Stylized Video Generation via Decoupled Data Synthesis and Gated Style Token Injection

Xin Wei[1], Yijie Fang[1], Yanjia Li[2], Liangyi Wu[2], Qin Yang[1], Mingrui Zhu[1], Nannan Wang[1⋆], and Xinbo Gao[1]

[1] Xidian University, Xi'an, China
[2] Xi'an Jiaotong University, Xi'an, China
{weixin,mrzhu,nnwang}@xidian.edu.cn,
{fangyijie,qinyang}@stu.xidian.edu.cn,
{yanjiali,748811623}@stu.xjtu.edu.cn, xbgao@mail.xidian.edu.cn

Fig. 1: Existing stylized video generation methods (*e.g.*, StyleMaster [50], StyleCrafter [21]) suffer from progressive motion degradation and poor fidelity to the reference style. Our method addresses both limitations via decoupled content-motion data synthesis and implicit gated style token injection.

Abstract. Stylized video generation requires synthesizing dynamic content that reflects the artistic characteristics of a reference image while maintaining semantic control and temporal coherence. Existing fine-tuning-based approaches suffer from two recurring limitations: motion degradation caused by training on static image stylization datasets, which lack temporal supervision and erode the temporal priors of pre-trained backbones; and a representation bottleneck from encoder-based style injection, which discards fine-grained stylistic details and bounds stylization quality by the encoder's representational capacity. We address both limitations through a unified pipeline that covers dataset construction, model

⋆ Corresponding author.

architecture, and training strategy. Specifically, we construct **MoStyle-5K**, a stylized video dataset of 5,000 videos spanning 27 artistic styles and 4,000 preference pairs, built via a decoupled content-motion synthesis pipeline that explicitly preserves temporal priors in the supervision signal. Building on this, we propose **Implicit Gated Style Token Injection** (IGST), which directly concatenates style patches as tokens to bypass the encoder bottleneck, disables RoPE for style tokens to remove spatial alignment bias, and applies head-wise gated attention to suppress residual structural correlations. The framework is optimized with a hybrid image-video objective and refined via Diffusion-DPO post-training using the structured preference pairs in MoStyle-5K. Experiments show that our method achieves superior style fidelity and motion dynamics compared with existing baselines. Project page: `https://sixi111.github.io/MoStyle/`.

Keywords: Stylized Video Generation · Video Diffusion Models · Stylized Video Dataset

1 Introduction

Text-to-video diffusion models [1,3,10,17,30,33,37,48] have demonstrated strong generative priors for dynamic visual content, enabling high-quality video synthesis from textual descriptions. Stylized video generation extends this capability by conditioning generation on a reference style image, with the goal of producing videos that reflect the artistic characteristics of the reference while adhering to textual semantic controls and maintaining temporal coherence [21, 46, 50]. Despite notable advances in reference-based image stylization [38,39,45], extending style conditioning to the video domain introduces qualitatively distinct challenges: videos require temporally consistent style propagation under dynamic motion, where existing methods exhibit characteristic failure modes including progressive motion degradation and poor fidelity to the reference style, as illustrated in Fig. 1.

We identify two limitations in existing methods, as illustrated in Fig. 2. First, **motion degradation from static training data**: existing methods [21, 50] fine-tune pre-trained video diffusion models on static image stylization datasets [35], which provide only frame-wise supervision without temporal constraints. The optimization objective consequently shifts from spatio-temporal modeling to per-frame appearance adaptation, gradually weakening the temporal modeling capability of the pre-trained backbone and leading to degraded motion coherence in generated videos. Second, **representation bottleneck from encoder-based style injection**: existing methods compress the reference image through a dedicated style encoder before injecting it via cross-attention, which tends to discard fine-grained stylistic statistics that are critical for stylization fidelity. More fundamentally, stylization quality is limited by the representational capacity of the encoder: any limitation in its expressive power imposes an upper bound on cross-style generalization, and increasing training scale alone cannot recover

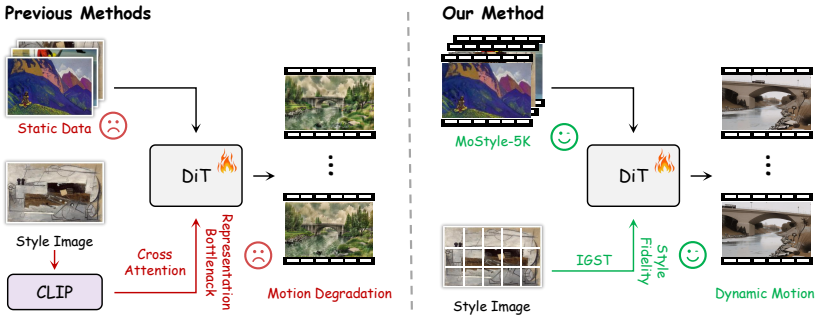


Fig. 2: Comparison of stylized video generation paradigms. Existing fine-tuning approaches rely on static stylized image datasets and encoder-based style injection, leading to motion degradation and representation bottlenecks. In contrast, our method constructs MoStyle-5K with explicit temporal supervision and introduces Implicit Gated Style Token Injection (IGST), which directly concatenates style patches as tokens and applies head-wise gated attention to improve style fidelity and motion coherence.

information already compressed away. While training-free alternatives such as FreeVis [46] sidestep these pitfalls, this approach typically relies on a multi-stage pipeline that introduces substantial inference overhead and limits practical scalability. Consequently, a fundamental challenge remains: how to overcome the motion and representation bottlenecks of fine-tuning-based stylization without sacrificing computational efficiency.

We argue that resolving these two issues requires rethinking the problem at two levels simultaneously: the training data must provide explicit temporal supervision to preserve the motion priors of the backbone, and the conditioning mechanism must operate directly on uncompressed style representations to avoid the information loss inherent in encoder-based injection. To this end, we construct **MoStyle-5K**, a stylized video dataset built via a decoupled content-motion synthesis pipeline. A large language model decomposes each text prompt into a static content description and a dynamic motion description; the former drives stylized image synthesis via Qwen-Image [43], while the latter conditions an image-to-video model to produce temporally coherent stylized videos. From over 27,000 synthesized candidates, we curate 5,000 videos spanning 27 artistic styles and 1,000 prompts, together with 4,000 structured preference pairs for preference-based optimization. By providing explicit temporal supervision, MoStyle-5K establishes the data foundation that makes implicit style token injection a viable training objective.

Building on this foundation, we introduce **Implicit Gated Style Token Injection** (IGST), as illustrated in Fig. 2. Rather than compressing the reference image with a dedicated encoder, IGST directly represents style image patches as tokens and concatenates them with video tokens in the transformer input space. This preserves fine-grained stylistic statistics at full token resolution. To prevent structural leakage from the reference image, IGST controls style propagation through two designs: disabling rotary positional embeddings for style tokens to

remove spatial alignment bias, and using lightweight head-wise gates to suppress attention heads that carry structural layout cues. These designs decouple structural cues from stylistic features during cross-token interaction. For training, we adopt a hybrid image-video training strategy that balances stylistic and temporal supervision signals, and apply Diffusion-DPO [29, 36] as a post-training stage to explicitly penalize structural artifacts and style degradation with the preference pairs in MoStyle-5K.

The main contributions of this work are as follows:

- We construct **MoStyle-5K**, a stylized video dataset of 5,000 videos across 27 artistic styles and 4,000 preference pairs, built via a decoupled content-motion synthesis pipeline that provides explicit temporal supervision absent in existing static stylization datasets.
- We propose **Implicit Gated Style Token Injection** (IGST), which directly concatenates style patches as tokens to preserve full-resolution stylistic information, disables rotary positional embeddings for style tokens to remove spatial alignment bias, and applies head-wise gated attention to suppress structural contamination from the reference image.
- We present a complete training pipeline combining a hybrid image-video fine-tuning objective with Diffusion-DPO post-training [36], which disentangles stylistic and temporal supervision signals and leverages the preference pairs in MoStyle-5K to suppress structural artifacts and style degradation.

2 Related Work

2.1 Stylized Video Generation

Stylized image generation has evolved from optimization-based feature statistics matching [8] and feed-forward arbitrary style transfer [6, 7, 13, 16, 22, 26] to diffusion-based reference conditioning and modular adapters [5,39,49]. Training-free composition methods [31,38,47] such as CSGO [45] further decouple content and style representations within diffusion features. Despite these advances in the image domain, extending stylization to videos introduces additional challenges beyond single-frame appearance transfer.

Stylized text-to-video generation must preserve motion coherence and long-range temporal consistency while enforcing style control. Training-free alternatives such as FreeVis [46] sidestep fine-tuning by combining base video generation, stylized image synthesis, and attention-based feature manipulation at inference time, but the resulting multi-stage pipeline substantially increases computational overhead and limits practical scalability. Fine-tuning-based methods couple pre-trained T2V backbones with explicit style conditioning: StyleCrafter [21] adopts a Q-Former [19] to extract style descriptions from a reference image, while StyleMaster [50] employs a decoupled dual-path extractor with global and local branches to capture overall appearance and fine-grained textures. However, these fine-tuning-based approaches share a common limitation: reliance on static training data and explicit style encoding jointly constrain stylization fidelity and motion coherence.

A complementary line studies video-to-video stylization and reference-based editing, where motion and spatial layout are inherited from a source video while appearance is transformed according to a reference [9,15,18,20,25,41,51]. In such settings, temporal coherence is largely preserved by construction. In contrast, stylized T2V generation must synthesize motion and appearance jointly from scratch, without access to external dynamic guidance.

2.2 Conditioning Paradigms in Video Diffusion Models

Controllable video generation requires incorporating external control signals into a pre-trained video diffusion backbone while preserving visual fidelity and temporal coherence. Existing approaches can be broadly categorized into explicit and implicit conditioning, depending on whether the control signal is first encoded by an additional module or directly integrated within the generative backbone.
Explicit Conditioning. Explicit conditioning methods introduce dedicated encoders, adapters, or auxiliary branches to extract control representations before injecting them into the generation backbone [4, 11, 12, 24, 42]. Although this paradigm has been widely adopted for controllable video generation, it introduces an intermediate representation separated from the native token space of the pre-trained backbone. As a result, generation quality and control fidelity depend on the expressiveness of the external module and its compatibility with the backbone representation. Such a design may prematurely constrain the control signal, discard fine-grained information, or introduce distributional mismatch during injection. For reference-based style control, compressing the reference image into an external style representation may weaken key style statistics, such as color palettes, brushstroke patterns, texture distributions, and local contrast.
Implicit Conditioning. Implicit conditioning methods incorporate control information through internal model design or parameter-efficient adaptation, without requiring a separate feature extraction stage. Tune-A-Video [44] and MotionDirector [52] enable motion customization via light-weight fine-tuning [44, 52], while AnimateDiff injects transferable motion priors in a plug-and-play manner [10]. CamCloneMaster [23] clones camera motion from a reference video through model-internal design, without explicit camera parameterization or test-time fine-tuning. These methods suggest that control signals can be modeled within the backbone, reducing reliance on externally compressed representations and preserving the native generative prior. Following this idea, we directly introduce reference style images into the transformer token space, enabling fine-grained style statistics to interact with video tokens.

3 Method

3.1 Overall Framework

Given a reference style image $\mathcal{I}_s \in \mathbb{R}^{H_s \times W_s \times 3}$ and a textual prompt $\mathcal{P}$, the goal is to synthesize a stylized video $\mathcal{V} \in \mathbb{R}^{F \times H \times W \times 3}$ that adheres to the semantic

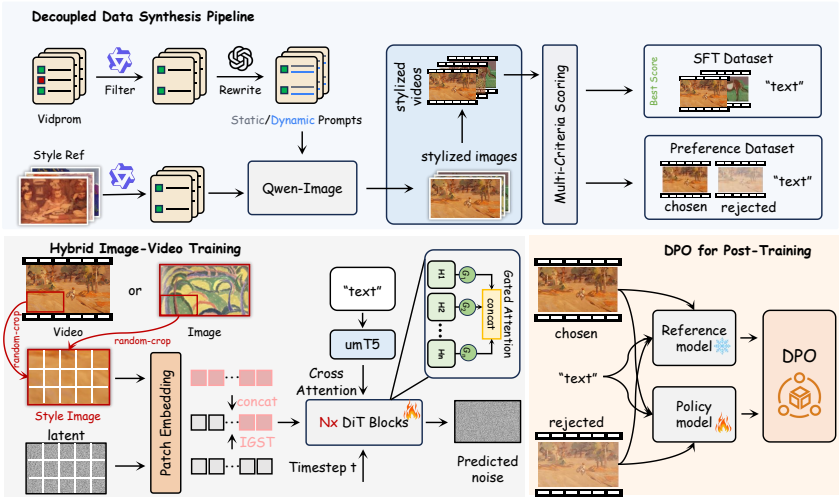


Fig. 3: Overall framework. **(Top)** MoStyle-5K is constructed by decomposing prompts into static and dynamic descriptions, synthesizing stylized videos, and curating SFT and preference datasets via multi-criteria scoring. **(Bottom left)** IGST directly concatenates style patches as tokens alongside video latents and applies head-wise gated attention to suppress structural contamination. **(Bottom right)** Diffusion-DPO post-training leverages the preference pairs in MoStyle-5K to suppress structural artifacts and style degradation.

content of $\mathcal{P}$, reflects the artistic characteristics of $\mathcal{I}_s$, and maintains coherent temporal dynamics:

$$\mathcal{V} = \mathcal{F}_\theta(\boldsymbol{z}_T, \mathcal{I}_s, \mathcal{P}), \tag{1}$$

where $\boldsymbol{z}_T$ denotes the initial Gaussian noise and $\mathcal{F}_\theta$ is our unified transformer-based video diffusion model. An overview is illustrated in Fig. 3.

Our framework addresses two recurring limitations in existing stylized video generation: representation bottlenecks from encoder-based style injection and motion degradation from static supervision. The pipeline operates across three levels. On the data side, **MoStyle-5K** (Sec. 3.2) provides stylized videos with explicit temporal supervision, avoiding reliance on static image datasets that undermine the temporal modeling capability of pre-trained backbones. On the architectural side, both the noisy video latents and $\mathcal{I}_s$ are patchified, linearly projected into token sequences, and concatenated along the sequence dimension, enabling compression-free style-content interaction through self-attention. To regulate how style information propagates into video content, **Implicit Gated Style Token Injection** (IGST) (Sec. 3.3) disables rotary positional embeddings for style tokens to remove spatial alignment bias, and applies head-wise gated attention to suppress structural layout transfer while retaining style-relevant statistics. On the training side, a hybrid image-video supervised fine-tuning objective decouples spatial style reinforcement from temporal modeling, followed by

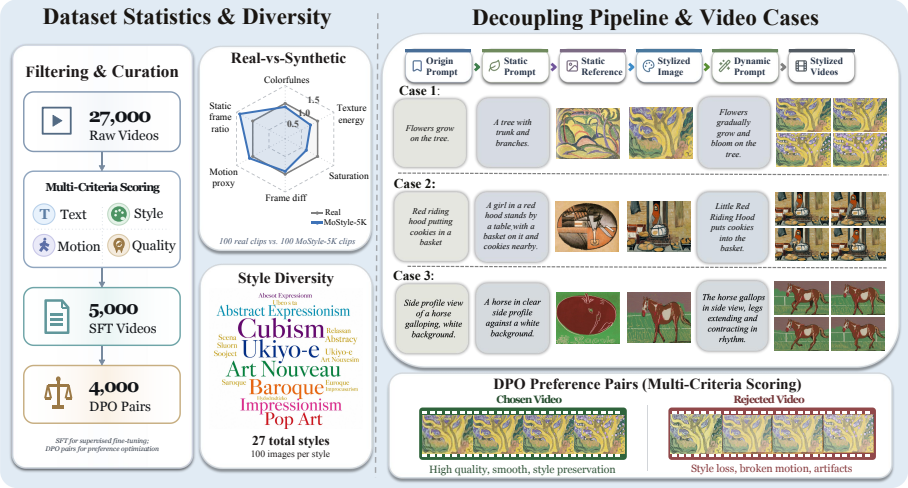


Fig. 4: Overview of the stylized video dataset. **Left:** dataset statistics and diversity. From 27,000 raw generations, we curate 5,000 high-quality SFT videos and 4,000 DPO preference pairs across 27 artistic styles and diverse semantic categories. **Right:** decoupled content–motion pipeline and example cases, where stylized images are first generated from static prompts and references, and then extended with dynamic prompts to produce stylized videos, followed by multi-criteria scoring for preference construction.

Diffusion-DPO post-training using the structured preference pairs in MoStyle-5K to suppress structural artifacts and style degradation (Sec. 3.4).

3.2 Dataset Construction: MoStyle-5K

To provide explicit temporal supervision for stylized video generation, we construct **MoStyle-5K**, a stylized video dataset designed to preserve dynamic content throughout the stylization pipeline. Additional construction details and dataset examples are provided in Appendix A of the Supplementary Material.

Dynamic Prompt Curation. We begin with the VidProM [40] text corpus and apply a Qwen-assisted [2] multi-stage filtering process to obtain high-quality motion-rich prompts. First, we remove invalid or low-quality entries, including meaningless texts, NSFW content, HTML artifacts, and other noisy samples, reducing the corpus from about 1M to 666K prompts. We then perform action-word pre-filtering using a manually curated vocabulary of around 200 action and camera-motion keywords, retaining 335K prompts with explicit dynamic cues. Finally, we use Qwen to perform rule-based filtering and style-description removal, discarding ambiguous, low-quality, weakly dynamic, or style-leaking descriptions. This results in 258K high-quality dynamic prompts that are style-agnostic, ensuring that artistic appearance is controlled by the reference style condition rather than textual style leakage. From this filtered prompt pool, we sample 1,000 prompts via semantic clustering and motion-pattern balancing.

Content-Motion Rewriting. Each curated prompt is rewritten by GPT-5.1[1] into three structured fields: a static caption, a subject-level motion description, and an optional camera-motion description. The static caption describes the subject, scene, appearance, and atmosphere, while the subject-level motion description captures actions in natural language without camera movement. This decomposition separates static visual content from dynamic behavior, enabling stylized image synthesis from static descriptions while preserving motion as an explicit supervision signal for video generation.

Style References. We construct a curated reference artwork pool based on widely used artistic style categories, comprising a balanced set of 2,700 images across 27 styles (*e.g.*, Abstract Expressionism, Cubism, Ukiyo-e), with 100 images per category to avoid optimization bias toward specific visual patterns.

Video Synthesis and Curation. Given a static prompt and a style reference, we first use Qwen-Image to synthesize a stylized anchor image, which is then combined with the dynamic prompt and fed into a pre-trained image-to-video model to synthesize candidate videos. Multiple candidates are generated per prompt-style pair, yielding over 27,000 raw sequences. Each candidate is evaluated by a multi-criteria scoring mechanism covering text alignment, dynamic degree, motion smoothness, aesthetic quality, and style consistency. Candidates above all quality thresholds form the SFT dataset of **5,000** videos. For DPO fine-tuning, we further select informative sub-threshold generations and pair them with the top-scoring candidates generated under identical prompt-style conditions, resulting in **4,000** preference pairs.

Dataset Statistics and Realism. As shown in Fig. 4, we compare 100 MoStyle-5K clips with 100 real artistic-animation clips under identical temporal sampling and spatial resizing. We compute generator-independent appearance and motion statistics, including colorfulness, texture energy, saturation, frame-difference magnitude, motion proxy, and static-frame ratio. The small distributional gaps across these metrics indicate that MoStyle-5K captures relevant color, texture, and short-term temporal statistics of real artistic animations.

3.3 Architectural Design: Implicit Gated Style Token Injection

MoStyle-5K makes token-level style conditioning viable: by providing stylized videos with explicit motion supervision, it ensures that style conditioning is learned jointly with temporal dynamics rather than at their expense. Built on this foundation, we perform direct token-level conditioning within the DiT backbone. Given a reference style image $\mathcal{I}_s$ and a noisy video latent $\boldsymbol{z}_t$, we first encode the reference image into a **clean latent** and set its **timestep to** 0 during denoising. Compared with a noisy latent, the clean latent retains substantially richer style information, thereby providing a more informative signal for learning fine-grained stylistic patterns. Setting the timestep to 0 further prevents the style reference from being corrupted by the diffusion process. We then project the

[1] `https://developers.openai.com/api/docs/models/gpt-5.1`

style latent and video latent using two separate patch embedding layers, producing style tokens $\mathbf{S} \in \mathbb{R}^{N_s \times d}$ and video tokens $\mathbf{Z} \in \mathbb{R}^{N_v \times d}$, where N_s and N_v denote the respective sequence lengths and d the embedding dimension. The two sequences are concatenated to form a unified input $\mathbf{X} = [\mathbf{S}; \mathbf{Z}] \in \mathbb{R}^{(N_s+N_v)\times d}$, which is processed by the 3D spatio-temporal self-attention blocks of the DiT backbone [27]. This formulation preserves style information at full spatial resolution and allows fine-grained style propagation across all transformer layers without compression loss.

However, direct token concatenation introduces a spatial alignment bias: style tokens inherit rotary positional embeddings (RoPE) [34] that encode the spatial layout of the reference image, potentially biasing attention toward spatial correspondence rather than stylistic similarity. To address this, we introduce **Implicit Gated Style Token Injection** (IGST), which regulates style-content interaction at two complementary levels.

At the positional level, IGST disables RoPE for style tokens after query and key projection. Let $\mathbf{q}_i$ and $\mathbf{k}_i$ denote the query and key of token i, and let $\mathbf{p}_i$ denote its 3D position. We apply RoPE only to video tokens:

$$\tilde{\mathbf{q}}_i = \begin{cases} \mathrm{RoPE}(\mathbf{q}_i, \mathbf{p}_i), & i \in \mathcal{V}, \\ \mathbf{q}_i, & i \in \mathcal{S}, \end{cases} \qquad \tilde{\mathbf{k}}_i = \begin{cases} \mathrm{RoPE}(\mathbf{k}_i, \mathbf{p}_i), & i \in \mathcal{V}, \\ \mathbf{k}_i, & i \in \mathcal{S}. \end{cases} \tag{2}$$

where $\mathcal{V}$ and $\mathcal{S}$ denote the sets of video tokens and style tokens, respectively. As a result, video tokens preserve complete spatio-temporal positional awareness, whereas style tokens interact with video tokens through feature-space similarity without carrying reference-image spatial phase information. This positional decoupling reduces structural leakage while retaining fine-grained style cues.

At the attention-head level, IGST introduces lightweight learnable gates to regulate style-token interactions. Each attention head is assigned a sigmoid-normalized scalar gate, which is shared across query tokens and applied to the style-token keys in self-attention:

$$\mathbf{X}_{\text{out}} = \mathrm{Attention}\left([\mathbf{Q}_{\text{vid}}, \mathbf{Q}_{\text{sty}}], [\mathbf{K}_{\text{vid}}, \mathbf{K}_{\text{sty}} \odot \boldsymbol{\alpha}], [\mathbf{V}_{\text{vid}}, \mathbf{V}_{\text{sty}}]\right), \tag{3}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^H$ denotes the sigmoid-normalized learnable gates for the H attention heads, and $\odot$ denotes head-wise multiplication. This design selectively amplifies or attenuates style-token contributions at the head level, encouraging the model to focus on discriminative style statistics while suppressing redundant or spatially biased interactions.

3.4 Hybrid Optimization and Preference Alignment

We optimize the model with hybrid image-video supervision to jointly improve style fidelity and temporal consistency, followed by preference alignment to further enhance generation quality. To establish reliable style adaptation, we begin with an image-level warm-up stage using stylistic image samples, which provide rich spatial supervision for learning fine-grained textures, color distributions, and local style statistics.

Image-Video Mixed Training. We then continue training with both stylized images and stylized videos. Image samples help retain the style fidelity learned in the first stage, while video samples introduce spatio-temporal supervision for motion-consistent stylization. This mixed training stage allows the model to preserve fine-grained style details while learning to propagate them coherently across frames.

Preference Alignment via Diffusion-DPO. Finally, we apply Diffusion-DPO [29, 36] as a post-training stage using the structured preference pairs in MoStyle-5K. Each pair contains a preferred sample and a rejected sample generated under the same style and text condition. By learning from these pairwise preferences, the model is further encouraged to produce outputs with better style fidelity, content preservation, and motion quality, while suppressing artifacts such as structural contamination and style degradation.

4 Experiments

4.1 Experimental Settings

Implementation Details. We adopt Wan-T2V 1.3B [37] as the base text-to-video diffusion backbone. We perform full-parameter fine-tuning of the diffusion transformer on MoStyle-5K (Sec. 3.2), with the learning rate and global batch size set to 2×10^{-5} and 8, respectively. For DPO post-training, we set the learning rate to 1.5×10^{-6} and β to 1. Both training stages are conducted on 8 NVIDIA H200 GPUs at a resolution of 480×832.

Testing Protocol. We construct a dedicated test set for style-image-guided text-to-video generation. The test set contains 27 content prompts and 40 style reference images, covering diverse content scenarios and representative artistic styles. To evaluate the model under varied content–style combinations, we randomly pair content prompts with style references, resulting in 462 prompt-style pairs in total.

Evaluation Metrics. We adopt multiple quantitative metrics for comprehensive evaluation. CLIP text score [28] measures text alignment, while style similarity (CSD) [32] is computed frame-wise between generated frames and the style reference. In addition, following StyleMaster [50] and StyleCrafter [21], we employ VBench metrics [14], including Dynamic Degree (DD), Dynamic Smoothness (DS), Image Quality (IQ), and Visual Quality (VQ), to assess motion strength, temporal smoothness, and perceptual fidelity in videos. To complement these automatic metrics, we conduct a blind human preference study with 30 participants, each assigned a different subset of 10 test cases. For each test case, raters are presented with shuffled results from four anonymized methods and asked to select the best and second-best results in terms of Text Alignment (TA), Motion Dynamics (MD), and Stylization Effect (SE). The selected best and second-best results are assigned scores of 2 and 1, respectively, while the remaining results receive 0, and the scores are normalized per case.

Comparison Methods. We compare against three baselines: StyleMaster [50], StyleCrafter [21], and VideoComposer [41], collectively covering dedicated styl-

Table 1: Quantitative comparison on the style-guided text-to-video generation task. Best results are **bold** and second-best are underlined.

Method	Content Alignment		Motion Quality		Visual Quality	
	Text ↑	CSD ↑	DD ↑	DS ↑	IQ ↑	VQ ↑
VideoComposer [41]	0.1840	0.4956	0.3312	0.9816	0.6390	0.5837
StyleCrafter [21]	0.2968	0.4414	0.2922	0.9788	0.5974	0.5622
StyleMaster [50]	**0.3054**	0.3039	0.1710	0.9898	**0.6735**	**0.6451**
Ours	0.3010	**0.5219**	**0.8247**	**0.9912**	0.6568	0.6071

Table 2: Human preference study on the style-guided text-to-video generation task. Raters compare anonymized results from four methods and select the best and second-best results in terms of Text Alignment (TA), Motion Dynamics (MD), and Stylization Effect (SE). Best results are **bold** and second-best are underlined.

Method	TA ↑	MD ↑	SE ↑	Overall ↑
VideoComposer [41]	0.0111	0.0261	0.1184	0.0519
StyleCrafter [21]	0.1144	0.1351	0.1816	0.1437
StyleMaster [50]	0.2684	0.2284	0.1184	0.2051
Ours	**0.6061**	**0.6103**	**0.5815**	**0.5993**

ized video generation and general text-to-video generation with style conditioning. FreeVis [46] is excluded from quantitative comparison as its official implementation is not publicly available at the time of submission. All baselines are evaluated under their recommended inference settings without additional tuning.

4.2 Main Results

Quantitative results are summarized in Tab. 1 and Tab. 2, and qualitative comparisons are shown in Fig. 5. We evaluate on 462 prompt-style pairs against three baselines: StyleMaster [50], StyleCrafter [21], and VideoComposer [41].

Quantitative Results. As shown in Tab. 1, our method achieves the best CSD, DD, and DS among all methods, while ranking second in Text, IQ, and VQ. Specifically, our CSD reaches 0.5219, outperforming VideoComposer, StyleCrafter, and StyleMaster by 0.0263, 0.0805, and 0.2180, respectively, demonstrating the advantage of full-resolution token-level style injection for preserving fine-grained stylistic statistics. For motion quality, our DD reaches 0.8247, approximately 2.8× that of StyleCrafter and 4.8× that of StyleMaster, while our DS remains the highest at 0.9912, indicating that stronger motion does not sacrifice temporal smoothness. Although our IQ and VQ are slightly lower than StyleMaster by 0.0167 and 0.0380, respectively, this gap is likely related to the tendency of automatic perceptual metrics to favor low-motion or near-static videos. Our Text score is also close to the best result, with only a 0.0044 margin.

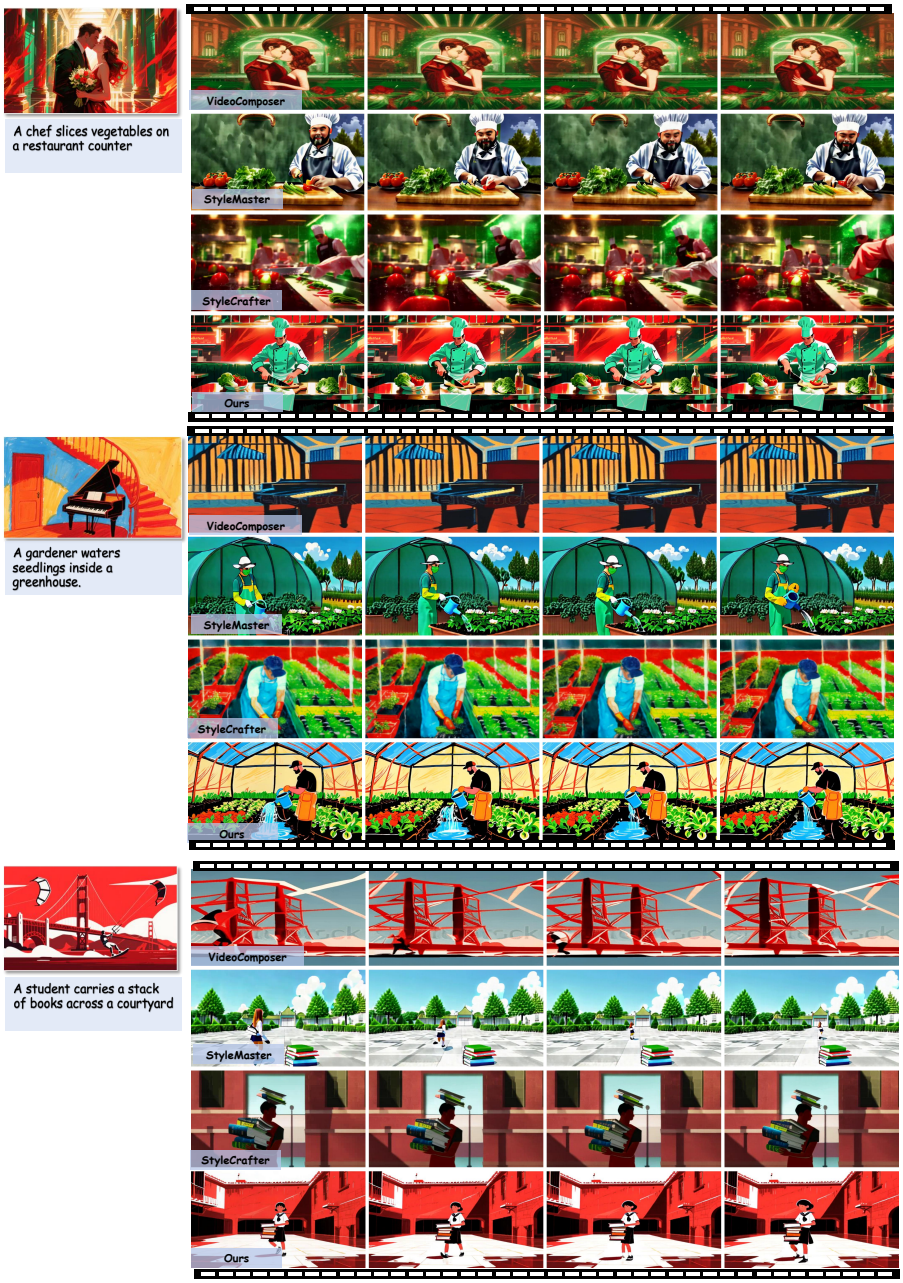


Fig. 5: Qualitative comparison with StyleMaster, StyleCrafter, and VideoComposer on style-guided text-to-video generation.

Table 3: Ablation results on style token injection design and Direct Preference Optimization. Best results are **bold** and second-best are underlined.

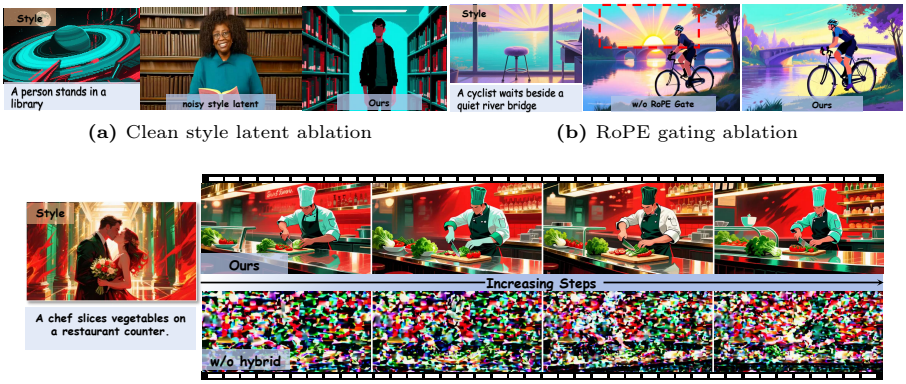
Setting	Content Alignment		Motion Quality		Visual Quality	
	CLIP ↑	CSD ↑	DD ↑	DS ↑	IQ ↑	VQ ↑
Explicit Injection	0.2992	0.3969	0.6472	0.9909	0.6981	**0.6557**
Noisy Latent	**0.3144**	0.1490	0.7619	0.9902	0.6678	0.6329
w/o GA	0.3037	0.4530	0.8225	0.9894	**0.6990**	0.6403
w/o RoPE Gating	0.2866	0.5031	0.8182	0.9888	0.6567	0.6136
w/o DPO	0.3063	0.4910	0.7792	0.9847	0.6649	0.6287
Ours	0.3010	**0.5219**	**0.8247**	**0.9912**	0.6568	0.6071

To complement automatic evaluation, we further conduct a blind human preference study in Tab. 2. Our method achieves the highest scores across TA, MD, SE, and Overall, with an Overall preference score of 0.5993, substantially surpassing StyleMaster, StyleCrafter, and VideoComposer. These results show that our method improves style preservation and motion dynamics while maintaining competitive semantic alignment and visual quality, and that the slight decrease in automatic IQ and VQ does not reduce perceptual preference.

Qualitative Results. Qualitative comparisons are shown in Fig. 5. Across diverse prompts and reference styles, existing methods exhibit distinct failure modes. VideoComposer tends to directly copy the content and spatial layout of the reference image, leading to weak prompt controllability and limited content diversity. StyleMaster shows relatively weak stylization and often produces noticeable wave-like motion artifacts, indicating that its generated dynamics are unstable and visually unnatural. StyleCrafter can introduce more visible motion in some cases, but often suffers from temporal inconsistency, object deformation, and unstable style rendering across frames. In contrast, our method preserves the color distribution, brushstroke patterns, and overall artistic characteristics of the reference image while generating coherent object and camera motion. These qualitative observations align with the improvements in CSD, DD, DS, and human preference scores in Tabs. 1 and 2, demonstrating that IGST supports stable style expression together with coherent motion modeling.

4.3 Ablation Study

Style Tokens Injection. We first evaluate whether MoStyle can improve existing style-conditioning paradigms by training only the cross-attention conditioning layers while keeping the StyleMaster encoder and backbone frozen, denoted as Explicit Injection. Compared with the original StyleMaster, Explicit Injection improves CSD from 0.3039 to 0.3969 and DD from 0.1710 to 0.6472, validating the effectiveness of temporally supervised stylized videos. However, it still lags behind Ours by 0.1250 in CSD and 0.1775 in DD, indicating that encoder-based explicit feature compression remains limited for fine-grained style preservation.



(a) Clean style latent ablation **(b)** RoPE gating ablation

(c) Hybrid training strategy ablation

Fig. 6: Qualitative ablation results. **(a)** Replacing the clean style latent with a noisy one weakens fine-grained style cues and destabilizes the stylized appearance. **(b)** Removing RoPE gating causes reference-specific structure to leak into the generated scene. **(c)** Video-only training leads to severe appearance collapse, while hybrid image–video training preserves temporal style consistency.

We then examine the role of clean style latents for token-level reference conditioning. As shown in Tab. 3, replacing the clean style latent with a noisy one sharply reduces CSD from 0.5219 to 0.1490, suggesting that diffusion noise corrupts fine-grained style information and weakens reference conditioning. Qualitative results in Fig. 6a further confirm that clean style latents provide more stable stylized appearance.

Finally, we analyze the architectural designs within IGST. Removing RoPE gating lowers the CLIP score from 0.3010 to 0.2866, indicating weaker content alignment when style tokens are assigned spatial rotary encoding. Qualitative results in Fig. 6b further show that this variant tends to copy reference-specific layouts or objects into the generated video. Removing head-wise gated attention reduces CSD from 0.5219 to 0.4530, despite competitive DD and IQ scores. These results show that positional decoupling and head-wise regulation play complementary roles: the former helps preserve content structure, while the latter improves faithful style statistics without sacrificing motion quality.

Training Strategy and Preference Alignment. We evaluate image-video mixed supervision by comparing it with a video-only variant. As shown in Fig. 6c, video-only training leads to unstable generations and severe appearance collapse, whereas mixed supervision preserves coherent content structures and strong style cues. This suggests that image-level style supervision serves as a distributional anchor and helps prevent style drift when temporal modeling is introduced.

We further compare our full model with a w/o DPO variant in Tab. 3. Removing Diffusion-DPO reduces CSD from 0.5219 to 0.4910 and DD from 0.8247 to 0.7792, showing that preference alignment provides complementary supervision beyond SFT.

5 Conclusion

We present a unified pipeline for stylized video generation that addresses motion degradation and encoder-based style compression through dataset construction, architectural design, and training strategy. MoStyle-5K provides explicit temporal supervision via decoupled content–motion synthesis, enabling stable style token injection without eroding the backbone's temporal priors. IGST bypasses the encoder bottleneck by directly concatenating style patches as tokens, with RoPE-gated style tokens and head-wise gated attention to suppress structural contamination. Hybrid image–video training and Diffusion-DPO post-training further enhance style fidelity and motion coherence. Experiments demonstrate that these components operate synergistically and collectively yield substantial improvements over prior baselines in both style consistency and dynamic quality.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62506280, U22A2096, 62576261 and 62571395, in part by the Young Talent Fund of Xi'an Association for Science and Technology under Grant 0959202613085, in part by the China Postdoctoral Science Foundation under Grant Number 2025M771559, in part by Scientific and Technological Innovation Teams in Shaanxi Province under grant 2025RS-CXTD-011, in part by the Shaanxi Province Core Technology Research and Development Project under grant 2024QY2-GJHX-11, in part by the Establishment Fund of the State Key Laboratory of Wakefulness/Sleep and Cognition(Chongqing Institute for Brain and Intelligence), in part by the Fundamental Research Funds for the Central Universities under Grant QTZX26138.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Chen, W., Ji, Y., Wu, J., Wu, H., Xie, P., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. arXiv preprint arXiv:2305.13840 (2023)
5. Chung, J., Hyun, S., Heo, J.P.: Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In: CVPR. pp. 8795–8805 (2024)
6. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: Stytr2: Image style transfer with transformers. In: CVPR. pp. 11326–11336 (2022)
7. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. In: ICLR (2017)
8. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR. pp. 2414–2423 (2016)
9. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. In: ICLR (2024)
10. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: ICLR (2024)
11. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for video diffusion models. In: ICLR (2025)
12. Hu, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: CVPR. pp. 8153–8163 (2024)
13. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV. pp. 1501–1510 (2017)
14. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR. pp. 21807–21818 (2024)
15. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: ICCV. pp. 17191–17202 (2025)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. pp. 694–711. Springer (2016)
17. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
18. Ku, M., Wei, C., Ren, W., Yang, H., Chen, W.: Anyv2v: A tuning-free framework for any video-to-video editing tasks. TMLR (2024)
19. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML. pp. 19730–19742. PMLR (2023)

20. Li, M., Chen, J., Zhao, S., Feng, W., Tu, P., He, Q.: Dreamstyle: A unified framework for video stylization. arXiv preprint arXiv:2601.02785 (2026)
21. Liu, G., Xia, M., Zhang, Y., Chen, H., Xing, J., Wang, Y., Wang, X., Shan, Y., Yang, Y.: Stylecrafter: Taming artistic video diffusion with reference-augmented adapter learning. ACM TOG **43**(6), 1–10 (2024)
22. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In: ICCV. pp. 6649–6658 (2021)
23. Luo, Y., Shi, X., Bai, J., Xia, M., Xue, T., Wang, X., Wan, P., Zhang, D., Gai, K.: Camclonemaster: Enabling reference-based camera control for video generation. In: SIGGRAPH Asia. pp. 1–10 (2025)
24. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: AAAI. vol. 38, pp. 4117–4125 (2024)
25. Ouyang, W., Dong, Y., Yang, L., Si, J., Pan, X.: I2vedit: First-frame-guided video editing via image-to-video diffusion models. In: SIGGRAPH Asia. pp. 1–11 (2024)
26. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: CVPR. pp. 5880–5888 (2019)
27. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
29. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: NeurIPS. vol. 36, pp. 53728–53741 (2023)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
31. Rout, L., Chen, Y., Ruiz, N., Kumar, A., Caramanis, C., Shakkottai, S., Chu, W.S.: Rb-modulation: Training-free stylization using reference-based modulation. In: ICLR (2025)
32. Somepalli, G., Gupta, A., Gupta, K., Palta, S., Goldblum, M., Geiping, J., Shrivastava, A., Goldstein, T.: Measuring style similarity in diffusion models. arXiv preprint arXiv:2404.01292 (2024)
33. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
34. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
35. Tan, W.R., Chan, C.S., Aguirre, H., Tanaka, K.: Improved artgan for conditional synthesis of natural image and artwork. IEEE TIP **28**(1), 394–409 (2019)
36. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: CVPR. pp. 8228–8238 (2024)
37. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
38. Wang, H., Spinelli, M., Wang, Q., Bai, X., Qin, Z., Chen, A.: Instantstyle: Free lunch towards style-preserving in text-to-image generation. arXiv preprint arXiv:2404.02733 (2024)
39. Wang, H., Xing, P., Huang, R., Ai, H., Wang, Q., Bai, X.: Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. arXiv preprint arXiv:2407.00788 (2024)

40. Wang, W., Yang, Y.: Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In: Advances in Neural Information Processing Systems (2024)
41. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. In: NeurIPS. vol. 36, pp. 7594–7611 (2023)
42. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH. pp. 1–11 (2024)
43. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
44. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: ICCV. pp. 7623–7633 (2023)
45. Xing, P., Wang, H., Sun, Y., Wang, Q., Bai, X., Ai, H., Huang, R., Li, Z.: Csgo: Content-style composition in text-to-image generation. In: Advances in Neural Information Processing Systems (2025)
46. Xu, J., Mei, Y., Zhang, K., Patel, V.M.: Freevis: Training-free video stylization with inconsistent references. In: International Conference on Learning Representations (2026)
47. Xu, R., Xi, W., Wang, X., Mao, Y., Cheng, Z.: Stylessp: Sampling startpoint enhancement for training-free diffusion-based method for style transfer. In: CVPR. pp. 18260–18269 (2025)
48. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Gu, X., Zhang, Y., Wang, W., Cheng, Y., Liu, T., Xu, B., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR (2025)
49. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
50. Ye, Z., Huang, H., Wang, X., Wan, P., Zhang, D., Luo, W.: Stylemaster: Stylize your video with artistic generation and translation. In: CVPR. pp. 2630–2640 (2025)
51. Zhang, S., Yang, X., Zi, B., Huang, H., Zhang, C., Li, X.: Telestyle: Content-preserving style transfer in images and videos. arXiv preprint arXiv:2601.20175 (2026)
52. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J.W., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. In: ECCV. pp. 273–290. Springer (2024)