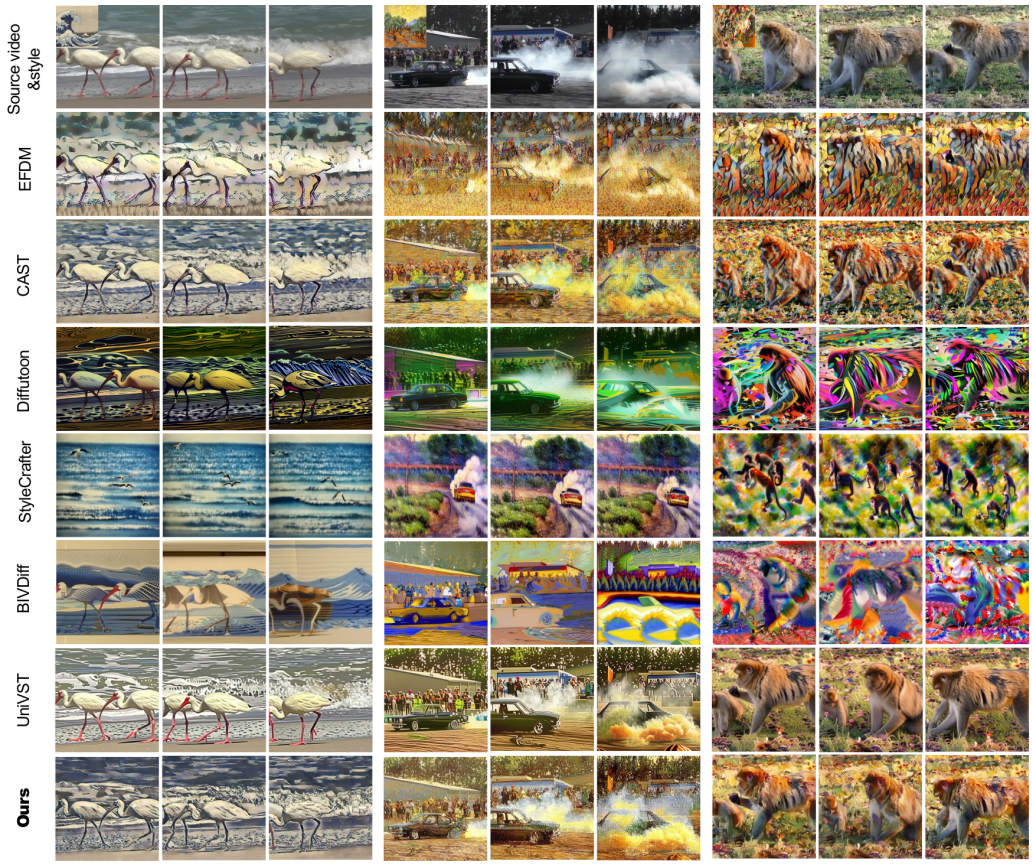

Figure 5: Qualitative comparison on three video–style pairs. The first row shows three consecutive source frames for each video, with the reference style image inset in the first frame. The remaining rows show the corresponding stylized frames from six baselines and FACT.

Table 1: Quantitative comparison of video stylization on 2,400 video–style pairs.

Method	**Overall Transfer**			**Source Preservation**		**Video Quality**	
	ArtFID↓	**FID↓**	**LPIPS↓**	**SSIM↑**	**CLIP-I↑**	**CLIP-F↑**	**VBench↑**
EFDM	36.54	21.95	0.592	0.935	0.933	0.964	71.14
CAST	40.00	23.07	0.662	0.948	0.936	0.972	71.17
Diffutoon	40.88	24.99	0.573	0.960	0.954	0.980	72.93
StyleCrafter	41.15	22.13	0.779	0.927	0.896	0.978	71.07
BIVDiff + InstantStyle	41.38	22.82	0.737	0.921	0.896	0.956	70.31
UniVST	33.89	23.63	**0.376**	0.986	0.991	0.981	74.25
FACT (Ours)	**31.96**	22.14	0.381	**0.989**	**0.994**	**0.986**	**75.43**

4.4 Ablation Study

Figure 6 compares the full model with each of its three modules removed in turn. Without latent frequency modulation, the bear and surrounding rock contours appear softer, with less fine texture than in the full result. This visual example supports the role of frequency modulation in retaining spatial detail. Table 2 reports the corresponding quantitative results for all three module ablations.

Table 2: Quantitative ablation of latent frequency modulation, LAS, and CTA on the evaluation set.

Configuration	Overall Transfer			Source Preservation		Video Quality	
	ArtFID↓	FID↓	LPIPS↓	SSIM↑	CLIP-I↑	CLIP-F↑	VBench↑
w/o LAS	33.57	23.05	0.396	0.986	0.991	0.985	74.75
w/o CTA	32.18	22.25	0.384	0.988	0.993	0.974	74.31
w/o Freq	32.62	22.40	0.394	0.987	0.993	0.985	74.87
FACT	**31.96**	**22.14**	**0.381**	**0.989**	**0.994**	**0.986**	**75.43**



Figure 6: Qualitative ablation over two consecutive frames. From left to right: input frames with an inset reference style image, results without CTA, without frequency modulation, without LAS, and the full FACT model.

Removing LAS raises ArtFID from 31.96 to 33.57, the largest increase among the three ablations, and the visible loss of source structure in Figure 6 confirms its role in content preservation. Without frequency modulation, ArtFID rises to 32.62 and fine contours become softer in the corresponding panel of Figure 6, consistent with the spatial-detail benefit of frequency modulation.

Removing CTA lowers CLIP-F from 0.986 to 0.974, a drop much larger than that caused by removing either of the other two modules. This supports CTA's central role in adjacent-frame consistency. The full model achieves the best result on all seven metrics among the four configurations.

5 Conclusion

We presented FACT, a training-free framework that combines latent frequency modulation with attention control for reference-guided video stylization. Frequency modulation preserves fine spatial details by attenuating the low-frequency latent component at each denoising step. LAS balances style injection with source-structure preservation by adapting attention strength across feature scales. CTA improves frame-to-frame consistency through confidence-weighted attention along tracked correspondences. Experiments show that FACT outperforms existing methods in style transfer quality, source preservation, and temporal consistency. Since FACT operates in the compressed latent space of Stable Diffusion, exploring higher-fidelity latent representations is a promising direction for recovering finer visual structures.

AI Use Statement

We used GPT-5 to generate reference-style descriptions for the Diffutoon baseline, as described in the experimental setup. We also used large language models to polish the manuscript's language and improve its clarity and readability. All AI-assisted content has been reviewed and verified by the authors, who take full responsibility for the final manuscript.

Ethics Statement

This work involves no human subjects, personal privacy, or sensitive personal data, and raises no additional ethical concerns.

Reproducibility Statement

The method and its components are specified in Section 3, and the evaluation protocol and implementation settings are described in Section 4.1. Additional ablations are in Appendix A, and the LAS scaling derivation is in Appendix B.

References

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7654–7664, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.527. URL https://aclanthology.org/2022.acl-long.527/.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8795–8805, 2024. doi: 10.1109/CVPR52733.2024.00840.

Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. StyTr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11326–11336, 2022. doi: 10.1109/CVPR52688.2022.01104.

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

Zhongjie Duan, Chengyu Wang, Cen Chen, Weining Qian, and Jun Huang. Diffutoon: High-resolution editable toon shading via diffusion models. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 7645–7653. International Joint Conferences on Artificial Intelligence Organization, August 2024. doi: 10.24963/ijcai.2024/846. URL https://doi.org/10.24963/ijcai.2024/846. AI, Arts & Creativity.

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. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1510–1519, 2017. doi: 10.1109/ICCV.2017.167.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 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.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. In *Computer Vision–ECCV 2024*, Lecture Notes in Computer Science, pp. 18–35. Springer, 2024. doi: 10.1007/978-3-031-73033-7_2.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Computer Vision–ECCV 2018*, volume 11219 of *Lecture Notes in Computer Science*, pp. 170–185. Springer, 2018. doi: 10.1007/978-3-030-01267-0_11.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, volume 30, pp. 386–396, 2017.

Junhao Lin, Lei Zhu, Jiaxing Shen, Huazhu Fu, Qing Zhang, and Liansheng Wang. ViDSOD-100: A new dataset and a baseline model for RGB-D video salient object detection. *International Journal of Computer Vision*, 132(11):5173–5191, 2024. doi: 10.1007/s11263-024-02051-5.

Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Ying Shan, and Yujiu Yang. StyleCrafter: Taming artistic video diffusion with reference-augmented adapter learning. *ACM Transactions on Graphics*, 43(6):251, 2024. doi: 10.1145/3687975. URL https://doi.org/10.1145/3687975.

Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. AdaAttN: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6649–6658, 2021. doi: 10.1109/ICCV48922.2021.00658.

OpenAI. GPT-5 system card. Technical report, OpenAI, 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf.

Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5880–5888, 2019. doi: 10.1109/CVPR.2019.00603.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. doi: 10.48550/arXiv.1704.00675.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 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. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *Pattern Recognition: 38th German Conference, GCPR 2016*, volume 9796 of *Lecture Notes in Computer Science*, pp. 26–36. Springer, 2016. doi: 10.1007/978-3-319-45886-1_3.

Dvir Samuel, Matan Levy, Nir Darshan, Gal Chechik, and Rami Ben-Ari. OmnimatteZero: Fast training-free omnimatte with pre-trained video diffusion models. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025. doi: 10.1145/3757377.3763917.

Fengyuan Shi, Jiaxi Gu, Hang Xu, Songcen Xu, Wei Zhang, and Limin Wang. BIVDiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7393–7402, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Quanjian Song, Mingbao Lin, Wengyi Zhan, Shuicheng Yan, Liujuan Cao, and Rongrong Ji. UniVST: A unified framework for training-free localized video style transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2820–2836, 2026. doi: 10.1109/TPAMI.2025.3630010.

Jianlin Su. Understanding the scale operation in attention from entropy invariance. https://spaces.ac.cn/archives/8823, 2021. Blog post in Chinese; original title: Cong shang bu bian xing kan Attention de Scale caozuo.

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. doi: 10.48550/arXiv.2404.02733.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

WikiArt. WikiArt: Visual art encyclopedia, n.d. URL https://www.wikiart.org/. Accessed July 16, 2026.

Matthias Wright and Björn Ommer. ArtFID: Quantitative evaluation of neural style transfer. In *Pattern Recognition: 44th DAGM German Conference, GCPR 2022*, volume 13485 of *Lecture Notes in Computer Science*, pp. 560–576. Springer, 2022. doi: 10.1007/978-3-031-16788-1_34.

Jiacong Xu, Yiqun Mei, Ke Zhang, and Vishal M. Patel. FreeViS: Training-free video stylization with inconsistent references. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=SiYNm21ifi.

Ruojun Xu, Weijie Xi, XiaoDi Wang, Yongbo Mao, and Zach Cheng. StyleSSP: Sampling startpoint enhancement for training-free diffusion-based method for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18260–18269, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8035–8045, 2022a.

Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–8, 2022b. doi: 10.1145/3528233.3530736.

A Additional Analysis

A.1 Attention Scaling

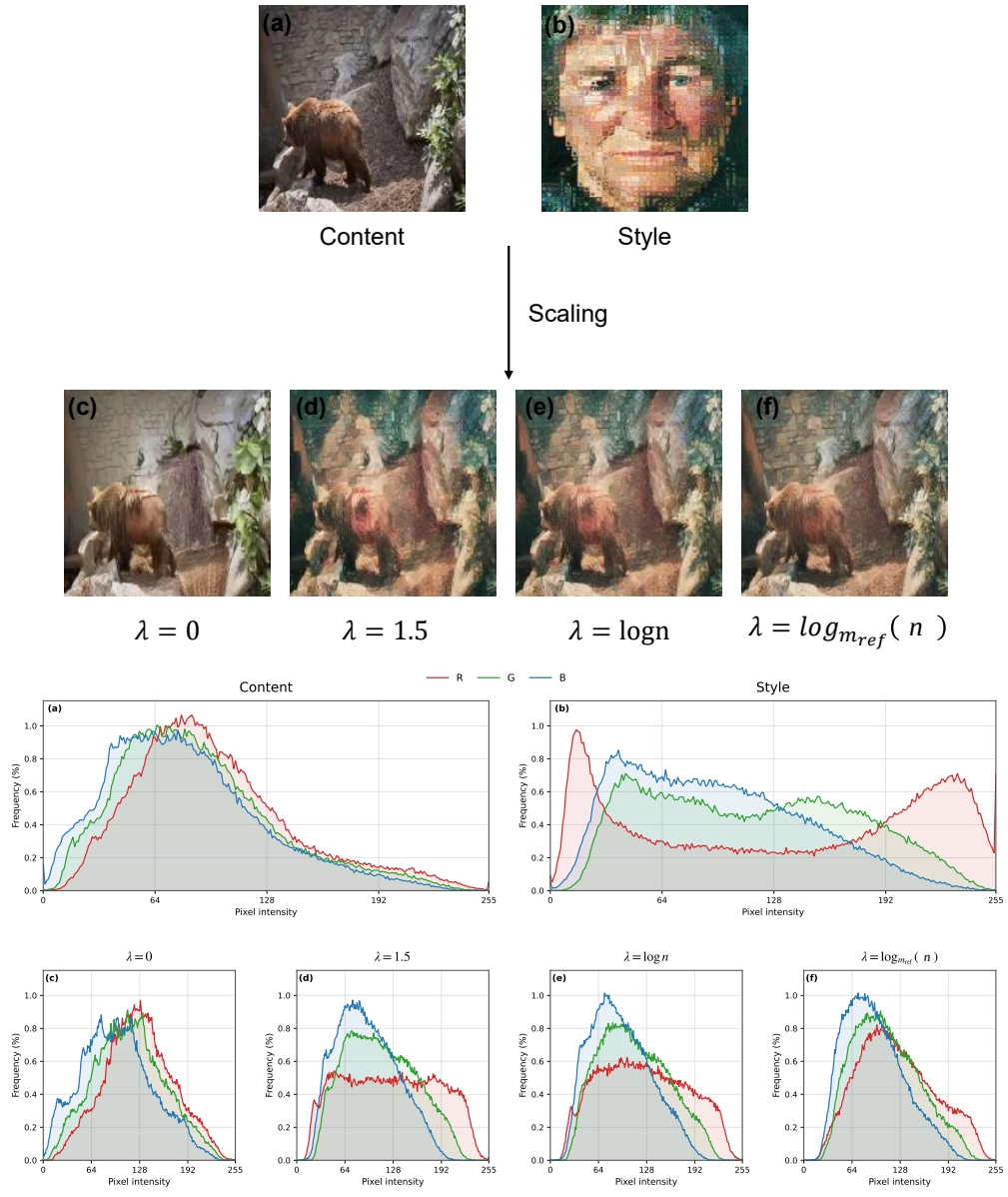

Figure 7: Effect of attention scaling on a content–style pair. The upper panels compare $\lambda = 0$, StyleID's fixed coefficient of 1.5 (Chung et al., 2024), sequence-length scaling $\lambda = \log n$ (Chiang & Cholak, 2022), and LAS scaling $\lambda = \log_{m_{\text{ref}}} n$. The lower panels show the RGB intensity distributions of the inputs and corresponding outputs.

Figure 7 compares four attention scaling settings on the same content–style pair. With $\lambda = 0$, the bear retains more of the source photograph's appearance. Fixed $\lambda = 1.5$ and $\lambda = \log n$ produce stronger appearance changes, but the bear's contours become less distinct. LAS adapts $\lambda = \log_{m_{\text{ref}}}(n)$ to the style-key length at each U-Net scale; its result keeps the bear and rock boundaries recognizable while introducing the reference appearance. The RGB distributions confirm that the four settings produce distinct tonal responses.

A.2 Low-Frequency Retention Coefficient

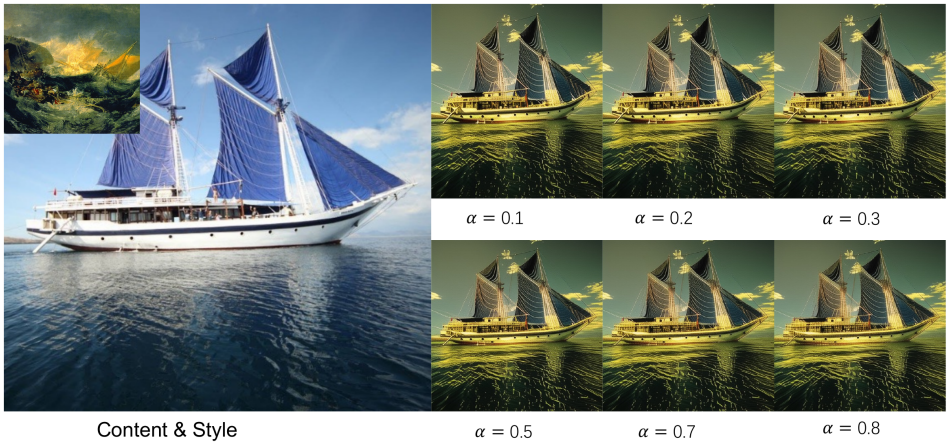

Figure 8: Effect of the low-frequency retention coefficient α on a stylized frame. Source frame and reference style image alongside results for $\alpha \in \{0.1, 0.2, 0.3, 0.5, 0.7, 0.8\}$.

Figure 8 compares six values of the low-frequency retention coefficient α on the same content frame and reference style image. In this example, $\alpha = 0.8$ retains fine details in the sail rigging, hull, and water texture while transferring the reference palette. We use $\alpha = 0.8$ in the main experiments based on this visual detail comparison.

A.3 CoTracker Sampling Density

Table 3 studies the effect of the proportion of tracked points sampled across all frames for CTA. CLIP-F increases from 0.974 with no sampled points to 0.980, 0.983, and 0.985 at 20%, 40%, and 60%, respectively. The 80% and 100% settings both reach 0.986. We use 80% in the main experiments because it matches the highest observed CLIP-F while sampling fewer points than the 100% setting.

Table 3: Effect of CoTracker point sampling ratio across all frames on CLIP-F.

Sampling Ratio	CLIP-F ↑
0%	0.974
20%	0.980
40%	0.983
60%	0.985
80%	**0.986**
100%	**0.986**

B Derivation of the LAS Scaling Factor

Let $a_j = e^{\beta q \cdot k_j} / \sum_{j'} e^{\beta q \cdot k_{j'}}$ be the attention weight on key j, where $q, k_j \in \mathbb{R}^d$ and β is the coefficient on the unnormalized dot product. Its entropy is

$$H = -\sum_{j=1}^{n} a_j \log a_j. \tag{14}$$

We seek a scale that limits the dilution of attention as the number of keys n grows.

Treating the keys as i.i.d. random vectors on the $\sqrt{d}$-radius sphere and applying a Laplace approximation to the partition function (Su, 2021) gives

$$H \approx \log n - c\beta d + O(1), \qquad c > 0. \tag{15}$$

Under this approximation, cancelling the leading dependence on $\log n$ gives $\beta \propto \log n/d$. Write $\beta = \kappa \log n/d$ for a constant κ.

We fix κ by requiring $\beta = 1/\sqrt{d}$ when $n = m_{\text{ref}}$, the reference-image sequence length:

$$\kappa \frac{\log m_{\text{ref}}}{d} = \frac{1}{\sqrt{d}} \implies \kappa = \frac{\sqrt{d}}{\log m_{\text{ref}}}. \tag{16}$$

Substitution yields

$$\beta = \frac{\sqrt{d}}{\log m_{\text{ref}}} \frac{\log n}{d} = \frac{\log_{m_{\text{ref}}} n}{\sqrt{d}}. \tag{17}$$

Since the attention formula in Eq. (6) already includes $1/\sqrt{d}$, its additional multiplicative coefficient is

$$\boxed{\lambda = \log_{m_{\text{ref}}} n,} \tag{18}$$

as in Eq. (5).