# VideoDPO: Omni-Preference Alignment for Video Diffusion Generation

## Supplementary Material

This supplementary material presents OmniScore details, additional analysis and experimental results. Section A enumerates the details of OmniScore, including the model each dimenson ultilizes and the corresponding weights. Section B compares the performance of single-dimensional, multi-dimensional settings, user study and aggregation methods, also examines the impact of training data scale on the results and provides the discussion of the limitations. Section C includes additional intra-frame and inter-frame qualitative results.

## A. OmniScore Implementation.

Inspired by the models used in [22], we build OmniScore by referencing these models and their corresponding weights to evaluate the quality of video samples. [22] aims to evaluate the quality of video generative models, whereas our OmniScore targets assessing the quality of video samples specifically for preference learning. Here we demonstrate the detailed composition of OmniScore:

**Motion smoothness.** We utilize the motion priors in the video frame interpolation model [29] to evaluate the smoothness of generated motions

**Temporal flickering.** We take static frames by RAFT [50] and compute the mean absolute difference across frames.

**Subject consistency.** For a subject(e.g., a person, a car, or a cat) in the video, we assess whether its appearance remains consistent throughout the whole video. To this end, we calculate the DINO [4] feature similarity across frames.

**Imaging quality.** Imaging quality refers to the distortion (e.g., over-exposure, noise, blur)presented in the generated frames, and we evaluate it using the MUSIQ [23] image quality predictor trained on the SPAQ [11] dataset.

**Aesthetic quality.** We evaluate the artistic and beauty value perceived by humans towards each video frame using the LAION aesthetic predictor [25].

**Dynamic degree.** We use RAFT [50] to estimate the degree of dynamics in synthesized videos.

**Text-video semantic alignment.** We use overall video-text consistency computed by ViCLIP [57].

The following dimensions are scaled to the range $[0, 1]$ based on the following values:

- **Subject consistency**: $\text{Min} = 0.1462, \text{Max} = 1.0$
- **Temporal flickering**: $\text{Min} = 0.6293, \text{Max} = 1.0$
- **Motion smoothness**: $\text{Min} = 0.706, \text{Max} = 0.9975$
- **Overall consistency**: $\text{Min} = 0.0, \text{Max} = 0.364$

The weights assigned to Motion Smoothness, Temporal Flickering, Subject Consistency, Imaging Quality, Aesthetic Quality and Dynamic Degree are all 4, and the weight for Text-Video Semantic Alignment is set to 1.

## B. Additional Analysis

**Single- vs. multi-dimensional score comparison.** In Table 5, we explore the results of training on a single-dimensional reward score compared to training on our OmniScore. The experimental results show that OmniScore achieves the best performance, highlighting the importance of a comprehensive score for our framework.

**Multi-dimensional score aggregation.** We explore two methods for multi-dimensional score aggregation: (1) selecting 10,000 pairs based on our OmniScore and (2) Combine preference pairs from individual dimensions into a larger dataset so that the VC2 model is trained on 40,000 pairs, with 10,000 pairs selected from each of the four dimensions: semantics, aesthetics, motion smoothness, and dynamic degree. The results indicate that the second approach significantly lowers performance to 78.26% on VBench-Total, showing that using our OmniScore can achieve better performance.

**User study.** We conducted a user study of OmniScore on videos with human preference labels from [22], showing strong agreement with human preferences as 78%. We also evaluated VideoDPO's performance through a user study involving 10 unaffiliated participants assessing 20 video sets, each with a prompt and videos generated by 4 baselines. This experiment demonstrates the efficacy of our method, with participants selecting preferred videos from each set yielding the following selection rates: 14% for VC2, 16% for SFT, 24% for VADER, and **46%** for VideoDPO, clearly indicating superior performance of our proposed approach.

**Effect of training scale on performance.** We compared the performance shown in Table 4 when using only half and 25% of the prompt data for training, observing a significant drop across all metrics. This result demonstrates that increasing the amount of prompt data in training yields substantially better performance. We attribute this to improved generalization, as the model aligns with a broader range of prompts. These experiments suggest that our method still has room for improvement, particularly with regard to the amount of data.

**Limitations of VideoDPO.** VideoDPO's limitation stems from computational demands, particularly due to OmniScore's use of multiple vision models and the base model's

| Data | VBench(%) | | | HPS (V) | PickScore |
|---|---|---|---|---|---|
| | Total | Quality | Semantic | | |
| 25% | 80.21 | 81.70 | 74.26 | 0.259 | 20.66 |
| 50% | 80.83 | 82.37 | 74.68 | 0.260 | 20.59 |
| Full(ours) | 81.93 | 83.07 | 77.38 | 0.261 | 20.65 |

Table 4. Scores for Different Dataset Sizes

inference speed for generating $N$ video candidates per prompt. Solutions include exploring faster vision models, choosing a wise $N$, and employing faster base models. Additionally, the base model's limited generation capabilities restrict the alignment effects as DPO depends on inherent model capabilities. Conversely, if the model inherently has the capability to sample high-quality outputs in certain aspects, then those can be improved well.

## C. Additional Qualitative Results

We present the results of inter-frame and intra-frame alignment before and after learning in Figure 6 and Figure 7, respectively, following the format of the main paper. The results demonstrate that our alignment method is effective across a wide range of prompts, improving temporal consistency, visual quality, and semantics.

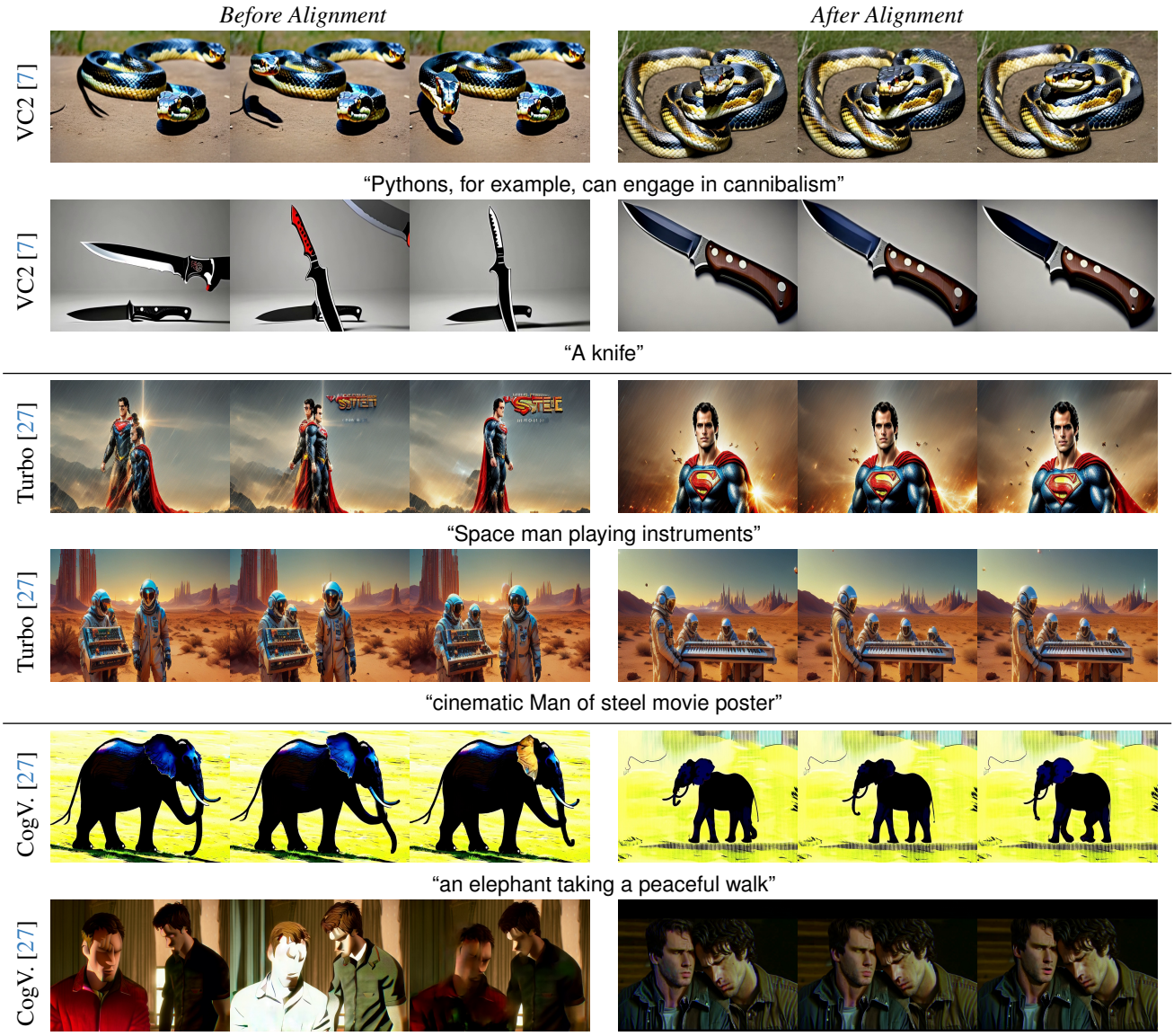


Figure 6. Additional inter-frame qualitative visualization.

| Score | VBench (%) | | | Subject Consis. | Aesthetic Quality | Overall Consis. |
|---|---|---|---|---|---|---|
| | Total | Quality | Semantic | | | |
| Overall Consis. | 80.20 | 81.57 | 74.74 | 95.61 | 62.94 | 78.76 |
| Aesthetic Quality | 79.65 | 81.67 | 71.57 | 97.13 | 63.27 | 76.98 |
| Subject Consis. | 77.05 | 79.00 | 69.28 | 94.25 | 58.23 | 73.35 |
| OmniScore (ours) | 81.93 | 83.07 | 77.38 | 95.69 | 63.18 | 78.43 |

Table 5. Scores for different training objectives include single-dimensional scores such as overall consistency, aesthetic quality, and subject consistency, as well as our multi-dimensional score, OmniScore. "Consis." is the abbreviation for "consistency."

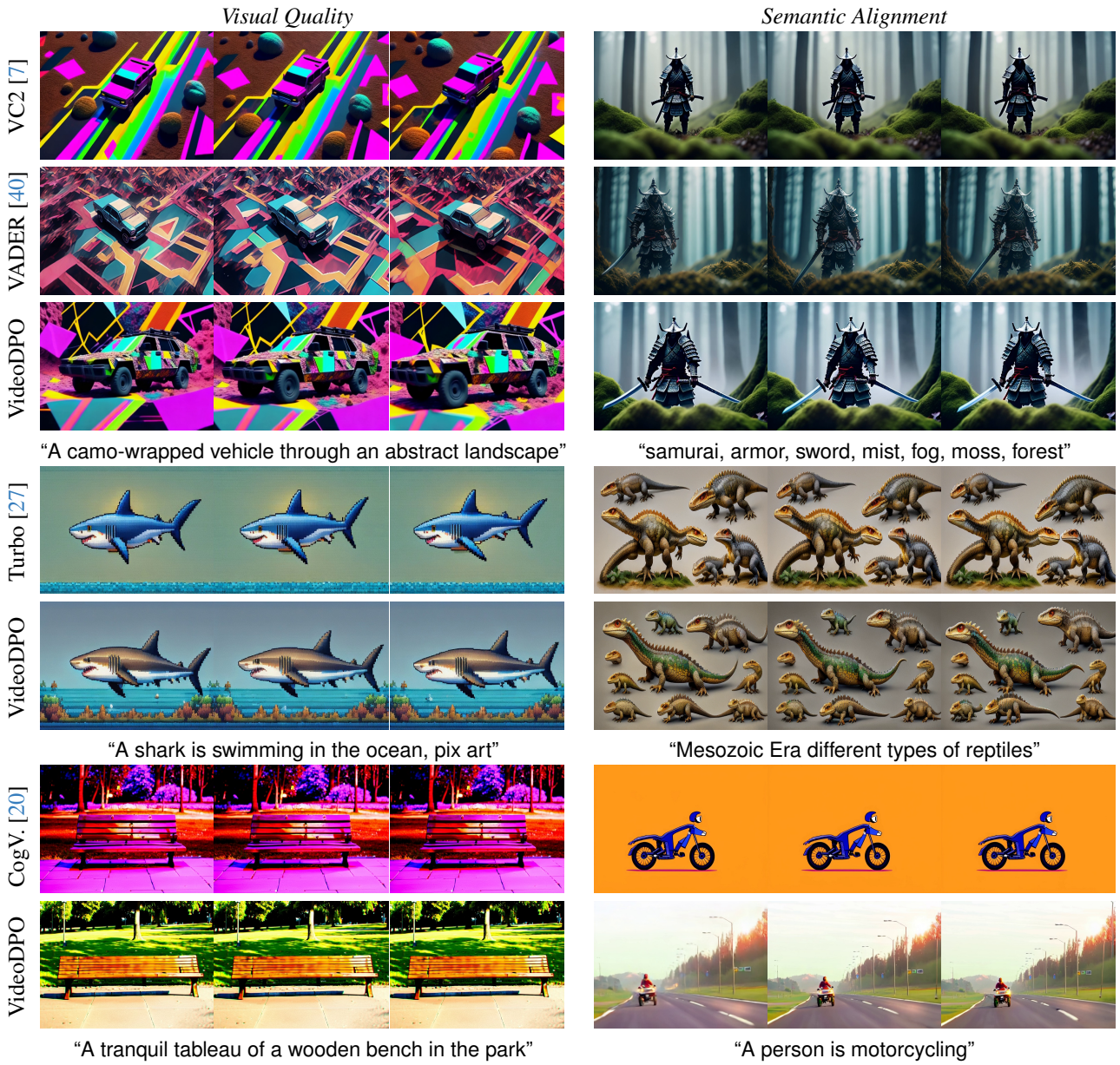


Figure 7. Additional intra-frame qualitative visualization.