

TOWARDS COHERENT VIDEO COLORIZATION: WHEN OPTICAL FLOW MEETS IMAGE DIFFUSION MODELS

Wen Si Yifan Li Luyao Zhang Shuai Yang* Jiaying Liu

Wangxuan Institute of Computer Technology, Peking University, Beijing, China



Fig. 1. Our model creates colorful and consistent video colorizations, which aligns with the veracity of reference frames.

ABSTRACT

Video colorization aims to generate semantically appropriate and temporally consistent colors for grayscale videos. Existing diffusion-based approaches often face challenges in high training resources or poor cross-frame consistency due to their reliance on stochastic denoising processes. To address these limitations, this paper proposes an optical flow-guided diffusion framework that explicitly utilizes inter-frame motion to enhance temporal coherence. We explicitly inject optical flow into the network during training and introduce an optical flow-guided cross-frame attention mechanism, which leverages precise optical flow correspondences to modulate attention between frames during sampling. This ensures stable and artifact-free color propagation under complex motion patterns. Moreover, our framework incorporates reference images as additional guidance to improve controllability and color fidelity. To reduce dependence on large-scale real-world video data, we design a hybrid data synthesis pipeline that first pre-trains the model using synthesized optical flow-image pairs and then fine-tunes it on a limited set of high-quality videos generated by modern video synthesis models. Experiments demonstrate that our method achieves superior performance in terms of training efficiency, temporal consistency, color realism, and user controllability compared to existing state-of-the-art approaches.

Index Terms— Video colorization, video consistency, optical flow, diffusion model

*Corresponding Author

1. INTRODUCTION

Video colorization aims at bring plausible colors back to grayscale videos and maintain temporal coherence, which enjoys wide applications in recoloring historical films, advertising and artistic creation. With the development of deep learning [1], early attempts utilize temporal convolutions [2], bidirectional feature fusion [3], or generative adversarial networks (GANs) [4,5] to learn the mapping from gray videos to their colorful counterparts with collected paired data. However, limited by data scale and expensive training cost, these approaches that learn gray-color mapping from scratch struggle to achieve stable cross-frame consistency and visually realistic color generation, producing obvious flickering, abrupt color changes, and irregular object hues.

Recently, diffusion models [6, 7] have demonstrated its superior generative ability in both image generation [8] and video editing [9]. In pursuit of taming this powerful generation backbone to video colorization, L-C4 [10] extends image diffusion models [8] to video domain by training additional temporal consistency modules and auxiliary inference-time optimization strategy. Although effective, L-C4 requires costly training data and resources, *i.e.*, six V100 GPUs and millions of video clips, making it impractical to be applied on resource-constraint scenarios efficiently. Furthermore, the inherent stochastic nature of diffusion models' denoising iteration is also harmful to cross-frame inconsistency [11].

Instead of extending image diffusion models to video domain, Color-Turbo [11] innovatively converts multi-step denoising models to one-step generator, maintaining original

generative prior and naturally mitigate unstable issues caused by stochastic noises. Therefore, such one step generation facilitates an intuitive per-frame colorization scheme, establishing a promising direction for video colorization. However, Color-Turbo heavily relies on image diffusion models to **implicitly** model the cross-frame relationship, which is challenge and unreliable under large and complex motion patterns. Since it lacks of video-specific designs to explicitly characterize temporal sequences, Color-Turbo sometimes produces color flickering and unrealistic artifacts on videos.

Based on the above analysis, we identify that the primary challenges of the video colorization task are **threefold**: (1) Generating semantically **appropriate and aesthetically convincing colors** given text prompts or reference images as conditions; (2) Maintaining **temporal coherence** across all frames, preventing visual artifacts such as flickering, abrupt color changes, or inconsistent object hues. (3) Providing **versatile controllability** for fine-grained color generation, in terms of not only hues, but also saturation and brightness conditioned on user demands.

To address these problems, we propose a novel optical flow-guided video colorization model that built upon image diffusion models [8], where optical flow serves as the explicit and precise representation of inter-frame correlation for temporal consistency, while the cross-modal generative prior of Stable Diffusion (SD) [8] is fully utilized for realistic and visually plausible color generation.

We utilize the optical flow in two aspects: In the training phase, we inject it through a simple auxiliary convolution layer at the very beginning of the network, inducing minimal computational overhead. In the inference phase, we exploit optical flows as a cross-frame binding ref-colored frame relationship in a training-free manner. Specifically, we propose an advanced Cross-Frame Attention (CFA) sampling strategy based on optical flows, namely Optical Flow-guided CFA (OFCFA). Although simple CFA [12] has demonstrated its effectiveness on maintaining video coherency, it is still challenge under complex motions and grayscale videos: original CFA heavily depends on the inherent ability of SD to match the similar patches, which is unstable, and sometimes wrong especially on the color-corrupted inputs, leading to color flickering, overflows and artifacts. To establish precise cross-frame correspondence, we exploit the optical flows as the explicit guidance to modulate the attention of CFA, further enhancing the inter-frame consistency and intra-frame detail preservation under the guidance of a reference frame.

Besides, we inject reference images as additional information to the network, which further improves user controllability and temporal stability. The fixed reference frames poses as a stable anchor with rich and user-required color guidance, potentially improving the coherency of video colorization results, and allowing more versatility for user requirements via providing a desired frame, as shown in Fig. 1. By synergistically integrating reference color images and optical flows

as external guidance, we are capable to leverage powerful diffusion prior and find precise correspondence between gray images and the reference, thereby achieving **superior motion coherency, color realism and versatile user controllability**.

Furthermore, unlike previous works [2, 10] that require large amounts of real-world video data for training, we propose an innovative **hybrid data synthesis strategy, eliminating the expensive requirements of real-world video data**. We reformulate the training as a two-stage procedure. In the first stage, we propose a novel data synthesis pipeline to create paired data, pre-training the image diffusion model to learn the basic integration of image and optical flows. We utilize existing large-scale image datasets and synthesize random optical flows as model inputs. In the second stage, we utilize modern video generation models (*i.e.*, Wan2.2 [9]) as the data engine to create a small set of high-quality colorful videos, fine-tuning the pre-trained model on more realistic and consistent optical flow patterns for advanced generalization on real-world scenes. Extensive experiments show the superior performance of our method.

The rest of the paper is organized as follows. Section 2 introduces our proposed framework with optical-guided diffusion-based colorization model. Details of our experimental results are shown in Section 3 and concluding remarks are given in Section 4.

2. METHOD

2.1. Motivations and Overview

Color-Turbo introduces a single-step image colorization framework based on LoRA-finetuned [13] Stable Diffusion model. By incorporating the skip connections between VAE encoder and decoder, and the one-step distillation without involving stochastic denoising iteration, it achieves superior detail preservation, stability, and efficiency and provides a **promising direction** for video colorization.

However, there still remains some essential challenges when applies Color-Turbo directly to video colorization in a per-frame manner, since it lacks of video-oriented designs for temporal consistency. Inspired by this, we build our framework based on the Color-Turbo, which contains two critical components: 1) Inter-frame consistency-aware colorization network, which involves optical flow as external inputs to improve temporal consistency, and 2) optical flow-guided cross-frame attention (OFCFA) mechanism, which leverage precise cross-frame correspondence to modulate temporal attention weighting. We will provide more details as following.

2.2. Inter-Frame Consistency-Aware Framework

Inspired by the superior generation performance of image diffusion models, we construct our framework based on a pre-trained Stable Diffusion model, as illustrated in Fig. 2. Given a grayscale video sequence $v_{in}^{gray} \in \mathbb{R}^{b \times H \times W \times 1}$, the corresponding optical flow $v_{fwd}, v_{bwd} \in \mathbb{R}^{b \times H \times W \times 2}$, and a text

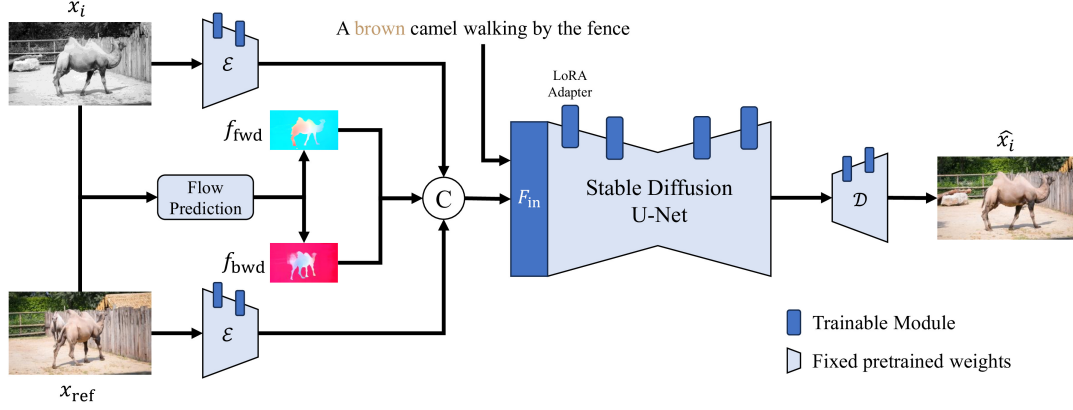


Fig. 2. Illustration of our proposed inter-frame consistency-aware framework. Our $\mathcal{E}$ and $\mathcal{D}$ refer to VAE encoders and decoder, x_i and x_{ref} denote the input gray frame and colored reference frame, f_{fwd} and f_{bwd} refer to optical flow of both directions (their visualizations are illustrated), and $\hat{x}_i$ denotes the colorized frame.

prompt $\mathcal{P}$ that describes scene and user-demanded color requirements of the desired colorized video, our colorization framework produces consistent and realistic colorized video results $v_{color} \in \mathbb{R}^{b \times H \times W \times 3}$.

In detail, we choose the median frame $x_{mid} \in \mathbb{R}^{H \times W \times 3}$ (specifically the $\lceil b/2 \rceil$ -th frame) of v_{in}^{gray} as the anchor, and utilize optical flow estimation models (e.g., GMFlow [14] and SEA-RAFT [15]) to calculate forward flow $f_{fwd} \in \mathbb{R}^{H \times W \times 2}$ and backward flow $f_{bwd} \in \mathbb{R}^{H \times W \times 2}$, which are then used as the input sequences v_{fwd}, v_{bwd} . In addition, x_{mid} is copied for each frame x_i in v_{in} and used as $v_{ref} \in \mathbb{R}^{b \times H \times W \times 3}$.

Specifically, we first utilize separate VAE encoders to extract the latent variables $z_{in}, z_{ref} \in \mathbb{R}^{b \times H/8 \times W/8 \times 4}$ of v_{in} and v_{ref} , respectively. v_{fwd}, v_{bwd} are then down-sampled to z_{fwd} and z_{bwd} to align spatially with the latents. In order to enhance the model’s functionality for colorization and optical flow understanding, in the diffusion process, we create an additional fusion layer F_{in} at the beginning of the U-Net to integrate all inputs, and obtain merged latent z_{merge} , the full operation can be formulated by

$$z_{merge} = F_{in}(z_{in}, z_{ref}, z_{fwd}, z_{bwd}). \quad (1)$$

To harness the inherent exceptional capabilities of text-to-image generation of the Stable Diffusion model, and maintain rich colorization performance, we initialize our model with a pre-trained Color-Turbo [11] checkpoint. To maintain the stability of training, we initialize the new weights in F_{in} corresponding to $z_{ref}, z_{fwd}, z_{bwd}$ to zero, which will be updated smoothly during optimization.

2.3. Optical Flow-guided Cross-Frame Attention

To maintain temporal coherence and reduce flickering, we propose an optical flow-guided cross-frame attention (OFCFA) mechanism. Simple cross-frame attention utilized in [12] allows every pixel to look around all other pixels globally, which often leads to “content drift” and ghosting artifacts,

even worse with large motion videos. In contrast, our approach leverages optical flow to explicitly model the trajectory of objects across the temporal axis.

For each feature map at resolution $h \times w$, we compute a sparse attention mask $\mathcal{M} \in \mathbb{R}^{(hw) \times (hw)}$ to regularize the computation of original cross-frame attention [12]. Intuitively, the mask $\mathcal{M}$ serves as the graph matrix that indicates the connected/isolated state of each two nodes in feature maps, which is initialized to small values as $-1e9$. If optical flow matches two consistent positions, then these two feature nodes are connected, and vice versa. Formally, let $\mathbf{p} = (i, j)$ denote the spatial coordinate in the current frame feature map, where $i \in \{1, 2, \dots, h\}, j \in \{1, 2, \dots, w\}$. The corresponding coordinate $\mathbf{p}'$ in the reference frame x_{ref} is obtained by adding the optical flow offset $\mathbf{d}_{\mathbf{p}}$:

$$\mathbf{p}' = \mathbf{p} + \mathbf{d}_{\mathbf{p}}, \quad (2)$$

where $\mathbf{d}_{\mathbf{p}} = (u_p, v_p)$ corresponds to the offset vector at position $\mathbf{p}$. In practice, we utilize absolute offsets where $\hat{u}_p, \hat{v}_p \in [0, 1]$ and $u_p = h \cdot \hat{u}_p, v_p = w \cdot \hat{v}_p$.

After obtaining this per-pixel mapping from current feature map $\mathbf{p} = (i, j)$ to reference feature map $\mathbf{p}'$, we can change the $(i \cdot w + j)$ -th row of the attention mask $\mathcal{M}$ from $-1e9$ to 0. In practice, to account for flow estimation inaccuracies and local deformations, we define a localized neighborhood $\mathcal{N} \in \mathbb{R}^{3 \times 3}$ around the warped coordinates:

$$\mathcal{M}[(i + u_p) \cdot w + (j + v_p), \quad i' \cdot w + j'] = 0, \quad (3)$$

where $(i', j') \sim \mathcal{N}$.

This mask is then utilized to modulate the original cross-frame attention weights. Specifically, we penalize the attention weights of irrelevant regions indicated by optical flows, and boost the weights for those within the flow-predicted neighborhood. We apply this mask to the softmax layer in the cross-frame attention block, ensuring that the hidden states of frame x_t are conditioned on the most contextually relevant

features from frame x_{ref} . Formally, let Q , K , and V denote the query (from current frame), key, and value (from reference frame) features, respectively. The optical flow-guided attention is computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + \mathcal{M}\right) \cdot V, \quad (4)$$

where d_k is the dimension of the key features.

By leveraging optical flow as the explicit inter-frame correlation, our proposed OFCFA effectively regularizes the model to preserve identity and texture along the estimated motion paths, significantly improving the stability.

2.4. Hybrid Training Scheme

There are inevitable domain gaps to handle optical flow conditions for SD, which is predominately tailored to process images. To effectively help the model integrate optical flow information, and enhance performance and realistic appliance, we design a two-stage hybrid training scheme.

Pretraining Stage: Random Optical Flow Synthesis. In order to scale up training data for fully utilization of the motion information emebded in optical flows, we propose a paired data synthesis pipeline with non-sequenced image data in the training stage. For an image $x \in \mathbb{R}^{H \times W \times 3}$, we get its luminance channel x^{gray} and create random optical flow $f_f \in \mathbb{R}^{H \times W \times 2}$. We then warp x according to the synthesized optical flow, obtaining image $x' = \text{warp}(x, f_f)$. After this, we calculate the backward flow $f_b \in \mathbb{R}^{H \times W \times 2}$ from x' to x . Finally, $[x^{\text{gray}}, x', f_f, f_b]$ performs a set of valid training data; we repeat this process to create a sufficient amount of random optical flow data. To ensure the diversity of training data, we generate our data on mass image datasets such as ImageNet [16] and MS-COCO [17].

Fine-tuning Stage: Virtual Video Data Generation. To further optimize the model's performance on realistic motion patterns, and maintain its text alignment properties, we utilize video generation models to generate high-quality video data based on detailed prompts. In this way we distill the temporal coherence ability from video generation models, and achieve better temporal consistency.

3. EXPERIMENTS

3.1. Implementation Details

Training. We utilize BLIP [18] and BLIP-2 [19] to generate a caption for each video sequence. In the pre-training stage, we utilize random optical flow data created on a subset of ImageNet [16]. In the fine-tuning stage, we generate our training data with video generation model Wan2.2-A14B [9] with detailed prompts and high-quality requirements. To ensure sufficient background motion (*e.g.*, camera motion) along with foreground motion, we crop the frames with a moving window of fixed size, which moves in a random direction and reflects when colliding with the edge of the frame. We train each stages for 10,000 iterations on an NVIDIA 4090 GPU.

Table 1. Quantitative results on DAVIS30 evaluation benchmark. $\uparrow$ ($\downarrow$) means the higher (lower) the better. Best performances are highlighted in **bold** and second best in *italic*.

Method	CDC $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
VCGAN	5.208	24.26	0.927	0.235
TCVC	3.855	25.04	0.922	0.224
BiSTNet	3.950	24.88	0.928	0.216
L-CoDe	4.410	24.97	0.861	0.219
L-CoDer	4.208	25.31	0.865	0.203
L-CoIns	3.946	23.04	0.871	0.236
L-CAD	4.407	22.70	0.886	0.212
I2VEdit	<i>3.423</i>	<i>25.59</i>	<i>0.935</i>	<i>0.207</i>
Ours	3.303	26.35	0.950	<i>0.207</i>

Testing. We test our model on DAVIS-val [20] and the old film dataset proposed by [12]. For DAVIS validation videos, we utilize BLIP-2 to generate captions for the colored ground-truth reference frame, with human adjustment to ensure compatibility. We then apply the caption to all frames in the video as colorization prompts. For historical film data, we use Nano-Banana [21] to obtain high-quality reference frames with refined detail generated from prompts provided by humans, then the prompts are input likewise. We adopt CDC [22] to evaluate the cross-frame colorization consistency, and measure PSNR [23], SSIM [24] and LPIPS [25] to evaluate visual quality and faithfulness to the given reference frame. To ensure a fair comparison, all metrics are measured upon the resolution of 512×512 .

Luminance-aware Post-process Strategy. For further structure consistency, we adopt a luminance-replaced strategy for post-processing. Specifically, we first transform the image into *CIELab* color space, and replace the luminance channel of the colorization results with the ones of the input grayscale frames, which aligns structural details more precisely.

3.2. Comparisons with SoTA

Quantitative Comparison. To assess our method, we compare our model with 3 automatic colorization approaches: VCGAN [5] and TCVC [22] and BiSTNet [3]; 4 language-guided colorization methods: L-CoDe [26], L-CoDer [27], L-CoIns [28] and L-CAD [29]; and 1 reference-guided video editing method I2VEdit [30], which we adapt to our colorization setting. The same luminance-aware post-processing strategy is used across compared methods. The results are shown in Table 1.

Our method achieves the best performance across all evaluation metrics, significantly outperforming both automatic and language-guided colorization baselines. Obtaining the highest PSNR, SSIMscore and second highest LPIPS, our model presents superior structural preservation and reconstruction accuracy. Compared to the recent reference-guided method I2VEdit, our model shows a notable improvement in color consistency, achieving the best CDC score, which suggests a more stable and natural color distribution across

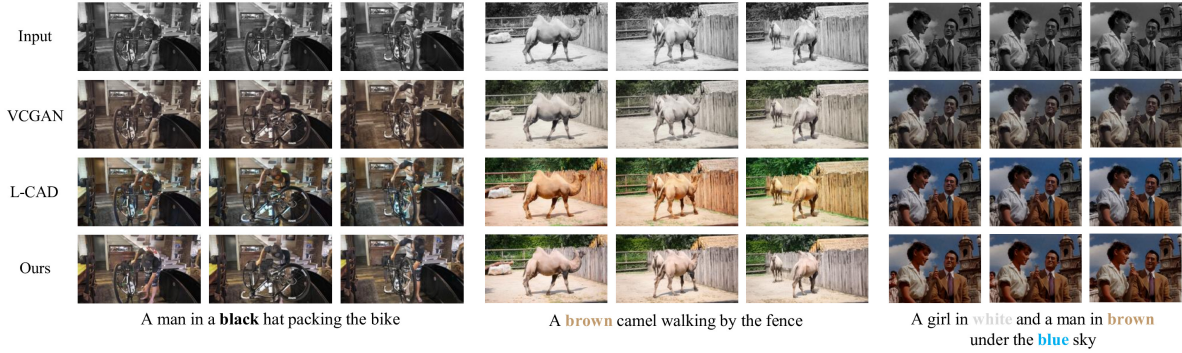


Fig. 3. Qualitative Comparison between our method and other colorization methods. Our colorized videos enjoy superior stability and faithfulness to given reference. Best viewed in color.

frames. These results validate the effectiveness of our proposed architecture in handling complex colorization tasks while maintaining high fidelity and temporal coherence.

Qualitative Comparison. As shown in Fig. 3, our method achieves better inter-frame color consistency, along with colorful and faithful reconstruction; on the left, the color of man’s hat and the background remain consistent in spite of movement; in the middle, the color of the camel and wooden fences are more stable and rich compared to other methods. In addition, our model enables better detail preservation, accurately colorizing the person in the right background, even without mentioning the object in the prompt.

3.3. Ablation Study

We conduct our ablation study starting from the baseline model, using the original U-Net from Color-Turbo [11], which only takes grayscale image latent input, output results are evaluated without performing OFCFA. Then we add components of our pipeline sequentially: first OFCFA, then a reference frame x_{ref} , followed by forward flow f_{fwd} and backward flow f_{bwd} .

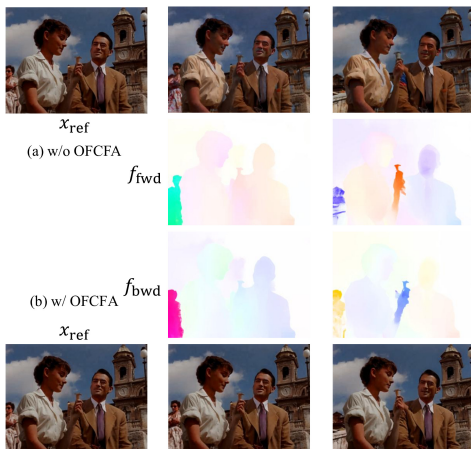


Fig. 4. Ablation study on OFCFA. Optical flows indicate the explicit cross-frame correspondence, improving coherency and detail preservation.

Table 2. Ablation studies on our key designs.

Variants	CDC↓	PSNR↑	SSIM↑	LPIPS↓
Baseline	3.916	19.80	0.862	0.215
w/ OFCFA	3.664	20.15	0.898	0.212
w/ x_{ref}	3.655	25.04	0.914	0.220
w/ f_{fwd}	3.415	24.97	0.936	0.213
w/ f_{bwd} (Ours)	3.303	26.35	0.950	0.207

The ablation without OFCFA significantly falls behind on CDC, indicating disability to generate temporally consistent colorized videos; adding OFCFA gives a surge to the model’s CDC score. As shown in Fig. 4, optical flow helps the model identify instances and colorize them correctly, as well as eliminate jittering between frames. The addition of the reference frame observably optimizes performance on PSNR, SSIM and LPIPS metrics, indicating more faithful reconstruction of frames. Forward and backward optical flow input improve the model’s performance in both aspects as well.

4. CONCLUSION AND DISCUSSION

In this paper, we propose a novel optical flow-guided diffusion framework designed for high-quality, temporally coherent video colorization. By explicitly integrating bidirectional optical flow into the training and sampling phases, our model establishes precise inter-frame correspondences that effectively mitigate common artifacts such as color flickering and content drift. We further incorporated reference images as anchors for rich color guidance. Furthermore, our hybrid two-stage training scheme allows the model to learn complex motion patterns without relying on massive real-world video datasets. While experiments demonstrate the superiority of our method, failure cases occur when handling complex scenes with multiple color assignments and rapid motion changes. Future work can focus on incorporating semantic guidance to further enhance integrity of complex structures.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62332010 and 62471009, and in part by the Phase I Open Research Projects, Yungang Academy under Grant 2025YGKFKT004.

5. REFERENCES

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2016.
- [2] Satoshi Iizuka and Edgar Simo-Serra, "DeepRemaster: Temporal Source-Reference Attention Networks for Comprehensive Video Enhancement," in *Proc. ACM SIGGRAPH ASIA*, 2019.
- [3] Yixin Yang, Zhongzheng Peng, Xiaoyu Du, Zhulin Tao, Jinhui Tang, and Jinshan Pan, "BiSTNet: Semantic Image Prior Guided Bidirectional Temporal Feature Fusion for Deep Exemplar-based Video Colorization," *arXiv:2212.02268*, 2022.
- [4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Nets," in *Proc. Advances in Neural Information Processing Systems*, 2014.
- [5] Yuzhi Zhao, Lai-Man Po, Wing-Yin Yu, Yasar Abbas Ur Rehman, Mengyang Liu, Yujia Zhang, and Weifeng Ou, "VC-GAN: Video Colorization With Hybrid Generative Adversarial Network," *IEEE Transactions on Multimedia*, vol. 25, pp. 3017–3032, 2022.
- [6] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," in *Proc. Advances in Neural Information Processing Systems*, 2020.
- [7] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising Diffusion Implicit Models," *arXiv:2010.02502*, 2020.
- [8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2022.
- [9] Wan Team, "Wan: Open and Advanced Large-Scale Video Generative Models," Tech. Rep., Alibaba Group, 2025.
- [10] Zheng Chang, Shuchen Weng, Huan Ouyang, Yu Li, Si Li, and Boxin Shi, "L-C4: Language-Based Video Colorization for Creative and Consistent Color," *arxiv: 2410.04972*, 2024.
- [11] Yifan Li, Shuai Yang, and Jiaying Liu, "Language-based Image Colorization: A Benchmark and Beyond," *arXiv: 2503.14974*, 2025.
- [12] Zichuan Huang, Yifan Li, Shuai Yang, and Jiaying Liu, "Cool-Color: Text-guided COherent OLd film COLOrization," in *Proc. ACM Int'l Conf. Multimedia in Asia*, 2024.
- [13] 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," in *Proc. Int'l Conf. Learning Representations*, 2022.
- [14] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezaatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger, "Unifying flow, stereo and depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13941–13958, 2023.
- [15] Yihan Wang, Lahav Lipson, and Jia Deng, "SEA-RAFT: Simple, Efficient, Accurate RAFT for Optical Flow," in *Proc. European Conf. Computer Vision*, 2024.
- [16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "ImageNet large scale visual recognition challenge," *Int'l Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.
- [17] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár, "Microsoft COCO: Common objects in context," *Proc. European Conf. Computer Vision*, 2014.
- [18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. IEEE Int'l Conf. Machine Learning*, 2022.
- [19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv:2301.12597*, 2023.
- [20] Federico Perazzi, Brian McWilliams, Markus Gross, Jordi Pont-Tuset, Luc Van Gool, and Alexander Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2016.
- [21] Gemini Team, "Gemini: A Family of Highly Capable Multimodal Models," Tech. Rep., Google DeepMind, 2024.
- [22] Yihao Liu, Hengyuan Zhao, Kelvin C.K. 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*, 2024.
- [23] Quan Huynh-Thu and Mohammed Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics Letters*, 2008.
- [24] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004.
- [25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2018.
- [26] Shuchen Weng, Hao Wu, Zheng Chang, Jiajun Tang, Si Li, and Boxin Shi, "L-CoDe: Language-based colorization using color-object decoupled conditions," in *Proc. AAAI Conference of Artificial Intelligence*, 2022.
- [27] Zheng Chang, Shuchen Weng, Yu Li, Si Li, and Boxin Shi, "L-CoDer: Language-based colorization with color-object decoupling transformer," in *Proc. European Conf. Computer Vision*, 2022.
- [28] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi, "L-CoIns: Language-based colorization with instance awareness," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2023.
- [29] Chang, Zheng and Weng, Shuchen and Zhang, Peixuan and Li, Yu and Li, Si and Shi, Boxin, "L-CAD: Language-based colorization with any-level descriptions using diffusion priors," in *Proc. Advances in Neural Information Processing Systems*, 2023.
- [30] Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan, "I2VEdit: First-frame-guided video editing via image-to-video diffusion models," *arXiv:2405.16537*, 2024.