Consistent Video Colorization via Palette Guidance

Han Wang, Yuang Zhang, Yuhong Zhang, Lingxiao Lu, Li Song

Shanghai Jiao Tong University

{esmuellert, zyayoung, rainbowow, lulingxiao, song_li}@sjtu.edu.cn

Abstract—**Colorization is a traditional computer vision task and it plays an important role in many time-consuming tasks, such as old film restoration. Existing methods suffer from unsaturated color and temporally inconsistency. In this paper, we propose a novel pipeline to overcome the challenges. We regard the colorization task as a generative task and introduce Stable Video Diffusion (SVD) as our base model. We design a palette-based color guider to assist the model in generating vivid and consistent colors. The color context introduced by the palette not only provides guidance for color generation, but also enhances the stability of the generated colors through a unified color context across multiple sequences. Experiments demonstrate that the proposed method can provide vivid and stable colors for videos, surpassing previous methods.**

Index Terms—**Video Colorization, Diffusion Models**

I. INTRODUCTION

Video colorization is essential for enhancing the visual experience of historical video materials and old films. This task involves transforming grayscale video sequences into vivid, full-color versions while maintaining temporal consistency across frames.

However, existing methods still face two primary challenges: **unsaturated colors** and **temporal inconsistency**. On one hand, the issue of **unsaturated colors** reflects a common challenge in image colorization, where colors often appear dull or lack diversity. To address this, researchers have successfully integrated generative models [1]–[4] and multi-modal priors [5], [6] into image colorization methods, achieving significant improvements in color vividness. For these methods, maintaining consistent colors across frames conflicts with the fact that each frame is colored independently. On the other hand, video colorization introduces the additional challenge of **temporal inconsistency**—the need to maintain consistent colors across frames. To tackle this, previous work [7]–[9] has employed techniques such as optical flow [10], [11]. While these methods process the video frame by frame, they struggle to achieve long-range coherence and are prone to cumulative errors, which further exacerbate the issue of temporal inconsistency. These limitations highlight the need for a more integrated approach that can balance color richness with temporal coherence, ensuring high visual quality in video colorization.

The rapid development of large-scale generative models has significantly advanced downstream vision tasks [12]–[14], including colorization [6], [15]–[20]. Among these, diffusion-based methods have become a cornerstone, using image-to-video diffusion priors to generate semantic reasonable colors. Recent studies, such as those in [18]–[21], have shown the
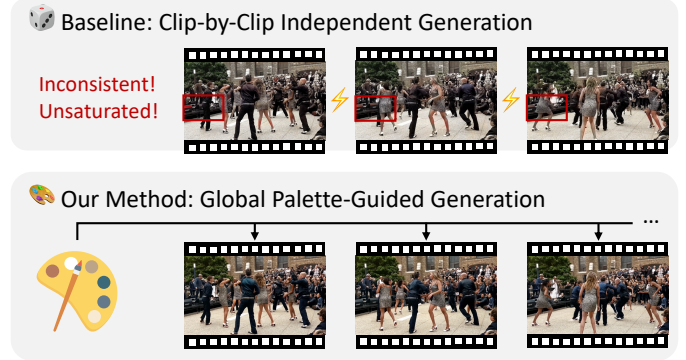

Fig. 1. Comparison with frame independent colorization framework TCVC [9]. Our method shows superiority in color vividness and temporal consistence.

potential of diffusion model (DM)-based approaches for video colorization. These methods address temporal consistency by integrating optical flow priors or cross-frame attention mechanisms into the image colorization backbone, enabling multi-frame colorization. For instance, methods like [20], [21] have introduced temporally deformable attention and cross-clip fusion to maintain long-term color consistency and prevent flickering or color shifts. However, these cross-frame attention mechanisms typically refer to only a limited number of adjacent frames, which may not be sufficient for ensuring long-range temporal consistency across the entire video. This limitation underscores the need for more robust and comprehensive approaches that can effectively address both color richness and temporal coherence in video colorization.

To address the challenges of video colorization, we propose a color palette-guided video diffusion framework, which enhances both color richness and temporal consistency by leveraging palette for global guidance and allowing for a diverse range of inputs for the palette. Our method is based on fine-tuning an image-to-video diffusion model. However, direct fine-tuning often results in generated frames that appear unsaturated and muted. We attribute this issue to two primary factors: (1) the model may fall into a conservative "shortcut" of merely recovering grayscale values, and (2) the output is highly sensitive to the color distribution of the training data, making it prone to biases.

To address these issues and obtain more saturated colors, we introduce a palette guidance mechanism. The color palette provides a rich color context that can significantly enhance the color saturation of the generated videos. Furthermore, we

identify that temporal consistency across denoising windows remains a significant challenge. Since previous models typically process only a limited number of frames simultaneously, maintaining globally consistent colors throughout longer video sequences becomes difficult. To address this, we utilize the palette functions as a global guide to ensure consistency across frames. Specifically, we process the global palette through a linear layer to obtain color embeddings that guide the denoising process. Besides, we found that simply using colors extracted from a single image as the palette condition may not always be suitable. This approach can sometimes lead to results that do not match reality. Our method, however, naturally allows for a wider variety of inputs to serve as the palette guidance. Therefore, we simply introduce more types of inputs. These include randomly sampling colors from a trained Mixture Model(GMM) [22] and using large language models like GPT [23] to extract colors based on the objects in the image. This helps to better guide the color generation of the image. In summary, the palette can effectively accommodate various types of inputs, converting different forms of colors into a unified domain condition.

Quantitative and qualitative experimental results indicate that our proposed automatic video colorization method outperforms the baseline methods in terms of color saturation and video quality. The innovations of our method are reflected in the following aspects:

1. We develop an integrated diffusion-based framework capable of simultaneously addressing color unsaturation and inter-frame color discontinuity issues.

2. We propose employing a palette as global guidance to resolve long-range instability problems in videos.

3. Our palette naturally accommodates various input formats, providing more solutions for color appropriateness.

II. RELATED WORKS

A. Image Colorization

Methods based on generative models, as opposed to those based on convolutional neural networks, offer richer colors and have thus become the predominant research direction in automatic image colorization. Both Generative Adversarial Networks (GANs) [1], [4], [24] and Transformers [2], [3], [17], [25] have made remarkable progress in this task. More recently, Latent Diffusion Models (LDMs) [26], as a superior alternative to GANs, have begun to attract attention in multimodal image colorization [6], [15], [16].

B. Video Colorization

Existing video colorization methods fall into two categories: 1) exemplar-based video colorization and 2) automatic video colorization.

Exemplar-based video colorization methods rely on a color exemplar image and propagate the colors to the video frames. Early methods [27]–[31] adopt networks to find the correspondence of gray frames and exemplar image in the deep feature domain. Then the colors are aligned according to the correspondence. Yang et al. [32] proposed a bidirectional

exemplar-based video colorization method to better propagate the reference colors and avoid the inaccurate matches. This category of methods requires the user to provide reference images that are highly relevant to the video content. In practical applications, frames from the video subsequent to the image colorization process may serve as exemplar. Nevertheless, methods relying on match propagation exhibit insufficient flexibility when confronted with the appearance of new objects.

Automatic video colorization methods operate independently of additional references and employ specialized modules to preserve consistency between frames. Previous GAN-based methods [7]–[9], [33], similar to the example-based approach, utilize optical flow [10], [11] to propagate motion information across frames. Even though these methods refer to inter-frame information, the accumulation of errors introduced by the optical flow methods causes color bleeding that impairs the visual effect. Recent DM-based methods [18]–[20] apply cross-frame attention mechanisms to the image colorization backbone to facilitate multi-frame colorization. Nevertheless, the image colorization backbone is not inherently capable of time-domain understanding, and cross-frame attention only refer to a limited number of adjacent frames.

C. Video Diffusion Models

With the significant success of diffusion models [26] in the field of image generation, diffusion-based video generation frameworks [34]–[41] have emerged. There are two main technical routes for video generation based on diffusion models: (1) adding a temporal layer on the basis of image generation diffusion models, and (2) training large-scale video diffusion models using massive data.

The representative work of Route 1 is AnimateDiff proposed by Guo et al. [34]. AnimateDiff extends the LDMs by incorporating a domain adapter, temporal transformer, and motion LoRAs [42], thereby adapting it for video generation tasks. Based on this flexibile architecture, motion-driven video generation has developed rapidly [43]–[46], and a large number of works based on motion modeling have also emerged [47]–[49]. However, although two motion-related modules were designed and video training was incorporated, the generation capability is constrained by the pre-trained image generation model, resulting in suboptimal dynamic generation, especially in long video scenarios.

The representative work of Route 2 is the large-scale trained video generation model. Stable Video Diffusion (SVD) [38], an open-source model with image-to-video generation capabilities, can generate 14 and 25 frame videos, providing high-quality video diffusion priors for downstream task research [21], [50]. DynamiCrafter [51] offers both text-to-video and image-to-video generation capabilities and provides a detailed analysis of the text-image control conditions.

III. METHOD

Our goal is to design a pipeline for automatic colorization of grayscale video sequence X^{gray} that can generate temporally
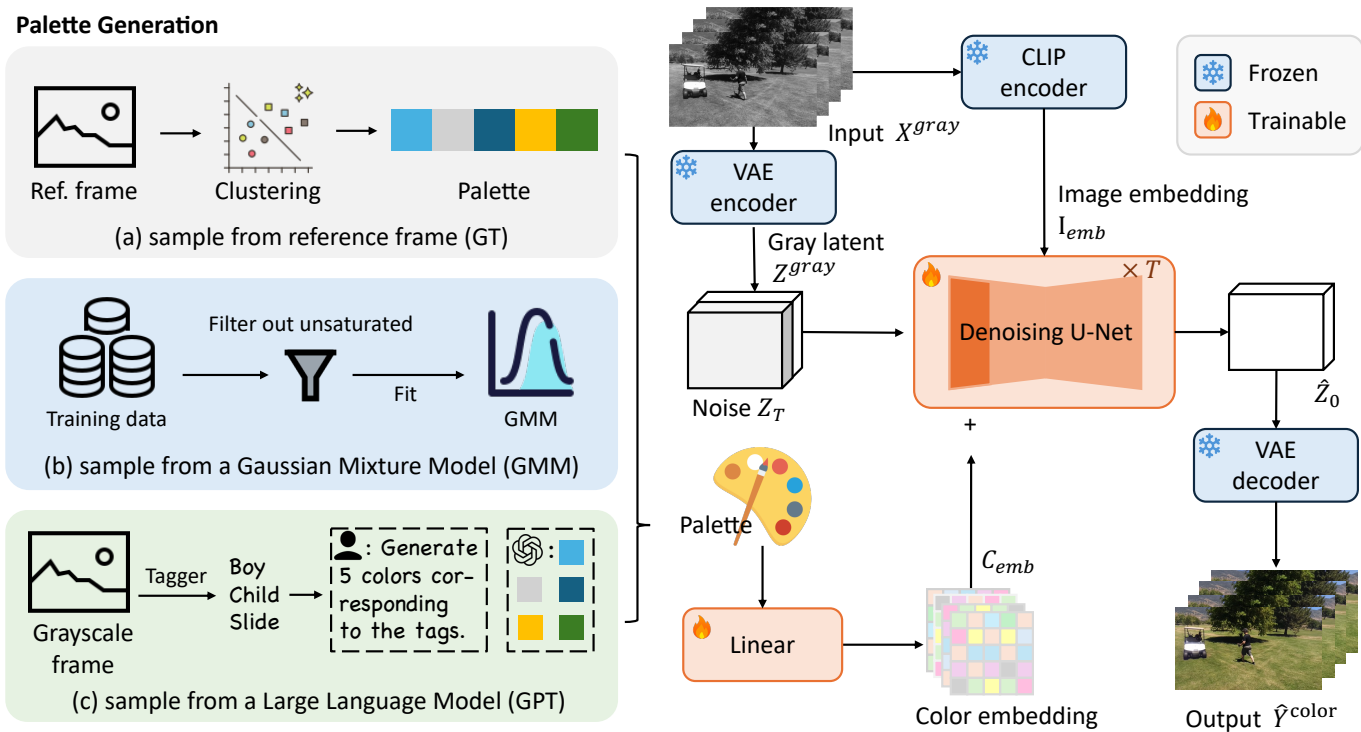
Fig. 2. Overview of the video colorization pipeline with palette-guidance. We re-purpose a pre-trained video diffusion model and augment it with palette guidance. The left side illustrates three methods for generating palettes. During training, the palette is generated in the manner depicted in (a), with the reference image being a randomly selected frame from each training clip.

consistent and color-saturated color sequences $\hat{Y}^{color}$. Firstly, in Section III-A, we discuss the model architecture, where we re-purpose the video generation model SVD [38] for the colorization task, enabling the model to generate videos guided by grayscale images. Once the model can achieve grayscale-controlled video generation, we introduce a global palette guidance to further refine the color control, supporting various flexible palette generation methods as described in Section III-B.

A. Video Diffusion Model for Video Colorization

Video diffusion models trained on large-scale image and video datasets exhibit excellent video quality and temporal consistency. We adopt the image-to-video model SVD as our backbone model. This choice is motivated by two aspects: (1) image-to-video models satisfy the requirements for grayscale fidelity in colorization tasks; (2) SVD is equipped with temporal modules in both the latent space and VAE, demonstrating stable performance across various downstream tasks. The framework is shown in Figure 2.

Specifically, during the training phase, we encode the grayscale input X^{gray} and ground truth color sequence Y^{color} to get the latent sequence Z^{gray} and Z^{color}. Subsequently, Z^{color} is subjected to the forward noise process to obtain the noisy latent code Z_t:

$$q(Z_t|Z^{color}) \sim \mathcal{N}(Z_t; \sqrt{\bar{\alpha}_t}Z^{color}, (1-\bar{\alpha}_t)\mathbf{I}) \quad (1)$$

where $\bar{\alpha}_t = \Pi_{i=1}^t \alpha_i$ and $\alpha_i = 1-\beta_i$. β_i represents the variance of the Gaussian noise added at the i-th diffusion step. The grayscale latent code Z^{gray} and Z_t are concatenated along the channel dimension to get the input Z_{in} for the denoising network:

$$Z_{in} = \{\text{concat}(z_{gray}^i, z_t^i)\}, \quad i = \{1, \cdots, N\} \quad (2)$$

where N denotes the input video length and t denotes the denoising timestep.

During the inference process, the input Z_{in} to the denoising U-Net consists of the encoded grayscale latent code Z^{gray} and the sampled Gaussian noise Z_T. Inspired by MimicMotion [50], we adopt the progressive approach for generating long videos with high temporal continuity. Specifically, we split the gray video sequence into segments with fixed length. Then, each segment is processed by the denoising U-Net. Finally, we obtain the color latent code by averaging the overlapped latent codes before sending it to the decoder.

B. Palette Guidance

Image-to-video models face two key challenges in video colorization. First, without explicit color guidance, they tend to produce conservative results with muted, unsaturated colors. Second, since these models typically process only a limited number of frames at a time, maintaining consistent colors throughout the video becomes difficult.

Based on the main challenges of automatic video colorization, color guidance should consider both color saturation

and cross-frame consistency. This requires color guidance to encourage saturated colors while maintaining the overall color style of the video, minimizing abrupt changes between frames.

As shown in Figure2, we design a global palette guide. The palette is composed of five distinct colors. Specifically, during the training process, we use the K-means clustering algorithm to categorize the colors within the reference frames into five palette colors. Subsequently, the palette vector $C_{palette} \in \mathbb{R}^{1 \times 15}$ constituted by these five colors is transformed through a straightforward color network composed of linear layers to obtain a color embedding $C_{emb} \in \mathbb{R}^{h \times w \times 320}$:

$$C_{emb} = W_{proj} C_{palette} \tag{3}$$

C_{emb} is aligned with the first-layer features $F^{(1)}$ of the denoising U-Net in the channel dimension. C_{emb} is then spatially broadcasted and fused with the first encoder layer:

$$\mathbf{F}^{(1)} \leftarrow \mathbf{F}^{(1)} + \Phi(\mathbf{C}_{emb}) \tag{4}$$

Where Φ denotes spatial replication of the color embedding across $H \times W$ positions. This integration ensures that color information is consistently incorporated into the feature representation of each sequence, thereby enhancing the performance of the denoising process by providing additional contextual cues related to palette.

During training, we fine-tune the denoising U-Net while keep the parameters of VAE and CLIP encoder frozen. The overall objective function is defined in a similar way to stable video diffusion:

$$\mathcal{L} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1), t, z_t, I_{emb}, C_{emb}}[||\epsilon - \epsilon_\theta(z_t; t, I_{emb}, C_{emb})||_2^2] \tag{5}$$

where ϵ is noise sampled from standard Gaussian. t denotes denoising timestep. I_{emb} represents the image embedding obtained by processing a randomly selected reference frame from the video through the CLIP image encoder [52].

C. Flexible Palette Generation

During the inference process, we provide various flexible methods for generating palette $C_{palette}$. For example, (a) by extracting from a colored reference image, (b) by randomly sampling the color space, or (c) by specifying colors provided by the user. Figure 2 illustrates 3 methods to get the palette.

Method (a) involves using K-means clustering on the provided colored reference image, following the same procedure as during training. The reference image can be any image or can be selected from frames processed by other colorization algorithms. We only extract the dominant colors from the reference image, and there are no strict requirements for the alignment between the reference image and the content of the video frames. In the practical application of colorizing old films, users can use images that reflect styles similar to the era of the current film as reference images for extracting color palettes. In this way, the model can adaptively align with the desired style.

Method (b), illustrated within the blue box, involves obtaining the palette through sampling from a trained GMM.

We leverage the training data to fit the GMM. Specifically, we randomly select 10 frames from each video in the training set. To avoid a preference of conservative colors, we remove pixels with an RGB value variance less than 50. We then fit the GMM using the filtered pixels. During inference, a palette can be obtained by randomly sampling colors from the GMM. If the user has no specific preferences, sampling the color palette in this way can facilitate rapid automatic colorization.

Method (c) involves the user directly specifying the colors of the palette. To reduce the user's workload, we propose the automated method illustrated in the green box. First, we use a tagging model [53] to extract content tags from the grayscale frame. Then, we employ a large language model [23] to generate colors corresponding to the extracted content. This approach can respond to objects within the video, offering more flexible and diverse color choices compared to sampling from a GMM.

A comprehensive analysis of various color palette generation approaches and their visual impacts will be presented in Section IV-D.

IV. EXPERIMENTS

We conduct three sets of experiments to evaluate our method: (1) comparison with state-of-the-art methods, (2) ablation studies, and (3) analysis of the palette-guidance capability.

A. Implementation

We leverage DAVIS2017 [54] benchmark to train and evaluate the proposed method. The training set contains 90 videos with an average length of 69 frames. During training, we resize the input video sequences to 1024×576 and set the learning rate to 10^{-5} with a linear warmup for the first 500 iterations. The pre-trained weights are from the stable video diffusion 1.1 image-to-video model.

We train the denoising U-Net and the linear color network while keeping the VAE and CLIP image encoder frozen.

B. Comparison with State-of-the-arts

We conduct qualitative comparison experiments with existing automantic colorization methods, including Deoldify [33] which is modified from an image colorization method, VCGAN [8] and TCVC [9]. The inference for the baseline methods employs the model weights and inference scripts released officially. The **visual comparison** is shown in Figure3. We select frames with a long temporal span for comparison to highlight the temporal consistency of the models. The frames from top to bottom correspond to the 1st, 21st, 41st, and 61st frames of the video.

In the simple scene (see the first video in Figure 3), Deoldify's results exhibit monotonous color in each individual frame, with insufficient contrast and saturation. Moreover, there is a significant color tone difference between frames that are farther apart, leading to a poor overall visual effect in the video. VCGAN shows extremely low color saturation, which can be considered a failure in colorization for this particular

Fig. 3. Visual comparison of colorization results for 2 videos. From left to right, the colorization results for Deoldify [33], VCGAN [8], TCVC [9], and ours.

Method	Colorful↑	FID↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓
Deoldify [33]	20.47	<u>54.28</u>	<u>648</u>	23.57	0.9937	0.1972
VCGAN [8]	14.46	74.52	889	22.85	0.9053	0.3090
TCVC [9]	<u>21.77</u>	71.25	776	**24.69**	<u>0.9973</u>	<u>0.1798</u>
Ours	**22.64**	**53.76**	**590**	<u>23.89</u>	**0.9993**	**0.1776**

case. TCVC performs relatively well but exhibits semantic errors in the second frame, where the person is not recognized and is assigned the same color as the surrounding rocks, indicating a lack of temporal stability. Our method provides stable, saturated, and semantically reasonable results in this case, with a natural overall visual effect.

In complex scene (see the second video in Figure 3), all methods are capable of coloring the green plants in the background. However, when it comes to coloring the foreground figures, Deoldify exhibits color bleeding at the junctions of different objects and fails to handle details such as the pendant on the woman's neck. VCGAN once again fails in the colorization task. TCVC, which outperforms Deoldify in simple scenes, encounters severe semantic errors in complex scenes. In the first frame, the skin tone of the foreground figure is incorrectly assigned green, although this issue is somewhat mitigated in subsequent frames, the colors of the background figures remain unsatisfactory. Moreover, due to color bleeding from the surrounding environment and the strange skin tone, the overall visual effect is poor. Additionally, TCVC suffers from the same problem as Deoldify in missing the color details of the pendant. These issues in baseline methods primarily stem from their reliance on GAN-based backbones, which struggle to generate diverse colors and fine-grained details. Furthermore, their single-frame processing approach lacks the capability for high-quality long-range perception, resulting in poor color consistency across frames.

Our method, on the contrary, provides vivid and reasonable colorization for both the foreground and background, with accurate handling of details. There is no noticeable color inconsistency between frames that are far apart, and the overall visual effect surpasses the compared methods. This superiority is attributed to our multi-frame processing framework, combined with a consistent global guidance mechanism, ensures high-quality long-range color consistency and temporal stability, addressing the limitations of baseline methods. Additionally, our colorization results do not rely on the ground truth color palette; the color palette used here is obtained through random sampling from the trained GMM.

To provide a comprehensive comparison between the proposed method and the baseline methods, we conduct **quantitative experiments**. We select the metrics widely used in recent years to evaluate image coloring tasks, the Colorful metric [55] and the Fréchet Inception Distance (FID) [56], to evaluate the color saturation and image quality of each frame, respectively. We also introduce the Fréchet Video Distance

(FVD) [57] to evaluate the video quality. Moreover, we also compare the commonly used metrics PSNR, SSIM and LPIPS [58], to provide a comprehensive evaluation.

TableI presents the results of our quantitative experiments. Our method demonstrates significant advantages, achieving the best quantitative results in all metrics except PSNR, where it is second only to TCVC. This aligns with the qualitative analysis in the previous section, indicating that our video colorization algorithm can provide semantically reasonable, saturated, and temporally stable color videos. The improvement in realism can be attributed to the introduction of an advanced diffusion model as the backbone, which enhances the generation of high-fidelity details. The enhancement in color saturation is primarily due to our innovative palette guidance mechanism, which ensures vibrant and visually appealing colorization. Furthermore, the overall video quality improvement stems from two key factors: (1) the integration of a video generation model for multi-frame processing, which ensures temporal consistency across frames, and (2) the consistent global guidance provided by our palette mechanism, which maintains coherence throughout the entire video sequence.

C. Ablation Studies

This section systematically evaluates two core innovations of our framework: (1) the necessity of global palette guidance and (2) the impact of different palette generation methods. We design four experimental configurations:

$\mathcal{A}$: Without palette guidance (retrained model)
$\mathcal{B}$: Random palette from trained GMM distribution
$\mathcal{C}$: LLM-assisted palette (GPT-4o as generator)
$\mathcal{D}$: Ground-truth first-frame extracted palette

All experiments maintain identical training configurations except for palette inputs. For fair comparison in Experiment $\mathcal{A}$, we completely remove the palette branch and retrain the model. Experiments $\mathcal{B} - \mathcal{D}$ preserve the original architecture but vary in palette sources: $\mathcal{B}$ samples colors from our trained Gaussian mixture model, $\mathcal{C}$ leverages GPT for semantic-aware color selection, while $\mathcal{D}$ adopts an oracle strategy using ground truth references.

As quantified in Table II, Experiment A demonstrates significant performance degradation across key metrics:
Colorfulness: 30.2% lower than average of $\mathcal{B} - \mathcal{D}$ ($\Delta = 8.36$).
Visual Quality (FID): 2.1 worse than average of $\mathcal{B} - \mathcal{D}$.
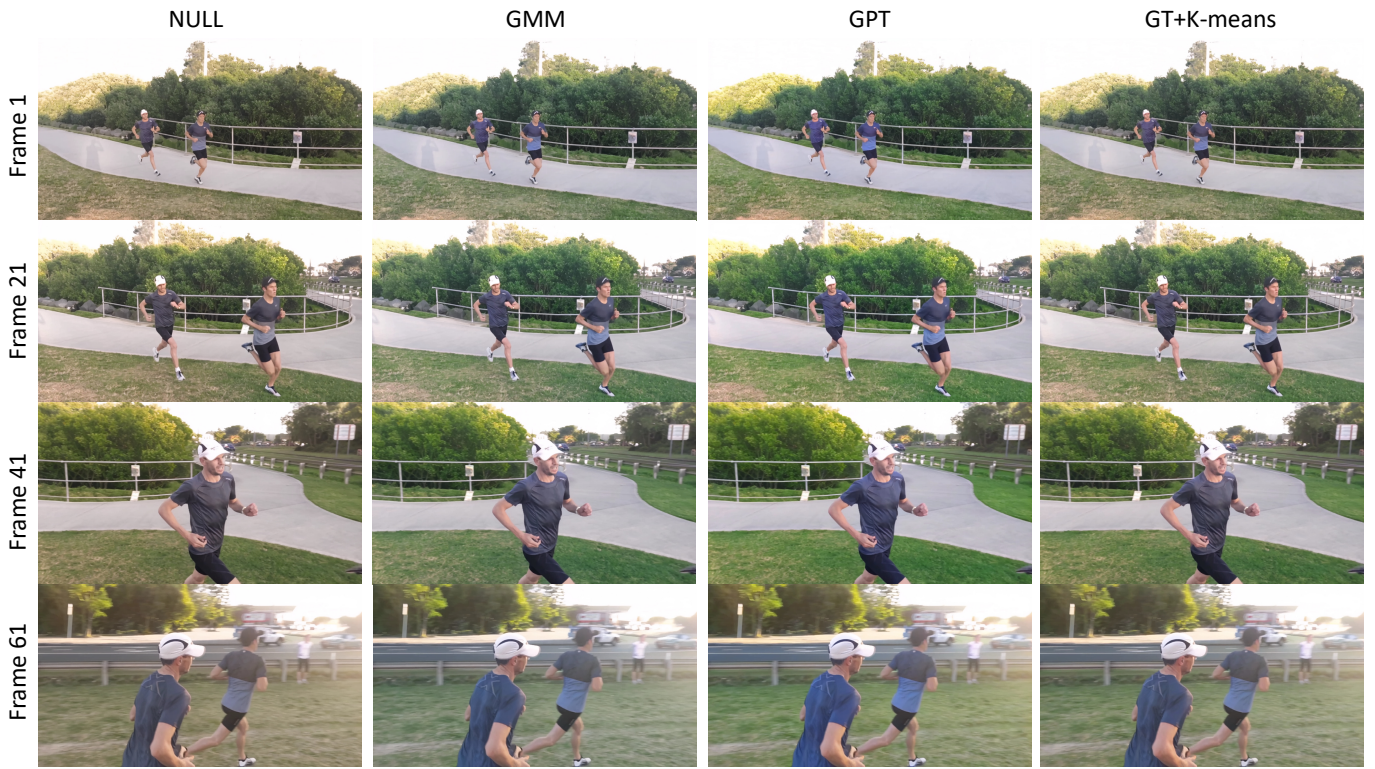**Temporal Consistency (FVD)**: 31.7 worse than average of $\mathcal{B} - \mathcal{D}$.

Fig. 4. Visual effects of ablation studies. The first column (NULL) shows the colorization results without palette guidance. The other columns display the colorization results of model variants with different palette generation methods.

Exp.	Palette	Colorful↑	FID↓	FVD↓
$\mathcal{A}$	NULL	19.34	54.63	628
$\mathcal{B}$	GMM	22.64	53.76	<u>590</u>
$\mathcal{C}$	GPT	**34.92**	<u>53.47</u>	637
$\mathcal{D}$	GT+K-means	<u>25.55</u>	**50.40**	**562**

This empirically confirms our hypothesis that palette guidance resolves ambiguity in ill-posed colorization tasks. The visual comparison in Figure 4 further reveals that $\mathcal{A}$ produces unsaturated results. In contrast, $\mathcal{B} - \mathcal{D}$ equipped with the palette guidance achieves saturated colors and better visual effects.

Our analysis of palette generation methods reveals distinct performance characteristics across key metrics: GMM-sampled palettes exhibit slightly lower color saturation (Colorfulness) compared to the other two approaches, while GPT-generated palettes achieve the highest color vibrancy. The GT-extracted palette demonstrates superior performance in both image quality (FID) and temporal consistency (FVD), indicating its effectiveness in maintaining photorealistic appearance and coherent video transitions. This result demonstrates that while GMM-based methods achieve better temporal consistency in automatic colorization tasks (as reflected in FVD metrics), LLM-driven approaches excel at creative color enhancement (evidenced by Colorfulness scores), whereas introducing color frame extraction for palette generation yields superior visual quality in both image (FID) and video (FVD) domains.

D. Flexibility of palette-guidance

We discuss the flexibility of palette guidance. Given that the colorization task is ill-posed, the color selection for many objects, particularly man-made ones, is not unique. There are various reasonable colors for the same image. We illustrate the colorization results corresponding to different palettes in Figure 5. We compare two different automatic palette generation methods, including random sampling from the trained GMM and GPT assisted generation, together with their corresponding colorization results. These two methods can automatically generate palettes without relying on external prompts. Focusing on man-made objects, the colors of the track in the left image and the toys as well as the child's clothing in the right image are significantly influenced by the palette. Even with substantial differences in palette colors, the generated images remain semantically reasonable. These cases demonstrate that our palette can achieve diverse and controllable colorization.

V. LIMITATION

While our method achieves effective global color style control by adaptively applying palette colors to video content,
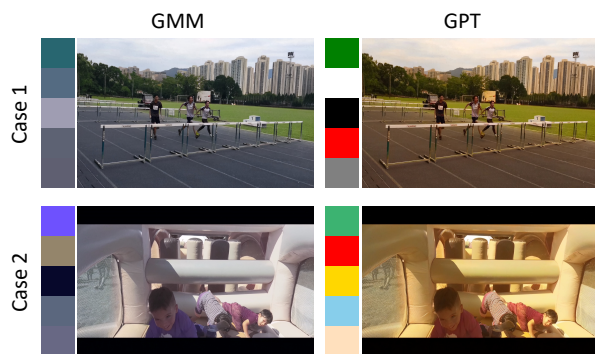
Fig. 5. Diverse colorization results with distinct palettes.

it currently lacks precise instance-level color manipulation capabilities. The adaptive palette propagation mechanism, though robust for holistic style transfer, cannot isolate and control colors for specific objects or regions (e.g., recoloring individual vehicles in traffic scenes or modifying clothing colors in human-centric videos). A promising extension of this work would involve integrating spatiotemporal object tracking and semantic-aware hint color guidance with our framework.

VI. Conclusion

In this paper, we propose an automatic video colorization method via palette guidance. To address the challenge of color inconsistency in video colorization, we introduce the image-to-video generation algorithm, Stable Video Diffusion, into the video colorization task. To tackle the challenge of color unsaturation, we design a global palette control. Our palette significantly enhances color saturation while maintaining video color consistency. Furthermore, our palette supports flexible generation methods, including automatic generation, user specification, and reference image extraction modes. Both quantitative and qualitative experimental results demonstrate that our proposed method can achieve highly saturated, high-quality, and consistent video colorization.

References

[1] Patricia Vitoria, Lara Raad, and Coloma Ballester, "Chromagan: Adversarial picture colorization with semantic class distribution," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2445–2454.

[2] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi, "Ct2: Colorization transformer via color tokens," in *ECCV*, 2022.

[3] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner, "Colorization transformer," in *International Conference on Learning Representations*, 2021.

[4] Geonung Kim, Kyoungkook Kang, Seongtae Kim, Hwayoon Lee, Sehoon Kim, Jonghyun Kim, Seung-Hwan Baek, and Sunghyun Cho, "Bigcolor: colorization using a generative color prior for natural images," in *European Conference on Computer Vision*. Springer, 2022, pp. 350–366.

[5] Zhitong Huang, Nanxuan Zhao, and Jing Liao, "Unicolor: A unified framework for multi-modal colorization with transformer," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–16, 2022.

[6] Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, Boxin Shi, et al., "L-cad: Language-based colorization with any-level descriptions using diffusion priors," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[7] Chenyang Lei and Qifeng Chen, "Fully automatic video colorization with self-regularization and diversity," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3753–3761.

[8] Yuzhi Zhao, Lai-Man Po, Wing-Yin Yu, Yasar Abbas Ur Rehman, Mengyang Liu, Yujia Zhang, and Weifeng Ou, "Vcgan: Video colorization with hybrid generative adversarial network," *IEEE Transactions on Multimedia*, vol. 25, pp. 3017–3032, 2023.

[9] Yihao Liu, Hengyuan Zhao, Kelvin CK Chan, Xintao Wang, Chen Change Loy, Yu Qiao, and Chao Dong, "Temporally consistent video colorization with deep feature propagation and self-regularization learning," *Computational Visual Media*, vol. 10, no. 2, pp. 375–395, 2024.

[10] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[11] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943.

[12] Yinhuai Wang, Jiwen Yu, and Jian Zhang, "Zero-shot image restoration using denoising diffusion null-space model," *The Eleventh International Conference on Learning Representations*, 2023.

[13] Tim Brooks, Aleksander Holynski, and Alexei A Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18392–18402.

[14] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool, "Diffir: Efficient diffusion model for image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13095–13105.

[15] Zhexin Liang, Zhaochen Li, Shangchen Zhou, Chongyi Li, and Chen Change Loy, "Control color: Multimodal diffusion-based interactive image colorization," *arXiv preprint arXiv:2402.10855*, 2024.

[16] Xiaoyan Cong, Yue Wu, Qifeng Chen, and Chenyang Lei, "Automatic controllable colorization via imagination," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2609–2619.

[17] Xiangcheng Du, Zhao Zhou, Xingjiao Wu, Yanlong Wang, Zhuoyao Wang, Yingbin Zheng, and Cheng Jin, "Multicolor: Image colorization by learning from multiple color spaces," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 6784–6792.

[18] Hanyuan Liu, Minshan Xie, Jinbo Xing, Chengze Li, and Tien-Tsin Wong, "Video colorization with pre-trained text-to-image diffusion models," *arXiv preprint arXiv:2306.01732*, 2023.

[19] Jiaxing Li, Hongbo Zhao, Yijun Wang, and Jianxin Lin, "Towards photorealistic video colorization via gated color-guided image diffusion models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10891–10900.

[20] Vukasin Bozic, Abdelaziz Djelouah, Yang Zhang, Radu Timofte, Markus Gross, and Christopher Schroers, "Versatile vision foundation model for image and video colorization," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[21] Zhitong Huang, Mohan Zhang, and Jing Liao, "Lvcd: reference-based lineart video colorization with diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 6, pp. 1–11, 2024.

[22] Douglas A Reynolds et al., "Gaussian mixture models.," *Encyclopedia of biometrics*, vol. 741, no. 659-663, 2009.

[23] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[24] Xin Jin, Zhonglan Li, Ke Liu, Dongqing Zou, Xiaodong Li, Xingfan Zhu, Ziyin Zhou, Qilong Sun, and Qingyu Liu, "Focusing on persons: Colorizing old images learning from modern historical movies," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1176–1184.

[25] Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie, "Ddcolor: Towards photo-realistic image colorization via dual decoders," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 328–338.

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffu-

sion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[27] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen, "Deep exemplar-based video colorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8052–8061.

[28] Satoshi Iizuka and Edgar Simo-Serra, "Deepremaster: temporal source-reference attention networks for comprehensive video enhancement," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–13, 2019.

[29] Duong Thanh Tran, Nguyen Doan Hieu Nguyen, Trung Thanh Pham, Phuong-Nam Tran, Thuy-Duong Thi Vu, Cuong Tuan Nguyen, Hanh Dang-Ngoc, and Duc Ngoc Minh Dang, "Swintexco: Exemplar-based video colorization using swin transformer," *Expert Systems with Applications*, vol. 260, pp. 125437, 2025.

[30] Siqi Chen, Xueming Li, Xianlin Zhang, Mingdao Wang, Yu Zhang, Jiatong Han, and Yue Zhang, "Exemplar-based video colorization with long-term spatiotemporal dependency," *Knowledge-Based Systems*, vol. 284, pp. 111240, 2024.

[31] Yixin Yang, Jiangxin Dong, Jinhui Tang, and Jinshan Pan, "Colormnet: A memory-based deep spatial-temporal feature propagation network for video colorization," in *European Conference on Computer Vision*. Springer, 2025, pp. 336–352.

[32] Yixin Yang, Jinshan Pan, Zhongzheng Peng, Xiaoyu Du, Zhulin Tao, and Jinhui Tang, "Bistnet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[33] Jason Antic, "DeOldify: A Deep Learning based project for colorizing and restoring old images (and video!)," .

[34] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.

[35] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai, "Sparsectrl: Adding sparse controls to text-to-video diffusion models," in *European Conference on Computer Vision*. Springer, 2025, pp. 330–348.

[36] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[37] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al., "Videocrafter1: Open diffusion models for high-quality video generation," *arXiv preprint arXiv:2310.19512*, 2023.

[38] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al., "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[39] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7310–7320.

[40] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al., "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.

[41] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al., "Hunyuanvideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.

[42] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[43] Li Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8153–8163.

[44] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan, "Dreamvideo: Composing your dream videos with customized subject and motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6537–6549.

[45] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu, "Champ: Controllable and consistent human image animation with 3d parametric guidance," in *European Conference on Computer Vision*. Springer, 2025, pp. 145–162.

[46] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo, "Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions," in *European Conference on Computer Vision*. Springer, 2025, pp. 244–260.

[47] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan, "Motionctrl: A unified and flexible motion controller for video generation," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[48] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al., "Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[49] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou, "Motiondirector: Motion customization of text-to-video diffusion models," in *European Conference on Computer Vision*. Springer, 2025, pp. 273–290.

[50] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou, "Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance," *arXiv preprint arXiv:2406.19680*, 2024.

[51] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong, "Dynamicrafter: Animating open-domain images with video diffusion priors," in *European Conference on Computer Vision*. Springer, 2025, pp. 399–417.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[53] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al., "Recognize anything: A strong image tagging model," *arXiv preprint arXiv:2306.03514*, 2023.

[54] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv preprint arXiv:1704.00675*, 2017.

[55] David Hasler and Sabine E Suesstrunk, "Measuring colorfulness in natural images," in *Human vision and electronic imaging VIII*. SPIE, 2003, vol. 5007, pp. 87–95.

[56] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[57] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly, "Fvd: A new metric for video generation," in *DGS@ICLR*, 2019.

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.