Figure I: VLM failure cases: "**golfing**", "**eating something using both hands**", and "**cast fishing pole**''. Despite VLM hallucination causing some text labels to misalign with action semantics, our method remains robust. Additionally, some VLM predictions and GT exhibit different motion directions (*e.g.*, , depth ambiguity of picking up from the front vs. right front in the first example), which may be interesting to consider in future work.

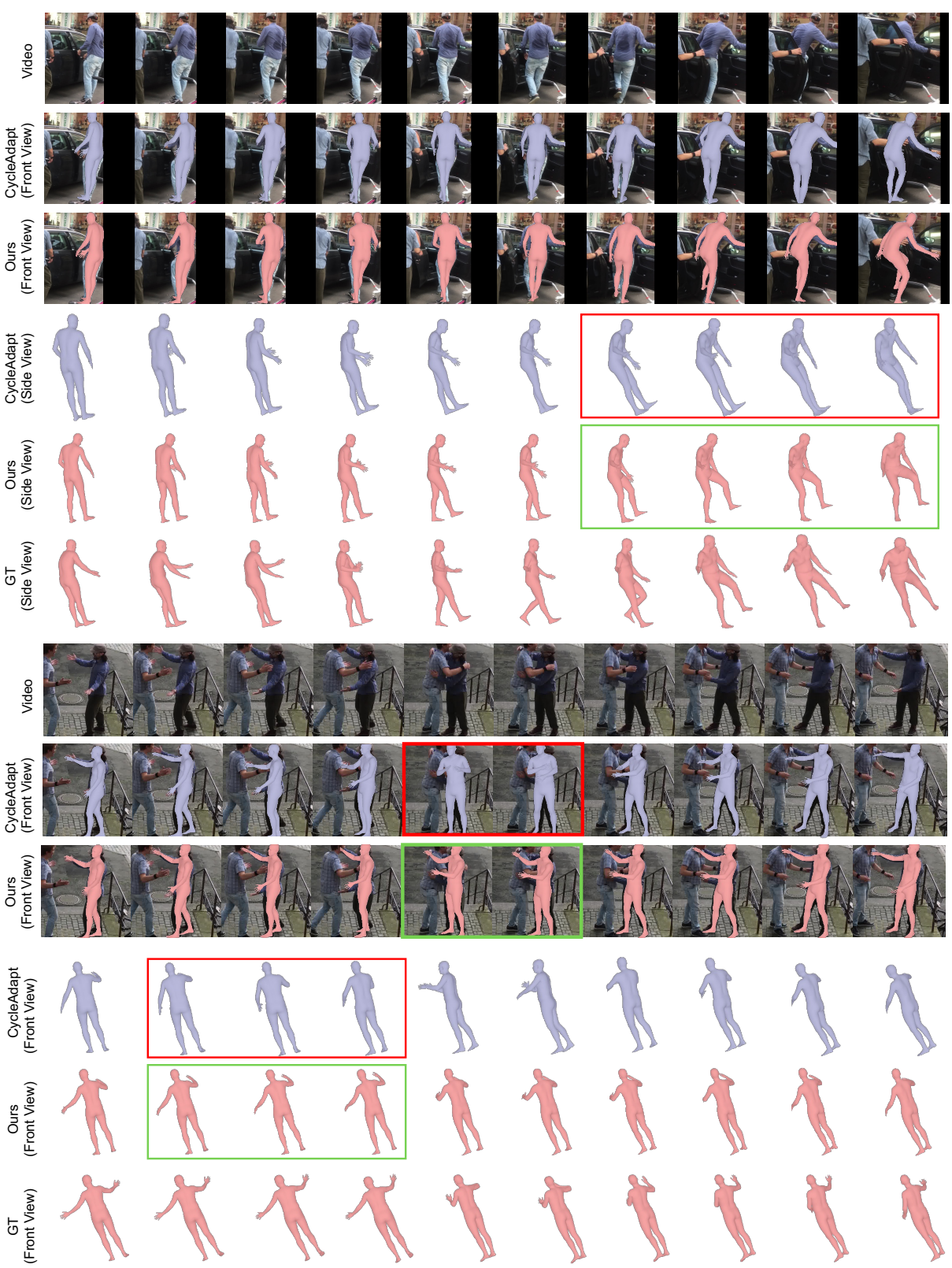

Figure II: Visualizations of challenging poses. We take sequences of "**open a door and sit**", "**hugging**" from 3DPW dataset (Von Marcard et al., 2018). Our method is more consistent with video semantics. For example, our prediction shows a reasonable stepping into the car. It is also robust in scenes such as hugging people, which has human-to-human interaction.

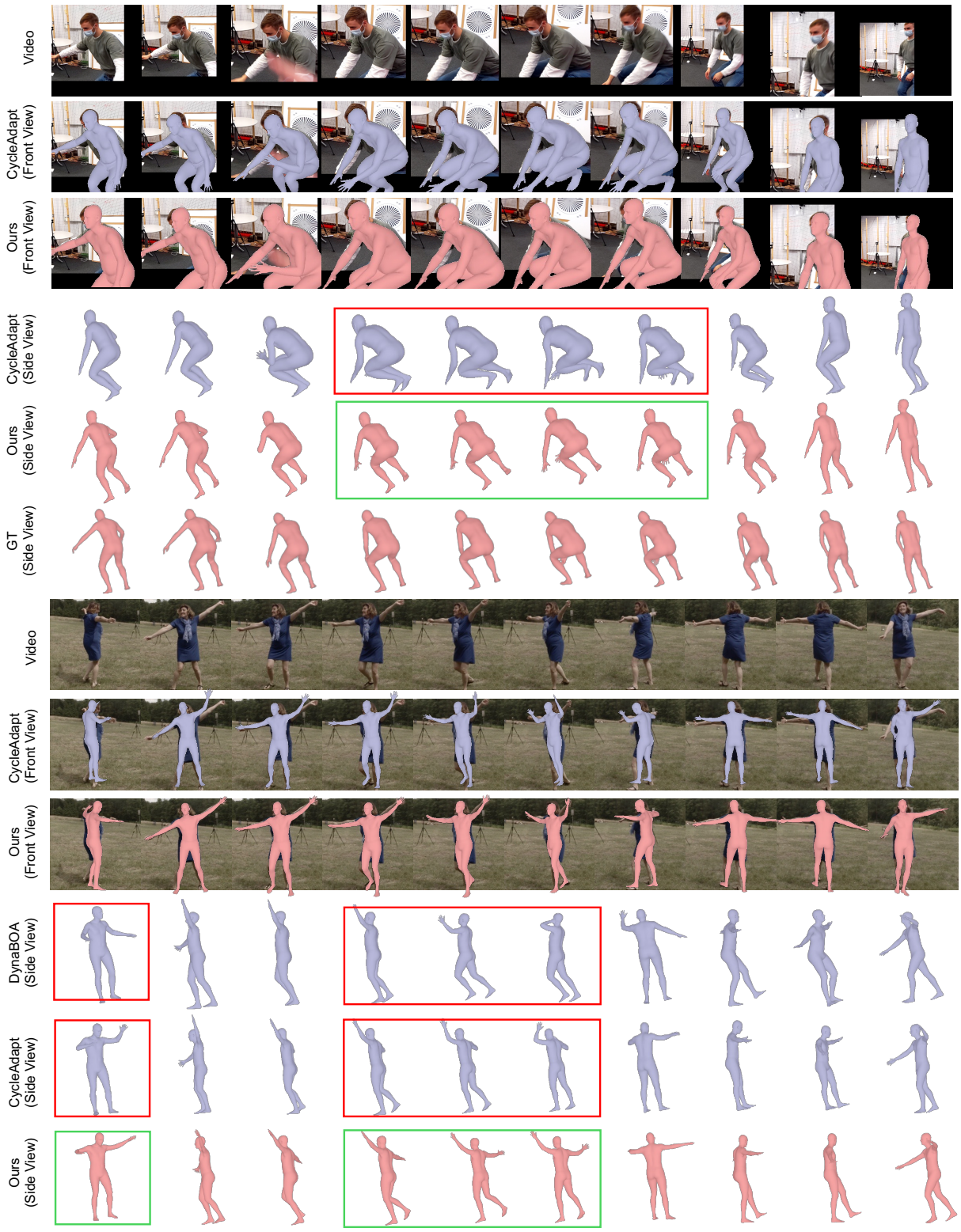

Figure III: Visualizations of challenging poses. We take sequences of "**squats series**" from EgoBody (Zhang et al., 2022) and "**spin around with right foot**" from 3DHP (Mehta et al., 2017). Our method is more consistent with video semantics. For example, CycleAdapt shows unnatural foot and knee positions during a squat, while ours is more realistic. Ours also can better capture the arm movements while spinning around.

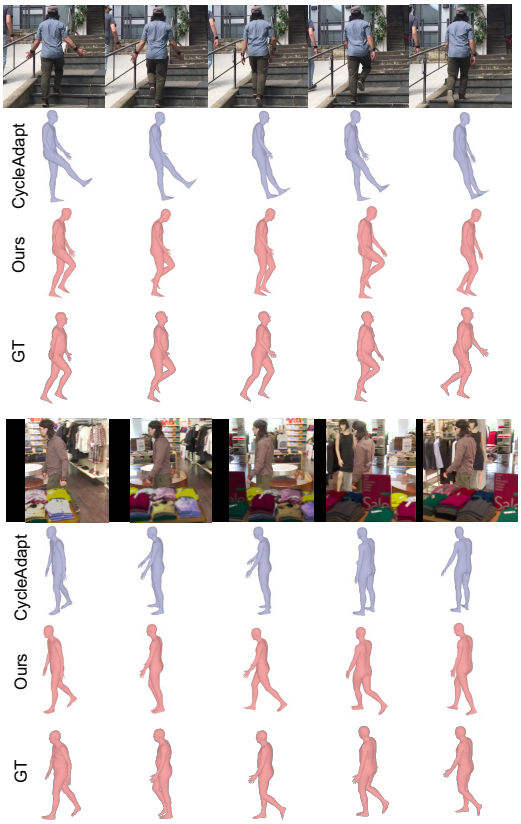

Figure IV: Compared to CycleAdapt method, our predictions show more aligned semantics (*i.e.*, , "**climbing-the-stairs**" and "**walking**") with the ground truth (GT).

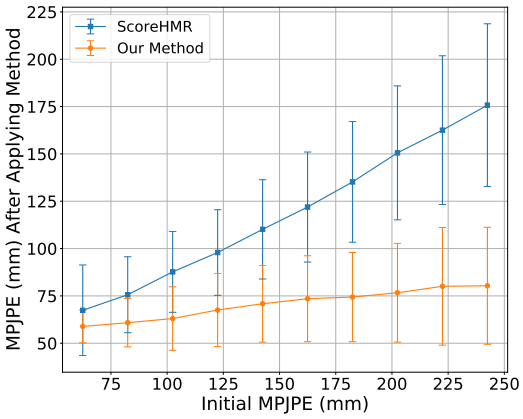

Figure V: Methods that optimize 3D pose outputs without model fine-tuning are heavily dependent on the quality of initial predictions. We demonstrate this limitation by comparing with ScoreHMR (Stathopoulos et al., 2024), which shows significant sensitivity to poor initial predictions. In contrast, ours maintains robust performance.

References

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., and Theobalt, C. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, pp. 506–516. IEEE, 2017.

Stathopoulos, A., Han, L., and Metaxas, D. Score-guided diffusion for 3d human recovery. In *CVPR*, pp. 906–915, 2024.

Von Marcard, T., Henschel, R., Black, M. J., Rosenhahn, B., and Pons-Moll, G. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018.

Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., and Tang, S. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, October 2022.