

UniColorFlow: A Unified Framework for Multimodal-Guided Video Colorization with Consistency Flow

Zheng Chang[†], Lihan Lin[†], Huan Ouyang, Shuchen Weng ,
Si Li , Boxin Shi

Received: date / Accepted: date

Abstract Video colorization aims to revitalize grayscale footage with vibrant and temporally consistent colors, yet it remains an inherently ill-posed problem. While user-guided methods offer controllability, existing multimodal frameworks struggle with significant limitations, including limited adherence to conditions due to luminance bias, potential conflicts among guidance signals, inconsistent identities across multi-shot videos, and prohibitive computational costs. To address these challenges, we introduce UniColorFlow, a unified framework for multimodal-guided video colorization. We propose a data augmentation strategy to decouple luminance from color, thereby mitigating inherent biases in the training data. We further design a unified condition network to integrate diverse spatial and language inputs. This module effectively resolves condition conflicts and enforces multi-shot consistency via a shot-aware attention mechanism. Finally, we employ a one-stage training strategy that unifies training and consistency learning, significantly reducing both training and inference overhead. Extensive experiments demonstrate that UniColorFlow outperforms state-of-the-art

methods, achieving superior generation quality with flexible controllability and high efficiency.

Keywords Video Colorization · Flow Matching Model · Multi-shot Video Editing

1 Introduction

Video colorization transforms grayscale videos into vibrant and colored versions, breathing new life into historical archives and enhancing creative expression in multimedia and artistic design (Lei and Chen, 2019; Li et al, 2024; Bozic et al, 2024). By maintaining temporal consistency and semantic accuracy, this technique creates a more immersive and visually appealing viewing experience.

However, video colorization is an inherently ill-posed problem, as a single grayscale value can map to multiple plausible colors, introducing ambiguity. This challenge has driven a shift from fully automatic approaches (Lei and Chen, 2019; Thasarithan et al, 2019; Zhao et al, 2023b) to user-guided interactive methods that offer greater control. Recent interactive methods typically fall into three main categories. Exemplar-based approaches (Zhang et al, 2019; Yang et al, 2024c,b) require users to provide a semantically similar reference image, from which the color distribution is propagated across all video frames. Scribble-based methods (Zhang et al, 2017; Sangkloy et al, 2017) enable users to define object colors through direct color hints, typically applied to a single frame and then propagated to reduce manual annotation effort. More recently, language-based colorization (Liu et al, 2025b; Li et al, 2024) emerges, allowing users to guide the process with natural language descriptions.

To achieve higher controllability, recent research has explored unified colorization frameworks that support multimodal guidance (Bozic et al, 2024; Fang et al, 2025). These frameworks build upon diffusion models, leveraging their

[†] Equal contribution.  Corresponding authors

Si Li · Shuchen Weng
E-mail: lisi@bupt.edu.cn & shuchenweng@pku.edu.cn

Zheng Chang · Lihan Lin · Huan Ouyang · Si Li
School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China.

Shuchen Weng
Beijing Academy of Artificial Intelligence, China.

Boxin Shi
State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, China.

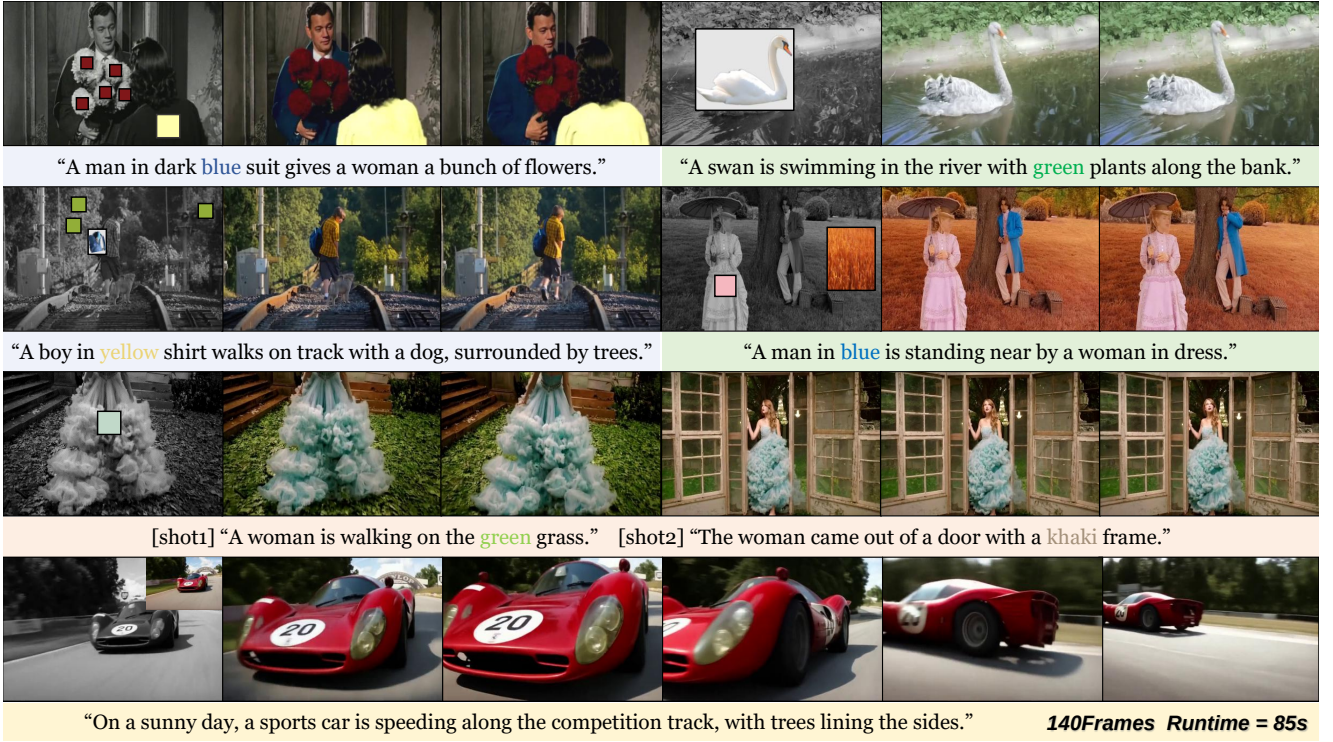


Fig. 1 Illustration of our UniColorFlow for multimodal-guided video colorization. **First row:** Our model faithfully adheres to the user-provided color conditions, accurately applying specified colors despite significant luminance differences from the source grayscale video. **Second row:** Our model demonstrates flexible control, effectively handling diverse multimodal conditions to produce high-fidelity colorization results. **Third row:** Our model is capable of handling multi-shot videos with significant variations in global color distributions while maintaining instance color consistency. **Fourth row:** Our model efficiently colorizes a 140-frame 480×832 grayscale video in 85 seconds, while maintaining its leading performance among relevant methods.

robust color priors to produce temporally consistent and visually compelling results. However, these methods often overlook the inherent correlation between luminance and color, struggling to generate faithful colors when user-provided color conditions have luminance levels that are underrepresented in the training set. Additionally, it remains challenging to integrate diverse multimodal inputs (*e.g.*, scribbles, reference images, and language descriptions) while effectively resolving potential conflicts among them. In the context of multi-shot videos, existing approaches often fall short in producing high-fidelity results, suffering from color bleeding and inconsistencies in object appearance across shots. Finally, training a unified framework compatible with multimodal guidance requires significant computational resources, and the inference process is prohibitively time-consuming.

In this paper, we propose UniColorFlow, a **Unified** framework for multimodal-guided video **Colorization** with consistency **Flow**. To ensure faithful adherence to multimodal conditions regardless of luminance, we propose a data augmentation strategy that decouples luminance from color. This approach effectively mitigates color assignment bias caused by training data imbalance while preserving fine-grained structural details (Fig. 1, first row). We then introduce a unified multimodal condition network to seamlessly integrate

diverse guidance signals. Unlike prior methods that rely on disjoint modules, we treat spatial conditions (*i.e.*, scribbles and reference images) as a unified color context, into which language descriptions are injected to resolve potential conflicts (Fig. 1, second row). Furthermore, to ensure multi-shot consistency, a shot-aware attention mechanism is employed to isolate temporal interactions and prevent color bleeding (Fig. 1, third row). Finally, to improve computational efficiency, we propose a unified one-stage training framework that jointly optimizes the condition network and consistency flow in a single stage. This process significantly reduces both training time and inference steps, enabling high-quality generation with minimal computational overhead (Fig. 1, fourth row). Our contributions are summarized as follows:

- We introduce a data augmentation strategy to decouple luminance from color, which ensures faithful adherence to color conditions regardless of video luminance.
- We design a unified network to integrate multimodal conditions and resolve their conflicts, where a shot-aware attention mechanism ensures multi-shot consistency.
- We propose an efficient one-stage training strategy, significantly reducing computational costs and accelerating the training and inference time.

2 Related work

Automatic video colorization. The primary objective of video colorization is to generate semantically coherent and aesthetically pleasing colors for grayscale footage, while critically ensuring temporal consistency. Early image colorization methods explored optimization-based propagation ((Levin et al, 2004)), example-based color transfer ((Irony et al, 2005)), intrinsic image priors ((Liu et al, 2008)), and semantic retrieval from Internet images (Chia et al, 2011). Pioneering deep-learning-based method (Cheng et al, 2015) employs deep neural networks trained on extensive visual datasets to infer color based on semantic information within the luminance channel. Following this, CIC (Zhang et al, 2016) tackles color under-saturation by adopting a per-pixel classification approach. To further minimize semantic color errors, CT² (Weng et al, 2022a) introduces the transformer architecture. Extending to the video domain, the research focus has shifted towards maintaining temporal consistency. Researchers have investigated various strategies to achieve this, including self-regularization techniques (Lei and Chen, 2019), generative adversarial learning (Zhao et al, 2023b), and optical flow estimation (Liu et al, 2024b). Nevertheless, without external guidance, automatic video colorization remains a fundamentally ill-posed problem.

User-guided video colorization. User guidance plays a pivotal role in achieving controllable and precise video colorization. Exemplar-based approaches leverage reference images to direct colorization (Wan et al, 2022; Zhang et al, 2019). To handle video dynamics, recent frameworks incorporate 3D CNNs for spatio-temporal feature extraction (Iizuka and Simo-Serra, 2019) or employ bidirectional fusion with semantic priors to mitigate occlusion and color bleeding (Yang et al, 2024c,b). However, the performance of these methods is heavily dependent on the reference image’s quality and semantic alignment, often lacking precise instance-level control. Alternatively, scribble-based methods utilize sparse user-drawn hints. To reduce annotation burden, networks are designed to propagate these sparse cues to global regions (Zhang et al, 2017; Sangkloy et al, 2017), ensuring temporal consistency via frame-to-frame propagation (Zhao et al, 2023a). More recently, language-based guidance has emerged as a highly flexible and intuitive modality. Despite progress in text-driven object mapping (Weng et al, 2022b; Chang et al, 2022; Liu et al, 2025b) and instance-level manipulation (Chang et al, 2023b,a, 2026), these methods often fail to capture nuanced color distinctions (*e.g.*, distinguishing between Klein blue and navy blue). UniColor (Huang et al, 2022) explores unified multimodal guidance for image colorization by converting different modalities (*e.g.*, exemplars and text) into a common hint-point representation. These preprocessing steps are not trained end-to-end with the colorization model and may be affected by errors from external matching

modules. These limitations underscore need for a unified framework capable of integrating multimodal guidance to address complex user requirements.

Diffusion-based video editing. Diffusion-based models have demonstrated superior performance in image and video generation (Rombach et al, 2022; Yang et al, 2025; WanTeam, 2025), motivating researchers to incorporate control architectures that adapt pre-trained generative priors for video editing. Early efforts extended image diffusion models by employing inter-frame global attention to model temporal dependencies in a training-free manner (Zhang et al, 2024; Wang et al, 2023a; Qi et al, 2023). Subsequently, alternative approaches integrated temporal modules (*e.g.*, temporal convolution (Carreira and Zisserman, 2017) and temporal attention (Bertasius et al, 2021)) into image diffusion models to enhance video generation capabilities (Wang et al, 2023b; Esser et al, 2023; Ho et al, 2022). The advent of the Diffusion Transformer (DiT) (Peebles and Xie, 2023) has further improved video editing performance by leveraging robust spatiotemporal modeling (Jiang et al, 2025; Weng et al, 2025; Bian et al, 2025). Specifically in video colorization, diffusion-based models have also driven significant advancements. VVFM (Bozic et al, 2024) introduces training-free attention fusion for image-to-video colorization, while Li *et al.* (Li et al, 2024) enhance long-range dependencies via sliding reference grids. Progress has also been made in exemplar-based line-art colorization through structure and appearance reference networks (Huang et al, 2024; Meng et al, 2025). More recently, ChromaFlow (Liu et al, 2025a) introduces flow matching for efficient end-to-end video colorization, while ColorFLUX (Li et al, 2026) leverages structure-color decoupling and generative priors to improve semantically faithful old-photo colorization. Despite these advances, generative colorization methods often compromise between high-fidelity colorization and computational efficiency, limiting their practical deployment.

3 Method

In this section, we present UniColorFlow, our unified framework for multimodal-guided video colorization. We begin by revisiting the preliminaries of video generation models and consistency distillation (Sec. 3.1). Next, we introduce the luminance-decoupled data augmentation strategy to improve color adherence (Sec. 3.2). We then detail the unified multimodal condition network, which effectively integrates diverse guidance signals (Sec. 3.3). Following this, we present the one-stage training strategy that ensures efficiency in both training and inference (Sec. 3.4). Finally, we provide implementation details (Sec. 3.5).

3.1 Preliminaries

Video generative model. We build our model upon Wan2.1 (WanTeam, 2025), leveraging its language-guided spatio-temporal generation prior. To enable scalable and efficient training, a 3D causal variational autoencoder (VAE) compresses RGB video $X^{\text{RGB}} \in \mathbb{R}^{F \times H \times W \times 3}$ into compact latent representations $z \in \mathbb{R}^{\frac{F}{4} \times \frac{H}{8} \times \frac{W}{8} \times D}$. A Diffusion Transformer (DiT) with N transformer blocks is employed as the backbone to process the latent codes. Each block integrates a self-attention mechanism, followed by a cross-attention mechanism that incorporates language description embeddings y^{tex} encoded by umT5 (Chung et al, 2023) to guide generation. The generation process is formulated as a rectified flow. The forward process linearly interpolates between the data latent z and random noise $\epsilon \sim \mathcal{N}(0, I)$, defined as $z_t = tz + (1 - t)\epsilon$, where $t \in [0, 1]$. The backward process predicts a velocity field $u_\theta(z_t, t, y^{\text{tex}})$ to transform noise into structured data, conditioned on the latent code z_t , language description y^{tex} , and timestamp t . With the ground truth velocity defined as $v_t = z - \epsilon$, the loss function is formulated as the mean squared error (MSE) between the predicted and ground truth velocities:

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{z, \epsilon, t} [\|u_\theta(z_t, t, y^{\text{tex}}) - v_t\|^2]. \quad (1)$$

Consistency distillation. Consistency Distillation (CD) (Song et al, 2023) is a technique that transforms iterative ODE-based generation models into highly efficient ones capable of few-step sample generation. The core principle is to enforce the self-consistency property. In the context of rectified flow (Yang et al, 2024a), self-consistency involves constraining both velocity and endpoint predictions, ensuring that different points (z_t at timestep t and z_s at timestep s) on the same probability flow ODE trajectory map to the same solution endpoint. The distillation loss is formulated as:

$$\mathcal{L}_{\text{CD}} = \mathbb{E} [\|\mathbf{f}_\theta(z_t, t) - \mathbf{f}_\theta(\hat{z}_s^\phi, s)\|^2 + \|u_\theta(z_t, t) - u_\theta(\hat{z}_s^\phi, s)\|^2], \quad (2)$$

where θ^- denotes the exponential moving average (EMA) of the parameter θ . The adjacent point $\hat{z}_s^\phi$ is estimated from z_t via a single step of an ODE solver, guided by the pre-trained teacher model’s velocity field u_ϕ :

$$\hat{z}_s^\phi = z_t + (s - t)u_\phi(z_t, t), \quad (3)$$

and $\mathbf{f}_\theta$ is the estimated data endpoint via straight flow:

$$\mathbf{f}_\theta(z_t, t) = z_t + (1 - t)u_\theta(z_t, t). \quad (4)$$

3.2 Luminance-decoupled data augmentation

Accurate color adherence is critical for multimodal-guided colorization. However, standard data preparation directly

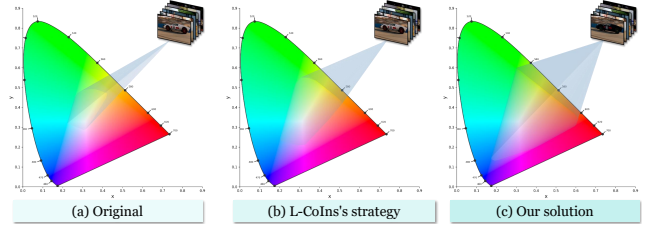


Fig. 2 Visualization of the feasible color range under extreme luminance levels ($L \in [220, 255]$) in CIE XYZ space. (a) Original color mapping. (b) L-CoIns’s strategy. (c) Our solution.

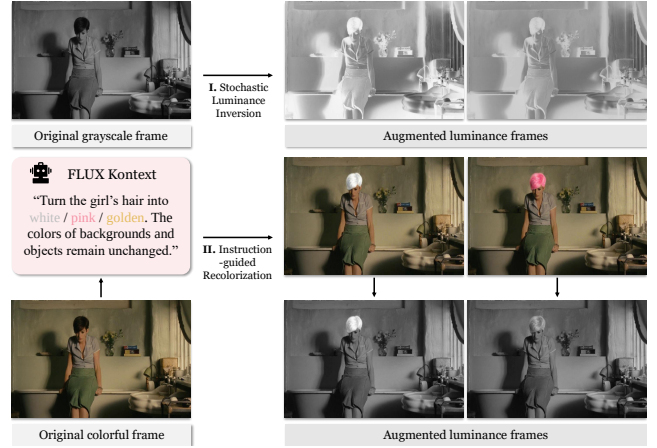


Fig. 3 Visualization of the two complementary operators in our luminance-decoupled data augmentation. **Top:** Stochastic luminance inversion probabilistically inverts luminance values. **Bottom:** Instruction-guided Recolorization alters specific object colors.

extracts the luminance channel from RGB video, introducing a strong correlation between luminance and color, evident in Fig. 2 (a). This learned bias fundamentally prevents the model from adhering to color conditions that violate natural statistics (e.g., applying a light color to a dark object).

While prior work (Chang et al, 2023b) attempts to mitigate this with random hue and gamma adjustments, this approach proves ineffective at extreme luminance levels. As shown in Fig. 2 (b), for luminance values within $[220, 255]$, this augmentation fails to effectively broaden the RGB gamut. To overcome this limitation, we introduce a luminance decoupled augmentation strategy to expand the feasible color range at extreme luminance levels, as illustrated in Fig. 2 (c). Specifically, our solution comprises two complementary operations, visualized in Fig. 3, which are applied stochastically to augment the training data.

Stochastic luminance inversion. This operation probabilistically inverts luminance values, mapping high-luminance ranges to low ones and vice versa, while preserving the target RGB unchanged. To prevent the model from learning a simple deterministic transformation, we implement the inversion

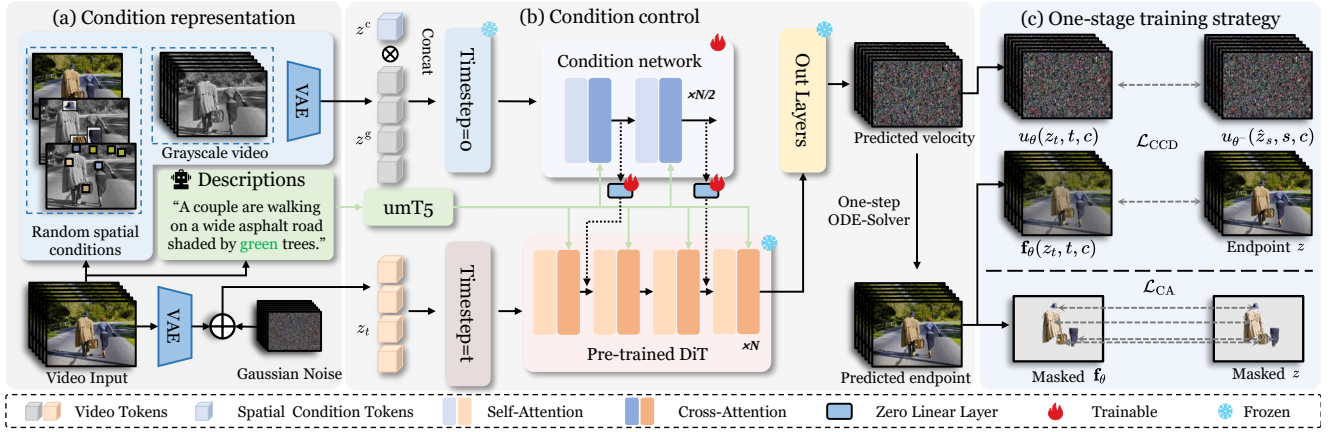


Fig. 4 An overview of the UniColorFlow pipeline. (a) The grayscale video and spatial condition are separately encoded by the causal VAE encoder to avoid unintended temporal dependencies, while the language descriptions are encoded by the text encoder (Chung et al, 2023). The resulting representations are then converted into token sequences. (b) Within the condition network, grayscale and spatial tokens are concatenated and injected into the pre-trained DiT blocks via zero-linear layers. In contrast, text embeddings are integrated via cross-attention layers to guide the colorization process. (c) The framework is trained using the one-stage training strategy. By simultaneously using consistency regularization as supervision and updating the parameters of the condition network, it significantly reduces training overhead.

as a randomized function:

$$f(x) = w \cdot (1 - x^a) + (1 - w) \cdot (1 - x^b), \quad (5)$$

where $x \in [0, 1]$ is the normalized luminance, and $w \in [0, 1]$, $a, b \in (0, 5]$ are randomly sampled parameters that control the weighting and curvature of the inversion.

Instruction-guided recolorization. We further leverage an instruction-guided editing model (Black-Forest-Labs et al, 2025) to recolorize objects with extreme luminance (e.g., white or black objects) in existing RGB images. To generate an unbiased dataset, the target color for each object is sampled randomly from a uniform distribution over the RGB space. To filter unwanted changes introduced by the editing model, we retain only samples whose background remains nearly unchanged outside the edited object, using $\text{LPIPS}_{\text{bg}} < 0.05$ and $\text{SSIM}_{\text{bg}} > 0.95$. We then rank the retained samples with MUSIQ (Ke et al, 2021) and select those with high image quality. We construct the training pairs using the original luminance as input and the edited RGB image as the target. This creates pairs that explicitly violate natural statistics (e.g., forcing the model to predict dark colors from high-luminance inputs). For compatibility with our video architecture, each static image is duplicated along the temporal dimension to form a pseudo-video clip.

3.3 Unified multimodal condition network

Our model generates multi-shot video colorization results guided by a grayscale input and multimodal user-provided conditions, including scribbles, reference images, and language descriptions. These representations are seamlessly integrated

within our unified multimodal condition network to form a comprehensive guidance signal.

Condition representation. As presented in Fig. 4 (a), we simulate diverse user inputs during training by randomly sampling spatial color conditions $X^c \in \mathbb{R}^{H \times W \times 3}$ from the following sources: (i) *Global color reference*. A randomly sampled video frame is cropped to serve as a holistic color palette. (ii) *Local instance reference*. Using the Grounding DINO detector (Liu et al, 2024a), we localize and crop instances via bounding boxes, providing high-fidelity object-level references. For samples generated by instance-guided recolorization, the local instance reference is extracted from the recolorized result. (iii) *Local color scribble*. We segment a frame into superpixels (Achanta et al, 2012), randomly select internal regions, and compute their dominant colors to mimic sparse user strokes. The grayscale frames X^g and the selected spatial color condition X^c are then encoded by the VAE encoder into latent representations $z^g \in \mathbb{R}^{\frac{F}{4} \times \frac{H}{8} \times \frac{W}{8} \times D}$ and $z^c \in \mathbb{R}^{1 \times \frac{H}{8} \times \frac{W}{8} \times D}$, respectively. The condition tokens z^c are treated as global anchors that provide consistent color information for the whole video. We arrange the input token sequence by first concatenating z^g and z^c along the temporal dimension, and then flattening the concatenated latent into a video token sequence as the input to the condition network.

For more flexible controllability, we use Qwen2.5-VL (Bai et al, 2025) to generate language descriptions focused on object colors and overall style. These descriptions are annotated for each shot, encoded into text embeddings y^{tex} using a pre-trained umT5 encoder (Chung et al, 2023), and injected into the corresponding shot of the DiT blocks following the methodology of Wan2.1 (WanTeam, 2025).

Condition control. As illustrated in Fig. 4 (b), our unified condition network fuses all color guidance signals. It first concatenates the latent representations of the grayscale input z^g and spatial conditions z^c , flattening them into a single token sequence. This sequence is processed by the condition network, which comprises $\frac{N}{2}$ transformer blocks mirroring the architecture of the main DiT backbone (Wan2.1 (Wan-Team, 2025)), where N is the total number of blocks in the backbone. The output of the i -th block in our condition network is projected via a linear layer to generate control signals, which are then injected into the $2i$ -th block of the main DiT. Unlike the main branch that processes noisy video tokens, the condition blocks operate on clean (noise-free) tokens.

Shot-aware attention. To process multi-shot videos, the condition network must handle significant shifts in global color statistics and scene compositions across hard cuts. If standard full attention is applied, the attention weights become suboptimally dispersed across visually divergent background features from adjacent shots. This introduces massive inter-shot noise that severely dilutes the computational focus on the actual control signals (*i.e.*, the global spatial anchors), ultimately degrading the semantic consistency of corresponding instances across shots. To prevent this *condition dilution* and ensure strict adherence to multimodal guidance, we propose a shot-aware attention mechanism that adaptively modulates interactions within the condition network through the use of distinct attention masks:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} + M \right) V. \quad (6)$$

In the self-attention layers, we operate on a concatenated token sequence, comprising grayscale video features z^g and spatial condition tokens z^c . To preserve shot-specific statistics, we maintain the interaction between grayscale video features z^g within the same shot while explicitly blocking cross-shot interactions. Meanwhile, the spatial condition tokens z^c are defined as global anchors to facilitate the propagation of consistent color information throughout the entire video. Guided by these principles, the shot-aware self attention mask M^{self} is designed as:

$$M_{ij}^{\text{self}} = \begin{cases} 0 & \text{if } i \in \mathcal{I}_c \text{ or } j \in \mathcal{I}_c, \\ 0 & \text{if } i, j \in \mathcal{I}_g \text{ and } \phi(i) = \phi(j), \\ -\infty & \text{otherwise,} \end{cases} \quad (7)$$

where $\mathcal{I}_g$ and $\mathcal{I}_c$ denote the index sets for video tokens and spatial condition tokens, respectively, and the function $\phi(\cdot)$ maps a video token index to its corresponding shot index.

In the cross-attention layers, to resolve potential conflicts between spatial color conditions and language descriptions, we employ a decoupled attention strategy. Gray video tokens z^g from k -th shot (acting as query Q) are constrained to attend exclusively to their corresponding text tokens y_k^{tex}

(acting as key K and value V). During calculation, the spatial tokens z^c are isolated from text embeddings. This design effectively decouples spatial guidance from text guidance, preventing conditional interference. The shot-aware cross attention mask M^{cross} is designed as:

$$M_{ij}^{\text{cross}} = \begin{cases} 0 & \text{if } i \in \mathcal{I}_g \text{ and } j \in \mathcal{J}_{\phi(i)}, \\ -\infty & \text{otherwise,} \end{cases} \quad (8)$$

where i and j are the indices for the query and key tokens, respectively. $\mathcal{J}_k$ denotes the k -th shot's index set of text tokens.

Algorithm 1 One-stage Train Strategy

- 1: **Input:** dataset $\mathcal{D}$, initial model parameter θ , learning rate η , decay factor λ , loss weight α
 - 2: $\theta^- \leftarrow \theta$ {Initialize the target model}
 - 3: **repeat**
 - 4: Sample clean data $z \sim \mathcal{D}$, noise $\epsilon \sim \mathcal{N}(0, I)$
 - 5: Sample time steps $t \sim \mathcal{U}[0, 1]$, $s \sim \mathcal{U}[0, 1]$, and $s < t$
 - 6: Construct noisy latent $z_t = tz + (1 - t)\epsilon$
 - 7: Estimate next state $\hat{z}_s \leftarrow z_t + (s - t)u_\theta(z_t, t, c)$, where $c = \{y^{\text{tex}}, z^g, z^c\}$
 - 8: Compute the loss $\mathcal{L} = \mathcal{L}_{\text{CCD}} + \alpha\mathcal{L}_{\text{CA}}$
 - 9: Update parameters $\theta \leftarrow \theta - \eta\nabla_\theta \mathcal{L}$
 - 10: Update target model $\theta^- \leftarrow \text{stopgrad}(\lambda\theta^- + (1 - \lambda)\theta)$
 - 11: **until** convergence
-

3.4 One-stage training strategy

In previous works (Lai et al, 2025; Zhuang et al, 2026), adapting a pre-trained diffusion model for controllable and efficient inference typically involves a two-stage process: training a condition network for controllability first, and then separately applying model distillation for inference acceleration. We argue that this sequential process is computationally inefficient.

As presented in Fig. 4 (c), given the grayscale input and color conditions, we propose a one-stage approach to train the condition network and reduce the denoising time step. In this strategy, the weights of the pre-trained DiT backbone remain frozen, and only the unified condition network is updated, which is detailed in Algorithm 1.

Conditional consistency loss.

Since pre-trained video generation model cannot generate videos following grayscale latents and color conditions, the EMA endpoint $\mathbf{f}_{\theta^-}$ in Eqn. (2) is not the desired condition-faithful colorization endpoint in early training. Therefore, we leverage the student model itself to estimate the trajectory. We first estimate the next state $\hat{z}_s$ using our conditional consistency flow θ :

$$\hat{z}_s = z_t + (s - t)u_\theta(z_t, t, c), \quad (9)$$

Table 1 Quantitative results on the combined benchmark (DAVIS30&Videvo20) and our constructed multi-shot dataset (MSColor). “Color.” denotes the Colorfulness score. “Ours-Ref”, “Ours-Scrb”, and “Ours-Mul” indicate our model conditioned on reference images, scribbles, and multimodal inputs, respectively. For better readability, the CDC values are scaled by 1000. $\uparrow$ ($\downarrow$) means higher (lower) is better. The best performance within each category is highlighted in **bold**.

Method	DAVIS30&Videvo20						MSColor						
	Color. $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$	CDC $\downarrow$	Color. $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$	CDC $\downarrow$	ISS $\uparrow$
VCGAN	13.65 $\pm$ 0.01	24.92 $\pm$ 0.01	0.938 $\pm$ 0.000	0.207 $\pm$ 0.000	3018.92 $\pm$ 3.39	4.525 $\pm$ 0.010	16.84 $\pm$ 0.02	23.33 $\pm$ 0.02	0.890 $\pm$ 0.000	0.208$\pm$0.000	3613.33 $\pm$ 3.76	10.562 $\pm$ 0.015	0.8008 $\pm$ 0.000
TCVC	18.75 $\pm$ 0.01	25.60 $\pm$ 0.01	0.945$\pm$0.000	0.195 $\pm$ 0.000	2823.52 $\pm$ 4.24	3.056$\pm$0.012	29.88$\pm$0.02	22.20 $\pm$ 0.02	0.888 $\pm$ 0.000	0.215 $\pm$ 0.000	2511.45$\pm$4.14	6.854 $\pm$ 0.016	0.8160 $\pm$ 0.0003
ColorSurge	25.71$\pm$0.02	26.72$\pm$0.02	0.936 $\pm$ 0.000	0.167$\pm$0.000	2621.23$\pm$7.06	3.167 $\pm$ 0.015	13.65 $\pm$ 0.03	24.92$\pm$0.03	0.893$\pm$0.000	0.211 $\pm$ 0.000	2680.23 $\pm$ 5.64	6.369$\pm$0.020	0.8208$\pm$0.000
DeepExemplar	25.00 $\pm$ 0.00	29.71 $\pm$ 0.00	0.955 $\pm$ 0.000	0.126 $\pm$ 0.000	2542.76 $\pm$ 0.00	3.039 $\pm$ 0.000	27.44 $\pm$ 0.00	24.28 $\pm$ 0.00	0.896 $\pm$ 0.000	0.163 $\pm$ 0.000	2302.44 $\pm$ 0.01	7.051 $\pm$ 0.000	0.8143 $\pm$ 0.000
DeepRemaster	16.45 $\pm$ 0.00	25.87 $\pm$ 0.00	0.937 $\pm$ 0.000	0.167 $\pm$ 0.000	2729.66 $\pm$ 0.00	4.658 $\pm$ 0.000	21.82 $\pm$ 0.00	25.35 $\pm$ 0.00	0.911 $\pm$ 0.000	0.147 $\pm$ 0.000	2228.93 $\pm$ 0.00	9.738 $\pm$ 0.000	0.8189 $\pm$ 0.0000
BiSTNet	25.83 $\pm$ 0.01	28.27 $\pm$ 0.01	0.950 $\pm$ 0.000	0.145 $\pm$ 0.000	2626.17 $\pm$ 4.24	3.185 $\pm$ 0.012	31.68 $\pm$ 0.02	27.71 $\pm$ 0.02	0.926 $\pm$ 0.000	0.134 $\pm$ 0.000	2160.03 $\pm$ 6.14	7.056 $\pm$ 0.016	0.8066 $\pm$ 0.000
ColorMNet	24.90 $\pm$ 0.00	32.05$\pm$0.00	0.966 $\pm$ 0.000	0.088$\pm$0.000	2452.38 $\pm$ 2.11	2.995 $\pm$ 0.002	31.33 $\pm$ 0.00	29.74$\pm$0.00	0.943 $\pm$ 0.001	0.089$\pm$0.000	1783.22$\pm$1.97	9.246 $\pm$ 0.000	0.8140 $\pm$ 0.0000
Ours-Ref	26.67$\pm$0.20	31.27 $\pm$ 0.12	0.968$\pm$0.002	0.109 $\pm$ 0.003	2426.71$\pm$96.07	2.830$\pm$0.14	32.36$\pm$0.26	29.37 $\pm$ 0.15	0.948$\pm$0.002	0.101 $\pm$ 0.003	1824.65 $\pm$ 72.46	6.338$\pm$0.15	0.8329$\pm$0.0042
SVCNet	26.89 $\pm$ 0.01	25.52 $\pm$ 0.01	0.931 $\pm$ 0.000	0.212 $\pm$ 0.000	2951.30 $\pm$ 4.24	3.295 $\pm$ 0.012	23.11 $\pm$ 0.02	23.38 $\pm$ 0.02	0.908 $\pm$ 0.000	0.205 $\pm$ 0.000	2445.70 $\pm$ 4.70	8.974 $\pm$ 0.018	0.8039 $\pm$ 0.000
Ours-Scrb	30.96$\pm$0.22	27.59$\pm$0.13	0.947$\pm$0.002	0.159$\pm$0.003	2594.67$\pm$113.02	2.848$\pm$0.16	27.58$\pm$0.24	24.95$\pm$0.13	0.920$\pm$0.002	0.162$\pm$0.003	2227.82$\pm$80.39	7.728$\pm$0.18	0.8140$\pm$0.0045
AnyV2V	20.52 $\pm$ 0.42	24.10 $\pm$ 0.26	0.880 $\pm$ 0.005	0.240 $\pm$ 0.008	3233.51 $\pm$ 320.39	3.215 $\pm$ 0.36	20.24 $\pm$ 0.46	21.35 $\pm$ 0.29	0.842 $\pm$ 0.005	0.256 $\pm$ 0.008	3292.77 $\pm$ 257.89	10.835 $\pm$ 0.40	0.7321 $\pm$ 0.0080
VACE	30.97 $\pm$ 0.38	27.35 $\pm$ 0.23	0.934 $\pm$ 0.004	0.177 $\pm$ 0.007	2789.28 $\pm$ 203.44	3.721 $\pm$ 0.33	26.20 $\pm$ 0.42	22.09 $\pm$ 0.26	0.860 $\pm$ 0.005	0.173 $\pm$ 0.007	2403.26 $\pm$ 144.73	7.576 $\pm$ 0.36	0.7977 $\pm$ 0.0072
Ours-Mul	31.33$\pm$0.26	31.20$\pm$0.15	0.968$\pm$0.002	0.107$\pm$0.004	2413.63$\pm$129.97	2.811$\pm$0.22	28.15$\pm$0.28	26.32$\pm$0.16	0.929$\pm$0.002	0.119$\pm$0.003	2163.98$\pm$95.54	7.140$\pm$0.27	0.8199$\pm$0.0060
Wan-Adapter	28.98 $\pm$ 0.23	30.25 $\pm$ 0.19	0.933 $\pm$ 0.003	0.149 $\pm$ 0.005	2577.50 $\pm$ 135.64	3.269 $\pm$ 0.21	27.90 $\pm$ 0.31	25.53 $\pm$ 0.17	0.910 $\pm$ 0.005	0.159 $\pm$ 0.002	2291.43$\pm$107.53	7.708 $\pm$ 0.21	0.8266$\pm$0.0071
W/o LDA	31.21$\pm$0.29	31.42$\pm$0.15	0.962 $\pm$ 0.002	0.110$\pm$0.004	2407.01$\pm$146.93	2.830$\pm$0.20	27.96$\pm$0.29	25.69$\pm$0.15	0.924$\pm$0.002	0.132$\pm$0.004	2384.76 $\pm$ 98.51	7.360 $\pm$ 0.20	0.8103 $\pm$ 0.0052
W/o SAM	29.77 $\pm$ 0.30	31.25 $\pm$ 0.16	0.967$\pm$0.003	0.115 $\pm$ 0.004	2456.38 $\pm$ 158.23	2.910 $\pm$ 0.21	26.89 $\pm$ 0.30	24.41 $\pm$ 0.16	0.913 $\pm$ 0.003	0.150 $\pm$ 0.004	2448.63 $\pm$ 104.81	7.549 $\pm$ 0.21	0.8089 $\pm$ 0.0050
W/o CAL	30.70 $\pm$ 0.27	30.43 $\pm$ 0.15	0.957 $\pm$ 0.002	0.118 $\pm$ 0.004	2515.79 $\pm$ 141.27	2.908 $\pm$ 0.19	26.55 $\pm$ 0.27	25.28 $\pm$ 0.15	0.919 $\pm$ 0.002	0.142 $\pm$ 0.004	2341.32 $\pm$ 95.19	7.256$\pm$0.19	0.8010 $\pm$ 0.0054
RF (2 Steps)	26.81 $\pm$ 0.31	28.91 $\pm$ 0.17	0.945 $\pm$ 0.003	0.124 $\pm$ 0.004	2586.62 $\pm$ 135.62	2.932 $\pm$ 0.20	25.92 $\pm$ 0.31	24.97 $\pm$ 0.17	0.917 $\pm$ 0.003	0.133 $\pm$ 0.004	2269.04 $\pm$ 90.73	7.565 $\pm$ 0.20	0.8071 $\pm$ 0.0051
RF (20 Steps)	31.42$\pm$0.26	31.44$\pm$0.14	0.969 $\pm$ 0.002	0.106$\pm$0.003	2401.30$\pm$124.32	2.789$\pm$0.17	28.80$\pm$0.26	26.85$\pm$0.14	0.937$\pm$0.002	0.104 $\pm$ 0.003	2078.18$\pm$82.05	7.083$\pm$0.17	0.8234$\pm$0.0046
RF+CD	30.81 $\pm$ 0.27	31.23 $\pm$ 0.15	0.970$\pm$0.002	0.108 $\pm$ 0.003	2476.30 $\pm$ 127.15	2.851 $\pm$ 0.18	28.07 $\pm$ 0.27	26.08 $\pm$ 0.15	0.926 $\pm$ 0.002	0.097$\pm$0.003	2091.54 $\pm$ 84.25	7.119 $\pm$ 0.18	0.8217 $\pm$ 0.0047

Table 2 Inference time comparison for colorizing a 140-frame video.

TCVC	VCGAN	ColorSurge	DeepExemplar	DeepRemaster	BiSTNet	ColorMNet	SVCNet	AnyV2V	VACE	Ours
50.70s	33.79s	14.93s	146.78s	40.56s	204.11s	39.68s	337.60s	239.58s	301.72s	85.23s

where $u_\theta(z_t, t, c)$ denotes the entire colorization network, and $c = \{y^{\text{tex}}, z^g, z^c\}$ represents the aggregated color conditions. To further stabilize the joint training of control and consistency, we modify the endpoint prediction target. Instead of distilling from a potentially unstable teacher $\mathbf{f}_\theta$, we directly utilize the ground-truth endpoint z as the supervision target. Accordingly, our conditional consistency training objective is defined as:

$$\mathcal{L}_{\text{CCD}} = \mathbb{E} [||\mathbf{f}_\theta(z_t, t, c) - z||^2 + ||u_\theta(z_t, t, c) - u_{\theta^-}(\hat{z}_s, s, c)||^2]. \quad (10)$$

Color adherence loss. As spatial color conditions are defined as global anchors, we additionally introduce a color adherence loss to ensure the colorization results faithfully follow them. We first generate a binary mask $m \in \mathbb{R}^{H \times W}$ to identify regions annotated with these conditions. Then, this mask is propagated across all video frames using feature matching (Lindenberger et al, 2023) to create a temporal mask sequence. Subsequently, we downsample this sequence via nearest-neighbor interpolation to match the latent space resolution, producing $m' \in \mathbb{R}^{\frac{F}{4} \times \frac{H}{8} \times \frac{W}{8}}$. The loss is designed to minimize the statistical discrepancy within the masked regions m' between the predicted clean latent code $\mathbf{f}_\theta(z_t, t, c)$ and the ground-truth latent code z . We implement it by aligning their channel-wise mean (μ) and standard deviation (σ)

with a Mean Squared Error (MSE) objective:

$$\mathcal{L}_{\text{CA}} = ||\mu(\mathbf{f}_\theta(z_t, t, c), m') - \mu(z, m')||^2 + ||\sigma(\mathbf{f}_\theta(z_t, t, c), m') - \sigma(z, m')||^2. \quad (11)$$

Finally, our total training objective is a combination of the conditional consistency loss and the color adherence loss:

$$\mathcal{L} = \mathcal{L}_{\text{CCD}} + \alpha \mathcal{L}_{\text{CA}}, \quad (12)$$

where the balancing weight $\alpha = 0.1$.

3.5 Training details

We construct our training dataset using 25K single-shot videos selected from OpenVid (Nan et al, 2025) and 10K multi-shot videos from the Condensed Movies Dataset (Bain et al, 2020). To ensure spatial robustness, we supplement this video data with 1K static images sampled from MS-COCO (Caesar et al, 2018) for joint image-video training. We train UniColorFlow with a total batch size of 8, which takes approximately 72 hours to converge. The video clip length is set to $F = 81$ frames with a resolution of 480×832 . We employ the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate $\eta = 1 \times 10^{-5}$ and momentum parameters $\beta_1 = 0.99, \beta_2 = 0.999$. The EMA decay rate λ follows the schedule proposed in (Song et al, 2023). All experiments are conducted on a cluster of 8 NVIDIA A100 GPUs using a fixed random seed of 42 for reproducibility.

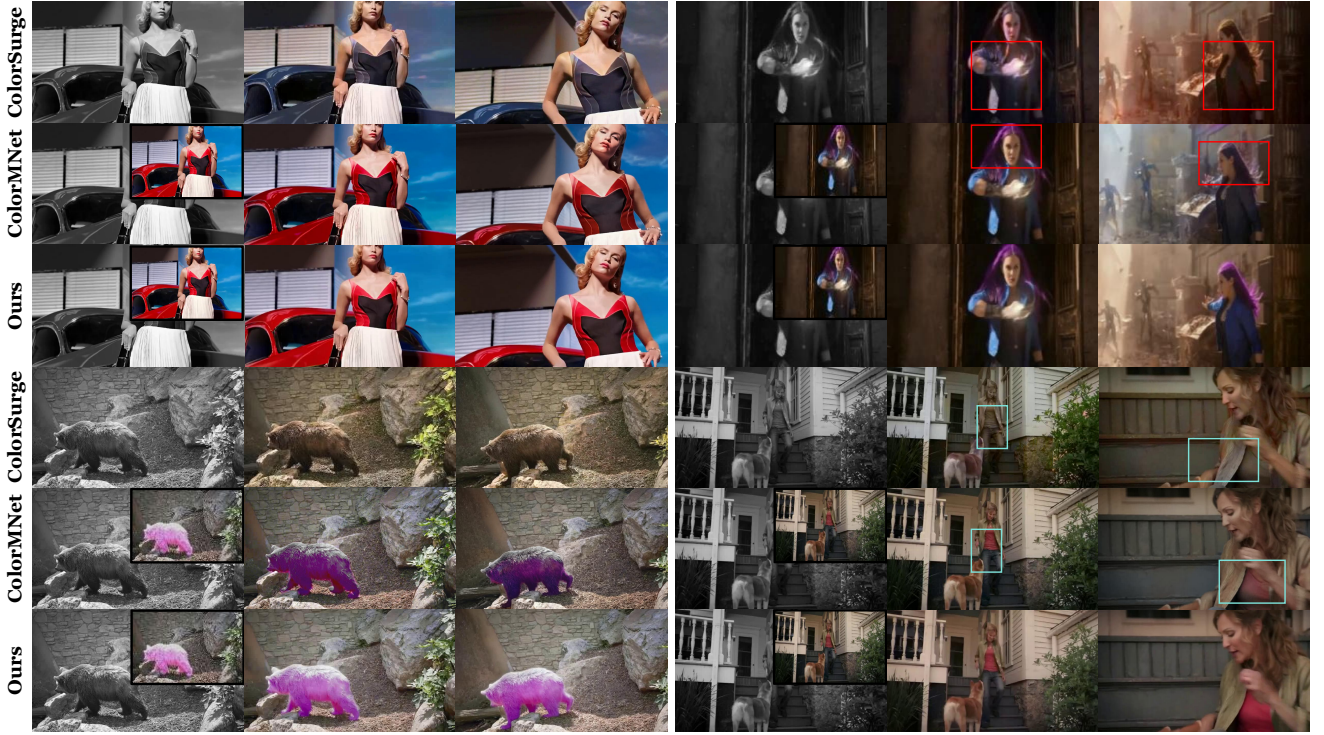


Fig. 5 Qualitative comparison with automatic (e.g., ColorSurge (Zhao et al, 2025)) and exemplar-based (e.g., ColorMNet (Yang et al, 2024b)) video colorization methods. Our UniColorFlow follows user instructions, effectively disentangling luminance from color while achieving multi-shot consistent colorization results.

4 Experiment

We perform evaluations on two widely used benchmarks: the DAVIS (Perazzi et al, 2016) and the Videvo (Lai et al, 2018) datasets. To evaluate performance on multi-shot videos, we additionally construct **MSColor**, a **M**ulti-**S**hot **C**olorization evaluation dataset that comprises 20 multi-shot videos across diverse scenes.

Evaluation metrics. At the frame level, we employ the Colorfulness score (Hasler and Suesstrunk, 2003) to assess the vividness of the colorization results. Additionally, we report PSNR (Huynh-Thu and Ghanbari, 2008), SSIM (Wang et al, 2004), and LPIPS (Zhang et al, 2018) to evaluate the fidelity and perceptual difference between the colorized frames and the ground truth. At the video level, we utilize the Fréchet Video Distance (FVD) (Unterthiner et al, 2019) to quantitatively assess the spatiotemporal realism of the generated videos by calculating the feature distribution distance between the results and the ground truth. We further measure temporal consistency using the Color Distribution Consistency (CDC) index (Liu et al, 2024b), scaled by 1000 for better readability. All metrics are computed at a resolution of 480×832 . Since the aforementioned metrics struggle to quantify instance color consistency across shot transitions, we propose Instance Similarity Score (ISS). Specifically, we utilize SAM2 (Ravi et al, 2025) to segment corresponding in-

stance regions in frames immediately adjacent to shot boundaries. We then calculate the cosine similarity of their visual feature embeddings extracted by CLIP (Radford et al, 2021) to measure semantic consistency.

4.1 Comparison with state-of-the-art methods

We compare UniColorFlow with representative open-source baselines to verify its effectiveness in producing accurate and conditionally consistent colorization. Our comparison spans four main categories: (i) *Automatic video colorization*, including TCVC (Liu et al, 2024b), VCGAN (Zhao et al, 2023b), and ColorSurge (Zhao et al, 2025). (ii) *Exemplar-based video colorization*, including DeepExemplar (Zhang et al, 2019), DeepRemaster (Iizuka and Simo-Serra, 2019), BiSTNet (Yang et al, 2024c), and ColorMNet (Yang et al, 2024b). (iii) *Scribble-based video colorization*, represented by SVCNet (Zhao et al, 2023a). (iv) *Multimodal video editing*, covering AnyV2V (Ku et al, 2024) and VACE (Jiang et al, 2025). To ensure a fair comparison, we align the input conditions for interactive methods: we use the same reference images (Ours-Ref), scribbles (Ours-Scrib), or multimodal conditions (Ours-Mul) as input for each corresponding interaction. All our results are generated with 2 sampling steps.

Quantitative comparisons. As shown in Table 1 (left), our method achieves the highest scores across most evaluation

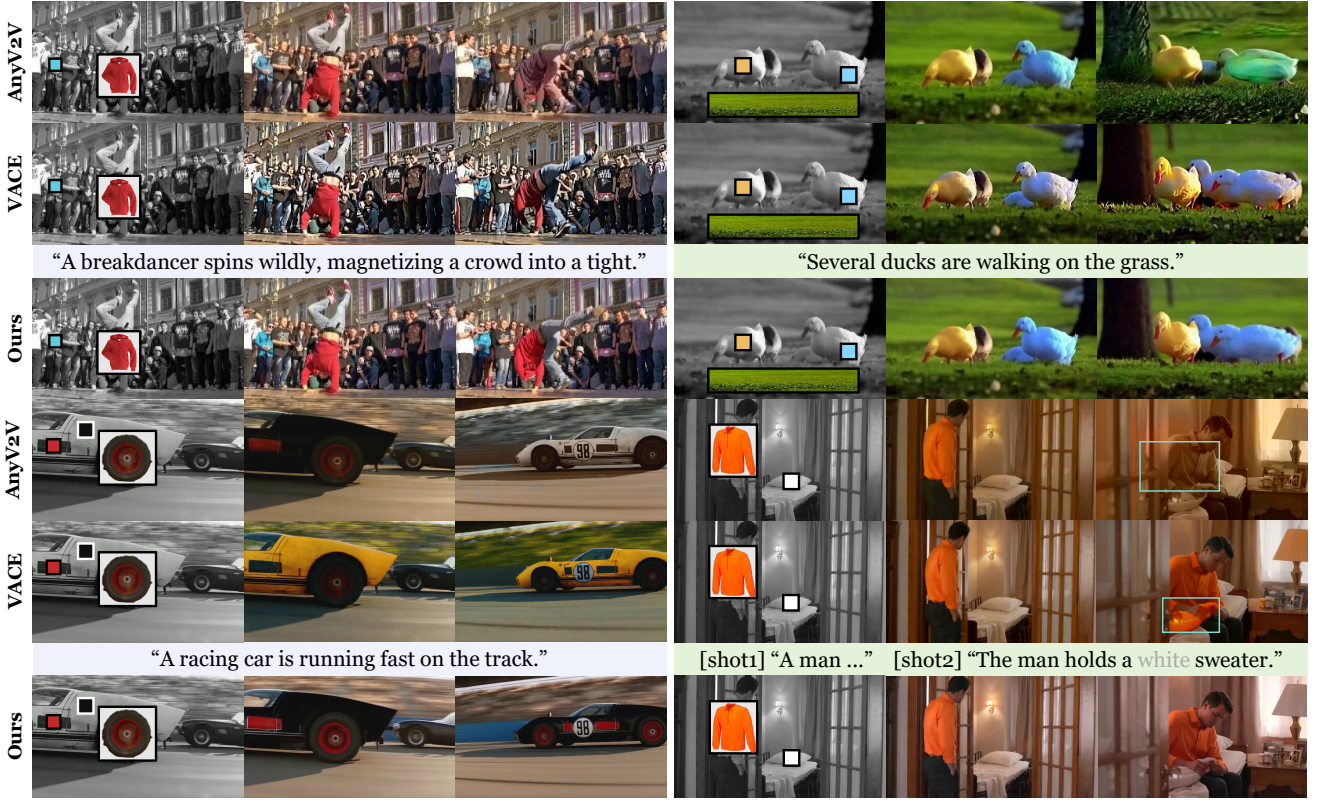


Fig. 6 Qualitative comparison with multimodal video editing methods (e.g., AnyV2V (Ku et al, 2024) and VACE (Jiang et al, 2025)). Our UniColorFlow overcomes the challenges of severe color deviation, limited color adherence, and instance color inconsistency across shots under multimodal conditions.

metrics on the DAVIS (Perazzi et al, 2016) and Videvo (Lai et al, 2018) datasets. The superior FVD and CDC scores demonstrate that our results exhibit both higher visual realism and improved temporal consistency compared to other methods. Furthermore, the higher SSIM and PSNR scores indicate that our method maintains better color fidelity, benefiting from the proposed color adherence loss. We further present quantitative comparisons for multi-shot colorization in Table 1 (right), where our method achieves the highest ISS, indicating superior performance in preserving the color consistency of instances across different shots.

Qualitative comparisons. We first compare UniColorFlow with the automatic method ColorSurge (Zhao et al, 2025) and the exemplar-based method ColorMNet (Yang et al, 2024b) in Fig. 5. Limited by the ill-posed nature of automatic colorization, ColorSurge often fails to satisfy specific user intent (e.g., the dark car and camisole in row 1, left; and the brown bear in row 4, left). While ColorMNet shows competitive quantitative results, it struggles to synthesize colors that defy luminance priors (e.g., the dark hair in row 2, right) and fails to maintain multi-shot instance color consistency (e.g., the red shirt in row 5, right). In Fig. 6, we compare our approach with multimodal video editing methods. Both AnyV2V (Ku et al, 2024) and VACE (Jiang et al, 2025) exhibit noticeable

artifacts, including severe color deviation (e.g., the degraded clothes in rows 1-2, left), limited color adherence (e.g., the blue duck in rows 1-2, right; the red car in rows 4-5, left), and instance color inconsistency across shots (e.g., the orange man and white bed in rows 4-5, right). Finally, as illustrated in Fig. 7, UniColorFlow demonstrates more precise control than SVCNet (Zhao et al, 2023a) when conditioned on scribbles, especially in challenging scenarios involving multiple instances (e.g., the blue car and green people in row 1, left) or large motion (e.g., the blue person in row 3, left). Collectively, these comparisons validate the superiority of our method in interpreting diverse multimodal inputs, ensuring faithful color adherence, and maintaining consistency across different shots.

Inference speed. We further evaluate inference time for colorizing a 140-frame video on a single NVIDIA A100 GPU. As shown in Table 2, UniColorFlow holds a significant speed advantage over other diffusion-based methods (i.e., AnyV2V (Ku et al, 2024) and VACE (Jiang et al, 2025)), requiring only 2 sampling steps to produce high-quality results. While some non-diffusion models (e.g., ColorSurge (Zhao et al, 2025) and ColorMNet (Yang et al, 2024b)) offer faster inference times, they lack the strong generative priors, resulting in inferior robust controllability across diverse color

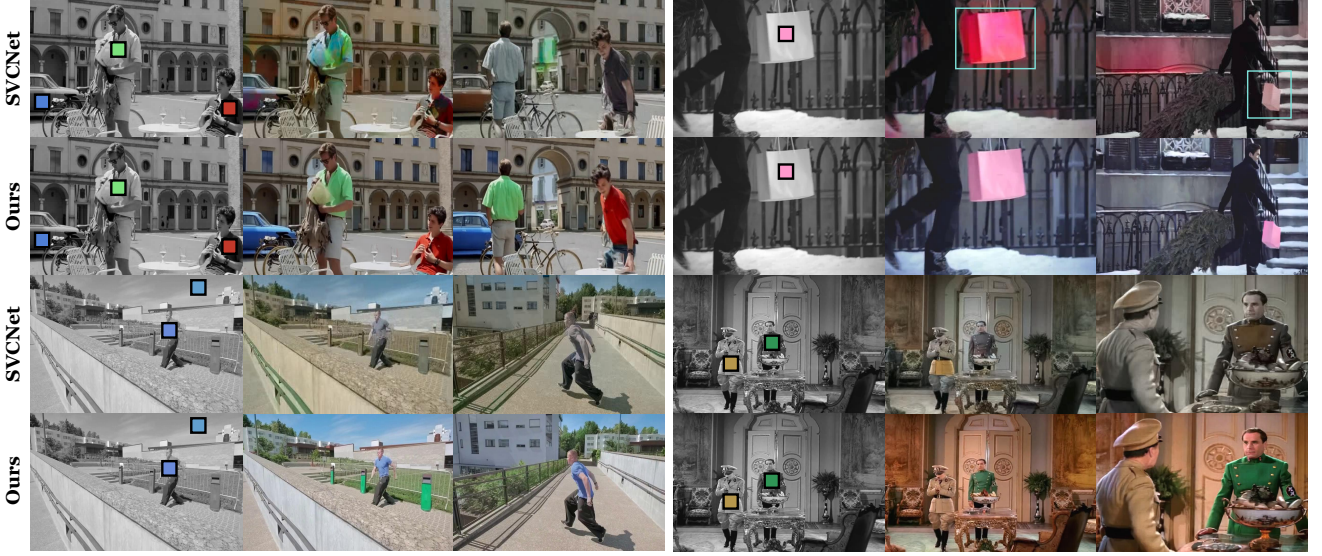


Fig. 7 Qualitative comparison with the scribble-based video colorization method (e.g., SVCNet (Zhao et al, 2023a)). Our UniColorFlow more accurately satisfies the scribble constraints in challenging scenarios involving multiple instances and large motion, ensuring faithful color adherence.

conditions. Consequently, our method strikes a compelling balance between controllability and speed.

4.2 Ablation

4.2.1 Effectiveness of modules

To validate the effectiveness of our proposed modules, we conduct an ablation study by creating three variants. The quantitative scores and qualitative results are presented in Table 1 and Fig. 8, respectively.

Wan-Adapter. To disentangle the contribution of the pre-trained video prior from the proposed designs, we construct a Wan2.1-based adapter baseline. Specifically, the grayscale latent and color-condition latent are projected by the linear conditioning module and added to the noisy latent z_t before the first transformer block. As shown in Fig. 8 (row 1), it exhibits degraded color adherence and cross-shot instance consistency.

W/o Luminance Decoupling Augmentation (LDA). We disable the luminance decoupling augmentation, relying solely on standard RGB-luminance video pairs during training. This causes the model to become overly reliant on learned statistical priors, failing to faithfully apply user-provided colors that defy these priors. As illustrated in Fig. 8 (row 2, left), even when a white color hint is provided for the woman’s dress, this variant renders a grayish tone instead.

W/o Shot-aware Attention Mechanism (SAM). We replace the shot-aware attention mechanism with vanilla self-attention and cross-attention. This removes the explicit boundary handling, causing the model to struggle with maintaining semantic consistency across shot transitions. As shown in Fig. 8

(row 3, right), when shot transitions lead to drastic scene changes, this variant fails to preserve the consistency of the character’s clothing.

W/o Color Adherence Loss (CAL). We remove the color adherence loss during training to verify its necessity. This discards the explicit constraints on channel-wise statistics (e.g., mean and variance), thereby compromising color fidelity. As observed in Fig. 8 (row 4, right), even with local color conditions, the model exhibits noticeable color deviation in the man’s pink suit.

4.2.2 Model Efficiency

Our proposed one-stage approach simultaneously trains the condition network and reduces the denoising time step, significantly reducing training overhead. To validate its effectiveness, we compare it against the conventional two-stage paradigm. The condition network contains 732M parameters. Under BF16 inference at 480×832 resolution with batch size 1, the complete model requires 10.16 GiB and 14.84 GiB of peak GPU memory for 81-frame and 161-frame videos, respectively. Quantitative and qualitative results are presented in Table 1 and Fig. 8, respectively.

RF. We first train the condition network using standard rectified flow loss (Eqn. (1)). This baseline (denoted as RF) requires numerous sampling steps to produce high-quality results, leading to inferior performance at 2-step inference.

RF+CD. We then append a separate consistency distillation phase to the RF baseline. Although this eventually achieves results comparable to our approach, it necessitates two full training cycles with significant computational overhead.

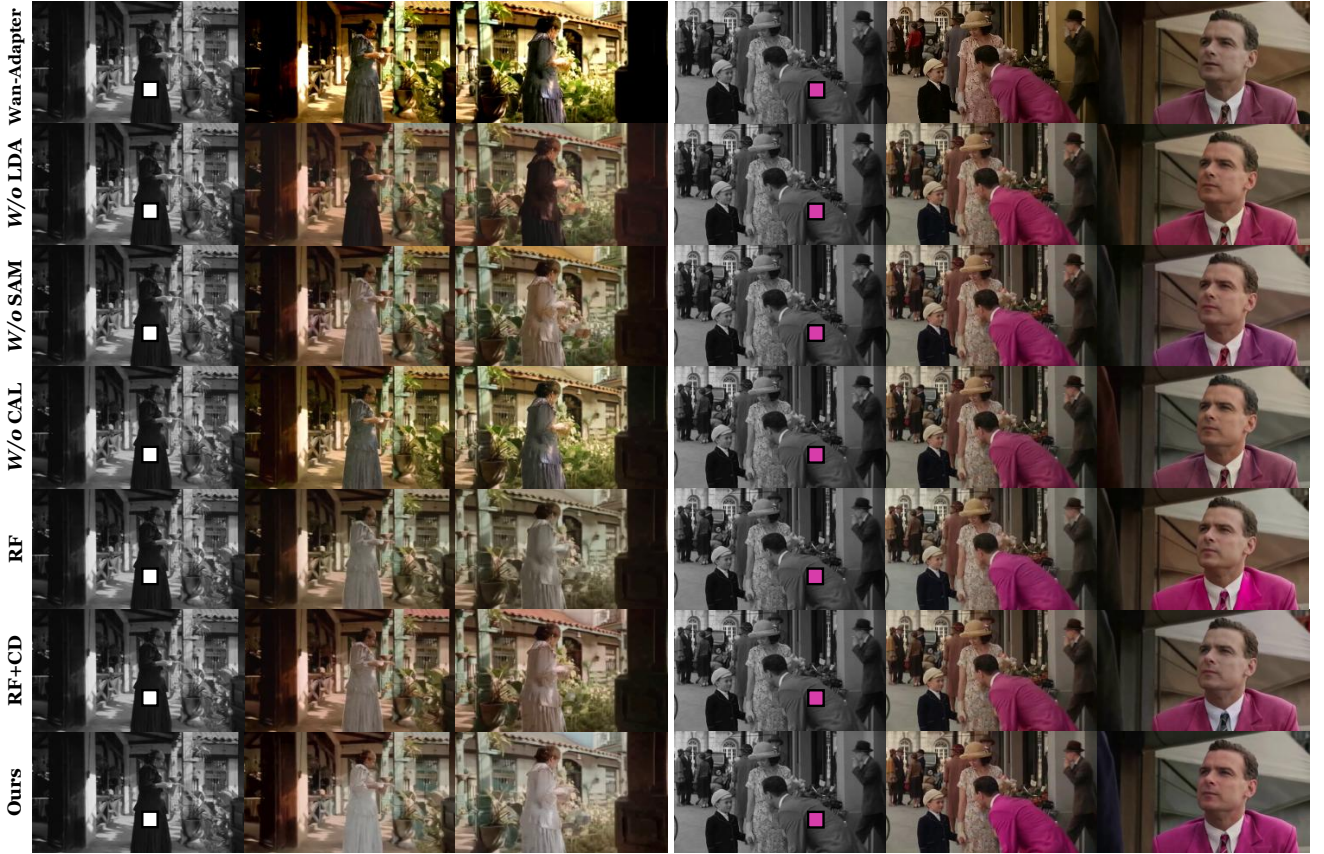


Fig. 8 Qualitative ablation study results. Compared to the Wan-Adapter baseline and variants without the proposed modules (*i.e.*, W/o LDA, W/o SAM, and W/o CAL) and the conventional two-stage paradigm (*i.e.*, RF and RF+CD), our UniColorFlow achieves superior temporal consistency and high-fidelity colorization, faithfully adhering to the user-provided color conditions.

Efficiency. Quantitatively, the conventional paradigm is computationally expensive, requiring approximately 72 hours for the first stage (RF) followed by an additional 48 hours for the consistency distillation stage (CD), totaling 120 hours. In contrast, our one-stage training strategy requires only 72 hours. While balancing the trade-off between synthesis quality and computational speed, our approach achieves a 40% reduction in training overhead, while directly securing high-fidelity outcomes with only two inference steps.

4.3 Analysis of Condition Conflicts

Our model integrates diverse spatial (*e.g.*, scribbles and reference images) and language descriptions within a unified condition network. The color adherence loss ($\mathcal{L}_{CA}$) explicitly penalizes regional statistical discrepancies to resolve conflicting guidance, inherently prioritizing local color conditions over global color conditions and language prompts. To demonstrate its robustness, we deliberately introduce scenarios with conflicting conditions.

Spatial-spatial conflicts. As shown in Fig. 9 (row 1), we use a reference image of green grass as the global condi-

tion, while simultaneously applying brown scribbles to locally specify the bear’s color. Despite the semantic conflict between these spatial conditions, our approach effectively generates harmonious results, correctly rendering a brown bear against a green grass background.

Language-spatial conflicts. As shown in Fig. 9 (row 2), we provide the language description “a white puppy” while placing brown scribbles over the corresponding instance. This introduces a direct conflict between language and spatial guidance. The result demonstrates that our model prioritizes the spatial condition, which we attribute to the color adherence loss that penalizes statistical discrepancies within user-specified regions.

4.4 Failure case.

In cross-shot colorization, UniColorFlow may assign the spatial color condition to an incorrect instance when the identity cues are weak or ambiguous across shots. As shown in Fig. 10, the user specifies the clothing color of a person viewed from the back in the previous shot, but the model fails to reliably associate this condition with the same person

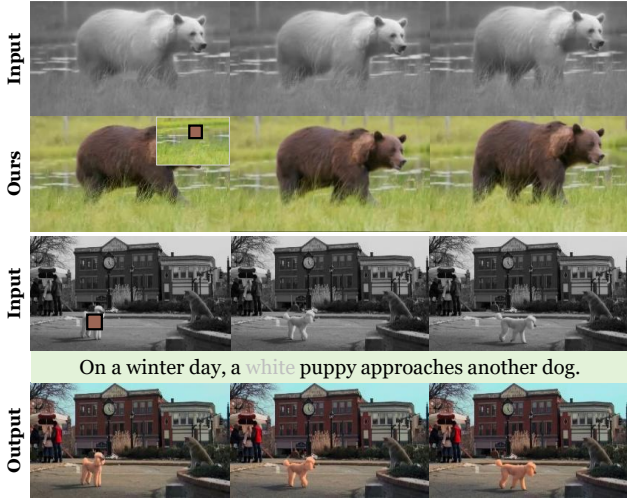


Fig. 9 Visualization of results under conflicting user-provided conditions. **First row:** For spatial-spatial conflicts, our model adaptively generates harmonious results. **Second row:** For language-spatial conflicts, our model prioritizes the spatial condition.



Fig. 10 Failure cases of UniColorFlow. Ambiguous identity information may lead to the failure of color condition propagation across shots.

viewed from the front in the next shot. This indicates that robust cross-shot color propagation still depends on sufficiently discriminative identity and appearance cues.

4.5 User Study

To assess the performance of our model from a human-centric perspective, we conduct three user study experiments. (i) *Color adherence*: Participants are instructed to select the video that best matches a given color condition when presented with results from our method and relevant baselines. (ii) *Overall visual quality*: Participants are asked to select the video they perceive as the most visually pleasing and realistic from a set containing results from our method and all competitors. (iii) *Cross-shot consistency*: Participants are asked to rate the consistency of instance colors across shot boundaries, focusing on whether the same object or subject preserves a stable color after a scene cut. For evaluation, we randomly select 50 sample videos (25 single-shot and 25 multi-shot) and invite 20 volunteers to perform independent assessments. As shown in Fig. 11, our method achieves the

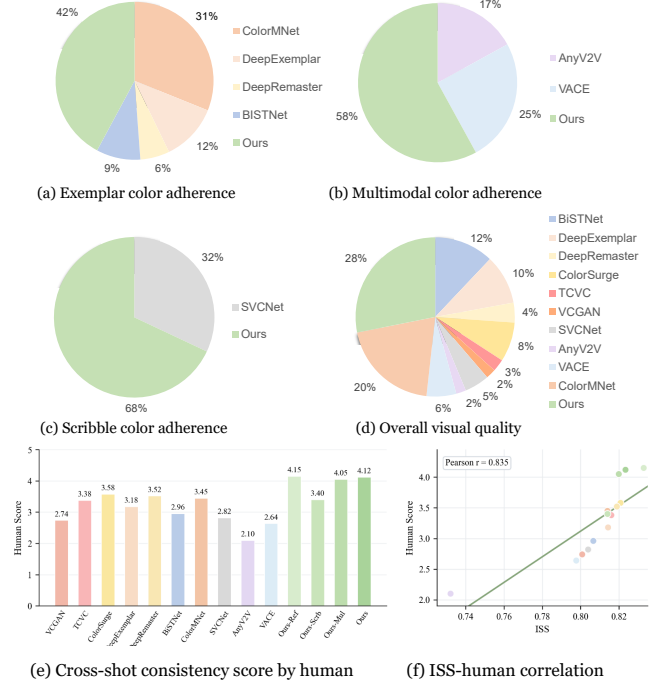


Fig. 11 User study results. UniColorFlow achieves the highest user preference scores for color adherence across (a) exemplar, (b) multi-modal, and (c) scribble conditions, while also ranking highest in (d) overall visual quality and (e) cross-shot consistency score. (f) shows a strong positive correlation between ISS and human judgments.



Fig. 12 Results of UniColorFlow in the resolution of 720×1280 .

highest preference scores for color adherence and overall visual quality. Furthermore, the cross-shot consistency scores are positively correlated with ISS, as shown in Fig. 11 (f), suggesting that ISS is aligned with human judgments and can serve as an indicator of cross-shot instance color consistency.

5 Applications

We present three typical application scenarios of UniColorFlow to demonstrate its versatility and generalization, each of which is detailed as follows:



Fig. 13 Top: Applications in diverse colorization. By leveraging diverse multimodal guidance for same grayscale video, UniColorFlow demonstrates its flexibility for creative video creation. **Bottom:** Applications in local recolorization. By leveraging a powerful segmentation model, UniColorFlow enables precise modification of local visual attributes, empowering workflows in film production, social media and advertising design.

Scalability. Our method can be extended to higher-resolution video colorization. As shown in Fig. 12, UniColorFlow still produces visually coherent and fine-grained results with a resolution of 720×1280 .

Diversity colorization. We utilize diverse spatial and language descriptions for the same grayscale video, enabling our model to render varied colorization results. As presented in Fig. 13 top, UniColorFlow effectively controls the colors of different instances according to user conditions, confirming its flexibility for creative applications.

Local recolorization. We leverage powerful video segmentation models (*e.g.*, SAM 2 (Ravi et al, 2025)) to detect instances within the video, enabling the model to recolorize arbitrary videos. As presented in Fig. 13 bottom, UniColorFlow can precisely modify local visual attributes, thereby empowering workflows in film production and advertising design.

Black-and-white film colorization. We apply our model to colorize historical black-and-white films to showcase its potential as a powerful tool for artistic restoration. As presented

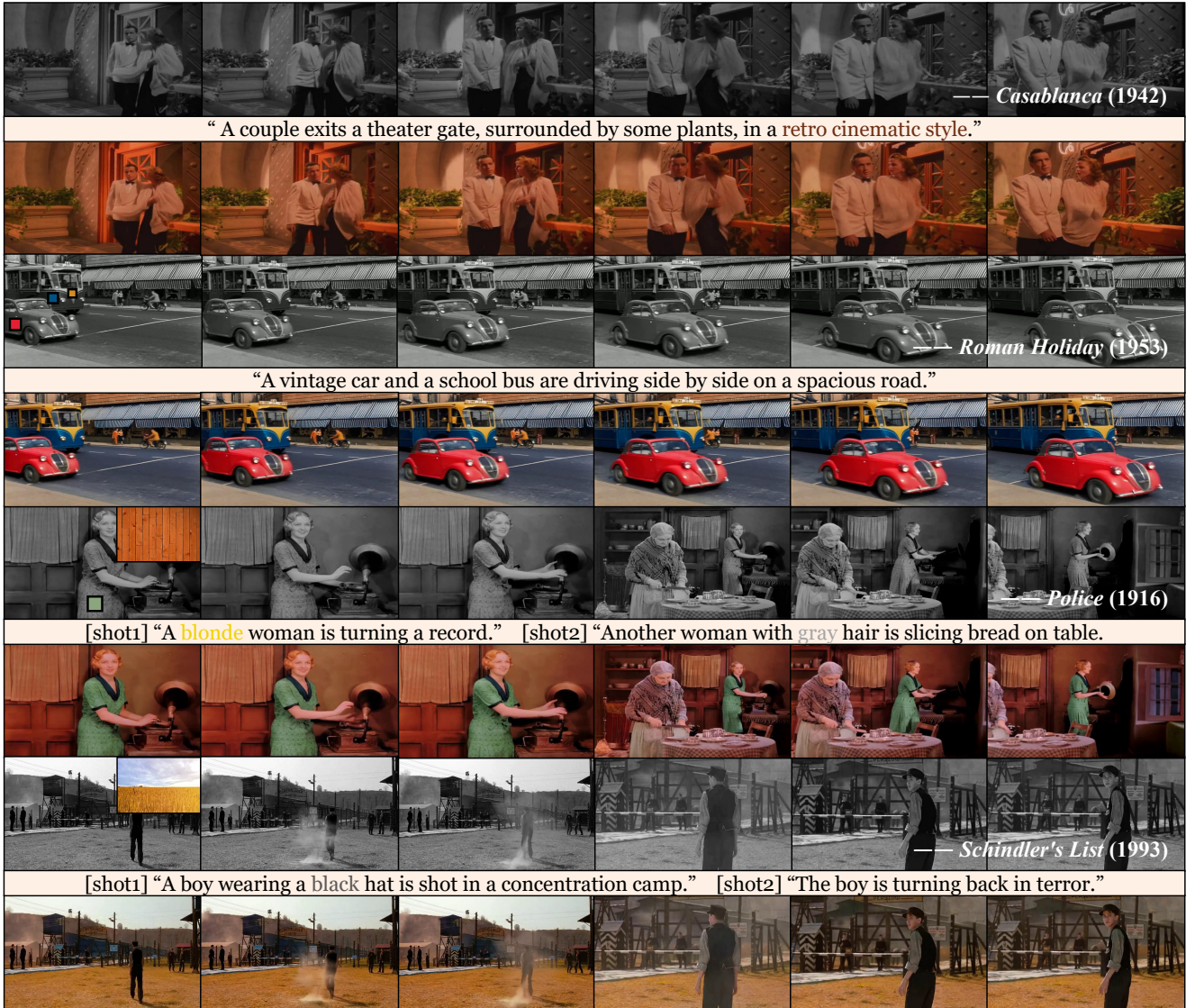


Fig. 14 Applications in black-and-white film colorization. Given historical films, UniColorFlow restores their colors, lowering the barrier for realizing artistic intentions.

in Fig. 14, UniColorFlow restores the colors of these old films, even across shot boundaries. This lowers the barrier for realizing user-guided artistic intention.

6 Conclusion

In this paper, we present UniColorFlow, a unified framework for controllable multimodal-guided video colorization designed to overcome the critical challenges of color adherence, condition integration, multi-shot consistency, and computational efficiency. By introducing a luminance-decoupled augmentation strategy, we mitigate the model’s inherent biases in the training data, ensuring faithful adherence to user-provided color conditions. Furthermore, we propose a unified condi-

tion network, equipped with a shot-aware attention mechanism, to seamlessly integrate diverse spatial and language conditions. This design effectively resolves condition conflicts and maintains multi-shot consistency. Finally, to reduce both training and inference overhead, we propose an efficient one-stage training strategy, which enables high-fidelity generation with only 2 sampling steps. Extensive experiments demonstrate that UniColorFlow outperforms state-of-the-art methods in terms of visual realism, color fidelity, and temporal consistency, offering a practical and efficient solution for real-world video creation workflows.

Acknowledgement. This work is supported by National Natural Science Foundation of China (Grant No. 62136001), Beijing Major Science and Technology Project (Grant No. Z251100008125009), Beijing High Innovation Plan (Con-

tract No. 202604841443), and Science and Technology Assistance Project of Beijing University of Posts and Telecommunications (Grant No. 2025KJBF02).

Data availability. Datasets used in this study will be publicly released upon publication.

References

- Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S (2012) Slic superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI*
- Bai S, Chen K, Liu X, Wang J, Ge W, Song S, Dang K, Wang P, Wang S, Tang J, Zhong H, Zhu Y, Yang M, Li Z, Wan J, Wang P, Ding W, Fu Z, Xu Y, Ye J, Zhang X, Xie T, Cheng Z, Zhang H, Yang Z, Xu H, Lin J (2025) Qwen2.5-vl technical report. [2502.13923](#)
- Bain M, Nagrani A, Brown A, Zisserman A (2020) Condensed movies: Story based retrieval with contextual embeddings. [2005.04208](#)
- Bertasius G, Wang H, Torresani L (2021) Is space-time attention all you need for video understanding. In: *ICML*
- Bian Y, Zhang Z, Ju X, Cao M, Xie L, Shan Y, Xu Q (2025) Videopainter: Any-length video inpainting and editing with plug-and-play context control. In: *ACM SIGGRAPH*
- Black-Forest-Labs, Batifol S, Blattmann A, Boesel F, Consul S, Diagne C, Dockhorn T, English J, English Z, Esser P, Kulal S, Lacey K, Levi Y, Li C, Lorenz D, Müller J, Podell D, Rombach R, Saini H, Sauer A, Smith L (2025) Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. [2506.15742](#)
- Bozic V, Djelouah A, Zhang Y, Timofte R, Gross M, Schroers C (2024) Versatile vision foundation model for image and video colorization. In: *ACM SIGGRAPH*
- Caesar H, Uijlings J, Ferrari V (2018) COCO-Stuff: Thing and stuff classes in context. In: *CVPR*
- Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: *CVPR*
- Chang Z, Weng S, Li Y, Li S, Shi B (2022) L-CoDer: Language-based colorization with color-object decoupling transformer. In: *ECCV*
- Chang Z, Weng S, Zhang P, Li Y, Li S, Shi B (2023a) L-CAD: Language-based colorization with any-level descriptions using diffusion priors. In: *NeurIPS*
- Chang Z, Weng S, Zhang P, Li Y, Li S, Shi B (2023b) L-CoIns: Language-based colorization with instance awareness. In: *CVPR*
- Chang Z, Weng S, Ouyang H, Lin L, Li Y, Li S, Shi B (2026) L-C4: Language-based video colorization for creative and consistent color. *Neurocomputing*
- Cheng Z, Yang Q, Sheng B (2015) Deep colorization. In: *ICCV*
- Chia AYS, Zhuo S, Gupta RK, Tai YW, Cho SY, Tan P, Lin S (2011) Semantic colorization with internet images. *ACM TOG*
- Chung HW, Constant N, Garcia X, Roberts A, Tay Y, Narang S, Firat O (2023) Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In: *ICLR*
- Esser P, Chiu J, Atighehchian P, Granskog J, Germanidis A (2023) Structure and content-guided video synthesis with diffusion models. In: *ICCV*
- Fang Z, Liu Z, Zhu K, Liu Y, Cheng KL, Zhai W, Cao Y, Zha ZJ (2025) Vangogh: A unified multimodal diffusion-based framework for video colorization. [2501.09499](#)
- Hasler D, Süsstrunk SE (2003) Measuring colorfulness in natural images. In: *Human vision and electronic imaging VIII*
- Ho J, Chan W, Saharia C, Whang J, Gao R, Gritsenko A, Kingma DP, Poole B, Norouzi M, Fleet DJ, Salimans T (2022) Imagen video: High definition video generation with diffusion models. [2210.02303](#)
- Huang Z, Zhao N, Liao J (2022) Unicolor: A unified framework for multi-modal colorization with transformer. *ACM TOG*
- Huang Z, Zhang M, Liao J (2024) Lvcd: Reference-based lineart video colorization with diffusion models. *ACM TOG*
- Huynh-Thu Q, Ghanbari M (2008) Scope of validity of PSNR in image/video quality assessment. *Electronics letters*
- Iizuka S, Simo-Serra E (2019) Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM TOG*
- Irony R, Cohen-Or D, Lischinski D (2005) Colorization by example. In: *Proceedings of the sixteenth eurographics conference on rendering techniques*
- Jiang Z, Han Z, Mao C, Zhang J, Pan Y, Liu Y (2025) Vace: All-in-one video creation and editing. In: *ICCV*
- Ke J, Wang Q, Wang Y, Milanfar P, Yang F (2021) Musiq: Multi-scale image quality transformer. In: *ICCV*
- Ku M, Wei C, Ren W, Yang H, Chen W (2024) Anyv2v: A tuning-free framework for any video-to-video editing tasks. *Transactions on Machine Learning Research*
- Lai WS, Huang JB, Wang O, Shechtman E, Yumer E, Yang MH (2018) Learning blind video temporal consistency. In: *ECCV*
- Lai Z, Sun K, Wang FY, Sagar D, Ding E (2025) Instantportrait: One-step portrait editing via diffusion multi-objective distillation. In: *ICLR*
- Lei C, Chen Q (2019) Fully automatic video colorization with self-regularization and diversity. In: *CVPR*
- Levin A, Lischinski D, Weiss Y (2004) Colorization using optimization. *ACM TOG*
- Li B, Wang Z, Li F, Xu J, Guo J, Pei R, Li X, Chen Z (2026) Colorflux: A structure-color decoupling framework for old photo colorization. In: *CVPR*
- Li J, Zhao H, Wang Y, Lin J (2024) Towards photorealistic video colorization via gated color-guided image diffusion models. In: *ACM MM*
- Lindenberg P, Sarlin PE, Pollefeys M (2023) Lightglue: Local feature matching at light speed. In: *ICCV*
- Liu H, Li C, Wong TT, Liu X (2025a) Chromaflow: End-to-end flow matching for efficient video colorization. In: *ACM SIGGRAPH Asia*
- Liu H, Xie M, Xing J, Li C, Leung CS, Wong TT (2025b) Colordiffuser: Video colorization with pretrained text-to-image diffusion models. In: *ACM MM*
- Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, Li C, Yang J, Su H, Zhu J, et al (2024a) Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: *ECCV*
- Liu X, Wan L, Qu Y, Wong TT, Lin S, Leung CS, Heng PA (2008) Intrinsic colorization. In: *ACM SIGGRAPH*
- Liu Y, Zhao H, Chan KC, Wang X, Loy CC, Qiao Y, Dong C (2024b) Temporally consistent video colorization with deep feature propagation and self-regularization learning. *CVM*
- Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: *ICLR*
- Meng Y, Ouyang H, Wang H, Wang Q, Wang W, Cheng KL, Liu Z, Shen Y, Qu H (2025) Anidoc: Animation creation made easier. In: *CVPR*
- Nan K, Xie R, Zhou P, Fan T, Yang Z, Chen Z, Li X, Yang J, Tai Y (2025) Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In: *ICLR*
- Peebles W, Xie S (2023) Scalable diffusion models with transformers. In: *ICCV*
- Perazzi F, Pont-Tuset J, McWilliams B, Van Gool L, Gross M, Sorkine-Hornung A (2016) A benchmark dataset and evaluation methodology for video object segmentation. In: *CVPR*
- Qi C, Cun X, Zhang Y, Lei C, Wang X, Shan Y, Chen Q (2023) Fatezero: Fusing attentions for zero-shot text-based video editing. In: *ICCV*
- Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al (2021) Learning transferable visual models from natural language supervision. In: *ICML*

- Ravi N, Gabeur V, Hu YT, Hu R, Ryali C, Ma T, Khedr H, Rädle R, Rolland C, Gustafson L, Mintun E, Pan J, Alwala KV, Carion N, Wu CY, Girshick R, Dollár P, Feichtenhofer C (2025) Sam 2: Segment anything in images and videos. In: ICLR
- Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: CVPR
- Sangkloy P, Lu J, Fang C, Yu F, Hays J (2017) Scribbler: Controlling deep image synthesis with sketch and color. In: CVPR
- Song Y, Dhariwal P, Chen M, Sutskever I (2023) Consistency models. In: ICML
- Thasaratthan H, Nazeri K, Ebrahimi M (2019) Automatic temporally coherent video colorization. In: CRV
- Unterthiner T, van Steenkiste S, Kurach K, Marinier R, Michalski M, Gelly S (2019) FVD: A new metric for video generation. In: ICLR
- Wan Z, Zhang B, Chen D, Liao J (2022) Bringing old films back to life. In: CVPR
- Wang W, Xie k, Liu Z, Chen H, Cao Y, Wang X, Shen C (2023a) Zero-shot video editing using off-the-shelf image diffusion models. [2303.17599](#)
- Wang X, Yuan H, Zhang S, Chen D, Wang J, Zhang Y, Shen Y, Zhao D, Zhou J (2023b) Videocomposer: Compositional video synthesis with motion controllability. In: NeurIPS
- Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: From error visibility to structural similarity. IEEE TIP
- WanTeam (2025) Wan: Open and advanced large-scale video generative models. [2503.20314](#)
- Weng S, Sun J, Li Y, Li S, Shi B (2022a) CT²: Colorization transformer via color tokens. In: ECCV
- Weng S, Wu H, Chang ZC, Tang J, Li S, Shi B (2022b) L-CoDe: Language-based colorization using color-object decoupled conditions. In: AAAI
- Weng S, Zheng H, Zhang P, Hong Y, Jiang H, Li S, Shi B (2025) Vires: Video instance repainting via sketch and text guided generation. In: CVPR
- Yang L, Zhang Z, Zhang Z, Liu X, Xu M, Zhang W, Meng C, Ermon S, Cui B (2024a) Consistency flow matching: Defining straight flows with velocity consistency. [2407.02398](#)
- Yang Y, Dong J, Tang J, Pan J (2024b) ColorMNet: A memory-based deep spatial-temporal feature propagation network for video colorization. In: ECCV
- Yang Y, Pan J, Peng Z, Du X, Tao Z, Tang J (2024c) BiSTNet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. IEEE TPAMI
- Yang Z, Teng J, Zheng W, Ding M, Huang S, Xu J, Yang Y, Hong W, Zhang X, Feng G, et al (2025) CogVideox: Text-to-video diffusion models with an expert transformer. In: ICLR
- Zhang B, He M, Liao J, Sander PV, Yuan L, Bermak A, Chen D (2019) Deep exemplar-based video colorization. In: CVPR
- Zhang R, Isola P, Efros AA (2016) Colorful image colorization. In: ECCV
- Zhang R, Zhu JY, Isola P, Geng X, Lin AS, Yu T, Efros AA (2017) Real-time user-guided image colorization with learned deep priors. ACM TOG
- Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR
- Zhang Y, Wei Y, Jiang D, Zhang X, Zuo W, Tian Q (2024) ControlVideo: Training-free controllable text-to-video generation. In: ICLR
- Zhao H, Li J, Zhang P, Xiao P, Lin J, Wang Y (2025) Colorsurge: Bringing vibrancy and efficiency to automatic video colorization via dual-branch fusion. In: ACM SIGGRAPH
- Zhao Y, Po LM, Liu K, Wang X, Yu WY, Xian P, Zhang Y, Liu M (2023a) Svcnet: Scribble-based video colorization network with temporal aggregation. IEEE TIP
- Zhao Y, Po LM, Yu WY, Rehman YAU, Liu M, Zhang Y, Ou W (2023b) VCGAN: Video colorization with hybrid generative adversarial network. IEEE TMM
- Zhuang J, Guo S, Cai X, Li X, Liu Y, Yuan C, Xue T (2026) Flashvsr: Towards real-time diffusion-based streaming video super-resolution. In: CVPR