EgoControl: Controllable Egocentric Video Generation via 3D Full-Body Poses

Supplementary Material

6. Additional Implementation Details

We applied classifier-free guidance using the input context frames for the Cosmos baseline and EgoControl. During training, the observed context frames were randomly dropped with a probability of 0.2, and we pre-extracted the latent embeddings to reduce computational overhead. We used a global batch size of 1024, learning rate of $1.1 \cdot 10^{-5}$, and trained our final version of EgoControl for 30.000 iterations on H100s GPUs. At inference time, we enabled video guidance with a guidance weight of 2, which resulted in better qualitative results. We conducted all experiments without applying the default Cosmos Guardrail to the conditioning frames or the generated videos, ensuring that it did not influence the results. For the body control alignment evaluation, we used the `small` version of SAM2. The trained model and evaluation scripts are available at `cvg-bonn.github.io/EgoControl`.

7. Global Motion Evaluation

While in the previously described experiments the global motion control evaluation is performed using VIPE to extract the camera motion from videos, we report in Tab. 6 results using VGGT [52] for extracting camera motion. This evaluation shows similar relative improvements across our experiments, further showing the positive effects of full-body conditioning for effective control.

Experiment	VIPE		VGGT	
	TransErr	RotErr	TransErr	RotErr
Base Cosmos	16.53	15.60	6.89	12.29
Finetuned	9.93	13.65	5.01	11.28
Head Control	<u>5.16</u>	<u>3.29</u>	<u>2.83</u>	<u>3.46</u>
Body Control	**4.90**	**2.96**	**2.40**	**2.76**

Table 6. Motion evaluation with VGGT instead of VIPE.

8. Additional Qualitative Results

We proceed describing some additional qualitative results.

Baseline comparison. In order to show the effect of our full-body pose control in the generated videos, we include a visual comparison of models using different control information in Fig. 7. More specifically, it can be seen that the simple finetuned version of Cosmos (row 2), i.e., with no other information than the past context frames, does not follow the body pose and camera view compared to the ground truth video. This behaviour is expected due to multiple possible feasible futures given only the context frames. In row 3, we show the generated video using only the head pose for control. The camera view starts to follow a similar path to the ground truth, however, the hand movements are still different from what they should be. Finally, by including the full-body pose (row 4), EgoControl manages to control both the camera view and body movements, precisely generating the motion of the arms going upward and the complex interaction with the sheet.

Different context. In Fig. 8, we show another example of applying the same sequence of poses, in this case extending the left arm while making a slight upper-body rotation, to three different initial context frames. The results show that regardless of the initial state of the person, the generated video is successfully controlled by the conditioning pose. It should be noted that the location of the hands in the 2D pixel space is not necessarily the same across different contexts. Depending on the initial head pose, the view of the body can be different, given the correctly followed pose.

Different control pose. The last example of control abilities of EgoControl is shown in Fig. 9 by applying different control poses to the same initial context frames. In the first row, we apply a pose that differs significantly from the starting one in the context frames. In this case, the model transitions smoothly toward the target pose, effectively bringing the hands together. The second and third row present other types of movements, including controlling only the left arm or both arms simultaneously. Interestingly, these examples show a physically plausible interaction with the object being held.

Higher resolution. We finetune the final version of EgoControl on videos at 960×960 resolution for additional 2.000 iterations. This short finetuning stage enables EgoControl to generate higher resolution videos. Examples are shown at `cvg-bonn.github.io/EgoControl`.

Longer video generation. To generate longer sequences, we run EgoControl autoregressively by iteratively feeding the last generated frames back as conditions. Fig. 10 shows two examples, each with a duration of 8 seconds.

Ego4D. To assess generalization, we also take a subset of 250 videos (200/50 train/val) of Ego4D [15]. The dataset, however, does not contain any body pose annotation. The final version of EgoControl is thus finetuned on a subset of Ego4D without body poses. With just 1.000 iterations,

the model generalizes, showing control abilities even with a new camera setting and body appearance (Fig. 11).

Failure Cases. Fig. 12 shows some failure cases caused by physically implausible sequences of poses. In the first row, the person walks through the table. In the second row, the person carries an object and walks through a window. In most cases, implausible control sequences can be recognized by strong artifacts. For instance when the person walks through the window, the object and arm disappear. Automatically detecting physically implausible sequences of poses is an interesting future work.

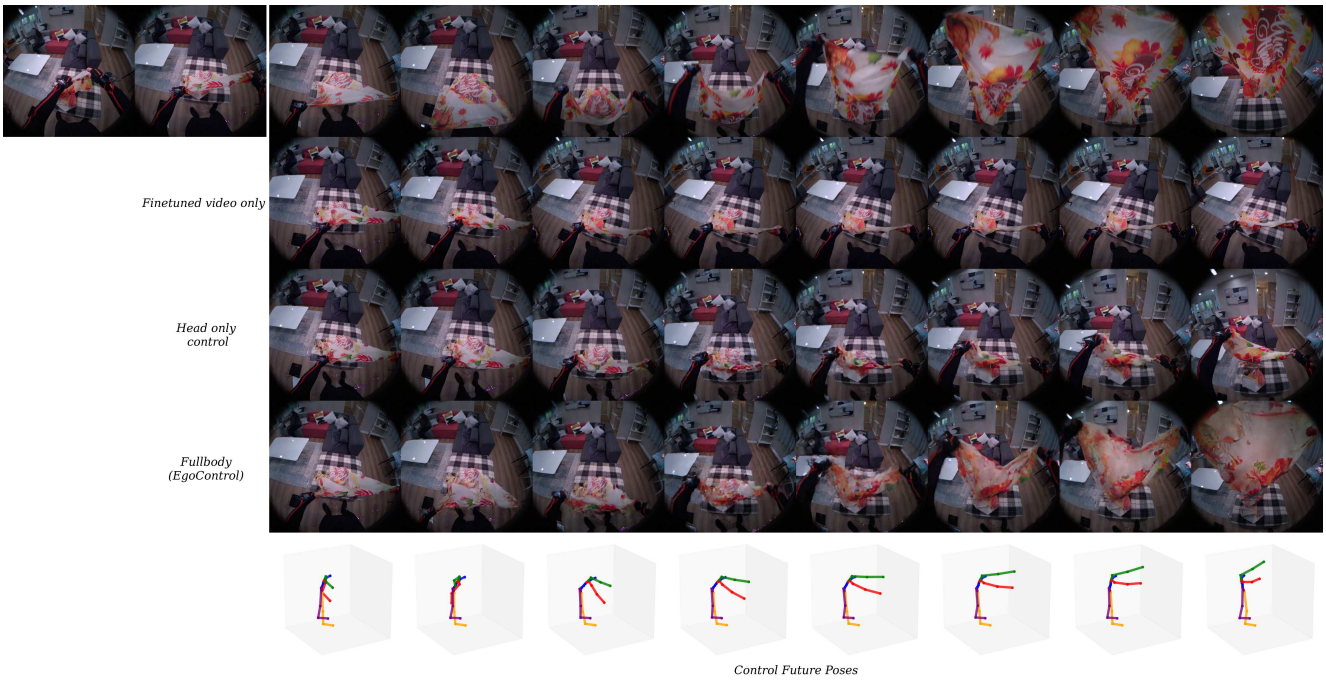

Figure 7. Comparing EgoControl (fourth row) to ground truth (first row), finetuning (second row), and head only control (third row).

Figure 8. Applying the same sequence of human poses to different context frames. EgoControl shows accurate pose alignment for all three scenarios.

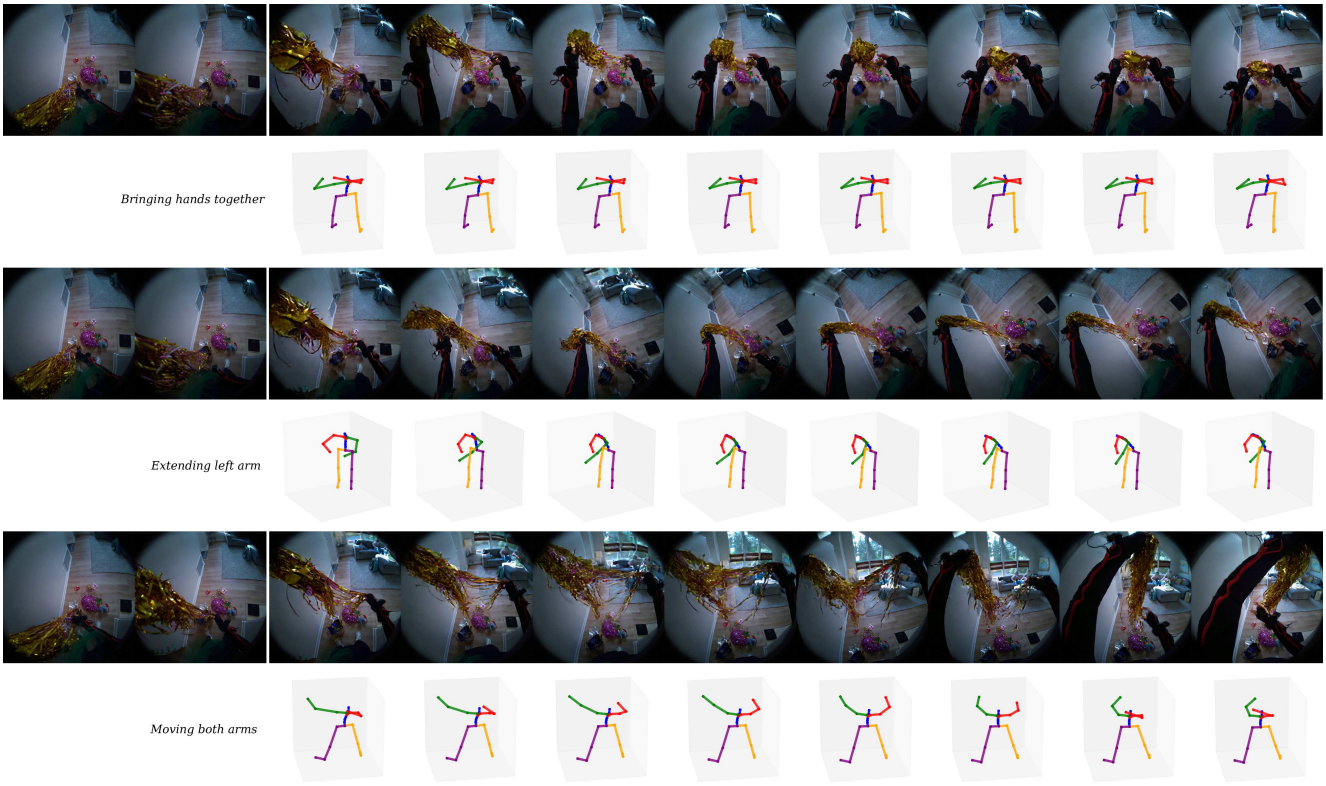

Figure 9. Applying different sequences of human poses to the same context frames. EgoControl is capable of generating videos following the different body movements.

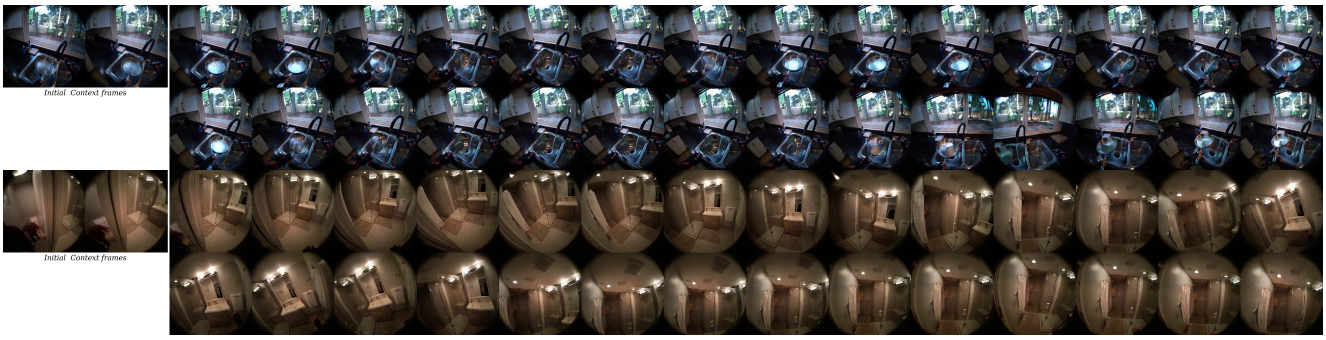

Figure 10. Videos of 8 seconds generated by EgoControl.

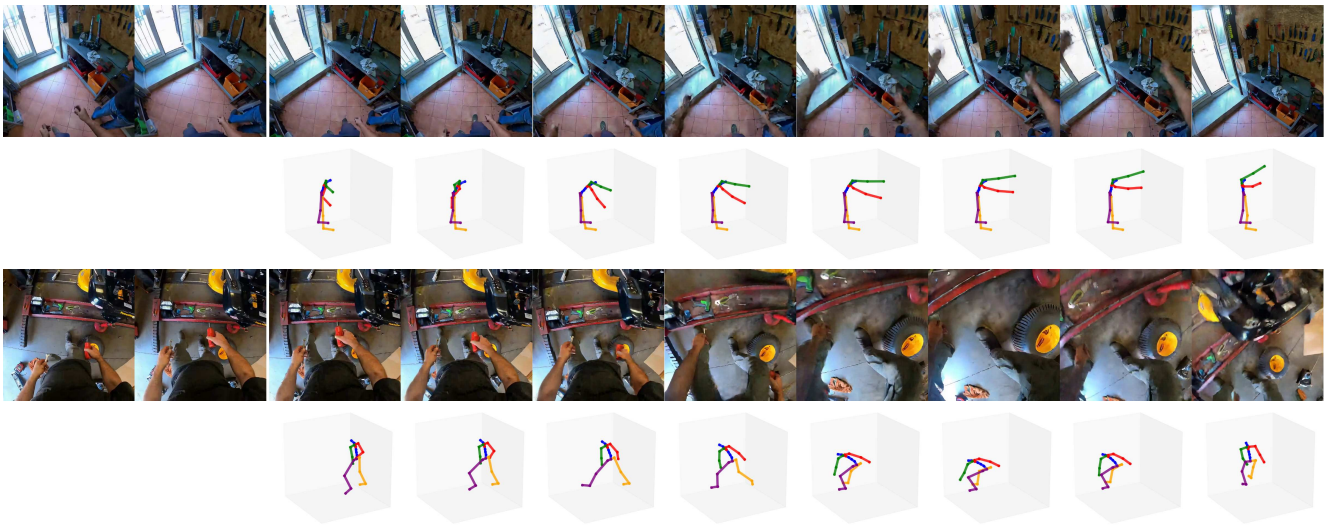

Figure 11. EgoControl shows control abilities on Ego4D, a dataset without any pose annotations.

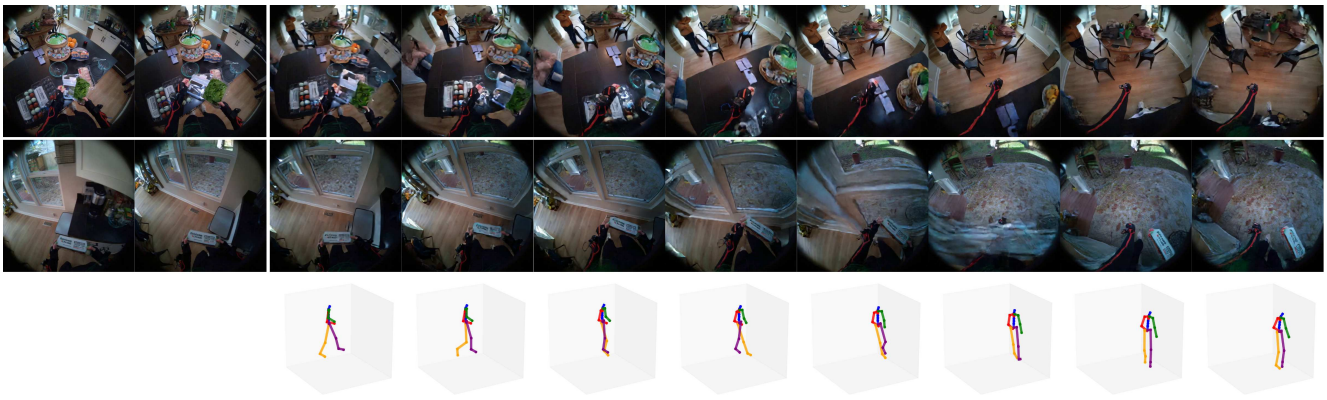

Figure 12. Failure cases when the provided sequence of poses is inconsistent with the environment.