

VIDiff: Translating Videos via Multi-Modal Instructions with Diffusion Models

Zhen Xing¹ Qi Dai² Zihao Zhang¹ Hui Zhang¹ Han Hu² Zuxuan Wu¹ Yu-Gang Jiang¹

¹ Fudan University ² Microsoft Research Asia



Figure 1. We introduce VIDiff, a generalist model for video translation tasks. Given an input video and human instructions, our unified model effectively accomplishes tasks such as video re-colorization, dahnazing, deblurring, editing, in-painting, and object segmentation.

Abstract

Diffusion models have achieved significant success in image and video generation. This motivates a growing interest in video editing tasks, where videos are edited according to provided text descriptions. However, most existing approaches only focus on video editing for short clips and rely on time-consuming tuning or inference. We are the first to propose Video Instruction Diffusion (VID-

iff), a unified foundation model designed for a wide range of video tasks. These tasks encompass both understanding tasks (such as language-guided video object segmentation) and generative tasks (video editing and enhancement). Our model can edit and translate the desired results within seconds based on user instructions. Moreover, we design an iterative auto-regressive method to ensure consistency in editing and enhancing long videos. We pro-

vide convincing generative results for diverse input videos and written instructions, both qualitatively and quantitatively. More examples can be found at our website <https://ChenHsing.github.io/VIDiff>.

1. Introduction

In recent years, the field of artificial intelligence has witnessed significant advancements, especially in Natural Language Processing (NLP), where Large Language Models (LLMs) like GPT [36] unify diverse tasks under one single framework. In contrast, the development of foundational models in computer vision is still far behind that in NLP, due to the natural diversities arising from vision tasks, *e.g.*, various output formats, and different model architectures.

Inspired by the success of GPT [36] in unifying NLP tasks, some foundational models have emerged that aim to unify visual tasks, primarily focusing on understanding tasks like recognition and retrieval [26, 49, 52]. Nonetheless, research on unified frameworks for generative tasks is relatively scarce. InstructDiffusion [13] explores generalizing diffusion models to both image editing and understanding tasks. Despite that, unifying tasks in the video domain is still challenging, since the data distribution and task variation are more complex than images. Few have endeavored to design a unified framework that addresses both video understanding and editing tasks.

Among generative modeling tasks [6, 16, 44, 60, 61, 63, 64, 69], Video-to-Video (V2V) translation possesses enormous potential in social media, advertising, promotional campaigns, television, *etc.* Presently, most methods rely on detailed textual descriptions, a strict requirement hinging on accurate descriptions of the original and target videos. In addition, the majority of methods depend on time-consuming training and inference processes such as DDIM [47] inversion. While instructional editing [6, 8, 12, 40] takes in user-friendly prompts, current techniques can only be applied to a very few editing scenarios. Besides, while instructional texts are able to relieve the need for professional prompts, producing precise and detailed descriptions of expected outputs sometimes require domain knowledge, *e.g.*, art or medicine. It is desirable to provide more effective instructions to enable effortless guidance.

To address these concerns, in this paper, we present a general video diffusion framework, VIDiff, for various conditional video-to-video translation that operates on multi-modal instructions. Our method accepts instructions together with a source video as input and generates a target video output. In addition to the textual instruction, we also leverage images that “worth a thousand words” as straightforward instruction without demanding expert knowledge. We therefore design a multi-modal condition injection mechanism for image and text-guided video edit-

ing. Our approach is trained with multiple stages to adapt a pre-trained T2I model [44] for V2V translation. We also design an iterative training and inference scheme to allow long video translation. We effortlessly extend our method to various tasks, building up a unified framework for video understanding and editing.

In summary, the main contributions of this paper can be summarized as follows:

- We are the first to employ a unified diffusion framework for both video understanding and video enhancement tasks.
- We design a multi-stage training method to seamlessly transfer the T2I model for multi-modal conditional video translation tasks.
- Our proposed iterative generation method is simple yet effective, allowing easy application in long video translation tasks.
- We conduct extensive experiments, showcasing several cases, proving the effectiveness of our approach both qualitatively and quantitatively.

2. Related Work

Video Language Foundation Model. Although image-language foundation models have been successfully applied to various tasks, including image recognition [41], image-text retrieval [41], visual question answering [2], and even image generation and editing [13], there has been limited research on video-language foundation models [7, 56, 59]. Existing methods are typically designed for understanding tasks like classification. Inspired by contrastive learning, Omnilv [49] explores cross-modal alignment for images, text, and videos, demonstrating effectiveness in video classification and retrieval tasks. Unmasked Teacher [27] combines masked auto-encoder with contrastive learning in a multimodal paradigm, making it applicable to diverse video-language tasks such as classification, retrieval, temporal detection, and video question-answering. Approaches like Unicorn [66] and OmniTracker [50] aim to unify video object segmentation and tracking tasks. However, there has been limited research on video translation tasks. We are the first to design multiple video translation tasks into a unified foundation model.

Text-guided Image Translation. Image editing is a complex process that involves modifying an image based on specific guidance, often provided by a reference image, rather than generating images without constraints. Various methods have been developed to address this task. One approach includes zero-shot image-to-image translation techniques, like SDEdit [33], which applies diffusion and denoising techniques to a reference image. Other methods incorporate optimization techniques to refine the editing process. For example, Imagic [21] utilizes textual inver-

sion concepts from [11]. Null-text Inversion [35] leverages Prompt-to-Prompt [16] to control cross-attention [25] behavior in the diffusion model. However, these methods require a time-consuming editing process due to the need for per-image optimization. Instead, Instruct Pix2Pix [6] achieves image editing by training on paired synthetic data. More recently, InstructDiffusion [13] unifies several vision tasks under this paradigm. In this paper, we focus on the video translation task, which is more challenging compared to images.

Text guided Video Editing. Video editing methods often require detailed textual descriptions of both the original and target videos, and then reconstruct the videos based on these descriptions for editing purposes. Tune-A-Video [57] and SimDA [63] fine-tune a single model to generate new videos with similar motion patterns. Video-P2P [30], Vid2Vid-Zero [53], and FateZero [39] leverage cross-attention maps to adjust videos. More recently, InstructVid2Vid [40] and InsV2V [8] attempt to build instruction-based video editing. While we share a similar structure, our focus is on unifying various video tasks in a generalist framework, and our approach is also applicable to editing long videos.

3. Method

In this section, we first introduce the preliminaries of the latent diffusion model (LDM) in Sec. 3.1. Next, we explain the problem definition in Sec. 3.2. Then, we present the collection of the dataset used in our approach in Sec. 3.3. Finally, we describe the architecture and training pipeline of our VIDiff in Sec. 3.4.

3.1. Preliminaries of LDM

Diffusion models [18, 47] model complex data distributions by two pivotal processes: diffusion and denoising. Given an input data sample \mathbf{x} from the distribution $p(\mathbf{x})$, the diffusion process adds random noise to transform the sample to $\mathbf{x}_t = \alpha_t \mathbf{x} + \sigma_t \epsilon$, where ϵ is sampled from a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This diffusion process is achieved by T steps, and the noise scheduler is parametrized by the parameters α_t and σ_t . In the denoising stage, the model employs ϵ -prediction and v -prediction methodologies to learn a denoiser function ϵ_θ , which is trained to minimize the mean square error loss as follows:

$$\mathbb{E}_{\mathbf{x}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} [||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||_2^2]. \quad (1)$$

Latent Diffusion Model (LDM) [44] utilizes a VAE [48] encoder \mathcal{E} to compress the input data in low-dimensional latent space. LDM conducts diffusion and denoising processes in both the training and inference stages. The optimizing objective is:

$$\mathbb{E}_{\mathbf{x}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} [||\epsilon - \epsilon_\theta(\mathcal{E}(\mathbf{x}_t), \mathbf{c}, t)||_2^2], \quad (2)$$

where \mathbf{c} is the text condition that is extracted by the pre-trained CLIP [41] ViT-L/14 model from the text prompt. LDM is a text-to-image model, and we adapt it to video-to-video translation tasks in this work.

3.2. Problem Definition

Video understanding [28, 51] and generative tasks differ in various aspects. However, we can reformulate each task and raise some commonalities. For most common video tasks, we can consider them as conditional video translation tasks. For instance, video object segmentation can be seen as translating raw video pixels into corresponding segmentation maps. Video recoloring task involves translating grayscale video pixels into colored video frames. As for video enhancement and video editing tasks, they are inherently video translation tasks as well.

We intend to design a unified model capable of addressing all these tasks simultaneously. Thus, we tackle the aforementioned tasks uniformly in an instructional video translation. Specifically, given a source video V_s and an instruction \mathbf{c} , the objective is to translate V_s into the corresponding target video V_t . To achieve this goal, during the training phase, we construct training video triplets $\langle V_s, V_t, \mathbf{c} \rangle$ for each task. In the inference phase, the method could translate a source video V_s to the target video V_t conditioned on instruction \mathbf{c} .

3.3. Training Data Construction

As previously mentioned, the training of a video-to-video translation model relies on the construction of triplets, which consist of $\langle V_s, V_t, \mathbf{c} \rangle$. In this section, we will discuss how to collect datasets for various tasks. The visualization of the triplet dataset can be found in Fig. 1.

Video Re-colorization and Inpainting For tasks like video re-colorization and video inpainting, we can easily construct training data using unlabeled videos. Any video can be converted into a grayscale version, and videos with missing parts can be generated by creating masks of arbitrary shapes [76]. As for the instruction, we can write phrases like “convert the grayscale clip into a colorful masterpiece” and “repair the video with missing parts.” This approach allows us to effortlessly obtain video triplets.

Video Dehazing and Deblurring For enhancement tasks like video denoising and dehazing, we can utilize commonly used datasets [43, 65] in this domain, all of which are annotated with corresponding input and ground-truth data. Therefore, we only need to write a few instruction phrases manually, such as “remove the applied haze from this video” and “enhance the clarity of this blurry video.”

Language-guided Video Object Segmentation For the language-guided video object segmentation task, the goal is to identify and segment objects within the video based

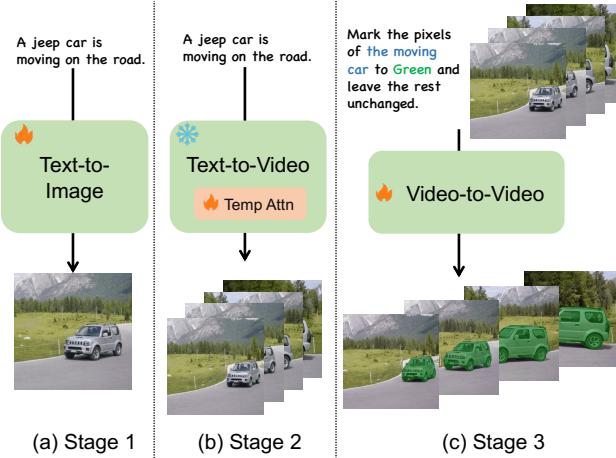


Figure 2. The overview of the training stage, which shows how to transfer a T2I model for V2V translation tasks. (a) Text-to-Image stage, (b) Text-to-Video stage, (c) Video-to-Video stage.

on natural language instructions. We utilize established datasets specified for this task [9, 22, 46] for training. As for the instruction, we can manually craft phrases such as “change the {object} pixels to {color}, while keeping the other pixels constant.” These kinds of instructions can yield superior visual results, as validated in [13].

Instruction-guided Video Editing For most diffusion-based video editing methods, detailed textual descriptions of both the source and target videos are required. Additionally, operations such as one-shot tuning [57] during training and DDIM Inversion [47] at the inference stage are time-consuming and resource-intensive. For a short video, this process requires several minutes, limiting its practicality. In contrast, our approach only requires the original video and editing instructions during the inference stage, enabling video editing within seconds. Nevertheless, constructing video editing datasets is challenging. We follow [6, 40] to utilize GPT [36] and the excellent video editing models [10, 57, 63, 67] to create triplet training data.

3.4. Unified Instructional Model for Video Tasks

In this subsection, we present how to design a unified instructional model to handle various video tasks and discuss how to transfer a pre-trained T2I model for general video translation tasks.

Architecture A common T2I model [44] contains a modified U-Net [45] comprising 4 downsample/upsample blocks, and 1 middle block. Each block typically consists of spatial 2D convolutional layers, self-attention layers, and cross-attention layers with the text condition. To cope with video inputs, we inflate 2D convolutional layers into 3D convolutions [19]. Additionally, we add a vanilla temporal attention [3, 5, 62] layer for motion modeling. Before

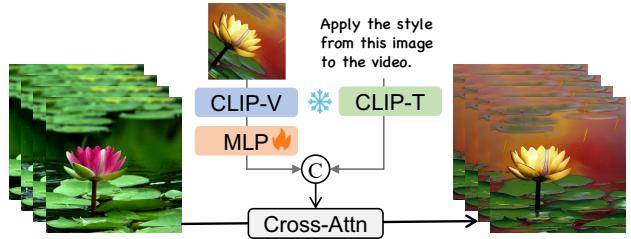


Figure 3. The overview of our multi-modal conditional method. During the training stage, we freeze the CLIP-Text and CLIP-Vision Encoder and only finetune the parameters of the MLP.

passing a video with f frames $[b, c, f, h, w]$ to the temporal module, we reshape it into $[(b \cdot h \cdot w), f, c]$. To seamlessly integrate the temporal module into the training process without causing any adverse effects, we adopt a zero initialization approach for the output projection layer of the temporal transformer following [14, 63, 71].

Training Stage As shown in Fig. 2, we design a multi-stage training method to transfer a T2I model for V2V translation. The first stage is exactly the original T2I [44] model training. In the second stage, we introduce the aforementioned temporal attention layer and inflate the U-Net from 2D to 3D. By fixing the parameters of the original T2I model, we tune the temporal module to achieve T2V generation with a video-text dataset [4]. Leveraging pre-training from the previous stage, the model learns temporal motion modeling well. In the final stage, we fine-tune the pre-trained network using the collected datasets to accomplish the video-to-video translation task.

Multi-Modal Condition Injection Mechanism Most previous video editing methods rely on provided textual descriptions or specific instructions [10, 30, 57, 63, 73]. Here, we introduce a straightforward multi-modal condition injection mechanism for video editing as shown in Fig. 3. For a given textual instruction, we use the CLIP-Text [41] Encoder to extract the embedding of the text. Additionally, we aim to incorporate images as visual instructions to learn editing patterns related to image styles. During training, we randomly select a frame from the target video and apply data augmentations such as flipping, rotation and cropping to create the image instruction. This image instruction is then processed through a pre-trained CLIP-Vision [41] Encoder and a newly added MLP layer. Subsequently, we concatenate the resulting image embedding and the text embedding along the channel dimension to form a joint instruction embedding. In this training setting, the CLIP vision and text encoders remain fixed, and only the MLP layer needs training. This approach allows effective image instruction, which eliminates the professional text instruction that demands expert knowledge.

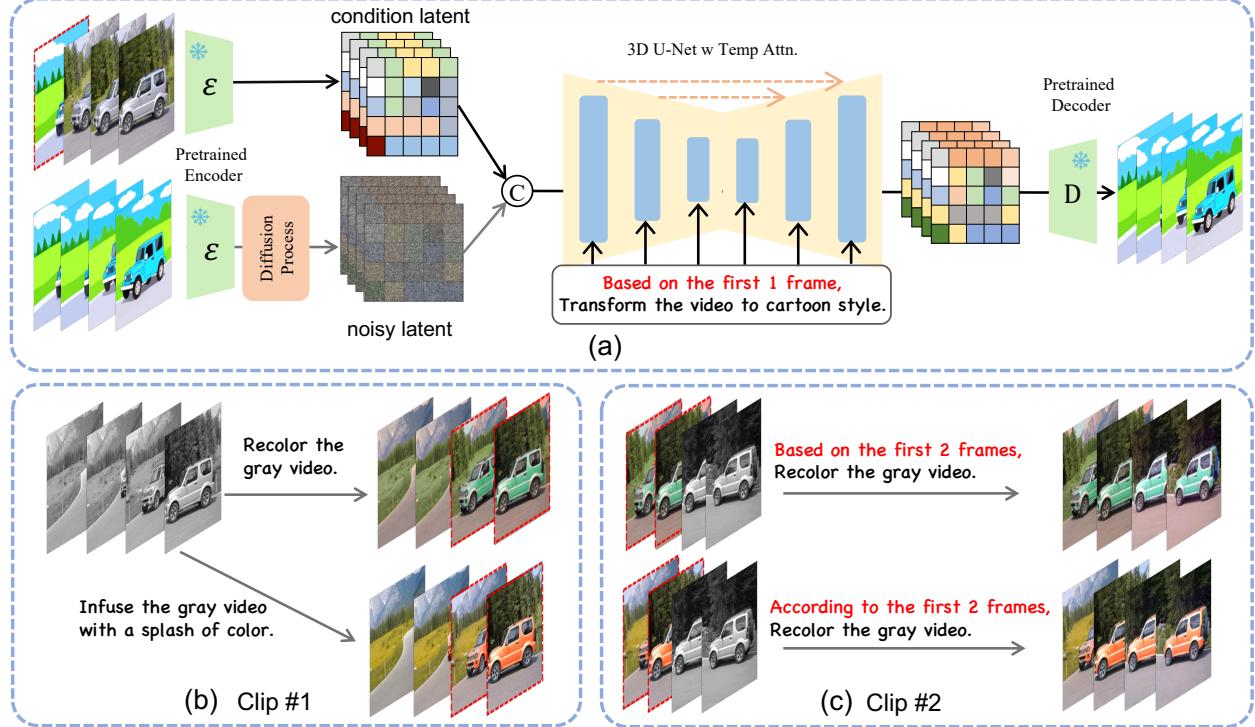


Figure 4. Pipeline of our VIDiff translation framework. We utilize the pre-trained auto-encoder as in Stable Diffusion [44] to obtain latent representation. (a) During training, we randomly decide to use the first several frames as the condition. (b) The inference of the first clip. (c) Utilize the last frames of the video as the condition, we can iteratively translate the long videos.

Training pipeline The detailed training pipeline for stage 3 is shown in Fig. 4 (a). First, the source video V_s and the target video V_t are both input into a pre-trained VAE [48] encoder to transform them to \mathbf{x}_s and \mathbf{x}_t in latent space. Subsequently, noise is added to the target video through the diffusion process. The noisy latent, along with the latent of the source video (*i.e.* condition latent), is concatenated in the channel dimension and fed into the inflated U-Net with temporal attention. During this process, the source video \mathbf{x}_s and the instruction c serve as conditions to control the denoising process to translate to the target video \mathbf{x}_t . We minimize the following latent diffusion objective:

$$\mathbb{E}_{\mathbf{x}_s, \mathbf{x}_t, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} [\|\epsilon - \epsilon_\theta([\mathbf{x}_t, \mathbf{x}_s], c, t)\|_2^2]. \quad (3)$$

In addition to translating short video clips based on instructions, our approach can also be extended to long video translation. The training process is also straightforward. When constructing training pairs, we randomly select the first n frames of the source video to align with the target video. Additionally, we modify the instruction to include phrases like “Based on the first n frames.” This allows the network to learn the correspondence between temporal modeling and the frame number n specified in the instruction. Additionally, this approach facilitates the maintenance of coherence in the translation across each clip by leveraging information from the n reference frames when computing temporal attention.

During training, we combine multiple tasks into a unified training paradigm, randomly selecting a specific task at each training step. We will demonstrate the effectiveness of this multi-task training paradigm and the integration of diverse tasks within a single model in Sec. 4.3.

Inference pipeline During inference, we employ an iterative inference approach for the long videos. As illustrated in Fig 4(b), for the first clip #1, we employ a regular inference approach. Based on the source video and instructions as conditions, the model progressively denoises from Gaussian noise to derive the target video. Once the first clip is obtained, we can use the last n frames from clip #1 as the condition. For the next clip #2, we replace the initial n frames of the source video with the corresponding frames from the preceding clip #1. Through this overlapping sampling and iterative inference method, our approach maintains consistency in the translation of videos with arbitrary lengths as shown in Fig. 4(c).

4. Experiments

4.1. Settings

Dataset Details The model training is based on the triplet of \langle source, target, instruction \rangle , which includes various video tasks mentioned above, such as video dehazing, deblurring, recoloring, inpainting, object segmentation, and video editing, *etc.* Specifically, for **dehazing and deblur-**

Table 1. Properties of different text-driven video editing methods. We compare the model type, additional control, tune time, and the application scopes. The reported tune time is from [8].

Method	ModelType	Additional Ctrl	Tune Time /Video	Need Latent Inversion	Prompt Type	Support Image-guided	Multi-Task	Support Long Video
Tune-A-Video [57]	Per-Vid-Per-Model	No	15 mins	Yes	Original&Target	No	No	No
Vid2Vid-Zero [53]	Per-Vid-Per-Model	Prompt-to-Prompt	12 mins	Yes	Original&Target	No	No	No
Video-P2P [30]	Per-Vid-Per-Model	Prompt-to-Prompt	10 mins	Yes	Original&Target	No	No	No
ControlVideo [74]	Per-Vid-Per-Model	ControlNet	15 mins	Yes	Original&Target	No	No	Yes
SimDA [63]	Per-Vid-Per-Model	No	5 mins	Yes	Original&Target	No	No	No
ReRender-A-Video [67]	One-Model-All-Vid	ControlNet	No Need	No	Original&Target	No	No	Yes
Instruct-Vid2Vid [40]	One-Model-All-Vid	No	No Need	No	Instruction	No	No	No
InsV2V [8]	One-Model-All-Vid	No	No Need	No	Instruction	No	No	Yes
VIDiff(Ours)	One-Model-All-Vid	No	No Need	No	Instruction	Yes	Yes	Yes

Table 2. Quantitative comparison with open-sourced evaluated baseline. We report both the quantitative and the user study results. The “Tuning” refers to the process of optimization. The “Inference” time includes both Inversion and Denoising times.

Method	Video-Text Alignment			Frame consistency		Runtime [min]		Support Long Video	Additional Control
	CLIPScore(\uparrow)	PickScore(\uparrow)	User Vote(\uparrow)	CLIPScore(\uparrow)	User Vote(\uparrow)	Tuning(\downarrow)	Inference(\downarrow)		
Tune-A-Video [57]	30.51	20.24	32.3%	91.21	30.5%	11.68	0.96	No	No
Vid2Vid-Zero [53]	30.28	20.09	30.1%	92.06	29.6%	4.32	2.67	No	No
ControlVideo [73]	31.03	<u>20.57</u>	36.8%	92.89	43.6%	3.75	2.75	Yes	ControlNet
Rerender-A-Video [67]	<u>31.09</u>	20.45	40.2%	90.87	37.9%	-	4.12	Yes	ControlNet
VIDiff (Ours)	31.15	20.73	-	<u>92.20</u>	-	-	0.54	Yes	No

ring tasks, we utilized the HazeWorld [65] and BSD [43] datasets, respectively. For **video editing** tasks, we followed Instruct-Pix2Pix [6], using GPT-4 [36] and advanced video editing model [53, 57, 67] to create the triplet data. Due to limitations in the generality of current video editing models, we only generated 8,000 pairs of data, mainly focusing on style and color editing tasks. For the **video object segmentation** task, we constructed the training set using the DAVIS-RVOS [22] and Refer-YoutubeVOS [46] datasets. During training, we formulated the target video as a semi-transparent mask. As for **video recoloring and inpainting** tasks, we believe any video dataset can train such networks. We employed the datasets mentioned above as well as part data from WebVid [4] for training these tasks.

Implementation Details We use Stable Diffusion v1.5 [44] as initialization to leverage the text-to-image generation prior. Additionally, we utilize the motion module from AnimateDiff [14] to initialize the temporal layer for better temporal modeling. The learning rate during training is set to $1e - 4$. The input video frames consist of 16 frames with a resolution of 256×256 . We validate that even with low and fixed resolution during training, our approach can easily be extended to arbitrary resolutions and aspect ratios during inference. Once the training is completed, our method can be effortlessly applied to various video translation tasks. For each training step, we randomly select a task. In this way, we only need to train a unified model. Due to the mutual learning of multiple tasks, we confirm the effectiveness of the unified model. We employ classifier-free diffusion guidance [17] and introduce two guidance scales, s_V and s_T . These scales can be adjusted to balance the degree to which

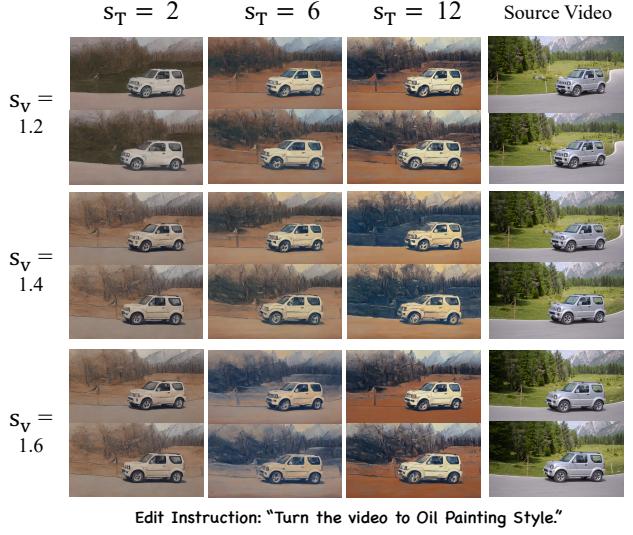


Figure 5. Classifier-free guidance weights over two conditional inputs. s_V controls similarity with the input video, while s_T controls consistency with the edit instruction.

the generated samples align with the input video and the extent to which they adhere to the editing instructions. We show the effects of these two parameters on generated samples in Fig. 5.

4.2. Experimental Results

In this section, we will compare our model with baseline methods across multiple different tasks.

Video Editing We first list several common video editing baseline methods in Table 1, comparing their various

Table 3. Quantitative results on the evaluation datasets from different methods. The best items and second best items are highlighted in bold and underlined respectively.

Method	FID(\downarrow)	Color(\uparrow)	PSNR(\uparrow)	SSIM(\uparrow)	LPIPS(\downarrow)	CDC(\downarrow)
AutoColor [24]	83.05	14.14	<u>24.41</u>	0.915	0.264	0.003734
Deoldify [1]	76.21	25.47	23.99	0.885	0.306	0.004901
DeepExemplar [68]	77.26	28.82	21.78	0.846	0.325	0.004006
DeepRemaster [20]	97.54	25.66	21.95	0.848	0.354	0.005098
TCVC [32]	74.94	21.72	25.17	<u>0.921</u>	0.239	0.003649
VCGAN [75]	70.29	15.89	23.90	0.910	0.247	0.005303
ColorDiffuser [29]	<u>69.51</u>	<u>29.13</u>	23.73	0.939	<u>0.213</u>	0.003607
VIDiff (Ours)	63.96	32.84	23.59	0.895	0.196	0.003994

attributes. It can be seen that most of these methods are based on one-shot tuning, meaning they require training a dedicated model for the specific video to be edited. This not only requires additional tuning time but also detailed textual descriptions of both the source and target videos, greatly limiting their generality. Our approach is similar to [8, 40] in that it does not require additional tuning. Additionally, our method supports multi-modal instructions, various tasks, and the editing of long videos.

We also follow the previous benchmark methods, replicating several open-source video editing techniques for comparison. We report CLIP Score [41], PickScore [23], and Frame Consistency between different frames following [58]. Additionally, we conducted a user study in which participants were presented with two sets of reports: one from our method and one from other methods. They were asked to choose the one with better matching of text and video as well as smoother video continuity. The experimental results are shown in Table 2. We also report the tuning and inference time on a single NVIDIA A100 GPU.

Video Re-colorization Regarding the video recoloring task, we conduct the experiment on a widely used benchmark. We followed previous studies and validated our approach on the validation set of the DAVIS [38] dataset. Evaluating video recoloring tasks generally involves assessing perceptual realism, color vividness, and temporal consistency. To evaluate the perceptual realism of colorized videos, we used the FID [37] (Fréchet Inception Distance) metric, which measures the similarity between the predicted colors and the ground truth distribution. To assess color vividness, we employed the Colorfulness [15] metric. Additionally, to evaluate temporal consistency, we utilized the Color Distribution Consistency [31](CDC) index. Furthermore, we also reported metrics such as PSNR [54], SSIM [55], and LPIPS [72] in our study. We compare our method with several automatic video colorization techniques as well as exemplar-based video colorization baselines. The quantitative results are presented in Table 3, demonstrating significant improvements in our approach based on perceptual evaluation metrics. Moreover, our method maintains a comparable performance in structural metrics.

Video Enhancement We evaluate the performance of our model on video enhancement tasks across several common benchmarks. We follow the evaluation method as [43, 65, 76], which includes evaluation datasets commonly used in this field such as the test set of BSD [43], Youtube [46], and DAVIS [38], among others. The quantitative results are presented in Table 4. We not only report the distortion metric PSNR [54] to measure the difference between the edited frame and the ground truth, but also follow the method described in [42] to calculate some aesthetic perceptual image quality metrics such as FID [37], LPIPS [72], and NIQE [34]. We compare our method with several mainstream open-source instructive editing techniques. It can be observed that our method outperforms others significantly across all metrics. Lastly, the performance of our model in image enhancement is constrained by the VAE [48] model, which introduces information loss. Therefore, we also report the results of VAE reconstructing the original image and compare it with the ground truth. This serves as an upper baseline, allowing us to measure the upper limit that methods based on the LDM [44] can achieve.

Visualizations Here we provide more visualization results for the VIDiff method of video translation. We present results of the video re-colorization task in Fig. 8, video dehazing and video in-paint in Fig. 9, video deblurring and language-guided video object segmentation in Fig. 10. We also show the multi-modal instruction guided video editing in Fig. 11. For fully rendered videos, we primarily refer the reader to our project page (<https://chenhsing.github.io/VIDiff>).

4.3. Ablation Study

The Effectiveness of Multi-task Training Presently, multi-task learning has become increasingly popular. It not only allows a single model to handle multiple related tasks but also enables the model to achieve better generalization performance. We conduct experiments to compare our multi-task learning model with individually trained single-task models. The performance differences are reported in Fig. 7. This comparison was made across four task-specific test datasets, demonstrating that our jointly trained model outperforms the specialized models. Clearly, the model trained jointly performs better. Additionally, we observed that this advantage extends to the field of video editing. We will present more qualitative comparison results in the supplementary materials.

The Benefit of Multi-Stage Transferring Learning As we know, most video editing methods rely on transferring pre-trained T2I model Stable Diffusion [44]. However, approaches like Instruct-Vid2Vid [40] and InsV2V [8] directly transfer T2I to V2V. The original model lacks motion information, resulting in poor temporal modeling. Fine-tuning makes the model focus more on temporal modeling, ne-

Table 4. Quantitative results on video enhancement. We also report an upper baseline, obtained by reconstructing the ground truth images using VAE, representing the performance upper bound achievable with the used VAE model.

Method	Deblurring				Dehazing				In-painting			
	PSNR(\uparrow)	FID(\downarrow)	LPIPS(\downarrow)	NIQE(\downarrow)	PSNR(\uparrow)	FID(\downarrow)	LPIPS(\downarrow)	NIQE(\downarrow)	PSNR(\uparrow)	FID(\downarrow)	LPIPS(\downarrow)	NIQE(\downarrow)
Instruct-Pix2Pix [6]	19.51	61.71	0.4260	12.65	14.24	89.29	0.6375	14.05	15.44	82.07	0.3313	24.94
MagicBrush [70]	23.45	38.05	0.2977	14.16	15.13	27.60	0.5314	14.20	16.15	56.14	0.2659	25.74
Instruct Diffusion [13]	25.62	14.05	0.1775	14.66	16.72	32.73	0.5188	13.97	20.85	38.01	0.2015	31.37
VIDiff (Ours)	27.68	14.17	0.1633	10.12	20.68	19.55	0.1319	10.42	23.17	19.21	0.1314	24.33
VAE Recon (Upper)	29.15	4.782	0.0587	13.25	27.34	6.313	0.0908	10.79	25.54	8.981	0.0847	26.54

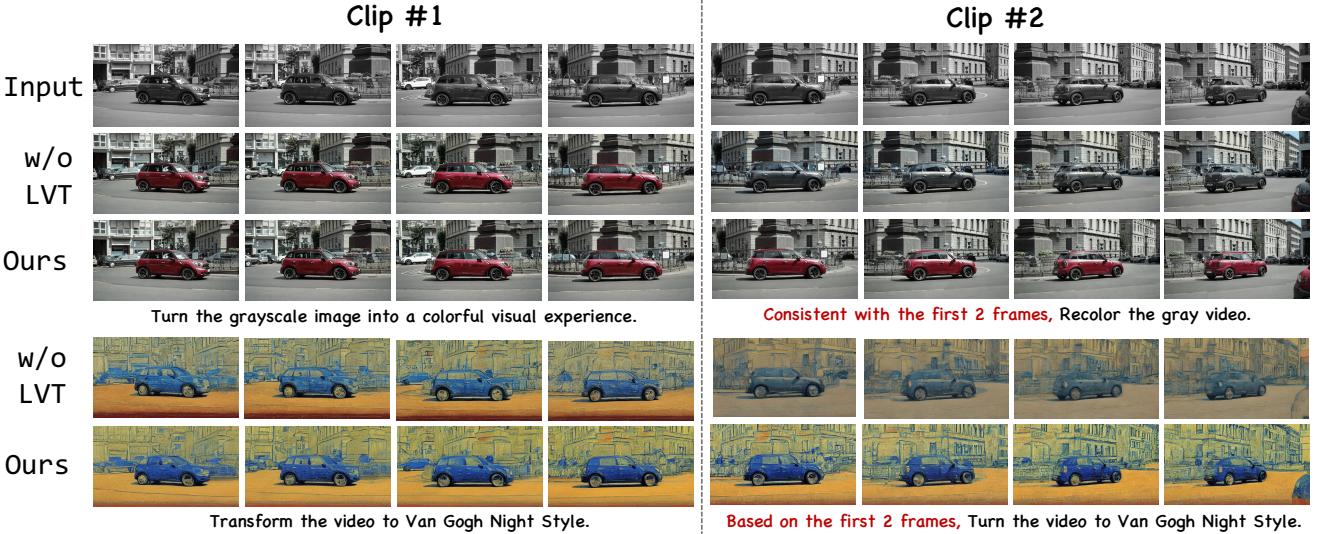


Figure 6. Ablation of our auto-regressive long video translation. VIDiff could maintain temporal consistency across different video clips.

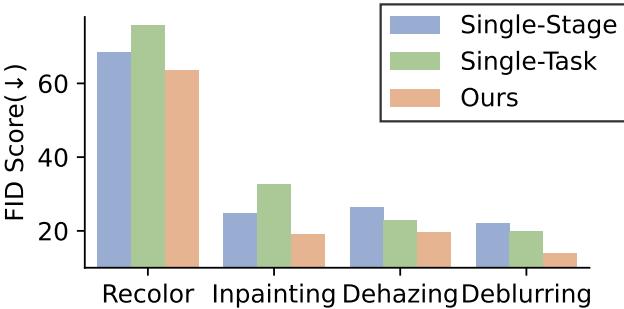


Figure 7. Ablation study on multi-task training and multi-stage transferring method. We evaluate our models on four tasks. It demonstrates that joint multi-task training and multi-stage transferring significantly enhance the performance of each task.

glecting the spatial transfer inherent in the model. We employed a multi-stage training approach that effectively mitigates this issue. Our ablation experiments, as illustrated in Fig. 7, demonstrated that the multi-stage transfer method yields significantly better image quality compared to direct fine-tuning.

The Effectiveness of Long Video Translation For tasks such as video re-colorization and video editing, ensuring consistency in long videos is a crucial challenge. Methods based on diffusion are primarily trained on short video

clips [57, 63, 74], and therefore, they can only edit relatively short video clips. The auto-regressive long video translation paradigm we propose effectively addresses this issue. In our comparative experiment shown in Fig. 6, it can be observed that a given long video, is typically divided into different clips. Without the Long Video Translation (LVT) constraints, these different clips are entirely inconsistent with each other. With our approach, the method can maintain excellent consistency between different clips in long videos.

5. Conclusion

In conclusion, this paper introduced Video Instruction Diffusion (VIDiff), a novel unified framework for aligning video tasks with human instructions. VIDiff treated various video understanding tasks as conditional video translation problems, allowing us to translate videos into desired outcomes based on instructions. We demonstrated the effectiveness of our approach across multiple tasks, with joint training enhancing the generalization capabilities of the model. This research marked a significant step in constructing a universal modeling interface for video tasks, paving the way for future advancements in the pursuit of artificial general intelligence in video understanding. In future work, we plan to further explore the performance and capabilities

of VIDiff, considering potential integration with large language models to enable more versatile unified video tasks such as video question answers and video contextual understanding.

References

- [1] J Antic. Deoldify—a deep learning based project for colorizing and restoring old images (and video!), 2019. [7](#)
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. [2](#)
- [3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. [4](#)
- [4] Max Bain, Arsha Nagrani, Gü̈l Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. [4, 6](#)
- [5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. [4](#)
- [6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. [2, 3, 4, 6, 8](#)
- [7] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videomm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023. [2](#)
- [8] Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset. 2023. [2, 3, 6, 7](#)
- [9] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2694–2703, 2023. [4](#)
- [10] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccredit: Creative and controllable video editing via diffusion models. *arXiv:2309.16496*, 2023. [4](#)
- [11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. [3](#)
- [12] Yulu Gan, Sungwoo Park, Alexander Schubert, Anthony Philippakis, and Ahmed M Alaa. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. *arXiv preprint arXiv:2310.00390*, 2023. [2](#)
- [13] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, et al. Instructdiffusion: A generalist modeling interface for vision tasks. *arXiv preprint arXiv:2309.03895*, 2023. [2, 3, 4, 8](#)
- [14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahu Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv:2307.04725*, 2023. [4, 6](#)
- [15] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, pages 87–95. SPIE, 2003. [7](#)
- [16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. [2, 3](#)
- [17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS*, 2021. [6](#)
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. [3](#)
- [19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. [4](#)
- [20] Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019. [7](#)
- [21] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. [2](#)
- [22] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 123–141. Springer, 2019. [4, 6](#)
- [23] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiania, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023. [7](#)
- [24] Chenyang Lei and Qifeng Chen. Fully automatic video colorization with self-regularization and diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3753–3761, 2019. [7](#)
- [25] Bei Li, Yi Jing, Xu Tan, Zhen Xing, Tong Xiao, and Jingbo Zhu. Transformer: Slow-fast transformer for machine translation. *arXiv preprint arXiv:2305.16982*, 2023. [3](#)
- [26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. [2](#)
- [27] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. *arXiv preprint arXiv:2303.16058*, 2023. [2](#)
- [28] Wenwu Li, Gang Pan, Chen Wang, Zhen Xing, and Zhenjun Han. From coarse to fine: Hierarchical structure-aware video summarization. *ACM TOMM*, 2022. [3](#)
- [29] Hanyuan Liu, Minshan Xie, Jinbo Xing, Chengze Li, and Tien-Tsin Wong. Video colorization with pre-trained text-to-image diffusion models. *arXiv:2306.01732*, 2023. [7](#)
- [30] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv:2303.04761*, 2023. [3, 4, 6](#)

- [31] Yihao Liu, Hengyuan Zhao, Kelvin CK Chan, Xintao Wang, Chen Change Loy, Yu Qiao, and Chao Dong. Temporally consistent video colorization with deep feature propagation and self-regularization learning. *arXiv preprint arXiv:2110.04562*, 2021. 7
- [32] Yihao Liu, Hengyuan Zhao, Kelvin CK Chan, Xintao Wang, Chen Change Loy, Yu Qiao, and Chao Dong. Temporally consistent video colorization with deep feature propagation and self-regularization learning. *arXiv preprint arXiv:2110.04562*, 2021. 7
- [33] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 2
- [34] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 7
- [35] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2022. 3
- [36] OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 2, 4, 6
- [37] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, pages 11410–11420, 2022. 7
- [38] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 7
- [39] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *ICCV*, 2023. 3
- [40] Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yuetong Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. *arXiv:2305.12328*, 2023. 2, 3, 4, 6, 7
- [41] 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 *ICML*, 2021. 2, 3, 4, 7
- [42] Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, and Peyman Milanfar. Multiscale structure guided diffusion for image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10721–10733, 2023. 7
- [43] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 184–201. Springer, 2020. 3, 6, 7
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4, 5, 6, 7
- [45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4
- [46] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 4, 6, 7
- [47] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 3, 4
- [48] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017. 3, 5, 7
- [49] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Lu-wei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnidvl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710, 2022. 2
- [50] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Xiyang Dai, Lu Yuan, and Yu-Gang Jiang. Omnitracker: Unifying object tracking by tracking-with-detection. *arXiv preprint arXiv:2303.12079*, 2023. 2
- [51] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Chuannxin Tang, Xiyang Dai, Yucheng Zhao, Yujia Xie, Lu Yuan, and Yu-Gang Jiang. Look before you match: Instance understanding matters in video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2268–2278, 2023. 3
- [52] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjork, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singh, Subhajit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2
- [53] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv:2303.17599*, 2023. 3, 6
- [54] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 7
- [55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 7
- [56] Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, and Yu-Gang Jiang. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *International Conference on Machine Learning*, pages 36978–36989. PMLR, 2023. 2
- [57] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023. 3, 4, 6, 8
- [58] Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jin-bin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei Huang, Yuanxi Sun, et al. Cvpr 2023 text guided video editing competition. *arXiv preprint arXiv:2310.16003*, 2023. 7

- [59] Zuxuan Wu, Zejia Weng, Wujian Peng, Xitong Yang, Ang Li, Larry S Davis, and Yu-Gang Jiang. Building an open-vocabulary video clip model with better architectures, optimization and data. *arXiv preprint arXiv:2310.05010*, 2023. 2
- [60] Zhen Xing, Yijiang Chen, Zhixin Ling, Xiangdong Zhou, and Yu Xiang. Few-shot single-view 3d reconstruction with memory prior contrastive network. In *ECCV*, 2022. 2
- [61] Zhen Xing, Hengduo Li, Zuxuan Wu, and Yu-Gang Jiang. Semi-supervised single-view 3d reconstruction via prototype shape priors. In *ECCV*, 2022. 2
- [62] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. In *CVPR*, 2023. 4
- [63] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. *arXiv:2308.09710*, 2023. 2, 3, 4, 6, 8
- [64] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*, 2023. 2
- [65] Jiaqi Xu, Xiaowei Hu, Lei Zhu, Qi Dou, Jifeng Dai, Yu Qiao, and Pheng-Ann Heng. Video dehazing via a multi-range temporal alignment network with physical prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18053–18062, 2023. 3, 6, 7
- [66] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *European Conference on Computer Vision*, pages 733–751. Springer, 2022. 2
- [67] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia*, 2023. 4, 6
- [68] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8052–8061, 2019. 7
- [69] Hui Zhang, Zuxuan Wu, Zhen Xing, Jie Shao, and Yu-Gang Jiang. Adadiff: Adaptive step selection for fast diffusion. *arXiv preprint arXiv:2311.14768*, 2023. 2
- [70] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *arXiv preprint arXiv:2306.10012*, 2023. 8
- [71] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 4
- [72] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7
- [73] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv:2305.13077*, 2023. 4, 6
- [74] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv:2305.17098*, 2023. 6, 8
- [75] 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*, 2022. 7
- [76] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10477–10486, 2023. 3, 7



Instruction: Convert the grayscale clip into a colorful masterpiece



Instruction: Introduce a range of colors to the gray video

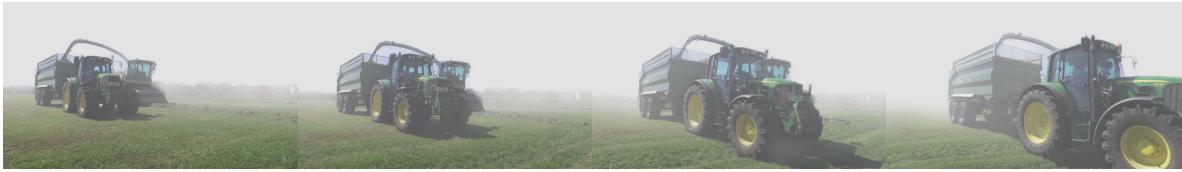


Instruction: Transform the gray video into color



Instruction: Give life to the gray video with beautiful colors

Figure 8. Results of extending our VIDiff to video re-colorization task.



Instruction: Remove the applied haze from this video



Instruction: Clear the haze from this video



Instruction: Apply inpainting algorithms to recover the missing video



Instruction: Restore the missing video content

Figure 9. Results of extending our VIDiff to video dehazing and in-painting task.



Instruction: For the white dog with gray patches, set its pixels to Green and let the others remain the same



Instruction: Mark the pixels of the girl riding the horse in Red and leave the rest unchanged



Instruction: Enhance the clarity of this blurry video



Instruction: Improve the quality of this fuzzy video

Figure 10. Results of extending our VIDiff to language guided video object segmentation and deblurring task.



Input Source Video



Instruction: Turn the video to Sketch Style.



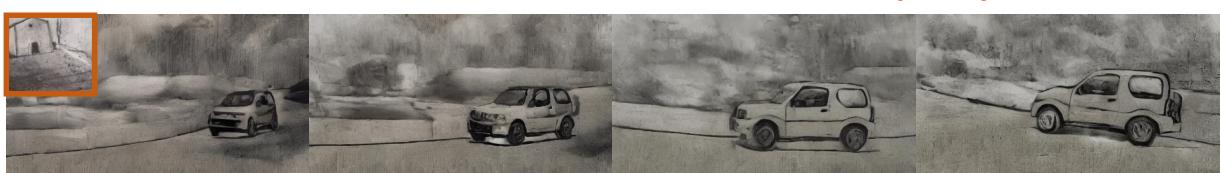
Instruction: Transform the video to Watercolor Style.



Instruction: Transform the video to Oil Painting Style.



Instruction: Adjust the video to match the style of the target image.



Instruction: Apply the style from this image to the video.



Instruction: Edit the video to reflect the style of the target image.

Figure 11. Results of extending our VIDiff to video editing task with both single-modal and multi-modal instructions.