FlowStyler: Artistic Video Stylization via Transformation Fields Transports

Yuning Gong
Sichuan University, China
20213230400003@stu.scu.edu.cn

Jiaming Chen
Sichuan University, China
2024226040003@stu.scu.edu.cn

Xiaohua Ren
University of Macau, China
xiaohua_ren@126.com

Yuanjun Liao
Sichuan University, China
20233230400029@stu.scu.edu.cn

Yanci Zhang *†
Sichuan University, China
yczhang@scu.edu.cn

Abstract

Contemporary video stylization approaches struggle to achieve artistic stylization while preserving temporal consistency. While generator-based methods produce visually striking stylized results, they suffer from flickering artifacts in dynamic motion scenarios and require prohibitive computational resources. Conversely, non-generative techniques frequently show either temporal inconsistency or inadequate style preservation. We address these limitations by adapting the physics-inspired transport principles from the Transport-based Neural Style Transfer (TNST) framework (originally developed for volumetric fluid stylization) to enforce inter-frame consistency in video stylization. Our framework employs two complementary transformation fields for artistic stylization: a geometric stylization velocity field governing deformation and an orthogonality-regularized color transfer field managing color adaptations. We further strengthen temporal consistency through two key enhancements to our field architecture: a momentum-preserving strategy mitigating vibration artifacts, and an occlusion-aware temporal lookup strategy addressing motion trailing artifacts. Extensive experiments demonstrate FlowStyler's superior performance across dual dimensions: Compared to generator-based approaches, we achieve 4× lower short-term warping errors, while maintaining comparable style fidelity; Against non-generative methods, FlowStyler attains 22% higher style fidelity with slightly improved temporal stability.

1. Introduction

Modern video stylization methods struggle with stylizing artistic features while maintaining temporal consistency. While diffusion-based models demonstrate remarkable image-to-image capabilities, their video stylization extensions ([15, 26]) face critical limitations: pronounced temporal inconsistency and blur artifacts in highly dynamic scenes. Their GPU memory demands also approximately scale linearly with video length, limiting practical use to short clips. Meanwhile, non-generative approaches face dual tradeoffs: those enforcing optical flow constraints (e.g., [19, 38]) exhibit residual flickering, while methods adopting relaxed style losses (e.g., [48, 50]) excessively restrict artistic expression, reducing outputs to simplistic filter-like effects.

These observations motivate us to address the challenge of achieving temporally consistent artistic stylization while maintaining reasonable GPU memory utilization.

To resolve these challenges, we present **FlowStyler** —a physics-inspired **non-generative** framework that extends the principles of Transport-based Neural Style Transfer (TNST) [22] (originally developed for volumetric fluid stylization) to artistic video stylization. Our core technical innovation reformulates video stylization as a synergistic evolution process of two transformation fields: a stylization velocity field that directs continuous geometric transformations, and an orthogonality-regularized color transfer field that performs localized spectral adaptations. This dual-field mechanism concurrently addresses the key limitations of conventional non-generative approaches in both *artistic stylization* and *temporal consistency*. By fundamentally eschewing diffusion models, our approach achieves *substantial GPU memory reduction* without compromising artistic expressiveness.

*National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation
†Corresponding author.

Building upon this field architecture, we further enhance temporal consistency through two critical augmentations:

- *Momentum-Preserving Fields Optimization*: While our formulation inherently supports stable evolution, discrete inter-frame optimization introduces vibration artifacts. Our solution transfers gradient momentum across temporal steps to dampen these vibrations while preserving valid deformation dynamics.
- *Occlusion-Aware Temporal Field Lookup*: Dynamic field propagation suffers parameter discontinuities at occlusion boundaries. This mechanism maintains parameter continuity by adaptively retrieving the last valid field parameters from visible historical data.

Experimental results demonstrate two key advantages: Compared to generator-based approaches, our method achieves significantly enhanced temporal consistency with at least 4× reduction in short-term warping errors while maintaining comparable stylistic fidelity. When evaluated against non-generative baselines, FlowStyler exhibits superior artistic expressiveness as quantified by DINO [3] metrics with 22% improvement, along with slightly improved temporal consistency of warping error.

2. Related Work

2.1. Non-Generative Video Stylization

Optical flow estimation [40, 42, 51] has been extensively used for video stylization to maintain temporal consistency. The seminal work of [38] pioneered flow-based propagation of warping stylized results between frames with temporal consistency losses, establishing the fundamental paradigm of warping-based temporal consistency constraint that remains influential today. Subsequent methods like [19] employed optical flow and handcrafted patch correspondences to enhance consistency. Alternative strategies include feed-forward networks with flow-based temporal constraints [5, 11, 18] or attempts to reduce the sensitivity of geometrical stylization [46, 48]. [10] proposes a luminance warping constraint for temporal consistency preservation, but necessitates per-style model retraining.

Several approaches attempt to circumvent optical flow computation entirely[11, 43]. [43] proposes to propagate stylized features from few keyframes using Few-Shot Patch-Based Training. [11] adopts two additional trained networks to improve temporal consistency. [6] distilled flow knowledge into test-time networks, while [8] employed multi-channel correlation. [50] addresses global consistency through local contrastive learning. Layer decomposition methods [20, 31, 53] separate videos into canonical and deformation fields. Recent innovations like using hash encoding and specialized regularization [34] achieve 200× faster convergence than [20]. However, these decomposition-based methods struggle with complex motions and scene dynamics, particularly failing in cases with significant motion in videos as identified in [30]. [14] proposes a unified domain interaction transformer that enables joint image-video style transfer through cross-domain contextual learning, eliminating the need for separate models. Recent research [13, 29, 39] attempts to convert videos into dynamic NeRF[33] or 3D Gaussian splatting[21] representations for stylization, as their 3D representations inherently guarantee consistency. However, they face difficulties in handling long videos and complex motion, as it is challenging to reconstruct them as dynamic NeRF or 3DGS models.

2.2. Generator-Based Video Stylization

Recent advances in diffusion models have significantly advanced deep learning-based video stylization. Early approaches like [4, 47, 49] established appearance consistency through cross-frame attention mechanisms that typically anchor anchoring to the first or previous frame.

Subsequent works enhanced temporal consistency through optical flow supervision, with methods like [7, 17, 52] explicitly aligning corresponding regions across frames using flow-guided constraints. Alternative approaches focus on enhancing consistency through different strategies. AnyV2V [26] proposes using an edited first frame and adapting an image-to-video model to stylize the video through temporal feature injection from the first frame. TokenFlow [12] enforces semantic correspondence through diffusion feature matching, demonstrating particular effectiveness in preserving object identities during drastic viewpoint changes. ControlVideo [55] extends image-level ControlNet to videos via full cross-frame attention.

2.3. Transport-Based Neural Style Transfer (TNST)

Originating from fluid style transfer [22], TNST employs stylization velocity fields that transfer with scene motion. The framework has evolved through Lagrangian formulations [23], commercial implementations [1, 16]. Besides, TSDN[13] further proposes to adopt the TNST to dynaic NeRF[9], enabling artistic geometrical stylization of dynamic NeRF scenes. Our work builds upon TNST's core philosophy of field-based stylization and transport, but introduces novel adaptations for video-specific temporal consistency and artistic flexibility.

3. Method

Our FlowStyler framework establishes a three-fold solution to enable artistic stylization while maintaining temporal consistency:

- *A complementary transformation field architecture* employs an orthogonality-regularized color transfer field and a stylization velocity field (Sec. 3.1) to enable artistic stylization, using TNST-inspired motion-aligned field propagation preserving temporal consistency (Sec. 3.2).

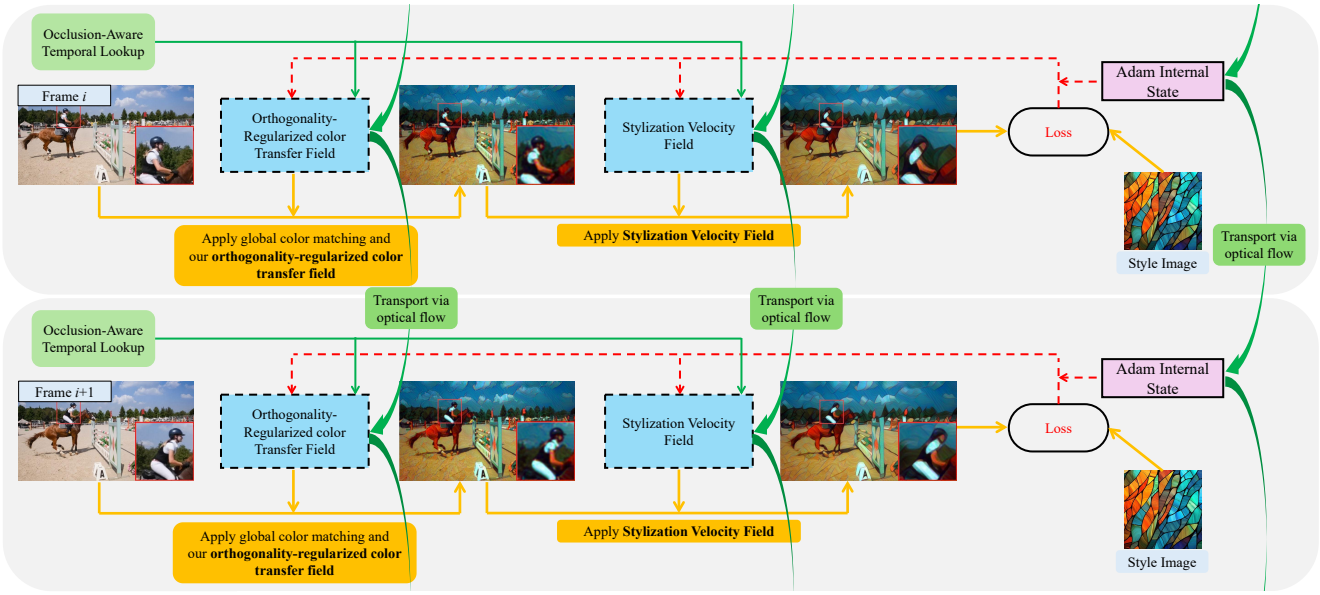

Figure 1. **FlowStyler Pipeline**: The pipeline begins by applying global color matching (Sec.3.1.3) combined with an orthogonality-regularized color transfer field (Sec.3.1.1) to the content frame, followed by processing through the stylization velocity field (Sec.3.1.2) to generate the final stylized results (orange lines). The outputs are then evaluated through loss functions, with gradients backpropagated to jointly optimize both transformation fields (red dot lines). Following iterative optimization (Sec.3.3), both the transformation fields and Adam optimizer states are propagated between frames (Sec.3.2) to enforce temporal consistency (green arrows). Finally, an occlusion-aware temporal lookup mechanism (Sec.3.4) retrieves occluded regions using historical data to eliminate trailing artifacts (green lines).

- *A momentum-preserving optimization paradigm* that maintains the Adam optimizer's momentum states across sequential frames via joint field propagation, effectively dampening vibration artifacts caused by per-frame optimizer reinitialization (Sec. 3.3).
- *An occlusion-aware temporal lookup strategy* that addresses information loss through backward tracing of historical frames with normalized aggregation for occluded regions to eliminate trailing artifacts (Sec. 3.4).

Fig. 1 shows the FlowStyler framework's processing flow.

3.1. Single Frame Stylization via Velocity and Color Fields

To simultaneously achieve artistic stylization and temporal consistency, we explicitly formulate the stylization process into two complementary transformations fields: a color transformation field (Sec.3.1.1) and a geometry transformation field (Sec.3.1.2). This decomposition strategy provides two key advantages over previous approaches:

First, our formulation systematically addresses the inherent misalignment problem in optical flow-dependent methods. Previous approaches [19, 38] relying on optical flow constraints struggle with temporal consistency when handling artistic geometric deformations, as the estimated flow becomes progressively misaligned with the distorted stylized features (visualized in supplemental material Figure 3). Our framework circumvents this issue by propagating the stylization fields rather than directly manipulating pixel values, thereby inherently maintaining cross-frame alignment consistency (visualized in supplemental material Figure 4).

Second, the explicit geometric transformation substantially reduces the need for compromised style relaxation loss [46]. Some of the previous methods [48] employ relaxed stylization losses to maintain temporal consistency, which inevitably sacrifices artistic expression. Our explicit geometric transformation formulation enables geometric deformations while ensuring temporal consistency through field propagation.

Let the input content video be represented as $\mathcal{X} = \{\mathbf{I}_t\}_{t=1}^{N}$ with content frame $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$. Our complementary transformation field stylization framework is as follows:

3.1.1. Orthogonality-Regularized Color Transfer Field

To establish an optimal framework for propagating color transformations across sequential frames, we first analyze an intuitive approach that defines an additive color transformation field $\mathbf{A}_t \in \mathbb{R}^{H\times W\times 3}$ through:

$$\mathbf{I}_t^{\text{colored}}(i,j) = \mathbf{A}_t(i,j) + \mathbf{I}_t(i,j) \tag{1}$$

However, this intuitive approach proves suboptimal due to unconstrained transformations that induce over-stylization over successive frames, as shown in Fig. 7.

Our key insight addresses this by formulating color transformations as orthogonality-regularized mappings through a learnable matrix field $\mathbf{M}_t \in \mathbb{R}^{H\times W\times 3\times 3}$:

$$\mathbf{I}_t^{\text{colored}}(i,j) = \mathbf{M}_t(i,j)\mathbf{I}_t(i,j) \tag{2}$$

The orthogonality regularization is imposed via:

$$\mathcal{L}_{\text{ortho}}(\mathbf{M}_t) = \sum_{i,j} \|\mathbf{M}_t(i,j)^\top \mathbf{M}_t(i,j) - \mathbf{I}\|_F^2 \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The advantages of this design can be analyzed from two perspectives: First, illumination consistency preservation through vector norm conservation $\|\mathbf{M}_t(i,j)\mathbf{c}\| \approx \|\mathbf{c}\|$, ensuring photometric consistency across video sequences. Second, bounded color variation between frames via the implicit condition number constraint $\kappa(\mathbf{M}_t) \approx 1$, which naturally arises from orthogonal transformations. The mathematical foundation stems from analyzing color variation after field propagation:

$$\|\delta\mathbf{c}'\| = \|\mathbf{M}_t\delta\mathbf{c}\| \leq \kappa(\mathbf{M}_t)\|\delta\mathbf{c}\|$$

where $\kappa(\mathbf{M}_t) \approx 1$ ensures the stylized color variation magnitudes $\|\delta\mathbf{c}'\|$ closely match the original color variation magnitudes $\|\delta\mathbf{c}\|$, thereby maintaining illumination changes consistent with the original content frames.

3.1.2. Stylization Velocity Field

Building upon previous Transport-based Neural Style Transfer (TNST) framework [1, 13, 22], we adopt a stylization velocity field to control geometric transformations in video stylization. Recent TNST approaches model content fields (e.g., NeRF representations or fluid simulations) as transported quantities guided by a stylization velocity field via semi-Lagrangian advection. A direct video adaptation of this paradigm would formulate the stylized frame as:

$$\mathbf{I}_t^{\text{style}} = \mathcal{A}(\mathbf{I}_t^{\text{colored}}, \mathbf{V}_t) \tag{4}$$

where $\mathbf{V}_t$ denotes the stylization velocity field and $\mathcal{A}$ represents the semi-lagrangian advection operator:

$$\mathcal{A}(\mathbf{E},\mathbf{V})(i,j) = \mathcal{S}\left(\mathbf{E},\ (i,j) - \mathbf{V}(i,j)\cdot\Delta t\right) \tag{5}$$

Here, $\mathbf{E}$ denotes the field to advect, $\mathbf{V}$ represents the velocity field, and $\mathcal{S}(\cdot)$ implements differentiable bilinear sampling. Notably, when $\Delta t = 1$, the semi-lagrangian advection operator $\mathcal{A}$ reduces to the backward warping operator [27]. However, this naive adaptation introduces critical artifacts in video applications: Geometric editing information becomes spatially decoupled from actual deformed locations, resulting in noise-prone displacement fields (visualized in supplemental material Figure 5). An example can be found in supplemental material Figure 8.

To overcome this challenge, we propose a lagrangian particle advection strategy that treats each pixel as a particle undergoing direct motion. More specifically, the geometric deformation is governed by a velocity field $\mathbf{V}_t \in \mathbb{R}^{H\times W\times 2}$, which induces pixel particle motion via forward advection. For a pixel $\mathbf{p}_t^{(i,j)}$, its stylized position $\hat{\mathbf{p}}_t^{(i,j)}$ evolves as:

$$\hat{\mathbf{p}}_t^{(i,j)} \leftarrow \mathbf{p}_t^{(i,j)} + \mathbf{V}_t(i,j)\cdot\Delta t \tag{6}$$

The virtual advection formulation allows flexible temporal scaling, though we fix $\Delta t = 1$ for implementation. Also, an adaptive splatting with kernel $K(\cdot)$ is used to enforce mass conservation:

$$\mathbf{I}_t^{\text{style}}(u,v) = \frac{\sum_{i,j} K\left(\|\hat{\mathbf{p}}_t^{(i,j)} - (u,v)\|\right)\cdot \mathbf{I}_t^{\text{colored}}(i,j)}{\sum_{i,j} K\left(\|\hat{\mathbf{p}}_t^{(i,j)} - (u,v)\|\right)} \tag{7}$$

This formulation also extends traditional forward advection with a large kernel support to suppress of high-frequency artifacts through spatial smoothing.

We also adopt a regularization to the velocity field $\mathbf{V}_t$ through total variation:

$$\mathcal{L}_{\text{TV}}(\mathbf{V}_t) = \sum_{i,j}\left(\|\nabla_x\mathbf{V}_t(i,j)\|_2^2 + \|\nabla_y\mathbf{V}_t(i,j)\|_2^2\right) \tag{8}$$

where ∇_x and ∇_y denote discrete spatial gradient operators along the image axes. This formulation suppresses high-frequency noise in the velocity field to maintain better temporal consistency.

3.1.3. Style Loss Formulation

Combining these components, we formulate the joint optimization objective:

$$\begin{aligned}\arg\min_{\mathbf{V}_t,\mathbf{M}_t} \mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Style}}(\mathbf{I}_t^{\text{style}}, \mathbf{S}) + \lambda_c\mathcal{L}_{\text{Content}}(\mathbf{I}_t^{\text{style}}, \mathbf{I}_t) \\ + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}}(\mathbf{M}_t) + \lambda_{\text{TV}}\mathcal{L}_{\text{TV}}(\mathbf{V}_t)\end{aligned} \tag{9}$$

where $\mathbf{S}$ denotes the style reference. The content preservation loss $\mathcal{L}_{\text{Content}}$ computes ℓ_2 distance between VGG-19 features of stylized and original frames, while $\mathcal{L}_{\text{Style}}$ employs Nearest Neighbor Feature Matching (NNFM) [25, 54]. We select NNFM over traditional Gram matrix-based loss because Gram matrices tends to produce blurry results in video sequences, whereas NNFM's local feature correspondence better preserves stylization sharpness. To help NNFM work better, we optionally apply global color matching methods, such as the linear color transformation [54] or color style transfer methods [28], as a preprocessing step (detailed in the supplemental material Section 8).

3.2. Temporal Consistency via Field Propagation

In the original TNST algorithm, temporal consistency is maintained by transferring the stylization fields via a physically simulated velocity field. However, video stylization

framework lacks such simulated velocity fields. To address this, we leverage optical flow to represent scene motion.

Given optical flow $\mathbf{F}_{t+1\to t} \in \mathbb{R}^{H\times W\times 2}$ between frames t and $t+1$, and computed visibility mask $\mathcal{M}^{\text{vis}}_{t+1\to t}$ (equivalently occlusion mask $\mathcal{M}^{\text{occ}}_{t+1\to t} = 1.0 - \mathcal{M}^{\text{vis}}_{t+1\to t}$) via forward-backward consistency [41], we propagate stylization fields through masked semi-lagrangian advection:

$$\mathbf{V}_{t+1} = \mathcal{M}^{\text{vis}}_{t+1\to t} \odot \mathcal{A}(\mathbf{V}_t, \mathbf{F}_{t+1\to t}) \tag{10}$$

$$\mathbf{M}_{t+1} = \mathcal{M}^{\text{vis}}_{t+1\to t} \odot \mathcal{A}(\mathbf{M}_t, \mathbf{F}_{t+1\to t}) \tag{11}$$

However, this basic propagation suffers from two key artifacts: vibration artifacts originating from inconsistent optimization direction across frames and motion trailing artifacts caused by information loss at occlusion boundaries. To solve these issues, we propose the following *momentum-preserving field optimization* (Sec.3.3) and *occlusion-aware temporal lookup strategy* (Sec.3.4). Due to space constraints, the complete algorithm workflow is provided in the supplemental material Algorithm 1.

3.3. Momentum-Preserving Field Optimization

We observe vibrating artifacts when sequentially finetuning stylization fields across video frames. A naive approach for stabilization spatially modulates learning rates by reducing rates for previously stylized regions while maintaining standard rates for new areas. However, this naive approach creates fundamental trade-offs: lower learning rates preserve temporal consistency but restrict style adaptation to varying viewpoints/distances. This challenge intensifies with camera movement, where persistent scene elements require consistent gradient updates for stable stylization. Simple rate reduction thus produces noisy stylized results.

To overcome such problems, our key insight stems from maintaining the temporal continuity of optimization direction in the Adam optimizer. Recall standard Adam's [24] update rule at iteration k of frame t operates as:

$$\begin{aligned}
\mathbf{m}_t^{(k)} &= \beta_1 \mathbf{m}_t^{(k-1)} + (1-\beta_1)\mathbf{g}_t^{(k)} \\
\mathbf{v}_t^{(k)} &= \beta_2 \mathbf{v}_t^{(k-1)} + (1-\beta_2)(\mathbf{g}_t^{(k)})^2 \\
\theta_t^{(k)} &= \theta_t^{(k-1)} - \eta \frac{\mathbf{m}_t^{(k)}}{\sqrt{\mathbf{v}_t^{(k)}} + \epsilon}
\end{aligned}$$

The vibration artifacts originate from two fundamental properties of Adam's optimization mechanics: (1) The momentum buffer $\mathbf{m}_t$ encodes historical gradient directions, guiding the stylization optimization direction in parameter space. (2) The variance estimate $\mathbf{v}_t$ adaptively scales update magnitudes based on gradient statistics. Resetting these states between frames disrupts the optimization continuity, leading to conflicting update directions and magnitudes across adjacent frames, resulting in vibrational artifacts.

To solve such problem, we propagate Adam's internal states between frames with our fields:

$$\mathbf{m}_{t+1}^{(0)} = \mathcal{M}^{\text{vis}}_{t+1\to t} \odot \mathcal{A}(\mathbf{m}_t^{(K)}, \mathbf{F}_{t+1\to t}) \tag{12}$$

$$\mathbf{v}_{t+1}^{(0)} = \mathcal{M}^{\text{vis}}_{t+1\to t} \odot \mathcal{A}(\mathbf{v}_t^{(K)}, \mathbf{F}_{t+1\to t}) \tag{13}$$

where K denotes the final optimization iteration in frame t. This propagation preserves both the momentum directionality and gradient variance statistics across consecutive frames. The momentum buffers $\mathbf{m}_t$ and $\mathbf{v}_t$ thereby accumulate gradient direction information continuously across both temporal and spatial dimensions.

For newly exposed regions (where $\mathcal{M}^{\text{occ}}_{t\to t-1}(i,j) = 1$), we implement a gradient attenuation strategy that preserves initialization stability while enabling gradual stylization:

$$\mathbf{g}_t^{(k)} \leftarrow \mathbf{g}_t^{(k)} \odot [\mathbf{1} - \alpha_t \mathcal{M}^{\text{occ}}_{t\to t-1}] \tag{14}$$

where α_t follows an exponential decay schedule, enabling smooth transition from initialization preservation to finetuning across optimization iterations.

3.4. Occlusion-Aware Temporal Lookup strategy

Motion trail artifacts originate from missing field parameters after occlusion. While 3D TNST preserves information via volumetric advection, our 2D parameterization suffers discontinuous propagation under occlusion. As moving objects occlude background elements, their deformation fields get discarded in subsequent frames, creating trailing artifacts along motion trajectories.

Our occlusion-aware reconstruction operates on pixels flagged by the occlusion mask $\mathcal{M}^{\text{occ}}_{t\to t+1}$. Through multi-frame temporal aggregation, we trace optical flow backwards to locate visible instances of occluded pixels in historical frames $k \in [t-K, t]$, validated by $\mathcal{M}^{\text{vis}}_{k\to t}(i,j) = 1$.

The critical design choice lies in the normalization factor: instead of dividing by the temporal window size K, we normalize by the sum of valid masks $\sum \mathcal{M}^{\text{vis}}_{k\to t}$. This adaptive normalization ensures that only *actually visible* historical samples contribute to the reconstruction. When multiple frames contain valid information (large mask sum), their contributions are averaged proportionally. When few frames are valid (small mask sum), the reconstruction automatically emphasizes available data while avoiding dilution by invalid samples. Combined with the ϵ-stabilized denom-

inator, the formulation achieves:

$$\mathbf{E}_t = \underbrace{\mathcal{M}_t^{\text{occ}} \cdot \left(\frac{\sum_{k=t-K}^{t} \mathcal{M}_{t\to k}^{\text{vis}} \cdot \mathcal{A}(\mathbf{E}_k, \mathbf{F}_{t\to k})}{\sum_{k=t-K}^{t} \mathcal{M}_{t\to k}^{\text{vis}} + \epsilon} \right)}_{\text{Occluded Regions}} + \underbrace{\mathcal{M}_t^{\text{vis}} \cdot \mathbf{V}_t}_{\text{Visible Regions}} \tag{15}$$

where $\epsilon = 10^{-6}$ prevents division-by-zero in completely occluded regions, $\mathbf{E}_t$ denotes the lookup field. We independently apply this computation both to $\mathbf{V}_t$ and $\mathbf{M}_t$.

4. Experiments

4.1. Comparison

To establish comparisons with two dominant algorithmic paradigms in video stylization, we evaluate against: (1) generative approaches including AnyV2V [26] and TokenFlow [12]; (2) non-generative methods represented by CAP-VSTNet [48] and UniST [14]. Our exclusion of CoDeF [34] stems from its limitations in dynamic motion reconstruction (see supplementary Fig. 6), where we observe blurry reconstruction during preliminary experiments. For AnyV2V, we employ their paper-recommended InstantStyle [45] for first-frame style initialization. Since TokenFlow operates as a text-prompt-driven method, we utilize GPT-4o for style-aligned text-prompt synthesis and use their officially suggested Plug-and-Play [44] integration scheme. Our evaluation dataset is drawn from the DAVIS 2017 [35] and 2019 [2] benchmarks, which comprises 60 image-video pairs with **moderate-to-strong dynamic motion with discernible camera movement**, designed to stress-test temporal consistency under challenging yet common scenarios. To enable fair comparison with generator-based methods, preventing their GPU memory overflows (common at HD resolutions), we spatially downsampled all content to 768×432 resolution. All experiments are conducted on NVIDIA A100 GPUs(80GB).

4.1.1. Qualitative comparison

Comparative results are visualized in Fig.2. While AnyV2V produces visually coherent results in scenes with limited camera motion (e.g., *Horse-jump-high*), it exhibits severe blurring artifacts under challenging camera movements as seen in *Parkour*. Our method maintains flicker-free and non-blurred results across both scenarios. TokenFlow effectively mitigates blurring artifacts but introduces temporal flickering and style misalignment (e.g., *Robot-battle* sequence), revealing inherent limitations of text-prompt-driven methods. Our approach overcomes these limitations, achieving superior temporal consistency with the reference style. Non-generative methods (CAP-VSTNet and UniST) demonstrate adequate temporal consistency but fail to preserve critical style characteristics, notably missing brushstroke details in *Surf* and geometric motifs in *Choreography*. In contrast, our method matches their temporal stability while preserving style elements through geometric stylization enabled by the stylization velocity field. Additional comparative examples are provided in the supplementary video material, which provide enhanced visualization of temporal consistency and style preservation characteristics.

4.1.2. Quantitative comparison

Our evaluation assesses three critical aspects in Tab. 1: (1) temporal consistency (measured at 1-frame and 10-frame intervals) through optical flow warping error [27], (2) style fidelity between stylized frames and the style image using DINO [3] and CLIP [36] similarity scores, (3) peak GPU memory allocation during 50-frame and 100-frame processing. We evaluate on aforementioned 60 challenging style-video pairs. We also conduct a user study following the Two-alternative Forced-Choice (2AFC) protocol [12]. Participants are presented with the style image and three synchronized videos: the original input sequence, our FlowStyler result, and a baseline methods' output. They are instructed to select which processed video better maintains temporal consistency while achieving the artistic stylization of the style image. From 328 participants, we collect around 2000 votes per baseline method, where preference rates (percentage favoring our method) are reported in Tab. 2.

Temporal Consistency: FlowStyler exhibits significantly lower warping errors compared to all baselines: AnyV2V shows 10×/2.2× higher short/long-term warping error, TokenFlow 4.4×/2.1×, and UniST 4.2×/2.1×, while CAP-VSTNet achieves the closest performance (1.6×/1.4×).

Style Fidelity: FlowStyler achieves superior style alignment via DINO/CLIP scores, demonstrating 18%/15% improvements over AnyV2V, 22%/31% enhancements versus UniST. CAP-VSTNet's relatively lower performance (24%/29% reductions) confirms that style relaxation losses critically degrade essential style elements. While TokenFlow reaches similar CLIP scores, its DINO deficit and qualitative comparisons exposes its suffering from the inherent limitations of text-guided paradigms.

GPU Efficiency: FlowStyler maintains a stable 3.6GB memory profile across 100-frame processing through its frame-wise processing paradigm. Our approach demonstrates 17.8× lower consumption than TokenFlow and 17.2× improvement over AnyV2V. While CAP-VSTNet (0.44GB) and UniST (1.1GB) achieve reduced footprints, their strategies fundamentally compromise style fidelity. Our practical memory consumption enables home deployment, whereas generator-based competitors' extreme mem-

Table 1. Quantitative Comparison

Method	Warping Error (10^{-3}) ↓		DINO Score ↑	CLIP Score ↑	Peak GPU Memory (GB) ↓	
	1-frame	10-frame			50 frames	100frames
AnyV2V	25.0	30.7	0.61	0.58	40	62
TokenFlow	11.0	29.5	0.66	0.65	24	64
CAP-VSTNet	4.1	19.8	0.58	0.52	**0.44**	**0.44**
UniST	10.6	29.8	0.59	0.51	1.1	1.1
FlowStyler	**2.5**	**14.2**	**0.72**	**0.67**	3.6	3.6

ory demands (>60GB) restrict them to data-center hardware.

User Preference: Our 2AFC study validates superior perceptual quality: 91% preference over AnyV2V, 79% against TokenFlow, 75% vs CAP-VSTNet, and 74% versus UniST. This demonstrates that FlowStyler achieves better quality-consistency tradeoff in video stylization.

Table 2. User Preference Comparison

Method	User Preference of our FlowStyler
AnyV2V	91%
TokenFlow	79%
CAP-VSTNet	75%
UniST	74%

4.2. Qualitative Ablation Studies

We present qualitative ablation studies in this section to visually demonstrate the impact of our core components. Due to space constraints, a detailed quantitative analysis is provided in the supplementary material.

4.2.1. Orthogonality-Regularized Color Transfer Field

To validate the effectiveness of our orthogonality-regularized color transfer field in preventing over-stylization, we conduct comparative studies in supplemental material Figure 7 between the naive additive transformation field (Eq. 1) and our proposed *orthogonality-regularized color transfer field* (Eq. 2). Qualitative comparisons demonstrate that the naive approach causes adjacent objects to merge after multiple frame iterations, inducing excessive artistic deformation that impairs scene comprehensibility. In contrast, our orthogonal-constrained method maintains clearer object boundaries while successfully retaining semantic integrity and avoiding over-stylization artifacts.

4.2.2. Lagragian advection

To investigate the noise suppression capability of our lagrangian advection, we perform a comparative analysis in supplemental material Figure 8 between conventional TNST semi-lagrangian advection Eq. 4 and our adopted lagrangian advection Eq. 6. Experimental results demonstrate that our advection formulation achieves superior noise reduction compared to the conventional TNST semi-lagrangian approach while preserving stylization fidelity.

4.2.3. Momentum-Preserving Field Optimization

To evaluate our momentum-preserving optimization, **we direct readers to the supplementary video for dynamic analysis, as static visualization is difficult for capturing temporal effects**. Comparative analysis in supplemental material Figure 9 reveals that the naive implementation lacking momentum preservation exhibit pronounced vibration artifacts during object motion. While simply reducing learning rates mitigates flickering, this compromise introduces undesirable blurring in stylized outputs. In contrast, our strategy maintains temporally Consistency stylized results across sequential frames.

4.2.4. Occlusion-Aware Temporal Lookup Strategy

To validate the efficacy of our occlusion-aware temporal lookup strategy in resolving motion trailing artifacts, we conduct an ablation study between the proposed method with and without this component. As revealed in supplemental material Figure 10, the naive implementation without temporal lookup reveals pronounced trailing artifacts along moving object trajectories. Our approach dynamically aggregates multi-frame historical data from visible regions, preserving appearance consistency in occluded areas.

5. Conclusions and Future Work

We present FlowStyler, a novel video stylization framework that achieves temporal consistency and artistic stylization through three key innovations: (1) a complementary transformation field architecture separating geometric and color transformations, (2) a momentum-preserving optimization strategy to avoid vibration, and (3) occlusion-aware temporal lookup to address trailing issues. Moving forward, we aim to enhance practical applicability through real-time optimization strategies, potentially via lightweight forward model approximators. While our style-image-based method

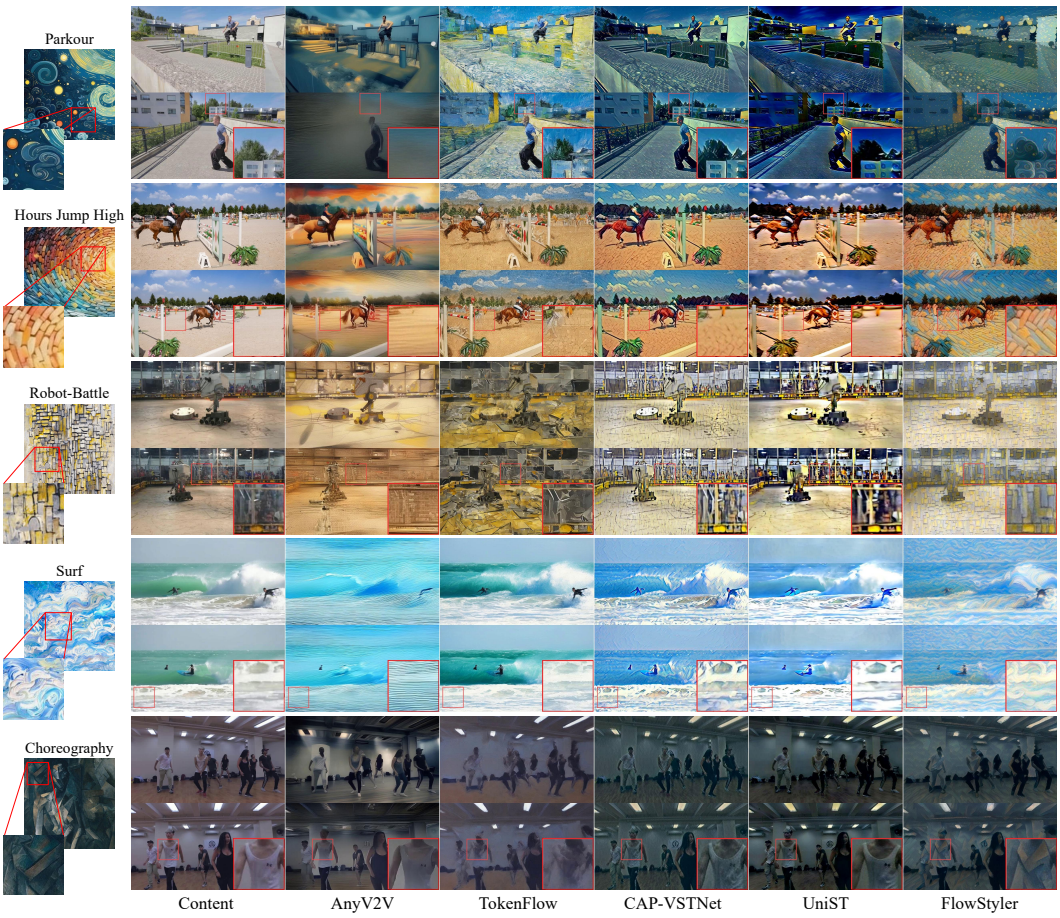

Figure 2. Qualitative comparison among AnyV2V [26], TokenFlow [12], CAP-VSTNet [48], UniST [14], and our FlowStyler. Our FlowStyler demonstrates superior style fidelity in capturing fine-grained artistic patterns (e.g., brushstrokes in *Surf*, geometric motifs in *Choreography*) while maintaining strict temporal consistency, outperforming state-of-the-art methods.

excels at texture and stroke transfer, we acknowledge its current limitations in text-driven stylization compared to CLIP-guided diffusion models. This stems from a design choice, as our VGG-based perceptual style loss targets low-level pattern matching. To expand our method's capabilities, we plan to integrate diffusion priors through Score Distillation Sampling (SDS) loss [32] to enable prompt-guided stylization. Our initial attempts produced blurry results due to gradient instability, a limitation we aim to resolve in future work.

6. Acknowledgments

This work was supported by the Open Research Project Program of the National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation.

References

[1] Joshua Aurand, Raphael Ortiz, Silvia Nauer, and Vinicius C Azevedo. Efficient Neural Style Transfer for Volumetric Simulations. *ACM Transactions on Graphics (TOG)*, 41(6): 1–10, 2022. Publisher: ACM New York, NY, USA. 2, 4

[2] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019. 6

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 6

[4] Duygu Ceylan, Chun-Hao Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. 2023. 2

[5] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. Coherent online video style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1105–1114, 2017. 2

[6] Xinghao Chen, Yiman Zhang, Yunhe Wang, Han Shu, Chunjing Xu, and Chang Xu. Optical flow distillation: Towards efficient and stable video style transfer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 614–630. Springer, 2020. 2

[7] Ernie Chu, Tzuhsuan Huang, Shuo-Yen Lin, and Jun-Cheng Chen. Medm: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1353–1361, 2024. 2

[8] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1210–1217, 2021. Issue: 2. 2

[9] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast Dynamic Radiance Fields with Time-Aware Neural Voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 2

[10] Chang Gao, Derun Gu, Fangjun Zhang, and Yizhou Yu. Reconet: Real-time coherent video style transfer network. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part VI 14*, pages 637–653. Springer, 2019. 2

[11] Wei Gao, Yijun Li, Yihang Yin, and Ming-Hsuan Yang. Fast video multi-style transfer. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3222–3230, 2020. 2

[12] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 6, 8

[13] Yuning Gong, Mingqing Song, Xiaohua Ren, Yuanjun Liao, and Yanci Zhang. TSDN: Transport-based Stylization for Dynamic NeRF. In *Pacific Graphics Conference Papers and Posters*. The Eurographics Association, 2024. 2, 4

[14] Bohai Gu, Heng Fan, and Libo Zhang. Two birds, one stone: A unified framework for joint learning of image and video style transfers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23545–23554, 2023. 2, 6, 8

[15] Samet Hicsonmez, Nermin Samet, Fidan Samet, Oguz Bakir, Emre Akbas, and Pinar Duygulu. Wait: Feature warping for animation to illustration video translation using gans. *arXiv preprint arXiv:2310.04901*, 2023. 1

[16] Jonathan Hoffman, Te Hu, Paul Kanyuk, Stephen Marshall, George Nguyen, Hope Schroers, and Patrick Witting. Creating Elemental Characters: From Sparks to Fire. In *ACM SIGGRAPH 2023 Talks*, pages 1–2. 2023. 2

[17] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023. 2

[18] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 783–791, 2017. 2

[19] Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sỳkora. Stylizing video by example. *ACM Trans. Graph.*, 38 (4):1–11, 2019. 1, 2, 3

[20] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Trans. Graph.*, 40(6):1–12, 2021. Publisher: ACM New York, NY, USA. 2

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2

[22] Byungsoo Kim, Vinicius Azevedo, Markus Gross, and Barbara Solenthaler. Transport-based neural style transfer for smoke simulations. *ACM Transactions on Graphics*, 38:1–11, 2019. 1, 2, 4

[23] Byungsoo Kim, Vinicius C. Azevedo, Markus Gross, and Barbara Solenthaler. Lagrangian Neural Style Transfer for Fluids. *ACM Trans. Graph.*, 39(4), 2020. Place: New York, NY, USA Publisher: Association for Computing Machinery. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] Nicholas Kolkin, Michal Kucera, Sylvain Paris, Daniel Sykora, Eli Shechtman, and Greg Shakhnarovich. Neural Neighbor Style Transfer, 2022. arXiv:2203.13215 [cs]. 4

[26] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. AnyV2V: A Tuning-Free Framework For Any Video-to-Video Editing Tasks. *Transactions on Machine Learning Research*, 2024. 1, 2, 6, 8

[27] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *European Conference on Computer Vision*, 2018. 4, 6

[28] Maria Larchenko, Alexander Lobashev, Dmitry Guskov, and Vladimir Vladimirovich Palyulin. Color style transfer with modulated flows. In *ICML 2024 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2024. 4, 2

[29] Xingyi Li, Zhiguo Cao, Yizheng Wu, Kewei Wang, Ke Xian, Zhe Wang, and Guosheng Lin. S-dyrf: Reference-based stylized radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20102–20112, 2024. 2

[30] Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8207–8216, 2024. 2

[31] Erika Lu, Forrester Cole, Tali Dekel, Weidi Xie, Andrew Zisserman, David Salesin, William T Freeman, and Michael Rubinstein. Layered neural rendering for retiming people in video. *arXiv preprint arXiv:2009.07833*, 2020. 2

[32] David McAllister, Songwei Ge, Jia-Bin Huang, David Jacobs, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Rethinking score distillation as a bridge between image distributions. *Advances in Neural Information Processing Systems*, 37:33779–33804, 2025. 8

[33] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[34] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8089–8099, 2024. 2, 6, 3

[35] 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:1704.00675*, 2017. 6

[36] 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*, pages 8748–8763. PmLR, 2021. 6

[37] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[38] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *Pattern Recognition: 38th German Conference, GCPR 2016, Hannover, Germany, September 12-15, 2016, Proceedings 38*, pages 26–36. Springer, 2016. 1, 2, 3

[39] Abhishek Saroha, Florian Hofherr, Mariia Gladkova, Cecilia Curreli, Or Litany, and Daniel Cremers. Zdyss–zero-shot dynamic scene stylization using gaussian splatting. *arXiv preprint arXiv:2501.03875*, 2025. 2

[40] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12469–12480, 2023. 2

[41] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European conference on computer vision*, pages 438–451. Springer, 2010. 5

[42] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2, 1

[43] Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menclei Chai, Sergey Tulyakov, and Daniel Sỳkora. Interactive video stylization using few-shot patch-based training. *ACM Trans. Graph.*, 39(4):73–1, 2020. 2

[44] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 6

[45] Haofan Wang, 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. 6

[46] Wenjing Wang, Jizheng Xu, Li Zhang, Yue Wang, and Jiaying Liu. Consistent video style transfer via compound regularization. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12233–12240, 2020. 2, 3

[47] Wen Wang, Yan Jiang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 2

[48] Linfeng Wen, Chengying Gao, and Changqing Zou. Cap-vstnet: Content affinity preserved versatile style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18300–18309, 2023. 1, 2, 3, 6, 8

[49] Bichen Wu, Ching-Yao Chuang, Xiaoyan Wang, Yichen Jia, Kapil Krishnakumar, Tong Xiao, Feng Liang, Licheng Yu, and Peter Vajda. Fairy: Fast parallelized instruction-guided video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8261–8270, 2024. 2

[50] Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai. Ccpl: Contrastive coherence preserving loss for versatile style transfer. In *European Conference on Computer Vision*, pages 189–206. Springer, 2022. 1, 2

[51] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022. 2

[52] Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. In *ACM SIGGRAPH Asia Conference Proceedings*, 2023. 2

[53] Vickie Ye, Zhengqi Li, Richard Tucker, Angjoo Kanazawa, and Noah Snavely. Deformable sprites for unsupervised video decomposition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[54] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *European Conference on Computer Vision*, pages 717–733. Springer, 2022. 4, 1

[55] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. ControlVideo: Training-free Controllable Text-to-Video Generation. *arXiv preprint arXiv:2305.13077*, 2023. 2