# Supplementary material for ConMo: Controllable Motion Disentanglement and Recomposition for Zero-Shot Motion Transfer

Jiayi Gao[1,2†] Zijin Yin[2] Changcheng Hua[1] Yuxin Peng[1] Kongming Liang[2]
Zhanyu Ma[2] Jun Guo[2] Yang Liu[1*]
[1] Wangxuan Institute of Computer Technology, Peking University
[2] Beijing University of Posts and Telecommunications
hcc@stu.pku.edu.cn {pengyuxin, yangliu}@pku.edu.cn
{gaojiayi, yinzijin2017, liangkongming, mazhanyu, guojun}@bupt.edu.cn

In the supplementary material, we provide additional information and experimental results relating to ConMo. We begin by providing more details about the experimental setup and user study (Sec.1). Then, we provide more experimental results comparing our method with our baseline DMT [3], focusing on the following three aspects: Multi-subject motion transfer, Fine-grained motion transfer and motion transfer with significant changes in shape (Sec.2). In Sec.3, We present additional results about applications focusing on repositioning and resizing. Finally we discuss the limitation of our method regarding the use of masks (Sec.4)

## 1. Implementation Details and User Study.

**Training details:** To ensure a fair comparison with DMT [3], we use the same parameter settings and feature selection. For the initial noise, we use the same initialization method as in DMT, which involves downsampling and upsampling operations, except for the resize and reposition processes, where we use randomly initialized noise.

**User study details:** For the user study on the right side of Table 1 in main manuscript, we investigated 25 participants to evaluate the effectiveness of ConMo and all the comparison methods on our dataset consists of 26 videos and 56 edited text-video pairs. The user study on the right side of Table 1 in main manuscript primarily assessed three aspects referencing VMC [1] and MotionClone [2]: the motion retention between the input video and the generated video, the motion quality of the generated video and the alignment between the target prompt and the generated video. The survey utilized a rating scale from 1 to 5. To evaluate motion preservation, the participants were asked: " To what extent is the motion from the input video retained in the generated video? " To assess motion quality, participants were asked: " Is the motion in the generated video sufficiently smooth? " To decide text alignment, participants were asked: "Does the generated video semantically align with the target prompt? " The result of Table 1 in main manuscript shows that our method outperforms the baselines in all three aspects.

## 2. More Results Comparing with DMT

In this section, we further illustrate our method through additional visualizations, primarily comparing it with our baseline DMT[3].

In Figure 1, We compare our method with the results generated by DMT[3] on multi-subject videos. In case (a), DMT[3] preserves holistic motion patterns but fails to distinguish individual subject trajectories when two cars share identical motion in the source video, it erroneously generates additional vehicles along the common trajectory rather than establishing precise correspondence between the synthesized SUV and reference race car. This limitation becomes more evident in case (b) involving fine-grained limb movements, where DMT's motion extraction strategy [3] based on compressed global feature only retains dominant foreground actions (the woman's motion) with degraded articulation details, whereas our decoupling strategy successfully preserves nuanced limb dynamics across all subjects. When handling conflicting motions as shown in (c), DMT's [3] entangled motion representation collapses into static outputs when reference subjects exhibit opposing movements, while our approach accurately reconstructs the collision physics through separated motion modeling. Furthermore, in scenario (e) containing subjects with varying motion saliency, DMT[3] tends to suppress subtle movements

† Jiayi Gao is jointly affiliated with Peking University and Beijing University of Posts and Telecommunications, both recognized as co-primary institutions.

∗ Corresponding author.

of less active subjects, whereas our separated representation learning ensures simultaneous preservation of both prominent and latent motions through explicit motion decomposition. Beyond these cases, our method consistently outperforms DMT[3] across other examples in terms of video quality and robustness, with significantly fewer visual distortions and artifacts.

In Figure 2, we compare our method's ability to preserve the original video's fine-grained motion against DMT[3]. In case (a), the duck's inconsistent motion direction and brief initial left-down motion cause DMT, which calculates motion globally, to overlook this process. In contrast, our method, which uses a fine-grained mask based approach, better retains the trajectory details. As a result, the generated video accurately preserves this part of the reference motion. In case (b), the smoke's motion in the original video affects the global motion extracted by DMT[3]. This leads to the car's left-turn process being "counteracted". The generated video shows the car moving in a straight line with many artifacts. In comparison, our method extracts the original drifting motion of the race car independently and transfers it well to the generated video. For cases (c) and (d), our method better preserves fine-grained human limb movements than DMT[3], whose results appear unnatural.

In Figure 3, we further demonstrate that motion can be transferred to subjects with very drastic shape changes (such as from an airplane to a hydrogen balloon, from a train to a person riding a bicycle, etc.) through soft guidance with larger $w_c$. In contrast, DMT[3] is limited by the shape-related information in the original motion. As a result, it often only achieves texture replacement for the generated subjects, failing to realize complete shape changes.

## 3. More Results about Applications

For the applications we proposed in the main text, we also present additional results here focusing on repositioning and resizing:

Regarding the repositioning task, as shown in Figure 4, we have successfully achieved the horizontal and vertical movement of the original subject's motion, making the generated video more aligned with the target prompt's description. Moreover, we have demonstrated that the corresponding repositioning strategy can be transferred to videos with multiple subjects.

For the resizing task, we further prove in Figure 5 that we can control the scaling of the target subject, from enlargement (man to giant) to reduction (man to boy), which is of significant importance for motion transfer that requires size control.

## 4. Limitation

Existing methods are limited by the mask segmentation process. If the mask input is incomplete or if the video contains effects caused by objects that cannot be annotated (e.g., large shadows), it may lead to the decoupled motion still containing information from other subjects, as shown in Figure 6. Such contaminated motion can negatively impact the generated videos.


Figure 1. **Multi-subject motion transfer.** We validate that our method achieves better motion retention for multi-subject videos. In each example, the results in the second row are from ConMo, and the results in the third row are from DMT [3].

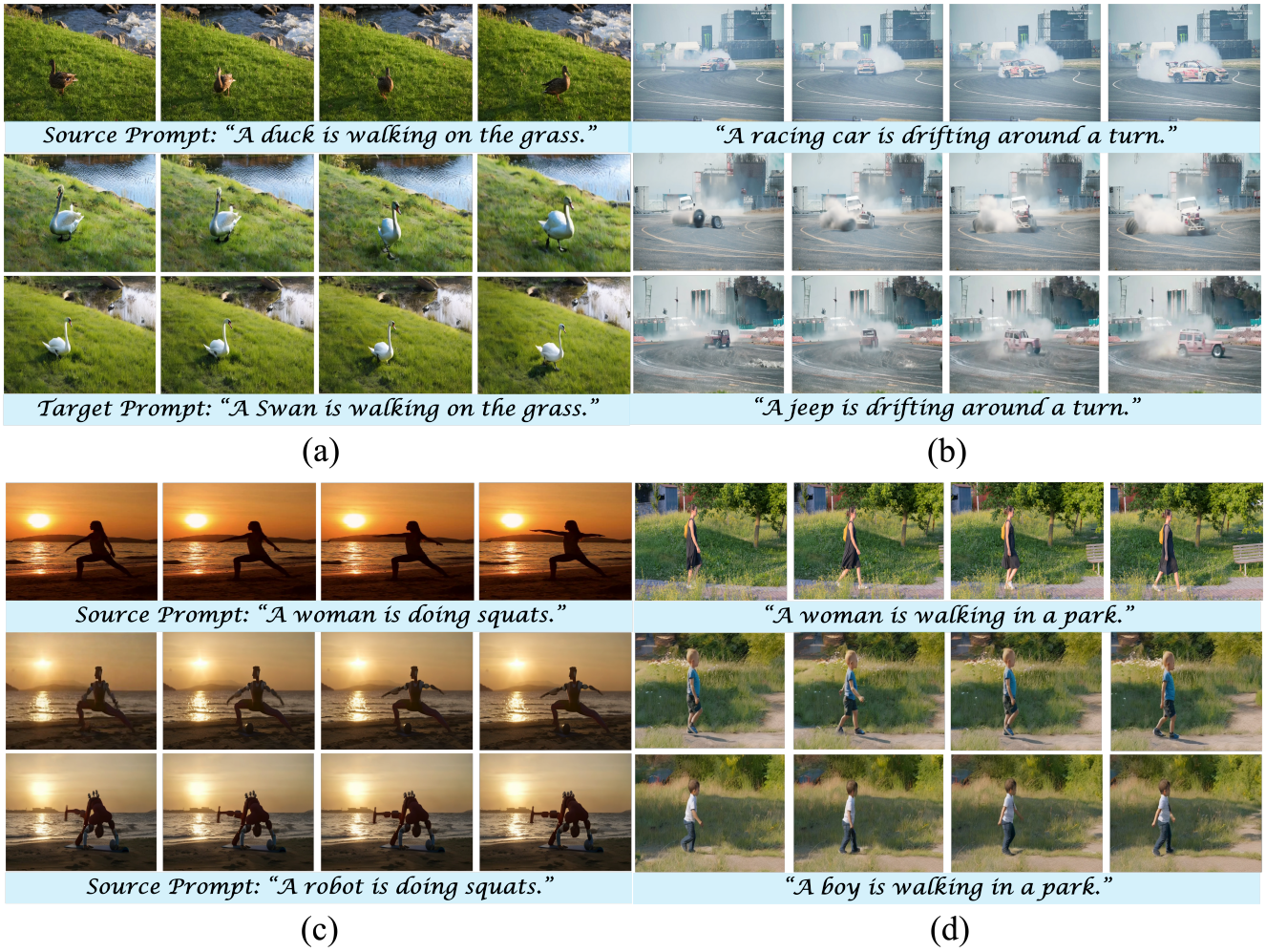



Figure 2. **Fine-grained motion transfer.** We demonstrates that our method effectively maintains fine-grained motion. In each example, the results in the second row are from ConMo, and the results in the third row are from DMT[3].

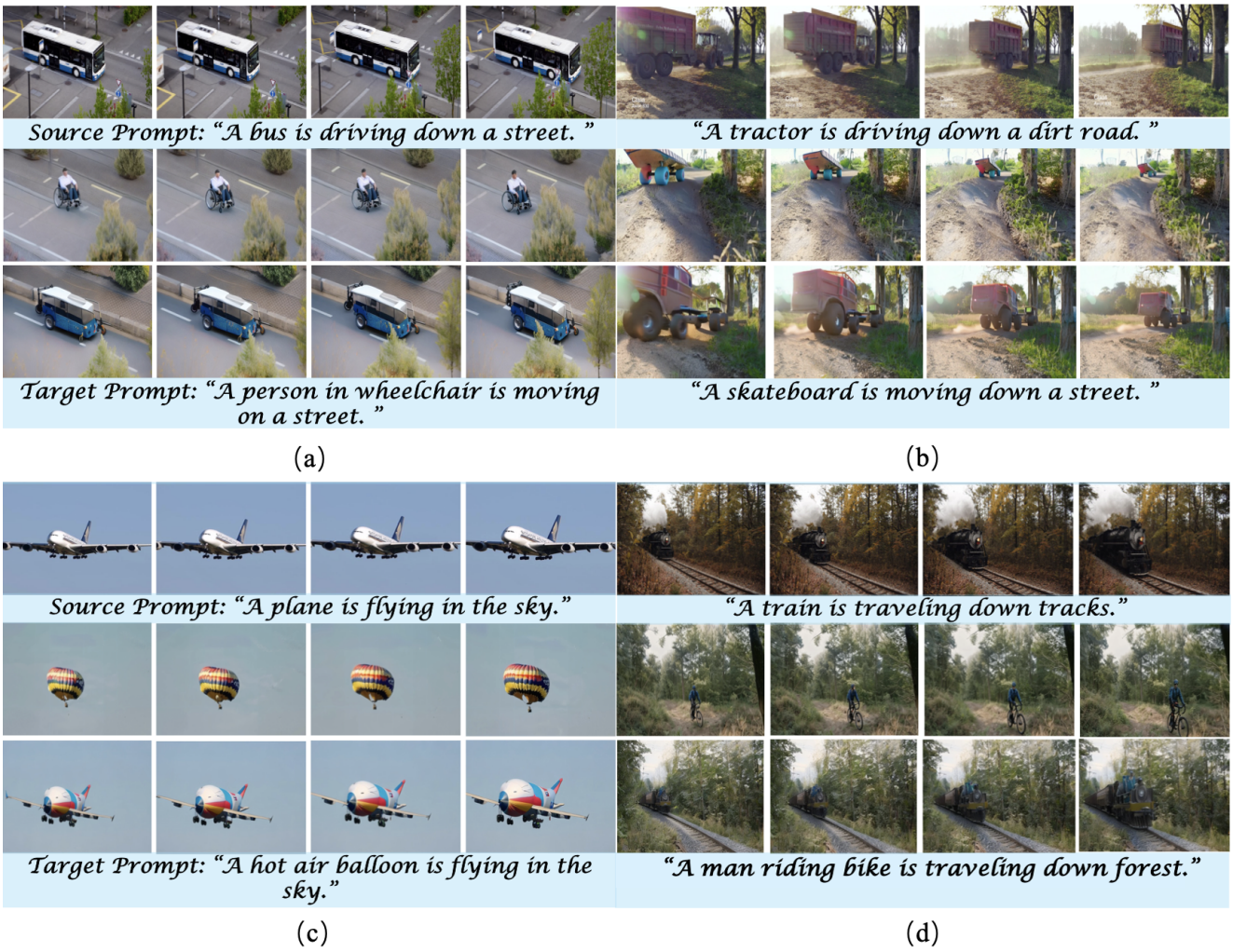
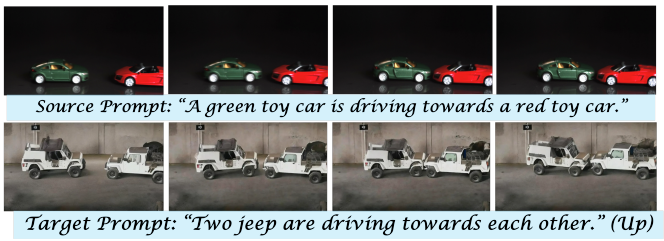



Figure 3. **Motion transfer with significant changes in shape.** We demonstrates the motion transfer results of ConMo compared to DMT [3] when there is a significant difference in shape between the target subject and the reference subject. In each example, the results in the second row are from ConMo, and the results in the third row are from DMT [3].


**(a) Single-Motion reposition**




**(b) Multi-motion reposition**

Figure 4. **Position Control.** In (a), we have demonstrated our ability to reposition the main subject to a specified location (moving left and up), and as shown in (b), this operation can be applied to videos with multiple subjects.


Figure 5. **Size Control.** We have demonstrated our control capabilities over size, which allows the moving subjects in the video to present a more semantically appropriate effect (with 'boy' corresponding to a smaller size and 'giant' corresponding to a larger size).


Figure 6. **Limitation.** In the process of removing the motion of the car on the left side of the original video, the segmentation model failed to account for the effects of the corresponding object, specifically the shadow in the video. As a result, the motion of the shadow can negatively impact the generated video.

## References


[1] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9212–9221, 2024. 1

[2] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024. 1

[3] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8476, 2024. 1, 2, 3, 4, 5