HOIDiNi: Human-Object Interaction through Diffusion Noise Optimization
Supplementary Material

Roey Ron Guy Tevet Haim Sawdayee Amit H. Bermano
Tel-Aviv University
roey1rg@gmail.com

1. Implementation Details

CPHOI is implemented as a Transformer decoder architecture with 8 layers and a hidden dimension of 512. Our pointwise object embedding network is a PointNet++ [9] fed with 512 randomly sampled vertices from the conditioned object. The model is trained with DDPM [4]; DDIM [12] is used at inference. More details per experiment can be found in Table 1. We use the Adam optimizer [5] for the noise optimization procedure. To reduce memory costs, we focus only on the MANO [10] subset of vertices of the SMPL-X human body and simplify it further to 1,100 vertices for each hand. The object meshes are simplified to 3,000 faces.

2. Contact Representation

Directly generating index pairs via diffusion is challenging. To address this, we adopt a more learnable representation. We define a fixed anchor set $\mathcal{A}$, consisting of a subset of MANO [11] hand vertices located on the palm (Figure 2). At each frame f, and for each anchor $a \in \mathcal{A}$, a binary variable b_a indicates whether the anchor is in contact. A corresponding position $p_a \in \mathbb{R}^3$ specifies the contact location on the object surface in its rest pose. This yields the contact representation:

$$F_{CP} = \left[p_1, \ldots, p_{|\mathcal{A}|}, b_1, \ldots, b_{|\mathcal{A}|}\right]$$

resulting in a per-frame contact feature of dimension $(3+1) \times |\mathcal{A}|$. The contact pairs figure in the main shows an example of contact pairs generated by HOIDiNi.

For the OMOMO [6] dataset, which lacks fingers' motion, we follow CHOIS and define the middle finger in each hand in the SMPL-X body model as the anchor, resulting in only two anchors for this benchmark.

3. Additional DNO Loses

3.1. Object-centric losses

Goal Loss. We encourage the object to reach a set of target poses at selected keyframes by penalizing both position and orientation errors. Concretely, we define

$$\mathcal{L}_{\text{Goal}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \Big(\|\hat{\mathbf{t}}_t - \mathbf{t}_t\|^2 + \mathcal{D}_{\text{rot}}\big(\hat{R}_t, R_t\big) \Big)$$

Where, $\mathcal{D}_{\text{rot}}$ measures the angular deviation between rotation matrices R_1 and R_2.

	Experiment	GRAB [2020]	OMOMO [2023]
Training parameters	# input prefix frames	15	1
	# generated frames	100	119
	diffusion steps (T)	8	14
	training steps	120K	50K
	batch size	64	64
DNO parameters	perturbation scale	10^{-6}	10^{-5}
	difference penalty	10^{-6}	10^{-5}
	λ_C	0.95	0.95
	λ_{Foot}	0.5	0.5
	λ_{Jitter}	10^{-5}	10^{-3}
	$\lambda_{P_{HO}}$	0.05	0.05
	$\lambda_{P_{HS}}$	0.2	–
	$\lambda_{P_{HH}}$	0.05	0.05
	$\lambda_{P_{OS}}$	1.2	0.05
	λ_{Goal}	0.5	0.9
	λ_{Static}	0.9	0.05
	$\lambda_{FeetFloorContact}$	–	0.5

Table 1. Hyper-parameters in use for each experiment.

Static Loss. To prevent unintended object motion, we penalize changes in position across frames where the object is not in contact with the human. We identify contiguous non-contact intervals $\{\mathcal{T}_s^{\text{nc}}\}_{s=1}^{S}$, and for each such segment s, we anchor the object pose to its value at the first frame, t_s^{start}. The loss is then computed as:

$$\mathcal{L}_{\text{Static}} = \frac{1}{\sum_{s=1}^{S} |\mathcal{T}_s^{\text{nc}}|} \sum_{s=1}^{S} \sum_{t \in \mathcal{T}_s^{\text{nc}}} \Big(\|\mathbf{t}_t - \mathbf{t}_{t_s^{\text{start}}}\|^2 + \mathcal{D}_{\text{rot}}(R_t, R_{t_s^{\text{start}}}) \Big),$$

where $\mathbf{t}_t$ and R_t are the object's translation and rotation at frame t, respectively, and $\mathcal{D}_{\text{rot}}$ measures the angular distance between two rotations. This encourages the object to remain static when not actively manipulated.

3.2. Human-centric losses

Feet-floor Contact Loss. For the OMOMO experiment, following CHOIS, we add a loss term that enforces accurate foot contact at the mesh level.

When reconstructing the human mesh with SMPL-X [8] using predicted root positions, joint rotations, and subject-specific body parameters, the generated feet may occasionally fail to touch the floor. To address this, we add a guidance term that encourages realistic feet-floor contact.

Let $\boldsymbol{J}_l$ and $\boldsymbol{J}_r$ denote the positions of the left and right toe joints. At each frame, the supporting foot is identified by comparing their z-coordinates. We further set a threshold height $h = 0.02$ meters, derived from analyzing foot heights in the ground truth motion. The guidance term is then defined as:

$$L_{\text{FeetFloorContact}} = |\min(\boldsymbol{J}_l^z, \boldsymbol{J}_r^z) - h|_2. \quad (1)$$

This measures the vertical deviation between the lower toe joint and the threshold height h.

4. Evaluation Metrics

For both benchmarks, GRAB and OMOMO, we evaluate our method along two dimensions: motion realism and interaction accuracy.

4.1. GRAB Evaluation

For the GRAB dataset [13] experiment, we follow IMoS [1], and compute realism metrics using embeddings from the final layer of an intent classifier. However, the classifier used in IMoS is limited to body joint positions and cannot capture fine-grained grasp dynamics. In contrast, our evaluation employs a more expressive classifier that takes as input body joints, hand joints, and object trajectories, allowing for a more comprehensive assessment of interaction quality.

FID. Fréchet inception distance measures the distance between the distributions of generated and ground-truth motions in a learned embedding space. Lower FID values indicate that the synthetic motion is closer in distribution to real motion data, capturing both realism and diversity.

Diversity. Evaluates how varied the generated motions are across different samples for the same input condition (e.g., prompt or object). It is computed as the average pairwise distance between multiple motion samples in the embedding space. A lower difference between ground truth and generated diversity scores suggests that the generated motions effectively capture the observed variability of human movement.

AVE. Measures the discrepancy between the variance of joint positions in generated motion and that of ground-truth motion. Specifically, it computes the average L^2 difference in per-joint positional variance across time. A lower AVE suggests that the model accurately captures the temporal dynamics and variability of natural human movement, avoiding overly rigid or overly jittery outputs.

IRA. Intent recognition accuracy quantifies how well the generated motions conveys the intended interaction or action. It is computed as the classification accuracy of the intent classifier on generated samples. High IRA indicates that the generated motions are semantically meaningful and align with their intended action labels, providing a measure of goal consistency and plausibility.

Multimodality. Assesses the model's capacity to produce distinct motions for the same conditioning intent. Unlike diversity, which measures sample variation globally, multimodality focuses on conditional variability by comparing multiple outputs conditioned on the same prompt. This metric is crucial for evaluating whether the model can express different plausible interaction strategies.

Penetration. Quantifies physical implausibility by measuring the extent to which the human mesh intersects with the object mesh. We compute the mean maximal penetration depth across frames where interpenetration occurs and the object is above table height. Lower penetration values indicate more physically valid interactions, particularly in grasping and manipulation scenarios where accurate surface contact is essential.

Floating. Captures the failure of hand-object interaction where the hand remains unnaturally far from the object surface. It is computed as the mean shortest distance between the body and object meshes, averaged across all motions (excluding frames with penetration and frames where the object is at table height). High floating values typically reflect unrealistic, disconnected grasping motion.

4.2. OMOMO Evaluation

For the OMOMO dataset [6] experiment, we follow the evaluation metrics defined by CHOIS [7]:

Condition Matching Metric. Those metrics calculate the Euclidean distance between the predicted and input object waypoints. It includes the start and end position errors (T_s, T_e), and waypoint errors (T_{xy}) measured in centimeters (cm).

Human Motion Quality Metric. Those metrics encompasses the foot sliding score (FS), foot height (H_{feet}), *Fréchet Inception Distance* (FID) and *R-precision* (R_{prec}). FS is the weighted average of accumulated translation in the xy plane, following prior work [3], measured in centimeters (cm). H_{feet} assesses the height of the feet, also in centimeters. R_{prec} and FID are computed following the text-to-motion task [2]. R_{prec} (top-3) measures whether the generated motion is consistent with the text. FID assesses the motion quality by computing the discrepancy between the distributions of ground truth and generated motions.

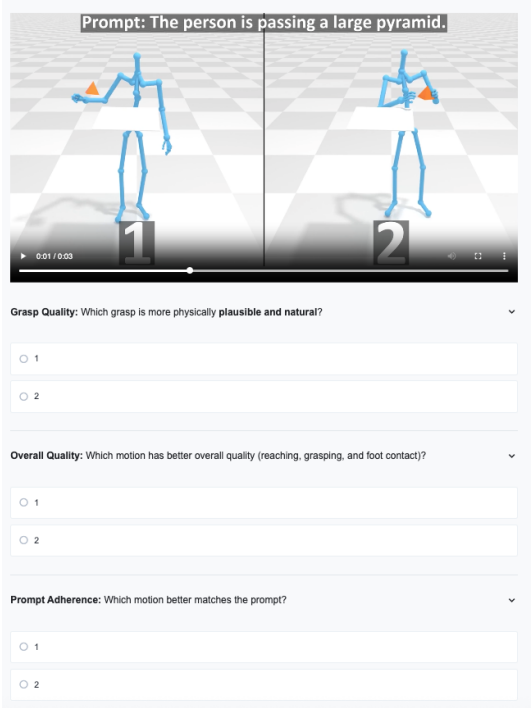

Figure 1. A screenshot from the user study.

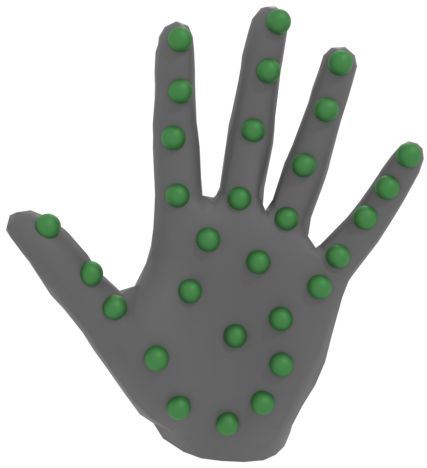

Figure 2. Palm anchor set $\mathcal{A}$ used by CPHOI for the GRAB [13] experiment.

Interaction Quality Metrics. Those metrics assess the accuracy of hand-object interactions, encompassing both contacts and penetrations. For contact accuracy, it employs precision (C_{prec}), recall (C_{rec}), and F1 score (C_{F_1}) metrics following prior work [6]. Additionally, it includes contact percentage $(C_{\%})$, determined by the proportion of frames where contact is detected. To compute the penetration score (P_{hand}), each vertex of the hand V_i is used to query the precomputed object's Signed Distance Field (SDF). This process yields a corresponding distance value d_i for each vertex. The penetration score is then derived by computing the average of the negative distance values (representing penetration), formalized as $\frac{1}{n}\sum_{i=1}^{n}|min(d_i, 0)|$, measured in centimeters (cm).

We note that CHOIS additionally measured the distance of the generated motion from the corresponding ground truth motion. Since HOIDiNi is a generative model, not aiming to reconstruct the ground truth, we find this metric irrelevant for our scope and omit it.

5. User Study

We conducted a user study for the GRAB dataset [13] with 24 participants, evaluating, in total, 12 side-by-side randomly selected samples of two models using the same inputs. We asked the user to evaluate the *grasp quality*, *prompt adherence*, and *overall quality*. As shown in the user study figure in the main paper, users preferred the results generated by our framework. A representative screenshot from the study interface is shown in Figure 1.

References

[1] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, pages 1–12. Wiley Online Library, 2023.

[2] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022.

[3] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. In *NeurIPS*, 2022.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[5] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[6] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

[7] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pages 54–72. Springer, 2024.

[8] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D

hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, Washington, DC, USA, 2019. IEEE Computer Society.

[9] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[10] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017.

[11] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017.

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, OpenReview.net, 2020. OpenReview.net.

[13] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020.