TRACKLUT: SEMANTIC COLOR TRACKING FOR ADAPTIVE VIDEO COLOR TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-based video color transfer remains challenging in dynamic scenes, where color transformations must adapt to evolving content and illumination while avoiding temporal flicker. Existing keyframe-based approaches provide limited adaptability, whereas independent frame-wise predictions often sacrifice temporal consistency. To address these limitations, we propose TrackLUT, a frame-adaptive framework that predicts a 3D lookup table (LUT) for each frame and exploits semantic correspondence for both spatial color transfer and temporal propagation. Specifically, semantic features establish reliable correspondences between target content and reference regions, while decoupled color features encode transferable appearance cues. This design enables selective color retrieval from semantically relevant regions, reducing erroneous mappings caused by spatial and appearance discrepancies. To maintain temporal coherence, we further align the historical color state with each incoming frame through semantic correspondence, allowing relevant appearance information to propagate as the scene evolves. The retrieved reference features and aligned temporal state are jointly decoded into a frame-specific LUT. Experiments demonstrate that TrackLUT adapts effectively to substantial content and illumination changes while preserving reference-style fidelity, structural consistency, and temporal stability, enabling efficient real-time video color transfer.

1 INTRODUCTION

Color retouching plays a pivotal role in visual storytelling and video production, as it shapes the perceived style, mood, and atmosphere of visual content. However, precise color adjustment requires substantial expertise and manual effort because of the complex dependencies among color channels and the need to maintain consistent appearance across dynamic scenes. Reference-based color transfer offers an efficient alternative by transferring the desired color characteristics from a reference image to an input video while preserving its underlying structure Reinhard et al. (2001); Ho & Zhou (2021); Ke et al. (2023). 3D Look-Up Tables (LUTs) are well suited to this task because they provide efficient nonlinear color mapping with negligible rendering overhead and preserve spatial structure Zeng et al. (2020); Yang et al. (2022); Li et al. (2025). However, existing LUT-based video color transfer methods typically follow a static one-video-one-LUT paradigm, where a LUT estimated from one or several representative frames is reused throughout the sequence Chen et al. (2023); Shin et al. (2025); Li et al. (2025). Although temporally stable, this strategy assumes that a single color mapping remains valid as scene content, subjects, and illumination change, which is often violated in real videos.

We therefore study real-time frame-adaptive LUT prediction for reference-based video color transfer, where the LUT is predicted online and adapts to the current frame. To the best of our knowledge, we are the first to study online frame-adaptive 3D LUT prediction with explicit temporal modeling for reference-based video color transfer. The key challenge is not simply to predict a different LUT for each frame, but to answer two questions jointly: what color state should the current LUT adapt to, and how should this state evolve over time? The first requires reliable correspondence between the current frame and the reference image, while the second requires temporal constraints across consecutive frames.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

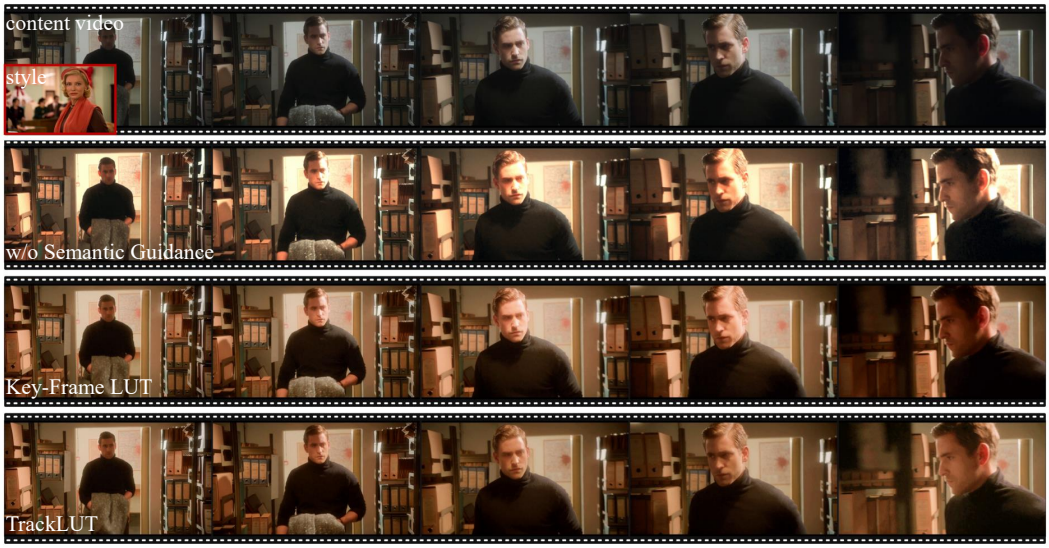

Figure 1: **Qualitative comparison of reference-based video color transfer.** Without semantic guidance, the transferred colors are less consistent across corresponding regions, while a fixed key-frame LUT cannot adapt to frame-wise content changes. TrackLUT combines semantic-guided color retrieval with temporal adaptation to achieve consistent and adaptive color transfer across frames.

For spatial adaptation, a standard 3D LUT defines a global RGB-to-RGB mapping and does not explicitly distinguish semantic regions such as people, sky, vegetation, or buildings Gong et al. (2025). When the current frame and reference differ in content or spatial layout, global color matching can therefore introduce incorrect color associations. We address this by decoupling semantic matching from color representation and retrieving reference color information according to semantic correspondence. For temporal adaptation, independently predicting a LUT for each frame can turn small feature variations into visible color flickering Lai et al. (2018); She & Xu (2022). We therefore maintain historical semantic representations together with their corresponding color states and retrieve temporally relevant historical information for the current frame. In this formulation, semantic correspondence determines the target of LUT adaptation, while historical information constrains how the color state evolves over time. As illustrated in Fig. 1, existing approaches either lack semantic correspondence for reliable reference-color transfer or rely on a fixed LUT that cannot adapt to frame-wise variations, whereas TrackLUT combines semantic-guided color retrieval with temporal adaptation.

Based on this idea, we propose TrackLUT, a real-time framework for semantic color tracking and frame-adaptive video color transfer. For each frame, TrackLUT retrieves content-relevant color information from the reference image through explicit semantic alignment and aligns historical color states with the current content to provide temporally consistent guidance. The retrieved reference features and aligned historical states are then jointly decoded into a frame-specific 3D LUT. By coupling semantic-guided reference retrieval with temporal color propagation, TrackLUT adapts to evolving content and illumination while maintaining temporal coherence. Our key contributions are summarized as follows:

- We introduce TrackLUT, an online reference-based video color transfer framework that uses semantic correspondence to guide the estimation and temporal updating of frame-adaptive global 3D LUTs.
- We propose semantic-guided reference retrieval that decouples correspondence estimation from color representation, providing semantically relevant reference color cues for predicting each frame's global color mapping.
- We develop a temporal update mechanism that aligns historical color states through semantic correspondence and incorporates gated fusion of LUT latents, encouraging temporally coherent color mappings while retaining adaptability to evolving content and illumination.

2 RELATED WORKS

Reference-based Color Transfer. Early color transfer methods match low-order statistics or complete color distributions between source and reference images Reinhard et al. (2001); Pitie et al. (2005); Rabin et al. (2014). Subsequent photorealistic style transfer methods employ feature-space transformations while explicitly preserving content structures Luan et al. (2017); Li et al. (2018); Wu et al. (2022); Wen et al. (2023). More recent approaches represent the transferred appearance through explicit presets or compact color mappings, enabling efficient and structure-preserving color manipulation Ho & Zhou (2021); Ke et al. (2023); Lin et al. (2023). Meanwhile, local and correspondence-aware methods exploit spatial attention or semantic relationships to establish more reliable content-reference associations Liu et al. (2021); Yang et al. (2024); Gong et al. (2025). Despite these advances, existing methods primarily focus on spatial adaptation for individual images, whereas our work investigates how reference-conditioned color mappings should dynamically adapt and evolve over time in videos.

LUT-based Video Color Transfer and Grading. LUT-based methods provide an efficient solution for video color transfer by representing appearance changes as compact color mappings. Existing approaches typically estimate a 3D LUT from representative video content and a reference image, then reuse it across the sequence. NLUT Chen et al. (2023) generates a stylized LUT from a selected keyframe and reference image, while D-LUT Li et al. (2025) derives reference-conditioned 3D LUTs for direct video application. VideoColorGrading Shin et al. (2025) further formulate reference-based video color grading as explicit LUT generation. Although effective for high-resolution video processing, these methods generally rely on a video-level transformation estimated from limited representative content, which may become suboptimal as scene composition, subjects, and illumination change.

Temporal Context and Memory Modeling in Video Processing. Recent video processing methods increasingly exploit cross-frame context rather than treating frames independently. Transformer-based approaches model short- and long-range dependencies through spatio-temporal attention or cross-frame feature interaction Guo et al. (2024); Li et al. (2024); Wu et al. (2024). Other methods explicitly maintain historical states or memory banks to selectively reuse informative past information during online processing Zhou et al. (2024); Wang et al. (2025). Such mechanisms have shown that effective temporal modeling depends not only on aggregating neighboring frames, but also on identifying which historical information remains relevant to the current content.

3 METHODS

3.1 OVERVIEW

Given a content video $\mathcal{I} = \{I_t\}_{t=1}^{T}$ and a reference I_s, our framework predicts frame-adaptive 3D LUTs $\mathbf{L}_t$ for structure-preserving appearance transfer. As shown in Fig. 2, frozen DINOv2 Oquab et al. (2023) and lightweight color encoders extract decoupled semantic and color features. Semantic-guided cross-attention retrieves reference colors, while semantic matching aligns historical color states for temporal fusion (Sec. 3.2). The fused state is decoded into a frame-specific LUT, where gated latent fusion further enforces temporal coherence (Sec. 3.3). Finally, trilinear interpolation applies the LUT to yield the stylized output Y_t.

3.2 SEMANTIC-GUIDED COLOR TRACKING

Our color extraction module features three stages: decoupled semantic and color encoding, semantic-guided reference retrieval, and history-aligned temporal fusion. This enables semantic-based reference retrieval while maintaining a temporally consistent color state.

Dual-Branch Feature Encoding. Given the current frame I_t and reference image I_s, we employ two complementary encoders to separate correspondence estimation from color representation. A pretrained and frozen DINOv2 encoder E_s extracts semantic features, while a lightweight trainable ViT Dosovitskiy et al. (2021) encoder E_c learns task-specific color features:

$$S_t = E_s(I_t), \quad S_s = E_s(I_s), \qquad C_t = E_c(I_t), \quad C_s = E_c(I_s). \tag{1}$$

Each feature sequence contains a global CLS token and spatial patch tokens.

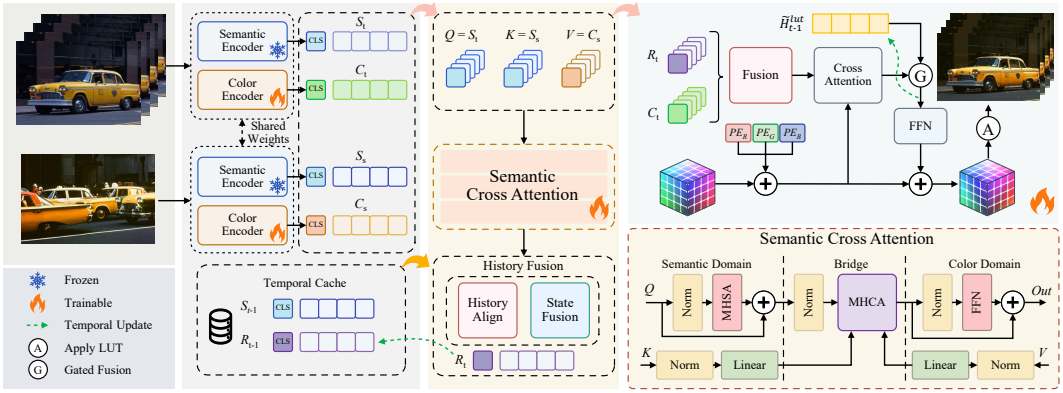

Figure 2: **Overview of TrackLUT.** Semantic and color encoders decouple correspondence estimation from appearance representation. Semantic cross-attention retrieves reference colors for the current frame, while history alignment and state fusion incorporate temporally consistent color cues from previous frames. The fused features are decoded into a frame-adaptive 3D LUT, with gated latent propagation further stabilizing its temporal evolution.

Semantic-Guided Reference Retrieval. We retrieve reference colors according to semantic correspondence using a semantic-guided decoder. The current semantic tokens are first contextualized by self-attention, yielding $\widetilde{S}_t$. For each attention head, cross-attention is then computed as

$$Q = W_Q\,\mathrm{LN}(\widetilde{S}_t), \qquad K = W_K\,\mathrm{LN}(S_s), \qquad V = W_V\,\mathrm{LN}(C_s), \tag{2}$$

$$A_t^s = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right), \qquad R_t^s = \mathrm{FFN}(A_t^s V). \tag{3}$$

Since Q and K are derived from semantic features while V originates from the reference color features, semantic correspondence determines where to retrieve and color features determine what appearance information is transferred.

History-Aligned Fusion. To mitigate temporal fluctuations from independent frame-wise retrieval, we maintain an accumulated color state $\bar{R}_t$. For each current-frame patch i, we find the most semantically relevant patch j_i^* in the previous frame:

$$j_i^* = \arg\max_j \cos(S_t^i, S_{t-1}^j). \tag{4}$$

The aligned historical feature is retrieved from the previous accumulated state as

$$H_t^i = \bar{R}_{t-1}^{j_i^*}. \tag{5}$$

Rather than assigning the same temporal weight to all tokens, we determine the history contribution from the raw semantic similarity of each matched pair:

$$m_t^i = \cos(S_t^i, S_{t-1}^{j_i^*}), \qquad \alpha_t^i = \alpha_p \sigma\left(\frac{m_t^i - \theta}{\gamma}\right), \tag{6}$$

where $\alpha_p \in [0,1]$ sets the maximum weight assigned to the aligned historical color feature, θ controls the semantic similarity threshold, and $\gamma > 0$ controls the transition sharpness. The color state is then updated through token-wise fusion:

$$\bar{R}_t^i = (1 - \alpha_t^i) R_t^{s,i} + \alpha_t^i H_t^i. \tag{7}$$

Reliable correspondences therefore preserve more historical information, whereas uncertain matches rely primarily on the current reference retrieval.

3.3 Color Fusion and Frame-Adaptive LUT Prediction

Given the current-frame color feature C_t and the temporally stabilized reference feature $\bar{R}_t$, we first combine them into a unified representation for LUT prediction. We then decode a frame-adaptive 3D LUT using learnable LUT queries and further regularize its temporal evolution in the latent space.

Color Feature Fusion. We fuse the current appearance and temporally aligned reference information as

$$F_t = \Phi\left(\mathrm{Concat}(C_t, \bar{R}_t)\right), \tag{8}$$

where Φ is a lightweight fusion network. The resulting representation jointly captures the current-frame appearance and the reference color information accumulated over time, and serves as the conditioning feature for LUT prediction.

Query-Based LUT Decoding. Following StatLUT Wang et al. (2026), we associate each lattice point of an identity 3D LUT with an individual query. Let $\mathbf{g}_i = (r_i, g_i, b_i)$ denote the i-th lattice point. Its query embedding is defined as

$$\mathbf{q}_i = W_g \mathbf{g}_i + \mathbf{p}^R_{r_i} + \mathbf{p}^G_{g_i} + \mathbf{p}^B_{b_i}, \tag{9}$$

where the factorized embeddings encode the position of each query along the three RGB axes. Collecting all queries as Q^{lut}, we condition them on F_t through cross-attention:

$$H^{\mathrm{lut}}_t = Q^{\mathrm{lut}} + \mathrm{MHA}\left(\mathrm{LN}(Q^{\mathrm{lut}}), \mathrm{LN}(F_t), \mathrm{LN}(F_t)\right). \tag{10}$$

This formulation allows different locations in the RGB lattice to extract transformation cues relevant to their corresponding color regions.

Temporal Gated LUT Fusion. To further improve temporal consistency, we adaptively fuse the current LUT latent with that of the previous frame. The temporal gate and fused representation are computed as

$$g_t = G\Big(H^{\mathrm{lut}}_t, \widetilde{H}^{\mathrm{lut}}_{t-1}\Big), \quad \widetilde{H}^{\mathrm{lut}}_t = (1 - g_t) \odot H^{\mathrm{lut}}_t + g_t \odot \widetilde{H}^{\mathrm{lut}}_{t-1}, \tag{11}$$

where G is a learnable projection. For the first frame, we set $\widetilde{H}^{\mathrm{lut}}_1 = H^{\mathrm{lut}}_1$. This adaptive fusion preserves useful historical transformations while allowing the LUT to respond to frame-wise appearance changes.

The LUT residual is decoded from the propagated latent representation:

$$\Delta\mathbf{L}_t = \mathrm{FFN}\left(\mathrm{LN}(\widetilde{H}^{\mathrm{lut}}_t)\right), \tag{12}$$

and added to the identity LUT:

$$\mathbf{L}_t = \mathrm{clip}\left(\mathbf{L}_{\mathrm{id}} + \Delta\mathbf{L}_t, 0, 1\right). \tag{13}$$

Finally, the predicted LUT is applied to the original frame using differentiable 3D LUT interpolation,

$$Y_t = \mathcal{G}(I_t, \mathbf{L}_t), \tag{14}$$

producing a frame-adaptive global color transformation while preserving the spatial structure of the input.

3.4 Training Strategies

Our training strategy consists of three stages designed to balance reference-conditioned color transfer with temporal stability.

Frame-wise Pretraining. We first train the image-based model using images from MIT-Adobe FiveK Bychkovsky et al. (2011) and PPR10K Liang et al. (2021), augmented online with a bank of 991 3D LUTs. By applying different LUTs to construct the content and reference appearances while keeping the reference and target transformations identical, the model learns to retrieve semantic color information and predict accurate frame-level transformations.

Temporal Training. Next, we train the temporal modules on synthetic sequences generated from two still images with smooth geometric and photometric perturbations. The content switches to the second image at a randomly sampled frame n, while the reference remains fixed. The frozen frame-wise predictor $\mathcal{F}_{\mathrm{img}}$ estimates a target LUT from the first frame of each segment:

$$\mathbf{L}^A = \mathcal{F}_{\mathrm{img}}(I_1, I_s), \qquad \mathbf{L}^B = \mathcal{F}_{\mathrm{img}}(I_n, I_s). \tag{15}$$

The target sequence is then generated via differentiable 3D LUT interpolation $\mathcal{G}$:

$$\hat{\mathbf{Y}}_t = \begin{cases} \mathcal{G}(I_t, \mathbf{L}^A), & t < n, \\ \mathcal{G}(I_t, \mathbf{L}^B), & t \geq n. \end{cases} \tag{16}$$

This supervision encourages stable mappings within each segment and adaptation to the new content when the source image switches.

Supervised Fine-tuning. Finally, we fine-tune the model on 1,000 curated content-reference-target triplets, which are converted into synthetic sequences using the aforementioned temporal augmentations. This stage adapts the model to high-quality aesthetic targets while preserving the temporal behavior learned in the previous stage.

3.5 Training Objectives

Given a target sequence $\{\hat{\mathbf{Y}}_t\}_{t=1}^N$, predictions $\{\mathbf{Y}_t\}_{t=1}^N$, and frame-wise LUTs $\{\mathbf{L}_t\}_{t=1}^N$, we optimize

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{lut}}. \tag{17}$$

Reconstruction Loss. We use a temporally weighted ℓ_1 loss for frame-wise reconstruction:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N} \sum_{t=1}^{N} w_t \left\| \hat{\mathbf{Y}}_t - \mathbf{Y}_t \right\|_1, \tag{18}$$

where w_t increases linearly over time and is normalized to unit mean. This slightly emphasizes later frames, which depend more on the accumulated temporal state.

LUT Consistency. To suppress unnecessary frame-to-frame variations in the predicted color transformation, we impose an ℓ_1 consistency loss between adjacent LUTs:

$$\mathcal{L}_{\text{lut}} = \frac{1}{N-1} \sum_{t=2}^{N} \|\mathbf{L}_t - \mathbf{L}_{t-1}\|_1. \tag{19}$$

This regularization encourages smooth temporal evolution of the predicted LUTs.

4 Experiments

4.1 Setup

Implementation Details. We implement our framework in PyTorch and conduct all experiments on a single NVIDIA L40S GPU. Trainable modules are optimized using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate 1×10^{-4}), while the DINOv2 ViT-S/14 semantic encoder remains frozen. To predict the frame-adaptive 33^3 LUT, input frames and reference images are downsampled to 224×224. During inference, the predicted LUT is applied directly to the original-resolution input via trilinear interpolation, thereby retaining the original spatial resolution.

Test Dataset. We construct a dedicated real-world test set consisting of 55 video-reference pairs through data collection and in-the-wild capture. The dataset covers diverse scenes, subjects, camera motions, illumination conditions, and color styles, and is used to evaluate reference-based video color transfer under practical scenarios.

Quantitative Metrics. Since traditional fidelity metrics such as PSNR and LPIPS Zhang et al. (2018) are not well suited for reference-based photorealistic color transfer, we follow the specialized evaluation protocol of Neural Preset Ke et al. (2023). We report *Style Similarity*, measured by a pre-trained style discriminator between each output frame and the reference image, and *Content Similarity*, computed as SSIM Wang et al. (2004) between LDC Soria et al. (2022) edge features of the output and the corresponding input frame. Both metrics are averaged over all video frames and normalized to $[0, 1]$, where higher values indicate better performance. To evaluate temporal color consistency, we further use a flow-aligned CIEDE2000 Sharma et al. (2005) metric:

$$E(O_i, O_j) = \frac{\sum_{\mathbf{x}} M_{i,j}(\mathbf{x})\, \Delta E\Big(O_i(\mathbf{x}), \hat{O}_{j \to i}(\mathbf{x})\Big)}{\sum_{\mathbf{x}} M_{i,j}(\mathbf{x})}, \tag{20}$$

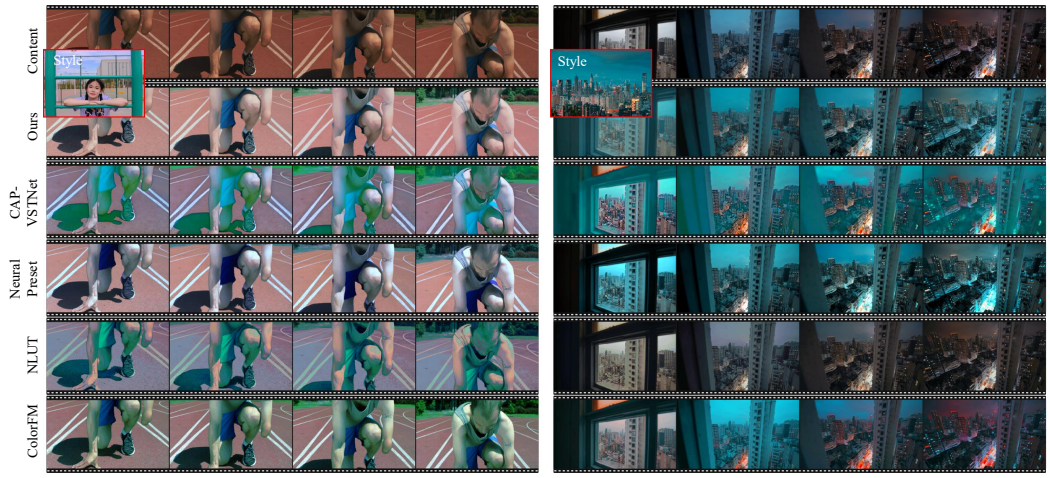

Figure 3: Qualitative comparison on reference-based video color transfer.

where $\hat{O}_{j\to i}$ denotes the output frame O_j warped to frame i using optical flow estimated from the corresponding original content frames, and $M_{i,j}$ denotes the valid non-occlusion mask obtained by forward–backward flow consistency. We use RAFT Teed & Deng (2020) for optical flow estimation and report ΔE at temporal intervals of 1, 5, and 35 frames to evaluate short-, medium-, and long-range temporal consistency, respectively. Lower values indicate better temporal color stability.

4.2 Comparisons

We compare our method with video-based approaches, including NLUT Chen et al. (2023), Video-ColorGrading Shin et al. (2025), and ColorFM He et al. (2026), as well as image-based methods, including WCT^2 Yoo et al. (2019), Neural Preset Ke et al. (2023), CAP-VSTNet Wen et al. (2023), and SA-LUT Gong et al. (2025). Since image-based methods lack explicit temporal modeling, we extend them to video by processing each frame independently with a fixed reference image. We use official pretrained models and default settings whenever available, except for Neural Preset, for which we adopt a publicly available unofficial implementation.

4.2.1 Qualitative Results

Figure 3 presents qualitative comparisons on two representative video scenes featuring stable lighting and substantial illumination changes, respectively. This allows us to evaluate both style fidelity and robustness to temporal variations.

In the stable scene, our method effectively transfers the reference style while preserving scene structure. In contrast, CAP-VSTNet introduces a noticeable greenish cast, and Neural Preset yields weak stylization, leaving regions near their original colors. While NLUT and ColorFM achieve stronger transformations, they tend to over-shift colors, reducing reference fidelity.

The differences are more pronounced in the second scene, which exhibits drastic indoor-to-outdoor illumination shifts. Since CAP-VSTNet and Neural Preset process frames independently, their outputs are sensitive to frame-wise illumination changes, leading to inconsistent color responses. NLUT and ColorFM improve temporal coherence by propagating transformations estimated from selected keyframes. However, such mappings cannot fully adapt to large intra-video illumination changes, resulting in under- or over-stylization when the current lighting differs from the keyframes. In particular, NLUT retains more of the original appearance in darker regions, while ColorFM shows noticeable color deviations under strong illumination changes.

Conversely, our method explicitly models temporal dependencies to predict frame-adaptive transformations. This enables robust responses to substantial lighting changes while maintaining temporal consistency. Overall, our approach achieves a superior balance of style fidelity, structural preservation, and temporal coherence.

Table 1: **Quantitative comparison of content similarity, style similarity, and runtime.** Dist. to Ideal denotes the Euclidean distance to the ideal point $(1, 1)$. Time denotes the inference time for a single 512×512 frame. Best and second-best results are **bolded** and underlined, respectively.

Method	Content Sim.↑	Style Sim.↑	Dist. to Ideal↓	Time (s)↓
WCT^2	0.414	0.832	0.609	0.780
Neural Preset	0.751	0.650	0.430	0.021
NLUT	**0.812**	0.376	0.652	6.256
CAP-VSTNet	0.646	**0.859**	0.381	0.048
SA-LUT	0.640	0.489	0.625	0.052
VideoColorGrading	0.520	0.573	0.643	5.648
ColorFM	0.744	<u>0.855</u>	<u>0.295</u>	<u>0.014</u>
Ours	<u>0.760</u>	0.851	**0.282**	**0.013**

Table 2: **Quantitative comparison of temporal color consistency.** Lower values indicate better temporal consistency. Methods marked with * perform frame-wise inference, while the remaining methods apply a globally inferred transformation to the entire video. Best results are shown in **bold**.

Type	Method	wΔE-1↓	wΔE-5↓	wΔE-35↓
Color-only mapping	WCT^{2*}	1.500	2.745	5.790
	NLUT	**1.224**	**2.291**	**4.940**
	Neural Preset*	1.287	2.318	5.244
	VideoColorGrading	1.485	2.668	5.900
Structure-semantic-aware	CAP-VSTNet*	1.437	2.695	5.809
	SA-LUT*	2.889	4.852	8.686
	ColorFM*	1.992	3.359	6.345
	ColorFM	1.326	2.380	5.464
	Ours*	**1.264**	**2.317**	**5.239**

4.2.2 Quantitative Results

Content and style similarity. Table 1 and Figure 4 report the quantitative results. Our method achieves the smallest distance to the ideal point $(1, 1)$, indicating the best balance between content preservation and style fidelity among the evaluated methods. Specifically, it improves style similarity by 30.9% over Neural Preset while also increasing content similarity, consistent with the more faithful stylization shown in Figure 3. Compared with CAP-VSTNet, our method achieves 17.6% higher content similarity, demonstrating stronger content preservation. It also maintains comparable style fidelity to ColorFM, the strongest baseline under this joint metric, while exhibiting greater adaptability to illumination changes in the qualitative comparisons.

Inference efficiency. Table 1 reports the inference time per frame at a resolution of 512×512. TrackLUT achieves the lowest runtime of 0.013 s, compared with 0.014 s for ColorFM and 0.021 s for Neural Preset, supporting real-time frame-adaptive video color transfer.

Temporal consistency. Table 2 reports weighted ΔE errors at 1-, 5-, and 35-frame intervals. While global methods like NLUT naturally yield the lowest errors by applying a static transformation, they fail to accommodate frame-level variations. In contrast, our frame-adaptive approach achieves the best consistency among semantic-aware methods across all intervals. Specifically, it reduces errors by 36.5%, 31.0%, and 17.4% compared to frame-wise ColorFM, and even outperforms ColorFM's global variant. This confirms that our history-aligned propagation effectively mitigates short-term fluctuations and long-term drift while maintaining responsiveness to temporal changes.

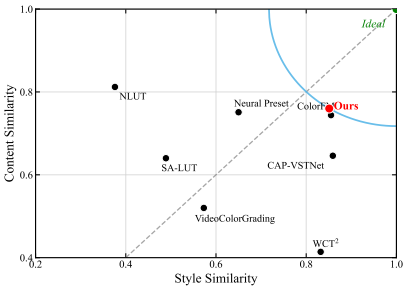

Figure 4: **Content and style similarity.** TrackLUT achieves the smallest Euclidean distance to the ideal point $(1, 1)$ among the compared methods. The blue curve indicates the equal-distance contour passing through our result.

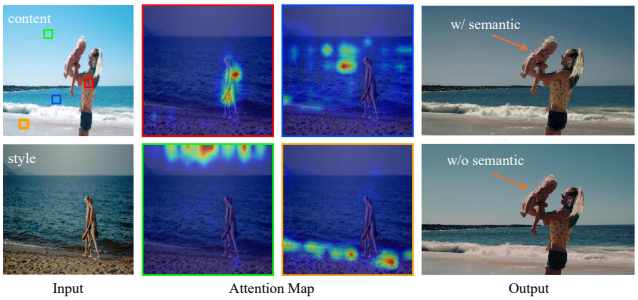

Figure 5: **Ablation of the Semantic Branch.** We compare results using a trainable ViT (w/o) and the frozen semantic encoder (w/). The attention maps visualize content-to-reference correspondence in the cross-attention module. Semantic guidance improves color consistency across semantically corresponding regions.

4.3 Ablation Study

Effectiveness of explicit semantic alignment. As shown in Fig. 5, the full model matches content queries to semantically corresponding reference regions despite substantial differences in spatial layout and appearance. To evaluate the role of the semantic encoder, we replace the frozen semantic encoder with a trainable ViT. This variant exhibits less accurate cross-image correspondence, causing colors to be retrieved from semantically unrelated regions and leading to less faithful color transfer, particularly on the foreground subject. These results suggest that the frozen semantic encoder provides more reliable correspondence cues for reference color retrieval.

Effectiveness of temporal consistency modeling. Table 3 demonstrates the efficacy of History-Aligned Fusion (HAF) and Temporal Gated LUT Fusion(LUT Gate). While each module independently improves temporal consistency, their joint application yields the lowest errors across all intervals. Mechanistically, HAF stabilizes color representation via history alignment, and LUT Gate regularizes the latent-space temporal transformation. Combined, they successfully drop the wΔE-1, wΔE-5, and wΔE-35 metrics from 1.436, 2.506, and 5.257 to 1.264, 2.317, and 5.239.

Table 3: Ablation study on History-Aligned Fusion (HAF) and Temporal Gated LUT Fusion (LUT Gate). Lower values indicate better performance.

HAF	LUT Gate	wΔE-1 ↓	wΔE-5 ↓	wΔE-35 ↓
×	×	1.436	2.506	5.257
✓	×	1.314	2.358	5.243
×	✓	1.310	2.385	5.254
✓	✓	**1.264**	**2.317**	**5.239**

5 Conclusion

This work revisits reference-based video color transfer from the perspective of dynamic color mapping. Our results show that a single video-level LUT is often insufficient for handling evolving content and illumination, while unconstrained frame-wise adaptation can compromise temporal stability. TrackLUT addresses this trade-off by introducing semantic correspondence and temporal color tracking into frame-adaptive LUT prediction. Extensive experiments demonstrate that this formulation improves the ability to follow appearance changes without introducing excessive temporal variation, while retaining the efficiency of LUT-based rendering. More broadly, our study suggests that effective video color transfer should treat color mapping as an evolving state rather than a fixed transformation, providing a useful direction for future adaptive video appearance modeling.

6 AI USE STATEMENT

We used ChatGPT during the preparation of this manuscript to assist with English editing, including grammar, clarity, and concision, and to identify potentially relevant literature. It also assisted in drafting and revising selected passages and figure captions based on technical descriptions and experimental results provided by the authors. The authors reviewed and revised all AI-assisted text, checked suggested references against the original publications, verified technical descriptions against the implementation, and checked numerical claims against the experimental records. The authors take full responsibility for the research and the final content of this paper.

REFERENCES

Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pp. 97–104. IEEE, 2011.

Yaosen Chen, Han Yang, Yuexin Yang, Yuegen Liu, Wei Wang, Xuming Wen, and Chaoping Xie. Nlut: Neural-based 3d lookup tables for video photorealistic style transfer. *arXiv preprint arXiv:2303.09170*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Zerui Gong, Zhonghua Wu, Qingyi Tao, Qinyue Li, and Chen Change Loy. Sa-lut: spatial adaptive 4d look-up table for photorealistic style transfer. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18294–18303. IEEE, 2025.

Zonghui Guo, Xinyu Han, Jie Zhang, Shiguang Shan, and Haiyong Zheng. Video harmonization with triplet spatio-temporal variation patterns. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19177–19186. IEEE, 2024.

Yuhang He, Kai Zhang, Xiaoming Li, Du Chen, and Jian Yang. ColorFM: An optimization-to-learning framework for color transfer via flow matching. *arXiv preprint arXiv:2607.07119*, 2026.

Man M Ho and Jinjia Zhou. Deep preset: Blending and retouching photos with color style transfer. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2112–2120. IEEE, 2021.

Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson WH Lau. Neural preset for color style transfer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14173–14182. IEEE, 2023.

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*, pp. 179–195. Springer, 2018.

Mujing Li, Guanjie Wang, Xingguang Zhang, Qifeng Liao, and Chenxi Xiao. D-lut: Photorealistic style transfer via diffusion process. In *2025 IEEE/CVF winter conference on applications of computer vision (WACV)*, pp. 9206–9214. IEEE, 2025.

Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7486–7495. IEEE, 2024.

Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *European Conference on Computer Vision*, pp. 468–483. Springer, 2018.

Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. PPR10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 653–661. IEEE, 2021.

Tianwei Lin, Honglin Lin, Fu Li, Dongliang He, Wenhao Wu, Meiling Wang, Xin Li, and Yong Liu. Adacm: adaptive colormlp for real-time universal photo-realistic style transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1613–1621, 2023.

Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6629–6638. IEEE, 2021.

Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6997–7005. IEEE, 2017.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Francois Pitie, Anil C Kokaram, and Rozenn Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pp. 1434–1439. IEEE, 2005.

Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE international conference on image processing (ICIP)*, pp. 4852–4856. IEEE, 2014.

Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

Gaurav Sharma, Wencheng Wu, and Edul N Dalal. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005.

Dongyu She and Kun Xu. An image-to-video model for real-time video enhancement. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1837–1846, 2022.

Seunghyun Shin, Dongmin Shin, Jisu Shin, Hae-Gon Jeon, and Joon-Young Lee. Video color grading via look-up table generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19141–19152. IEEE, 2025.

Xavier Soria, Gonzalo Pomboza-Junez, and Angel Domingo Sappa. Ldc: Lightweight dense cnn for edge detection. *IEEE access*, 10:68281–68290, 2022.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Meiao Wang, Xuejing Kang, Yaxi Lu, and Jie Xu. Retinexmcnet: A memory controller dominated network for low-light video enhancement based on retinex. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9716–9725. IEEE, 2025.

Yifan Wang, Zhixiang Hao, Yu Wang, and Congchao Zhu. Multimodal 3d lut generation via StatLUT with statistical features for photorealistic style transfer. *arXiv preprint arXiv:2607.08227*, 2026.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Linfeng Wen, Chengying Gao, and Changqing Zou. CAP-VSTNet: Content affinity preserved versatile style transfer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18300–18309. IEEE, 2023.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8261–8270. IEEE, 2024.

Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai. Ccpl: Contrastive coherence preserving loss for versatile style transfer. In *European conference on computer vision*, pp. 189–206. Springer, 2022.

Canqian Yang, Meiguang Jin, Xu Jia, Yi Xu, and Ying Chen. Adaint: Learning adaptive intervals for 3d lookup tables on real-time image enhancement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17501–17510. IEEE, 2022.

Wenjie Yang, Ning Xu, and Yifei Fan. Latent-guided exemplar-based image re-colorization. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4238–4247. IEEE, 2024.

Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9035–9044. IEEE, 2019.

Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2058–2073, 2020.

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

Junbao Zhou, Ziqi Pang, and Yu-Xiong Wang. Rmem: Restricted memory banks improve video object segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18602–18611. IEEE, 2024.