CoolColor: Text-guided COherent OLd film COLORization

Zichuan Huang
Wangxuan Institute of Computer Technology, Peking University
CN
huangzichuan@hotmail.com

Yifan Li
Wangxuan Institute of Computer Technology, Peking University
CN
2100012520@stu.pku.edu.cn

Shuai Yang*
Wangxuan Institute of Computer Technology, Peking University
CN
williamyang@pku.edu.cn

Jiaying Liu
Wangxuan Institute of Computer Technology, Peking University
CN
liujiaying@pku.edu.cn

Figure 1: Our *CoolColor* produces plausible and coherent colorized videos with different text guidance precisely.

Abstract

With the increasing demand for movie-watching, the significance of colorizing classic black-and-white films is gradually growing. In this paper, we introduce a text-guided old film colorization method that uses natural language descriptions to guide the process, offering precise control over colorization. Focusing on maintaining video consistency, we address the unique challenges throughout old film colorization, such as color flicker and motion blur. We employ a data augmentation strategy to enhance the robustness and stability of the model against motion across frames. Additionally, we implement a training-free sampling strategy to enhance correlation and reduce instability through successive frames. Moreover, we utilize a post-processing strategy to maintain the structural integrity of the original frames. Extensive experimental results demonstrate that our method could provide a realistic and controlled solution to old film colorization, enhancing the viewing experience for audiences.

CCS Concepts

• **Computing methodologies** → **Reconstruction**; **Image processing**; **Computational photography**.

Keywords

Video colorization, video consistency, text guidance, diffusion model

ACM Reference Format:
Zichuan Huang, Yifan Li, Shuai Yang, and Jiaying Liu. 2024. CoolColor: Text-guided COherent OLd film COLORization. In *ACM Multimedia Asia (MMASIA '24), December 03–06, 2024, Auckland, New Zealand.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3696409.3700173

1 Introduction

As a classical narrative formation of art, the film industry has experienced dramatic development since its outset, with color being a pivotal element. While rich in storytelling, those classic movies, captured in black and white, have witnessed technological limitations on hues nowadays. Restoring the colors of these monochrome

*Corresponding author

MMASIA '24, December 03–06, 2024, Auckland, New Zealand

ACM ISBN 979-8-4007-1273-9/24/12
https://doi.org/10.1145/3696409.3700173

films is of high artistic value, aiming to bridge the gap between the present and the past and to enhance the viewing experience.

The challenge of colorizing old movies lies in maintaining a delicate balance between preserving the fidelity of the original content and introducing colors authentically. Traditional automatic colorization methods [29, 37] suffer from uncertainty and less-controllability when facing various scenes, leading to distortion of the directors' intentions. Meanwhile, automatic colorization methods always have limitations in terms of under-saturation, over-fitting to specific datasets, and lack of general colorization ability, especially for historical old-film scenes. Therefore, we introduce a text-guided COherent OLd film COLORization method (**CoolColor**), using flexible and efficient natural language descriptions to guide the diffusion model [9, 24] with more detailed and accurate colorization constraints. Thus, our method enjoys higher saturation visual representation and enhances user controllability compared with previous automatic colorization methods.

Different from image colorization, **old film colorization** needs to pay more attention to video consistency except for the quality of a single frame. As the colorization problem is ill-posed, diverse reasonable color candidates may cause severe jitters across consecutive frames in a video. Besides, most old films have complex degradation patterns, such as motion blur, noise, and compression artifacts, leading to a more challenging task to colorize an old movie. Thus, we propose a novel training scheme, called the progressive equivariant training scheme, alleviating color overflow and artifacts. We observe that even a little motion disturbance may cause large variations across consecutive frames. To reduce such unstable affections, we propose to construct pairs of origin images and warped images with a series of common geometric transformations, such as rotation, shearing, and scaling. Then, the model is constrained to maintain the same results after alignment on such pairs. To stabilize the training process, we further propose a scale factor to control such equivariant optimization in a progressive way.

Furthermore, we propose a training-free frame correlation-aware sampling strategy, to reduce the interference of unstable factors such as color-flickering and motion-blurring via inter-frame collaboration without extra training. Specifically, we introduce additional frame-correlated connections between adjacent frames during the inference at the origin self-attention modules. We utilize rich semantic features of keyframes as anchors and conduct a cross-frame attention mechanism to align consecutive frames in a batch.

Finally, we introduce a luminance-aware post-process strategy, which maintains the brightness of pixels unchanged when adding vivid color information to the grayscale images, preserving the structural consistency between the original movie scene and colorized results. Such a simple yet effective post-process strategy can enhance the integration and harmony between generated colors and original grayscale values, creating a visual experience that respects original aesthetics while introducing creative new colors.

Extensive experiments have demonstrated our old-film colorization framework's effectiveness compared to the state-of-the-art baselines. Our contributions could be summarized as follows:

- We propose a simple yet effective text-guided old film colorization framework CoolColor, which fully utilizes the strong generative prior of pre-trained diffusion models, and adapts it to videos with a progressive equivariant training scheme and a keyframe correlation-aware sampling strategy.
- We propose a new progressive equivariant training scheme to reduce the sensitivity of the model to input, which steadily improves its robustness against flickering.
- To improve video consistency, we present a training-free keyframe correlation-aware sampling strategy to enhance the overall visual quality and coherence with the guidance of the best-colorized keyframe.

2 Related Works

2.1 Image Colorization

With the development of deep neural network, Cheng *et al.* [7] propose the first deep-based colorization method, and Zhang *et al.* [37] propose to optimize colorization as a classification problem in quantized *CIELab* color space, producing more colorful results. InstColor [25] utilizes detection boxes to reduce color overflow and color incompletion. BigColor [13] utilizes the pre-trained generative prior of BigGAN [3] to colorize a grayscale image, while ChormaGAN [27] and DeOldify [1] directly optimize a GAN from scratch. However, GAN-based colorization methods are still limited by unpleasant artifacts due to unstable training.

ColTran [15] proposes the first transformer-based [26] colorization method, which builds a probability model and incorporates a multi-stage colorization strategy. Inspired by the color cross-entropy loss proposed by CIC [37], CT2 [29] also considers colorization as a classification problem, and feeds image patches and color tokens together into a ViT-based network. Given the strong ability of query-based vision transformer, DDColor [11] presents a color decoder, which merges the grayscale structure and the generated color information with the cross-attention mechanism effectively.

2.2 Video Colorization

Compared with image colorization, video colorization not only aims to colorize each frame vividly, but also need to maintain the overall coherency. An intuitive idea is to apply image colorization methods frame-by-frame, and optionally utilize post-processing techniques to promote temporal consistency [16], which rely on optical-flow estimation and mapping.

Different from optical-flow-based methods that concentrate on detecting motion through pixel variations, RNN-based methods [34, 35] pay more attention to general sequence learning, which can include various types of temporal pattern recognition beyond motion. Chen *et al.* [6] propose a 3D-convolution-based video colorization method, which directly applies 3D convolutions to a stack of consecutive video frames to capture temporal consistency. However, these methods always struggle to achieve an optimal balance between model complexity and performance quality.

2.3 Diffusion Models for Colorization

Recently, the pre-trained text-to-image diffusion models have been increasingly employed to tackle a wide range of low-level vision tasks [12, 18, 21, 28]. As for image colorization, Liu *et al.* [20] propose a piggybacked diffusion model and a shortcut between the VAE encoder and decoder to maintain structural consistency. Liang *et al.* [19] build a multi-modal colorization framework based on

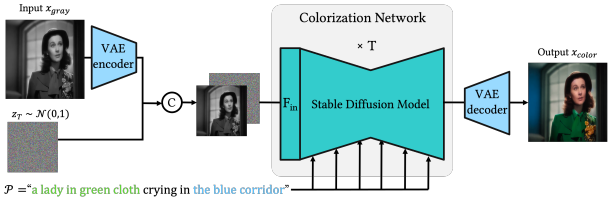

Figure 2: Overview of the proposed CoolColor framework.

ControlNet [36], which is an efficient and effective architecture to finetune a diffusion model. L-CAD [5] creates new blocks to insert grayscale structure information into denoising U-Net and proposes a luminance-aware image compression module to maintain the basic structure of the colorized image and reduce color overflow. However, this method is also limited to a time-consuming diffusion sampling process and low result resolution.

Compared with the above methods, our method not only fully utilizes the generative potential of diffusion models, but also is capable of high-efficiency training and inference, and decreases color artifacts and color overflow. Our method can also produce more consistent video-colorized results.

3 CoolColor Method

3.1 Preliminaries

DDPM. Diffusion models [9, 24] learn to represent natural image distribution with a forward process and a backward process. During the forward process, Gaussian noises are gradually added to the clean image x_0, producing noisy images $x_t, t \in [1, T]$,

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

where t is the time step, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and α_t are a set of hyper-parameters, ϵ is a random sampled standard Gaussian noise. Note that x_T can be approximately treated as a pure Gaussian noise, the backward process utilizes a neural network ϵ_θ parameterized by θ, to recover a clean image x_0 from a random sampled Gaussian noise x_T by performing denoising process iteratively following

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t z, \tag{2}$$

where z is a randomly sampled Gaussian noise, σ_t is a hyper-parameter to control the intensity of noises. Such a method to generate new images is called Denoising Diffusion Probabilistic Models (*DDPM*).

DDIM. In practice, T is always set to 1,000, and timesteps cannot be skipped because z in Eq. (2) is non-negligible. Therefore, generating images from a Gaussian noise within 1,000 steps of denoising is time-consuming. Luckily, Song *et al.* [24] presents an efficient and effective way to sample real images within as few as 20 timesteps, called Denoising Diffusion Implicit Models (*DDIM*). This method adjusts DDPM denoising in Eq. (2) to

$$\begin{aligned} x_{t-1} =& \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}\right) \\ &+ \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(x_t, t) + \sigma_t z, \end{aligned} \tag{3}$$

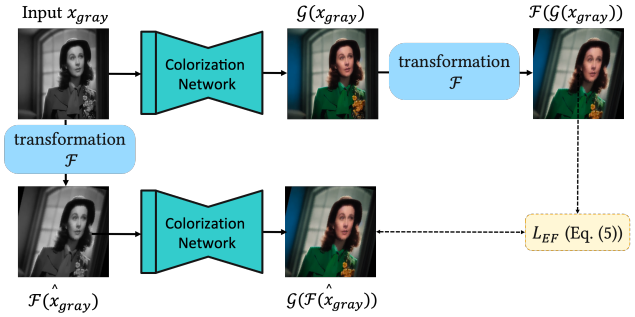

Figure 3: Illustration of the equivariant training.

which can be deterministic when σ_t is set to 0, enabling skipping timesteps to save time cost while keeping image quality.

Latent Diffusion Models. The requirements for high-resolution image generation bring new challenges: directly applying diffusion models to 512 × 512 image resolution brings too much computational cost. To balance the trade-off between image quality and memory/time cost, Latent Diffusion Models (LDM) [23] proposes to perform the above diffusion process in the latent space constructed by Variational AutoEncoder (VAE), rather than in image space, to save both training and inference cost. In specific, the VAE Encoder compresses a 512 × 512 × 3 image to a 64 × 64 × 4 latent code, and the VAE Decoder is trained to reconstruct the original image from the latent code precisely. In our work, we utilize pre-trained model weights of Latent Diffusion Models [2] to empower strong generative priors on old film colorization tasks.

3.2 Coherent Old Film Colorization Framework

Inspired by the superior performance of conditional diffusion models on image generation and image editing, we construct our CoolColor framework based on a pre-trained Stable Diffusion model, as illustrated in Fig. 2. Given a grayscale old film sequence $v_{gray} \in \mathbf{R}^{b\times H\times W\times 1}$ where b denotes the number of frames in a sequence, and a text prompt $\mathcal{P}$ that describes scene and concrete color requirements of input video, our colorization framework can produce a vivid and plausible colorized video results $v_{color} \in \mathbf{R}^{b\times H\times W\times 3}$. Specifically, we first utilize VAE to extract the latent variable $z_{gray} \in \mathbf{R}^{b\times H/8\times W/8\times 4}$ of v_{gray}. In order to extend the model's functionality to include image colorization tasks, we create an additional convolutional layer F_{in} at the beginning of the U-Net (ϵ_θ in Sec. 3.1) to merge gray latent variables z_{gray} and noisy latents z_t during the diffusion process mentioned in Sec.3.1, leading to a merged input z_{merge}. Such operation can be formulated by

$$z_{merge} = F_{in}(z_t, z_{gray}). \tag{4}$$

To harness the inherent exceptional capabilities of text-to-image generation of the Stable Diffusion model, we commence by initializing our model's weights with a checkpoint from the pre-trained v1.5 model. To maintain the stability of training, we initialize F_{in} to zero, which will be updated smoothly during optimization. Such simple merge strategy based on convolution blocks can not only introduce grayscale conditions into the generation process effectively, but also maximally preserve the original generative ability of Stable Diffusion, ensuring realism and fidelity simultaneously.

3.3 Progressive Equivariant Training Scheme

We experimentally find that even small motions or changes of an object in consecutive frames will lead to significant flickering during the colorization process. The networks tend to amplify subtle pixel variations and hinder their video applications. Inspired by the flicker suppression loss [31, 33], we propose an equivariant training scheme for consistent video colorization.

Specifically, as shown in Fig. 3, for a grayscale image x_{gray} from the training set, we create a warped image $\hat{x}_{gray}$ by random geometric transformations including rotation, translation, scaling, shearing, and resized crop. We denote such transformation by $\mathcal{F}$. $\mathcal{F}$ augments a single grayscale image to a pair of consecutive frames, which are expected to be colorized in the same manner. Thus, we colorize origin grayscale image x_{gray} and the warped grayscale frame $\hat{x}_{gray} = \mathcal{F}(x_{gray})$, resulting corresponding approximate colorized results, and reduce their L2 distance after alignment, which can be formulated as:

$$L_{EF}(x_{gray}) = \mathbf{E}_{t\sim[0,T]}||\mathcal{G}(\mathcal{F}(x_{gray}), t) - \mathcal{F}(\mathcal{G}(x_{gray}, t))||_2, \quad (5)$$

where we denote the mapping which maps input to colorized results as $\mathcal{G}$, t is a randomly sampled timestep and T is a hyper-parameter which we set to 1,000. To construct $\mathcal{G}$ that maps a grayscale input to an approximate colorized result, we seek power based on DDIM sampling. Thanks to the elegant deterministic characteristic of DDIM sampling, we can predict a denoised clean image $\hat{x}_0$ at any timestep t by

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\alpha_t}\cdot\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}}. \quad (6)$$

Therefore, $\mathcal{G}(x_{gray}, t) = \hat{x}_{color}$ is a combination of DDPM forward process (Eq. (1)) and DDIM sampling (Eq. (6)). Specifically, we first use Eq. (1) to obtain x_t from $x_0 = x_{gray}$, then use Eq. (6) to compute an approximate colorized result $\hat{x}_{color}$.

Formally, our training objective is

$$\begin{aligned} L &= \lambda \cdot L_{EF}(x_{gray}) + L_{recon}, \\ L_{recon} &= \mathbf{E}_{t\sim[0,T],\epsilon,\mathcal{P}}||\epsilon - \epsilon_\theta(x_t, t, \mathcal{P})||_2, \end{aligned} \quad (7)$$

where L_{recon} is the reconstruction loss to train Diffusion models, λ is a hyperparameter to control the intensity of equivariant loss.

We experimentally find that increasing λ progressively can improve performance effectively. As we will show later in Sec. (4.3), our methods after equivariant training experienced significant enhancements in the robustness of coherent video generation.

3.4 Frame Correlation-aware Sampling Strategy

The poor video consistency by frame-by-frame colorization severely damages the visual experience of users. Inspired by the cross-frame attention mechanism proposed by Tune-A-Video [32], we propose a novel frame correlation-aware sampling strategy as shown in Fig. 4, which effectively combines with our diffusion-based framework and does not need any additional training, fine-tuning, or optimization. Cross-frame attention is a simple but effective method that leverages the inherent relationships between consecutive frames to enhance overall video consistency performance. The core principle of cross-frame attention is to dynamically allocate computational resources based on the relevance of each frame to the task at hand. The main difference lies in that origin cross-frame attention used

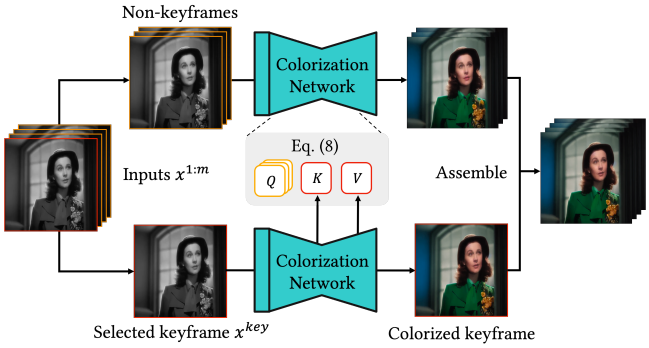

Figure 4: Illustration of the proposed frame correlation-aware sampling strategy.

by [32] simply utilizes the first frame or the previous frame to guide the generation process, while our method flexibly selects key frames that are well-colorized as anchors.

Specifically, we first select a keyframe denoted by x^{key} and replace self-attention blocks with our cross-frame attention blocks at the last two blocks of the U-Net decoder. Each attention layer receives m-frame inputs: $x^{1:m} = [x^1, \cdots, x^m] \in \mathrm{R}^{m\times h\times w\times c}$. Hence, the linear projection layers produce m queries, keys, and values, denoted by $Q^{1:m}$, $K^{1:m}$, and $V^{1:m}$ respectively. During the self-attention operation of each frame in a batch, we replace their original *Keys* and *Values* by the feature map of keyframe denoted by K^{key} and V^{key}, as shown in Eq. (8), for $k = 1, 2, \cdots, m$. By utilizing cross-frame attention, the appearance and structure of the objects and background as well as identities are carried over from the keyframe to subsequent frames, significantly increasing the temporal consistency of the generated frames.

$$\mathrm{Attn_{CF}}(\mathrm{Q^k}, \mathrm{K^{1:m}}, \mathrm{V^{1:m}}) = \mathrm{softmax}\left(\frac{\mathrm{Q^k}(\mathrm{K^{key}})^\mathrm{T}}{\sqrt{\mathrm{d}}}\right)\mathrm{V^{key}} \quad (8)$$

As for the keyframe selection, we first colorize the whole grayscale video clip frame-by-frame and measure the overall quality of each frame quantitatively with CLIP score and colorfulness metric. We will introduce those metrics in 4.1. Then we choose the best frame as the keyframe to achieve a coherent and high-quality video colorization. Moreover, to reduce time costs and further cater to diverse users' appetites, we also provide an option for users to choose their favorite frames as the keyframes. To obtain the best results, we adopt the second strategy.

3.5 Luminance-aware Post-process Strategy

To further maintain a stable structure consistency between grayscale inputs and colorized results, we additionally bring in a luminance-replaced strategy as a post-processing strategy for old-film frame colorization. Specifically, we first transform the image into *CIELab* color space. Considering the luminance channel contains the structure information, while the chrominance channels carry the color information, we combine the luminance (L) channel of grayscale inputs and the chrominance (ab) channels of colorized results. To put it more straightforwardly, we replace the luminance channel of the colorization result with that from grayscale inputs. This results in the final colorized outputs that retain both the original structure and details information and the vivid color information produced

Figure 5: Qualitative comparison between *CoolColor* and other text-guided colorization methods. Our colorized videos enjoy higher stability (indicated by yellow boxes) and more plausible visual effects. Zoom for better visualization.

by our CoolColor method, ensuring the colorization appears natural and harmonious with the original image.

4 Experiments

4.1 Implementation Details

Training. We use a subset of SA-1B dataset [14] as our training data. We first filter out under-saturated images using colorfulness metric [8], then utilize BLIP [17] to obtain captions as text guidance during training, and further filter out images that hold no color words in the caption. We finally got about 220,000 images for training. We use the pre-trained Stable Diffusion v1.5 checkpoint as the starting point, and then train our CoolColor with a single NVIDIA A40 GPU for 8k iterations with a batch size of 32. We use AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate of 10^{-4}. We linearly increases λ from 0 to 0.001.

Evaluation. We select video clips from the classic old movies *Stagecoach* (1939), *Waterloo Bridge* (1940), *Roman Holiday* (1953), and *The Sound of Music* (1965) for validation. The original *Stagecoach*, *Waterloo Bridge*, and *Roman Holiday* are monochrome, with random degradation (such as flickering, compression artifacts, or noises) throughout the frames, and can be colorized directly. The original *The Sound of Music* is chromatic, as a reference of colorization, and can be colorized after a grayscale preprocessing.

We adopt CLIP Score[22] to measure the color accuracy, which calculates cosine similarity between the CLIP features of prompts and colorization results, and Colorfulness [8] to measure the overall colorization vividness. For video colorization, we mainly utilize the preference rates of the user study as the evaluation metrics.

4.2 Comparison

Qualitative Comparison. As shown in Fig. 5, we make comparisons with state-of-the-art text-guided colorization methods L-CoDe [30] and L-CoDer [4]. As for single-frame colorization, our CoolColor has obvious advantages in colorizing old movie scenes with both realism and vividness. L-CoDe generates a red-biased color tone throughout the given scenes, while L-CoDer exhibits inflexibility and inaccuracy in response to the changes in input text prompts. Significant color overflow appears in the results of both L-CoDe and L-CoDer, which is effectively suppressed in our method. As for the whole video, our method can produce more consistent results than L-CoDe and L-CoDer. Specifically, our CoolColor has strong uniformity in the regions whose colors have been specified in the text prompt (such as skin complexion and clothing) and good coherence in the image details that could not be covered by the text (such as tie and jewelry) in the meanwhile. Besides, our method also performs well on prompts that indirectly indicate colors (such as the green colors of the rock implied by the word *vegetated*).

Table 1: Quantitative comparison and user preference rates.

Method	Objective Metrics		User Study			
	Colorfulness↑	CLIP Score↑	Accuracy	Stability	Realism	Overall
DeOldify	25.63	-	10.6%	3.5%	8.8%	14.1%
HistoryNet	18.51	-	1.2%	7.1%	4.1%	5.3%
L-CoDe	**58.96**	23.95	0.0%	0.0%	0.0%	0.0%
L-CoDer	26.02	24.61	1.2%	3.5%	2.4%	3.5%
Ours	38.53	**25.50**	**87.1%**	**85.9%**	**84.7%**	**77.1%**

Quantitative Comparison. As shown in Table 1, we make comparisons with two unconditional methods, DeOldify[1] and HistoryNet[10], to demonstrate our superior visual effect, and two text-guided methods, L-CoDe and L-CoDer, to justify the flexibility and accuracy in text control of our method. Our method achieves the highest CLIP Score and second-best Colorfulness under the users' autonomous selecting strategy of the keyframes, indicating our method can not only ensure that the generated colors comply with text control but also balance the relatively high saturation and visual naturalness. When we choose keyframe automatic selection strategies with higher colorfulness and CLIP Score preferences respectively, our method performs better numerically than the given manual selection strategy, namely 72.01 for colorfulness and

(a) Without L_{EF} .

(b) With progressive L_{EF} .

(c) With constant L_{EF} .

Figure 6: Effect of the proposed progressive equivariant training scheme.

(a) Without Cross-frame Attention.

(b) With Cross-frame Attention.

Figure 7: Effect of the proposed frame correlation-aware sampling strategy.

27.32 for CLIP Score. By comparison, although L-CoDe achieves the highest Colorfulness, it suffers from over-saturation as also qualitatively verified by the red-biased color tone in Fig. 5.

User Study. Since different users have different aesthetic requirements, we further conducted a user study to show the superior performance of our method. We invite 17 users to evaluate 5 video clips. In each question, participants are asked to measure several dimensions and select the *best*-colorized result among several compared methods. We define best based on the following four aspects: (1) consistency with the text descriptions (*Accuracy*); (2) uniformity of the colors throughout the frames (*Stability*); (3) realism of the frames independent from the given text prompt (*Realism*); (4) comprehensive personal preference (*Overall*). The statistics are summarized in Table 1, which shows our method outperforms all the other comparison methods.

4.3 Ablation Study

We provide qualitative ablations for the progressive strategy of our equivariant fine-tuning loss in Fig. 6 and the inter-frame stabilization effect of the cross-frame attention blocks in Fig. 7.

Progressive equivariant training scheme. As mentioned in Sec. 3.3, we stabilize the training of the model while improving the model's robustness against flickering by adjusting the hyperparameter λ of equivariant fine-tuning loss. We train three colorization networks with different λ growth modes to study the impact of L_{EF}. In the adjacent three frames shown in Fig. 6, there is a noticeable flicker in the color of the man's tie and the passerby's clothes when $\lambda = 0$, which means without L_{EF}, even very subtle movement of the objects will lead to severe color inconsistency. Although this phenomenon is alleviated when $\lambda = 0.001$, it still yields unstable facial colors of the passerby. When λ linearly increases to 0.001, which is the final strategy we use, these flickers are well suppressed.

Frame correlation-aware sampling strategy. We remove the cross-frame attention layers to study their role in maintaining inter-frame coherency and stability. The changes in object brightness caused by movements can lead to differences in grayscale values, and further affect the color stability in adjacent frames. In Fig. 7, the woman's white clothes, the man's tie, and the background building's color exhibit different color distributions in adjacent frames. By using the best colorization result as the keyframe (indicated by the red frame border), the scenery details in the remaining frames can be better colorized coherently, further demonstrating the effectiveness of our proposed frame correlation-aware sampling strategy.

5 Conclusion and Discussion

In this paper, we propose a novel diffusion-based framework for colorizing old film frames using natural language descriptions. Our CoolColor ensures temporal consistency and superior video synthesis through anchor-based cross-frame attention and equivariant training. Extensive experiments show the superiority of our method in accuracy, realism, and controllability.

Limitations and Future Work. Although our method generates plausible and coherent colorized videos, the frame correlation-aware sampling struggles with large motion variations due to its reliance on intrinsic feature similarity from Stable Diffusion. In future work, we will explore more effective ways to vividly colorize long videos with diverse motions.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62332010 and 62471009, in part by CCF-Tencent Rhino-Bird Open Research Fund, and in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

References

[1] Jason Antic. 2019. DeOldify. https://github.com/jantic/DeOldify.

[2] Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. 2022. Latent-diffusion. https://github.com/CompVis/latent-diffusion.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv preprint arXiv:1809.11096* (2019).

[4] Zheng Chang, Shuchen Weng, Yu Li, Si Li, and Boxin Shi. 2022. L-CoDer: Language-based Colorization with Color-object Decoupling Transformer. In *Proc. European Conf. Computer Vision.*

[5] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. 2023. L-CAD: Language-based Colorization with Any-level Descriptions using Diffusion Priors. In *Advances in Neural Information Processing Systems.*

[6] Siqi Chen, Xueming Li, Xianlin Zhang, Mingdao Wang, Yu Zhang, Jiatong Han, and Yue Zhang. 2024. Exemplar-based Video Colorization with Long-term Spatiotemporal Dependency. *Knowledge-Based Systems* 284 (2024), 111240.

[7] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. 2015. Deep Colorization. In *Proc. Int'l Conf. Computer Vision.*

[8] David Hasler and Sabine E Suesstrunk. 2003. Measuring Colorfulness in Natural Images. In *Human vision and electronic imaging VIII.*

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Pobabilistic Models. In *Advances in Neural Information Processing Systems.*

[10] Xin Jin, Zhonglan Li, Ke Liu, Dongqing Zou, Xiaodong Li, Xingfan Zhu, Ziyin Zhou, Qilong Sun, and Qingyu Liu. 2021. Focusing on Persons: Colorizing Old Images Learning from Modern Historical Movies. In *Proc. ACM Int'l Conf. Multimedia.*

[11] Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. 2023. Ddcolor: Towards Photo-realistic Image Colorization via Dual Decoders. In *Proc. Int'l Conf. Computer Vision.*

[12] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. 2022. Denoising Diffusion Restoration Models. In *Advances in Neural Information Processing Systems.*

[13] Geonung Kim, Kyoungkook Kang, Seongtae Kim, Hwayoon Lee, Sehoon Kim, Jonghyun Kim, Seung-Hwan Baek, and Sunghyun Cho. 2022. Bigcolor: Colorization Using a Generative Color Prior For Natural Images. In *Proc. European Conf. Computer Vision.*

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. In *Proc. Int'l Conf. Computer Vision.*

[15] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. 2021. Colorization Transformer. In *Proc. Int'l Conf. Learning Representations.*

[16] Chenyang Lei, Yazhou Xing, Hao Ouyang, and Qifeng Chen. 2022. Deep Video Prior for Video Consistency and Propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2022), 356–371.

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proc. IEEE Int'l Conf. Machine Learning.*

[18] Yifan Li, Yuhang Bai, Shuai Yang, and Jiaying Liu. 2024. COCO-LC: Colorfulness Controllable Language-based Colorization. In *Proc. ACM Int'l Conf. Multimedia.*

[19] Zhexin Liang, Zhaochen Li, Shangchen Zhou, Chongyi Li, and Chen Change Loy. 2024. Control Color: Multimodal Diffusion-based Interactive Image Colorization. *arXiv preprint arXiv:2402.10855* (2024).

[20] Hanyuan Liu, Jinbo Xing, Minshan Xie, Chengze Li, and Tien-Tsin Wong. 2023. Improved Diffusion-based Image Colorization via Piggybacked Models. *arXiv preprint arXiv:2304.11105* (2023).

[21] Ozan Özdenizci and Robert Legenstein. 2023. Restoring Vision in Adverse Weather Conditions with Patch-Based Denoising Diffusion Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2023), 10346–10357.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proc. IEEE Int'l Conf. Machine Learning.*

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution Image Synthesis With Latent Diffusion Models. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition.*

[24] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *Proc. Int'l Conf. Learning Representations.*

[25] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. 2020. Instance-aware Image Colorization. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition.*

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems.*

[27] Patricia Vitoria, Lara Raad, and Coloma Ballester. 2020. ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution. In *Proc. IEEE Winter Conf. Applications of Computer Vision.*

[28] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin Chan, and Chen Change Loy. 2024. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *Int'l Journal of Computer Vision* (2024).

[29] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. 2022. CT 2: Colorization Transformer via Color Tokens. In *Proc. European Conf. Computer Vision.*

[30] Shuchen Weng, Hao Wu, Zheng Chang, Jiajun Tang, Si Li, and Boxin Shi. 2022. L-CoDe: Language-based Colorization Using Color-object Decoupled Conditions. In *Proc. AAAI Conference of Artifical Intelligence.*

[31] Bichen Wu, Ching-Yao Chuang, Xiaoyan Wang, Yichen Jia, Kapil Krishnakumar, Tong Xiao, Feng Liang, Licheng Yu, and Peter Vajda. 2024. Fairy: Fast Parallelized Instruction-Guided Video-to-Video Synthesis. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition.*

[32] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-A-Video: One-shot Tuning of Image Diffusion Models for Text-to-video Generation. In *Proc. Int'l Conf. Computer Vision.*

[33] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. 2022. VToonify: Controllable High-Resolution Portrait Video Style Transfer. *ACM Transactions on Graphics* 41 (2022), 1–15.

[34] Yixin Yang, Jinshan Pan, Zhongzheng Peng, Xiaoyu Du, Zhulin Tao, and Jinhui Tang. 2024. Bistnet: Semantic Image Prior Guided Bidirectional Temporal Feature Fusion for Deep Exemplar-based Video Colorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46 (2024), 5612–5624.

[35] Bo Zhang, Mingming He, Jing Liao, Pedro Sander, Lu Yuan, Amine Bermak, and Dong Chen. 2019. Deep Exemplar-based Video Colorization. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition.*

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proc. Int'l Conf. Computer Vision.*

[37] Richard Zhang, Phillip Isola, and Alexei Efros. 2016. Colorful Image Colorization. In *Proc. European Conf. Computer Vision.*