

# Supplementary Materials of MoTrans: Customized Motion Transfer with Text-driven Video Diffusion Models

Anonymous Authors

## 1 DETAILS OF BENCHMARKS

To the best of our knowledge, there are currently no unified benchmarks for motion customization tasks. Most representative motion customization methods [6–8] typically select 8–12 different types of motions from UCF101 [3], UCF Sports Action [2], NTU RGB+D [1] for evaluation, covering a wide range of human-centric sports and daily activities. Following this setting, we have constructed our benchmark. Considering that the existing datasets contain many simple motions with small movement amplitudes, such as walking frontally and snapping fingers, we do not directly use them for evaluation. Instead, we select videos with larger movement amplitudes and higher quality to enhance the diversity and complexity of our benchmark. We have carefully selected 12 motions including bowing, clapping, skateboarding, drinking water, lifting weights, kicking something, playing golf, riding a bicycle, pointing to something, playing the guitar, waving hand, and wiping face.

## 2 ANALYSIS AND DISCUSSIONS

### 2.1 Discussion on Motion Fidelity Metric

Considering the diverse composition of our benchmark, which includes not only sports actions but also various limb movements that accentuate dynamic processes, conventional metrics focused on single frame-to-text alignment, such as Textual Alignment (CLIP-T) and Entity Alignment (CLIP-E), may not provide sufficient measurement for motion quality. For instance, given the prompt "A tiger is drinking water in the forest", existing video foundation models often generate videos showing a tiger merely standing by a lake. While such outputs might achieve high CLIP-T and CLIP-E scores and ostensibly align with the textual description, they frequently misrepresent the specific motion in reference videos. To address this challenge, we have introduced a novel metric named Motion Fidelity (MoFid). This metric leverages the advanced video understanding capabilities of VideoMAE [4] to quantitatively assess how well the motion in generated videos matches the motion observed in the training dataset.

Considering the different action types covered by VideoMAE and our method, this mismatch potentially leads to inaccuracies in motion type prediction. For instance, a video generated in response to the prompt "an alien is bowing" might be interpreted by VideoMAE as depicting "robot dancing", while a video produced for "a cat is wiping its face" could be erroneously categorized as "cat petting". Such discrepancies underscore the limitations of using straightforward classification accuracy to measure Motion Fidelity. Instead, we employ the cosine similarity of video representations to measure Motion Fidelity as mentioned in the main manuscript.

### 2.2 Tradeoff between Text/Entity Alignment and Motion Fidelity

We aim to explore the correlation between Text/Entity Alignment and Motion Fidelity, as depicted in Fig. 1. When a single reference

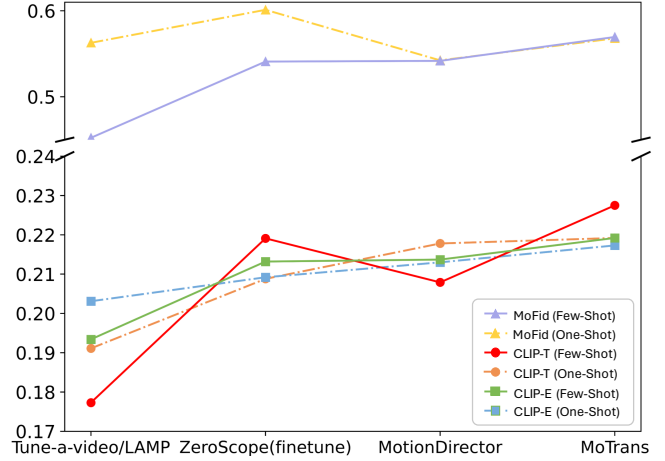


Figure 1: Relationship between Text/Entity Alignment and Motion Fidelity. A high MoFid score coupled with low CLIP-T and CLIP-E scores indicate that the synthesized video’s appearance is excessively fitted to the reference video, resulting in a lack of appearance diversity.

Table 1: Quantitative comparison results of motion customization on multiple videos.

Methods	CLIP-T (↑)	CLIP-E (↑)	TempCons (↑)	MoFid (↑)
DreamVideo	0.1791	0.2208	0.9680	0.4243
MoTrans	<b>0.2168</b>	<b>0.2225</b>	<b>0.9776</b>	<b>0.5386</b>

image is provided, the finetuned ZeroScope achieves the highest Motion Fidelity score, yet its CLIP-T and CLIP-E metrics score the lowest. These results indicate a high similarity in both appearance and motion between the synthesized video and the reference video. In other words, while finetuned ZeroScope may accurately model the motion in the reference videos, it fails to generate the new context or entity implied by the prompts, thus limiting the overall creativity and diversity of the generated content.

Additionally, the videos synthesized by the few-shot methods align more closely with the text prompt, as evidenced by higher CLIP-E score. This suggests that the few-shot setting, compared to the one-shot setting, is better at synthesizing the subject specified in the prompt and avoids replicating appearances from the reference video. Based on the analysis above, it can be concluded that although Motion Fidelity is a useful metric for assessing how consistently a video’s motion matches that of the reference, it only provides a limited view of overall performance. High Motion Fidelity coupled with very low CLIP-T and CLIP-E scores typically indicates an overfitting to the reference’s appearance. In contrast, our method

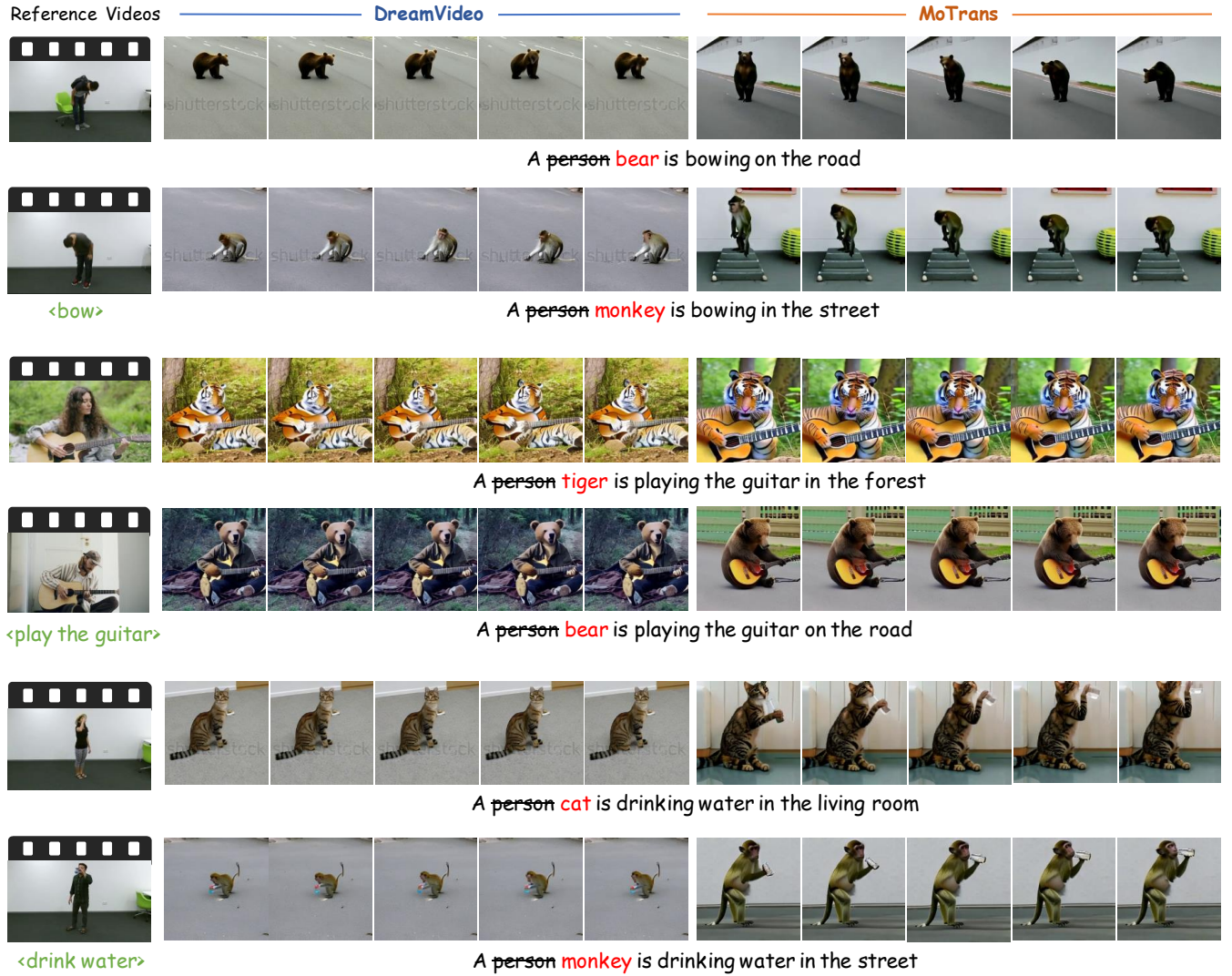


Figure 2: Qualitative comparisons between MoTrans and DreamVideo.

consistently achieves high CLIP scores and Motion Fidelity in both one-shot and few-shot settings, demonstrating its effectiveness in modeling the motion pattern in reference videos without overfitting to their appearances.

### 3 ADDITIONAL RESULTS

#### 3.1 Comparisons with DreamVideo

We conduct additional qualitative and quantitative comparisons with DreamVideo [5] to further demonstrate the superiority of our proposed MoTrans. DreamVideo employs an updated ModelScopeT2V model, which has been fine-tuned on its internal dataset at a resolution of 256. This fine-tuned version has not been made available to the public. Consequently, for our comparative analysis, we utilize the originally released ModelScopeT2V. DreamVideo is

optimized for generating videos at a resolution of 256x256. For consistency and to enable a direct comparison, we also synthesize videos at this resolution. As shown in Table 1, DreamVideo can generate subjects specified by the prompt but often struggles to synthesize specific motions contained in the reference videos and the context specified by the prompt. Correspondingly, while its CLIP-E score is relatively high and comparable to our method, there is a significant difference in its CLIP-T and MoFid scores when compared to ours. The visual results in Fig. 2 further confirm these observations. Compared to DreamVideo, our method demonstrates superior motion modeling capabilities.

#### 3.2 More Qualitative Results

Fig. 3 and 4 respectively present additional video examples synthesized by our approach, given a single reference video and multiple

reference videos. We further compare our approach with other baselines, as demonstrated in Fig. 5. Supplementary results from ablation studies are presented in Fig. 6.

REFERENCES

[1] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 2020. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 10 (2020), 2684–2701.

[2] Khurram Soomro and Amir R Zamir. 2015. Action recognition in realistic sports videos. In *Computer vision in sports*. Springer, 181–208.

[3] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[4] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.

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

[6] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. 2023. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint arXiv:2310.10769* (2023).

[7] Yuxin Zhang, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023. MotionCrafter: One-Shot Motion Customization of Diffusion Models. *arXiv preprint arXiv:2312.05288* (2023).

[8] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. 2023. MotionDirector: Motion Customization of Text-to-Video Diffusion Models. *arXiv preprint arXiv:2310.08465* (2023).





Reference Video

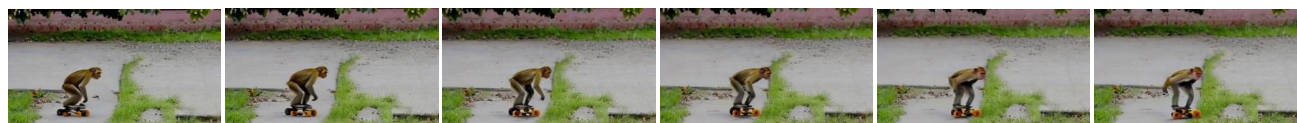
A ~~person~~ **panda** is skateboarding under the treeA ~~person~~ **tiger** is skateboarding in the forestA ~~person~~ **bear** is skateboarding on the roadAn ~~alien~~ **alien** is skateboarding on MarsA ~~person~~ **monkey** is skateboarding in the street

Figure 3: Results of motion customization of the proposed MoTrans on single reference video. The first row specifies the reference video, showcasing a woman performing a skateboarding tic-tac action. Then the motion is transferred to various subjects specified by the new prompt.

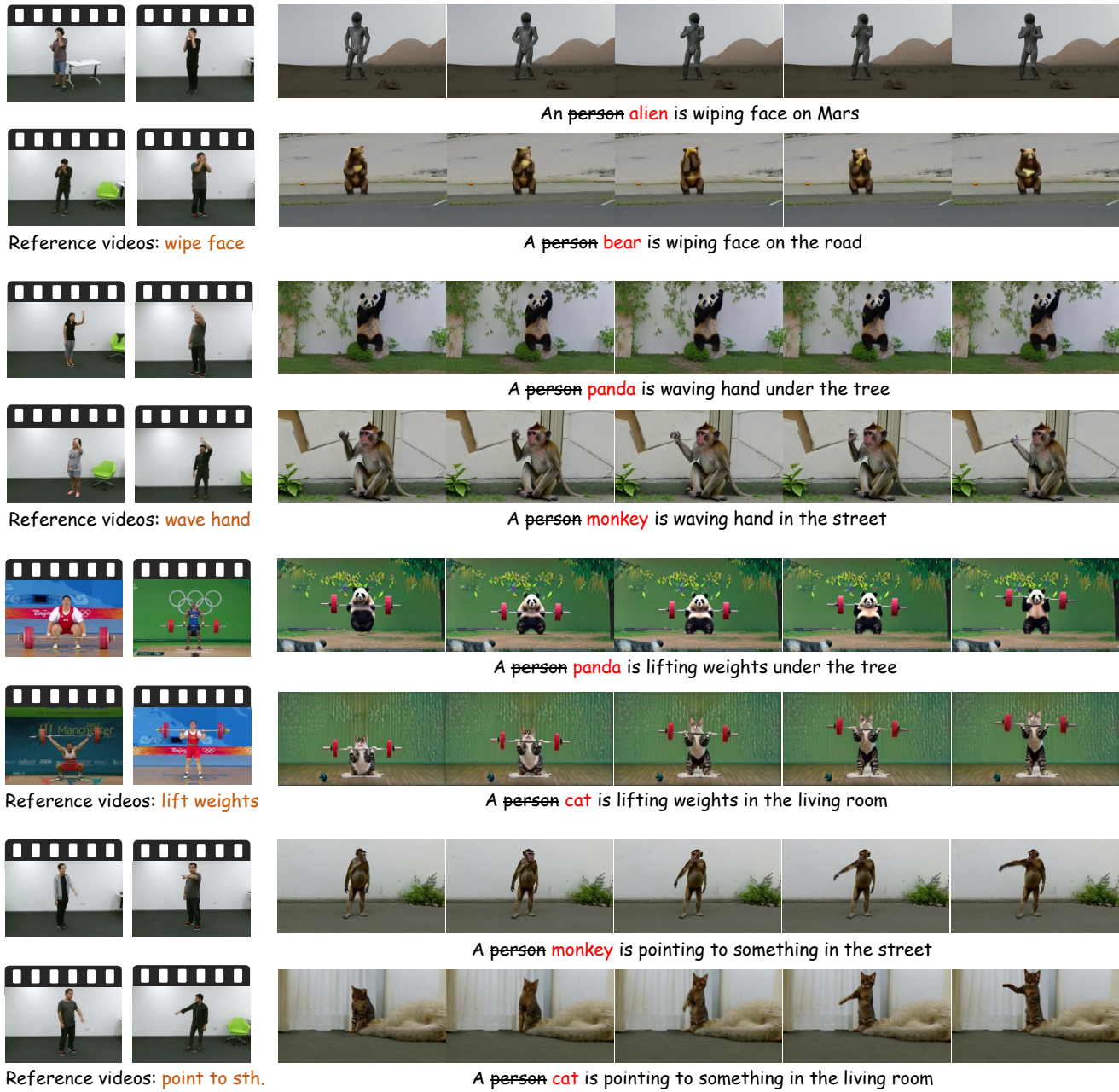
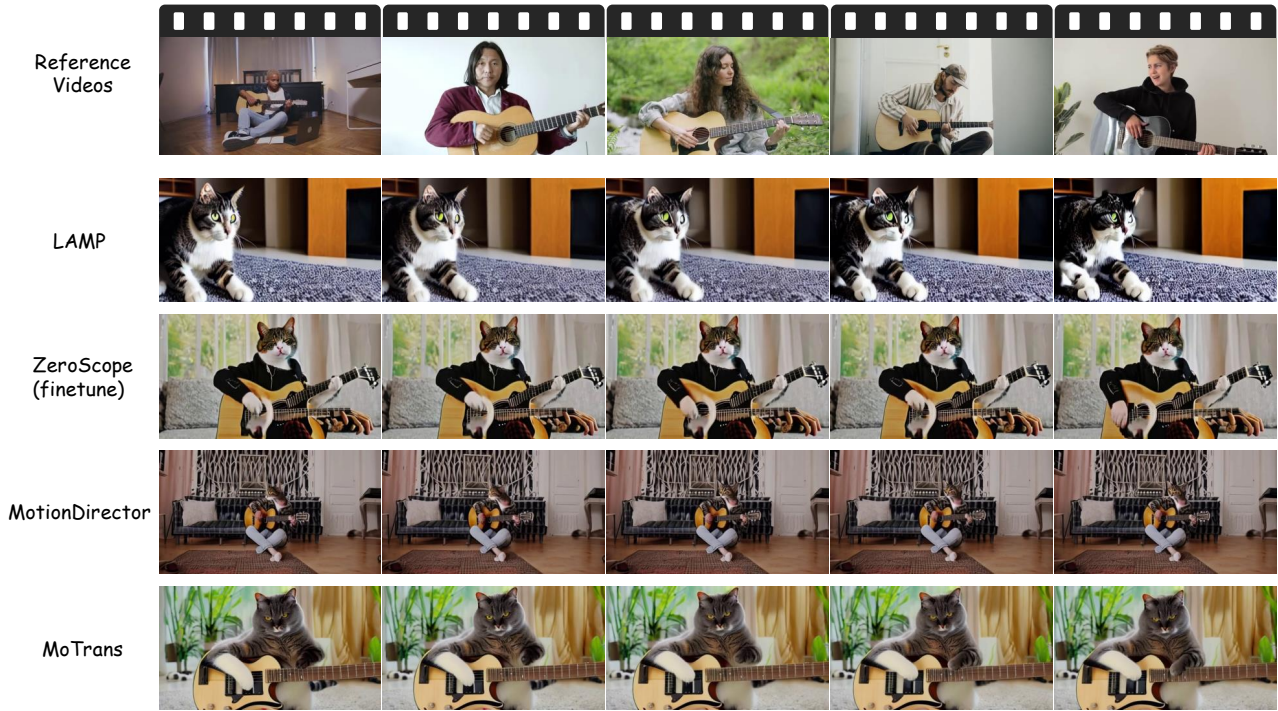


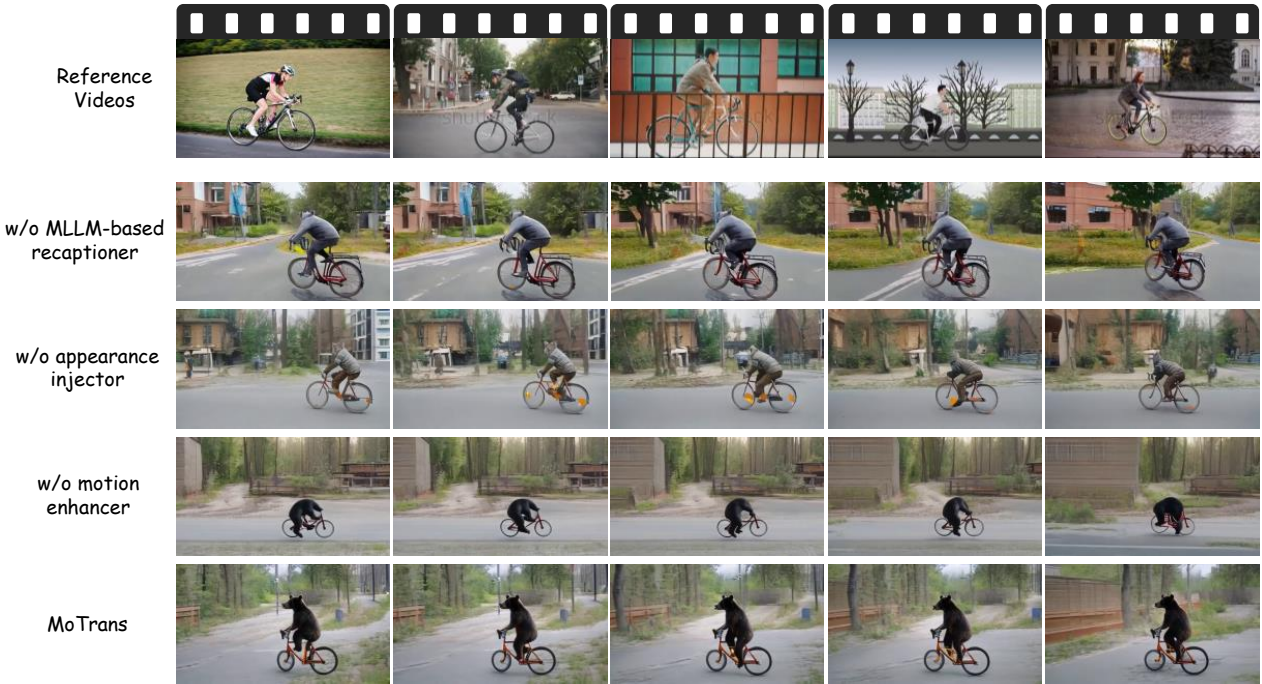
Figure 4: Results of motion customization of the proposed MoTrans on multiple reference videos. The left side displays the reference video, while the right side shows the results of transferring the motion from the reference video to new subjects.





A ~~person~~ cat is playing the guitar in the living room

Figure 5: Additional qualitative comparisons on customized motion transfer given multiple reference videos.



A ~~person~~ bear is riding a bicycle on the road

Figure 6: Additional ablation study.