Align Video Diffusion Model with Online Video-Centric Preference Optimization (*Supplementary Materials*)

We provide more details about our method, experiment, and additional experiment results to facilitate a comprehensive understanding of our work.

1. Pseudo Code of the OnlineVPO

The complete procedure of the proposed OnlineVPO is summarized in Algorithm.1. Practically, we convert the original direct preference optimization (DPO) [8] objective to the diffusion loss format following [11].

2. Extended Implementation Details

Training and Evaluation We implement OnlineVPO with UNet-based VideoCrafter v2 [2] and DiT-based OpenSora v1.2 [14] to demonstrate the performance of our approach and also for the sake of comparison with existing methods. Here, we provide more details about model training and evaluation. By default, all experiments are conducted on a machine with 8× NVIDIA L40 GPUs, with a total batch size of 8 (1 per GPU). We specify the details of our training with two base VDMS as follows:

VideoCrafter v2: We adopt the official implementation of T2V generation[1] version. More specifically, we build upon OnlineVPO on its T2V VideoCrafter v2 model that was released on the HuggingFace[2]. During training, we use text prompts from WebVid-10M [12] to generate online video candidates. For each prompt, six 12-frame videos at 320 × 512 resolution are generated. We then rank these samples using VideoScore's temporal dimension prediction score and select the best and worst candidates for training. The AdamW with a learning rate of 1e-4. Following [7], we utilize the LoRA [4] to achieve efficient fine-tuning, and the LoRA rank is set to 16. We train VideoCrafter with OnlineVPO for 4000 steps.

OpenSora v1.2: Our implementation is based on the official Open-Sora release[3]. During training, we use text prompts from WebVid-10M [1] to generate online video candidates (note that we do not use the associated videos). For online sample generation, we take the inherent rectified flow scheduler with 30 sampling steps, with a logit-normal sampling method and a CFG scale of 4.0. For each prompt, six 34-frame online samples at 240p with a respect ratio of 9:16 (i.e., height: width = 240 × 424) are generated. We then rank these samples using VideoScore's temporal dimension prediction score and select the best and worst candidates for training. The AdamW with a learning rate of 2e-6 is used, and the warmup step is set to 500. To avoid gradient explosion, we multiply the OnlineVPO loss by a grad scale of 0.01. We train OpenSora with OnlineVPO for 5000 steps.

Evaluation Details We assess our method using VBench [5], a comprehensive benchmark for video generation evaluation. From its curated set of approximately 1,000 prompts, we randomly select 100 prompts for efficient evaluation. We evaluate performance across six key dimensions that align well with human perception: dynamic degrees, subject consistency, background consistency, aesthetic quality, image quality, and motion smoothness. Furthermore, we compute an overall quality score following VBench's standard protocol[4] to provide a comprehensive performance summary.

User Study To evaluate the subjective quality of generated videos, we employ a two-alternative forced choice (2AFC) test comparing outputs from different video diffusion models (VDMs). We curate 100 commonly used prompts from existing literature. Then, for each model to be evaluated $M = \{m_1, m_2, ..., m_N\}$, we generate three candidate videos per prompt $\{p_1, p_2, p_3 | m_i\}$. During evaluation, we randomly select two models m_i and m_j for a given prompt. Then, the comparison videos p_{m_i} and p_{m_j} are randomly sampled from their corresponding candidate set. After that, the participants are presented with this pair and asked to select the superior video (no tie choice) based on three criteria: (1) more natural motion, (2) more aesthetic appearance, and (3) better alignment with the prompt.

[1] https://github.com/AILab-CVC/VideoCrafter

[2] https://huggingface.co/VideoCrafter/VideoCrafter2/blob/main/model.ckpt

[3] https://github.com/hpcaitech/Open-Sora/tree/opensora/v1.2/opensora

[4] https://github.com/Vchitect/VBench?tab=readme-ov-file#how-to-calculate-total-score

Algorithm 1: Online Video Preference Optimization (OnlineVPO)

```
Input: Prompt set $M = \{x_0, \cdots, x_n\}$, video diffusion model $\pi_\theta$, video reward model $r(\cdot)$, curriculum update interval $K$, online
       sample candidate number $N$
Output: Preference-aligned video diffusion model $\pi^*(\cdot)$
step $\leftarrow 0$
for $x_i$ in $M$ do
    // Online Preference Sample Generation
        $V \leftarrow \{v_1, v_2, \cdots, v_N\} \sim \pi_\theta(x_i)$
        $S \leftarrow \{s_1, s_2, \cdots, s_N \mid s_i = r(v_i)\}$
        $(v^w, v^l) \leftarrow (v_i, v_j)$, where $i = \arg\max_i s_i, j = \arg\min_j s_j$
        // Noise and Timestep Sampling
        $\epsilon^w, \epsilon^l \leftarrow \epsilon \sim \mathcal{N}(0, I), t \sim \mathcal{U}(0, T)$
        $y_t^w \leftarrow \text{AddNoise}(v^w, \epsilon^w)$
        $y_t^l \leftarrow \text{AddNoise}(v^l, \epsilon^l)$
        // OnlineVPO Loss Calculation
        $\mathcal{L}_{\text{OnlineVPO}} = -\mathbb{E}\Big[\log\sigma\big(\beta\log(||\epsilon^w - \pi_\theta(y_t^w, x_i, t)||_2^2 - ||\epsilon^w - \pi_{\text{ref}}(y_t^w, x_i, t)||_2^2)$
                $- (||\epsilon^l - \pi_\theta(y_t^l, x_i, t)||_2^2 - ||\epsilon^l - \pi_{\text{ref}}(y_t^l, x_i, t)||_2^2)\big)\Big]$
        // Model Update
        $\pi_\theta \leftarrow \pi_\theta + \nabla_{\pi_\theta}\mathcal{L}_{\text{OnlineVPO}}$
        step $\leftarrow$ step $+ 1$
        // Curriculum Update Reference
        if (step mod $K$) $= 0$ then
            $\pi_{\text{ref}} \leftarrow \pi_\theta$
```

3. Extended Results

3.1. More Ablation Study

We present more ablation studies to facilitate a more comprehensive understanding of our method.

Number of candidates N**.** OnlineVPO generates multiple video candidates and exploits the video reward model to determine the preference pair in an online manner. We analyze the impact of varying the number of video candidates N in Tab.1(a). It can be seen that increasing the number of video candidates leads to better performance. Generally, more video candidates can result in more diverse data samples, facilitating more robust and effective preference sample selection. However, performance gains tend to diminish once the number of candidates reaches 6 or more. Therefore, we opt to set the number of video candidates at 6 for our study.

Number of Curriculum Interval K OnlineVPO employs a curriculum-based approach to iteratively update the reference model, enhancing the efficiency of preference optimization. We investigate the impact of varying the frequency of reference model updates. As shown in Table.1(b), both more frequent and less frequent updates yield inferior performance. Frequent updates to the reference model can prevent it from serving as a stable baseline for optimizing targets, leading to suboptimal optimization. Conversely, infrequent updates may struggle to realign an already biased model. The optimal performance is observed with an update interval of $K = 200$, striking a harmonious balance between these two scenarios.

Other Evaluation Metrics In our main paper, we primarily utilize the VBench [5] to assess the effectiveness of our approach. Additionally, we also incorporate the FVD [10] and Video Dynamic Quality [6] metrics to conduct a comprehensive evaluation of our method. Specifically, we compute FVD on the UCF-101 [9] dataset following the methodology of [13]. However, FVD is criticized for its focus on individual frame quality. Therefore, we utilize the improved FVD implementation by [3] and compare our method's performance with others. Moreover, Video Dynamic Quality is a metric tailored to evaluate the dynamic characteristics of generated videos. We employ this metric to assess how well our method optimizes video dynamics. The results are summarized in Tab.1(c). Our approach demonstrates superior performance compared to other methods based on these two metrics. Taking the dynamic quality as an example, despite some improvement observed in InstructVideo in terms of FVD, it suffers from performance degeneration in dynamic quality. This is attributed to the naive application of the image-based reward model in a frame-wise manner, which enhances the frame image quality but can lead to temporal inconsistency. The effort made by the VADER only brings marginal improvement in dynamic quality due to the lack

Number	Dynamic Degree	Subject Consist.	Aesthetic Quality
2	42.4	96.09	49.16
4	42.0	97.09	54.33
6	43.0	**97.58**	**55.37**
8	**43.1**	97.21	54.75

(a) Online candidate number.

Interval	Dynamic Degree	Subject Consist.	Aesthetic Quality
100	42.7	97.02	53.74
200	43.0	**97.58**	**55.37**
400	41.0	96.64	51.26
600	**45.0**	96.11	51.88

(b) Reference curriculum interval.

Method	FVD↓	Dynamic Quality ↑
OpenSora	316.21	60.12
InstructVideo	296.50	58.37
VADER	244.78	61.89
Ours	**201.51**	**65.74**

(c) More metrics results.

Table 1. **OnlineVPO Extended Ablations**. We perform further ablations on (a) the number of video candidates when constructing the online preference pair, (b) the curriculum interval to update the reference model, and (c) the performance comparison with more evaluation metrics.

of a targeted video reward model. In contrast, OnlineVPO, leveraging a video reward model and online video preference learning, showcases significant enhancements in dynamic quality compared to the baseline.

3.2. More Visualization

We showcase more visualization examples of the generated performance of our approach and other methods with Open-Sora and VideoCrafter as base models, respectively, in Fig.2 and Fig.3. It can be observed that our approach outperforms the existing method in generating not only high-quality but also temporally coherent frame images.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1

[2] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024. 1

[3] Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *CVPR*, 2024. 2

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1

[5] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 1, 2

[6] Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. In *NeurIPS*, 2024. 2

[7] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024. 1

[8] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. 2024. 1

[9] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

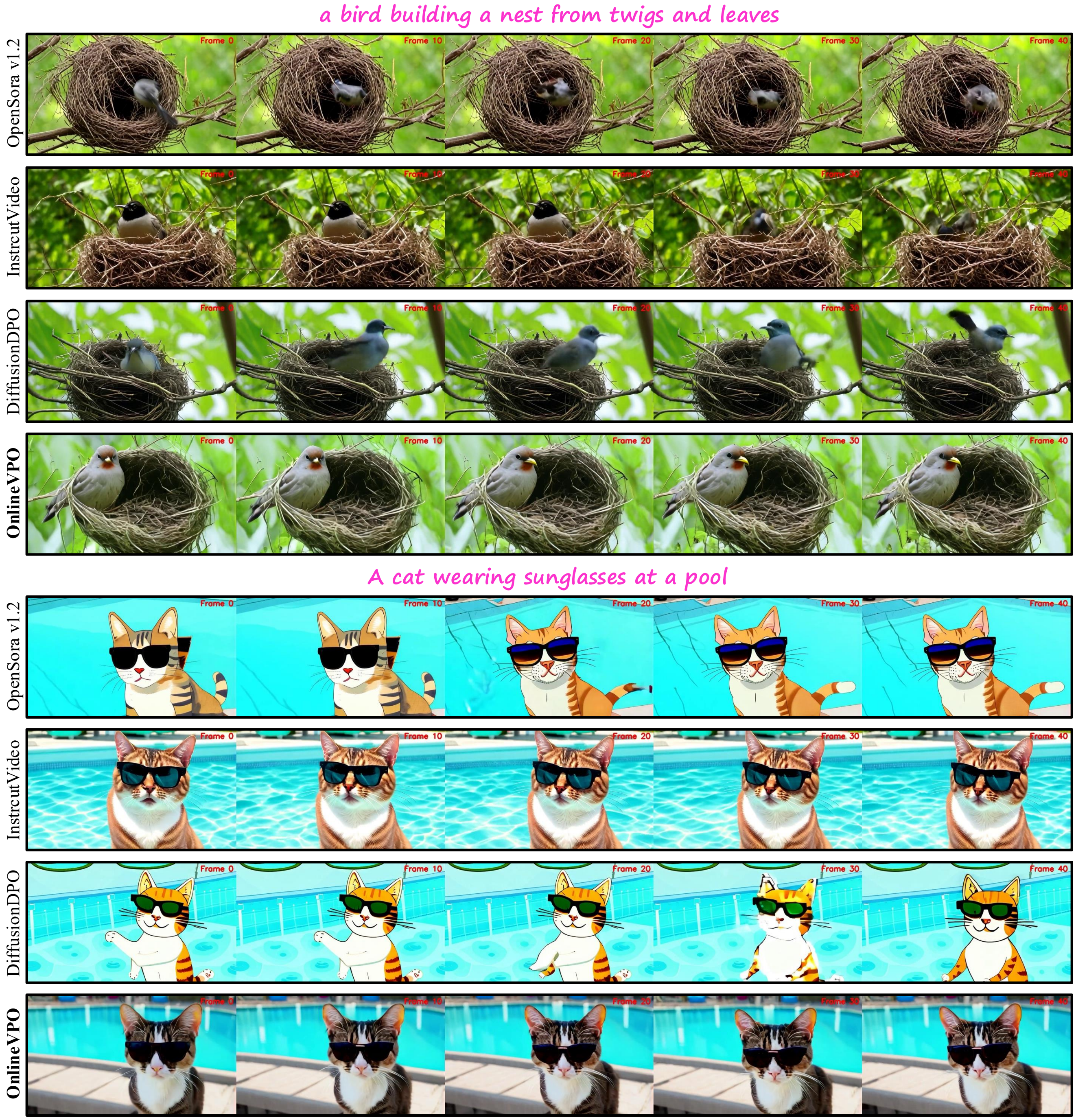
[10] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 2

[11] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, 2024. 1

[12] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 1

[13] Zhixing Zhang, Yanyu Li, Yushu Wu, Yanwu Xu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Dimitris Metaxas, Sergey Tulyakov, and Jian Ren. Sf-v: Single forward video generation model. *arXiv preprint arXiv:2406.04324*, 2024. 2

[14] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 1


a bird building a nest from twigs and leaves
OpenSora v1.2
Instrcut Video
DiffusionDPO
OnlineVPO
Frame 0
Frame 10
Frame 20
Frame 30
Frame 40
A cat wearing sunglasses at a pool
OpenSora v1.2
Instrcut Video
DiffusionDPO
OnlineVPO

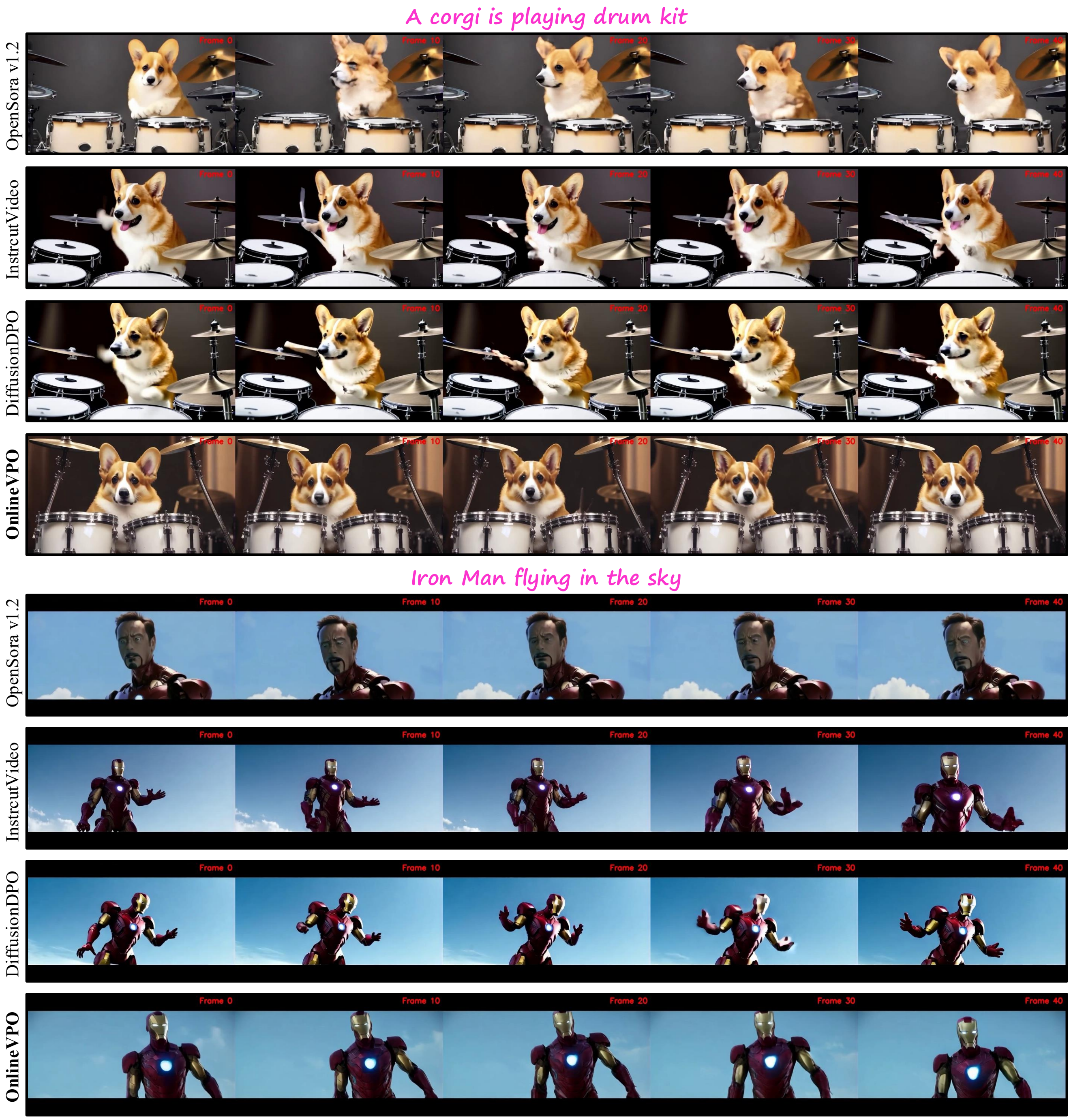
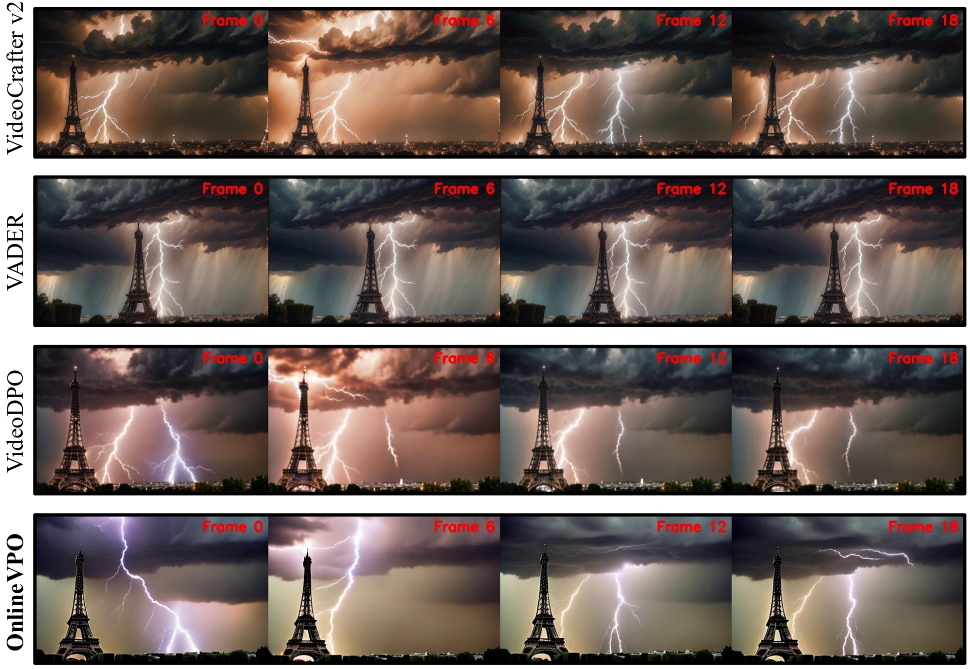
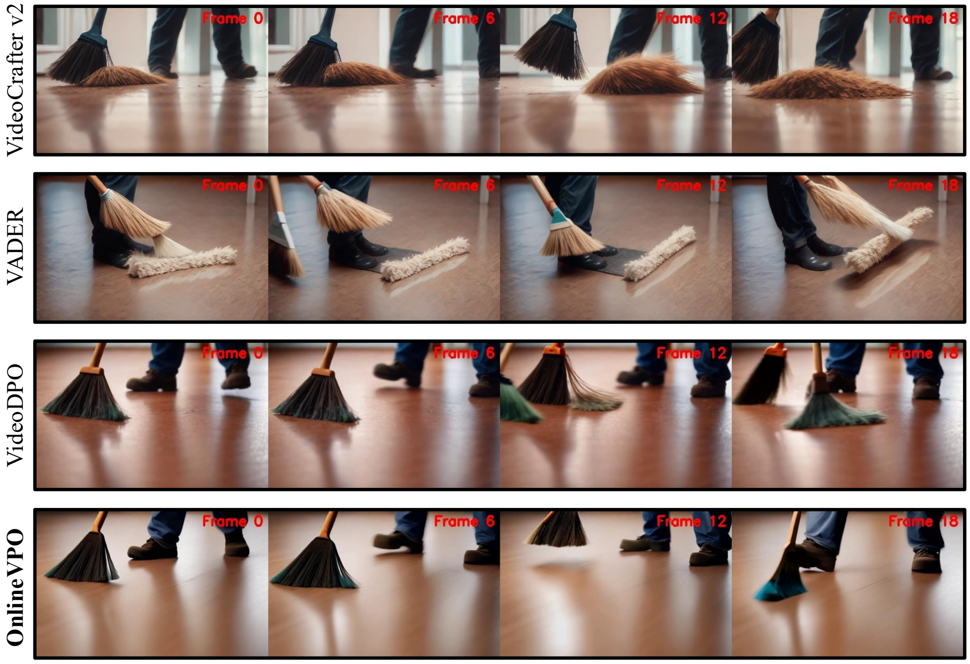


Figure 2. Visualization of the generation results of different video preference learning methods based on OpenSora v1.2.

A lightning striking atop of eiffel tower, dark clouds in the sky



A person is sweeping floor



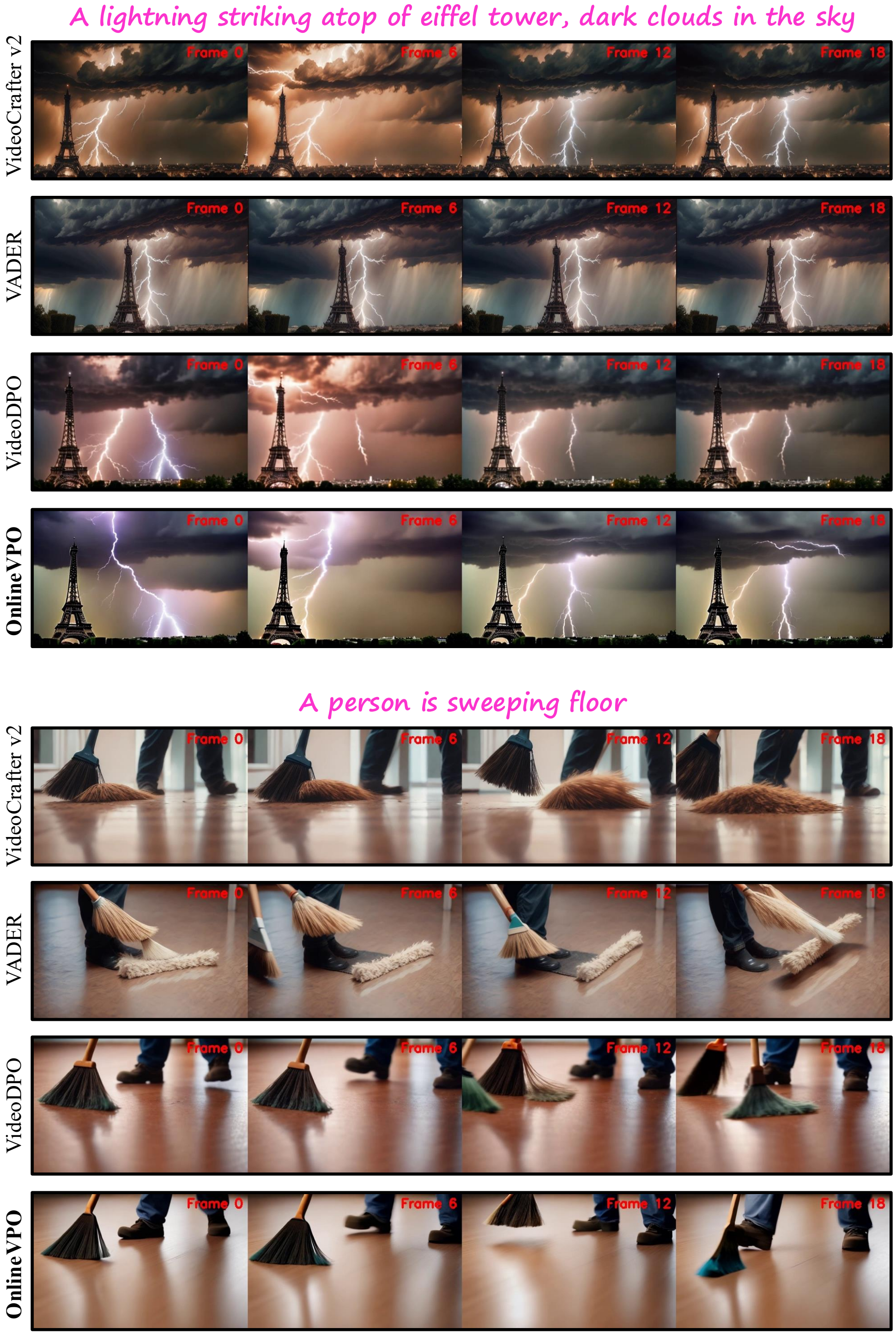


Figure 3. Visualization of the generation results of different video preference learning methods based on VideoCrafter v2.