InstructVideo: Instructing Video Diffusion Models with Human Feedback

Hangjie Yuan[1*] Shiwei Zhang[2] Xiang Wang[2] Yujie Wei[2] Tao Feng[3]
Yining Pan[4] Yingya Zhang[2] Ziwei Liu[5] Samuel Albanie[6] Dong Ni[1†]

[1]Zhejiang University [2]Alibaba Group [3]Tsinghua University
[4]Singapore University of Technology and Design [5]S-Lab, Nanyang Technological University
[6]CAML Lab, University of Cambridge

Abstract

Diffusion models have emerged as the de facto paradigm for video generation. However, their reliance on web-scale data of varied quality often yields results that are visually unappealing and misaligned with the textual prompts. To tackle this problem, we propose `InstructVideo` *to instruct text-to-video diffusion models with human feedback by reward fine-tuning.* `InstructVideo` *has two key ingredients: **1)** To ameliorate the cost of reward fine-tuning induced by generating through the full DDIM sampling chain, we recast reward fine-tuning as editing. By leveraging the diffusion process to corrupt a sampled video,* `InstructVideo` *requires only partial inference of the DDIM sampling chain, reducing fine-tuning cost while improving fine-tuning efficiency. **2)** To mitigate the absence of a dedicated video reward model for human preferences, we repurpose established image reward models, e.g., HPSv2. To this end, we propose Segmental Video Reward, a mechanism to provide reward signals based on segmental sparse sampling, and Temporally Attenuated Reward, a method that mitigates temporal modeling degradation during fine-tuning. Extensive experiments, both qualitative and quantitative, validate the practicality and efficacy of using image reward models in* `InstructVideo`*, significantly enhancing the visual quality of generated videos without compromising generalization capabilities. Code and models can be accessed through our project page* `https://instructvideo.github.io/`.

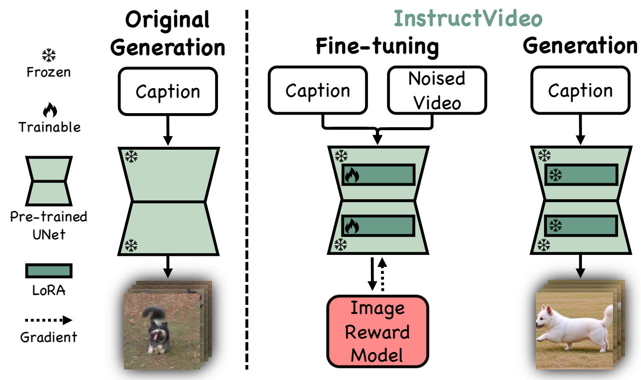

Figure 1. **Overview of the `InstructVideo` framework**. Our method performs efficient fine-tuning on sampled video-text pairs, instructed by human preferences in image reward models.

1. Introduction

The emergence of diffusion models [32, 68, 70, 98] has significantly boosted generation quality across a wide range of media content [34, 40, 62, 76]. This generation paradigm has shown promise for video generation [5, 33, 34, 76, 79], despite the challenges of working with high-dimensional data. While diffusion models are one factor driving progress, the scaling of training datasets has also played a key role [66, 81]. However, despite recent progress, the visual quality of generated videos still leaves room for improvement [76, 85]. A significant contributing factor to this issue is the varying quality of web-scale data employed during pre-training [3, 65], which can yield models capable of generating content that is visually unappealing, toxic and misaligned with the prompt.

While *aligning model outputs with human preferences* has proven highly effective for control [13], text generation [2, 43, 55, 56, 73] and image generation [42, 87, 92], it remains a notion unexplored in video diffusion models. The most widely-adopted methods for aligning models with human preferences include off-line reinforcement learning (RL) [4, 44, 56] and direct reward back-propagation [16, 58]. Typically, this entails training a reward model on manually annotated datasets that is then subsequently used to fine-tune the pre-trained generative model.

Two major challenges arise when seeking to align video generation models with human preferences: **1)** The optimization process for optimizing human preferences is computationally demanding, often requiring video generation from textual inputs. While video generation pre-training us-

*Work conducted during his research internships at DAMO Academy.
†Corresponding author.

ing DDPM [32] requires only a single-step inference for every iteration, reward optimization requires a 50-step DDIM inference [69]. **2)** The curation of a large annotated dataset to capture human preferences of videos is labor-intensive, while the computation- and memory-intensive demands of utilizing ViT-H [86] or ViT-L [92]-based computational alternatives to evaluate the entire video are high.

To surmount these mentioned challenges, we propose `InstructVideo`, a model that efficiently instructs text-to-video diffusion models to follow human feedback, as illustrated in Fig. 1. Regarding the first challenge of the demanding reward fine-tuning process caused by generating through the full DDIM sampling chain, we *recast the problem of reward fine-tuning as an editing procedure*. This reformulation requires only partial inference of the DDIM sampling chain, thereby reducing computational demands while improving fine-tuning efficiency. Drawing inspiration from established editing workflows in diffusion models [6, 7, 17, 26, 49, 52, 54, 59, 101], where primary visual content is initially corrupted with noise and then reshaped by a target prompt, our method focuses on refining coarse and structural videos into more detailed and nuanced outputs. This contrasts with previous methods [4, 16, 42] that generate results directly from text. Such blurry and structural videos, serving as the starting point for reward fine-tuning, are procured by a simple diffusion process with negligible cost. During generation, the optimized model retains the capability to produce videos directly from texts. In conjunction with back-propagation truncation of the sampling chain, we make reward fine-tuning on text-to-video diffusion models computationally attainable and effective.

Regarding the second challenge (the lack of a reward model tailored for video generation), we postulate that the visual excellence of a video is tied to both the quality of its individual frames and the fluidity of motion across consecutive frames. To this end, we resort to off-the-shelf image reward models, *e.g.*, HPSv2 [86], to ascertain frame quality. Drawing inspiration from temporal segment networks [77], we propose Segmental Video Reward (SegVR), which strategically evaluates video quality based on a subset of sparsely sampled frames. By providing sparse reward signals, SegVR offers dual benefits: it not only ameliorates computational burden but also mitigates temporal modeling collapse. On the other hand, although LoRA [36] is adopted by default to retain the capability to generate temporally smooth videos, SegVR still leads to videos with visual artifacts, such as structure twitching and color jittering. To mitigate this, we propose Temporally Attenuated Reward (TAR), which operates under the hypothesis that central frames should be assigned paramount importance, with emphasis tapering off towards peripheral frames. This strategic allocation of importance across frames ensures a more stable and visually coherent video generation process.

As part of our pioneering effort to align video diffusion models with human preferences, we conduct extensive experiments to assess the practicality and efficacy of integrating image reward models within `InstructVideo`. Our findings reveal that `InstructVideo` markedly enhances the visual quality of generated videos without sacrificing the model's generalization capabilities, setting a new precedent for future research in video generation.

2. Related Work

Video generation via diffusion models. Early efforts at video generation focused on GANs [24, 35, 47, 57, 67, 74, 84, 95] and VAEs [47, 53, 93]. However, due to the complexity of jointly modeling spatio-temporal dynamics, generating videos from texts remains an unresolved challenge. Recent methods for video generation aim to mitigate this by utilizing the de facto generation method, *i.e.*, diffusion models [32, 68, 69, 98], for generating videos with diversity and fidelity [1, 5, 8–10, 12, 19, 25, 28–30, 33, 34, 37, 48, 50, 51, 60, 64, 72, 78, 80–82, 88–90, 94, 96, 97, 99, 102–104] and scaling up the pre-training data or model architecture [33, 35, 66, 81, 83]. VDM [34] represents a pioneering work that extended image diffusion models to video generation. Owing to the computation-intensive nature of diffusion models, follow-up research sought to reduce overhead by leveraging the latent space [62], *e.g.*, ModelScopeT2V [76], Video LDM [5], MagicVideo [103] and SimDA [90], *etc.*. To enable more controllable generation, further efforts introduce spatio-temporal conditions [11, 19, 41, 79, 88, 94], *e.g.*, VideoComposer [79], Gen-1 [19], DragNUWA [94], *etc.*. However, generating videos that adhere to human preferences remains a challenge.

Human preference model. Understanding human preference in visual content generation remains challenging [39, 42, 45, 86, 87, 92]. Some pioneering works target solving this problem by annotating a dataset with human preferences, *e.g.*, Pick-a-pic [39], ImageReward [92] and HPD [86, 87]. Language-image models, *e.g.* CLIP [61] and BLIP [46], are then fine-tuned on the resulting annotated data. As such, the fine-tuned models represent a data-driven approach to modelling human preferences. However, the annotation process for capturing human preferences is highly labor- and cost-intensive. Thus, in this paper, we adopt off-the-shelf image preference models to improve the quality of generated videos.

Learning from human feedback. Learning from human feedback was first studied in the context of reinforcement learning and agent alignment [14, 43] and later in large language models [56, 71], enabling them to generate helpful, honest and harmless textual outputs. This goal of learning from human feedback is also desirable in visual content generation. In image generation and edit-

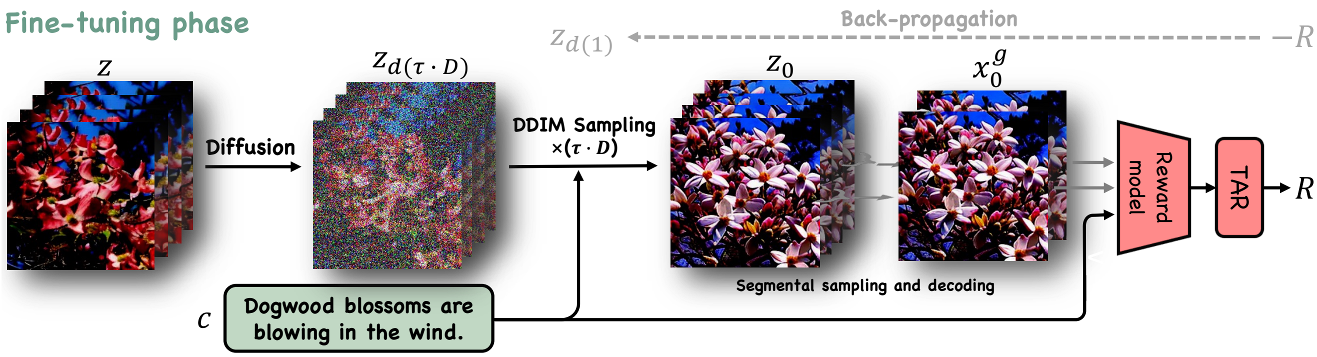

Figure 2. **The reward fine-tuning framework of `InstructVideo`.** During fine-tuning, we sample video-text pairs and apply a diffusion process to corrupt the videos to a noise level τ. Subsequently, we perform partial inference of the DDIM sampling chain to obtain the human preference edited videos. By utilizing SegVR and TAR, we can leverage image reward models to perform reward fine-tuning for video generation. The VQGAN encoder and decoder are omitted for clarity. In this example, the blurry video $\boldsymbol{z}$ is edited guided by human preferences, producing a result that highlights the vibrancy and structure of the dogwood blossoms.

ing, a key objective is to align generated images with the prompt [4, 16, 21, 58, 75, 100], preventing surprising and toxic results [42, 87, 92]. Lee et al. utilizes reward-weighted regression on a manually collected dataset, aiming to mitigate misalignment with respect to factors such as count, color and background. DDPO [4] and DPOK [21] propose to use policy gradients on a multi-step diffusion model [20], demonstrating improved rewards of aesthetic quality, image-text alignment, compressibility, *etc.*. DRaFT [16] and AlignProp [58] achieve feedback optimization by back-propagating the gradients of a differentiable reward function through the sampling procedure via gradient checkpointing [27]. However, learning from human feedback for video diffusion models remains under-explored owing to its prohibitive cost. `InstructVideo` aims to fill this gap, providing a solution for more efficient reward fine-tuning.

3. Methodology

In this section, we commence with preliminaries. Next, we delve into the details of `InstructVideo`, which encompasses: **1)** A reformulation of reward fine-tuning as editing that ensures computational efficiency and efficacy. **2)** Segmental Video Reward (SegVR) and Temporally Attenuated Reward (TAR) that enable efficient reward fine-tuning with image reward models.

3.1. Preliminaries

Text-to-video diffusion models. Text-to-video diffusion models aim to map textual input into a distribution representing video data via a reverse diffusion process [32, 68]. These models typically operate in a latent space to handle complex video data [62, 76, 79]. During pre-training, a sampled video $\boldsymbol{x}$ is processed by a fixed encoder [18] to derive its latent representation $\boldsymbol{z} \in \mathbb{R}^{F \times h \times w \times 3}$, where the video's spatial dimensions are compressed by a factor of 8. Next, random noise is injected into the sampled video by the forward diffusion process according to a predetermined schedule $\{\beta_t\}_{t=1}^{T}$. This process can be described as $\boldsymbol{z}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{z} + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\epsilon \in \mathcal{N}(0,1)$ is random noise with identical dimensions to $\boldsymbol{z}$, $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$ and $\alpha_t = 1 - \beta_t$. A UNet [15, 63] ϵ_θ is adopted to perform denoising, enabling the generation of videos through a reverse diffusion process, conditioned on the video caption $\boldsymbol{c}$. Optimisation employs the following reweighted variational bound [32]:

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{z},\epsilon,\boldsymbol{c},t}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{z} + \sqrt{1-\bar{\alpha}_t}\epsilon, \boldsymbol{c}, t)\|_2^2\right] \quad (1)$$

During inference, we adopt the DDIM sampling [69] method for realistic video generation.

Reward fine-tuning. Reward fine-tuning aims to optimize a pre-trained model to enhance the expected rewards of a reward function $r(\cdot,\cdot)$. In our case, we target optimizing the parameters θ of a text-to-video diffusion model to enhance the expected reward of the generated videos given the text distribution:

$$\mathcal{L}_r(\theta) = \mathbb{E}_{\mathbb{P}(\boldsymbol{c})}\mathbb{E}_{\mathbb{P}_\theta(x_0|\boldsymbol{c})}[-r(\boldsymbol{x}_0, \boldsymbol{c})] \quad (2)$$

where $\boldsymbol{x}_0$ is the video generated from the sampled text $\boldsymbol{c}$ via the diffusion model through the DDIM sampling chain. The reward function $r(\cdot,\cdot)$ is typically a pre-trained model to assess the quality of the model output.

3.2. `InstructVideo`

In Fig. 2, we illustrate `InstructVideo`'s fine-tuning pipeline and elaborate on the technical contributions below.

3.2.1 Reward Fine-tuning as Editing

Reward fine-tuning with diffusion models is costly due to the iterative refinement process during generation using DDIM [69]. During generation, initial steps are crucial for shaping coarse, structural aspects of videos, with subsequent steps refining the coarse videos. Understanding that the essence of reward fine-tuning is not to drastically alter the model's output but to subtly adjust it in line with human preferences, we propose to reinterpret reward fine-tuning as a form of editing [17, 52, 85]. This perspective shift allows us to perform partial inference of the DDIM sampling chain, reducing computational demands and easing optimization.

To implement this idea, we first curate a small amount of fine-tuning data from pre-training data for reward fine-tuning. For each video-text pair $(\boldsymbol{x}, \boldsymbol{c})$, we acquire the video's latent embedding $\boldsymbol{z}$ as stated in Sec. 3.1. We aim to smooth out the video to eliminate undesirable artifacts and distortions [52]. To achieve this, we leverage the diffusion process rather than DDIM inversion [17, 52] to enable efficient editing. If we denote the number of DDIM steps as D and the number of pre-training DDPM steps as T, we define a mapping $d : \{1, \ldots, D\} \rightarrow \{1, \ldots, T\}$ that maps DDIM step index to the DDPM step index[1], formulated as:

$$d(i) = \frac{T}{D} \cdot (i-1) + 1 \tag{3}$$

Given the noise level τ, the targeted diffusion step t_{noi} for injecting noise is formulated as:

$$t_{\text{noi}} = d(\tau \cdot D) \tag{4}$$

This allows us to obtain the starting point for reward fine-tuning via the diffusion process:

$$\boldsymbol{z}_{t_{\text{noi}}} = \sqrt{\bar{\alpha}_{t_{\text{noi}}}}\boldsymbol{z} + \sqrt{1-\bar{\alpha}_{t_{\text{noi}}}}\epsilon \tag{5}$$

Based on $\boldsymbol{z}_{t_{\text{noi}}}$, we can perform $\tau \times D$ steps of DDIM sampling [69] along the DDIM sub-sequence to obtain the edited result $\boldsymbol{z}_0$, which consumes τ of the computation of the full sampling chain. Utilizing the decoder [18], we decode $\boldsymbol{z}_0$ in the latent space to $\boldsymbol{x}_0$ in the RGB space.

3.2.2 Reward Fine-tuning with Image Reward Models

Since curating large datasets to capture human preferences for training video reward models is prohibitively expensive, we resort to off-the-shelf image reward models $r(\cdot, \cdot)$, *e.g.*, HPSv2 [86]. HPSv2 is trained on 430k pairs of images, which are annotated by humans for text-image alignment and image quality. Given that videos are natural extensions of images, we posit that these human preferences are also applicable to videos. However, initial experiments with applying dense reward fine-tuning produced degraded motion continuity. Taking inspiration from temporal segment networks [77], given a video $\boldsymbol{x}_0 \in \mathbb{R}^{F\times H\times W\times 3}$ generated from its caption $\boldsymbol{c}$, we evenly divide it into S segments. Within each segment, we perform random frame sampling, obtaining a sparse set capturing the essence of the video $\boldsymbol{x}_0^g = \{\boldsymbol{x}_0^{g(1)}, \ldots, \boldsymbol{x}_0^{g(S)}\}$. Here, $g(i) = \text{Uniform}\left((i-1)\cdot\frac{F}{S}, i\cdot\frac{F}{S}-1\right)$ denotes a uniform sampling of index within ith segment. Utilizing $r(\cdot,\cdot)$, we compute the reward score R with respect to $\boldsymbol{x}_0$ as follows:

$$R = \text{Agg}_i[r(\boldsymbol{x}_0^{g(i)}, \boldsymbol{c})], \quad i = 1, \ldots, S \tag{6}$$

where Agg_i denotes the aggregation function along index i. To consider the impact of all frames in $\boldsymbol{x}_0^g$, an intuitive implementation of Agg_i is the mean function.

However, the simple aggregation function leads to noticeable visual artifacts in the generated videos, such as structural twitching and color jittering. This issue arises because the mean function places equal weight on all frames, disregarding the inherent dynamic nature of videos where the reward scores of frames can vary throughout the sequence. To address this, we introduce TAR that strategically emphasizes central frames, with the emphasis tapering off towards the peripheral frames, thereby avoiding uniformly optimizing all frames' reward scores to be equally high. We define the temporally attenuated coefficient as:

$$f_i = \mathrm{e}^{-\lambda_{\text{tar}}|g(i)-\frac{F}{2}|} \tag{7}$$

where λ_{tar} controls the degree of the attenuating rate. We set $\lambda_{\text{tar}} = 1$ by default. Incorporating this coefficient, we rewrite the reward score R:

$$R = \frac{1}{S}\sum\nolimits_{i=1}^{S} f_i \cdot r(\boldsymbol{x}_0^{g(i)}, \boldsymbol{c}) \tag{8}$$

The optimization objective in Eq. (2) can be rewritten as:

$$\mathcal{L}_r(\theta) = \mathbb{E}_{\mathbb{P}(\boldsymbol{c})}\mathbb{E}_{\mathbb{P}_\theta(x_0|\boldsymbol{c})}[-R] \tag{9}$$

3.3. Reward Fine-tuning and Inference

Data preparation and evaluation metric. We follow DDPO [4] to experiment on prompts describing 45 animal species. In contrast to DDPO, since `InstructVideo` relies on video-text data, we select video-text pairs as the *fine-tuning data* from the base model's pre-training dataset, *i.e.*, WebVid10M, ensuring that no extra data is introduced. It is worth noting that we **do not apply** any quality filtering method to ensure that the selected videos are of high quality. Specifically, we select about 20 video-text pairs for each animal species. To evaluate the model's ability to optimize the reward scores, we also collect *evaluation data* comprising about 6 prompts for each animal. We use HPSv2 score to measure the optimization performance of reward fine-tuning on the first frames of all segments.

Reward fine-tuning. We adopt the publicly available text-to-video diffusion model ModelScopeT2V [76] as our base model. ModelScopeT2V is trained on WebVid10M [3] with

[1] For example, if $D = 20$ and $T = 1000$, then the DDIM step subsequence is $\{1, 51, \ldots, 901, 951\}$, *i.e.*, $d(2) = 51$.

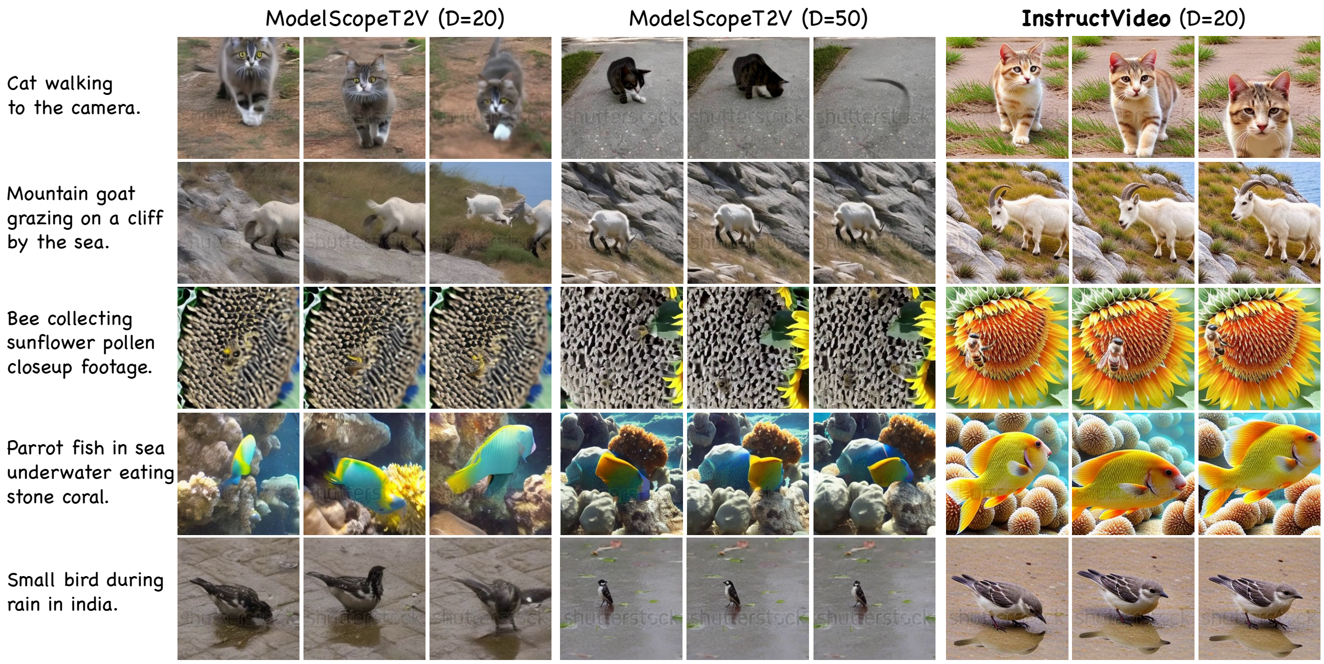

Figure 3. **Comparion of `InstructVideo` with the base model ModelScopeT2V**. ModelScopeT2V utilizes 20 and 50 DDIM steps.

$T = 1000$ and is able to generate videos of $16 \times 256 \times 256$ resolution, which we divide into $S = 4$ segments. By default, we adopt the differentiable HPSv2 [86] as the reward model and perform 20-step DDIM inference, *i.e.*, $D = 20$. Classifier-free guidance [31] is adopted by default. Directly back-propagating the reward loss to the diffusion models can be computationally intensive and risks catastrophic forgetting [22, 23, 38]. To circumvent these issues, we incorporate LoRA [36] by default. To further accelerate fine-tuning, we truncate the gradient to only back-propagate the last DDIM sampling step following [16]. Experiments are conducted on 4 NVIDIA A100s, with the batch size set to 8 and the learning rate set to 1×10^{-5}. To strike a cost-performance balance, we fine-tune `InstructVideo` with default parameters for 20k steps if not otherwise stated.

Inference. After reward fine-tuning, we merge the LoRA weights into the ModelScopeT2V parameters to ensure that `InstructVideo`'s inference cost is identical to ModelScopeT2V [36]. For text-to-video generation, `InstructVideo` uses 20-step DDIM inference.

4. Experiments

4.1. Effectiveness of `InstructVideo`

Comparison with the base model ModelScopeT2V. To verify the efficacy of `InstructVideo`, we compare it with ModelScopeT2V [76] utilizing 20 and even 50 DDIM steps in Fig. 3. Examining the examples, we observe that the quality of videos generated by `InstructVideo` consistently outperforms the base model by a margin. Specifically, notable enhancements include **1)** clearer and more coherent structures and scenes even if the animal is moving, exemplified by the walking cat and the swimming fish; **2)** more appealing coloration, exemplified by the sunflower, the bee and the mountain goat; **3)** an enhanced delineation of scene details, exemplified by the rock and grass on the cliff, and the texture of all the animals; and **4)** improved video-text alignment, exemplified by the distinct portrayal of sunflowers and the bird's reflections on the water. Remarkably, these advancements are achieved without compromising motion fluidity and the resultant videos can often surpass the video quality of the WebVid10M dataset. Notably, `InstructVideo` even attenuates watermarks present in WebVid10M. These qualitative leaps, consistently favored by human annotators, are attributed to the reward fine-tuning process, which effectively refines the video diffusion model.

Comparison with other reward fine-tuning methods. This aims to validate the efficacy of reward fine-tuning conceptualized as an editing process. We compare with other representative reward fine-tuning methods, including policy gradient algorithm, DDPO [4], reward-weighted regression, RWR [42] and direct reward back-propagation method, DRaFT [16]. For DRaFT, we adopt DRaFT-1 for efficient fine-tuning. SegVR and TAR are employed for all compared methods to standardize reward signals. The comparative analysis on the evaluation set, presented in Fig. 5(a), leads to two findings: **1)** Both RWR and DDPO exhibit a performance plateau after about 11 hours of fine-tuning, with further optimization failing to enhance or even deteriorating performance. **2)** Direct reward back-propagation methods, including `InstructVideo` and DRaFT, ini-

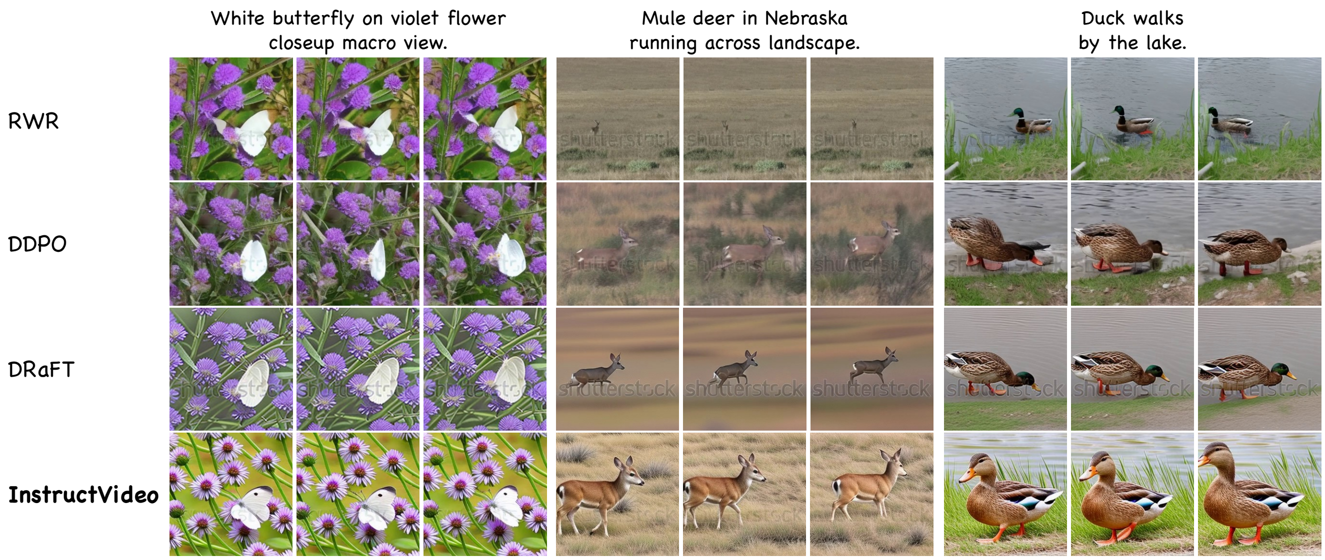

Figure 4. **Comparison of `InstructVideo` with other reward fine-tuning methods**. We set $D = 20$ for all methods.

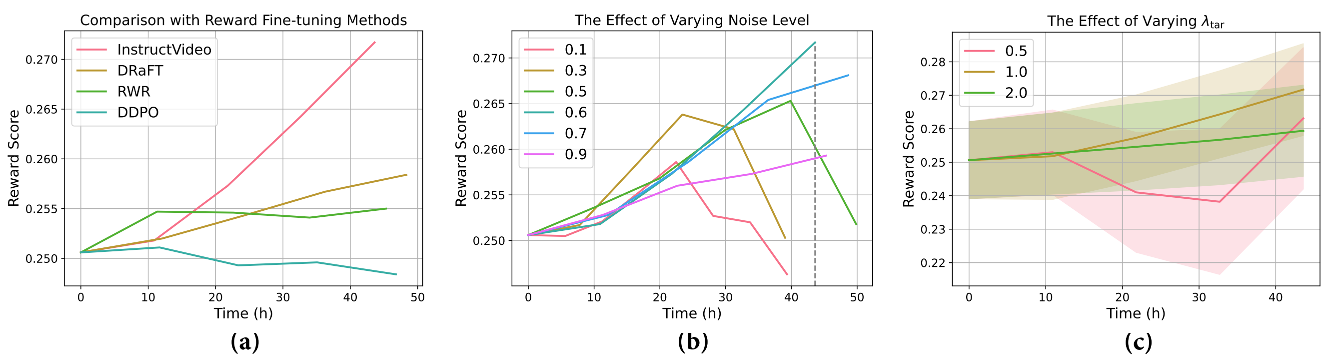

Figure 5. **(a) Comparison with other reward fine-tuning methods**. **(b) The effect of varying noise level** τ. The vertical dashed line indicates 20k steps of optimization for $\tau = 0.6$. **(c) The effect of varying** λ_{tar}.

Method	In-domain	New Animals	Non-animals
ModelScopeT2V†	0.2542 ± 0.0122	0.2541 ± 0.0109	0.2610 ± 0.0158
ModelScopeT2V	0.2506 ± 0.0155	0.2502 ± 0.0138	0.2557 ± 0.0177
DDPO [4]	0.2511 ± 0.0114	0.2524 ± 0.0112	0.2564 ± 0.0171
RWR [42]	0.2550 ± 0.0166	0.2517 ± 0.0101	0.2625 ± 0.0146
DRaFT [16]	0.2584 ± 0.0123	0.2561 ± 0.0098	0.2644 ± 0.0174
`InstructVideo`	**0.2717** ± 0.0137	**0.2645** ± 0.0125	**0.2682** ± 0.0202

Table 1. **Generalization to unseen text prompts**. † denotes the model utilizes $D = 50$ while others adopt $D = 20$. 'In-domain' denotes in-domain animal prompts from the evaluation data.

tially lag during the first 11 hours but subsequently demonstrate fine-tuning efficiency, especially `InstructVideo`. To further validate the efficacy of our method, we provide visual comparisons in Fig. 4, where we adopt the optimal fine-tuned checkpoint for each method in Fig. 5(a). The examples reflect `InstructVideo` 's superiority, evident in: **1)** the clarity and coherence of structural and scenic elements, **2)** the vibrancy of colors, **3)** the precision in depicting intricacies, and **4)** enhanced video-text alignment.

Generalization to unseen text prompts. We assessed the model's generalization capabilities using two distinct sets of prompts: **1)** those describing new animals and **2)** those related to non-animals, none of which are present in the fine-tuning data. For this purpose, we curate about 4 prompts each for 6 new animal species following [58] and 46 prompts for non-animals. Our comparative analysis of `InstructVideo`, the base model ModelScopeT2V and other reward fine-tuning methods is presented in Tab. 1. It is worth noting that during fine-tuning, SegVR and TAR are adopted by default to provide reward signals. Our observations are threefold: **1)** Increasing the number of DDIM steps enhances the video quality. **2)** Compared to the base model ModelScopeT2V in the second row, all methods improve reward scores for these unseen prompts. **3)** Among all methods, `InstructVideo` outperforms other alternatives, affirming its superior generalization capabilities. To further demonstrate this, we provide visual comparisons in Fig. 6, featuring a set of unseen animal species, various sceneries, and human figures. The presented examples display an enhanced quality and exhibit an improvement in video-text alignment.

User study. To further qualitatively compare the videos

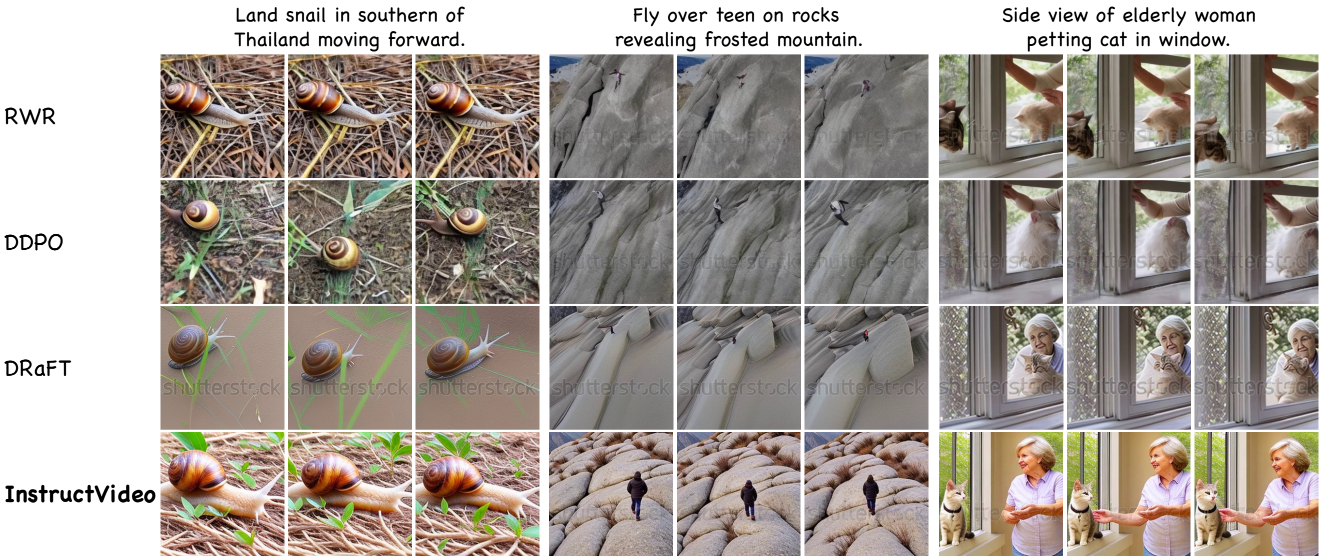

Figure 6. **Comparison of `InstructVideo`'s generalization capabilities with other methods**. We set $D = 20$ for all methods.

`InstructVideo` vs	Quality			Alignment		
	Ours	Tie	Other	Ours	Tie	Other
ModelScopeT2V	78.7%	13.2%	8.1%	29.1%	57.1%	13.8%
DRaFT	75.7%	11.6%	12.7%	29.0%	52.7%	18.3%

Table 2. **User study.** 'Tie' indicates instances where annotators think two videos are of comparable quality. 'Quality' and 'Alignment' represent video quality and video-text alignment.

generated by `InstructVideo` and other methods, we conduct a user study comparing our methods with the base model ModelScopeT2V and the reward fine-tuned model DRaFT in Tab. 2. We recruited five participants to assess the quality of videos in terms of video quality and video-text alignment. To simplify the annotation process, participants were presented with pairs of videos and asked to identify which video was superior or if both were of equal quality. To ensure a comprehensive comparison, we chose 60 prompts from the 45 fine-tuning animal species, 20 prompts from the 6 new animal species, 20 prompts describing non-animals and 200 prompts from MSR-VTT [91] (300 in total). More details are presented in the Appendix. We observe that our method consistently outperforms other methods. Specifically, improvements in video quality, a noted shortcoming of the base model, are more pronounced than improvements in video-text alignment.

4.2. Ablation Study

The effect of varying noise level τ. To determine the optimal choice for noise level τ, we vary its value and evaluate its impact on reward scores using the evaluation data. We illustrate the results in Fig. 5(b). An increase in τ from 0.1 to 0.5 correlates with a progressive enhancement in the highest reward scores achieved by `InstructVideo`. However, excessively prolonged fine-tuning precipitates a sharp decline in generative performance. This phenomenon can be attributed to the limited edited space available to the model at lower noise levels, which constricts its ability to find the optimal output space as directed by the reward scores. When we further increase τ from 0.6 to 0.9, we observe that the reward score enhancement per hour becomes minor, suggesting challenges associated with generating from an extended sampling chain. Optimally, a noise level of $\tau = 0.6$ strikes a balance, providing a feasible starting point for editing that still allows for a substantial exploration of the edited space. After 20k steps, more optimization leads to over-optimization [58], meaning that further steps can degrade the visual quality of the output despite potential increases in the reward score. Thus, we finalize on $\tau = 0.6$ with 20k steps of fine-tuning.

The effect of varying λ_{tar}. To determine the optimal choice for λ_{tar}, we vary its value and evaluate its impact. We illustrate the results in Fig. 5(c). A relatively high value λ_{tar}, such as 2.0, results in f_i decaying exponentially faster towards those border frames, thus providing diminished reward signals. A relatively low value λ_{tar}, such as 0.5, leads to f_i decaying more gently towards those border frames, thus strengthening the reward signals. This equalized weighting across frames can destabilize fine-tuning, leading to a precipitous decline in reward scores. Subsequent increases in scores do not necessarily indicate improved video quality but indicate quality degradation, as revealed by the rising variance. Thus, an appropriate coefficient to ensure stable fine-tuning is imperative and we finalize on $\lambda_{\text{tar}} = 1.0$.

Ablation on SegVR and TAR. To qualitatively verify the efficacy of SegVR and TAR, we present illustrative results in Fig. 7. Removing either SegVR or TAR results in a noticeable reduction in temporal modeling capabilities. This suggests that overly dense or excessively strong reward sig-

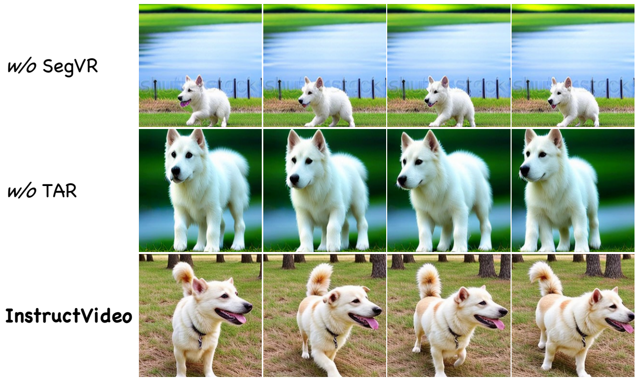

Figure 7. **Ablation study on SegVR and TAR**. This video shows a white dog walking in the park in slow motion.

nals can lead to generation collapse. The degradation of modeling temporal dynamics often leads to the degraded quality of the individual frames due to the intertwined nature of spatial and temporal parameters. These observations underscore the critical roles of SegVR and TAR in maintaining fine-tuning stability.

4.3. Further Analysis

The evolution of the generated videos during reward fine-tuning. To elucidate how the reward fine-tuning works, we present a visual progression in Fig. 8. The top row depicts a video generated without fine-tuning. All the frames in this video exhibit a lack of the dog's fur texture. Moreover, a notable blurriness characterizes the third frame due to sudden and unanticipated motion, while the fourth frame suffers from a loss of facial clarity. As reward fine-tuning proceeds, we observe a noticeable enhancement in terms of all aspects mentioned above. Surprisingly, watermarks, which are consistently present across the dataset, also gradually fade. The resultant video is full of clear details and aesthetically pleasing coloration.

Impact of fine-tuning data quality on the fine-tuning results. To investigate this, we self-collect a dataset comprising an equivalent number of video-caption pairs for 45 animal species, which are employed for fine-tuning. The results are illustrated in Fig. 9. We employ horizontal dashed lines to indicate the quality of different data, inferred from reward scores. While the variance of the generated videos are comparable for two kinds of fine-tuning data, the employment of higher-quality data, *i.e.*, WebVid10M, yields superior average reward scores compared to that obtained using the lower-quality counterpart. This suggests that superior fine-tuning data can facilitate reward fine-tuning.

Constraints of fine-tuning data on resultant video quality. Fig. 9 showcases that `InstructVideo` is capable of generating videos that achieve reward scores significantly exceeding those of the fine-tuning data itself, as denoted by the horizontal dashed lines. This observation leads us to conclude that the quality of the fine-tuning data does not

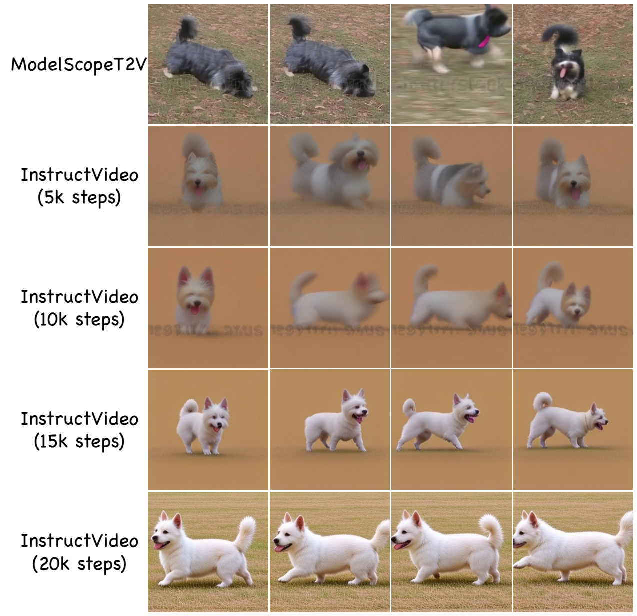

Figure 8. **The evolution of generated videos** during fine-tuning. The video shows a bobtail dog walking.

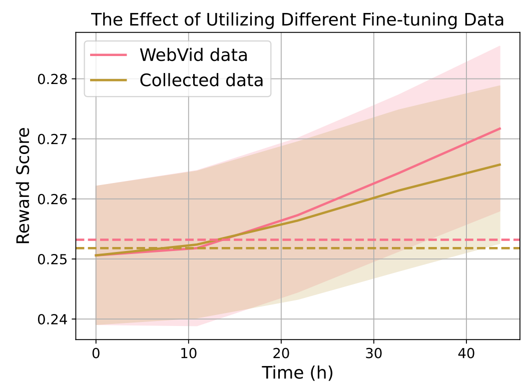

Figure 9. **The effect of utilizing different fine-tuning data.** The colored horizontal dashed lines denote the reward scores of different fine-tuning data, matched by the color of the respective curve.

impose a ceiling on the potential quality of the fine-tuned results. Our fine-tuning pipeline has the propensity to surpass the initial data quality, thus facilitating the generation of videos with substantially enhanced reward scores.

5. Conclusion

In this paper, we introduce `InstructVideo` that pioneers instructing video diffusion models with human feedback. We recast reward fine-tuning as an editing process and repurpose image reward models to provide human feedback on generated videos. Extensive experiments validate that `InstructVideo` not only elevates visual quality but also maintains robust generalization capabilities.

Acknowledgments. This research was backed by the National Natural Science Foundation of China under Grant No. 62173298 and the Alibaba Research Intern Program.

References

[1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 2

[2] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021. 1

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 1, 4

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 1, 2, 3, 4, 5, 6

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 1, 2

[6] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 2

[7] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023. 2

[8] Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, and Wenwu Zhu. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. *arXiv preprint arXiv:2311.00990*, 2023. 2

[9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023.

[10] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023. 2

[11] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models, 2023. 2

[12] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint arXiv:2310.20700*, 2023. 2

[13] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 1

[14] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 2

[15] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 3

[16] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 1, 2, 3, 5, 6

[17] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2, 4

[18] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3, 4

[19] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 2

[20] Ying Fan and Kangwook Lee. Optimizing ddpm sampling with shortcut fine-tuning. *arXiv preprint arXiv:2301.13362*, 2023. 3

[21] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *arXiv preprint arXiv:2305.16381*, 2023. 3

[22] Tao Feng, Mang Wang, and Hangjie Yuan. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9427–9436, 2022. 5

[23] Tao Feng, Hangjie Yuan, Mang Wang, Ziyuan Huang, Ang Bian, and Jianzhou Zhang. Progressive learning without forgetting. *arXiv preprint arXiv:2211.15215*, 2022. 5

[24] Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10681–10692, 2023. 2

[25] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22930–22941, 2023. 2

[26] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2

[27] Audrunas Gruslys, Rémi Munos, Ivo Danihelka, Marc Lanctot, and Alex Graves. Memory-efficient backpropagation through time. *Advances in neural information processing systems*, 29, 2016. 3

[28] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023. 2

[29] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022.

[30] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023. 2

[31] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 3

[33] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 2

[34] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 1, 2

[35] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[36] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 5

[37] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2

[38] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 5

[39] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, 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. 2

[40] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. 1

[41] Ariel Lapid, Idan Achituve, Lior Bracha, and Ethan Fetaya. Gd-vdm: Generated depth for better diffusion-based video generation. *arXiv preprint arXiv:2306.11173*, 2023. 2

[42] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 1, 2, 3, 5, 6

[43] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018. 1, 2

[44] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. 1

[45] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *arXiv preprint arXiv:2306.04717*, 2023. 2

[46] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022. 2

[47] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[48] Binhui Liu, Xin Liu, Anbo Dai, Zhiyong Zeng, Zhen Cui, and Jian Yang. Dual-stream diffusion net for text-to-video generation. *arXiv preprint arXiv:2308.08316*, 2023. 2

[49] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 2

[50] Shijie Ma, Huayi Xu, Mengjian Li, Weidong Geng, Meng Wang, and Yaxiong Wang. Optimal noise pursuit for augmenting text-to-video generation. *arXiv preprint arXiv:2311.00949*, 2023. 2

[51] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 2

[52] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 4

[53] Gaurav Mittal, Tanya Marwah, and Vineeth N Balasubramanian. Sync-draw: Automatic video generation using deep recurrent attentive architectures. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1096–1104, 2017. 2

[54] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 2

[55] OpenAI. GPT-4 technical report, 2023. 1

[56] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 1, 2

[57] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1789–1798, 2017. 2

[58] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 1, 3, 6, 7

[59] 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. *arXiv preprint arXiv:2303.09535*, 2023. 2

[60] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. *arXiv preprint arXiv:2312.04483*, 2023. 2

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3

[63] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[64] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023. 2

[65] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[66] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1, 2

[67] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3626–3636, 2022. 2

[68] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Computer Vision*, pages 2256–2265. PMLR, 2015. 1, 2, 3

[69] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3, 4

[70] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[71] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020. 2

[72] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*, 2023. 2

[73] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 1

[74] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 2

[75] Dimitri von Rütte, Elisabetta Fedele, Jonathan Thomm, and Lukas Wolf. Fabric: Personalizing diffusion models with iterative feedback. *arXiv preprint arXiv:2307.10159*, 2023. 3

[76] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1, 2, 3, 4, 5

[77] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36. Springer, 2016. 2, 4

[78] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 2

[79] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao,

and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 1, 2, 3

[80] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023. 2

[81] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 1, 2

[82] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. *arXiv preprint arXiv:2312.04433*, 2023. 2

[83] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 2

[84] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European Conference on Computer Vision*, pages 720–736. Springer, 2022. 2

[85] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 1, 4

[86] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 2, 4, 5

[87] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 1, 2, 3

[88] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *arXiv preprint arXiv:2306.00943*, 2023. 2

[89] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. 2023.

[90] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710*, 2023. 2

[91] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 7

[92] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 1, 2, 3

[93] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[94] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 2

[95] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*, 2022. 2

[96] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18456–18466, 2023. 2

[97] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation, 2023. 2

[98] Qinsheng Zhang, Molei Tao, and Yongxin Chen. gddim: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564*, 2022. 1, 2

[99] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2

[100] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. *arXiv preprint arXiv:2303.09618*, 2023. 3

[101] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2023. 2

[102] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023. 2

[103] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2

[104] Junchen Zhu, Huan Yang, Huiguo He, Wenjing Wang, Zixi Tuo, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, and Jianlong Fu. Moviefactory: Automatic movie creation from text using large generative models for language and images. *arXiv preprint arXiv:2306.07257*, 2023. 2