MotionDiff: Training-free Zero-shot Interactive Motion Editing via Flow-assisted Multi-view Diffusion
Supplementary Material

In the supplementary material, we provide more method details about MotionDiff in Section 1, along with explanations to facilitate readers in deepening understanding our work. In Section 2, we present more motion editing results. In Section 3, we analyze the limitation of MotionDiff and consider potential solutions.

1. More Details about MotionDiff

1.1. Interactive User Operations

In Figure 1, we demonstrate how a user can obtain optical flow from a given single-view image. Specifically, we first use the mouse to select the object by drawing a Region of Interest (ROI), and then apply SAM [3] to obtain the segmented object. Following the pipeline principle of Motion Guidance [1], we can then obtain optical flow according to corresponding motion modes, such as translation, scaling, rotation, stretching, and so on.

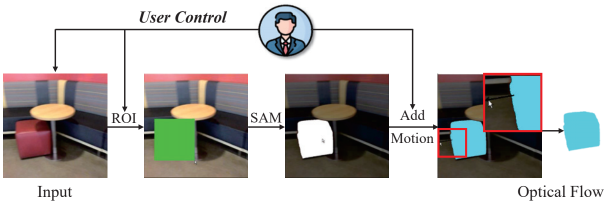


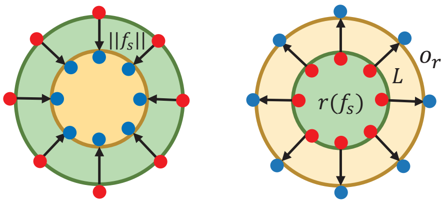
Figure 1. The interactive user operation pipeline of GUI.

1.2. More Formulas Details about PKM

In the following, we provide more details about PKM.

1) For translation, we aim to utilize $\boldsymbol{P}_{\boldsymbol{so}}$ and $\boldsymbol{P}_{\boldsymbol{sm}}$ obtained from a single-view image and its optical flow $\boldsymbol{f}_{\boldsymbol{s}}$ to estimate the 3D offset $\boldsymbol{p}_{\boldsymbol{off}}$. Therefore, in Eq. 6 of the main paper, we calculate the average offset between the $\boldsymbol{P}_{\boldsymbol{so}}$ and $\boldsymbol{P}_{\boldsymbol{sm}}$, and take it as the motion offset between the 3D points $\boldsymbol{P}_{\boldsymbol{o}}$ and $\boldsymbol{P}_{\boldsymbol{m}}$.

2) For shrinkage, we aim to use $\boldsymbol{f}_{\boldsymbol{s}}$ to calculate the scaling factor s_f. As shown in Figure 2(a), the shrinkage area must be contained within the original area. Thus, we can calculate the average magnitude of $\boldsymbol{f}_{\boldsymbol{s}}$, and divide it by the maximum optical flow magnitude to obtain the s_f, as demonstrated in Eq. 7 of the main paper.



(a) shrinkage (b) enlargement

Figure 2. The diagram of scaling motion.

3) Enlargement motion generates novel region outwards, as shown in Figure 2(b). Thus, it is challenging to represent this motion solely based on the magnitude of optical flow. Therefore, we use the linear sampling function to expand the original region $r(\boldsymbol{f}_{\boldsymbol{s}})$ based on $\boldsymbol{f}_{\boldsymbol{s}}$ and take the union with it to obtain $\boldsymbol{o}_{\boldsymbol{r}}$. Then, we divide the region $\boldsymbol{o}_{\boldsymbol{r}}$ by $r(\boldsymbol{f}_{\boldsymbol{s}})$ to obtain the s_f for the enlargement motion. The aforementioned process is demonstrated in Eq. 8 of the main paper.

2. More Experimental Results

In this Section, we provide more motion editing results to further showcase the performance of MotionDiff, as shown in Figure 3 to 6. Figure 3 follows Figure 7 in the main paper, and demonstrates more results about how MotionDiff can progressively achieve motion editing as the diffusion steps increase.

3. limitation and Analysis

Despite achieving high-quality results in complex multi-view motion editing tasks, MotionDiff still possesses minimal limitations. **1)** Specifically, for small objects, like the scaling motion examples in Figure 5, the motion object may lose a small amount of edge details. This is due to the inevitable degradation of texture details when an image is encoded to 64 × 64 and then recovered using VAE (including the edge details of the image and the jagged edges of the mask). This issue is common in diffusion models based on latent space. In contrast, for larger objects (as shown in Figure 6), this situation is improved. **2)** Additionally, for other motions, establishing corresponding models in PKM is required, while we believe it is relatively straightforward to implement. For example, for wind blowing through plants, we can apply wind fields with varying magnitudes and directions to 3D points. Then, PKM can obtain multi-view optical flows, enabling subsequent editing. Moreover, for popular 3DGS, it is theoretically to use PKM to change the position of the gaussian, thereby editing 3DGS.

In the future, we will try using higher resolution or more general motion representation, while also improving the corresponding guidance strategies. In addition, fine-tuning LoRA [2] for enhancing the details of diffusion models might also be a good option. These may help resolve the issue mentioned above.

References

[1] Daniel Geng and Andrew Owens. Motion guidance: Diffusion-based image editing with differentiable motion estimators. In *The Twelfth International Conference on Learning Representations*.

[2] 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. In *International Conference on Learning Representations*, 2022.

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

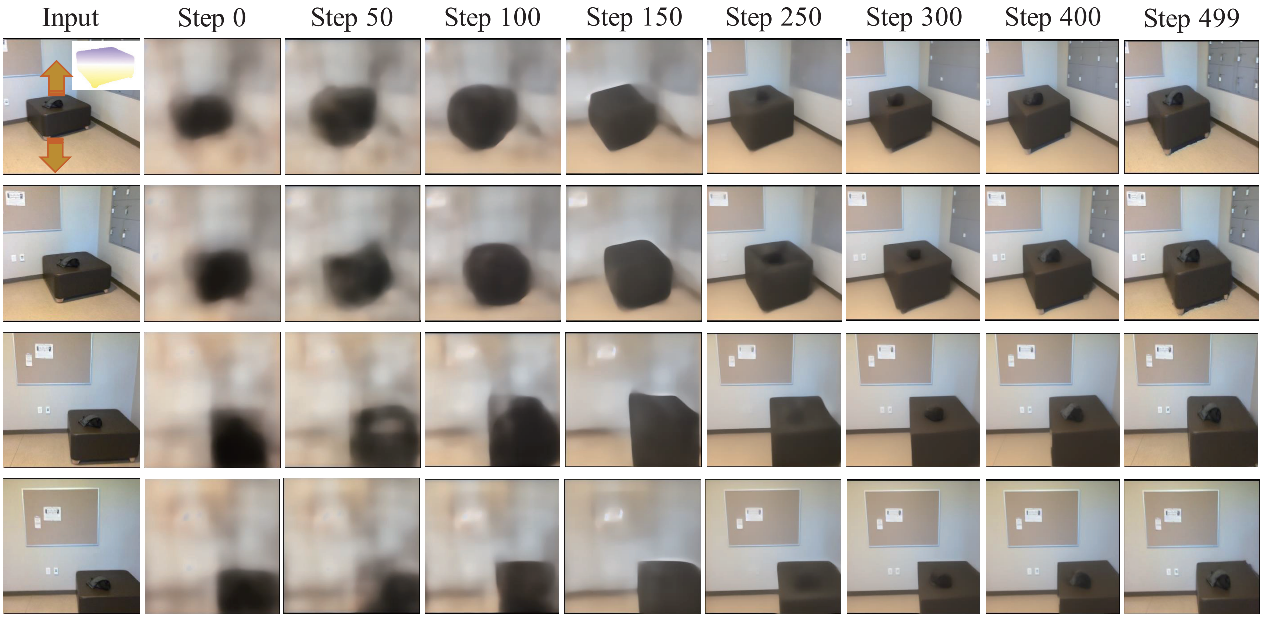


Figure 3. Visualization of different diffusion steps, demonstrating under our effective guidance strategies, MotionDiff gradually produces motion results.

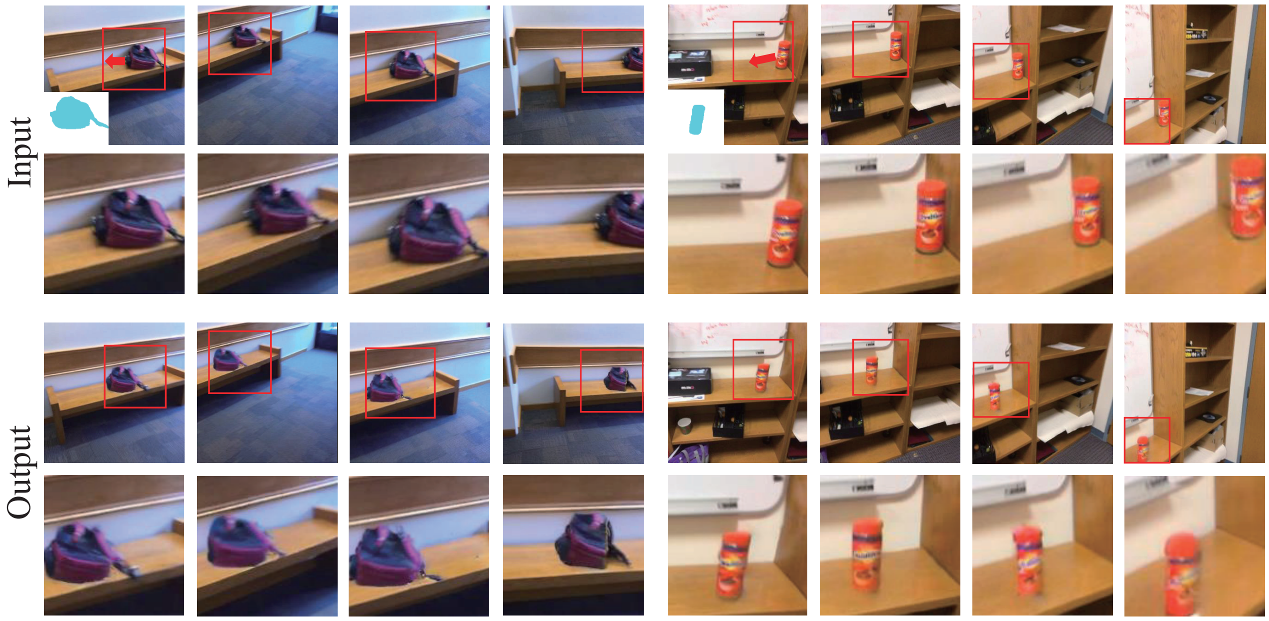


Figure 4. More motion editing results of MotionDiff.

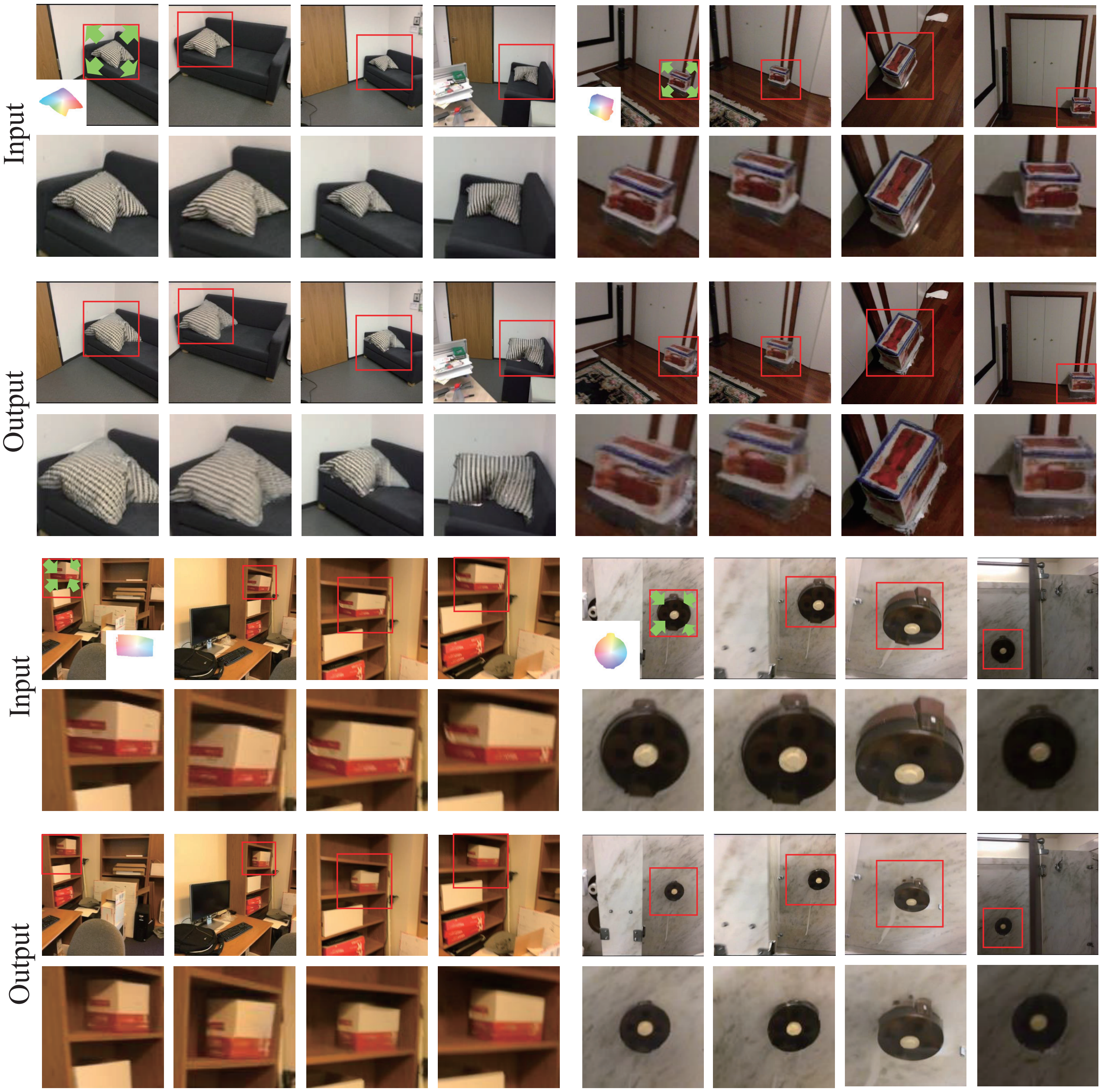


Figure 5. More motion editing results of MotionDiff.

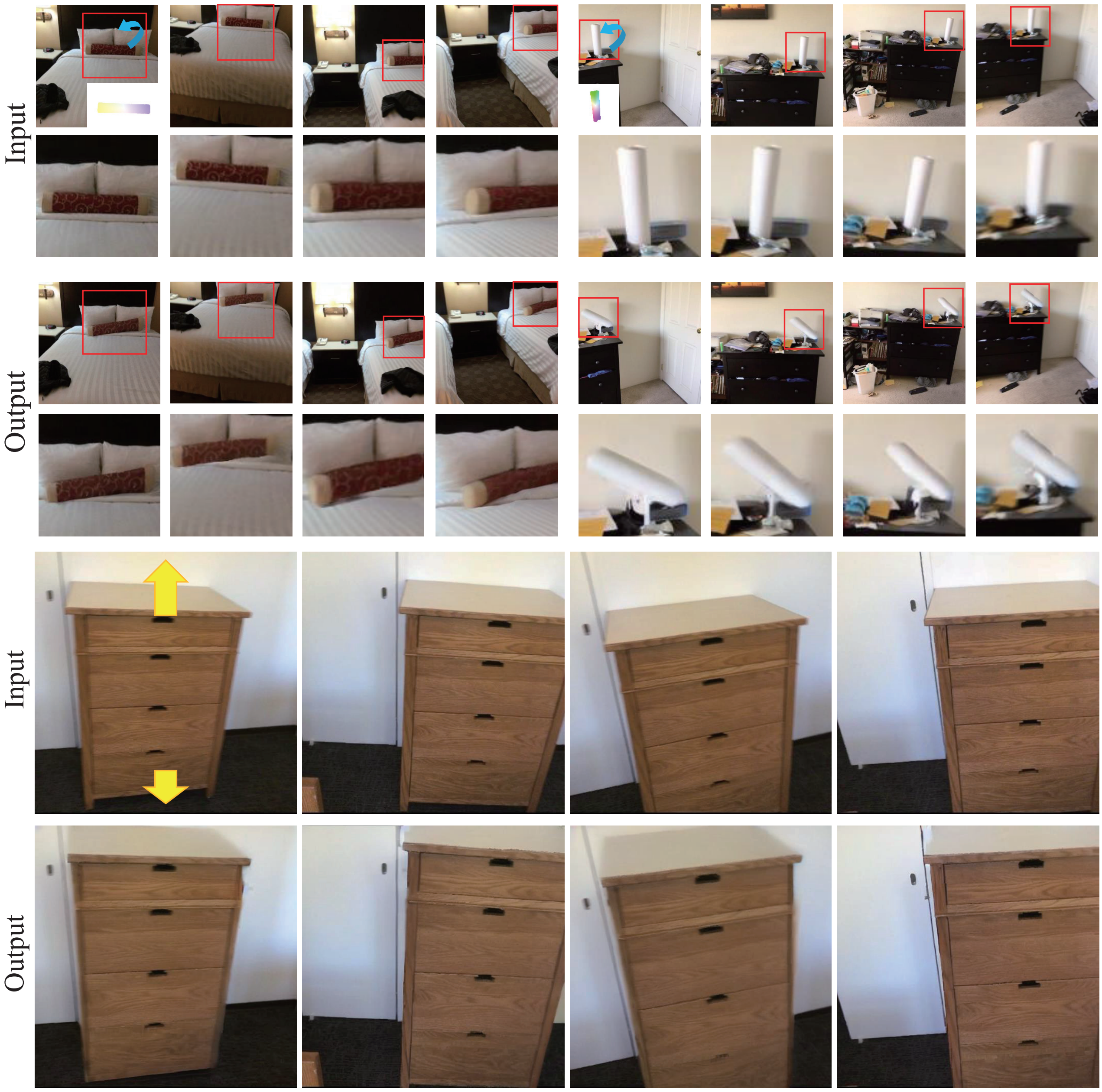


Figure 6. More motion editing results of MotionDiff.