InstructVideo: Instructing Video Diffusion Models with Human Feedback

Supplementary Material

In this Appendix, we first detail on the potential societal impact (Appendix A), and limitations and potential future work (Appendix B). Subsequently, we provide additional details about implementing LoRA (Appendix C) and user study (Appendix D). Moreover, we provide more visualization results to demonstrate the efficacy of InstructVideo (Appendix E). Next, we provide experiments to validate the efficacy of generation with 50-step DDIM inference (Appendix F), reward fine-tuning with 50-step DDIM inference (Appendix G) and InstructVideo's adaptation to other reward functions (Appendix H). Finally, we present more ablation studies to demonstrate the necessity of SegVR and TAR (Appendix I), the extreme case of TAR (Appendix J) and more comprehensive evaluation using CLIPSIM and Temporal Consistency (Appendix K).

A. Potential Societal Impact

InstructVideo, as the pioneering effort in instructing video diffusion models with human feedback, prioritizes users' preferences for AI-generated content. We conducted this research, motivated by the varied quality of generated videos induced by the varied quality of the curated web-scale datasets. Pre-training models on such unfiltered data can lead to outputs that deviate from human preferences. In the context of the broader research community, we advocate that video generation systems, akin to other generative models like language models [7, 8], should prioritize ethical considerations and human values.

Moreover, conventional video generation systems might not always resonate with all users in terms of aesthetic style and often struggle with accurately reflecting textual prompts. InstructVideo steps in as a human-centered technology, efficiently addressing these issues in a data- and computation-efficient way and opening up possibilities for commercial applications, particularly in sectors like education and entertainment.

However, as InstructVideo primarily targets research, aiming at investigating the practicality of aligning video diffusion models with human preferences, its deployment to any circumstance beyond research should be approached with thorough oversight and evaluation to ensure responsible and ethical use.

B. Limitation and Future work

We recognize that InstructVideo, as an initial endeavor in this area, comes with its limitations. Although we validate the efficacy of image reward models, we anticipate that specialized video reward models capturing human preferences might be even more superior since they evaluate one generated video as a whole. Additionally, as a common issue mentioned in previous works [1, 2, 4, 9], reward fine-tuning carries a risk of over-optimization, meaning that excessive optimization steps will result in the degradation of the video quality despite potential increases in the reward score. Addressing these aspects presents avenues for future research, including the development of a more advanced video reward model and the design of strategic mechanisms to identify and ameliorate over-optimization.

C. More Details about Implementing LoRA

To instantiate LoRA [5] for efficient tuning, we adopt the implementation used in Diffusers[1]. Specifically, we configure the intrinsic rank within LoRA to 4 to ensure fast processing. LoRA modifications are applied to every Transformer [10] layer within our model, targeting the linear layers responsible for query, key, value, and output projections. ModelScopeT2V [11] contains 1,347.44M parameters, whereas the additional parameters introduced by adding LoRA amount to only 1.58M – approximately **0.1%** of the total ModelScopeT2V parameters.

D. More Details about User Study

In the main paper, we present a user study to demonstrate the effectiveness of InstructVideo. This study involves a comparative analysis of videos generated by InstructVideo and other methods, focusing on two key aspects: video quality and video-text alignment. For video quality, we asked annotators to evaluate: **1)** The overall visual quality of the videos, **2)** Alignment with general human aesthetic preferences, such as pleasing visuals, texture and details, and **3)** The smoothness and consistency in terms of structural and color transitions within the video. Regarding video-text alignment, annotators are tasked with determining the extent to which the generated videos accurately and clearly represent the content of the provided text prompts. This assessment included evaluating the depiction of entities, attributes, relationships, and motions as described in the prompts. To simplify the evaluation process, annotators are asked to perform pairwise comparisons between videos, thereby streamlining their task to direct contrasts rather than isolated assessments.

[1] https://github.com/huggingface/diffusers/blob/main/src/diffusers/models/lora.py

Method	In-domain	New Animals	Non-animals
ModelScopeT2V	0.2506 $\pm$ 0.0155	0.2502 $\pm$ 0.0138	0.2557 $\pm$ 0.0177
ModelScopeT2V†	0.2542 $\pm$ 0.0122	0.2541 $\pm$ 0.0109	0.2610 $\pm$ 0.0158
InstructVideo	**0.2717** $\pm$ **0.0137**	**0.2645** $\pm$ **0.0125**	**0.2682** $\pm$ **0.0202**
InstructVideo†	**0.2736** $\pm$ **0.0125**	**0.2664** $\pm$ **0.0131**	**0.2739** $\pm$ **0.0210**

Table A.1. **Generation with 50-step DDIM inference** after fine-tuning with 20-step DDIM inference. † denotes the model utilizes $D = 50$ while others adopt $D = 20$. 'In-domain' denotes in-domain animal prompts from the evaluation data.

E. More Visualization Results

We provide more visualization results to exemplify the conclusions we draw in the main paper, including: **1)** More results demonstrating how the generated videos evolve as the fine-tuning process proceeds as shown in Fig. A.2; **2)** More results showcasing the comparison between InstructVideo and the base model ModelScopeT2V as illustrated in Fig. A.3; **3)** More results exemplifying the comparison between InstructVideo and other reward fine-tuning methods as shown in Fig. A.4; **4)** More results showing the InstructVideo's generalization capabilities to unseen text prompts as shown in Fig. A.5.

F. 50-Step Generation with **InstructVideo**

To showcase the adaptability and effectiveness of InstructVideo, we conduct experiments using a 50-step DDIM inference for generation after initial fine-tuning with 20-step DDIM inference. The results are shown in Tab. A.1. We observe that InstructVideo, despite being fine-tuned with a 20-step protocol, remains effective under a longer 50-step DDIM inference protocol, as demonstrated by the boosted reward scores. We present several cases to further illustrate InstructVideo's efficacy as shown in Fig. A.6. We observe that both inference schemes can significantly improve over the base model and adopting more inference steps can occasionally lead to better results.

G. 50-Step Reward Fine-tuning

To assess the adaptation of InstructVideo to different DDIM steps, we experiment on reward fine-tuning with the commonly-used 50-step DDIM inference and evaluate its 20-step generation quality for a fair comparison. We present the results in Fig. A.1. The results demonstrate that InstructVideo could be optimized towards higher reward scores in both settings. However, utilizing 50 steps degrades the fine-tuning efficiency, likely due to the increased computation brought by longer sampling chains.

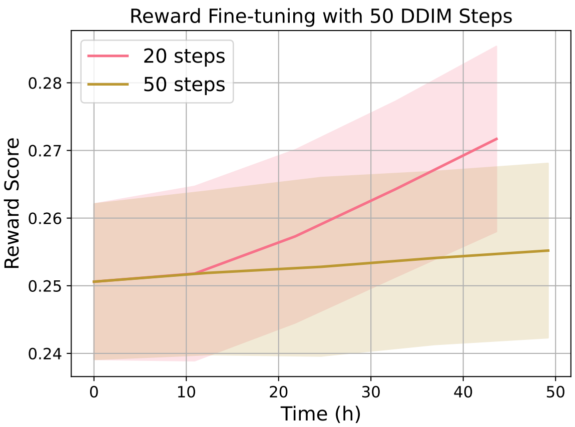


Figure A.1. **Reward finetuning with 50-step DDIM inference.**

H. Adaptation to Other Reward Functions

In the main paper, we focus on the utilization of HPSv2 [12]. To further validate the generalization of our method to other reward functions, we explore the application of ImageReward [13] as our reward model. ImageReward is a general-purpose text-to-image human preference reward model, fine-tuned on BLIP [6]. We perform reward fine-tuning as HPSv2 and present results in Fig. A.7. We observe that the quality of the videos are generally boosted in terms of structures, color vibrancy and details, despite that the stylistic aspects of the videos differ from those fine-tuned with HPSv2.

I. More Ablation Studies on SegVR and TAR

We present more qualitative comparisons in Fig. A.8. The figures indicate that removing both components leads to generation degradation. We also present a quantitative comparison in Fig. A.9(a). The curve indicates that removing both components can accelerate fine-tuning at an early stage, but the generation ability falters around the turning point (marked by red circle) and cannot be recovered, with variance increased. This is consistent with the main paper that overly dense or excessively strong reward signals can lead to generation collapse.

J. The Extreme Case of TAR

One extreme case of TAR is that only the central frame is utilized. This can be achieved by setting $\lambda_{\text{tar}} = +\infty$, which ensures the exclusive use of the central frame. We present the quantitative results (fine-tuning curves) in Fig. A.9(b). With only the central frame providing supervision, the reward signal is relatively weak, leading to low fine-tuning efficiency.

Method	Reward Score ↑	CLIPSIM ↑	Temporal Consistency ↑
ModelScopeT2V	0.2513	0.2961	0.9395
DDPO	0.2519	0.2976	0.9419
RWR	0.2558	0.3010	0.9692
DRaFT	0.2591	**0.3024**	0.9624
`InstructVideo` **(20k)**	**0.2707**	0.2998	**0.9848**
`InstructVideo` **(17k)**	0.2675	0.3020	0.9780

Table A.2. **Quantitative evaluation with CLIPSIM and Temporal Consistency .**

K. Evaluation with CLIPSIM and Temporal Consistency

We use the reference-free metric CLIPSIM and Temporal Consistency from Gen-1 [3] to evaluate `InstructVideo`. We evaluate on all in-domain and unseen prompts from Tab. 1 of the main paper, and present the results in Tab. A.2. The default setting (20K) achieves the best on Reward Score and Temporal Consistency. If we perform less fine-tuning (17k), we could obtain higher CLIPSIM, indicating that CLIPSIM is not correlated with the objective of human preferences.

References

[1] 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] Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023. 1

[3] 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. 3

[4] 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. 1

[5] 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. 1

[6] 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

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

[8] 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

[9] 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

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 1

[11] 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

[12] 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

[13] 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. 2

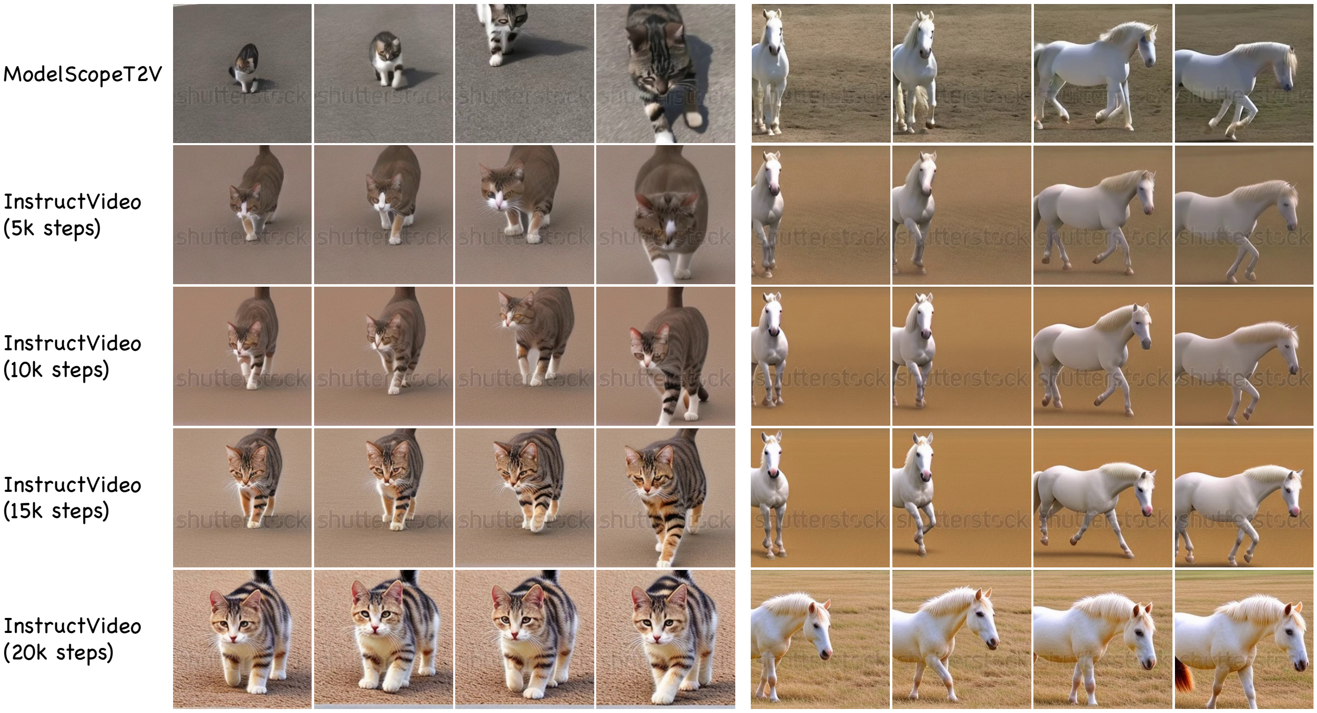

ModelScopeT2V
InstructVideo (5k steps)
InstructVideo (10k steps)
InstructVideo (15k steps)
InstructVideo (20k steps)
shutterstock

Figure A.2. More examples showing **the evolution of generated videos** during fine-tuning.

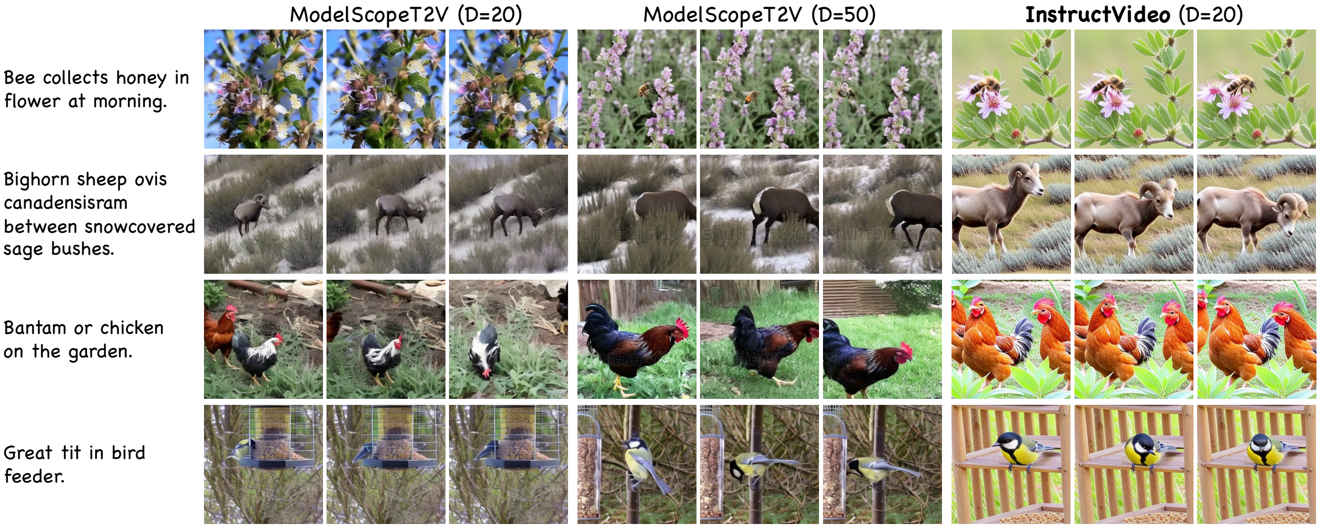

ModelScopeT2V (D=20)
ModelScopeT2V (D=50)
InstructVideo (D=20)
Bee collects honey in flower at morning.
Bighorn sheep ovis canadensisram between snowcovered sage bushes.
Bantam or chicken on the garden.
Great tit in bird feeder.

Figure A.3. More examples showing the **comparion of `InstructVideo` with the base model ModelScopeT2V**. ModelScopeT2V utilizes 20 and 50 DDIM steps.

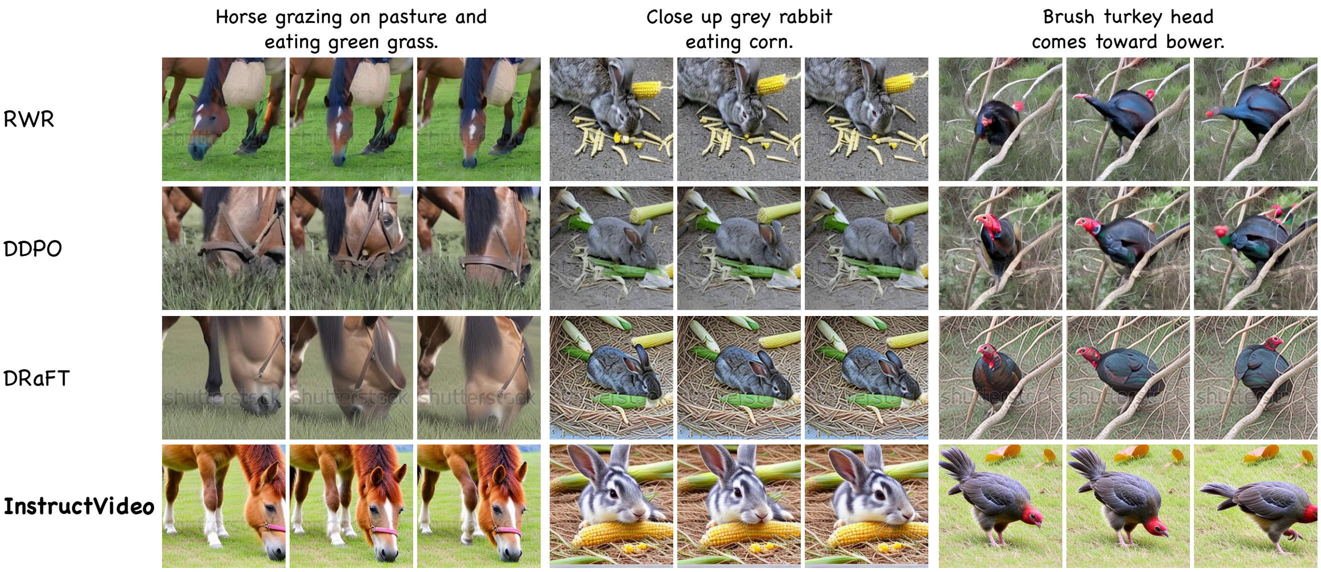

Horse grazing on pasture and eating green grass.
Close up grey rabbit eating corn.
Brush turkey head comes toward bower.
RWR
DDPO
DRaFT
InstructVideo

Figure A.4. More examples showing the **comparison of `InstructVideo` with other reward fine-tuning methods**. We set $D = 20$ for all methods.

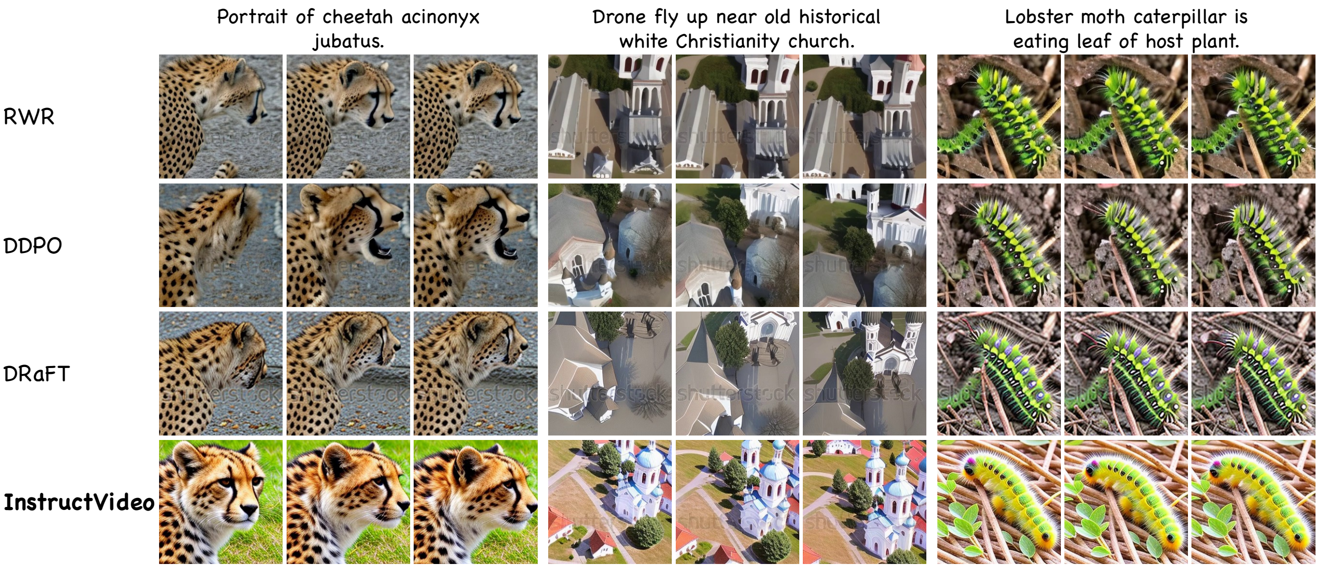

Portrait of cheetah acinonyx jubatus.
Drone fly up near old historical white Christianity church.
Lobster moth caterpillar is eating leaf of host plant.
RWR
DDPO
DRaFT
InstructVideo

Figure A.5. More examples showing the **comparison of `InstructVideo`'s generalization capabilities with other methods**. We set $D = 20$ for all methods.

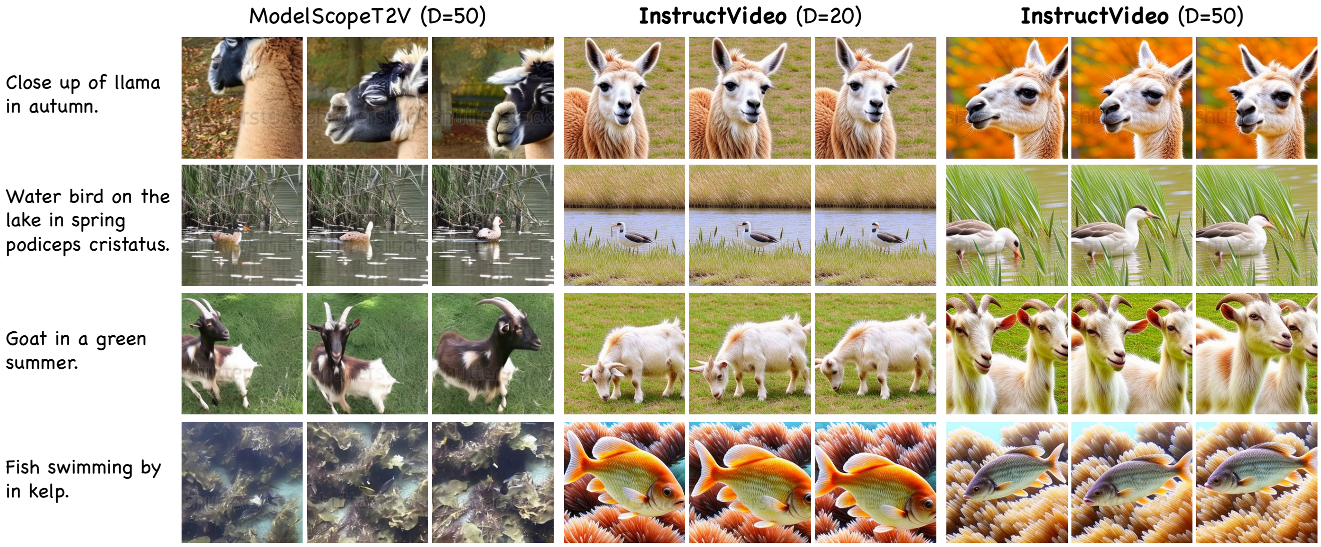

ModelScopeT2V (D=50)
InstructVideo (D=20)
InstructVideo (D=50)
Close up of llama in autumn.
Water bird on the lake in spring podiceps cristatus.
Goat in a green summer.
Fish swimming by in kelp.

Figure A.6. **Generation with 50-step DDIM inference** after fine-tuning with 20-step DDIM inference.

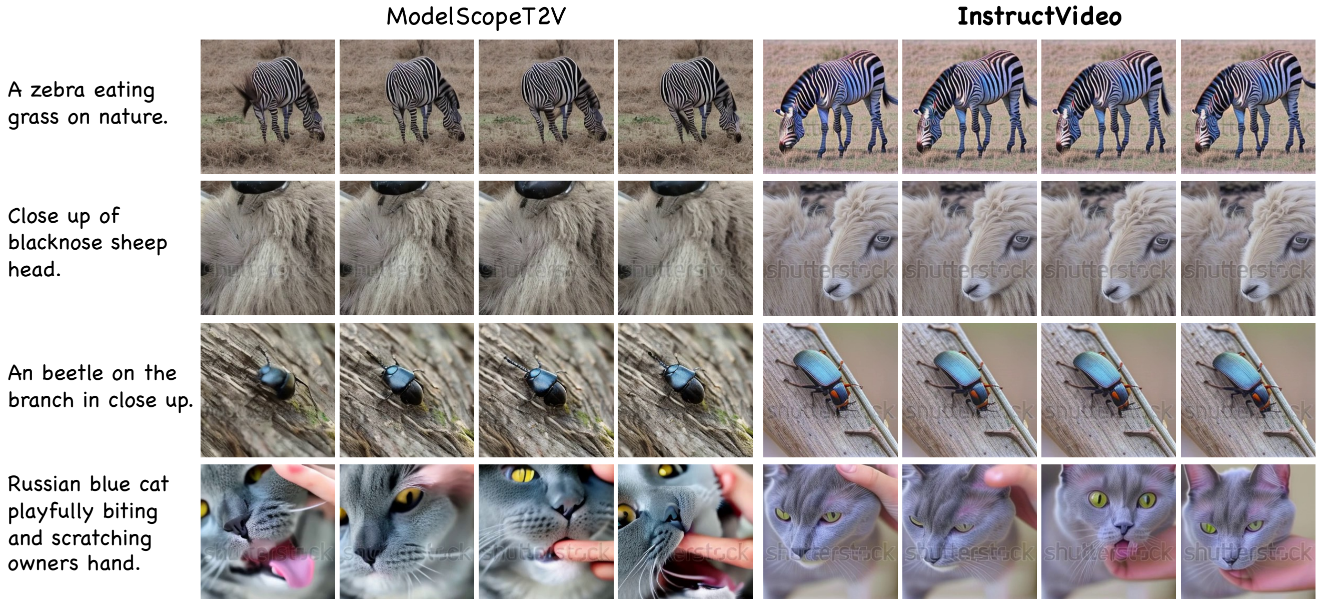

ModelScopeT2V
InstructVideo
A zebra eating grass on nature.
Close up of blacknose sheep head.
An beetle on the branch in close up.
Russian blue cat playfully biting and scratching owners hand.

Figure A.7. **Comparison of `InstructVideo` fine-tuned using ImageReward with the base model ModelScopeT2V. We set** $D = 20$ **for two methods.**

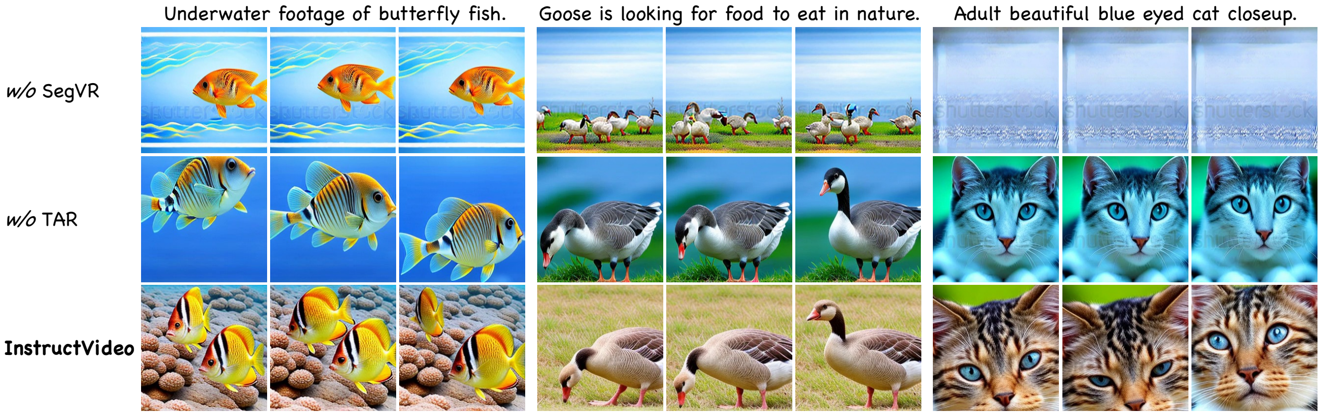
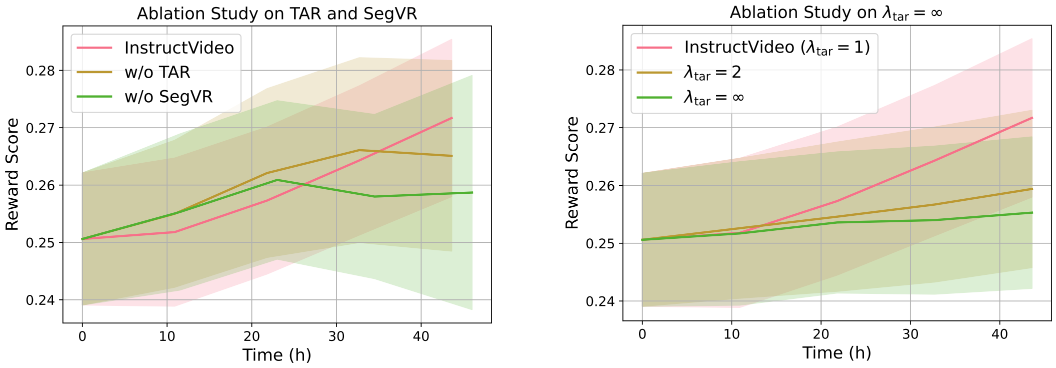

Underwater footage of butterfly fish.
Goose is looking for food to eat in nature.
Adult beautiful blue eyed cat closeup.
w/o SegVR
w/o TAR
InstructVideo

Figure A.8. **More ablation studies on SegVR and TAR**.



(a) **(b)**

Figure A.9. **(a) The quantitative analysis of the ablation study on SegVR and TAR. (b) The quantitative analysis of the extreme case of TAR**, where only the central frame is used.