Perception-as-Control: Fine-grained Controllable Image Animation with 3D-aware Motion Representation

Supplementary Material

A. More Results of Perception-as-Control

Perception-as-Control can achieve precise camera motion control adhering to any designated camera trajectory, multi-instance object motion control with adaptive control granularity, and harmonious collaborative motion control.

A.1. Camera-only Motion Control

To demonstrate the superiority of Perception-as-Control in camera-only motion control, we show the generated results under different camera trajectories in Fig. 9. Our approach places no restrictions on the camera trajectories and supports camera motion control under basic camera trajectories such as Pan Left and Dolly In, their combinations such as Pan Left + Zoom In, and any user-provided camera trajectories. Please refer to our webpage: *Fine-grained Collaborative Motion Control - Camera-only Motion Control* for the generated videos.

As shown in Fig. 9, our framework can generate videos that adhere to any designated camera trajectories. The generated videos maintain stable and consistent spatial structures owing to the world envelope in our 3D-aware motion representation, which demonstrates the direction and magnitude of camera movements and enhances overall perception. Moreover, by adding several unit spheres to essential object parts, the generated videos are aware of depth information and can accurately reflect perspective effects under different focal lengths, as shown in Fig. 10.

A.2. Object-only Motion Control

For object-only motion control, we show the generated results in Fig. 11 to demonstrate the superiority of our approach in achieving flexible and precise multi-instance object motion control with adaptive control granularity. Please refer to our webpage: *Fine-grained Collaborative Motion Control - Object-only Motion Control*.

For visualization clarity, we project the unit sphere centroids onto the reference image to show how they change over time, using colors to represent the direction of movement. As shown in the left part of Fig. 11, we control an increasing number of hot air balloons in the reference image, and the generated videos precisely reflect the movements of each one. In the right part of Fig. 11, when additional controls are applied to the bamboo, the proposed method can adaptively determine the control range for each unit sphere and achieve harmonious collaborative control results.

A.3. Collaborative Motion Control

Our framework also supports collaborative control of both camera and object motions. Please refer to our webpage: *Fine-grained Collaborative Motion Control - Collaborative Motion Control* for the generated videos.

As illustrated in Fig. 12, we control both camera and object motions and visualize the corresponding 3D-aware motion representation. Our method can handle both large object motions and delicate object motions. The generated videos adhere to the specified camera motion and accurately present the adapted object motions.

B. More Details of ++ Versions

To demonstrate that our 3D-aware motion representation is orthogonal to base model architectures and effective for different base model architectures, we incorporate 3D-aware motion representation into the SVD [2] version of CameraCtrl [9] and MotionCtrl [37] to empower them with fine-grained collaborative motion control capability. Without making major changes to their architectures, we successfully incorporate our motion representation into them, as shown in Fig. 13. The experimental results verify the effectiveness of our 3D-aware motion representation and the superiority of our framework.

C. More Details of Ablation Study

In Fig. 14, we illustrate the three training strategies included in our ablation study (Table 4). *One-stage* refers to directly optimizing all the modules mentioned using mixed data and *Two-stage* refers to merging the last two stages into one.

D. More Details of Applications

By carefully interpreting user intentions into manipulation on 3D-aware motion representations, our framework can support various potential motion-related applications. In this paper, we describe four kinds of motion-related applications, including motion generation, motion clone, motion transfer, and motion editing. The intuitive illustration is shown in Fig. 15. Please refer to our webpage: *Potential Applications* for the generated videos.

Motion Generation refers to the task setting of interactive image animation that allows users to draw 2D or 3D point trajectories on the reference image and set an arbitrary camera trajectory with a carefully designed user interface. In this application, we employ an off-the-shelf met-

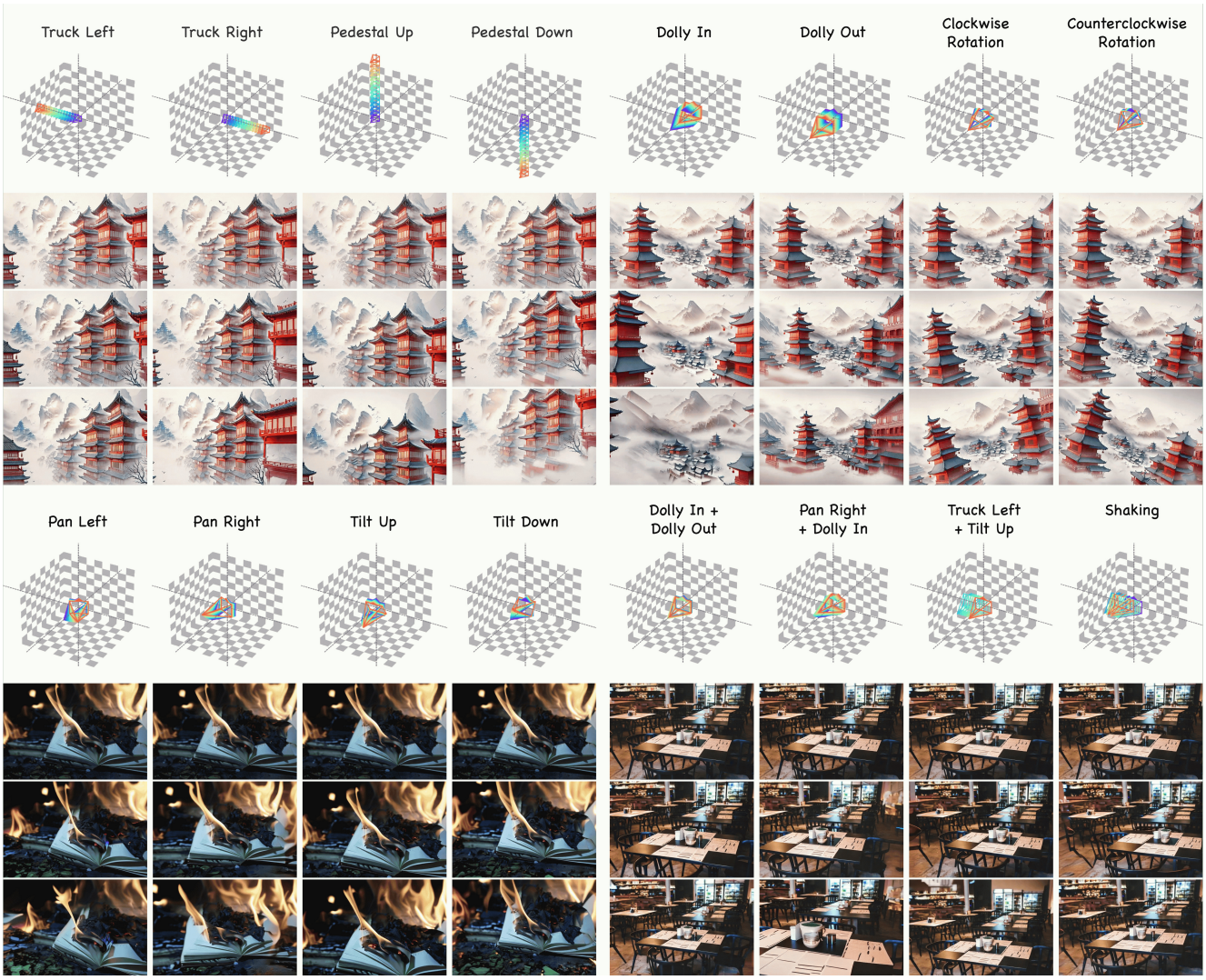

Figure 9. Results of camera-only motion control using our Perception-as-Control.

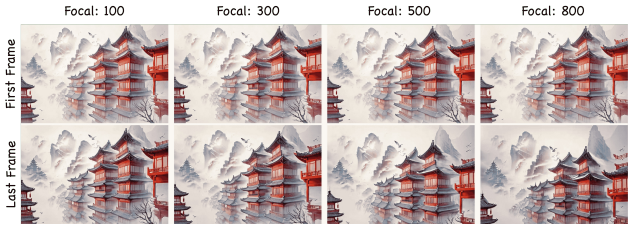

Figure 10. Results with the same camera extrinsic sequence (Pedestal Up) under different focal lengths.

ric depth estimation algorithm on the reference image, construct our 3D-aware motion representation based on the estimated results. We manipulate unit spheres corresponding to the starting point of each trajectory, and perceive the representation based on user-designated camera trajectory. The perception results are used as motion control signals. Ideally, the object parts corresponding to the starting point of each trajectory should move along the trajectory in the generated video. The difficulty lies in correctly understanding the control range of each trajectory point and generating reasonable and accurate motion videos under the control of multiple trajectories.

Motion Clone refers to extracting all motions from a source video and cloning them for driving a reference image. Users begin by providing a source video, which is used to construct our 3D-aware motion representation. They then edit the first frame of the source video to create a reference image. After that, we use the constructed 3D-aware motion representation and the reference image as inputs of the proposed framework to obtain the motion clone results.

Motion Transfer, similar to motion clone, refers to

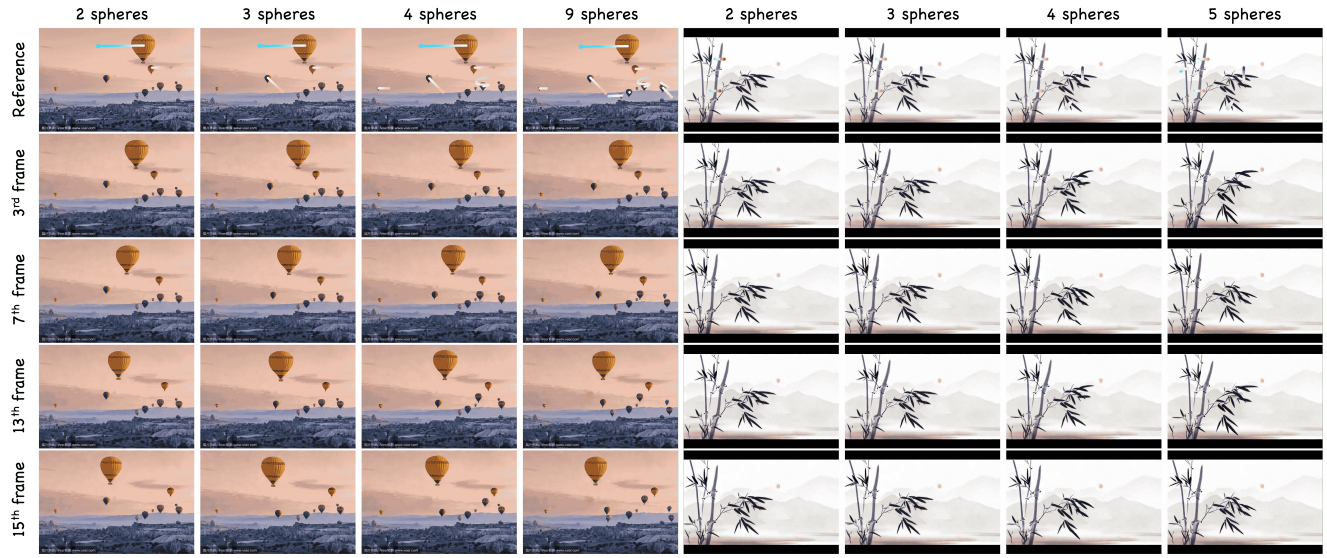

Figure 11. Results of object-only motion control using our Perception-as-Control.

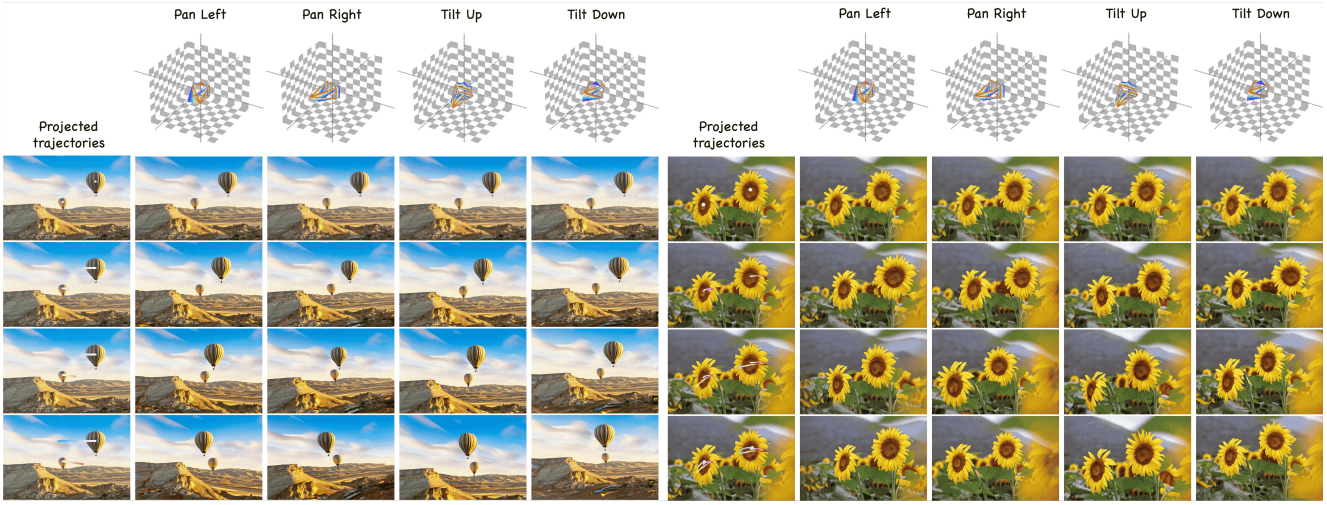

Figure 12. Results of collaborative motion control.

transferring local object motions from the source video to the reference image. In this case, the object may differ from that in the source video in terms of scale and location. To meet the needs of this application, we first find the semantic correspondences between the first frame of the source video and the reference image. Then, we construct our 3D-aware motion representation from the source video and locate essential unit spheres based on semantic correspondence points on the first frame. after that, we adaptively relocate and rescale them to match those semantic correspondence points on the reference image. Finally, the manipulated 3D-aware motion representation can be used to animate the reference image.

Motion Editing has greater control freedom than the above settings. Based on their source video, users provide segmentation masks and instructions to edit motions inside or outside these masks, *i.e.*, editing fine-grained local motions in user-specified regions. In this way, users could freeze or modify motions of several objects or the background, *etc*. To achieve such operations, we first construct 3D-aware motion representation from the source video, and select unit spheres inside the regions to be edited. Then, we manipulate those selected unit spheres according to user intentions and use the manipulated representation to reanimate the first frame of the source video.

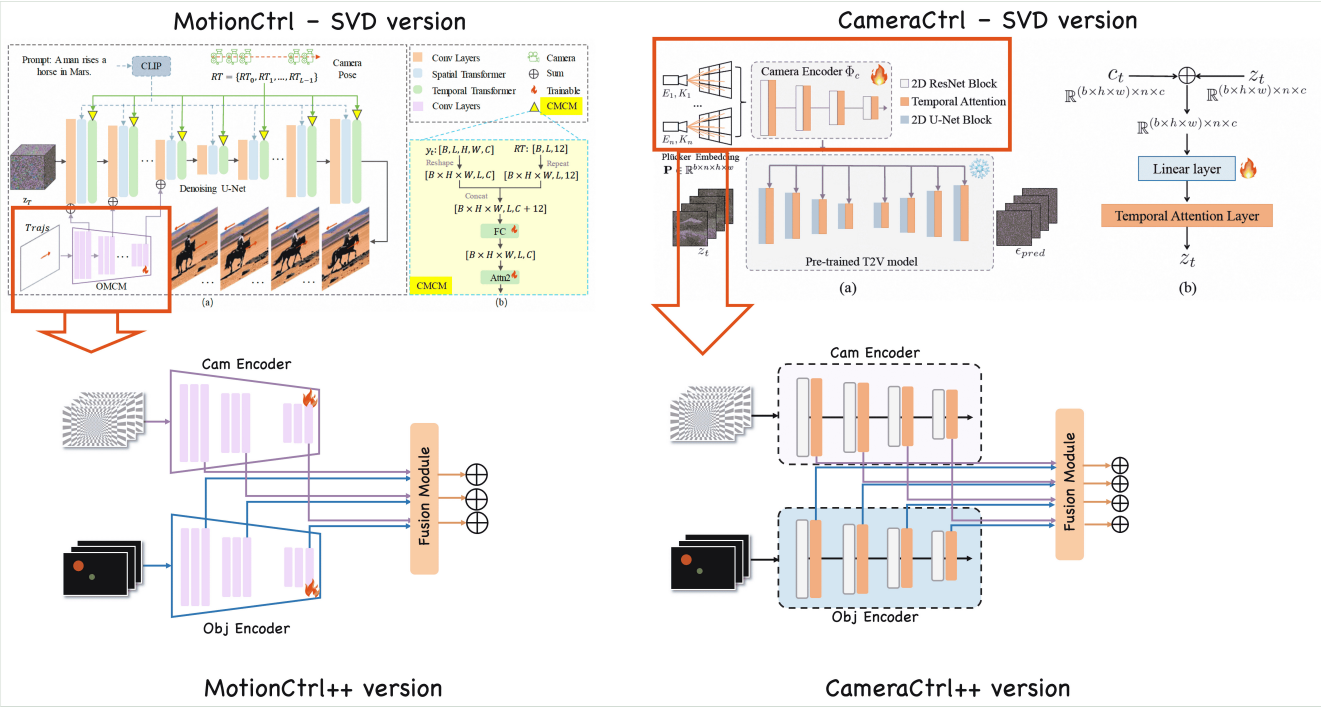

Figure 13. The frameworks of MotionCtrl++ and CameraCtrl++.

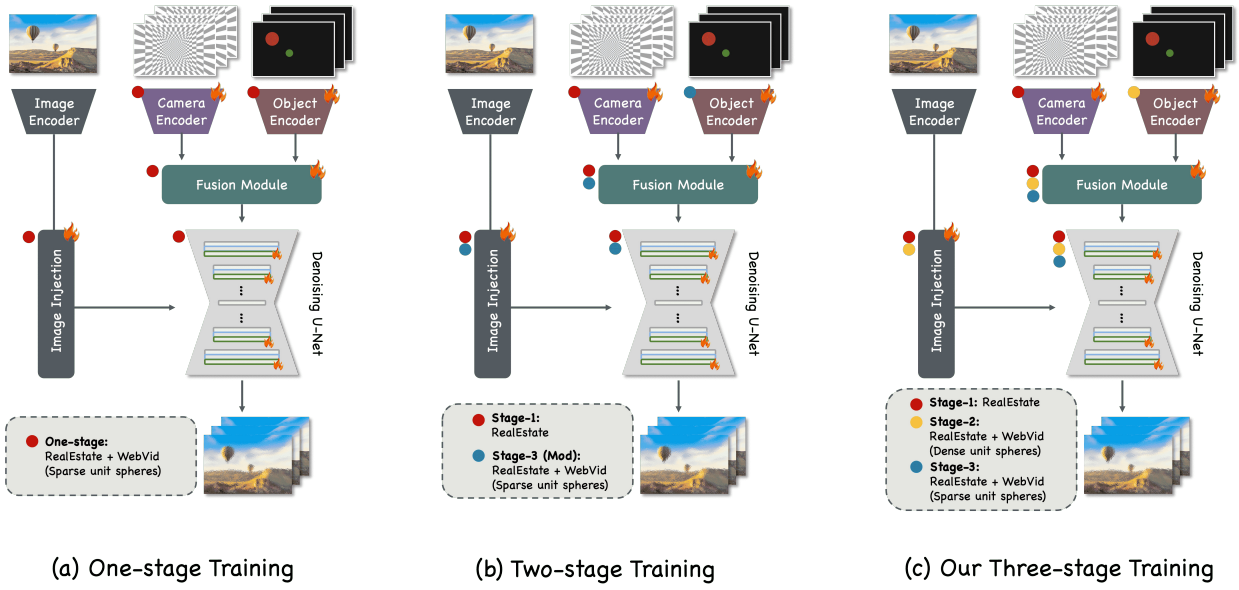

Figure 14. Illustration of different training strategies in ablation study.

E. Limitations and Future Work

Our work introduces a novel 3D-aware motion representation for motion-controllable image animation task. Although our image animation framework that incorporates 3D-aware motion representation achieves promising and insightful results in fine-grained collaborative motion control, it still has some limitations.

First, we simplify object parts as unit spheres, which can hardly reflect object rotation. As shown in the first row of Fig. 16, the relative positions between the two billiards and the markers on them remain unchanged, *i.e.*, those bil-

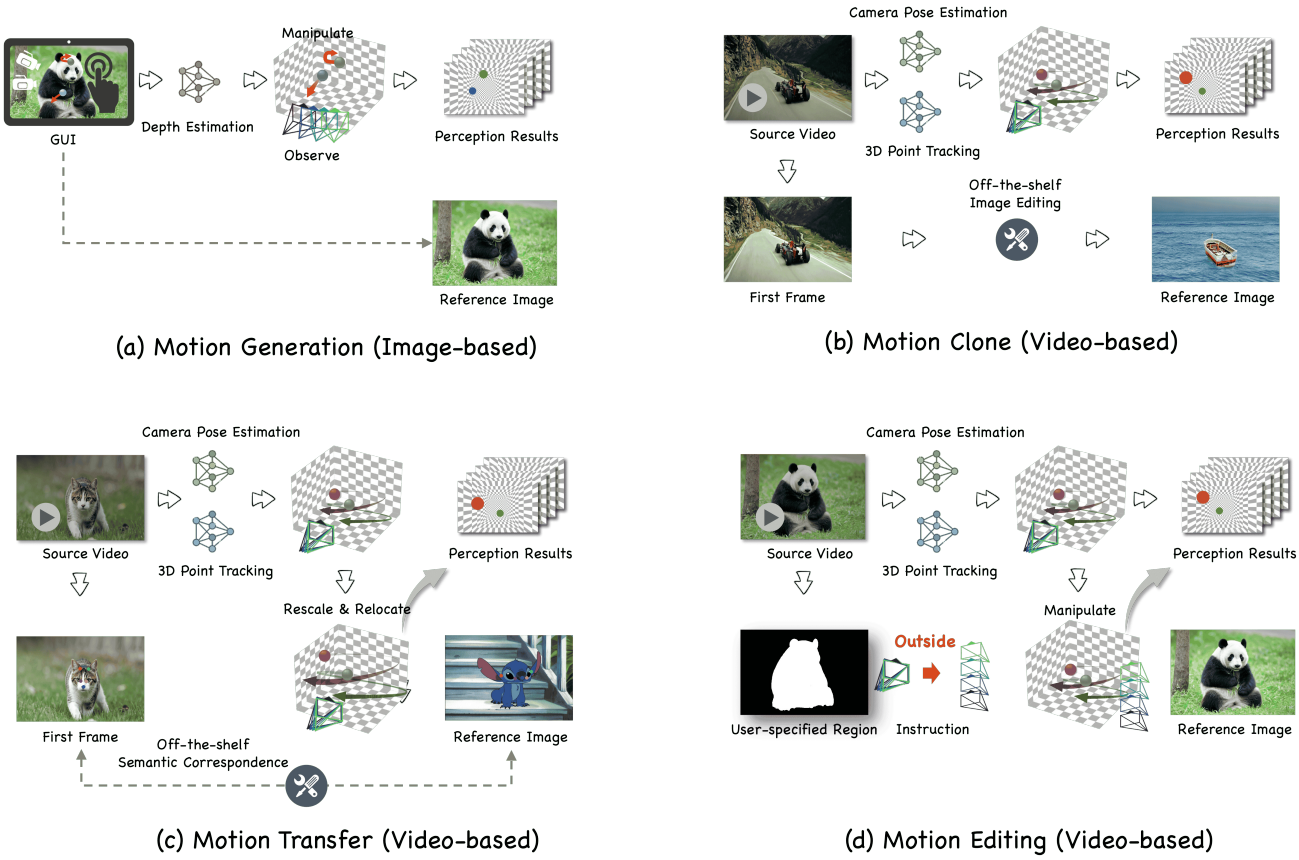

Figure 15. Illustration of different motion-related applications.

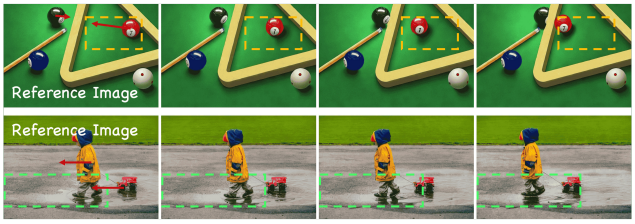

Figure 16. Limitations of Perception-as-Control.

liards only move in translation without rotation. In future work, exploring more powerful alternatives may help express more comprehensive forms of movement.

Second, our approach is constrained by base model performance and cannot handle human-related motions satisfactorily. As shown in the second row of Fig. 16, the boy is moving forward without reasonable stepping movements. It is promising to incorporate the proposed 3D-aware motion representation into more powerful base model architectures.

Also, those off-the-shelf algorithms for constructing our 3D-aware representations influence the performance of our approach. Future work could employ state-of-the-art deep visual SLAM works for more precise estimation.