Diffusion-based 3D Hand Motion Recovery with Intuitive Physics

Supplementary Material

In this supplementary material (Appendix as referred to in the main paper), we provide additional evaluation of our proposed approach, including

- Appendix A: Evaluation on Model-specific Training
- Appendix B: Evaluation on Complex Bimanual Hand-object Manipulation Dataset
- Appendix C: Additional Qualitative Evaluation

A. Evaluation on Model-specific Training

Method	*Rec. Error*		*Phys. Plausibility*		
	MJE	P-MJE	ACCL	KIN	STA
HaMer [44]	18.9	4.4	7.95	22.49	1.08
VIBE [29]	17.0	6.4	-	-	-
TCMR [12]	16.0	6.3	-	-	-
Deformer [16]	13.6	5.2	-	-	-
BioPR [12]	12.9	-	-	-	-
Ours (A)	17.5	4.1	1.01	**0.00**	**0.00**
Ours (S)	**12.4**	**3.9**	**0.80**	**0.00**	**0.00**

Table 4. **Quantitative Comparison of Our Model-Specific Approach against SOTA Methods.** The evaluation is conducted on the DexYCB dataset. Results of other methods are obtained from their published papers. "**Ours** (A)" refers to our model-agnostic approach discussed in the main manuscript, while "**Ours** (S)" represents our model-specific variant trained to enhance HaMer. For all the values, the smaller, the better.

In the main paper, we focused on a challenging model-agnostic setting where access to specific frame-wise reconstruction models is not assumed. Here, we explore the potential of our approach through model-specific training, where we adapt our refinement model to enhance a specific frame-wise reconstruction method. Taking the leading approach HaMer as our baseline, we train our model on paired sequences $(\mathbf{x}_{1:T}, \mathbf{y}_{1:T})$ consisting of HaMer's frame-wise predictions and ground truth motion data. Table 4 shows the evaluation results on the DexYCB dataset. Our model-specific variant (Ours (S)) achieves substantial improvements over HaMer in both reconstruction accuracy (reducing MJE from 18.9 to 12.6 mm) and physical plausibility (reducing ACCL from 7.95 to 0.8 mm/frame^2). Moreover, our model-specific variant significantly outperforms recent video-based methods in 3D hand reconstruction accuracy. Specifically, compared to leading method BioPR, our approach achieves better accuracy with an MJE of 12.4 mm versus their 12.9 mm. These improvements further demonstrate the effectiveness of our physics-augmented diffusion-based refinement approach. While our model-agnostic approach provides flexibility across different reconstruction methods, the model-specific training can achieve superior performance when targeting a specific frame-wise reconstruction model.

B. Evaluation on Complex Bimanual Hand-object Manipulation Dataset

Method	P-MJE	ACCL
A:HaMer [44]	9.7	9.88
A+PoseBERT [4]	11.0	3.58
A+**Ours**	**8.9**	**0.81**

Table 5. **Quantitative Evaluation on Complex Bimanual Hand-object Manipulation Dataset TACO [38].** For all the values, the smaller, the better.

We here extend the evaluation to a recent dataset featuring more complex hand-object interaction sequences to further demonstrate the effectiveness of our approach. Specifically, we consider the TACO dataset, a large-scale benchmark for bimanual hand-object interactions. Captured through a multi-view setup, it encompasses a diverse range of tool-action-object compositions representative of daily human activities. Without any model fine-tuning or retraining, we integrate our method on top of HaMer and evaluate it on the S1 testing split of the TACO dataset. We present the evaluation results in Table 5. As illustrated, our approach consistently improves upon the leading frame-wise reconstruction HaMer (8.9 vs. 9.7 mm P-MJE, and 0.81 vs. 9.88 mm/frame^2 ACCL) and outperforms the leading motion refinement model PoseBERT (11.0 mm P-MJE, and 3.58 mm/frame^2 ACCL). These improvements in significant dynamic scenarios involving complex bimanual hand-object manipulation further validate the robustness of our physics-augmented diffusion-based approach.

C. Additional Qualitative Evaluation

Our qualitative evaluation in the main paper demonstrates improved accuracy and physical plausibility in hand motion recovery compared to SOTA approaches. In this section, we further highlight our advantages through evaluation on stable grasping sequences. We present a qualitative

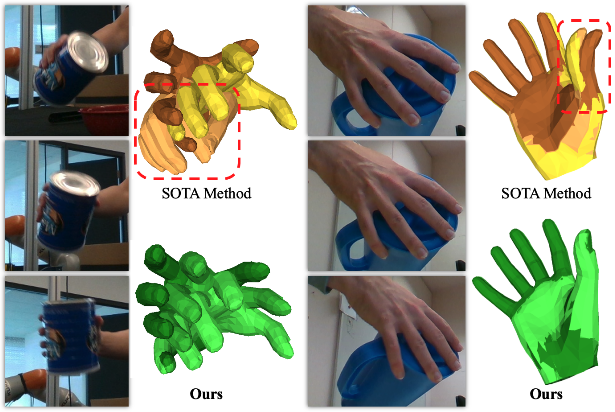

Figure 6. **Qualitative Evaluation on Stable Grasping Sequences.** The testing sequences are from DexYCB (left) and HO3Dv2 (right). To better illustrate the enhanced reconstruction of our model on grasping poses, root rotation is removed in the results on the right.

comparison with HaMer in Figure 6. Despite being trained on large-scale data with a high-capacity model, HaMer still produces degraded reconstructions even for simple, static hand poses when partially observed—such as when holding a bottle (highlighted by red rectangles). In contrast, our model achieves temporally consistent motion recovery by capturing an intuitive understanding of how the hand naturally interacts with objects. In particular, the example on the right shows that our method maintains stable hand poses over time during prolonged grasping.