DA-CoLoR: Scalable Text-Guided Long-Video Colorization with Dual Attention and Video Diffusion

Rintaro Kimura[1] and Keiji Yanai[1]

The University of Electro-Communications, Tokyo, Japan
kimura-r@mm.inf.uec.ac.jp, yanai@cs.uec.ac.jp

Abstract. We propose DA-CoLoR, a training-free video colorization framework that balances flexible text-prompt-based color control with temporal consistency in long videos. Based on a Text-to-Reference-to-Video paradigm, our method combines reference image generation, Dual Attention-based color injection into sparse anchor frames, bidirectional Stable Video Diffusion for anchor-to-anchor interpolation, and Lab-space luminance restoration, without retraining pretrained image or video diffusion models. This design decouples spatial color transfer from temporal color propagation, enabling scalable and stable colorization for long video sequences. On DAVIS30 and Videvo, DA-CoLoR achieves the lowest reported LPIPS among the compared methods. Evaluations over 240–720 frames further show stable LPIPS and CDC with constant peak GPU memory under segment-wise processing.

Keywords: Video colorization · Diffusion models · Temporal consistency · Text-guided colorization

1 Introduction

Historical materials and archival footage often remain in monochrome, and there is a high demand for preserving and utilizing them through colorization. While recent advances in diffusion models have enabled natural, text-driven colorization, extending them to long videos remains a major challenge. Generating dense frames directly with heavy diffusion models not only leads to computational explosion but also suffers from severe temporal drift and flicker as the sequence lengthens.

To address these challenges, we argue that robust long-video colorization requires fundamentally decoupling spatial color transfer from temporal propagation. Based on this design principle, we propose a novel framework, Dual-Attention Controllable Long-video Recolorization (DA-CoLoR). Instead of relying on unstable frame-by-frame text-to-video generation, we introduce a Text-to-Reference-to-Video paradigm. First, we utilize a text-guided image editing model to generate a single, high-quality reference image from a text prompt and the center frame. Crucially, we employ a Dual Attention mechanism not as a

naive video adapter, but as a key enabler to strictly anchor this global color representation onto sparse keyframes (spatial transfer). Finally, we propagate these colors to intermediate frames using Stable Video Diffusion (SVD) guided by bidirectional inference and structural ControlNet (temporal propagation).

The main contributions of this paper are summarized as follows:

- We propose DA-CoLoR, a novel divide-and-conquer framework that decouples spatial color transfer from temporal propagation, providing a favorable balance between perceptual quality and temporal consistency in long videos.
- We introduce a Text-to-Reference-to-Video paradigm, utilizing a text-guided image editing model for global color anchoring and positioning Dual Attention as a critical mechanism to strictly enforce this color representation across sparse keyframes.
- Our method achieves strong PSNR and LPIPS results on DAVIS30 and Videvo, with constant peak GPU memory across the evaluated sequence lengths under segment-wise processing.

2 Related Work

2.1 Types and Challenges of Video Colorization Methods

Image colorization technology has drastically improved with the advent of CNNs, GANs, and diffusion models. When extending this to video, the greatest challenge is ensuring temporal consistency. Conventional methods can be broadly classified into five categories.

- **Optical flow-based**: Methods that maintain temporal consistency by calculating the movement of pixels and features between adjacent frames (optical flow) and propagating color information based on it [2, 14].
- **Scribble-based**: Methods where a user specifies colors in specific regions within a video frame, and the colors are propagated to the target object as a whole and to subsequent frames [7, 15, 22].
- **Reference-based**: Approaches that extract corresponding colors from a colored reference image and transfer them to grayscale video frames [4, 10, 24, 25].
- **Fully automatic**: Methods that colorize end-to-end without requiring reference images or user scribbles [13, 18, 27].
- **Prompt-based**: Methods that freely control colors using natural language via text prompts during colorization, which is adopted by many of the latest studies [3, 5, 16, 17].

2.2 Positioning of This Study and Foundational Technologies

This research is positioned within prompt-based video colorization. However, unlike conventional methods that struggle with temporal drift by individually generating each frame from text, we introduce a "Text-to-Reference-to-Video"

paradigm. This paradigm fundamentally decouples the colorization process into spatial color transfer and temporal propagation.

As foundational technologies, we adopt Qwen-Image-Edit [23] for generating reference images, which excels in the consistency of text understanding and image generation. For transferring color information in anchor frames, we introduce the Dual Attention mechanism proposed by Kosugi [11] to accurately inject the color information of the reference image according to the structure of the input image. Additionally, we utilize the video generation model Stable Video Diffusion (SVD) [1] for interpolating intermediate frames between anchor frames. Here, based on the framework of Liu et al. [19], we combine bidirectional inference from the start and end points with structural control using a multimodal ControlNet utilizing edges and optical flow. This suppresses the drift phenomenon characteristic of diffusion models and achieves high-quality and temporally smooth colorization even for long videos.

3 Proposed Method

In this paper, we propose a novel video colorization framework, Dual-Attention Controllable Long-video Recolorization (DA-CoLoR), which takes a grayscale video and a text prompt as input and achieves both temporal consistency and high perceptual quality.

3.1 Pipeline Overview

The overall pipeline of the proposed DA-CoLoR framework is illustrated in Fig. 1. DA-CoLoR is a training-free inference framework: all image and video diffusion backbones are kept frozen, and the proposed design connects them through reference anchoring, sparse color transfer, bidirectional temporal propagation, and Lab-space integration. To realize our design principle of decoupling spatial color transfer from temporal propagation, the framework is structured into four sequential stages: (1) **Reference Image Generation**, which creates a global color anchor from a text prompt; (2) **Anchor Frame Colorization**, which precisely transfers this color representation to sparse keyframes using a Dual Attention mechanism; (3) **Intermediate Frame Interpolation**, which propagates the colors temporally via bidirectional Stable Video Diffusion (SVD) and structural ControlNet; and (4) **Integration in Lab Space**, which restores the original high-frequency details. The detailed formulation of each stage is described in the following subsections.

3.2 Reference Image Generation

To achieve consistent colorization throughout the video, our method extracts the center frame v_{mid} from the input grayscale video and colorizes it based on text conditions as the reference image. We use Qwen-Image-Edit [23] as the generation model. With Qwen2.5-VL as the condition encoder and MMDiT as

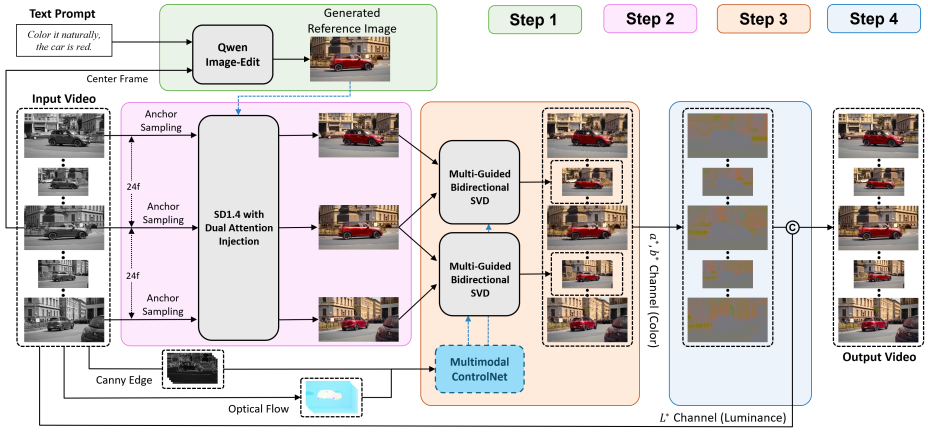


Fig. 1: Overview of the proposed DA-CoLoR framework. The framework decouples spatial color transfer and temporal propagation to achieve both high perceptual quality and temporal consistency. (Step 1) A reference image is first generated from the center frame using a text-guided image editing model. (Step 2) Dual Attention is then used to inject the reference color representation into sparse anchor frames within a Stable Diffusion backbone. (Step 3) Intermediate frames are subsequently interpolated using bidirectional Stable Video Diffusion guided by structural constraints from Canny edges and optical flow through a multimodal ControlNet. (Step 4) Finally, the chrominance channels of the generated frames are integrated with the luminance channel of the original grayscale video in the Lab color space to restore fine structural details.

the diffusion backbone, it enables color editing faithful to the text instructions while preserving the input structure. Specifically, we input v_{mid} and a prompt T containing color specifications to generate a structural-preserving color reference image I_{ref}. The obtained I_{ref} functions as the sole color reference in the subsequent anchor frame colorization, supporting consistent color propagation aligned with the prompt's intent.

3.3 Anchor Frame Colorization

In this process, using Stable Diffusion 1.4 (SD1.4) [21] as the base model, we apply the Dual Attention mechanism proposed by Kosugi [11] in the high-resolution latent space to accurately transfer the color of the reference image according to the structure of the input frame. The process consists of three steps: feature extraction by inverting the diffusion process, color transfer using Dual Attention, and refinement through statistical correction.

Inversion of Diffusion Process and Feature Extraction Given an input grayscale anchor frame and a color reference image, we obtain their latent representations using the SD1.4 VAE encoder: z_0^{in} and z_0^{ref}. We also construct a

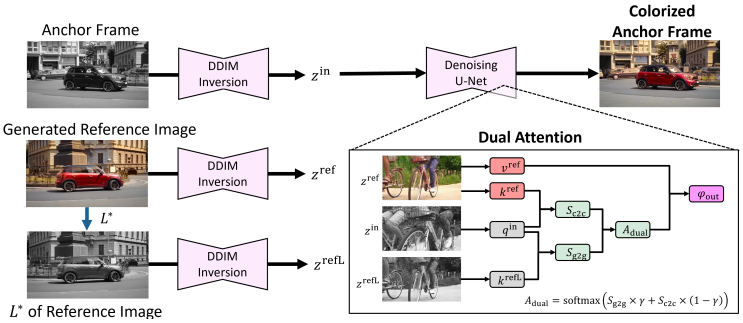


Fig. 2: Schematic diagram of anchor frame colorization.

luminance-only version of the reference and encode it as z_0^{refL} to capture structural correspondence independent of chrominance. We then apply DDIM inversion to these latents to estimate the corresponding initial noises z_T. During inversion, we cache the self-attention features required for Dual Attention at each diffusion step t, namely the query/key from the input and the key/value (and luminance-key) from the reference. These cached features are reused during the subsequent denoising process to inject reference-aligned colors while preserving the input structure.

Dual Attention Mechanism The generation of the colorized anchor frame is performed by removing noise through DDIM sampling.

For the initial latent variable z_T, AdaIN (Adaptive Instance Normalization) [9] is applied, adjusting its mean and variance to match the features of the reference image, thereby preparing the initial state for color generation.

In the update step of the diffusion model, inference is performed by replacing the normal Self-Attention layer with a Dual Attention mechanism as shown in Fig. 2. Here, two attention maps are created using the saved features, and by integrating them, structure preservation and color transfer are achieved simultaneously.

First, to align the structure of the input image and the reference image, Gray-to-Gray Attention is calculated using the saved input query q_t^{in} and the grayscale reference key k_t^{refL}.

$$S_{g2g} = \frac{q_t^{\text{in}}(k_t^{\text{refL}})^T}{\sqrt{d}} \tag{1}$$

Second, Colorized-to-Color Attention is calculated using the query q_t^{out} in the current generation process and the reference key k_t^{ref}.

$$S_{c2c} = \frac{q_t^{\text{out}}(k_t^{\text{ref}})^T}{\sqrt{d}} \tag{2}$$

These are mixed with weights to obtain the final attention map A_{dual}.

$$A_{\text{dual}} = \text{softmax}(S_{g2g} \times \gamma + S_{c2c} \times (1-\gamma)) \tag{3}$$

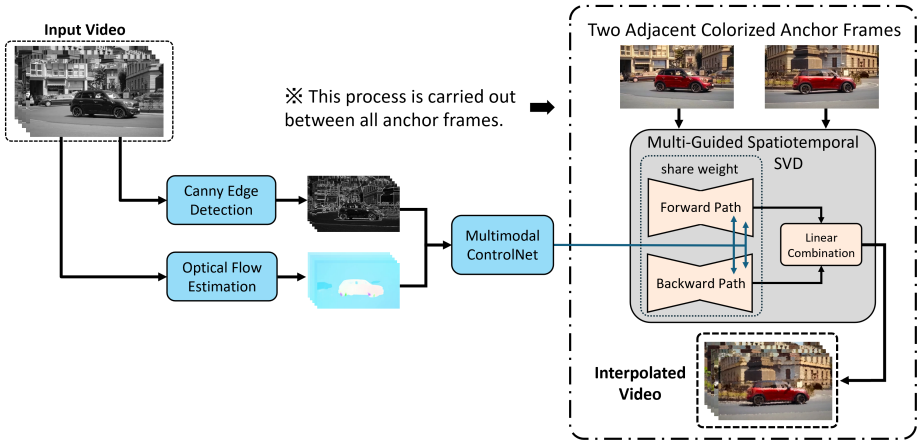


Fig. 3: Schematic diagram of intermediate frame interpolation.

Based on this map, the reference value features v_t^{ref} are injected into the input features, transferring colors from semantically corresponding regions. We use layer-wise fixed values of γ: intermediate U-Net layers balance gray-to-gray structural alignment and colorized-to-color matching, while later layers rely more on gray-to-gray alignment to retrieve reference color features.

Furthermore, in noise prediction, Classifier-free Colorization Guidance is applied. By executing both conditional inference with color transfer via Dual Attention and unconditional inference, and updating by amplifying the difference between them, an image that reflects the vividness of the reference image is generated while preserving the structure of the input.

Statistical Correction and Iterative Refinement After decoding the generated latent z_0^{out} to pixel space, we apply lightweight post-corrections to suppress minor tone inconsistencies. Specifically, we reconstruct the final anchor frame by combining the input luminance with the generated chrominance in the Lab space, which helps recover fine structural details that may be attenuated by diffusion. We additionally perform a simple chrominance alignment to keep the overall color tone close to the reference. For quality improvement, we optionally run iterative refinement by re-encoding the current output and repeating inversion and denoising; in our experiments, we use three refinement iterations.

3.4 Intermediate Frame Interpolation

Since the colorized anchor frames generated in the previous stage are temporally sparse, it is necessary to use them as boundary conditions to highly accurately generate the frames in between and reconstruct a temporally smooth color video. In our method, we perform conditional video generation using Stable Video Diffusion (SVD) [1] (Fig. 3).

Extraction of Structural and Motion Constraints In simple video interpolation, frames with different movements and structures from the original video's intermediate frames may be output. Therefore, to interpolate only colors while fully preserving the shape and motion of the original grayscale video, we adopt the AnchorSync [19] framework, which controls structure and motion with a multimodal ControlNet.

First, Canny edge detection and optical flow estimation are applied to each frame. To process these conditional inputs, the conditional features are encoded using a multimodal ControlNet fine-tuned by Liu et al. [19].

The extracted features are injected into each block of the frozen SVD U-Net, allowing the specified color information from the anchor frames to be properly propagated while faithfully reproducing the dynamics of the original grayscale video.

Bidirectional Inference by Video Diffusion Model For intermediate frame generation, we use SVD, an image-conditioned video generation model, as the foundation. Standard SVD generates sequentially in the forward temporal direction conditioned only on the starting frame, causing accumulated errors as the generation sequence lengthens, resulting in a drift phenomenon where the sequence becomes discontinuous with the end anchor frame [19].

To solve this problem, our method introduces bidirectional temporal frame fusion. Specifically, using two adjacent anchor frames f'_i (start) and f'_{i+1} (end), the following two generation processes are executed in parallel.

1. **Forward Pass:** The start f'_i is input as the condition image, and diffusion model inference is performed in the forward temporal direction. The latent representation obtained at each time step is denoted as Z^{forward}.
2. **Backward Pass:** The end f'_{i+1} is input as the condition image, and inference is performed in the backward temporal direction (or on the reversed sequence). The obtained latent representation is denoted as Z^{reverse}.

At each denoising step, these two latent representations are linearly synthesized to obtain the fused latent representation Z_j^{fuse}.

$$Z_j^{\text{fuse}} = \alpha \cdot Z_j^{\text{forward}} + (1 - \alpha) \cdot Z_j^{\text{reverse}} \tag{4}$$

Following [19], the fusion weight α is linearly scheduled within each anchor segment, assigning larger weights to the forward latent near the start anchor and to the reverse latent near the end anchor. This differs from BiSTNet [24], which fuses correspondence-warped color maps from external reference exemplars; here, we fuse forward and reverse SVD latent states generated from two colorized anchor boundaries to suppress diffusion drift during anchor-to-anchor interpolation. Through this fusion process, the generated sequence of intermediate frames strongly reflects the features of the start anchor near the beginning and the end anchor near the end. As a result, a color video segment with a smooth transition that is consistent with the anchor frames at both ends and less prone to artifacts is generated. This process is applied to all anchor segments and concatenated to complete the colorization of the entire video.

3.5 Integration in Lab Space

While intermediate frame interpolation yields a temporally consistent color video, the generation process of diffusion models can cause degradation in textures or minor artifacts. On the other hand, the input grayscale video retains its resolution and high-frequency components, making its structural information highly reliable. Therefore, in our method, we separate the color information of the generated video and the structural information of the input video and recombine them to improve the final quality. Specifically, the generated video sequence V_{gen} and the input grayscale video sequence V_{in} are converted to the Lab space, extracting the color difference components a^*, b^* from V_{gen} and the luminance component L^* from V_{in}. Then, $L^*(V_{\text{in}})$ and $a^*, b^*(V_{\text{gen}})$ are integrated to reconstruct the Lab image, which is then inversely converted to RGB to obtain the final video V_{final}.

This luminance replacement recovers detailed structures easily lost during diffusion generation while maintaining color consistency, realizing high-quality colorization results.

4 Experiments

4.1 Implementation Details

For all experiments, the common prompt *"Colorize it naturally with colorful tone."* was used unless otherwise specified. The anchor frame interval was set to $K = 24$ following [19] as a practical default that balances interpolation quality and computational cost; smaller intervals increase the number of anchor colorization steps, whereas larger intervals make long-range interpolation more challenging. We used 30 sampling steps for reference image generation, 50 DDIM steps for anchor frame colorization, and 25 denoising steps for SVD interpolation. For Dual Attention, the attention mixing weights were set to $\gamma_1 = 0.5$ for the 6th–8th U-Net layers and $\gamma_2 = 1.0$ for the 9th–11th layers, with the colorization guidance scale $w = 10.0$. During SVD interpolation, the guidance scale was linearly increased from 1.0 to 8.0, the motion bucket ID was 127, and the ControlNet strength was set to 1.0. We fixed the reference-generation seed to 42, applied three rounds of iterative refinement with AdaIN initialization, and used feature-injection thresholds of $\tau_{attn} = 0.44$ and $\tau_f = 0.65$. Intermediate frames were generated at the SVD resolution of 576×320, and final outputs were reconstructed in Lab space by combining the input luminance channel with the generated chrominance channels. Detailed model configurations (Qwen-Image-Edit, SD1.4, and SVD with multimodal ControlNet) and hyperparameters are provided in the supplementary material.

4.2 Experimental Settings

Datasets To verify the effectiveness of the proposed method, we used the DAVIS30 dataset [20], widely used as a benchmark for video colorization, and the Videvo [12] dataset, which contains relatively long videos.

Evaluation Metrics To evaluate the generated color videos from multiple perspectives, we use standard image fidelity and perceptual metrics, including **PSNR**, **SSIM**, **LPIPS** [26], and **FID** [8]. We also report **CDC (Color Distribution Consistency)** [18], which evaluates temporal stability of color distributions with smaller values indicating less flicker, and **Colorfulness** [6], which quantifies the richness of generated colors with larger values indicating more vivid results.

Compared Methods To position the performance of the proposed method, we compared it with the following representative video colorization methods. As conventional methods, we adopted the automatic colorization AutoColor [13], the reference-based DeepExemplar [25] and DeepRemaster [10], the temporally consistent TCVC [18], and the GAN-based VCGAN [27]. We also compared with the latest diffusion model and prompt-based methods, ColorDiffuser [17], L-C4 [3], and VanGogh [5].

4.3 Quantitative Evaluation

The quantitative evaluation results on the DAVIS30 and Videvo datasets are shown in Table 1 and Table 2. The values for ColorDiffuser [17], L-C4 [3], and VanGogh [5] are cited from their respective papers, while the values for other compared methods are cited from [17]. For metrics other than Colorfulness, ↑/↓ denote higher/lower is better; Colorfulness ↑ denotes greater vividness only. **Bold** indicates the best result, and underline indicates the second best. Note that data marked with "-" are not reported in the cited papers or public releases.

Table 1: Quantitative evaluation results on DAVIS30.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CDC ↓	Colorfulness ↑
AutoColor (CVPR2019) [13]	24.41	0.915	0.264	83.05	0.003734	14.14
DeepExemplar (CVPR2019) [25]	21.78	0.846	0.325	77.26	0.004006	28.82
DeepRemaster (TOG2019) [10]	21.95	0.848	0.354	97.54	0.005098	25.66
TCVC (CVM2024) [18]	25.17	0.921	0.239	74.94	0.003649	21.72
VCGAN (TMM2023) [27]	23.90	0.910	0.247	70.29	0.005303	15.89
ColorDiffuser (ACMMM2025) [17]	23.73	**0.939**	0.213	69.51	0.003607	29.13
VanGogh (arXiv2025) [5]	23.20	**0.939**	0.191	–	–	**60.09**
L-C4 (Neurocomputing 2026) [3]	25.69	0.933	0.209	–	**0.003114**	29.33
DA-CoLoR (Ours)	**25.91**	**0.939**	**0.114**	**61.47**	0.004754	47.76

DA-CoLoR achieves the highest reported PSNR and lowest reported LPIPS on both datasets, with LPIPS scores of 0.114 on DAVIS30 and 0.103 on Videvo. Colorfulness is 47.76 on DAVIS30 and 38.94 on Videvo, exceeding the respective GT values of 35.47 and 30.83; this indicates greater vividness, not necessarily

Table 2: Quantitative evaluation results on Videvo.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CDC ↓	Colorfulness ↑
AutoColor (CVPR2019) [13]	<u>25.90</u>	0.925	0.277	76.28	0.001668	13.23
DeepExemplar (CVPR2019) [25]	20.63	0.831	0.348	76.63	0.002011	32.44
DeepRemaster (TOG2019) [10]	21.88	0.856	0.358	86.23	0.003607	28.72
TCVC (CVM2024) [18]	25.18	0.929	0.273	76.02	0.001629	18.89
VCGAN (TMM2023) [27]	24.67	0.919	0.276	**63.83**	0.002998	14.90
ColorDiffuser (ACMMM2025) [17]	25.27	**0.951**	0.205	66.11	<u>0.001591</u>	20.73
L-C4 (Neurocomputing 2026) [3]	25.17	0.939	<u>0.198</u>	–	**0.001572**	<u>32.59</u>
DA-CoLoR (Ours)	**27.48**	<u>0.949</u>	**0.103**	<u>65.79</u>	0.002033	**38.94**

more natural colors. Outputs can be overly vivid. Our CDC is not the lowest among the reported methods; within-pipeline ablations and nested-window results instead provide evidence of temporal stability in the evaluated settings.

4.4 Long-Video Scalability

Although Videvo contains relatively long videos, dataset-level averages alone do not isolate the effect of sequence length. We therefore evaluate nested temporal windows from the same clips with 240, 480, and 720 frames, while keeping the prompt, center-frame reference, anchor interval ($K = 24$), SVD resolution (576×320), and Lab integration fixed. Runtime includes the full pipeline except model loading, and peak VRAM is measured for the inference process. We could not perform matched long-video comparisons with ColorDiffuser, L-C4, or VanGogh because runnable implementations and checkpoints were unavailable. The conventional baselines considered here do not support prompt-guided color control.

Table 3: Long-video scalability of DA-CoLoR on nested temporal windows. LPIPS/CDC/time are averaged over clips; peak VRAM is the maximum over runs.

Frames	LPIPS ↓	CDC ↓	Time (s)	s/frame	Peak VRAM
240	0.1072	0.000973	1009	4.206	24.4GB
480	0.1030	0.000976	1862	3.880	24.4GB
720	0.1071	0.001034	2719	3.776	24.4GB

Across the evaluated 240–720-frame windows, LPIPS and CDC remain stable, runtime grows approximately linearly, and peak VRAM remains at 24.4GB (Table 3). These results support scalable long-video colorization with bounded peak GPU memory through independent processing of anchor segments.

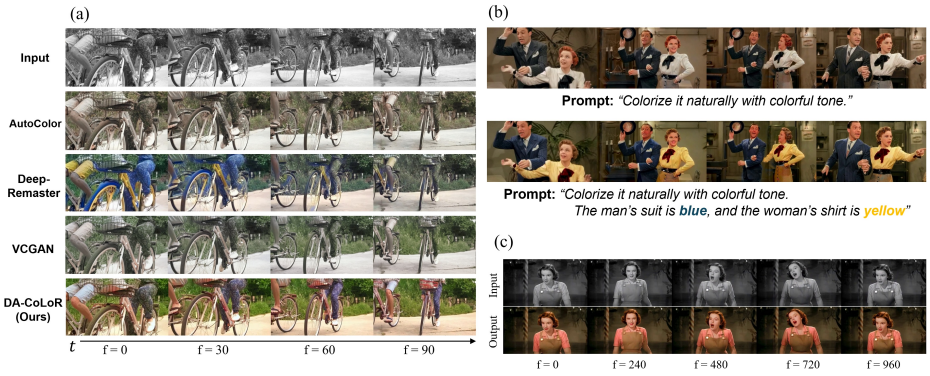


Fig. 4: (a) Comparison of colorization results between conventional methods and our proposed method. (b) Comparison of generation results using different prompts. (c) Colorization results of a long video.

4.5 Qualitative Evaluation

Fig. 4(a) compares methods on a bicycle video. Recent large-model prompt-guided methods are omitted from this qualitative comparison because their implementations and checkpoints are unavailable; Table 1 and Table 2 instead provide quantitative comparisons using values reported in their papers. Methods such as AutoColor [13] and VCGAN [27] generally have a tendency toward low saturation and color bleeding, and in DeepRemaster [10], color consistency is lost over time. On the other hand, the proposed method (Ours) accurately reflects the color information of the reference image through the Dual Attention mechanism, maintaining color consistency across anchor frames. Furthermore, structural constraints by ControlNet are effective in intermediate frames, realizing smooth colorization without edge breakdown even in scenes with motion. The supplementary material includes a small pilot user study.

Prompt Controllability Since our method adopts control via text prompts during reference image generation in the first stage, it allows not only restoration to the ground truth data but also free colorization according to the user's intent. Fig. 4(b) shows the results when different prompts are provided for the same grayscale video.

In the upper images, colorization is performed with a generic prompt *"Colorize it naturally with colorful tone."* On the other hand, in the lower images, by adding a prompt specifying colors such as *"The man's suit is blue, and the woman's shirt is yellow"* to the generic prompt, colorization aligns with the specified colors.

Application to Long Videos Fig. 4(c) shows a long-video result exceeding 300 frames. DA-CoLoR processes anchor segments independently, producing tempo-

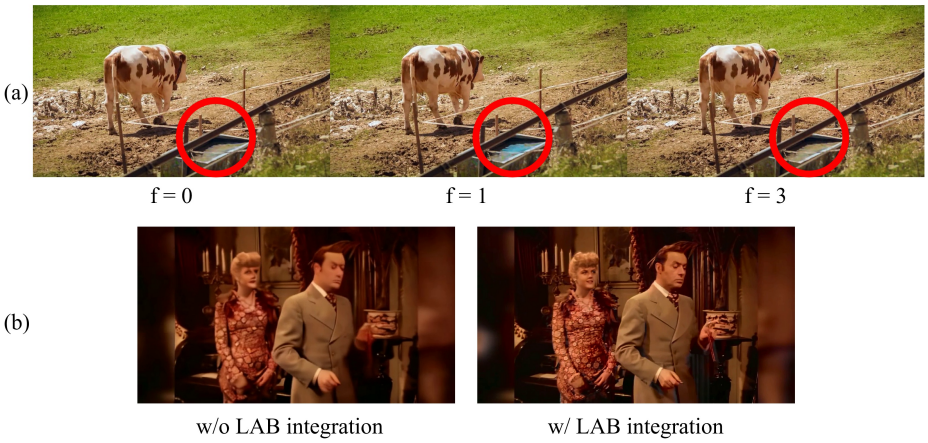


Fig. 5: (a) Result when anchor frame colorization is applied to all frames. (b) Comparison between without and with integration in the Lab space.

rally coherent colors in this example; the scalability evaluation reports constant peak VRAM of 24.4GB across the tested sequence lengths.

5 Ablation Study

5.1 Impact of Intermediate Frame Interpolation

We evaluated the impact of omitting intermediate frame interpolation by applying the anchor frame colorization to all frames ($K = 1$).

When all frames are colorized independently, minute differences in latent representations between adjacent frames can cause flicker, compromising temporal continuity. Fig. 5(a) shows that the water bucket color changes between adjacent frames.

Furthermore, since Stable Diffusion is computationally expensive, inference for all frames increases execution time. Specifically, anchor colorization for all frames ($K = 1$) took 724 seconds on DAVIS30, whereas our method ($K = 24$), which interpolates sparse anchor frames, took 476 seconds. Thus, the baseline without interpolation requires more than 1.5 times the duration of our proposed method (excluding the constant reference image generation time). This demonstrates that our approach of interpolating sparse anchor frames using a video generation model excels in both computational efficiency and temporal consistency.

5.2 Impact of Bidirectional Inference

We verified the necessity of bidirectional inference in intermediate frame interpolation. Fig. 6(A) shows a comparison between forward generation conditioned

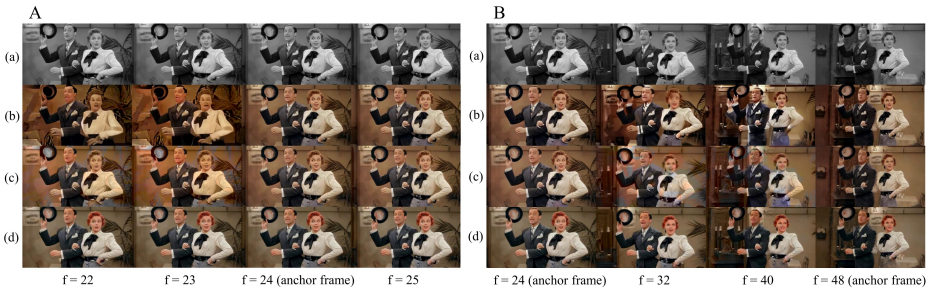


Fig. 6: (A) Comparison between applying only forward generation and applying bidirectional inference. (B) Comparison between not using and using the multimodal ControlNet.

only on the starting anchor frame and bidirectional inference utilizing both start and end points.

Fig. 6(A-b) is the output video before Lab integration in the case of only forward generation. Although global motion is somewhat captured due to the introduction of ControlNet, a significant divergence in color tone and structure is visible between the anchor frame $f = 24$ and its preceding frame $f = 23$. As a result, even in the final output after integrating in the Lab space (Fig. 6(A-c)), the continuity of color and structure is broken between the generated $f = 23$ and the highly accurately colorized $f = 24$, causing flicker during video playback.

On the other hand, in the proposed method utilizing bidirectional inference (Fig. 6(A-d)), information in the reverse direction from the end anchor frame ($f = 24$) is also fused in the latent space; thus, an image smoothly connecting to $f = 24$ is already generated by $f = 23$, leading to a colorization result with a natural transition even after Lab integration. This suppresses unnatural structural changes and color jumps around anchor frames, ensuring high temporal consistency. These results indicate that bidirectional inference based on AnchorSync is important for achieving smooth connections between segments when dividing and processing long videos.

5.3 Impact of Multimodal ControlNet

We verified the effectiveness of structural control via multimodal ControlNet using Canny edges and optical flow in intermediate frame interpolation, the third stage of our proposed method. Fig. 6(B) shows a comparison between the results without using ControlNet and the proposed method using it.

Fig. 6(B-b) shows the pre-Lab output without ControlNet, where body structures and backgrounds break down between anchor frames ($f = 32, 40$). This occurs because SVD interpolates from only the start and end anchors, ignoring the accurate motions and shapes of the original grayscale video. Thus, even after Lab integration, color-luminance inconsistency remains in Fig. 6(B-c), causing colors to shift from the subjects.

With ControlNet (Fig. 6(B-d)), the intermediate frames better preserve the input structure and motion (Fig. 6(B-a)), reducing color misalignment after Lab integration in this example. From these results, we show that structural constraints by multimodal ControlNet are important for preserving original video dynamics and preventing color misalignment during temporal interpolation between anchor frames.

5.4 Impact of Integration in Lab Space

We evaluated the impact of using integration in the Lab color space during the final stage.

When using intermediate frames generated by SVD as direct output (left side of Fig. 5(b)), although global motion and colors are generally correct, high-frequency components such as facial expressions and fabric textures are lost due to the compression-restoration process by the VAE of the video diffusion model, leading to an overall blurred appearance.

Conversely, using the L channel of the input grayscale video (right side of Fig. 5(b)), sharp edges and fine textures innate to the original footage are fully restored, confirming a significant enhancement in perceptual resolution.

6 Conclusion

In this study, we proposed DA-CoLoR, a scalable text-guided video colorization framework that balances flexible style control with high spatiotemporal consistency. To address the computational cost growth and flicker of diffusion-based video generation, DA-CoLoR combines reference image generation, anchor-frame colorization, bidirectional interpolation, and Lab-space integration. DA-CoLoR achieves the highest reported PSNR and lowest reported LPIPS among the compared methods on both datasets; ablations support reduced SVD-specific drift through bidirectional inference, and nested-window evaluation shows constant peak GPU memory over 240–720 frames. These results indicate that separating spatial color transfer from temporal propagation is effective for scalable prompt-guided video colorization.

Several challenges remain for diverse real-world videos: SVD's fixed aspect ratio and resolution can introduce artifacts for inputs far from its training distribution; reliance on edge and optical-flow guidance may still cause color bleeding or flicker under fast thin structures, occlusion, and crowded motion; and shot changes may break the single-reference assumption. Representative failure cases are provided in the supplementary material. Future work will incorporate scene detection, dynamic anchor selection, and multiple references to improve robustness for long video colorization.

Acknowledgements This work was supported by JST CRONOS JPMJCS24K4 and JSPS KAKENHI 26K23875.

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets (2023), arXiv:2311.15127
2. Bonneel, N., Tompkin, J., Sunkavalli, K., Sun, D., Paris, S., Pfister, H.: Blind video temporal consistency. ACM Trans. Graph. (SIGGRAPH) **34**(6), 196–204 (2015)
3. Chang, Z., Weng, S., Ouyang, H., Lin, L., Li, Y., Li, S., Shi, B.: L-C4: Language-based video colorization for creative and consistent color. Neurocomputing **665**, 132199 (2026)
4. Fang, X., Dai, L., Tang, J.: OmniFusion: Exemplar-based video colorization using omnimotion and diffusion priors. In: Proc. of Asian Conference on Computer Vision. pp. 1215–1232 (2024)
5. Fang, Z., Liu, Z., Zhu, K., Zhai, W., Cao, Y., Liu, Y., Cheng, K.L., Zha, Z.J.: VanGogh: A unified multimodal diffusion-based framework for video colorization (2025), arXiv:2501.09499
6. Hasler, D., Suesstrunk, S.E.: Measuring colorfulness in natural images. In: Human Vision and Electronic Imaging VIII. vol. 5007, pp. 87–95. SPIE (2003)
7. Heu, J.H., Hyun, D.Y., Kim, C.S., Lee, S.U.: Image and video colorization based on prioritized source propagation. In: Proc. of IEEE International Conference on Image Processing. pp. 465–468 (2009)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. vol. 30, pp. 6626–6637 (2017)
9. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proc. of IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)
10. Iizuka, S., Simo-Serra, E.: DeepRemaster: Temporal source-reference attention networks for comprehensive video enhancement. ACM Trans. Graph. (SIGGRAPH) **38**(6), 176 (2019)
11. Kosugi, S.: Leveraging the powerful attention of a pre-trained diffusion model for exemplar-based image colorization. IEEE Transactions on Circuits and Systems for Video Technology **35**(10), 10059–10069 (2025)
12. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proc. of European Conference on Computer Vision. pp. 179–195. Cham (2018)
13. Lei, C., Chen, Q.: Fully automatic video colorization with self-regularization and diversity. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 3748–3756 (2019)
14. Lei, C., Xing, Y., Ouyang, H., Chen, Q.: Deep video prior for video consistency and propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 356–371 (2023)
15. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. ACM Trans. Graph. (SIGGRAPH) **23**(3), 689–694 (2004)
16. Li, J., Zhao, H., Wang, Y., Lin, J.: Towards photorealistic video colorization via gated color-guided image diffusion models. In: Proc. of ACM International Conference Multimedia. pp. 10891–10900 (2024)
17. Liu, H., Xie, M., Xing, J., Li, C., Leung, C.S., Wong, T.T.: ColorDiffuser: Video colorization with pretrained text-to-image diffusion models. In: Proc. of ACM International Conference Multimedia. pp. 8891–8900 (2025)

18. Liu, Y.H., Zhao, H.Y., Chan, K.C., Wang, X.T., Loy, C.C., Qiao, Y., Dong, C.: Temporally consistent video colorization with deep feature propagation and self-regularization learning. Computational Visual Media **10**(2), 375–395 (2024)
19. Liu, Z., Wang, Y., Wei, T., Ma, C.: AnchorSync: Global consistency optimization for long video editing. In: Proc. of ACM International Conference Multimedia. pp. 4494–4503 (2025)
20. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 724–732 (2016)
21. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
22. Sheng, B., Sun, H., Magnor, M., Li, P.: Video colorization using parallel optimization in feature space. IEEE Transactions on Circuits and Systems for Video Technology **24**(3), 407–417 (2014)
23. Wu, C., Li, J., et al.: Qwen-Image technical report (2025), arXiv:2508.02324
24. Yang, Y., Pan, J., Peng, Z., Du, X., Tao, Z., Tang, J.: BiSTNet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5612–5624 (2024)
25. Zhang, B., He, M., Liao, J., Sander, P.V., Yuan, L., Bermak, A., Chen, D.: Deep exemplar-based video colorization. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 8044–8053 (2019)
26. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 586–595 (2018)
27. Zhao, Y.Z., Po, L.M., Yu, W.Y., Rehman, Y.A.U., Liu, M.Y., Zhang, Y.J., Ou, W.F.: VCGAN: Video colorization with hybrid generative adversarial network. IEEE Transactions on Multimedia **25**, 3017–3032 (2023)