

# VanGogh: A Unified Multimodal Diffusion-based Framework for Video Colorization

Zixun Fang<sup>1</sup> Zhiheng Liu<sup>2</sup> Kai Zhu<sup>1</sup> Yu Liu<sup>4</sup> Ka Leong Cheng<sup>3</sup>  
Wei Zhai<sup>1†</sup> Yang Cao<sup>1</sup> Zheng-Jun Zha<sup>1</sup>

<sup>1</sup>USTC <sup>2</sup>HKU <sup>3</sup>HKUST <sup>4</sup>Independent Researcher

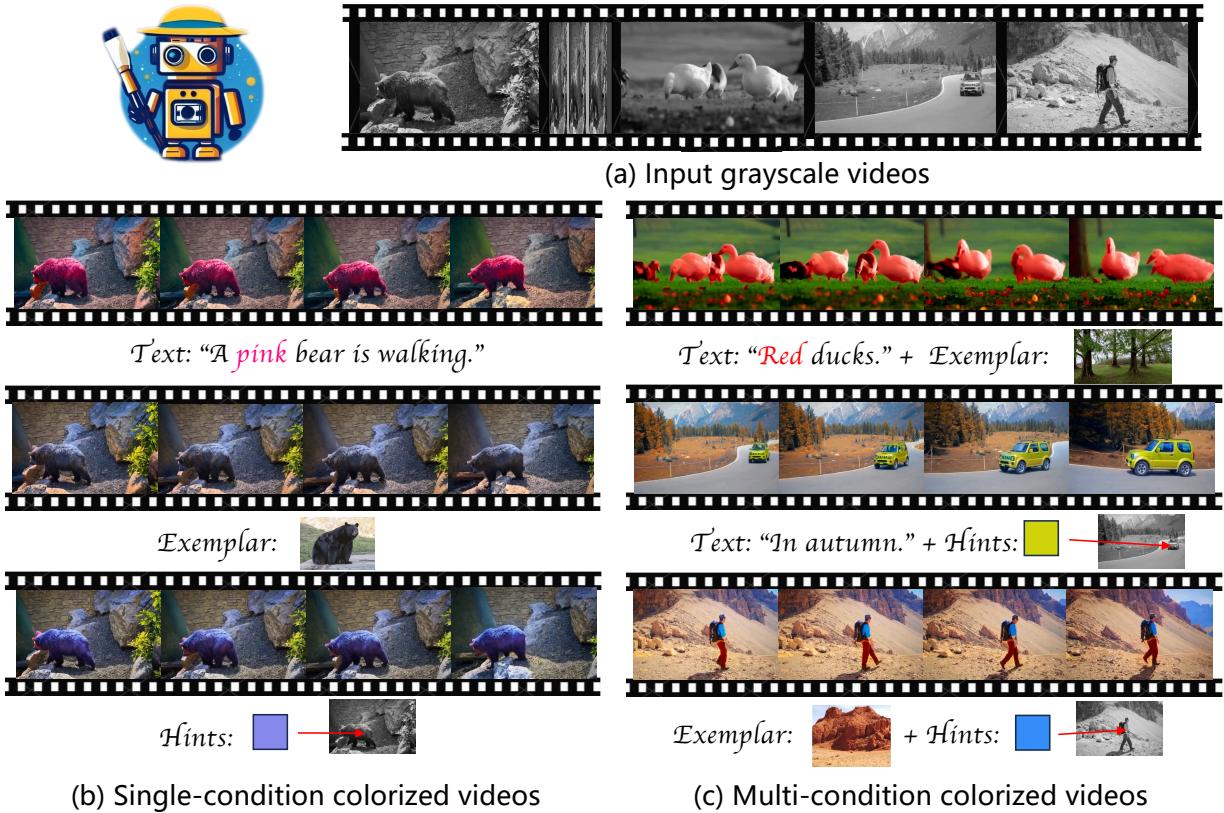


Figure 1. We present **VanGogh**, a novel multimodal video colorization method that accepts one or more conditional inputs to generate high-quality, vivid colorization results. Given the grayscale video in (a), our method can take a single condition as input to produce the result shown in (b), or accept multiple conditions for joint control, as demonstrated in (c). Please zoom in for more details.

## Abstract

Video colorization aims to transform grayscale videos into vivid color representations while maintaining temporal consistency and structural integrity. Existing video colorization methods often suffer from color bleeding and lack comprehensive control, particularly under complex motion or diverse semantic cues. To this end, we introduce **VanGogh**, a unified multimodal diffusion-based framework for video colorization. **VanGogh** tackles these challenges using

a *Dual Qformer* to align and fuse features from multiple modalities, complemented by a depth-guided generation process and an optical flow loss, which help reduce color overflow. Additionally, a color injection strategy and luma channel replacement are implemented to improve generalization and mitigate flickering artifacts. Thanks to this design, users can exercise both global and local control over the generation process, resulting in higher-quality colorized videos. Extensive qualitative and quantitative evaluations, and user studies, demonstrate that **VanGogh** achieves su-

terior temporal consistency and color fidelity. Project page: <https://becauseimbatman0.github.io/VanGogh>.

## 1. Introduction

Video colorization [16, 27, 30, 32, 42, 50, 51, 55, 60, 61] aims to transform a grayscale video into a colorized version while maintaining temporal consistency across frames and preserving the original structural information. This technology enhances the expressive quality and aesthetic appeal of videos, with broad applications in areas such as vintage video restoration and artistic creation.

Despite previous advancements, two significant challenges persist: the lack of comprehensive controllability and color bleeding. Automatic colorization techniques [24, 60] generate natural results based on model priors, but they struggle to align with user-provided conditions. Text [27, 30] and exemplar-based [16, 32, 42, 50, 51, 55] methods offer a global, semantic-level guidance, but they lack user-interactive, fine-grained control capabilities. Recent studies introduce multi-modal image colorization [4, 15, 29] by integrating conditions such as text, exemplars, and hints. This approach allows for accurate color allocation and spatial orientation, fostering greater user interaction and markedly elevating overall quality. How to effectively integrate multi-modal guidance in video coloring while maintaining the inherent properties of video remains an open question.

Recently, diffusion-based video generative models [3, 28, 46] show significant advancements due to their extensive training on large datasets of text-video pairs. These models possess rich priors and exhibit strong cross-modal capabilities. In this work, we choose Stable Video Diffusion (SVD) [3] as the base model. To better align the text and image conditions, we design a lightweight Dual Qformer structure that extracts features from text and image respectively and fuses them in a shared feature space. Notably, to enable the model to better learn color distribution from reference images, we introduce a color injection strategy that decouples color distribution from structural information, thus enhancing generalization capability. For hint conditions, we directly mark hints in grayscale videos and feed them into the UNet, avoiding the complexity of a ControlNet-style [57] approach.

However, we find that directly applying this method to colorize videos results in significant color bleeding and color overflow. Furthermore, we observe that color overflow significantly impacts optical flow estimation, thereby degrading video quality. To address this issue, we design a flow-based perceptual loss that randomly selects two predicted video frames for optical flow estimation [40, 47], followed by loss calculation against the ground truth to

reduce color overflow at object edges. We also incorporate video depth as an aid to enhance color-spatial consistency. Additionally, to mitigate the issue of imprecise reconstruction in certain detail areas of current video VAEs [12, 59], which leads to flickering artifacts, we utilize known grayscale videos for luma channel replacement in *Lab* color space during inference, thereby improving video quality.

In summary, our primary contributions are as follows: 1) We propose a novel multimodal video colorization framework that supports not only automatic colorization but also allows for combinations of one or more of the following types of guidance: text, exemplars, and hints, effectively enhancing the flexibility and interactivity of video colorization. 2) We employ a Dual Qformer to integrate images and text, along with a Depth Guider and optical flow loss to alleviate the color overflow effect. Additionally, we utilize luma channel replacement to address the inherent flickering artifacts in video VAEs, thereby enhancing the visual quality of the generated results. 3) We design qualitative and quantitative experiments, as well as user studies, to demonstrate the effectiveness and superiority of our method.

## 2. Related Work

**Image colorization.** Image colorization approaches [2, 4, 6–9, 15, 17, 18, 23, 29, 39, 44, 53, 54, 58] can be divided into two categories: automatic colorization and conditional colorization. Automatic colorization [6–9, 17, 18, 23, 39, 44, 58] aims to convert a grayscale image into a natural, vivid color image without user intention. Previous researchers [58] train a CNN to map from a grayscale image to a color distribution as they treat the problem as multinomial classification. Instance-aware colorization [39] is achieved by leveraging off-the-shelf models to detect object features, resulting in smooth outputs. ColorFormer [17] proposes a novel framework that utilizes hybrid attention and color memory to extract contextual semantics and diverse color acquisition. DDColor [18] employs dual decoders: one learns color queries from visual features, while the other provides multi-scale semantic representations. A more recent work [7] introduces an imagination module based on powerful diffusion models to synthesize diverse and colorful outcomes. Although the methods mentioned above can achieve natural and pleasing results, shortcomings, including grayish tones and color bleeding, still exist. Additionally, the inability to receive user intervention as guidance limits the application of automatic colorization. As a result, many studies explore conditional colorization [2, 4, 5, 10, 15, 26, 29, 43, 45, 48, 53, 54], aiming to generate high-quality color images that align with given conditions such as text, exemplar images, and hints. Diffusing Colors [54] analyzes the color properties of the VAE latent space and achieves superior control

through textual cues for color guidance. For exemplar-based methods, [2] builds a more accurate correspondence between a coarse result and the reference image to produce more detailed results with lower computation cost. Ke *et al.* [20] treat the methods in reference-guided colorization as style transfer works [11, 34, 52]. iColoriT [53] utilizes a ViT for hints-interactive colorization and achieves real-time performance. More recently, [4, 15, 29] integrate text, exemplar, and scribble to perform colorization in a multi-modal manner, significantly enhancing the flexibility and interactivity of the colorization process. Although the aforementioned methods achieve great success in image colorization, applying them to the video domain can lead to flaws such as temporal inconsistency, error accumulation, and color bleeding.

**Video colorization.** Due to the additional time dimension, video colorization [16, 24, 27, 30, 32, 42, 50, 51, 55, 60, 61] requires not only that the color of each individual frame is visual-appealing, but also that the relationships between frames are considered. Lei *et al.* [24] propose a self-regularized approach to automatic video colorization. However, it still tends to wash out the colors and fails to respond to user intent. Text-based video colorization methods utilize human language to guide the coloring process. [27, 30] employ text-to-image diffusion models to combine text conditions with grayscale videos. Despite many efforts, there are still issues with color overflow and temporal incoherence. Moreover, text conditions cannot achieve fine-grained local control and only provide semantic-level color descriptions such as “red” or “yellow”. Similar to text-based methods, exemplar-based video colorization aims to generate colorized videos that align with the target exemplars. ColorMNet [51] develops a memory-based feature propagation module that captures temporal features from long-range videos while reducing memory usage. Although exemplars can provide richer semantic information, exemplar-based approaches tend to lose the correlation between frames and exemplars when videos are lengthy or the motion is extensive, resulting in color artifacts. To achieve stable and precise color control, SVCNet [61] proposes scribble-based video colorization, which includes CPNet for precise colorization and SSNet for temporal smoothing. User-given color scribbles ensure precise region guidance and color control at the RGB level. However, scribbles can not express high-level semantics, such as “In winter...” or “At sunset...”. Therefore, relying solely on one modality—be it text, exemplars, or scribbles—makes it challenging to meet the demands of the task. To this end, we introduce a multi-modal solution that achieves high-quality and flexible video colorization by integrating multiple conditions.

### 3. Method

#### 3.1. Preliminaries

Latent diffusion models (LDMs) [38] are widely used in the image and video generation research community. LDMs learn to represent the distribution of images from large datasets by conducting the diffusion process and the denoising process in latent space. The given image is first encoded into latent space by a VAE [22] encoder, after which Gaussian noise is added to the latent code. The noisy latent code is then fed into a UNet as input, which learns to reconstruct the clean image distribution at timestep  $t$ . During this process, external conditions such as text are incorporated through the cross-attention mechanism to guide the denoising direction. Finally, a VAE decoder decodes the clean latent code back to pixel space. The overall training loss can be formulated as follows:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathcal{E}(\mathbf{x}), \mathbf{y}, \epsilon \sim \mathcal{N}(0, 1), t} [\|\epsilon - \epsilon_{\theta}(\mathbf{z}_t, t, \tau_{\theta}(\mathbf{y}))\|_2^2], \quad (1)$$

where  $\mathcal{E}(\cdot)$  is the VAE encoder, and  $\mathbf{x}$ ,  $\mathbf{y}$ ,  $\tau_{\theta}(\cdot)$  and  $\mathbf{z}_t$  represent the image, the text, the CLIP text encoder, and the noisy latent code at each timestep  $t$ , respectively.

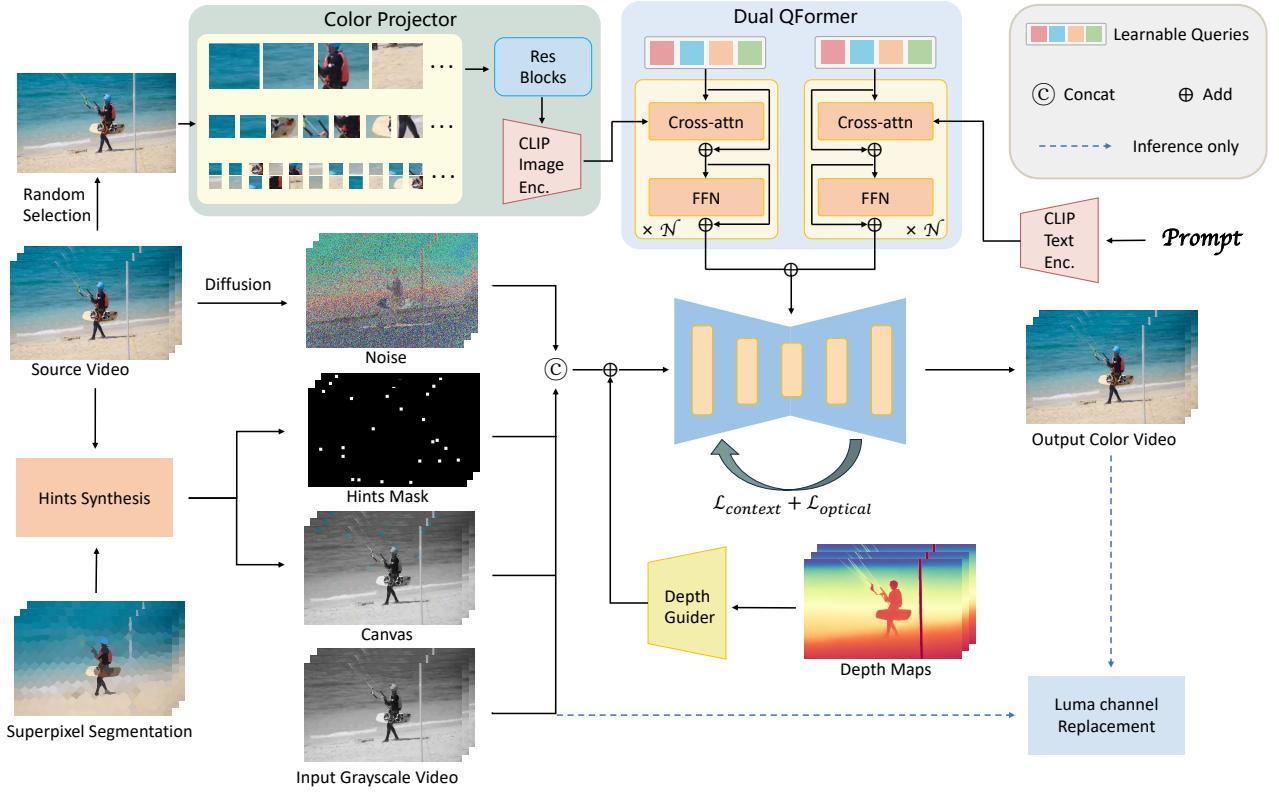
#### 3.2. Network Design

Given a grayscale video  $\mathbf{I}_g^{1:N} = [I_g^1, \dots, I_g^N]$ , our goal is to synthesize a color video  $\mathbf{I}_c^{1:N} = [I_c^1, \dots, I_c^N]$ , guided by the provided conditions such as hints, text, images, or a combination of these clues.

As shown in Fig. 2, we first convert the source color video  $\mathbf{I}_{\text{gt}}^{1:N} = [I_{\text{gt}}^1, \dots, I_{\text{gt}}^N]$  to obtain the grayscale video  $\mathbf{I}_g$ , then we randomly select one frame from the source video as an exemplar and input it into the Color Projector. The Color Projector output, along with the encoded prompt, is sent to the Dual QFormer. Meanwhile, we calculate the superpixel segmentation of the source video to synthesize hints and utilize depth maps to enhance spatial-temporal consistency. During inference, we replace the luma channel of the output video with that of the grayscale video to eliminate flickering artifacts caused by the video VAE.

**Hints injection.** Hints offer local guidance, enabling interactive control over colorization details. Users only need to mark one or several hints in one frame of the grayscale video and specify the colors, and the entire video will be colorized based on the provided hints.

To achieve this, we first perform superpixel segmentation on the color video using the simple linear iterative clustering (SLIC) [21] algorithm, representing each superpixel by its mean color (denoted as  $\mathbf{I}_{\text{sp}} \in \mathbb{R}^{3 \times F \times H \times W}$ ). This approach retains color information while reducing fine-grained structural detail, resulting in a more uniform color distribution when sampling hints. After that, we randomly select  $K$  points on the first frame of the video and use



**Figure 2. Overall pipeline.** We omit the depiction of the VAE encoder and decoder for simplicity. Given a source color video  $\mathbf{I}_{gt}^{1:N}$ , we first randomly select one frame as the exemplar image and feed it into the Color Projector, where the exemplar is divided into three groups of patches and then passed through the ResBlocks and CLIP image encoder to obtain the color features. The color features are sent to the Dual Qformer along with the encoded prompt, and the calculated features are injected into the UNet through cross attention. For hints injection, we leverage superpixel techniques to synthesize the hints mask  $\mathbf{M}^{1:N} = [M^1, \dots, M^k, \dots, M^N]$ , where  $M^k \in \mathbb{R}^{H \times W}$ , and  $M^k[i, j] = 1$  denotes that a hint exists at the coordinate  $(i, j)$  of the  $k$ -th frame; otherwise,  $M^k[i, j] = 0$ . We then compute the canvas as  $\mathbf{I}_{\text{canvas}} = \mathbf{M} \times \mathbf{I}_{\text{sp}} + (1 - \mathbf{M}) \times \mathbf{I}_g$ . Finally,  $\mathbf{I}_{\text{canvas}}$  is encoded into latent space by a VAE encoder, and  $\mathbf{M}$  is downsampled and sliced with a stride of 4 to align with the shape of the latent code. Their concatenation serves as the injection of hints of information.

**Color Projector.** During training, a randomly selected frame from the video as an exemplar inherently contains structural information similar to the whole video. However, we aim to allow the color distribution of an arbitrary image to be used as a reference, independent of structural similarity. To address this coupling of structure and color distribution in the training data, we design a Color

Projector that effectively extracts the color features of the exemplar while weakening its structural information. We begin by randomly selecting a frame from the video and then dividing it into smaller patches at various scales, specifically at resolutions of  $\frac{1}{4}$ ,  $\frac{1}{8}$ , and  $\frac{1}{16}$ . After that, the multi-scale patches are sent to the Color Projector, where their patch embeddings are extracted using ResBlocks and then concatenated into a sequence. To further enhance the color awareness between patch embeddings, we input this sequence into the CLIP [37] image encoder and use the output as color features.

**Dual Qformer.** To bridge the natural gap between text and color feature modalities, we employ a Dual Qformer to map both the text and color representations into a shared feature space, facilitating their adaptation to SVD [3]. Specifically, text is first encoded into text embeddings by the CLIP text encoder. Subsequently, these text embeddings and color features are separately fed into the Qformer, which consists of a learnable query,  $N$  stacked feed-forward network (FFN) layers, and cross-attention layers, allowing

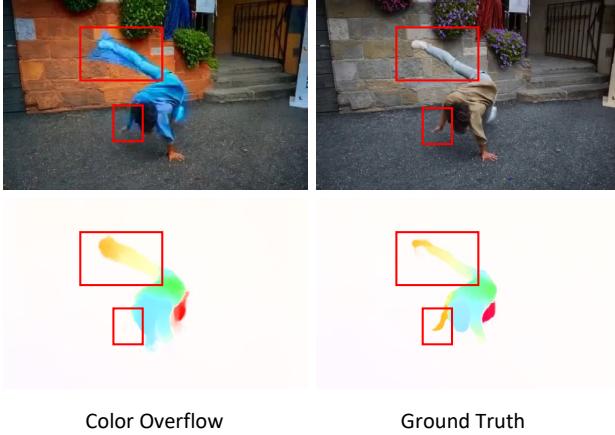


Figure 3. Color overflow caused by large motion results in optical flow estimation errors.

for effective cross-modal representation learning through the cross-attention mechanism. After obtaining the learned queries  $\mathbf{l}_{\text{text}}$  and  $\mathbf{l}_{\text{color}}$ , we apply two scaling factors,  $\lambda_1$  and  $\lambda_2$ , and fuse the queries into  $\mathbf{l}_{\text{fuse}} = \lambda_1 \times \mathbf{l}_{\text{text}} + \lambda_2 \times \mathbf{l}_{\text{color}}$ , which is then fed into the cross-attention layers of SVD.

**Depth Guider.** To reduce color overflow and improve spatial consistency in colorized results, we design a lightweight Depth Guider composed of several convolutional layers. We use Depth Anything V2 [49] to estimate depth information, which is then passed through the Depth Guider. The depth features are added to the latent code as input for the denoising UNet.

### 3.3. Training Strategy

To achieve a more efficient training process, we adopt a two-stage training strategy. In the first stage, we train the model on image datasets to ensure that SVD acquires the prior knowledge that is necessary for colorization. In the second stage, we fine-tune the model on video datasets to enhance temporal consistency.

**Image stage.** In the image training stage, we use the input image itself as the exemplar image and freeze the temporal modules of SVD to retain the knowledge of modeling the temporal dimension. Following Zhang *et al.* [55] and Liang *et al.* [29], we employ contextual loss to further enhance the

consistency of color distribution between the output image and the exemplar image. Specifically, we have

$$\begin{aligned} d^l(i, j) &= \text{sim}_{\cos}(\phi_{I_{\text{gt}}}^l(i), \phi_{I_c}^l(j)), \\ A^l(i, j) &= \text{softmax}_j(1 - \tilde{d}^l(i, j)/h), \\ \mathcal{L}_{\text{context}} &= \sum_l w_l [-\log(\frac{1}{N_l} \sum_i \max_j(A^l(i, j)))] \end{aligned} \quad (2)$$

where  $\phi^l$  represent the feature maps extracted at the *relu1\_2* layer from the VGG19 network,  $\tilde{d}^l(i, j)$  denotes the normalized cosine similarity  $d^l(i, j)$  of paired feature points,  $A^l(i, j)$  is the pairwise affinities between features from the  $l$ -th layer from the VGG19 network, and  $h$  and  $w_l$  are hyperparameters.

**Video stage.** After shifting SVD to the task of colorization, we fine-tune the temporal modules of SVD on the video dataset.

As shown in Fig. 3, we observe that when color overflow occurs in the predicted video, even with the structural information preserved intact, the areas of color overflow still exhibit optical flow estimation errors. To address this, we introduce an optical flow loss to mitigate color overflow caused by large motion. Specifically, we select the  $i$ -th frame and the  $(i+1)$ -th frame from the source color video, and use GMFlow [47] to compute the optical flow  $\mathbf{V}_{\text{gt}}$  between the two frames. We then calculate the optical flow  $\mathbf{V}_{\text{pre}}$  from the predicted video using the same method. After that, we compute the MSE loss between the optical flows:

$$\mathcal{L}_{\text{optical}} = \gamma \|\mathbf{V}_{\text{pre}} - \mathbf{V}_{\text{gt}}\|_2, \quad (3)$$

where  $\gamma$  is the weight hyperparameter.

### 3.4. Luma Channel Replacement

As shown in Fig. 4, we observe that video reconstruction with video VAEs often suffers from issues like pixel drifting in high-frequency areas, leading to noticeable flickering artifacts. Inspired techniques in image colorization [18, 29, 54], we leverage the inherent stability of the grayscale video to maintain structural consistency in VAE reconstructions. Thus, we combine the luma channel of the grayscale video with the *ab* channels of the predicted color video to produce more visually stable and pleasing results.

## 4. Experiments

### 4.1. Implementation Details

We collect an in-house image dataset containing 1.9M image-text pairs and choose OpenVid-1M [35] as our video dataset. We utilize SVD-img2vid-xt [1] as the base model and CV-VAE [59] as the video VAE to generate longer videos with fewer computational resources. During training, we resize the input data to  $320 \times 512$  and use

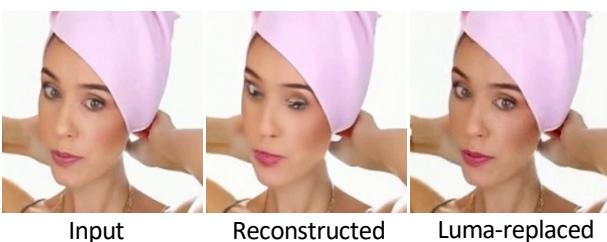
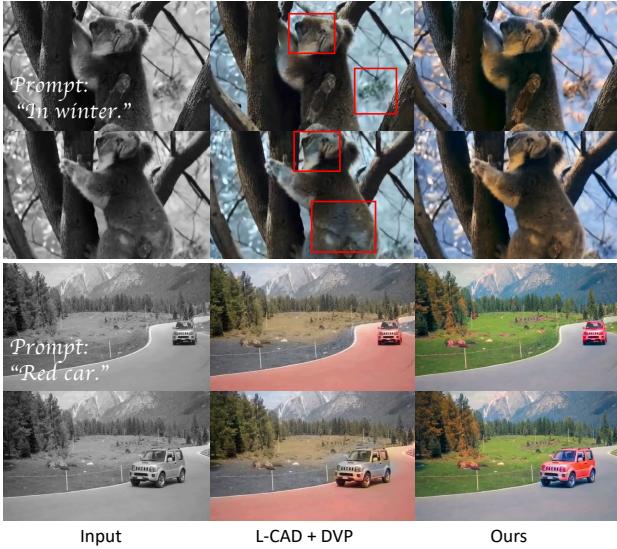


Figure 4. The reconstruction results of the video VAE exhibit flickering artifacts in high-frequency areas. Replacing the luma channel in *Lab* color space can significantly improve the visual quality.



**Figure 5. Comparison results for automatic video colorization.** VCGAN and SVCNet exhibit severe grayish issues. L-CAD suffers from flickering artifacts; even though it is post-processed by DVP, color bleeding still persists. ColorMNet heavily relies on the colored exemplar frame, and error accumulation occurs, as we can see the top of the car turning black. In contrast, our model can generate temporal-coherent, vivid color videos.

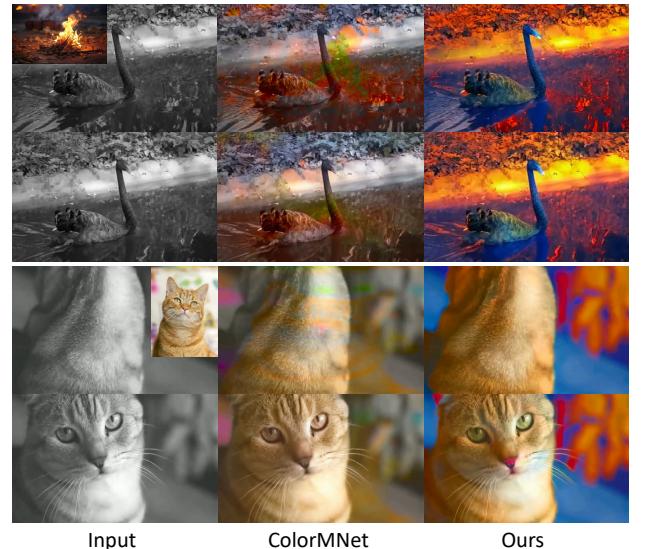


**Figure 6. Comparison for text-based video colorization.** L-CAD+DVP exhibits color bleeding and temporal incoherence. In contrast, our method can generate vivid and natural results that align with given prompts.

AdamW [33] as the optimizer, with a learning rate set to  $1 \times 10^{-5}$ . We train our model on 4 Nvidia A800 GPUs. In the image stage, we set the batch size to 8 and the training steps to 130k. In the video stage, we set the batch size to 1, the video clip length to 61 frames, and the number of iterations to 50k.

For hints sampling, we randomly select  $K$  points ranging from 0 to 150 and set the side length of the cells to be between 10 and 20. During training, we set the scaling factors  $\lambda_1 = 1$  and  $\lambda_2 = 1$ . For the contextual loss, we set  $h = 0.1$  and  $w_l$  to 8, 4, 2 for  $l = 5, 4, 3$ . For the optical flow loss, we set  $\gamma = 1$ .

During inference, we optionally utilize SD-DINO [56] to map the exemplar image to the canvas in a frame-wise manner, thereby achieving strong semantic matching for the exemplar.



**Figure 7. Comparison for exemplar-based video colorization.** ColorMNet fails to produce exemplar-aligned videos. Our method is capable of capturing color distribution and semantic correspondence from exemplars.

## 4.2. Qualitative Comparison

We conduct qualitative comparison experiments with existing colorization methods, including automatic (VCGAN [60]), text-based (L-CAD [45]), exemplar-based (ColorMNet [51]), and hints-based (SVCNet [61]) solutions.

As shown in Fig. 5, regarding automatic colorization, both VCGAN [60] and SVCNet [61] suffer from severe grayish issues. Directly applying L-CAD [45] frame by frame leads to significant temporal inconsistency and color flickering. After using Deep-Video-Prior (DVP) [25] to post-process the results from L-CAD, although temporal consistency is improved, color overflow still persists. Since ColorMNet [51] does not support automatic coloring, we first utilize L-CAD to obtain a colored exemplar frame, which is then fed to ColorMNet to enable automatic colorization capability. The results, however, heavily depend on the quality of the exemplar frame and can lead to error accumulation. In contrast, our method can produce natural, color-rich videos without any assistance.

Table 1. **Quantitative comparison for automatic colorization.** Our method is the most versatile while achieving state-of-the-art performance on most metrics.

	Automatic	Text	Exemplar	Hints	SSIM↑	LPIPS↓	PSNR↑	Colorfulness↑	FVMD↓	Colorfulness / FVMD↑
VCGAN	✓	✗	✗	✗	0.9196	0.1912	22.5359	13.5129	645.6010	0.0209
ColorMNet	✗	✗	✓	✗	<b>0.9517</b>	0.1956	22.2794	27.1427	632.0929	0.0429
L-CAD	✓	✓	✗	✗	0.9184	0.2493	21.9502	34.5885	994.6452	0.0347
L-CAD+DVP	✓	✓	✗	✗	0.9325	0.2354	22.5300	26.2798	779.8531	0.0336
SVCNet	✓	✗	✗	✓	0.9294	0.1951	23.1570	14.2890	678.0615	0.0213
Grayscale	-	-	-	-	0.9371	0.1962	23.0636	0.0000	<b>596.1689</b>	0.0000
Ours	✓	✓	✓	✓	0.9393	<b>0.1908</b>	<b>23.2013</b>	<b>60.0881</b>	662.9929	<b>0.0906</b>



Figure 8. **Comparison for hints-based video colorization.** SVCNet suffers from grayish issues and fails to align with the given hints, while our method can synthesize diverse results that align with the provided hints.

For text-based colorization, we compare our method with L-CAD + DVP. As shown in the first two rows of Fig. 6, for the prompt “*In a snowy day, a koala is climbing the tree.*”, L-CAD [45] exhibits color bleeding issues, with abrupt color changes on the koala’s face and unnatural color transitions in the thigh area due to the color bleeding effect created by the background colors. In the third and fourth rows, for the prompt “*A red car*”, L-CAD generates unpleasant results with severe color bleeding. In contrast, our method generates text-aligned and temporally coherent results.

For exemplar-based colorization, our comparison results with ColorMNet are shown in Fig. 7. In the first two rows, ColorMNet [51] fails to generate normal colorized results and suffers from significant artifacts. In the last two rows, ColorMNet exhibits grayish issues and color bleeding effects as the cat’s eyes turn yellow. In contrast, our method can produce frames that align with the exemplar image and maintain good temporal consistency.

For hints-based colorization, we first select one or several hints in the first frame of the grayscale video.

The hints are then propagated throughout the entire video (using CPNet for SVCNet [61] and CoTracker [19] for our method). As shown in the first two rows of Fig. 8, we select three hints on the boat and the sky; although SVCNet can generate natural color videos, it cannot respond to the given hints. In the last two rows, however, with only one hint as input, SVCNet exhibits grayish effects. In contrast, our method can effectively respond to the provided hints and achieves diverse coloring of various regions with only a minimal number of hints.

### 4.3. Quantitative Comparison

To ensure a fair comparison, we conduct a quantitative evaluation of our method against existing methods on the DAVIS dataset [36] in an unconditional manner, using SSIM, LPIPS, PSNR, Colorfulness [13], and Fréchet video motion distance (FVMD) [31] to assess video quality. Additionally, we argue that colorfulness and temporal consistency should be evaluated as a unified metric for the task of video colorization, as this task not only requires temporal consistency but also demands diversity and richness in color. If Colorfulness and FVMD are used separately to assess video quality, it can lead to the following issues: 1) Each frame may exhibit very good visual quality, but the inter-frame relationships could be completely ignored; 2) Grayish videos tend to achieve higher FVMD scores. As observed in our experiments, the FVMD metric for grayscale videos is the best, yet it does not satisfy the task of colorization. Therefore, we propose a new metric, denoted as Colorfulness / FVMD, which effectively takes into account both color and temporal relationships. As shown in Tab. S1, our method is not only the most versatile but also achieves state-of-the-art performance across multiple metrics.

Table 2. **Quantitative comparison for text-based colorization.** Our method surpasses L-CAD in both CLIP score and Colorfulness.

	Colorfulness↑	CLIP score↑
L-CAD	43.0182	62.9091
Ours	<b>62.1264</b>	<b>65.0381</b>

For text-based comparison, we first leverage video captioning models to obtain descriptions of colored objects



Figure 9. **Ablation studies.** Removing the Color Projector results in inaccurate color allocation. Without the optical flow loss, color overflow effects occur. Eliminating the Depth Guider leads to spatial-temporal inconsistency. In contrast, our full method achieves smooth and condition-aligned videos compared to other designs.

Table 3. **Quantitative comparison for exemplar-based colorization.** We outperform ColorMNet in both LPIPS and Colorfulness, demonstrating the robustness of our method.

	LPIPS↓	Colorfulness↑
ColorMNet	0.6437	38.9738
Ours	<b>0.5693</b>	<b>65.9782</b>

Table 4. **Quantitative comparison for hints-based colorization.** Our method achieves better performance in hints-guided colorization compared to SVCNet.

	SSIM↑	Colorfulness↑
SVCNet	0.9302	19.9826
Ours	<b>0.9418</b>	<b>63.6238</b>

for the DAVIS[36] validation set, then we employ CLIP score [14] to evaluate the coherence between the provided text and the generated videos. As demonstrated in Tab. 2, our method achieves a higher CLIP score and Colorfulness compared to L-CAD[45]. For exemplar-based methods, we collect a set of diverse images as exemplars, and we compute the LPIPS between the exemplar images and synthesized videos. As shown in Tab. 3, our method surpasses ColorMNet[51] in both LPIPS and Colorfulness, demonstrating the robustness of our algorithm. For hints-based approaches, we compare our method with SVCNet[61]. Specifically, we first conduct color augmentation for the source color videos to avoid automatic colorization priors; then we randomly sample hints from the augmented videos as input for SVCNet and our method. As illustrated in Tab. 4, we outperform SVCNet in both SSIM and Colorfulness.

#### 4.4. Ablation Study

To validate the effectiveness of our design, we conduct ablation studies on the components of the model. As shown in Fig. 9, given the prompt “A boy wearing a blue T-shirt.” along with an exemplar image describing autumn maple

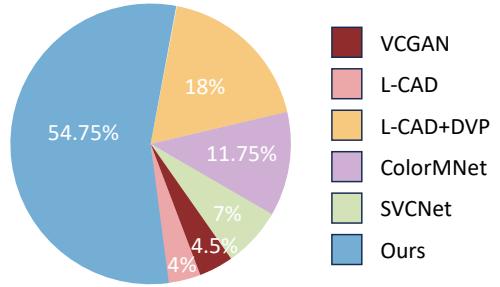


Figure 10. **User study results.** The vast majority of users prefer our model.

leaves. Removing the Color Projector results in ineffective extraction of color information from the exemplar. Furthermore, omitting the optical flow loss leads to color overflow during significant motion, as seen with the arms turning blue. Without the Depth Guider, color bleeding also occurs. Our complete method, however, can generate results that align with the input conditions and exhibit good temporal consistency. We also examine the effectiveness of the Dual Qformer, please refer to the supplementary materials for more details on the ablation studies.

#### 4.5. User Study

To further assess the quality of the generated results, we conduct a user study, as subjective evaluations from users are often a more reasonable standard in the field of colorization. Specifically, we prepare 20 questions that encompass several dimensions: color richness, temporal consistency, aesthetic preference, and the alignment between the given conditions (text, exemplars, and hints) and the synthesized videos. In each question, participants are asked to choose the best result from various compared methods. We receive 20 questionnaires, and the results are shown in Fig. 10. Our method acquires the preference of the majority of users, followed by L-CAD+DVP and ColorMNet, while L-CAD and VCGAN are the least voted.

## 5. Conclusion

In this paper, we present VanGogh, a unified diffusion-based framework for multimodal video colorization. Comprehensive experiments and user studies demonstrate that our method is capable of generating visually appealing results that align with given conditions, thus significantly enhancing the flexibility and interactivity of the video colorization process. We believe our methodology can serve as a valuable reference for researchers in the field of video colorization.

## References

- [1] Stable-Video-Diffusion-img2vid-xt. <https://huggingface.co/stabilityai/stable-video-diffusion-img2vid-xt>, 2023. 5, 1
- [2] Yunpeng Bai, Chao Dong, Zenghao Chai, Andong Wang, Zhengzhuo Xu, and Chun Yuan. Semantic-sparse colorization network for deep exemplar-based colorization. In *Eur. Conf. Comput. Vis.*, pages 505–521. Springer, 2022. 2, 3
- [3] 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. 2, 4, 1
- [4] 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*, pages 1–11, 2024. 2, 3
- [5] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. L-CoIns: Language-based colorization with instance awareness. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19221–19230, 2023. 2
- [6] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *Int. Conf. Comput. Vis.*, pages 415–423, 2015. 2
- [7] Xiaoyan Cong, Yue Wu, Qifeng Chen, and Chenyang Lei. Automatic Controllable Colorization via Imagination. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2609–2619, 2024. 2
- [8] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *Int. Conf. Comput. Vis.*, pages 567–575, 2015.
- [9] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth. Learning diverse image colorization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6837–6845, 2017. 2
- [10] Faming Fang, Tingting Wang, Tieyong Zeng, and Guixu Zhang. A superpixel-based variational model for image colorization. *IEEE Trans. Vis. Comput. Graph.*, 26(10):2931–2943, 2019. 2
- [11] Leon A Gatys. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 3
- [12] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *Eur. Conf. Comput. Vis.*, pages 102–118. Springer, 2022. 2
- [13] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, pages 87–95. SPIE, 2003. 7, 1
- [14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 8, 1
- [15] Zhitong Huang, Nanxuan Zhao, and Jing Liao. Unicolor: A unified framework for multi-modal colorization with transformer. *ACM Trans. Graph.*, 41(6):1–16, 2022. 2, 3
- [16] Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM Trans. Graph.*, 38(6):1–13, 2019. 2, 3
- [17] Xiaozhong Ji, Boyuan Jiang, Donghao Luo, Guangpin Tao, Wenqing Chu, Zhifeng Xie, Chengjie Wang, and Ying Tai. Colorformer: Image colorization via color memory assisted hybrid-attention transformer. In *Eur. Conf. Comput. Vis.*, pages 20–36. Springer, 2022. 2
- [18] Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. DDCColor: Towards photo-realistic image colorization via dual decoders. In *Int. Conf. Comput. Vis.*, pages 328–338, 2023. 2, 5, 1
- [19] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 4, 7
- [20] Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson WH Lau. Neural preset for color style transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14173–14182, 2023. 3
- [21] Salar Hosseini Khorasgani, Yuxuan Chen, and Florian Shkurti. SLIC: Self-supervised learning with iterative clustering for human action videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16091–16101, 2022. 3
- [22] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3
- [23] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Eur. Conf. Comput. Vis.*, pages 577–593. Springer, 2016. 2
- [24] Chenyang Lei and Qifeng Chen. Fully automatic video colorization with self-regularization and diversity. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3753–3761, 2019. 2, 3
- [25] Chenyang Lei, Yazhou Xing, Hao Ouyang, and Qifeng Chen. Deep video prior for video consistency and propagation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):356–371, 2022. 6, 1
- [26] Bo Li, Yu-Kun Lai, Matthew John, and Paul L Rosin. Automatic example-based image colorization using location-aware cross-scale matching. *IEEE Trans. Image Process.*, 28(9):4606–4619, 2019. 2

- [27] Jiaxing Li, Hongbo Zhao, Yijun Wang, and Jianxin Lin. Towards Photorealistic Video Colorization via Gated Color-Guided Image Diffusion Models. In *ACM Int. Conf. Multimedia*, 2024. 2, 3
- [28] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 2
- [29] Zhixin 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. 2, 3, 5, 1
- [30] 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. 2, 3
- [31] Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet Video Motion Distance: A Metric for Evaluating Motion Consistency in Videos. *arXiv preprint arXiv:2407.16124*, 2024. 7, 1
- [32] Sifei Liu, Guangyu Zhong, Shalini De Mello, Jinwei Gu, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Switchable temporal propagation network. In *Eur. Conf. Comput. Vis.*, pages 87–102, 2018. 2, 3
- [33] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [34] Yifang Men, Yuan Yao, Miaomiao Cui, Zhouhui Lian, Xuansong Xie, and Xian-Sheng Hua. Unpaired cartoon image synthesis via gated cycle mapping. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3501–3510, 2022. 3
- [35] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 5
- [36] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 7, 8, 1
- [37] 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 *Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021. 4
- [38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 3
- [39] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7968–7977, 2020. 2
- [40] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2
- [41] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *ArXiv*, abs/1812.01717, 2018. 1
- [42] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Eur. Conf. Comput. Vis.*, pages 391–408, 2018. 2, 3
- [43] Hanhang Wang, Deming Zhai, Xianming Liu, Junjun Jiang, and Wen Gao. Unsupervised deep exemplar colorization via pyramid dual non-local attention. *IEEE Trans. Image Process.*, 2023. 2
- [44] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. CT 2: Colorization transformer via color tokens. In *Eur. Conf. Comput. Vis.*, pages 1–16. Springer, 2022. 2
- [45] Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, Boxin Shi, et al. L-CAD: Language-based colorization with any-level descriptions using diffusion priors. *Adv. Neural Inform. Process. Syst.*, 36, 2024. 2, 6, 7, 8, 1
- [46] 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 *Eur. Conf. Comput. Vis.*, pages 399–417. Springer, 2025. 2
- [47] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. GMFlow: Learning optical flow via global matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8121–8130, 2022. 2, 5
- [48] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. Stylization-based architecture for fast deep exemplar colorization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9363–9372, 2020. 2
- [49] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. *arXiv preprint arXiv:2406.09414*, 2024. 5
- [50] 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 Trans. Pattern Anal. Mach. Intell.*, 2024. 2, 3
- [51] Yixin Yang, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. ColorMNet: A Memory-based Deep Spatial-Temporal Feature Propagation Network for Video Colorization. In *Eur. Conf. Comput. Vis.*, pages 336–352. Springer, 2025. 2, 3, 6, 7, 8
- [52] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Int. Conf. Comput. Vis.*, pages 9036–9045, 2019. 3
- [53] Jooyeol Yun, Sanghyeon Lee, Minho Park, and Jaegul Choo. iColoriT: Towards propagating local hints to the right region in interactive colorization by leveraging vision transformer. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 1787–1796, 2023. 2, 3
- [54] Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. Diffusing Colors: Image Colorization with Text Guided Diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 2, 5, 1

- [55] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8052–8061, 2019. 2, 3, 5
- [56] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Adv. Neural Inform. Process. Syst.*, 36, 2024. 6
- [57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, pages 3836–3847, 2023. 2
- [58] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Eur. Conf. Comput. Vis.*, pages 649–666. Springer, 2016. 2
- [59] Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. CV-VAE: A Compatible Video VAE for Latent Generative Video Models. *arXiv preprint arXiv:2405.20279*, 2024. 2, 5, 1
- [60] 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 Trans. Multimedia*, 25:3017–3032, 2022. 2, 3, 6
- [61] Yuzhi Zhao, Lai-Man Po, Kangcheng Liu, Xuehui Wang, Wing-Yin Yu, Pengfei Xian, Yujia Zhang, and Mengyang Liu. SVCNet: Scribble-based video colorization network with temporal aggregation. *IEEE Trans. Image Process.*, 2023. 2, 3, 6, 7, 8

# VanGogh: A Unified Multimodal Diffusion-based Framework for Video Colorization

## Supplementary Material

In this supplementary material, we first describe the specific details of our implementation in App. A, including the parameter settings of the network during training and the analysis of luma channel replacement. Then, in App. B, we present more detailed ablation experiments and an analysis of the Dual Qformer. Finally, in App. C, we show more visualization results.

### A. Implementation Details

#### A.1. Training details

We use SVD-img2vid-xt [1] to initialize the UNet. The original SVD [3] has 4 input channels, but since we concatenated the hints mask, canvas, and grayscale video to the original latent code, the input channels are expanded to 13 (4 for the original latent code, 1 for the hints mask, 4 for the canvas, and 4 for the grayscale video). During training, we modify the input channels of the UNet *conv\_in* layer with *in\_channel* = 13 and randomly initialize the parameters.

For the selection of control conditions, we set a random variable  $X \sim U(0, 1)$ . When  $X \in [0, 0.3]$ , we only use the image modality as the condition, and both the hints information and the text are set to null. Similarly, when  $X \in [0.3, 0.5]$ , we only use text as the condition, and when  $X \in [0.5, 0.8]$ , we only keep the hints guidance. For multi-condition control, when  $X \in [0.8, 0.83]$ , we preserve the image and hints; when  $X \in [0.83, 0.86]$ , we keep the text and hints; when  $X \in [0.86, 0.9]$ , we maintain the image and text. When  $X \in [0.9, 0.95]$ , all three modalities are retained, and finally, when  $X \in [0.95, 1]$ , all three modalities are discarded.

#### A.2. Luma channel replacement

The *Lab* color space (often referred to as *CIELAB* or simply *LAB*) is a color model that is designed to be a more perceptually uniform representation of colors than other color spaces like RGB or CMYK. The key components of the *Lab* color space are:

- 1) Luma: This represents the lightness of the image, ranging from 0 (black) to 100 (white).
- 2) a: This axis represents the color's position between green and red. Negative values indicate green, and positive values indicate red.
- 3) b: This axis represents the position between blue and yellow. Negative values indicate blue, and positive values indicate yellow.

In the task of colorization, the luma channel of the grayscale input retains the structural information intact, which is why many image-based colorization methods [18, 29, 54] replace the luma channel of the colorized output with the luma channel of the grayscale input to enhance image quality. However, we find that this technique can also address the flickering artifacts caused by the VAE video VAE [59], thus not only improving image quality but also enhancing temporal consistency. As shown in Fig. 4, we observe that the reconstruction results of the VAE show obvious artifacts. After replacing the luma channels, the quality improves significantly. Thus, we argue that the human eye is more sensitive to structural artifacts than to color artifacts.

### B. Ablation study

#### B.1. Quantitative results

We conduct ablation studies on the components of the mode across the following metrics: SSIM, LPIPS, PSNR, Colorfulness [13], FVMD [31] and Colorfulness / FVMD. Note that we do not adopt FVD [41] because FVD can only measure whether the appearance of the generated video is consistent with the reference video. However, video colorization is inherently an ill-posed problem, and evaluating the quality of the generated results based on whether the appearance aligns with the ground truth is biased. Furthermore, our experiments reveal that after post-processing with DVP [25], the results of L-CAD [45]+DVP perform worse on the FVD metric than L-CAD, which colorizes each frame individually. This is clearly unreasonable.

We evaluate our designs on the DAVIS validation dataset [36] in an automatic manner. As shown in Tab. S1, the full method outperforms other designs in all metrics.

#### B.2. Analysis on the Dual QFormer

We propose the Dual QFormer to fill the gap between image modality and text modality and force the adaption of SVD [3] to the task of colorization. The learnable queries extract color semantics from both the text and the image effectively, and fuse them into a shared feature space, enhancing the alignment between the colorization results and the given conditions. For text-based colorization, we leverage CLIP score[14] to assess the alignment between the colorization videos and the provided text. As shown in Tab. S2, Dual QFormer offers better perceptual capabilities in text description. For exemplar-based colorization,

Table S1. **Quantitative comparison for ablation studies.** The full method outperforms other designs across all the metrics.

	SSIM↑	LPIPS↓	PSNR↑	Colorfulness↑	FVMD↓	Colorfulness / FVMD↑
w/o contextual loss	0.9154	0.1990	22.4264	56.2335	674.9781	0.0833
w/o optical flow loss	0.9023	0.2271	21.2794	57.8907	721.2509	0.0802
w/o Color Projector	0.9278	0.2053	22.8653	53.9126	671.0116	0.0803
w/o Depth Guider	0.9124	0.2123	21.1537	58.3792	698.4211	0.0835
w/o Dual QFormer	0.9302	0.2019	22.5060	55.7230	683.8873	0.0814
Full method	<b>0.9393</b>	<b>0.1908</b>	<b>23.2013</b>	<b>60.0881</b>	<b>662.9929</b>	<b>0.0906</b>



Figure S1. **Scaling factor control.** By modifying the text scale and image scale, our method can generate results with varying degrees of alignment between text and image. When the text scale increases, the duck turns red, while when the image scale increases, the background turns green.

we compare the full method with both the Dual QFormer-removed version and the Color Projector-removed version. As shown in Tab. S3, removing either the Dual QFormer or the Color Projector results in the insufficient extraction of the color semantics of the exemplar, thereby compromising the alignment between the coloring results and the exemplar images.

Table S2. **Quantitative comparison for text-based ablation.** The Dual QFormer enhances performance in both CLIP score and Colorfulness.

	Colorfulness↑	CLIP score↑
w/o Dual QFormer	60.3253	61.0973
Full method	<b>62.1264</b>	<b>65.0381</b>

Table S3. **Quantitative comparison for exemplar-based ablation.** The Dual QFormer and Color Projector both enhance performance in LPIPS score and Colorfulness.

	LPIPS↓	Colorfulness↑
w/o Color Projector	0.5912	52.9806
w/o Dual QFormer	0.6023	50.4281
Full method	<b>0.5693</b>	<b>65.9782</b>

As depicted in Sec. 3.2, we set two scaling factors,  $\lambda_1$  and  $\lambda_2$ , to fuse the extracted color features and text features into a shared feature space. With this design, users can control the intensity of the given exemplar image and text, thereby enhancing the flexibility of multi-condition colorization. As shown in Fig. S1, given the prompt "Red ducks" and the exemplar image, modifying the text scale ( $\lambda_1$ ) and the image scale ( $\lambda_2$ ) in different ratios

yields diverse colorized results, demonstrating that Dual QFormer offers a unified feature space and thus improves the interactivity and flexibility.

## C. More Qualitative Results

### C.1. Automatic colorization results

As shown in Fig. S2, we present more automatic colorization results. Our method can directly generate realistic, natural, and richly colored videos without any conditions.

### C.2. Single-condition colorization results

We also include the colorization videos under single-condition guidance. As shown in Fig. S3, Fig. S4, and Fig. S5, we display the results for text, exemplar, and hints, respectively. Our method can synthesis condition-aligned color videos.

### C.3. Multi-condition colorization results

Our model can accept multiple conditions to achieve global-local joint control, as shown in Fig. S6. Given various conditions, our method can generate diverse results that align with the given conditions, significantly improving the visual quality, interactivity, and flexibility of video colorization.



Figure S2. **Automatic colorization results.** Each pair of two rows contains a gray scale video and a colorized video. Our method is capable of generating natural, vivid color videos.



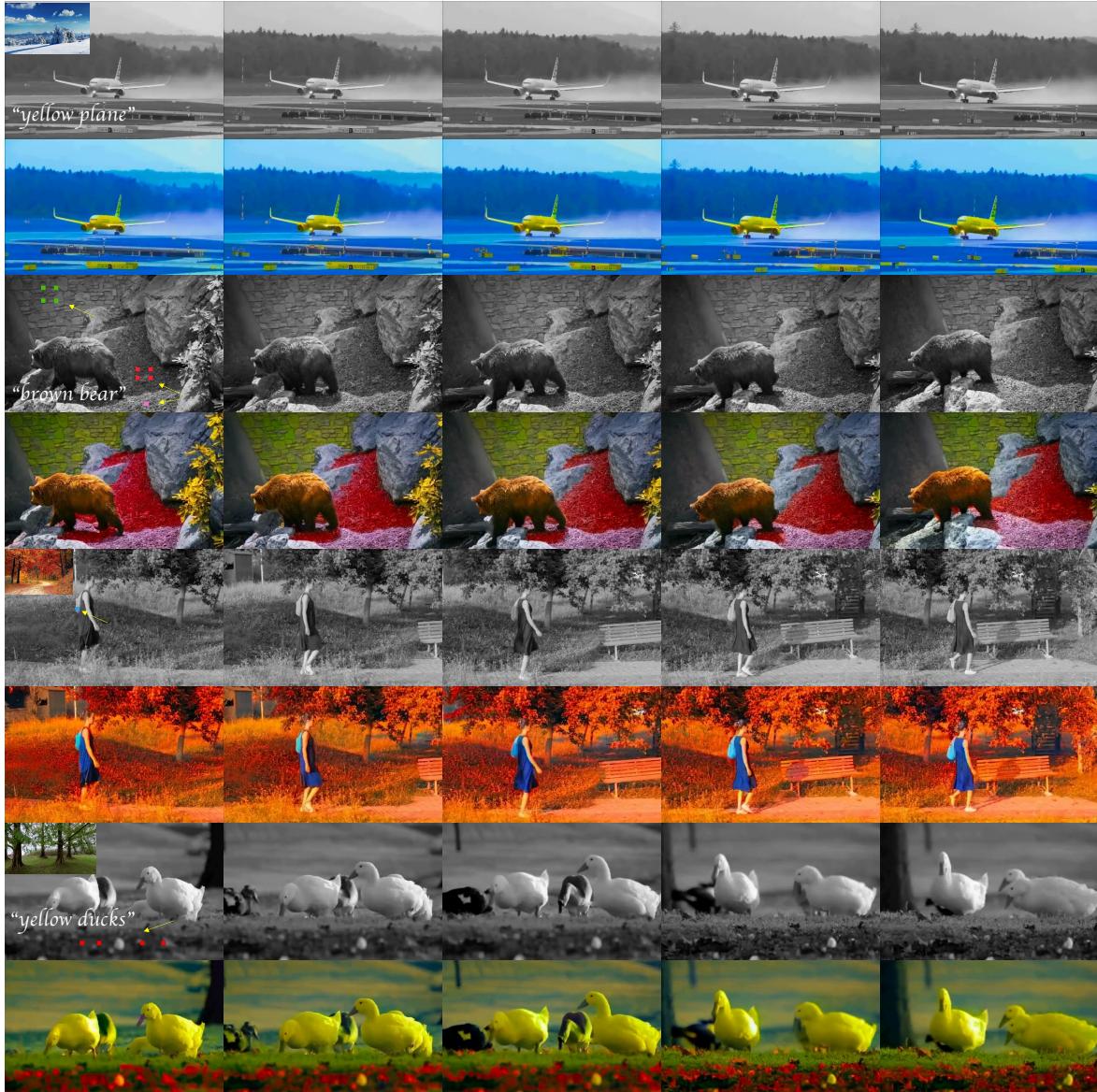
Figure S3. **Text-based colorization results.** In every two rows, the first row contains a grayscale video and a prompt, while the second row shows the colorized result. Our method can generate color videos that align with the text.



Figure S4. **Exemplar-based colorization results.** In every two rows, the first row contains a grayscale video and an exemplar image, while the second row shows the colorized result. Our method can generate color videos that align with the exemplars.



Figure S5. **Hints-based colorization results.** In every two rows, the first row contains a grayscale video along with the provided hints (see yellow arrow(s)), while the second row shows the colorized result. Our method can generate color videos that align with the hints.



**Figure S6. Multi-condition colorization results.** The first two rows display the colorization results of text + exemplar. The third and fourth rows show the results of text + hints, the fifth and sixth rows display the results of exemplar + hints, and the last two rows present the colorization videos of text + exemplar + hints.