ReGenHOI: Unifying Reconstruction and Generation for 3D Human–Object Interaction Understanding - *Supplement Materials*

1. Supplementary Material

1.1. Diffusion Bridge Refinement

This section provides an extended mathematical formulation of the proposed gravity-field–based diffusion bridge for refining human–object interactions. The goal is to transform a coarse SMPL-X–based human configuration $\mathcal{H}_0$ into a physically valid configuration $\mathcal{H}^*$ while maintaining anatomical plausibility and contact consistency. Note that the current formulation assumes rigid objects, and thus it is not designed to handle deformable or articulated (e.g., hinge-based) objects.

Potential manifold and smoothness assumptions. Let $\mathcal{H}_t \in \mathbb{R}^{N_H \times 3}$ denote the SMPL-X human point cloud at diffusion time t. Following GravityDB [4], we construct a multi-scale gravitational potential $\varphi(\mathcal{H})$ using the predicted interaction region, where the object-contact manifold is modeled as a zero-potential surface. We assume that φ, the SMPL-X structural loss $\mathcal{L}_{\text{SMPL-X}}$, and the normal-consistency loss $\mathcal{L}_{\text{normal}}$ are continuously differentiable with bounded local curvature, ensuring stable integration of the diffusion dynamics.

SDE refinement flow. The refinement process is modeled by the following Itô SDE:

$$\begin{aligned} d\mathcal{H}_t = & -\alpha \nabla \varphi(\mathcal{H}_t)\, dt - \lambda_1 \nabla \mathcal{L}_{\text{SMPL-X}}(\mathcal{H}_t)\, dt \\ & - \lambda_2 \nabla \mathcal{L}_{\text{normal}}(\mathcal{H}_t)\, dt + g(\mathcal{H}_t)\, dW_t, \end{aligned} \tag{1}$$

where W_t is a standard Wiener process and $g(\cdot)$ is a bounded diffusion scale. The three deterministic drifts correspond to the three terms used in the main method: attraction toward the interaction manifold, SMPL-X structural regularization, and normal-consistency refinement.

Well-posedness. Under the above smoothness assumptions, the drift and diffusion components of Eq. (1) satisfy local Lipschitz continuity and linear growth, guaranteeing the existence and uniqueness of a strong solution. Thus, the stochastic refinement trajectory $\{\mathcal{H}_t\}$ is well-defined.

Energy descent property. Ignoring stochasticity, the drift of Eq. (1) corresponds to the gradient flow of the total energy:

$$\mathcal{E}(\mathcal{H}) = \alpha\varphi(\mathcal{H}) + \lambda_1 \mathcal{L}_{\text{SMPL-X}}(\mathcal{H}) + \lambda_2 \mathcal{L}_{\text{normal}}(\mathcal{H}). \tag{2}$$

Taking expectation over the SDE yields:

$$\frac{d}{dt}\mathbb{E}[\mathcal{E}(\mathcal{H}_t)] = -\mathbb{E}\big[\|\nabla \mathcal{E}(\mathcal{H}_t)\|^2\big] + \tfrac{1}{2}\mathrm{Tr}(gg^\top \nabla^2 \mathcal{E}), \tag{3}$$

indicating monotonic energy decrease for sufficiently small or annealed noise, consistent with the coarse-to-fine refinement behavior.

Variational interpretation. The diffusion bridge can be interpreted as constructing a stochastic transport path from the distribution of coarse human poses $\mathbb{P}_0$ to the distribution $\mathbb{P}^*$ of physically valid human–object interactions:

$$\min_{\mathbb{Q}} \ \mathrm{KL}(\mathbb{Q}\|\mathbb{W}) + \int_0^1 \mathbb{E}_{\mathbb{Q}}[\mathcal{E}(\mathcal{H}_t)]\, dt, \tag{4}$$

where $\mathbb{W}$ is the Wiener measure. This provides a principled interpretation of the refinement as an entropically regularized diffusion bridge.

Normal-consistency term. Let $\mathcal{C} \subset \mathcal{H} \times \mathcal{O}$ denote the set of human–object contact pairs inferred from the predicted contact map $\hat{C}$. For each pair $(i, j) \in \mathcal{C}$, we denote by $\mathbf{n}_i^h$ and $\mathbf{n}_j^o$ the unit outward normals of the human hand surface and the object surface, respectively. The normal-alignment loss used in our method is defined as

$$\mathcal{L}_{\text{normal}} = 1 - \frac{1}{|\mathcal{C}|} \sum_{(i,j)\in\mathcal{C}} \big(\mathbf{n}_i^h \cdot \mathbf{n}_j^o\big)_+, \tag{5}$$

where $(a)_+ = \max(a, 0)$ ensures that only positively aligned normals contribute.

For two unit normals $\mathbf{n}_i^h, \mathbf{n}_j^o \in \mathbb{S}^2$, the term

$$(\mathbf{n}_i^h \cdot \mathbf{n}_j^o)_+ \in [0, 1]$$

penalizes deviations from parallel orientation while ignoring antiparallel contributions, which is consistent with

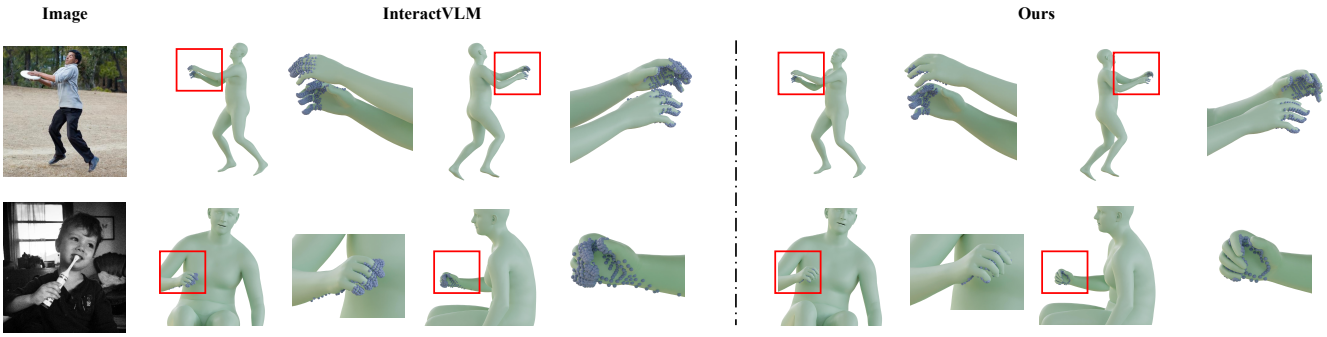

Figure 1. Qualitative comparison of interaction regions. Our method shows a clear advantage when the contact area is small.

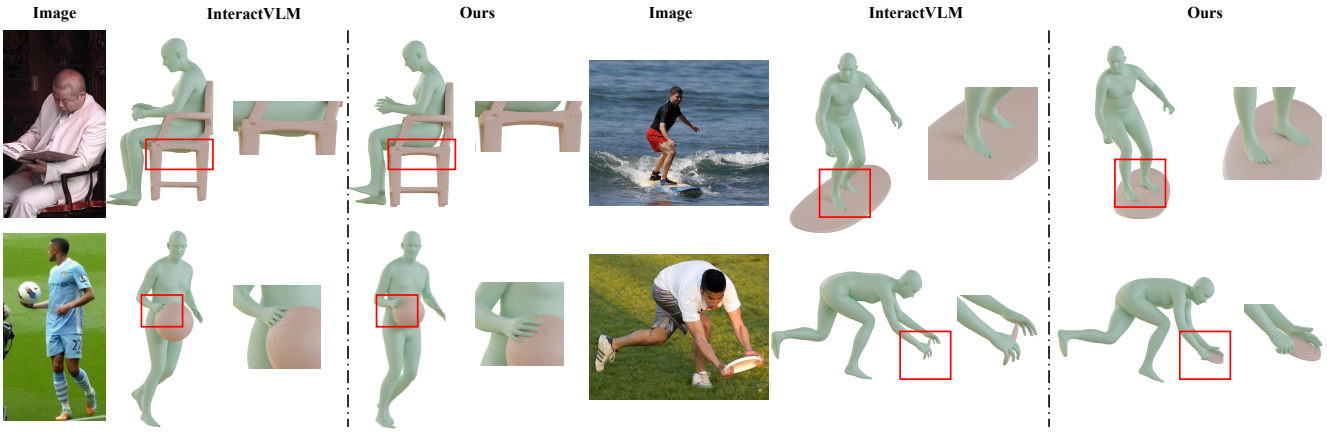

Figure 2. More qualitative human-object interaction reconstruction results.

physical hand–object contact where only inward-facing normals are meaningful.

Since the dot product is smooth and the hinge operator is piecewise linear, $\mathcal{L}_{\text{normal}}$ is subdifferentiable everywhere. Moreover, when contact pairs (i, j) are induced via closest-point projection

$$j = \Pi_{\mathcal{O}}(i),$$

and the object surface is C^2-smooth, the projection map $\Pi_{\mathcal{O}}$ is differentiable almost everywhere, guaranteeing a well-defined gradient $\nabla_{\mathcal{H}}\mathcal{L}_{\text{normal}}$.

Discretization and stopping criterion. Euler–Maruyama discretization gives:

$$\mathcal{H}_{t+1} = \mathcal{H}_t - \eta\nabla\mathcal{E}(\mathcal{H}_t) + g(\mathcal{H}_t)\sqrt{\eta}\,\varepsilon_t. \tag{6}$$

We stop refinement when:

$$\|\nabla\varphi(\mathcal{H}_t)\| < \epsilon, \quad \text{(gravity equilibrium)} \tag{7}$$

$$\|\nabla\mathcal{L}_{\text{normal}}(\mathcal{H}_t)\| < \epsilon_n, \quad \text{(normal consistency)} \tag{8}$$

$$t > T_{\max}. \quad \text{(time budget)} \tag{9}$$

The refined configuration $\mathcal{H}^*$ produced by the above dynamics enjoys several desirable geometric properties: it eliminates interpenetration through the combined effects of attraction and normal-consistency constraints, achieves normal alignment in regions predicted to be in contact, and converges to a locally optimal solution of the overall energy $\mathcal{E}$, while the term $\mathcal{L}_{\text{SMPL-X}}$ further ensures anatomical plausibility consistent with the SMPL-X prior.

Overall, the diffusion bridge defines a mathematically well-posed, energy-reducing stochastic flow that lifts coarse SMPL-X predictions onto a geometrically and physically consistent human–object interaction manifold.

1.2. Additional Experiments and Visual Results

In this section, we provide additional qualitative results that complement the experiments in the main paper. These visualizations further demonstrate the effectiveness, robustness, and generalization capability of our unified framework for 3D human–object interaction (HOI) reconstruction and generation.

1.2.1. Comparison of Contact Region Regression

Figure 1 compares our predicted contact regions with those obtained from InteractVLM [2]. For clearer comparison, we reposed the standard human model of InteractVLM to match the same pose as ours. Due to its reliance on lift-

Table 1. Aditional quantitative comparisons on the PICO dataset of Procrustes-aligned (PA) Chamfer Distance (CD) for human, object, and joint human–object reconstruction.

Method	$PA\text{-}CD_h$ $(cm, \downarrow)$	$PA\text{-}CD_o$ $(cm, \downarrow)$	$PA\text{-}CD_{h+o}$ $(cm, \downarrow)$
PHOSA [5]	12.38	13.21	12.79
CHORE [3]	10.71	12.64	11.52
InteractVLM [2]	6.38	13.91	9.02
PICO [1]	6.66	13.34	8.36
Ours	**5.42**	**12.68**	**7.62**

ing 2D contact cues into 3D, InteractVLM often struggles with small objects and fine-grained contact areas, leading to incomplete or spatially imprecise predictions. In contrast, our method directly infers 3D-aware contact regions conditioned on spatial priors and reasoning-based cues, producing more accurate and localized contact estimates, especially for small or thin interaction regions. It is worth noting that InteractVLM performs interaction reasoning in a canonical pose. For clearer visualization, we adjust its results to match the pose in the image.

1.2.2. Additional Reconstruction Results

Figure 2 presents a diverse set of reconstruction examples across various human poses, object categories, and interaction types. Our method consistently produces anatomically plausible human meshes, accurately aligned object poses, and stable contact configurations. Even in challenging cases involving strong occlusions or extreme viewpoints, our approach maintains geometric integrity and avoids artifacts such as severe interpenetration or unrealistic separations. These results highlight the reliability of our unified latent representation in capturing fine-grained spatial relationships. Additional comparison results of Procrustes-aligned (PA) Chamfer Distance (CD) for human, object, and joint human–object are provided in Table 1. PHOSA [5] and CHORE [3] are SMPL-based methods without explicit hand modeling, which limits their performance on fine-grained interactions.

1.2.3. Additional Generation Results

We include extended motion generation sequences in Figure 3. Given only textual prompts, our model synthesizes temporally coherent motions that exhibit smooth transitions, semantically aligned hand–object interactions, and stable contact over time. The generated motions demonstrate meaningful variations and adhere closely to the intent expressed in the language instructions. These examples further validate the effectiveness of our unified latent space in bridging static geometry understanding and dynamic motion synthesis.

Figure 3. More interaction generation results for various actions.

1.3. Limitations.

Despite its effectiveness, our approach has several limitations. First, the method relies on the accuracy of upstream predictions and requires reasonably reliable initialization. When the estimated human pose, object geometry, or contact priors are inaccurate, the subsequent optimization may converge to suboptimal interaction states. Second, challenging visual conditions such as severe occlusion or ambiguous language instructions can degrade the quality of the predicted contact regions, leading to unstable interaction reconstruction. Moreover, the current framework is unable to recover interaction scenarios where the human and the object are not initially in contact, as the refinement process assumes the existence of plausible contact priors. Finally, the current formulation assumes rigid objects and therefore cannot properly model interactions with deformable or articulated (e.g., hinged) objects. Addressing these limitations will be an important direction for future work.

References

[1] Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun S Lakshmipathy, Agniv Chatterjee, Michael J Black, and Dimitrios Tzionas. Pico: Reconstructing 3d people in contact with objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1783–1794, 2025. 3

[2] Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J Black, and Dimitrios Tzionas. Interactvlm: 3d interaction reasoning from 2d foundational models. In *Proceedings of the Computer*

Vision and Pattern Recognition Conference, pages 22605–22615, 2025. 2, 3

[3] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision*, pages 125–145. Springer, 2022. 3

[4] Miao Xu, Xiangyu Zhu, Xusheng Liang, Zidu Wang, Jinlin Wu, and Zhen Lei. Towards realistic hand-object interaction with gravity-field based diffusion bridge. *arXiv preprint arXiv:2509.03114*, 2025. 1

[5] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European conference on computer vision*, pages 34–51. Springer, 2020. 3