OMNIHOI: GENERALIZABLE HUMAN-OBJECT INTERACTION GENERATION THROUGH EXPLICIT INTERACTION MODELING

Anonymous authors
Paper under double-blind review

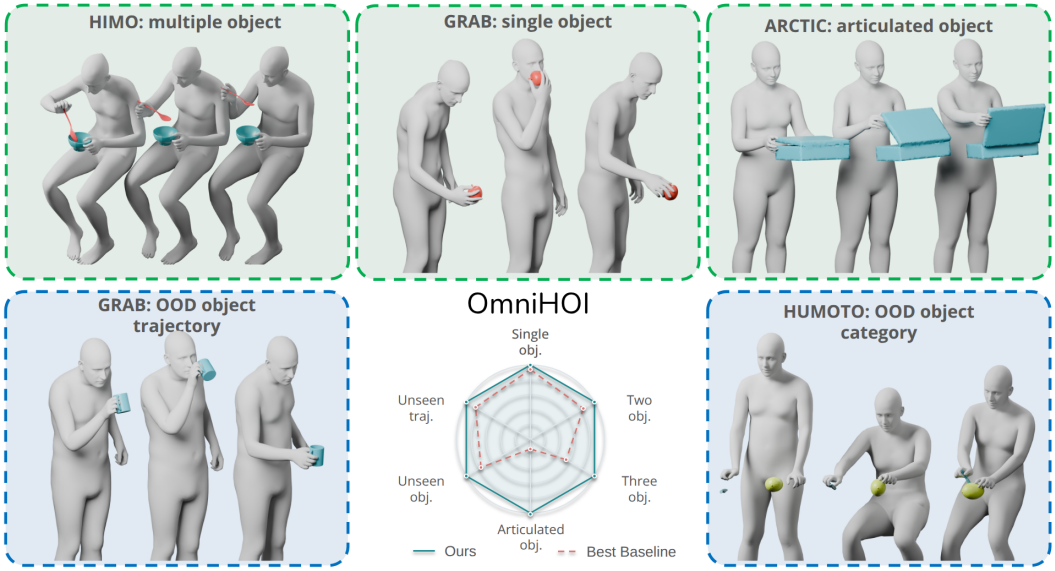


Figure 1: OmniHOI across HOI settings. Generated sequences for single-object (GRAB), articulated (ARCTIC), multi-object (HIMO), unseen-trajectory (GRAB), and unseen-category (HUMOTO) interaction. Center: normalized FID vs. the strongest per-setting baseline. OmniHOI outperforms prior methods across all settings.

ABSTRACT

Synthesizing human-object interaction (HOI) is valuable for applications in robot imitation learning and AR/VR animation. However, existing methods are ineffective in handling multiple-object interactions, or otherwise specialized to particular object configurations. Meanwhile, they fail to generalize to object categories or trajectories unseen during training. We present OmniHOI, a single HOI framework that jointly tackles single-object, articulated-object and multiple-object interactions. It devises a hierarchical generation pipeline, consecutively outputting object motion from text instructions, hand motion from object motion, and full-body motion from both object and hand motion. To reinforce interaction information, OmniHOI introduces interaction-aware self-conditioning and interaction-informed loss for diffusion models. OmniHOI also introduces multi-object test-time optimization to enhance hand contacts with multiple objects. OmniHOI sets a new state of the art on GRAB, ARCTIC and HIMO dataset, achieving stronger motion quality, text alignment, and hand–object interaction accuracy. Conditioned on unseen object categories or object trajectories, OmniHOI demonstrates stronger out-of-distribution generalization than prior work. Figure 1 summarizes our findings.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Learning to synthesize human-object interaction (HOI) has recently attracted growing interest (Ghosh et al., 2023). The task captures how humans plan object movements to accomplish the task while coordinating hand and body motion to achieve target object poses. It therefore supports downstream applications such as visual imitation learning in robotics (Luo et al., 2024; Xie et al., 2026) and interaction synthesis for AR/VR.

Despite recent progress, current HOI generation methods still fall short of downstream application needs, and we identify two challenges. The first lies in multiple-object interactions. Such tasks require deciding, at each moment, which object to target, which hand to engage, and what action to perform. General generative methods (Tevet et al., 2023; Ghosh et al., 2023) fail to handle this complexity; whereas, methods that build in inductive biases, such as architectures customized for two-object manipulation (Pratikaki et al., 2026), cannot generate interaction with more objects. The second challenge lies in generalization to out-of-distribution (OOD) object categories and trajectories. Current methods learn the joint distribution of object, hand, and body motion without reasoning about how the hands and body should move to manipulate individual objects. As a result, they tend to overfit to the training distribution and generalize poorly (Di et al., 2026).

HOI generation frameworks typically apply denoising diffusion models to generate object, hand and body motion (Cha et al., 2024), and use test-time optimization to refine output motion (Ye et al., 2026). To tackle multi-object interactions and enhance OOD generalization, our key idea is to reinforce the interaction information from the model's input, its training loss and the optimization procedure. First, we introduce interaction-aware self-conditioning for diffusion models. At each denoising step, we extract interaction features (Prokudin et al., 2019) from previous estimate of motion, and feed them back as conditioning input (Chen et al., 2023a), allowing the model to correct its own errors as generation proceeds. Second, we introduce interaction-informed regularization: in addition to the standard denoising objective, the model learns to reconstruct the interaction features of ground-truth motion. Third, we introduce test-time optimization to refine interaction with multiple objects. We learn to detect each interacting object and the engaging hand from the generated trajectories, then optimize the motion to bring the hand into contact with the predicted object part. These components together provide direct reasoning about object-hand-body interaction.

We introduce OmniHOI, a hierarchical HOI model that consecutively generates target object trajectories from text instructions, hand motion from object trajectories, and full-body movements from both generated results. The decomposed generation pipeline allows us to apply individual self-conditioned diffusion models, interaction-informed losses and test-time optimizations, specialized for object, hand or body motion, at the corresponding generation stage. With these components, OmniHOI effectively handles multiple-object interactions, setting a new state of the art in single-object, articulated-object and multiple-object HOI generation (Figure 1). On GRAB (Taheri et al., 2020) and ARCTIC (Fan et al., 2023), benchmarks for single-object and articulated-object interaction, OmniHOI outperforms the second best baselines by 6.4% and 86.4% in FID, and by 13.5 and 1.9 percentage points in retrieval precision, respectively (Table 2). On HIMO (Lv et al., 2024), a benchmark for two-object and three-object manipulation, OmniHOI outperforms the second best baselines by 17.5% and 52.3% in FID, and by 13.2 and 18.0 percentage points in retrieval precision, respectively (Table 3).

OmniHOI also demonstrates strong generalization to OOD object categories and trajectories. OmniHOI substantially outperforms baselines in both cases (Table 4). On OOD trajectories in GRAB and OOD object categories in HUMOTO (Lu et al., 2025), OmniHOI improves FID by 14.6% and 23.1%, and retrieval precision by 5.6 and 2.6 percentage points, respectively.

2 RELATED WORK

Controllable Human Motion Generation. Text-conditioned human motion generation has advanced through diffusion and discrete-token models (Tevet et al., 2023; Zhang et al., 2024; Chen et al., 2023b; Zhang et al., 2023; Guo et al., 2024). Generalist models further unify motion tasks and conditioning modalities within a single architecture (Jiang et al., 2023; Li et al., 2025; Guo et al., 2025; Jiang et al., 2026). For example, GENMO (Li et al., 2025) supports motion estimation and generation conditioned on video, text, or audio. Another line of work generates full-body motion from sparse signals: EgoAllo (Yi et al., 2025) uses head trajectories, while PriorMDM, GMD, and

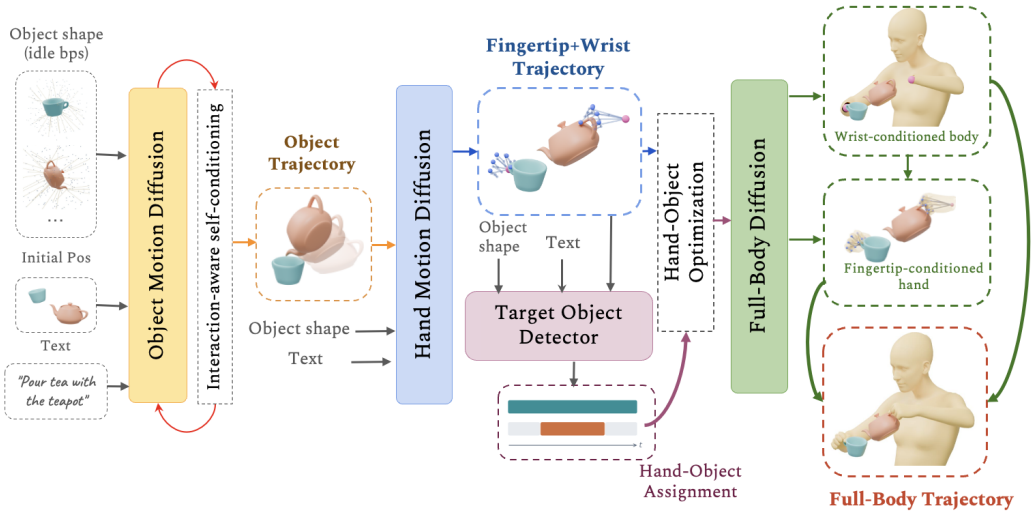


Figure 2: **Overview of OmniHOI framework.** From left to right, OmniHOI generates object trajectories, predicts fingertip and wrist motion with target-object detection and hand–object optimization, and finally synthesizes full-body motion conditioned on the generated object and hand trajectories. This hierarchical design enables explicit interaction modeling at different stages while using a unified pipeline across diverse HOI settings.

OmniControl (Shafir et al., 2024; Karunratanakul et al., 2023; Xie et al., 2024) constrain selected joints to follow target trajectories. These approaches demonstrate flexible motion control, but do not model how the body interacts with objects.

Human–Object Interaction. Most HOI generation methods primarily focus on a single object (Taheri et al., 2022; Ghosh et al., 2023; Xu et al., 2023). OMOMO (Li et al., 2023) generates full-body motion conditioned on object trajectories, while HOIDiNi (Ron et al., 2026) generates full-body motion with dexterous hand manipulation. To improve contact, several methods apply test-time optimization: CHOIS (Li et al., 2024) and CG-HOI (Diller & Dai, 2024) guide the sampling of jointly generated human and object motion, whereas HOIDiNi and CoDA (Pi et al., 2025) optimize the initial diffusion noise through the sampling process. These objectives assume a known target object, and noise optimization back-propagates through the entire sampling chain, substantially increasing inference cost.

Recent methods extend HOI generation to more complex object configurations. CoDA (Pi et al., 2025) addresses articulated objects, while HIMO, Dex2HOI, and PAMotion (Lv et al., 2024; Pratikaki et al., 2026; Di et al., 2026) address multi-object interaction. Dex2HOI uses object-specific interaction streams and supports one or two objects. These methods each target a particular object configuration. OmniHOI instead uses one architecture for single-object, articulated-object, and multi-object HOI generation, and generalizes to unseen object trajectories and categories without fine-tuning.

3 METHOD

We propose OmniHOI, a hierarchical HOI framework that consecutively generates object, hand and full-body trajectories from text instructions, generated object motion and object-hand motion. Similar to prior work, OmniHOI adopts denoising diffusion models as backbone generative models, and refine generated motion with test-time optimization. To make generated HOI more physically plausible, OmniHOI incorporates interaction-aware self-conditioning and interaction-informed regularizations for diffusion models. To handle the complexity of multiple-object interaction, OmniHOI learns to detect which object to target and which hand to engage from generated trajectories, and optimizes the motion of detected hand-object pairs to achieve plausible contacts.

We describe the problem formulation in Sec. 3.1 and the hierarchical generation framework in Sec. 3.2. We detail our proposed interaction-aware self-conditioning, interaction-informed regularizatons, and multi-object test-time optimization in Sec. 3.3, Sec. 3.4 and Sec. 3.5, respectively. Finally, we present an overview of OmniHOI in Figure 2.

3.1 Problem Formulation

The task of human-object interaction (HOI) generation aims to produce a sequence of object, hand and full-body motion, corresponding to conditioning text instructions and object instances. Each object instance is represented as a posed 3D mesh, or also the joint configuration for an articulated object. Next, we describe the representations of object, hand and full-body motion, respectively.

Object motion representation. We represent an object's motion as a sequence of rigid-body 3D poses and contact states paired at each moment. Each contact state is a binary value, indicating if the object is contacted by the human. In the next stage, such information helps the model to generate hand motion more consistently with object movements. Formally, we write object i's motion across time horizon T as: $\mathbf{o}^i = ([p_1^i, R_1^i, q_1^i], \cdots, [p_T^i, R_T^i, q_T^i])$, where p_1^i and R_1^i denote the translation and the rotation matrix of the 3D pose, and q_t^i is the contact state at time t. For articulated objects, we additionally append the joint configuration, such as the joint angle, to each pose-contact pair of the sequence. To encompass both single-object and multi-object interaction, we generally denote $O = \{\mathbf{o}^i\}_{i=1}^N$ as the set of N object motion, which the model is tasked to generate.

Hand motion representation. Because both fingers and the wrist must be controlled to manipulate objects, we represent a hand's motion with the movements of five fingers and the wrist. At time t, we describe finger i's movements with three terms: (1) its basis point set (BPS) representation b_t^i with respect to the target object (Prokudin et al., 2019), (2) its fingertip 3D position k_t^i, and (3) its contact state q_t^i. Meanwhile, we describe a wrist movements with its translation p_t^{wrist} and rotation matrix R_t^{wrist}. Combining them together, we denote a hand's motion over time horizon T as $\mathbf{h} = ([b_1^{1:5}, k_1^{1:5}, q_1^{1:5}, p_1^{\text{wrist}}, R_1^{\text{wrist}}], \cdots, [b_T^{1:5}, k_T^{1:5}, q_T^{1:5}, p_T^{\text{wrist}}, R_T^{\text{wrist}}])$. In this work, OmniHOI is tasked to predict both the left and right hand's motion $H = (\mathbf{h}^l, \mathbf{h}^r)$.

Full-body motion representation. Following prior work (Pi et al., 2025), we represent human body with SMPL-X model (Pavlakos et al., 2019) thereby describing full-body motion with a sequence of SMPL-X parameters. Across time horizon T, we denote full-body motion as $B = ([\Theta_1, p_1^{\text{root}}, R_1^{\text{root}}], \cdots, [\Theta_T, p_T^{\text{root}}, R_T^{\text{root}}])$, where Θ_t is the set of SMPL-X shape and pose parameters, p_T^{root} is the body root's translation and R_T^{root} is the body root's rotation matrix at time t.

Among these output representations, only the SMPL-X parameters, the body's root translation, the body root's rotation matrix, each object's translation and each object's rotation matrix are used to determine the final object-hand-body motion. Others are intermediate representations either for enriching the diffusion model's conditioning inputs or for guiding the test-time optimization. We detail the usage of every output motion representation in the Appendix A.1.

3.2 Hierarchical HOI Generation Framework

Conditioned on a text instruction l and a set of posed 3D object meshes V, HOI generation methods typically model the joint distribution of object, hand and full-body motion: $P(O, H, B \mid l, V)$. However, it is difficult to apply regularizations or optimizations on individual generated motion based on this formulation. Following Pi et al. (2025), we instead hierarchically model the distribution of object, hand and full-body motion, decomposing the generation pipeline into three stages:

$$P(O, H, B \mid l, V) = \underbrace{P(O \mid l, V)}_{\substack{\text{object-motion} \\ \text{generation}}} \underbrace{P(H \mid O, l, V)}_{\substack{\text{hand-motion} \\ \text{generation}}} \underbrace{P(B \mid H, O, l, V)}_{\substack{\text{body-motion} \\ \text{generation}}}. \tag{1}$$

OmniHOI first generates target object motion from conditioning text instructions and object instances, then hand motion from generated object motion, and finally full-body motion from generated object and hand motion. At each stage, a denoising diffusion model is trained to generate the motion. More details of the model architecture and training parameters are provided in the Appendix A.

3.3 INTERACTION-AWARE SELF-CONDITIONING

Self-conditioning (Chen et al., 2023a) offers a feedback mechanism that allows a diffusion model to correct its own generation error by conditioning on its output estimate, generated from the previous denoising step. We are inspired to use self-conditioning to enhance object-motion generation, that is highly dynamic and challenging to model. Rather than feeding back motion estimate, we propose to feed interaction features, extracted from motion estimate, back as conditioning input. Because such features capture object-wise interactions more directly than the motion estimate itself, our interaction-aware self-conditioning enables the diffusion model to correct motion errors more effectively.

Self-conditioning for object-motion generation. At each denoising step, the diffusion model estimates clean object motion based on the noisy motion input and the predicted noise (Ho et al., 2020). Let $\hat{O} = \{\hat{\mathbf{o}}^i\}_{i=1}^N$ denote the motion estimate of all N objects, with $\hat{\mathbf{o}}^i$ as object i's motion estimate. Our goal is to featurize interaction among objects according to the motion estimate.

To capture this, we calculate the relative geometry between any pair of objects. We first define function f^{pose} that returns the relative pose of a pair of rigid-body poses:

$$f^{\text{pose}}(R_a, p_a, R_b, p_b) = [R_a^\top (p_b - p_a),\ R_a^\top R_b,\ |p_b - p_a|_2], \tag{2}$$

where R_a and p_a are the rotation matrix and the translation of object a's pose. The three terms in Eqn. 2 represent object a's relative position, orientation and distance to object b, respectively. Here, we use the function to calculate object j's relative pose to object i at time t: $f^{\text{pose}}(\hat{R}_t^i, \hat{p}_t^i, \hat{R}_t^j, \hat{p}_t^j)$, where $\hat{R}_t^i$ and $\hat{p}_t^i$ are the rotation matrix and translation vector of object i's motion estimate at time t. However, relative poses are coarse-grained geometric descriptions, ignoring each object's semantics such as shape. To capture how object parts engage in interaction, we additionally extract fine-grained geometric descriptions with basis points associated to each object. We define function f^{point} that captures geometric structures among two sets of points U, V:

$$f^{\text{point}}(U, V) = [\text{NN}(u_1, V) - u_1, \cdots, \text{NN}(u_K, V) - u_K], \tag{3}$$

where u_k and K denote the k-th point and the number of points in U, and function NN returns the nearest neighbor of point u_i from V. Let $\hat{\mathbf{v}}_t^j$ denote the posed 3D mesh of object j according to its motion estimate $\hat{\mathbf{o}}^j$, and $\hat{\mathbf{v}}_t^{j\to i}$ be the same mesh described in object i's reference frame at time t. We derive object j's fine-grained geometric features relative to object i as: $f^{\text{point}}(\mathbf{u}^i, \hat{\mathbf{v}}_t^{j\to i})$, with basis points $\mathbf{u}^i$ associated to object i. We combine these two features for each object pair across time, and flatten which into a vector. The concatenated vectors of every object pair are fed back as the diffusion model's conditioning input via cross attentions. Details are provided in the Appendix A.1.

3.4 INTERACTION-INFORMED REGULARIZATIONS

In addition to the standard denoising objective, training diffusion models to predict high-quality interaction can further enhance motion generation. We propose to supervise diffusion models for reconstructing the interaction features of ground-truth motion. In this work, the interaction-informed loss is introduced at the stage of hand-motion generation.

Interaction-informed hand-motion loss. We extract hand-object interaction feature from the object motion generated at the previous stage, and the ground-truth / estimated hand motion at this stage. Following our proposed interaction-aware self-conditioning, we form the interaction features with both the relative pose, calculated by function f^{pose}, and the fine-grained geometric features, calculated by function f^{point}, for each hand-object pair.

Let $\hat{H} = (\hat{\mathbf{h}}^l, \hat{\mathbf{h}}^r)$ be the left- and right-hand motion estimate, and $O = \{\mathbf{o}^j\}_{j=1}^N$ be the generated object motion with N objects. First, we calculate object j's relative pose to hand i at time t: $f^{\text{pose}}(\hat{R}_t^{i,\text{wrist}}, \hat{p}_t^{i,\text{wrist}}, R_t^j, p_t^j)$, where $\hat{R}_t^{i,\text{wrist}}$ and $\hat{p}_t^{i,\text{wrist}}$ are the wrist's rotation matrix and translation vector. Second, we define the five fingertips and the wrist's root position as the set of basis points $\mathbf{u}^i$ associated to hand i, and derive object j's fine-grained geometric features relative to hand i as: $f^{\text{point}}(\mathbf{u}^i, \hat{\mathbf{v}}_t^{j\to i})$. Here, $\hat{\mathbf{v}}_t^{j\to i}$ is the posed 3D mesh of object j, expressed in hand i's reference frame using estimated hand motion at time t. Likewise, we obtain both terms using the ground-truth hand motion H: $f^{\text{pose}}(R_t^{i,\text{wrist}}, p_t^{i,\text{wrist}}, R_t^j, p_t^j)$ and $f^{\text{point}}(\mathbf{u}^i, \mathbf{v}_t^{j\to i})$. Together, they

yield an interaction-informed loss for the diffusion model:

$$\mathcal{L}_{\text{interact}} = \sum_{i\in\{l,r\}}\sum_{j=1}^{N}\sum_{t=1}^{T} \|f^{\text{point}}(\mathbf{u}^i, \hat{\mathbf{v}}_t^{j\to i}) - f^{\text{point}}(\mathbf{u}^i, \mathbf{v}_t^{j\to i})\| \tag{4}$$
$$+ \sum_{i\in\{l,r\}}\sum_{j=1}^{N}\sum_{t=1}^{T} \|f^{\text{pose}}(\hat{R}_t^{i,\text{wrist}}, \hat{p}_t^{i,\text{wrist}}, R_t^j, p_t^j) - f^{\text{pose}}(R_t^{i,\text{wrist}}, p_t^{i,\text{wrist}}, R_t^j, p_t^j)\|.$$

3.5 Multi-object Test-time Optimization

Because diffusion models are simulation-free, they inevitably may generate physically implausible artifacts, such as floating, sliding or penetration (Liu & Yi, 2024). Test-time optimization is therefore commonly applied to remove such artifacts from generated motion.

Hand-object motion optimization. Following Pi et al. (2025), we optimize generated hand motion with two objectives. The first term brings the hand into contact with the predicted object part. Given generated hand motion H, it is written as:

$$\mathcal{L}_{\text{contact}}(H) = \mathbb{E}_{m,t,i,j}\big[\,\Big|\|k_t^{m,i} - u_j\|_2 - b_t^{m,i}[j]\Big|\,\big], \tag{5}$$

where u_j denotes the j-th basis point of the target object, and $k_t^{m,i}$ and $b_t^{m,i}[j]$ are hand m's i-th fingertip predicted 3D position and its predicted distance to u_j at time t. However, this loss term fails in multiple-object interaction: the selection of the target object in Eqn. 5 becomes ambiguous when multiple objects co-occur in the scene. To address this, we learn a lightweight network to detect the target object for each hand at every time step. More details of the model architecture and training parameters are provided in the Appendix A.

The second term regularizes the optimized hand motion to prevent the optimization from introducing unrealistic hand configurations. Specifically, it penalizes large changes from the generated fingertip positions, preserves the pairwise distances between fingertips and wrist, and encourages smooth fingertip trajectories over time. The final hand motion is obtained by

$$H^* = \underset{H}{\text{argmin}}\mathcal{L}_{\text{contact}}(H) + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}(H). \tag{6}$$

Detailed formulations and optimization settings are provided in the Appendix B.2.

Full-body motion optimization. Following Yi et al. (2025), we further refine the generated full-body motion to reduce foot skating. Optimization details are provided in the Appendix B.3

4 Experiments

Our experiments test whether a single framework can cover the full range of HOI settings. We ask two questions: (Q1) Can a unified architecture, without modification, handle single-object, articulated-object, and multi-object interactions? (Q2) Does the model generalize to OOD object trajectories and object categories? We also ablate each design choice and report inference cost.

4.1 Experimental Setup

Datasets. We evaluate on three benchmarks spanning a spectrum of HOI: GRAB (Taheri et al., 2020) for single rigid objects, ARCTIC (Fan et al., 2023) for articulated objects, and HIMO (Lv et al., 2024) for interaction with two or three objects. We train one model per benchmark with the same architecture; on HIMO, a single model is trained jointly on 2-object and 3-object sequences. For generalization, we use a held-out test set of GRAB with object trajectories unseen during training, and HUMOTO(Lu et al., 2025), a held-out benchmark with novel object categories.

Baselines. As summarized in Tab. 1, each prior method is designed for, and evaluated on, a subset of these interaction settings. When a baseline does not natively support a setting, we adapt it and retrain it on the same data; these baselines are marked with †. PAMotion (Di et al., 2026) is omitted because its official code and pretrained models are not publicly available at the time of submission.

Metrics. We report FID, R-Precision (R@3), Diversity (Div.), and Foot Skating following standard motion-generation evaluation protocols. For interaction accuracy, we report HOI-ADD (Ye et al., 2026), which aligns the generated and ground-truth sequences by the hand trajectory and measures the resulting object alignment error. We further introduce OO-ADD for multi-object interaction, which analogously aligns by one object and measures the alignment error of the remaining object(s). Both metrics summarize the percentage of object points falling within distance thresholds up to 15 cm. We additionally report contact ratio and penetration depth as complementary physical-plausibility metrics. Detailed definitions are provided in Appendix C.

Table 1: Interaction settings demonstrated in each method's original paper.

Method	Single	Artic.	2-obj	3-obj
MDM	✗	✗	✗	✗
IMoS	✓	✗	✗	✗
OMOMO	✓	✗	✗	✗
HOIDiNi	✓	✗	✗	✗
CoDA	✓	✓	✗	✗
HIMO-Gen	✗	✗	✓	✓
Dex2HOI	✓	✗	✓	✗
OmniHOI (ours)	✓	✓	✓	✓

4.2 State of the Art Across Diverse HOI Settings (RQ1)

A single architecture achieves state-of-the-art performance in every interaction setting we evaluate. On single-object (GRAB), articulated-object (ARCTIC), and multi-object (HIMO) interaction, OmniHOI achieves the best FID, R@3, and HOI-ADD (Tabs. 2 and 3): it generates the most realistic motion, follows the text instruction most closely, and places the hands most accurately relative to the objects. Baselines, in contrast, are competitive mainly in the setting they were designed for. The strongest competitor changes from benchmark to benchmark, and methods specialized for articulated or multi-object interaction, such as CoDA and Dex2HOI, fall well behind OmniHOI on single-object interaction. OmniHOI needs no such specialization.

Table 2: **Comparison on the ARCTIC and GRAB datasets.**

Method	FID ↓	R@3 ↑	Div. →	Foot Skat. (m/s) ↓	HOI-ADD ↑	Contact ↑	Pen. (mm) ↓
ARCTIC							
Real	-	35.34%	25.629	0.002	-	7.14%	2.87
IMoS†	91.25	14.29%	23.331	0.401	6.80%	0.62%	**0.79**
MDM†	82.06	17.67%	23.891	0.018	16.39%	5.04%	6.31
OMOMO†	26.25	31.20%	25.090	0.016	24.26%	5.92%	7.02
HOIDiNi †	100.11	22.56%	22.69	0.028	24.23%	7.11%	6.57
CoDA	16.63	32.33%	**25.633**	0.003	23.17%	**10.15%**	6.81
OmniHOI (ours)	**2.26**	**34.21%**	25.522	**0.003**	**28.95%**	9.52%	8.44
GRAB							
Real	-	45.49%	23.719	0.009	-	5.55%	4.76
IMoS	354.15	4.83%	16.216	0.272	2.06%	0.00%	**0.00**
MDM†	169.66	15.52%	20.878	0.151	13.53%	0.72%	2.86
OMOMO	143.93	26.90%	22.247	0.139	11.85%	0.75%	3.58
CoDA	136.27	34.14%	22.474	0.048	9.23%	4.44%	6.33
HOIDiNi	96.80	37.24%	23.167	0.063	23.19%	**6.95%**	9.49
Dex2HOI	148.57	42.62%	22.288	0.070	23.98%	5.81%	4.88
OmniHOI (ours)	**90.62**	**56.15%**	**23.616**	**0.028**	**34.45%**	6.11%	7.19

4.3 Generalization to OOD Trajectories and Object Categories (RQ2)

We evaluate OOD generalization under two setups: conditioning on seen object categories with OOD trajectories, evaluated on a held-out GRAB test set, and conditioning on text instructions for manipulating OOD object categories, evaluated on HUMOTO. OmniHOI outperforms all baselines in both.

OOD trajectories. We condition generation on object trajectories that are not seen during training (Tab. 4, top). Because object motion is an explicit intermediate in our hierarchy, a given trajectory directly replaces the output of the object stage, without retraining or architectural change. OmniHOI

Table 3: **Comparison on the HIMO dataset with two and three objects.**

Method	FID ↓	R@3 ↑	Div. →	Foot Skat. (m/s) ↓	HOI-ADD ↑	OO-ADD ↑	Contact ↑	Pen. (mm) ↓
HIMO 2o								
Real	-	80.00%	22.378	0.010	-	-	4.16%	5.60
IMoS†	105.80	26.45%	20.469	0.305	16.77%	19.80%	1.50%	3.36
MDM†	69.33	42.15%	21.208	0.036	15.25%	11.62%	2.28%	5.23
HIMO-Gen	109.31	40.50%	19.801	0.059	15.77%	22.53%	4.22%	7.62
Dex2HOI	46.38	54.96%	21.960	0.023	21.40%	15.26%	4.73%	6.87
OmniHOI (ours)	**38.25**	**68.18%**	**22.095**	**0.013**	**22.89%**	**27.10%**	**4.84%**	**4.22**
HIMO 3o								
Real	-	79.00%	22.147	0.009	-	-	4.13%	10.84
IMoS†	258.02	37.00%	17.187	0.283	12.51%	**29.38%**	2.24%	11.82
MDM†	167.84	17.00%	20.858	0.017	7.51%	2.79%	2.99%	16.72
HIMO-Gen	219.47	30.00%	18.872	0.101	4.20%	11.06%	2.04%	13.05
OmniHOI (ours)	**80.06**	**55.00%**	**21.397**	**0.010**	**17.34%**	24.29%	**4.36%**	**7.86**

outperforms OMOMO, which is designed specifically to generate human motion from object motion, in both motion quality and interaction accuracy.

OOD object categories. We train all methods on HIMO and evaluate them on HUMOTO without fine-tuning (Tab. 4, bottom). Although this cross-dataset setting is challenging for every method, OmniHOI achieves the best FID, R@3, Foot Skating, and HOI-ADD. Because the human generator interacts with objects only through the hand–object interface produced by the earlier stages, it does not need to have seen an object category to realize interactions with it.

Table 4: **OOD generalization results.** We evaluate generalization to unseen object trajectories on GRAB and unseen object categories on HUMOTO.

Method	FID ↓	R@3 ↑	Div. →	Foot Skat. (m/s) ↓	HOI-ADD ↑	Contact ↑	Pen. (mm) ↓
OOD Trajectory Generalization on GRAB							
Real	-	45.49%	23.719	0.009	-	5.55%	4.76
MDM†	143.23	30.58%	22.778	0.171	6.43%	**0.18%**	0.51
OMOMO	86.10	27.93%	23.715	0.179	16.29%	1.08%	3.37
OmniHOI (ours)	**73.49**	**36.21%**	**23.843**	**0.038**	**26.56%**	**4.25%**	5.31
OOD Object Category Generalization on HUMOTO							
Real	-	14.47%	22.038	0.008	-	4.73%	1.67
IMoS†	364.56	1.64%	14.060	2.444	0.07%	0.02%	2.16
MDM†	336.17	1.97%	12.091	0.034	6.95%	5.16%	2.16
HIMO-Gen	231.23	2.30%	15.025	0.063	3.01%	4.72%	2.90
Dex2HOI	219.55	2.63%	13.946	0.026	6.63%	**6.39%**	**1.82**
OmniHOI (ours)	**168.79**	**5.26%**	**16.679**	**0.014**	**8.21%**	5.01%	2.59

Qualitative comparison. In the teapot example (Fig. 3a), OmniHOI coordinates both hands: one tilts the teapot while the other holds the cup beneath the spout. Dex2HOI pours with the spout away from the cup. With unseen objects (Fig. 3b), OmniHOI preserves the bimanual structure of the task, holding the mango in one hand and bringing the peeler to its surface with the other, whereas Dex2HOI's peeler separates from the hand.

4.4 Ablation and Inference Cost

Ablation study. Tab. 5 ablates the three main components of OmniHOI on the HIMO 2-object subset. Removing the hierarchical generation framework causes the largest degradation in both hand–object and object–object interaction accuracy, confirming the importance of decomposing object, hand, and body generation. Interaction-aware self-conditioning primarily improves object–object coordination, while the interaction-informed regularization improves hand–object interaction. Together, these results show that the stage-specific interaction modeling introduced in Secs. 3.3– 3.4 provides complementary benefits.

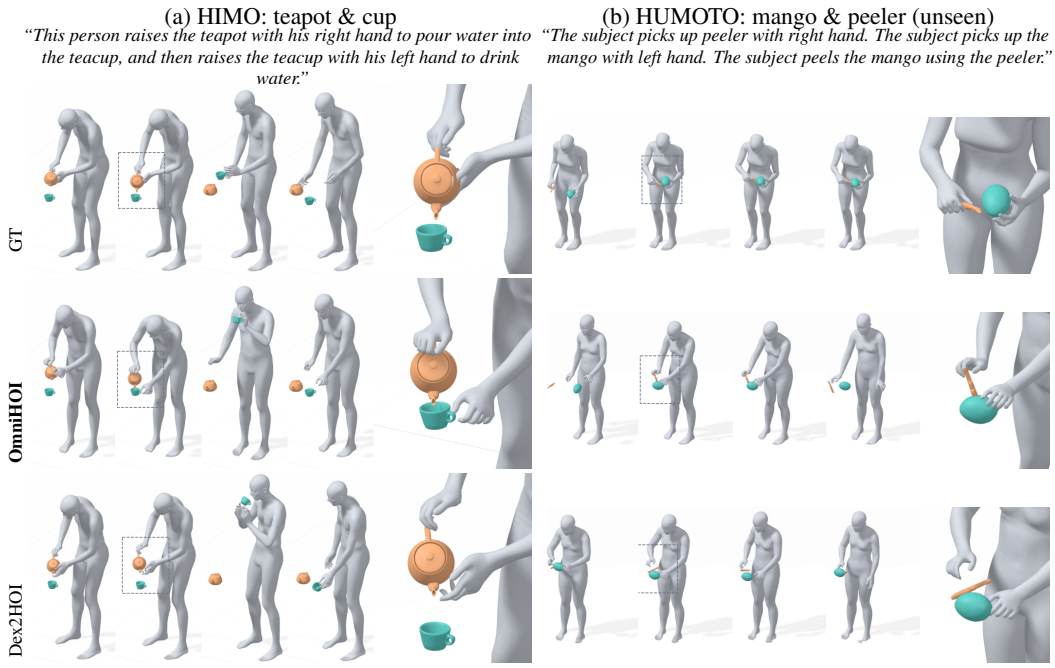


Figure 3: **Qualitative comparison**. Each row shows four keyframes followed by an enlarged view of the boxed frame. (a) Two-object interaction from HIMO. (b) Objects from HUMOTO that are unseen during training.

Table 5: **Ablation of the main components of OmniHOI on the HIMO 2-object subset.**

Method	HOI-ADD ↑	HOI-ADD-S ↑	OO-ADD ↑	OO-ADD-S ↑
OmniHOI w/o Hierarchical Generation	16.88%	37.89%	11.64%	25.80%
OmniHOI w/o Interaction-Aware Self-Conditioning	–	–	25.53%	38.21%
OmniHOI w/o Interaction-Informed Regularization	20.56%	37.21%	–	–
OmniHOI (ours)	**22.89%**	**38.84%**	**27.10%**	**39.57%**

Inference cost. OmniHOI has 409.4M parameters and generates an average 310-frame sequence in 3.77 s on a single AMD Instinct MI100 GPU, dominated by whole-body generation (2.01 s). The lightweight target detector and BPS refinement together add only 0.13 s, so explicit interaction reasoning is nearly free at inference time. The optional foot guidance adds 9.5 s; a per-stage breakdown is given in the Appendix D.

5 Conclusion

We presented OmniHOI, a hierarchical framework that generates object, hand, and full-body motion in sequence and explicitly models the interactions most relevant to each stage. With a single architecture, OmniHOI achieves state-of-the-art results on single-object, articulated-object, and multi-object benchmarks, and generalizes to unseen object trajectories and categories without fine-tuning, suggesting that explicit intermediate motion and geometry provide an effective interface for general HOI generation.

Limitations. OmniHOI is still trained separately per dataset, and stable joint training across heterogeneous HOI datasets remains open. It also assumes rigid or articulated objects; supporting deformable objects will require deformation-aware object representations and interaction features.

ETHICS STATEMENT

We have reviewed the ICLR Code of Ethics and conducted our research in accordance with its guidelines. Our work does not involve human subjects or sensitive personal data, and we are not aware of any significant ethical concerns associated with this research.

REPRODUCIBILITY STATEMENT

We will release the complete codebase used in our paper upon publication to facilitate reproducibility. The codebase will include implementation details, experimental configurations, and supporting analyses necessary to reproduce and verify our results.

AI USE STATEMENT

Generative AI tools were used during the preparation of this work for language editing, code generation and debugging assistance. AI-generated suggestions were reviewed, modified, and validated by the authors before use. All experimental results were produced using the authors' implementation and independently verified by the authors. The authors take full responsibility for the content of the paper and its accompanying code.

REFERENCES

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. 2023a.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023b.

Yan Di, Yuheng Li, Yaoxing Wang, Mengge Liu, Shan Gao, and Xiangyang Ji. Pamotion: Physics-aware motion generation for full-body interaction with multiple objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *CVPR*, 2024.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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*, volume 42, pp. 1–12. Wiley Online Library, 2023.

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 *CVPR*, 2022.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, 2024.

Ziyan Guo, Zeyu Hu, De Wen Soh, and Na Zhao. Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. In *ICCV*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *NeurIPS*, 2023.

Nan Jiang, Yunhao Li, Lexi Pang, Zimo He, Siyuan Huang, and Yixin Zhu. Motionmaster: Generalizable text-driven motion generation and editing. In *CVPR*, 2026.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *ICCV*, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics*, 2023.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024.

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. In *ICCV*, 2025.

Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. *ACM Transactions on Graphics*, 39(4), 2020.

Xueyi Liu and Li Yi. Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. In *International Conference on Learning Representations (ICLR)*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. Humoto: A 4d dataset of mocap human object interactions. In *ICCV*, 2025.

Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Omnigrasp: Grasping diverse objects with simulated humanoids. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xintao Lv Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. Himo: A new benchmark for full-body human interacting with multiple objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2023.

Huaijin Pi, Zhi Cen, Zhiyang Dou, and Taku Komura. Coda: Coordinated diffusion noise optimization for whole-body manipulation of articulated objects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Chrysa Pratikaki, Pablo Ruiz-Ponce, Jiankang Deng, Stefanos Zafeiriou, and Rolandos Alexandros Potamias. Dex2hoi: Dexterous bimanual two-object interaction generation. *arXiv preprint arXiv:2605.30444*, 2026.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019.

Roey Ron, Guy Tevet, Haim Sawdayee, and Amit Bermano. Hoidini: Human-object interaction through diffusion noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *ICLR*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

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

Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *CVPR*, 2022.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Cohenor Daniel, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023.

Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, et al. Grail: Generating humanoid loco-manipulation from 3d assets and video priors. *arXiv preprint arXiv:2606.05160*, 2026.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *ICLR*, 2024.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023.

Yufei Ye, Jiaman Li, Ryan Rong, and C. Karen Liu. WHOLE: World-grounded hand-object lifted from egocentric videos. In *Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7072–7084, 2025.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

A Architecture and Training Details

A.1 Model Architecture

OmniHOI consists of three diffusion-based generation stages for object motion, hand motion, and full-body motion. All three stages use Transformer-based denoisers.

Object-Motion Generator. The object-motion generator jointly models all objects in the scene. Each object is represented as a temporal token stream, while interactions across objects are modeled through object-wise attention together with the interaction-aware geometric features introduced in Sec. 3.3.

The model conditions on the text instruction, initial object configuration, object scale, and static BPS shape descriptors (Prokudin et al., 2019). Variable object counts are handled using object masks.

The diffusion backbone uses a hidden dimension of 512, a feed-forward dimension of 2048, four attention heads, and six encoder and six decoder layers.

Hand-Motion Generator. The hand-motion generator predicts both hands conditioned on the generated object trajectories, object geometry, and text instruction.

Its output follows the representation in Sec. 3.1, including wrist poses, fingertip positions, fingertip contact states, and object-relative BPS representations (Prokudin et al., 2019).

The model uses a Transformer backbone with hidden dimension 512, feed-forward dimension 2048, and four attention heads.

Full-Body Generator. The full-body generator predicts SMPL-X motion (Pavlakos et al., 2019) conditioned on the generated hand and object motion.

The generated wrist trajectories provide the primary interaction constraint, while the generated object trajectories describe the surrounding dynamic scene.

The model uses a Transformer diffusion backbone with hidden dimension 512, feed-forward dimension 2048, four attention heads, and six encoder and six decoder layers.

A.2 Training Details

The three generation stages are trained separately.

For HIMO (Lv et al., 2024), the object-motion and hand-motion generators are trained jointly on both the 2-object and 3-object training sets. Object masking allows the same model to process varying numbers of objects.

The main training settings are summarized in Tab. 6.

Table 6: **Training settings of OmniHOI.**

Component	Batch Size	Learning Rate	Sequence Length
Object-motion generator	32	1×10^{-4}	300
Hand-motion generator	32	1×10^{-4}	300
Target-object detector	16	1×10^{-4}	300
Full-body generator	256	1×10^{-4}	128

All diffusion models are trained with AdamW (Loshchilov & Hutter, 2019) and a 1,000-step cosine diffusion schedule. We use gradient clipping with maximum norm 1 and exponential moving average with decay 0.995, updated every ten optimization steps.

The object-motion and hand-motion generators use a 1,000-step learning-rate warmup and conditioning dropout of 0.1.

At inference time, the object-motion and hand-motion generators use DDIM sampling (Song et al., 2021) with 100 steps, $\eta = 0.8$, and guidance scale 1.0. Full-body motion is generated using 30 sampling steps on overlapping 128-frame windows with stride 96.

B Multi-Object Test-Time Optimization

B.1 Target-Object Detector

For multi-object interaction, the target object associated with each hand must be determined before hand-object optimization. We therefore train a lightweight *target-object detector* operating on the generated object and hand trajectories.

The detector contains four Transformer blocks with hidden dimension 256, four attention heads, and feed-forward dimension 1024. Each block applies temporal self-attention, object-wise self-attention, text cross-attention, and a feed-forward layer.

For each hand and frame, the detector predicts (1) a binary hand-engagement state and (2) a categorical distribution over valid objects.

The detector is trained with

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{engage}} + \mathcal{L}_{\text{target}}, \tag{7}$$

where $\mathcal{L}_{\text{engage}}$ is binary cross-entropy and $\mathcal{L}_{\text{target}}$ is cross-entropy over the target-object labels on engaged frames.

During training, Gaussian noise with standard deviation sampled uniformly from $[0, 0.05]$ is added to normalized object states.

At inference time, target assignments are temporally stabilized to avoid rapid switching between objects. The detected target object determines the object coordinate frame and geometry used by hand-object optimization.

B.2 Hand-Object Motion Optimization

Following Sec. 3.5 and prior BPS-based interaction optimization (Pi et al., 2025), hand-object optimization refines the generated wrist and fingertip trajectories while keeping object motion fixed.

Let $k_t^{h,i}$ denote the i-th keypoint of hand h at frame t, expressed in the coordinate frame of the detected object.

Given the predicted BPS distances $b_t^{h,i}[j]$, the contact term is

$$\mathcal{L}_{\text{contact}} = \text{mean}_{t,h,i,j} \left| \|k_t^{h,i} - u_j\|_2 - b_t^{h,i}[j] \right|, \tag{8}$$

where u_j denotes the j-th basis point. Only valid and engaged hand-object pairs contribute to the loss.

We optimize the hand motion in two stages.

Rigid Hand Adjustment. The first stage applies a rigid translation and rotation correction to each hand:

$$\mathcal{L}_{\text{rigid}} = \mathcal{L}_{\text{contact}} + 0.05\mathcal{L}_{\text{trans}} + 0.1\mathcal{L}_{\text{rot}} + \mathcal{L}_{\text{temp}}. \tag{9}$$

The translation and rotation priors penalize the magnitude of the correction. Temporal regularization penalizes changes in the correction parameters between adjacent engaged frames:

$$\mathcal{L}_{\text{temp}} = \text{mean}_{t,h} \left(\|\tau_{th} - \tau_{t-1,h}\|_2 + \|\omega_{th} - \omega_{t-1,h}\|_2 \right), \tag{10}$$

where τ and ω are the translation and axis-angle rotation corrections.

We optimize using Adam (Kingma & Ba, 2014) with learning rate 0.05 for 300 steps. The learning rate uses a short warmup followed by cosine decay. Translation corrections are bounded by 0.25 m and rotation corrections by 30°.

Residual Keypoint Adjustment. The second stage allows small residual changes to individual wrist and fingertip positions:

$$\mathcal{L}_{\text{res}} = \mathcal{L}_{\text{contact}} + 0.02\mathcal{L}_{\text{pos}} + 0.3\mathcal{L}_{\text{shape}}. \tag{11}$$

The position term constrains the optimized keypoints to remain close to the rigid-stage result,

$$\mathcal{L}_{\text{pos}} = \text{mean}_{t,h,i} \, \|k_t^{h,i} - k_t^{0,h,i}\|_2, \tag{12}$$

where k^0 denotes the output of the rigid adjustment stage.

The shape term preserves the relative configuration of the wrist and fingertips:

$$\mathcal{L}_{\text{shape}} = \text{mean}_{t,h} \frac{1}{15} \sum_{i<r} \left| \|k_t^{h,i} - k_t^{h,r}\|_2 - \|k_t^{0,h,i} - k_t^{0,h,r}\|_2 \right|. \tag{13}$$

We optimize this stage using Adam (Kingma & Ba, 2014) with learning rate 0.05 for 800 steps with cosine decay. Each keypoint is constrained to move by at most 1 cm from the rigid-stage result.

After optimization, wrist orientations are refitted to remain consistent with the refined fingertip configuration.

B.3 Full-Body Motion Optimization

Following prior foot-guidance optimization (Yi et al., 2025), we optionally refine the generated full-body motion to reduce foot skating.

We use the predicted contact confidences of four foot and ankle joints. For adjacent frames, the contact confidence is averaged as

$$\bar{\psi}_j^t = \frac{1}{2} \left(\psi_j^t + \psi_j^{t-1} \right). \tag{14}$$

The contact residual is

$$r_{j,t}^{\text{skate}} = 30 \bar{\psi}_j^t \left(p_{\text{world},j}^t - p_{\text{world},j}^{t-1} \right). \tag{15}$$

The optimization additionally uses the residual terms summarized in Tab. 7.

Table 7: **Residual weights used for full-body motion optimization.**

Residual	Weight
Rotation prior	1
Torso position prior	5
Correction smoothness	10
Rotation smoothness	1
Rotational-velocity smoothness	5
Foot-contact residual	30

Local body rotations are optimized while the generated root translation, root orientation, and body shape remain fixed. We solve the resulting nonlinear least-squares problem using Levenberg–Marquardt optimization with trust-region updates.

C Evaluation Details

C.1 Dataset Splits

GRAB. We use GRAB (Taheri et al., 2020) with the subject-disjoint split of subjects s1–s8 for training, subject s9 for validation, and subject s10 for testing, containing 845, 101, and 145 sequences, respectively.

For the OOD-trajectory experiment, ground-truth object trajectories from the held-out test set directly replace the output of the object-motion generation stage. The subsequent hand and body stages remain unchanged.

HIMO. We use HIMO (Lv et al., 2024) with the splits summarized in Tab. 8.

Table 8: **Dataset splits for HIMO.**

Setting	Train	Validation	Test
2-object	2,233	132	121
3-object	793	47	50

The object-motion and hand-motion generators are trained jointly on both object-count settings. All sequences are converted to a common coordinate frame and evaluated at 30 Hz.

HUMOTO. For cross-dataset generalization, OmniHOI is trained on HIMO (Lv et al., 2024) and directly evaluated on the held-out two-object HUMOTO (Lu et al., 2025) sequences without generator fine-tuning.

A dataset-specific text-motion evaluator is trained separately for metric computation; HUMOTO data is not used to train the HOI generator.

C.2 Baseline Adaptation

When a baseline does not natively support the corresponding interaction setting, we adapt and retrain it using the same training split. These methods are marked with † in the main paper.

We preserve the original model design whenever possible and modify only the representation or input/output dimensions required by the target interaction setting.

For baselines requiring an initial human pose, we provide one observed initial frame. No future ground-truth human motion is used during generation.

C.3 Standard Evaluation Metrics

FID, R@3, and Diversity are evaluated in a TMR-style text-motion embedding space (Petrovich et al., 2023), following the standard text-to-motion evaluation protocol (Guo et al., 2022).

FID is computed from the generated and ground-truth motion embeddings. R@3 is the average of text-to-motion and motion-to-text top-3 retrieval accuracy over the full evaluation gallery. Diversity is computed as the mean pairwise Euclidean distance among generated motion embeddings.

Foot skating follows the standard motion-generation evaluation protocol (Ling et al., 2020) and measures the horizontal velocity of the lower foot between adjacent frames, reported in m/s.

For physical interaction metrics, a hand vertex is considered in contact when its distance to the closest object surface is below 5 mm. Penetration depth is computed as the maximum hand-object penetration depth per frame and averaged over the sequence.

C.4 HOI-ADD and HOI-ADD-S

Following the interaction evaluation protocol of WHOLE (Ye et al., 2026), HOI-ADD evaluates whether the relative configuration between the generated hands and objects matches the ground truth.

Generated and ground-truth sequences are aligned using the bilateral hand trajectories. The resulting transformation A is applied to the generated object vertices.

For object i,

$$d_{it}^{\mathrm{ADD}} = \frac{1}{K_i} \sum_k \left\| A(\hat{x}_{itk}) - x_{itk} \right\|_2 . \tag{16}$$

For symmetry-aware evaluation, HOI-ADD-S instead uses the nearest ground-truth vertex,

$$d_{it}^{\mathrm{ADD-S}} = \frac{1}{K_i} \sum_k \min_r \left\| A(\hat{x}_{itk}) - x_{itr} \right\|_2 . \tag{17}$$

We report normalized AUC up to $\tau = 0.15$ m:

$$\mathrm{AUC}_{15} = 100\,\mathrm{mean}_{i,t}\,\max\left(0, 1 - \frac{d_{it}}{0.15}\right). \tag{18}$$

Higher values indicate more accurate hand-object interaction geometry.

C.5 OO-ADD AND OO-ADD-S

We further introduce OO-ADD to evaluate relative geometric consistency among multiple objects.

For every frame, each object is used in turn as an alignment anchor. Given anchor object a, we compute the rigid transform that aligns the generated anchor pose to its ground-truth pose. The same transform is then applied to every other object $b \neq a$, and ADD is evaluated against its corresponding ground-truth pose.

All directed non-self object pairs are included. A scene with N objects therefore contains $N(N-1)$ evaluated pairs.

The final metric is

$$\text{OO-AUC}_{15} = 100\,\mathrm{mean}_{a\neq b,t}\,\max\left(0, 1 - \frac{d_t^{a\to b}}{0.15}\right). \tag{19}$$

OO-ADD-S follows the same procedure using nearest-neighbor vertex distances.

D INFERENCE COST BREAKDOWN

We measure inference time on a single AMD Instinct MI100 GPU with batch size one.

The benchmark uses 20 HIMO 2-object sequences with an average length of approximately 310 frames. Model loading, disk I/O, evaluation, and visualization are excluded.

Table 9: **Per-stage inference cost on HIMO 2-object sequences.**

Component	Time (s / sequence)
Object-motion generation	0.79
Hand-motion generation	0.32
Target-object detection	0.06
Hand-object optimization	0.07
Full-body generation	2.01
Hand-condition construction	0.10
Articulated-hand generation	0.42
Total	**3.77**
Optional full-body optimization	+9.50

OmniHOI contains approximately 409.4M parameters.

The target-object detector and hand-object optimization together require approximately 0.13 s per sequence, while full-body generation accounts for the largest portion of inference time.

Optional full-body optimization adds approximately 9.5 s and is excluded from the reported 3.77 s generation time.