ScoreHOI: Physically Plausible Reconstruction of Human-Object Interaction via Score-Guided Diffusion

Supplementary Material

In this appendix, we describe the DDIM sampling loop in our methods in pseudo-code in Section A. We also provide additional detailed implementations and qualitative comparisons in Section B and Section D. Furthermore, we append rendered 3D object models and the source code along with this appendix.

A. DDIM Refinement Loop

For a deeper understanding of the DDIM sampling loop, we illustrate a pseudo-code implementation in Algorithm A. The algorithm describe the DDIM_Loop in detail.

B. Detailed Implementations

We employ the PointNeXt [42] model, pre-trained on the ModelNet40 [55] dataset, as our affordance-aware regressor. ModelNet40 is a extensively utilized benchmark dataset for 3D shape classification and retrieval tasks, comprising 12,311 CAD models across 40 distinct object categories, including airplanes, chairs, tables, and cars. PointNeXt maintains the fundamental hierarchical architecture of PointNet++ [41] while integrating contemporary deep learning techniques to augment performance and efficiency. Through the incorporation of point cloud awareness, the model is capable of comprehending object geometry across a variety of human-object interaction scenarios. For the contact predictor and mesh regressor, we leverage the contact estimation transformer and the contact-based refinement methodology proposed by [37]. The contact predictor receives human and object feature tokens, along with estimated human and object mesh vertices, as inputs to generate contact masks through two symmetrical 4-layer transformer blocks. During our contact-driven iterative refinement process, we iteratively update the human and object meshes to enhance contact interactions, thereby refining the contact prediction results with each iteration. After N iterations of optimization, we obtain the $\boldsymbol{x}_0^N$ for final refinement. The mesh regressor, also constructed with two analogous 4-layer transformer branches, takes the updated human and object feature tokens and contact masks as input to further refine the mesh results.

C. Ablation Studies

About Diffusion Priors. We demonstrate the generation results starting from different noise levels without physical guidance, as shown in the Figure A. The IG-Adapter successfully constructs plausible human-object interaction patterns, confirming that the diffusion model has acquired a valid prior distribution.

Table A. Efficiency comparision with template-free methods.

	HDM	InterTrack	ScoreHOI	ScoreHOI-F
FPS↑	0.0047	0.0012	0.2895	**2.0080**

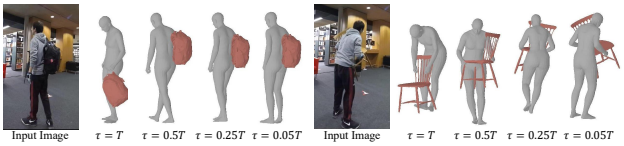

Figure A. Generation results starting from different noise levels.

Comparison with template-free methods. We compared ScoreHOI with template-free methods, including HDM [57] and InterTrack [61]. As shown in the Table A, ScoreHOI surpasses these methods in efficiency. Additionally, these methods output point clouds, which are less practical for applications like robotic data collection due to the additional time and uncertainty introduced when fitting SMPL parameters.

D. More Qualitative Results

To further substantiate the performance of our ScoreHOI, we present additional qualitative results of human-object interaction reconstruction in Figure B. Our methodology exhibits superior performance across diverse interaction patterns, such as sitting, carrying, grasping, and lifting. By incorporating physical constraints, our ScoreHOI achieves a higher degree of accuracy and physical fidelity in the reconstruction of human and object meshes.

E. Geometry Model Demos

The demo 3D geometry mesh results are available under the **demo** directory. Reviewers are able to assess the performance from any perspective utilizing software such as MeshLab or Blender.

Algorithm A Score-Guided DDIM refinement loop

Input: latent parameters $\boldsymbol{x}_0^n$ at step n, denoising model ϵ_ϕ, image features $\boldsymbol{c}_\mathrm{I}$, geometry features $\boldsymbol{c}_\mathrm{G}$, gradient step size ρ, noise level τ, DDIM step size Δt, estimated caontact masks $\{\mathbf{M}_i\}_{i\in\{\mathrm{h,o,f}\}}$
Output: latent parameters $\boldsymbol{x}_0^{n+1}$ for next sampling step $n+1$
$\boldsymbol{x}_\tau = \mathrm{DDIMInvert}(\boldsymbol{x}_0^n, \boldsymbol{c}_\mathrm{I}, \boldsymbol{c}_\mathrm{G})$ ▷ Run DDIM inversion until noise level τ
for $t = \tau$ to Δt with step size Δt **do**
    $\tilde{\epsilon} \leftarrow \epsilon_\phi(\boldsymbol{x}_t^n, t, \boldsymbol{c}_\mathrm{I}, \boldsymbol{c}_\mathrm{G})$ ▷ Predict noise
    Initialize computational graph for $\boldsymbol{x}_t^n$
    $\hat{\boldsymbol{x}^n}_0 \leftarrow \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x}_t^n - \sqrt{1-\alpha_t}\tilde{\epsilon})$ ▷ Predict one-step denoised result
    $L_\mathcal{P} \leftarrow \mathrm{PhysicalGuidance}(\hat{\boldsymbol{x}}_0^n, \{\mathbf{M}_i\}_{i\in\{\mathrm{h,o,f}\}})$ ▷ Compute physical guidance loss
    $\tilde{\epsilon}' \leftarrow \tilde{\epsilon} + \rho\sqrt{1-\alpha_t}\nabla_{\boldsymbol{x}_t^n} L_\mathcal{P}$ ▷ Compute modified noise after score-guidance
    $\hat{\boldsymbol{x}^n}_0' \leftarrow \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x}_t^n - \sqrt{1-\alpha_t}\tilde{\epsilon}')$ ▷ Predict one-step denoised result with modified noise
    $\boldsymbol{x}_{t-\Delta t}^n \leftarrow \sqrt{\alpha_{t-\Delta t}}\hat{\boldsymbol{x}}_0^{n\prime} + \sqrt{1-\alpha_{t-\Delta t}}\tilde{\epsilon}'$ ▷ DDIM sampling step
end for
$\boldsymbol{x}_0^{n+1} \leftarrow \hat{\boldsymbol{x}}_0^{n\prime}$ ▷ Update $\boldsymbol{x}_0^n$ for next generation
return $\boldsymbol{x}_0^{n+1}$

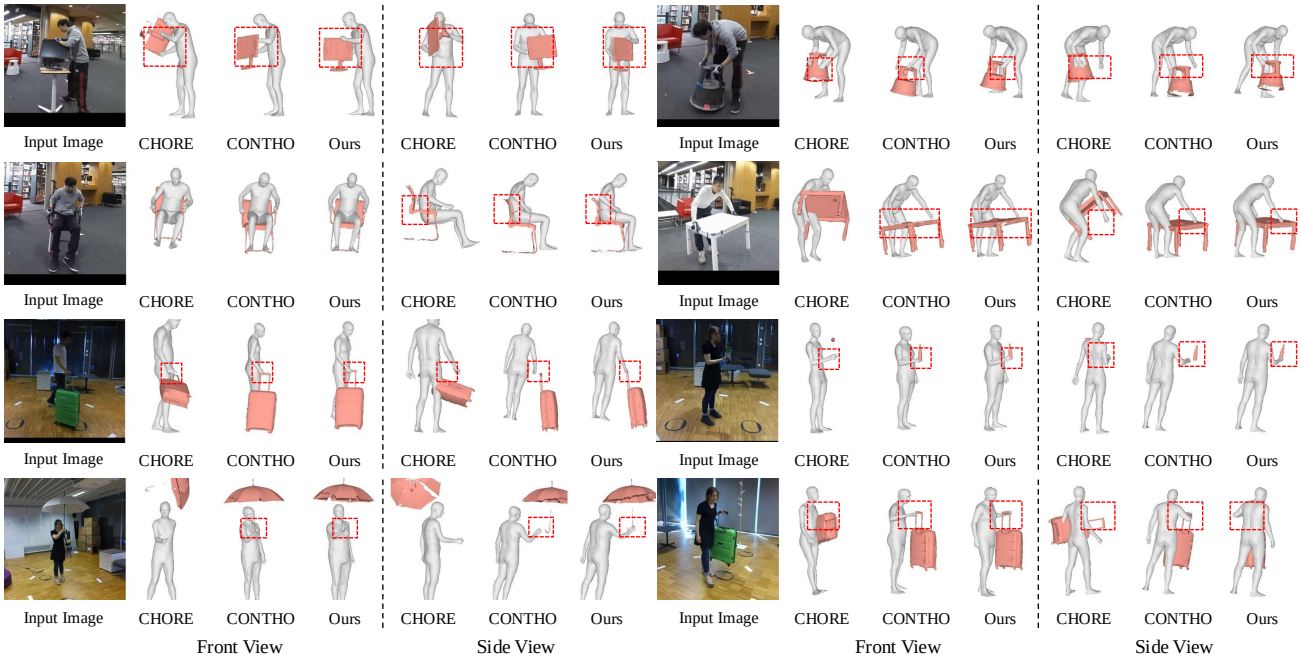

Figure B. **Extra Qualitative comparisons.** We highlight the contact interaction within each picture.