BUMP: BODY-DRIVEN UNIFIED MULTI-OBJECT PHYSICS FOR HOI SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating human-object interactions (HOIs) with multiple objects is challenging because the resulting motions arise from different underlying causes. Human motion reflects semantic intent, intentionally manipulated objects must remain coordinated with the human, while many surrounding objects move primarily as physical consequences of contact and collision. We present BUMP, a text-driven framework for multi-object HOI generation based on a simple principle: learn intentional motion and simulate physical consequences. Our method first generates full-body human motion from the text and scene context, then predicts the trajectories of a small number of intentionally manipulated *driver* objects conditioned on the human motion. Rather than explicitly generating every moving object, we obtain the motions of physically induced *responder* objects through simulation, allowing the framework to handle varying numbers and configurations of surrounding objects without increasing the learned trajectory output space. To address the scarcity of paired full-body multi-object HOI data, we further combine heterogeneous supervision from human-only motion, full-body HOI, and hand-centric manipulation data augmented with plausible body motion. Finally, we introduce outcome-conditioned interaction editing, which infers a minimal modification of the causal driver trajectory from a desired responder behavior without requiring an explicit target driver trajectory. Extensive experiments demonstrate that BUMP outperforms baseline methods on both single and multiple object interaction generation. We further showcase diverse scenarios where inter-object relationships are crucial (*e.g.*, stacking box), highlighting the advantage of our unified model formulation.

1 INTRODUCTION

Human activities in the real world rarely involve a single object in isolation. Everyday behaviors such as preparing food, organizing a workspace, or playing sports often require a person to intentionally manipulate one or two objects while simultaneously interacting with several surrounding objects. While most HOI generation methods focus on a single manipulated object (Li et al., 2023; 2024; Xu et al., 2023; Peng et al., 2025; Xu et al., 2024; Wu et al., 2025), recent work has begun to address multi-object interactions (Lv et al., 2024; Lu et al., 2025; Di et al., 2026; Pratikaki et al., 2026). However, explicitly generating every interacting object couples human motion, multiple object trajectories, contacts, and inter-object relationships within the learned generation problem.

We argue that these sources of complexity need not be modeled symmetrically. The semantic structure of an interaction is often expressed primarily by the human motion and the intentionally manipulated objects, rather than by every surrounding object that may subsequently participate. For example, the action of swinging a baseball bat is defined by the human motion and the intentionally controlled bat, while the subsequent motion of a struck ball, or another object displaced by that ball, is largely determined by contact dynamics. This motivates a simple principle: *learn intentional motion and simulate physical consequences*.

Based on this principle, we propose BUMP, a text-conditioned framework that decomposes an interaction into the human, intentionally manipulated *driver objects*, and physically induced *responder objects*. We first generate full-body human motion from text and scene context, then generate a small number of driver-object trajectories conditioned on the human. Responder-object trajectories are

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.

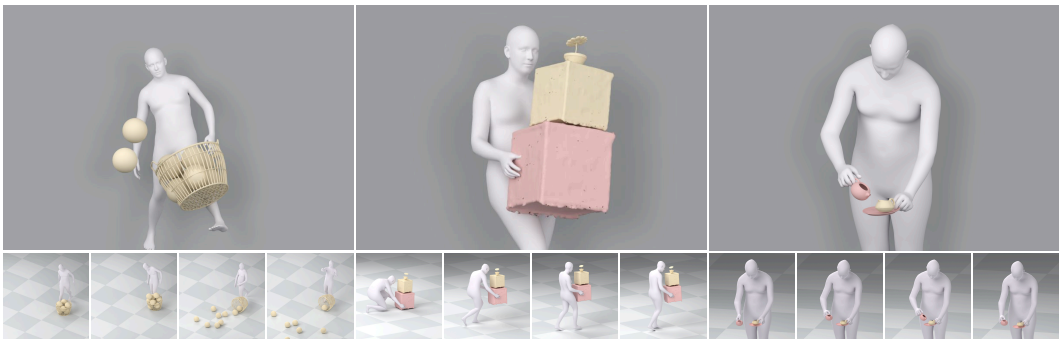


Figure 1: **Overview of BUMP.** Given a text prompt and an initial scene, BUMP generates intentional human and driver-object motion while obtaining physically induced responder motion through simulation. The same causal decomposition also enables outcome-conditioned interaction editing by modifying the driver object to realize a desired responder behavior.

instead obtained through physics simulation from contacts with the human, driver objects, or other responder objects. Thus, surrounding objects can condition the learned motion without requiring their trajectories to become learned outputs, so the learned trajectory space depends primarily on the number of driver objects rather than the total number of responder objects.

To address limited full-body multi-object HOI supervision, we combine human-only motion, captured full-body HOI, and fine-grained manipulation data. Hand-centric and partial-body sequences are lifted to a common full-body representation through offline fitting and retargeting, with reconstructed body regions treated as weaker supervision. The same causal decomposition also enables *outcome-conditioned interaction editing*: given a desired responder behavior, we locally modify the causal driver-object motion and physically replay the interaction to realize the requested outcome.

In summary, our main contributions are:

- We propose BUMP, a **factorized multi-object HOI framework** that learns human and driver-object motion while producing responder motion through physics simulation, decoupling the learned trajectory space from the number of responders.
- We introduce a **heterogeneous full-body interaction data construction strategy** that lifts fine-grained hand-centric and partial-body manipulation data to a common full-body representation, providing additional dexterous and multi-object supervision.
- We introduce **outcome-conditioned interaction editing** that infers a local modification of the causal driver object from a desired responder behavior, without an explicit target driver trajectory.

Code and models will be released.

2 Related Work

Human-Object Interaction Modeling. Prior work has modeled human–object affordances, contacts, and geometric relationships from 3D observations (Deng et al., 2021; Yang et al., 2023; 2024; Tripathi et al., 2023; Kim et al., 2024), providing useful priors for interaction locations. Full-body HOI generation methods such as OMOMO (Li et al., 2023), CHOIS (Li et al., 2024), InterDiff (Xu et al., 2023), HOIDiff (Peng et al., 2025), AffordMotion (Wang et al., 2024), TesMo (Yi et al., 2024), and InterDreamer (Xu et al., 2024) further synthesize temporally coordinated human and object motion. Recent approaches improve generalization or physical plausibility, e.g., DAViD (Kim et al., 2025a) adapts a pretrained human-motion prior to HOI, while HOIfHLI (Wu et al., 2025) combines kinematic generation with physics-based tracking. These methods primarily focus on a single manipulated object.

Multi-Object HOI Generation. Recent work extends HOI synthesis to multiple objects. HIMO (Lv et al., 2024) generates full-body two- and three-object interactions with a dual-branch diffusion model,

while Dex2HOI (Pratikaki et al., 2026) coordinates multiple object-motion streams for dexterous one- and two-object manipulation. PAMotion (Di et al., 2026) further introduces physics-aware objectives for joint generation of human and multi-object trajectories. In these approaches, interacting object motions remain explicit outputs of learned generative models. Our work instead separates intentionally manipulated driver objects from physically induced responder objects, learning only human and driver-object motion while obtaining responder trajectories through simulation. This allows additional objects without expanding the learned trajectory output space.

HOI Data and Heterogeneous Supervision. Full-body multi-object HOI data remains limited. HIMO (Lv et al., 2024), HUMOTO (Lu et al., 2025), and ParaHome (Kim et al., 2025b) capture multi-object interactions, while GRAB (Taheri et al., 2020), InterCap (Huang et al., 2024), BEHAVE (Bhatnagar et al., 2022), ARCTIC (Fan et al., 2023), and InterAct (Xu et al., 2025) provide complementary HOI supervision. We additionally exploit broader human-motion data and fine-grained manipulation datasets such as TACO (Liu et al., 2024b) and OakInk2 (Zhan et al., 2024), lifting heterogeneous sources into a common full-body representation.

Physics-Based Interaction and Editing. Existing physics-aware HOI methods use physics to regularize, refine, or track motions produced by learned models (Di et al., 2026; Wu et al., 2025). In contrast, our simulator directly produces responder-object trajectories, making physics part of the generative process itself. We further exploit this causal structure for outcome-conditioned editing by modifying the causal driver motion from a desired responder outcome rather than requiring a prescribed object trajectory.

3 METHOD

Given a text prompt and an initial scene, BUMP generates human motion together with the motion of multiple objects. Our key idea is to separate objects by their causal role. *Driver objects* are those the human intentionally manipulates, and their motion is generated by our model. *Responder objects* move only when the human or a driver object pushes, hits, or carries them. Rather than generating their motion, we obtain it through physics simulation. This separation lets a learned model handle the parts of the interaction that depend on intent, while a physics simulator handles the parts governed by physics alone. Formally, given a text prompt ℓ, an initial human pose $\mathbf{h}_1$, and an initial scene $\mathcal{O}^0$ consisting of object geometries and poses, BUMP produces the human trajectory $\mathcal{H}$, the driver object trajectories $\mathcal{D}$, and the responder object trajectories $\mathcal{R}$:

$$(\mathcal{H}, \mathcal{D}, \mathcal{R}) = \mathrm{BUMP}(\ell, \mathbf{h}_1, \mathcal{O}^0, \mathcal{S}_R^0). \tag{1}$$

The outputs are produced in three stages that follow the causal order, with each stage conditioned on the outputs of the stages before it:

$$\mathcal{H} = G_H(\ell, \mathbf{h}_1, \mathcal{O}^0), \tag{2}$$

$$\mathcal{D} = G_D(\mathcal{H}, \mathcal{C}, \ell, \mathcal{O}^0), \tag{3}$$

$$\mathcal{R} = \mathrm{Sim}(\mathcal{H}, \mathcal{D}, \mathcal{S}_R^0; \psi). \tag{4}$$

$$\mathcal{C} = P_C(\mathbf{Z}_H, \mathcal{Q}_D), \tag{5}$$

Here G_H and G_D are diffusion transformers that generate human and driver-object motion, respectively. Their architectures follow prior work (Rempe et al., 2026; Li et al., 2024), with the input and output layers adapted to our motion representation. Here $\mathbf{Z}_H$ denotes the frame-wise hidden features of G_H, and $\mathcal{Q}_D$ denotes the object queries of the selected driver objects used by the contact head. $\mathcal{S}_R^0$ denotes the responder initialization used by the simulator, including the object geometry and its initial pose, linear velocity, and angular velocity. P_C predicts the hand-driver contacts $\mathcal{C}$ from $\mathbf{Z}_H$ and $\mathcal{Q}_D$, while Sim is a rigid-body physics simulator with physical parameters ψ. See Figure 2 for an overview.

3.1 PROBLEM SETUP AND NOTATION

Human. The human trajectory is a sequence of T frames, $\mathcal{H} = \{\mathbf{h}_t\}_{t=1}^T$, where each frame is

$$\mathbf{h}_t = \left[\mathbf{p}_t^{\mathrm{root}},\ \boldsymbol{\theta}_t^{\mathrm{body}},\ \boldsymbol{\theta}_t^{\mathrm{hand}}\right] \in \mathbb{R}^{3+22\times6+30\times6}. \tag{6}$$

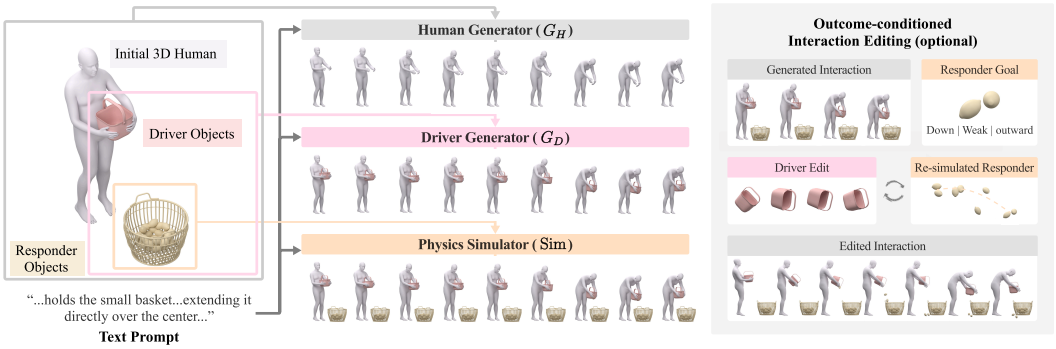


Figure 2: **Overview of the generation and editing pipeline.** We first generate human motion and predict hand-driver contacts. The driver model then generates the manipulated object trajectories conditioned on the human motion and contacts, while physics simulation produces responder motion. For outcome-conditioned editing, we locally modify the driver trajectory, validate the resulting responder motion through simulation, and align the human motion with the edited driver while preserving the original motion.

Here $\mathbf{p}_t^{\text{root}} \in \mathbb{R}^3$ is the root translation, $\boldsymbol{\theta}_t^{\text{body}} \in \mathbb{R}^{22\times 6}$ stacks the global rotations of the 22 body joints, and $\boldsymbol{\theta}_t^{\text{hand}} \in \mathbb{R}^{30\times 6}$ stacks the local rotations of the 30 finger joints (15 per hand). All rotations use the 6D representation (Zhou et al., 2019). The world coordinate frame is defined by the root joint frame of the initial pose $\mathbf{h}_1$. We denote the final frame-wise hidden features of the human generator by $\mathbf{Z}_H = \{\mathbf{z}_{H,t}\}_{t=1}^T$, where $\mathbf{z}_{H,t} \in \mathbb{R}^{1024}$.

Input scene and object roles. The scene contains N objects, $\mathcal{O}^0 = \{\mathbf{o}_i\}_{i=1}^N$, each described by

$$\mathbf{o}_i = \left[\mathbf{g}_i,\ \mathbf{p}_i^0,\ \mathbf{R}_i^0\right] \in \mathbb{R}^{256+3+9}, \tag{7}$$

where $\mathbf{g}_i \in \mathbb{R}^{256}$ is an encoded feature of a 1,024-point Basis Point Set (BPS) representation (Prokudin et al., 2019; Li et al., 2024), $\mathbf{p}_i^0 \in \mathbb{R}^3$ is the initial translation, and $\mathbf{R}_i^0 \in \mathbb{R}^9$ is the flattened initial rotation matrix. The objects are partitioned into driver objects $\mathcal{O}_D$ and responder objects $\mathcal{O}_R$, with at most two driver objects per scene. In the training data, roles are assigned automatically from hand proximity and object motion (Sec. 3.6). At inference, they are given together with the scene. The model takes four objects as input. Scenes with more objects are handled by passing the driver objects and the objects closest to the human to the model, up to four in total, and leaving the remaining objects to the physics simulator as responder objects. When a scene contains fewer than four objects, the unused slots are masked in all stages.

For contact prediction, each selected driver object forms a 3088-D query by concatenating its raw 1024×3 BPS descriptor, 12-D initial pose, and 4-D physical-slot one-hot encoding. We denote the set of selected driver queries by $\mathcal{Q}_D$.

Output trajectories and contacts. The pose of driver object i at frame t is $\mathbf{d}_t^i = (\mathbf{p}_t^i, \mathbf{R}_t^i) \in SE(3)$, and $\mathcal{D} = \{\mathbf{d}_t^i\}$ denotes the driver trajectories, with at most two driver objects per scene. The contact prediction $\mathcal{C}$ contains, for each hand and frame, a categorical distribution over no contact and the two driver objects. Similarly, $\mathbf{r}_t^j \in SE(3)$ denotes the pose of responder object j, and $\mathcal{R} = \{\mathbf{r}_t^j\}$ denotes the responder trajectories.

3.2 Stage 1: Human Motion Generation

We first generate the human trajectory $\mathcal{H}$, which serves as a fixed condition for the subsequent stages (Eq. 2). The human generator G_H is a diffusion transformer trained to recover the clean sequence $\mathcal{H}$ from its noised version, conditioned on the text ℓ, the initial pose $\mathbf{h}_1$, and the scene $\mathcal{O}^0$. Each object $\mathbf{o}_i$ is embedded as a separate token, giving $N \leq 4$ scene tokens. The scene condition $\mathcal{O}^0$. allows the generated human to approach and reach the objects it will manipulate, although G_H does not produce any object motion itself.

Optional human guidance. For qualitative visualization, we optionally guide reverse diffusion toward contact and force-closure grasp targets while penalizing penetration and temporal artifacts; this guidance is not used for quantitative evaluation.

3.3 Stage 2: Contact Prediction and Driver Generation

Driver object motion is represented relative to the manipulating hand. We therefore first predict hand–driver contacts $\mathcal{C}$ (Eq. 5) and use their hard categorical assignments to determine the hand-relative coordinate frame for driver decoding.

Hand-driver contact prediction. The lightweight contact head P_C combines the frozen frame-wise human features $\mathbf{Z}_H$ with the selected driver object queries $\mathcal{Q}_D$ (Eq. 5). For each hand and frame, it predicts a categorical distribution over no contact and the two driver objects. We train P_C after freezing G_H using $\mathcal{L}_C = \mathcal{L}_{\text{CE}} + 0.01\mathcal{L}_{\text{temp}}$, where $\mathcal{L}_{\text{temp}}$ encourages temporal consistency when the ground-truth assignment remains unchanged.

Driver Motion generator. The driver motion generator G_D is a diffusion transformer conditioned on the human joint positions and wrist orientations recovered from $\mathcal{H}$, the contacts $\mathcal{C}$, the text ℓ, and the scene $\mathcal{O}^0$ (Eq. 3). We represent the driver trajectory as offsets from the hand anchors rather than absolute positions in the world frame. For each driver object i and frame t, G_D predicts

$$\bar{\mathbf{d}}_t^i = \left[\Delta\mathbf{p}_{\text{L},t}^i,\ \Delta\mathbf{p}_{\text{R},t}^i,\ \bar{\mathbf{R}}_t^i\right] \in \mathbb{R}^{3+3+9}, \tag{8}$$

which consists of translation offsets from the left and right hands and an unconstrained rotation. This gives a 30-dimensional output per frame for the two driver slots. For decoding, we convert the predicted contact probabilities into discrete assignments indicating which driver, if any, each hand contacts. These assignments determine which hand-relative position candidate is used to reconstruct each driver's position. When both hands contact the same driver, the two candidates are averaged; when neither hand contacts it, a fixed fallback hand is used. Finally, $\bar{\mathbf{R}}_t^i$ is projected onto $SO(3)$ and combined with the reconstructed position to obtain the final driver pose $\mathbf{d}_t^i$.

We first train G_D with the standard diffusion objective on ground-truth human motion paired with ground-truth contact labels, and then fine-tuned with an additional cubic B-spline residual loss. At inference, the ground-truth human motion and contacts are replaced by the generated $\mathcal{H}$ and the predicted $\mathcal{C}$. The human motion $\mathcal{H}$ stays fixed throughout driver sampling.

Optional object guidance. For qualitative visualization, we optionally apply differentiable guidance during driver reverse diffusion. At each guided step, we update the predicted clean driver trajectory using contact- and grasp-consistency objectives together with temporal smoothness and penetration penalties, before computing the posterior transition. Hard categorical contacts determine the hand-relative COM interpretation, while soft contact probabilities modulate the strength of contact guidance. This guidance is applied only during sampling and is not used for quantitative benchmark results.

3.4 Stage 3: Responder Objects by Physics Simulation

Instead of generating responder object motion, we replay the human and the driver objects kinematically in a rigid-body physics simulator and let the responders react to the resulting contacts. We build a MuJoCo (Todorov et al., 2012) scene from the object meshes and their initial poses. Rather than using a single convex hull, each mesh is decomposed into multiple convex collision parts with CoACD. Driver objects follow prescribed kinematic trajectories, while responder objects are simulated as dynamic rigid bodies. All responder objects share a uniform density and the same contact parameters specified in MJCF, so no per-object mass or inertia needs to be measured. When the human is expected to touch responder objects directly, the generated human motion drives capsule and sphere collision proxies attached to the body. Otherwise, human collisions are disabled. In both cases, the simulator never modifies the prescribed human and driver motion.

We then run the simulation (Eq. 4) to obtain the responder trajectories $\mathcal{R}$. The human and driver trajectories are defined at 30 fps. Between consecutive frames, we linearly interpolate driver translations, interpolate their orientations with SLERP, and advance the simulator by 96 substeps. The resulting responder motion follows rigid-body dynamics and can be triggered by the human, by a driver object, or by another responder object. The complete interaction is the union of the prescribed human and driver trajectories and the simulated responder trajectories.

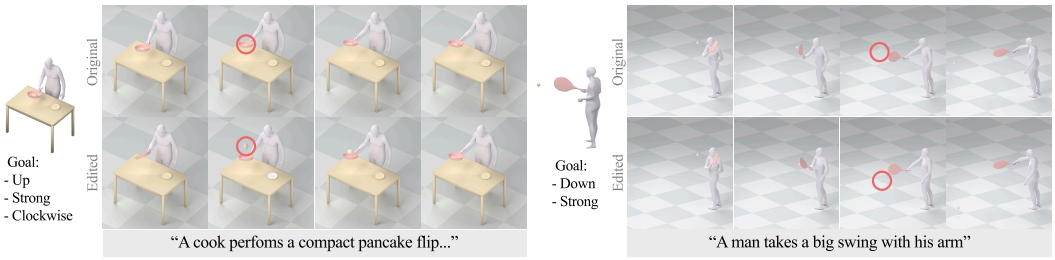


Figure 3: **Outcome-conditioned interaction editing.** We compare the original and edited interactions under different responder-level edit conditions. The edited trajectories produce the requested changes in the physical outcomes while preserving the overall interaction.

3.5 Outcome-Conditioned Interaction Editing

Given a desired responder outcome g_R, we modify the causal driver motion rather than directly editing the responder trajectory. The goal specifies a responder-level outcome, such as a human-relative motion direction and response strength with an optional rotation direction, or a spatial relation to preserve.

For motion goals, the direction category is converted to a unit vector $\mathbf{u} \in \mathbb{R}^3$ at the interaction event. The horizontal directions are defined relative to the human yaw, vertical directions follow the world vertical axis. The response strength determines a directional speed interval $[s_{\min}, s_{\max}]$, and α denotes the allowed angular deviation from $\mathbf{u}$.

For the responder's post-contact velocity $\mathbf{v}_r \in \mathbb{R}^3$, we define the feasible response set

$$\mathcal{V}(g_R) = \left\{ \mathbf{v}_r \in \mathbb{R}^3 \;\middle|\; s_{\min} \le \mathbf{u}^\top \mathbf{v}_r \le s_{\max},\ \left\| \mathbf{v}_r - (\mathbf{u}^\top \mathbf{v}_r)\mathbf{u} \right\|_2 \le \tan(\alpha)\, \mathbf{u}^\top \mathbf{v}_r \right\}. \tag{9}$$

The first constraint specifies the requested velocity component along the semantic direction, while the second restricts the velocity to a cone of half-angle α around that direction. The motion-goal error is measured as the Euclidean distance from the simulated velocity to this feasible set. A motion goal may additionally specify a semantic rotation direction. We define this direction in the human frame at the interaction event and require the responder's post-contact angular velocity to follow the requested axis.

For relation goals, we instead measure violations of the requested relation over the simulated trajectory. For example, a support-preservation goal constrains the responder to remain near its initial driver-relative pose and can additionally require persistent contact.

Local driver correction. We first replay the original interaction and select a local interaction event relevant to the requested outcome. The goal violation at this event is converted into a requested responder velocity change $\Delta\mathbf{v}_r$. For motion goals, this change moves the measured post-contact velocity toward the feasible set $\mathcal{V}(g_R)$. For relation goals, it is derived from the corresponding relation residual. When a rotation goal is present, we additionally construct a local rotational correction using the rotational components of the same SE(3) controls.

We modify the driver object only within a short temporal window around the selected event. The local edit is parameterized by a control vector $\delta \in \mathbb{R}^{5\times(3+3)}$ containing five temporal controls, each with translational and rotational corrections. These controls are smoothly interpolated over the edit window using compactly supported basis functions whose value and first two derivatives vanish at the boundaries, leaving the rest of the driver trajectory unchanged.

At the selected contact, $\Delta\mathbf{v}_r$ is converted to a requested contact impulse and projected onto the local Coulomb friction cone. This gives a feasible change $\Delta\mathbf{v}_c$ in the driver contact-point velocity. For small local edits, we linearize the driver contact-point velocity with respect to the control vector δ and solve

$$\delta^* = \arg\min_{\delta} \left\| \frac{\partial \mathbf{v}_c}{\partial \delta}\, \delta - \Delta\mathbf{v}_c \right\|_2^2 + \lambda_p \|\delta_p\|_2^2 + \lambda_R \|\delta_R\|_2^2, \tag{10}$$

where δ_p and δ_R denote the translational and rotational components of the local driver controls. The first term steers the contact-point velocity toward the requested physical response, while the

remaining terms favor small changes to the original driver trajectory. The resulting local driver edit is used as a candidate, whose actual physical effect is verified through MuJoCo replay.

Simulation-guided selection and human retargeting. We evaluate scaled versions of each candidate correction through full MuJoCo replay and retain only edits that improve the responder objective. Accepted corrections are accumulated and the procedure is repeated until the desired response is reached, no improving edit is found, or the iteration budget is exhausted. During this process, the driver object follows the edited kinematic trajectory while the responder evolves as a dynamic rigid body.

After driver editing, we retarget the contacting human limbs to remain consistent with the modified driver trajectory. We transfer the driver object's SE(3) modification to the corresponding wrist targets and optimize smooth cubic B-spline residual rotations of the arms. When necessary, torso joints are also adjusted using a coarser and more strongly regularized control lattice. The edited driver object remains fixed, while the root translation, finger articulation, and non-edited joints are preserved. The retargeted human, edited driver object, and physically simulated responder trajectory together form the final edited interaction.

3.6 Heterogeneous Interaction Data Construction

Data collection and unification. We construct a heterogeneous training corpus by combining human-only motion, captured full-body HOI, and fine-grained manipulation data. Human-only motion comes from SeqZero, Motion-X++, AIST++, BEAT2, OpenBiomechanicsProject, and LAFAN1 (Yang et al., 2022; Zhang et al., 2025; Li et al., 2021; Liu et al., 2024a; Wasserberger et al., 2022; Harvey et al., 2020). Full-body interactions are obtained from OMOMO, HuMoTo, GRAB, and InterAct (Li et al., 2023; Lu et al., 2025; Taheri et al., 2020; Xu et al., 2025), while TACO and OakInk2 provide detailed hand–object manipulation (Liu et al., 2024b; Zhan et al., 2024). We map all human motions to a common neutral-body model and express the associated object trajectories in the same coordinate system. All sequences are resampled to 30 fps and segmented into 120-frame windows.

Object role assignment. During training-data construction, we automatically assign object roles independently within each interaction window based on object motion and hand proximity. An object is considered a driver candidate if it moves and its sampled surface comes within 2 cm of a hand joint in at least one frame. Others are labeled as *responders*. When more than two objects qualify as driver candidates, we use the two highest-scoring candidates as the generated driver objects.

Full-body fitting from incomplete observations. For TACO, which provides hand and object motion without full-body observations, we keep the observed finger articulation and fit a neutral SMPL-X body to the observed wrist trajectories. We regularize the unobserved body toward a neutral standing posture while encouraging temporal smoothness. For OakInk2, we retarget the observed upper body and hands to the neutral body mdeol, then complete the lower body with grounded, stationary feet while limiting changes to the observed upper body. The fitting objectives are detailed in App. A.6.

Canonicalization and supervision. For each window, we express the human and object trajectories in a canonical coordinate frame defined by the initial human heading and horizontal root position. Hand–object contact labels are derived from the same 2 cm surface-proximity criterion used for driver identification. Dataset-specific supervision masks account for unavailable or reconstructed body channels; in particular, observed manipulation signals are treated as more reliable than synthetically completed body motion.

4 Experiments

We evaluate BUMP on single- and multi-object HOI generation, followed by quantitative comparisons and ablations of the proposed components.

4.1 Datasets

Benchmark evaluation. For single-object interaction, we use InterAct, while for multi-object interaction we use the two- and three-object subsets of HIMO as separate evaluation settings. Within each setting, BUMP and all compared methods are trained using the same benchmark-specific training

Table 1: **Quantitative comparison across single- and multi-object HOI settings.** We report generation quality and interaction quality. Contact F1 is reported in percentage. Metrics are comparable only within each interaction setting.

		Generation				Interaction
Setting	Method	FID↓	R@1/2/3↑	MM-Dist↓	Div. (GT)	Contact F1↑
1 Object (InterAct)	CHOIS	0.2636	55.81/76.48/85.62	5.4210	**1.3509** (1.3539)	26.09
	MDM	**0.1834**	24.31/68.35/79.27	8.7214	1.3593 (1.3539)	16.07
	Dex2HOI	0.8662	50.83/71.34/81.30	5.9722	1.0300 (1.3539)	9.20
	Ours	0.4871	**57.18/77.18/86.04**	**5.4181**	1.2207 (1.3539)	**28.11**
2 Objects (HIMO)	HIMO	0.3114	29.89/48.51/60.74	5.3877	1.3179 (1.4013)	12.73
	Dex2HOI	0.4389	33.32/54.24/65.70	4.1991	1.2794 (1.4013)	43.03
	Ours	**0.2253**	**35.55/58.89/69.34**	**4.0410**	**1.3519** (1.4013)	**44.95**
3 Objects (HIMO)	HIMO	0.3557	27.87/45.78/59.31	4.2359	**1.3621** (1.3805)	12.49
	Ours	**0.3074**	**29.34/48.09/61.28**	**4.1501**	1.3438 (1.3805)	**41.42**
Ablation (HUMOTO)	w/o Lifted Data	0.2475	11.64/22.50/32.19	**3.6145**	6.3578 (6.1040)	32.22
	w/o Pretraining	0.2616	11.88/20.86/30.08	3.7236	6.1744 (6.1040)	31.93
	Ours	**0.2378**	**13.12/22.81/32.34**	3.6384	6.3291 (6.1040)	**34.91**

data and evaluated under the same split and evaluation protocol. The heterogeneous corpus introduced in Sec. 3.6 is not used for these apples-to-apples benchmark comparisons.

Heterogeneous-supervision evaluation. We separately evaluate the effect of the heterogeneous training corpus from Sec. 3.6, which combines human-only motion, captured full-body HOI, and lifted manipulation data. For these ablations, we keep the evaluation set fixed to the held-out HUMOTO split and vary only the available training supervision. The complete corpus contains 57,434 human-only training windows and 65,237 HOI training windows, including 13,545 lifted TACO and OakInk2 windows.

4.2 Evaluation Metrics

We evaluate both the quality of the generated motion distribution and the consistency of the resulting human–object interaction. All metrics are computed independently for each benchmark setting; therefore, scores are compared only among methods evaluated on the same dataset and object-count.

Generation quality and text alignment. Following the standard text-to-motion evaluation protocol (Guo et al., 2022), we report Fréchet Inception Distance (FID), R-Precision (R@1/2/3), Multimodal Distance (MM-Dist), and Diversity. FID measures the discrepancy between the feature distributions of generated and ground-truth motions, with lower values indicating better distributional agreement. R-Precision evaluates text–motion retrieval accuracy, while MM-Dist measures the distance between matched text and motion embeddings. Diversity measures variation among generated motions. Because larger diversity is not necessarily better, we additionally report the ground-truth diversity in parentheses and interpret generated values by their proximity to this reference.

Interaction quality. We evaluate hand-object interaction consistency using pairwise Contact F1. For each frame, contact is determined separately for every hand-object pair using a 2 cm proximity threshold. Contact F1 is computed over these hand-object-frame assignments, such that a prediction is correct only when both the contacting hand and the contacted object match the ground truth. We report Contact F1 as a percentage.

4.3 Results

Quantitative comparisons. Table 1 compares BUMP with prior methods across single- and multi-object interaction settings. BUMP achieves competitive generation quality on InterAct while showing improved hand-object contact consistency. The gains become more evident in the multi-object HIMO settings, where BUMP maintains strong generation quality while substantially improving Contact F1 over existing methods. These results support our factorized formulation for coordinated multi-object interaction generation.

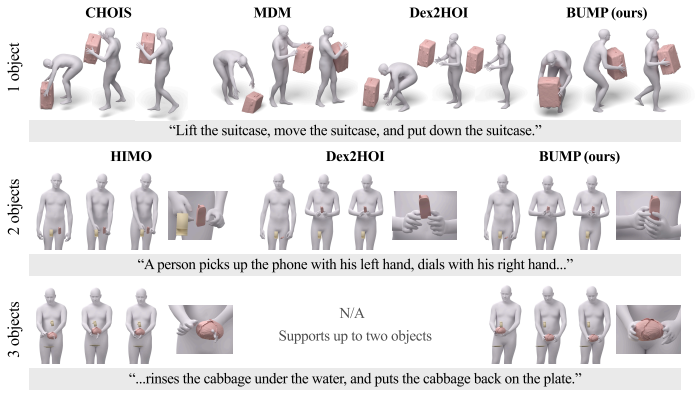


Figure 4: **Qualitative comparison across different numbers of interacting objects.** We compare representative generations on InterAct and HIMO using temporal snapshots and hand-object close-ups.

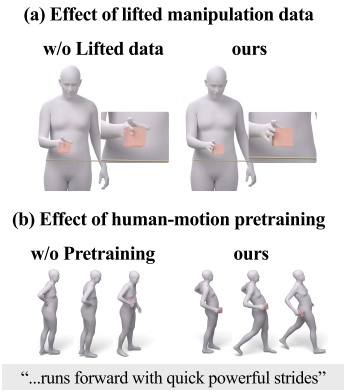


Figure 5: **Qualitative ablation of heterogeneous supervision.**

Qualitative comparisons. Figure 4 shows representative generations across the single-, two-, and three-object settings. BUMP produces coordinated full-body and object motion while maintaining the hand-object relationships required by each interaction. In multi-object examples, the generated motions additionally preserve the coordination among simultaneously manipulated objects.

4.4 Ablation Study

Lifted manipulation data. We first study the effect of augmenting full-body HOI training with the lifted TACO and OakInk2 sequences. As shown in Table 1, removing the lifted data degrades both generation quality and Contact F1 on the held-out HUMOTO evaluation. This suggests that the additional hand-centric manipulation data provides useful supervision for learning detailed multi-object interaction patterns. Qualitatively, Fig. 5(a) further shows improved hand and finger coordination on a held-out GRAB sample.

Human-motion pretraining. We next evaluate the contribution of human-only motion pretraining. Pretraining improves FID, text alignment, and Contact F1, while also increasing the diversity of the generated motions. Interestingly, the resulting Diversity is higher than that of the held-out HUMOTO distribution, suggesting that the human-motion prior broadens the motion space beyond the relatively limited dynamics present in the HOI training data. As shown in Fig. 5(b) on a held-out HUMOTO scene, this effect is particularly visible for prompts requiring larger and more dynamic full-body motion.

5 Discussion and Conclusion

We presented BUMP, a text-driven multi-object HOI framework that separates intentional motion from physical consequences. BUMP generates the human and a small number of manipulated driver objects, while responder motion is obtained through physics simulation. We further leverage heterogeneous supervision from human-only, full-body HOI, and lifted manipulation data, and introduce outcome-conditioned editing that modifies the causal driver trajectory from a desired responder behavior.

Limitations. Our current model supports at most two learned driver objects and assumes rigid-body responders with prescribed human and driver trajectories during simulation. The editing process is local and contact-driven, and part of our training data relies on fitted full-body motion from hand-centric observations. Extending the framework to more drivers, richer physical interactions, and long-horizon editing remains future work.

AI USE STATEMENT

We used generative AI tools to discuss and refine the research methodology and to assist with method implementation and debugging. We also used these tools for literature exploration, manuscript drafting, and editing for clarity and readability. The authors reviewed the AI-assisted content, checked relevant references against their original sources, and assessed implementation changes through testing and experiments. We take full responsibility for the final manuscript, including its methodology, code, experimental results, and claims.

ETHICS STATEMENT

This work studies human-object interaction synthesis using existing motion and interaction datasets, with offline fitting and retargeting to complete missing body motion. These reconstructed motions are approximations rather than additional observations of the original participants. The source datasets may contain biases in body characteristics, action coverage, and interaction settings, which can be inherited by the learned models. Generated interactions may also be used to create misleading depictions of human behavior and should be clearly identified as synthetic when presented outside a research context. Furthermore, prescribed human and driver trajectories and simplified rigid-body simulation do not guarantee that generated interactions are safe or executable in the real world. Use and redistribution of the source data and derived assets remain subject to their original licenses and access conditions.

REPRODUCIBILITY STATEMENT

Section 3 describes the motion representations, generation pipeline, heterogeneous data construction, and outcome-conditioned editing procedure. Appendix A provides additional details on dataset composition, split construction, preprocessing, role assignment, contact labels, and full-body fitting. Appendices B and C describe the human and driver generators, while Appendices D and E describe physics replay and interaction editing. Section 4 specifies the benchmark settings, evaluation metrics, and ablation protocol, distinguishing benchmark-specific comparisons from heterogeneous-supervision experiments. We will release code and pretrained models to support reproduction and further research.

REFERENCES

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15914–15925. IEEE, 2022.

Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13445–13454. IEEE, 2021.

Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1778–1787. IEEE, 2021.

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)*, pp. 30661–30674, June 2026.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12954. IEEE, 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5142–5151. IEEE, 2022.

Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction from multi-view rgb-d images. *International Journal of Computer Vision*, 132(7):2551–2566, 2024.

Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models. In *European Conference on Computer Vision*, pp. 400–419. Springer, 2024.

Hyeonwoo Kim, Sangwon Baik, and Hanbyul Joo. David: Modeling dynamic affordance of 3d objects using pre-trained video diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10330–10341. IEEE, 2025a.

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1816–1828. IEEE, 2025b.

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

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*, pp. 54–72. Springer, 2024.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021. URL https://arxiv.org/abs/2101.08779.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1144–1154. IEEE, 2024a.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21740–21751. IEEE, 2024b.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. Humoto: A 4d dataset of mocap human object interactions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10886–10897. IEEE, 2025.

Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, et al. Himo: A new benchmark for full-body human interacting with multiple objects. In *European Conference on Computer Vision*, pp. 300–318. Springer, 2024.

Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2869–2879. IEEE, 2025.

Chrysa Pratikaki, Pablo Ruiz-Ponce, Jiankang Deng, Stefanos Zafeiriou, and Rolandos Alexandros Potamias. Dex2hoi: Dexterous bimanual two-object interaction generation, 2026. URL https://arxiv.org/abs/2605.30444.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3072–3081. IEEE, 2019.

Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, et al. Kimodo: Scaling controllable human motion generation. *arXiv preprint arXiv:2603.15546*, 2026.

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*, pp. 581–600. Springer, 2020.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7967–7979. IEEE, 2023.

Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 433–444. IEEE, 2024.

KW Wasserberger, AC Brady, DM Besky, BR Jones, and KJ Boddy. The openbiomechanics project: The open source initiative for anonymized, elite-level athletic motion capture data. 2022.

Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11176–11186. IEEE, 2025.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14882–14894. IEEE, 2023.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *Advances in Neural Information Processing Systems*, 37: 52858–52890, 2024.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, et al. Interact: Advancing large-scale versatile 3d human-object interaction generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7048–7060. IEEE, 2025.

Jingfeng Yang, Haoming Jiang, Qingyu Yin, Danqing Zhang, Bing Yin, and Diyi Yang. Seqzero: Few-shot compositional semantic parsing with sequential prompts and zero-shot models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 49–60, 2022.

Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10871–10881. IEEE, 2023.

Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16284–16295. IEEE, 2024.

Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*, pp. 246–263. Springer, 2024.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 445–456. IEEE, 2024.

Yuhong Zhang, Jing Lin, Ailing Zeng, Guanlin Wu, Shunlin Lu, Yurong Fu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x++: A large-scale multimodal 3d whole-body human motion dataset. *arXiv preprint arXiv:2501.05098*, 2025.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5738–5746. IEEE, 2019.

Table 2: **Composition of the heterogeneous corpus.** Each training/validation cell gives processed sequences / canonical windows. The final column gives held-out test sequence counts; canonical HOI test windows were not materialized in this inventory.

Dataset	Train (seq./win.)	Val. (seq./win.)	Test seq.
Human-only			
SeqZero	2,558 / 13,500	592 / 3,205	1,825
Motion-X++	3,411 / 10,630	362 / 1,222	466
AIST++	99 / 1,008	5 / 52	45
BEAT2	5,464 / 21,856	530 / 2,120	1,325
OBP	515 / 1,228	59 / 138	68
LAFAN1	2,303 / 9,212	290 / 1,160	138
Total	*14,350 / 57,434*	*1,838 / 7,897*	*3,867*
Full-body HOI			
OMOMO	3,747 / 25,847	605 / 3,367	534
HuMoTo	558 / 5,954	70 / 730	70
GRAB	1,070 / 8,442	120 / 886	145
InterAct	1,182 / 11,449	168 / 1,515	175
Hand-centric or partial-body HOI			
TACO	1,854 / 10,228	231 / 1,474	231
OakInk2	891 / 3,317	82 / 309	314
HOI total	*9,302 / 65,237*	*1,276 / 8,281*	*1,469*

A Appendix

You may include other additional sections here.

A Additional Data Construction Details

This section describes the offline preprocessing used to construct the heterogeneous corpus in Sec. 3.6. We distinguish processed sequence rows from motion windows: a source recording may be divided into several processed clips, and each clip can yield overlapping windows. The prepared data are organized as 120-frame windows at a nominal 30 fps.

A.1 Dataset Composition

We combine human-only motion from SeqZero, Motion-X++, AIST++, BEAT2, OpenBiomechanicsProject (OBP), and LAFAN1 (Yang et al., 2022; Zhang et al., 2025; Chung et al., 2021; Li et al., 2021; Liu et al., 2024a; Wasserberger et al., 2022; Harvey et al., 2020); full-body HOI from OMOMO, HuMoTo, GRAB, and InterAct (Li et al., 2023; Lu et al., 2025; Taheri et al., 2020; Xu et al., 2025); and hand-centric or partial-body HOI from TACO and OakInk2 (Liu et al., 2024b; Zhan et al., 2024). Table 2 reports the retained processed sequences and canonical windows. The counts refer to the selected all-source human-first training data, rather than the number of original recordings. In particular, BEAT2 and LAFAN1 contain pre-cut clips before canonical window extraction.

Splits are established before overlapping windows are formed. For OMOMO, subjects 1–13/14–15/16–17 form the train/validation/test partitions; GRAB reserves subjects s7/s10 for validation/test; and InterAct reserves NeuralDome subjects 10/01, respectively. HuMoTo uses disjoint action–object-category groups, TACO uses disjoint capture-date/session groups, and OakInk2 reserves actors A008/A007 for validation/test. AIST++ and BEAT2 retain their source split families, OBP is split by athlete, and LAFAN1 by theme and take; SeqZero and Motion-X++ retain their materialized split assignments. The legacy `interact__` records in the combined data belong to OMOMO and are distinct from the separately added InterAct source.

A.2 Motion Unification and Windowing

We map the human poses to a common neutral-body representation and apply the same coordinate transformation to each associated object track. Source-specific temporal conversion precedes fitting when needed; in particular, OakInk2 is sampled from 120 to 30 Hz before retargeting.

The canonical frame of an interaction window is defined by the initial human heading and horizontal root position, leaving its vertical placement unchanged. The same rigid transformation is applied to all object trajectories.

For a window length of 120 frames, the cache advances by 30 frames. A short sequence or a trailing window with fewer than 120 valid frames is zero-padded to length 120, and padded frames are excluded from the training loss. The resulting 120-frame horizon corresponds to four seconds at 30 fps when all frames are valid. Stage-specific training statistics are used to normalize the position channels; held-out windows use the saved training statistics.

A.3 Object Role Assignment

Roles are assigned separately for each interaction window from its observed motion. Let $\tilde{\mathbf{p}}_{i,t}$ be the normalized object-center position used by the role ranker, and let $\mathbf{r}_{i,t} \in \mathbb{R}^9$ be its stored rotation representation. The ranker computes

$$M_i = \max_t \|\tilde{\mathbf{p}}_{i,t} - \tilde{\mathbf{p}}_{i,1}\|_2 , \qquad Q_i = \mathrm{mean}_{t,k} \, |r_{i,t,k} - r_{i,1,k}| . \tag{11}$$

An object is considered moving when $M_i > 0.02$ or $Q_i > 0.01$. The positional threshold is measured in normalized object-position coordinates and should not be interpreted as a 2 cm physical displacement.

For each moving object, we count the frames in which one of the four body hand/wrist joints (indices 20–23) or 30 finger joints lies within 2 cm of its reconstructed BPS surface points. If at least one such frame exists, the object is considered a driver candidate. Other moving objects are responders, while non-moving objects are static.

The driver generator contains two output slots. When more than two candidates occur, candidate i is ranked by

$$s_i = 10C_i + M_i + 0.1Q_i, \tag{12}$$

where C_i denotes the number of frames containing hand–surface contact. The two highest-scoring candidates are assigned to the generated driver slots. Ties retain the original cache ordering. Candidates outside the two selected slots retain their driver identity in the role metadata and are not relabeled as responders. In the selected cache, 115 of 73,518 interaction windows contain more than two driver candidates: 112 contain three and three contain four.

The role layout is derived from the reference interaction window and supplied during the evaluated generation procedure. The contact head described in Sec. C predicts temporal hand–driver contact but does not infer these discrete object roles.

A.4 Hand–Object Contact Labels

We construct the contact head's framewise training labels after assigning the driver slots. For each selected driver, we reconstruct the SMPL-X hand meshes and measure the minimum distance between each hand's vertices and sampled points on the object surface. The left and right hand masks contain 772 and 763 vertices, respectively. The object side uses all 1,024 per-frame BPS nearest-surface samples with stride one; transformed mesh vertices are used as a fallback when BPS samples are unavailable.

A hand–driver pair is labeled as being in contact when the minimum distance is at most 0.02 m. Unlike the role-selection procedure above, which measures proximity using hand/wrist and finger joints, the contact-head supervision is computed using hand-mesh vertices. The two procedures therefore share the same 2 cm proximity threshold but operate on different point sets.

A.5 Object Geometry Representation

We follow the Basis Point Set (BPS) representation used in CHOIS (Li et al., 2024). A fixed set of 1,024 basis points lies within a radius-1 m ball centered at the object. For each basis point, we encode the displacement to its nearest point in the sampled object geometry. The raw descriptor is

$$\mathbf{G}_i \in \mathbb{R}^{1024\times 3}. \tag{13}$$

Its 3,072 values are flattened and passed through a learned encoder to obtain the 256-D object-geometry feature used for conditioning. The 256-D feature denotes the encoder output, rather than the number of BPS points.

A.6 TACO Full-Body Fitting

TACO provides articulated hand and object trajectories but no complete full-body motion. We therefore recover auxiliary full-body motion through offline SMPL-X fitting while preserving the observed finger articulation.

For each frame t and hand $h \in \{L, R\}$, let $\widehat{\mathbf{p}}_{t,h}$, $\widehat{\mathbf{R}}_{t,h}$, and $\widehat{\boldsymbol{\phi}}_{t,h}$ denote the preprocessed wrist position, wrist orientation, and finger articulation, respectively. We use a neutral SMPL-X model with zero shape parameters, $\boldsymbol{\beta} = \mathbf{0}$, and unit body scale. The observed finger articulation is fixed throughout optimization.

If the mean object-center height lies below 0.85 m, the hand and object tracks are shifted upward together before fitting, preserving their relative placement.

The fitted joints and joint rotations are obtained as

$$(\mathbf{J}_t, \mathbf{R}_t) = \mathcal{M}_{\text{SMPL-X}}\left(\boldsymbol{\beta} = \mathbf{0},\, \boldsymbol{\theta}_t,\, (0, \psi, 0),\, \mathbf{c},\, \widehat{\boldsymbol{\phi}}_{t,L},\, \widehat{\boldsymbol{\phi}}_{t,R}\right), \tag{14}$$

where $\boldsymbol{\theta}_t$ denotes the body pose, ψ is a sequence-level root yaw, and $\mathbf{c}$ is a sequence-level global translation.

We optimize

$$\{\boldsymbol{\theta}^*_{1:T}, \psi^*, \mathbf{c}^*\} = \arg\min_{\boldsymbol{\theta}_{1:T},\psi,\mathbf{c}} \; 3000\,\mathcal{L}_{\text{wpos}} + 12\,\mathcal{L}_{\text{wrot}} + \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{stand}} + \mathcal{L}_{\text{temp}} + \mathcal{L}_{\text{aux}}. \tag{15}$$

The wrist-position term is

$$\mathcal{L}_{\text{wpos}} = \frac{1}{6T}\sum_{t=1}^{T}\sum_{h\in\{L,R\}} \|\mathbf{J}_{t,h} - \widehat{\mathbf{p}}_{t,h}\|_2^2, \tag{16}$$

and wrist orientation is constrained by

$$\mathcal{L}_{\text{wrot}} = \frac{1}{2T}\sum_{t=1}^{T}\sum_{h\in\{L,R\}} \left[1 - \text{clip}\left(\frac{\text{tr}\left(\mathbf{R}_{t,h}^{\mathsf{T}}\widehat{\mathbf{R}}_{t,h}\right) - 1}{2}, -1, 1\right)\right]. \tag{17}$$

The MANO finger poses $\widehat{\boldsymbol{\phi}}_{t,h}$ are treated as observations and are not optimized.

The fitting variables are the 21 per-frame body-pose rotations, one sequence-level root yaw, and one sequence-level 3D translation. Adam jointly optimizes these variables for 400 iterations, after which the body poses alone are optimized for another 100 iterations while the root yaw and translation remain fixed. The learning rates are 0.02 for body pose and 0.01 for root yaw and translation.

Distance-aware temporal regularization. Because body regions farther from the observed hands are less directly constrained by the input, we regularize them more strongly in time.

For a sequence $\mathbf{X}_t \in \mathbb{R}^D$, we define acceleration and velocity energies as

$$A(\mathbf{X}) = \frac{1}{(T-2)D}\sum_{t=2}^{T-1} \|\mathbf{X}_{t+1} - 2\mathbf{X}_t + \mathbf{X}_{t-1}\|_2^2, \tag{18}$$

$$V(\mathbf{X}) = \frac{1}{(T-1)D}\sum_{t=1}^{T-1} \|\mathbf{X}_{t+1} - \mathbf{X}_t\|_2^2. \tag{19}$$

The temporal fitting objective is

$$\begin{aligned}\mathcal{L}_{\text{temp}} = 5A(\boldsymbol{\theta}) + 250\,[A(\boldsymbol{\theta}_{\mathcal{T}}) + 0.10V(\boldsymbol{\theta}_{\mathcal{T}})] \\ + 12\,[A(\boldsymbol{\theta}_{\mathcal{A}}) + 0.015V(\boldsymbol{\theta}_{\mathcal{A}})] \\ + 1000\,[A_w(\mathbf{J}_{\mathcal{D}}) + 0.05V_w(\mathbf{J}_{\mathcal{D}})]\,,\end{aligned} \tag{20}$$

where

$$\mathcal{T} = \{3, 6, 9, 12, 15\}, \tag{21}$$

$$\mathcal{A} = \{13, 14, 16, 17, 18, 19, 20, 21\}, \tag{22}$$

$$\mathcal{D} = \{0, 3, 6, 9, 12, 13, 14, 15, 16, 17\} \tag{23}$$

denote the torso, arm, and distance-regularized joint sets, respectively.

For joint j, its distance to the closest fitted wrist is

$$d_{t,j} = \min_{h \in \{L,R\}} \|\mathbf{J}_{t,j} - \mathbf{J}_{t,h}\|_2\,. \tag{24}$$

We map this distance to a smooth temporal weight as

$$u_{t,j} = \text{clip}\left(\frac{d_{t,j} - 0.18}{0.55 - 0.18}, 0, 1\right), \tag{25}$$

$$w_{t,j} = \text{sg}\left(u_{t,j}^2(3 - 2u_{t,j})\right), \tag{26}$$

where sg denotes stop-gradient. Thus, the temporal weight increases smoothly between 18 and 55 cm from the nearest wrist.

For second- and first-order temporal differences, respectively, we use

$$\bar{w}_{t,j}^{(2)} = \frac{w_{t+1,j} + 2w_{t,j} + w_{t-1,j}}{4}, \tag{27}$$

$$\bar{w}_{t,j}^{(1)} = \frac{w_{t+1,j} + w_{t,j}}{2}. \tag{28}$$

The distance-weighted acceleration term is

$$A_w(\mathbf{J}_{\mathcal{D}}) = \frac{\displaystyle\sum_{t=2}^{T-1}\sum_{j\in\mathcal{D}} \bar{w}_{t,j}^{(2)} \|\mathbf{J}_{t+1,j} - 2\mathbf{J}_{t,j} + \mathbf{J}_{t-1,j}\|_2^2}{3\max\left(\displaystyle\sum_{t=2}^{T-1}\sum_{j\in\mathcal{D}} \bar{w}_{t,j}^{(2)}, 10^{-8}\right)}, \tag{29}$$

and the weighted velocity term is

$$V_w(\mathbf{J}_{\mathcal{D}}) = \frac{\displaystyle\sum_{t=1}^{T-1}\sum_{j\in\mathcal{D}} \bar{w}_{t,j}^{(1)} \|\mathbf{J}_{t+1,j} - \mathbf{J}_{t,j}\|_2^2}{3\max\left(\displaystyle\sum_{t=1}^{T-1}\sum_{j\in\mathcal{D}} \bar{w}_{t,j}^{(1)}, 10^{-8}\right)}. \tag{30}$$

This weighting suppresses unsupported motion in the torso and other weakly observed regions while retaining sufficient arm flexibility to match the observed hand trajectories.

Pose and standing priors. We further regularize the recovered body toward plausible standing poses:

$$\begin{aligned}\mathcal{L}_{\text{pose}} = 0.05\ \text{mse}(\boldsymbol{\theta}) + 28\ \text{mse}(\boldsymbol{\theta}_{\mathcal{T}}) \\ + 700\ \text{mean}_{t,j\in\mathcal{T}}\left[(\|\boldsymbol{\theta}_{t,j}\|_2 - 0.22)_+^2\right] \\ + 0.7\ \text{mse}(\boldsymbol{\theta}_{\mathcal{H}}) + 5\ \text{mse}(\boldsymbol{\theta}_{\mathcal{L}}) \\ + 0.05\ \text{mse}\big((0, \psi, 0)\big) + 0.2\ \text{mse}(\mathbf{c} - \mathbf{c}_0),\end{aligned} \tag{31}$$

where

$$\mathcal{H} = \{12, 15\}, \tag{32}$$
$$\mathcal{L} = \{1, 2, 4, 5, 7, 8, 10, 11\}. \tag{33}$$

The standing objective is

$$\mathcal{L}_{\text{stand}} = 50\mathcal{L}_{\text{ground}} + 50\mathcal{L}_{\text{foot-static}} + 40\mathcal{L}_{\text{lower-target}}, \tag{34}$$

where

$$\mathcal{L}_{\text{ground}} = \frac{1}{T}\sum_{t=1}^{T}\left(\min_{j\in\mathcal{F}} J_{t,j,y}\right)^2, \tag{35}$$

$$\mathcal{L}_{\text{foot-static}} = \frac{1}{8T}\sum_{t=1}^{T}\sum_{j\in\mathcal{F}} \left\|\Pi_{xz}\left(\mathbf{J}_{t,j} - \mathbf{J}_{1,j}\right)\right\|_2^2, \tag{36}$$

with

$$\mathcal{F} = \{7, 8, 10, 11\}. \tag{37}$$

The lower body is therefore not fixed to a neutral pose but optimized under the standing and planted-foot priors.

Auxiliary geometric priors. The complete objective additionally includes

$$\begin{aligned}\mathcal{L}_{\text{aux}} = {} & 25\mathcal{L}_{\text{facing}} + 12\mathcal{L}_{\text{gaze}} + 1.5\mathcal{L}_{\text{torso-look}} \\ & + 10\mathcal{L}_{\text{reach}} + 2\mathcal{L}_{\text{root-place}}.\end{aligned} \tag{38}$$

These terms stabilize global body placement and orientation relative to the observed interaction while leaving the measured hand trajectories unchanged.

Overall, the fitting procedure optimizes only the missing body motion, sequence-level root yaw, and global translation. The observed finger articulation is preserved, the body shape is fixed to neutral ($\boldsymbol{\beta} = \mathbf{0}$), and body scale is fixed to one. The resulting fitted human and observed hand/object trajectories receive one common ground-height correction before export.

A.7 OakInk2 Retargeting and Lower-Body Completion

OakInk2 is sampled from its 120-Hz source at a stride of four before fitting. We reconstruct its partial body and hands, then retarget them to a neutral, zero-shape SMPL-X model.

The first stage matches source joint positions for

$$\mathcal{J}_U = \{0, 3, 6, 9, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21\} \cup \{25, \ldots, 54\}, \tag{39}$$

with additional emphasis on

$$\mathcal{J}_H = \{20, 21\} \cup \{25, \ldots, 54\}. \tag{40}$$

We optimize per-frame translation, both hand-pose vectors, and body-pose offsets at zero-based indices $(2, 5, 8, 12, 13, 15, 16, 17, 18, 19, 20)$; the source root orientation and lower-body pose remain fixed. With E_U, E_H denoting mean squared joint-position errors on the two sets, and Δ denoting unscaled frame differences, the objective is

$$\begin{aligned}\mathcal{L}_{\text{retarget}} = {} & E_U + 2E_H + 0.5E_{\Delta U} + 0.25E_{\Delta^2 U} \\ & + 0.03E_{\delta\theta_U} + 0.02(E_{\theta_L} + E_{\theta_R}) + 0.05E_{\delta t}.\end{aligned} \tag{41}$$

The final three terms penalize changes to the optimized body pose, the two hand poses, and translation. Adam runs for 45 iterations, using learning rates 8×10^{-4} for body pose and translation and 10^{-3} for the hand poses.

In the second stage, we initialize planted-foot anchors using the coordinate-wise temporal medians of the stage-one left/right ankle and toe trajectories, corresponding to joint indices $7, 10, 8, 11$. We

optimize lower-body pose offsets at body-pose indices $(0, 1, 3, 4, 6, 7, 9, 10)$, small torso offsets at $(2, 5, 8)$, and translation offsets. The fitted finger and arm rotations remain fixed.

The completion objective is

$$\begin{aligned}\mathcal{L}_{\text{OakInk2}} = 80E_{\text{anchor}} + 20E_{\Delta\text{foot},xz} + 3E_U + 6E_H \\ + 0.03E_{\delta\text{lower}} + 0.2E_{\Delta\delta\text{lower}} + 3E_{\delta\text{torso}} + 5E_{\delta t}.\end{aligned} \tag{42}$$

All E terms denote mean squared errors. The first line constrains planted-foot positions and horizontal foot displacement while retaining the source upper-body and hand positions. The second line regularizes lower-body, torso, and translation changes.

Adam runs for 80 iterations, with learning rate 0.02 for the lower body and 8×10^{-4} for torso and translation. After fitting and converting to the common up-axis, a single sequence-level ground-height shift is computed from the median of per-frame minimum human-vertex heights and applied equally to the human and object trajectories.

A.8 Dataset-Specific Supervision

The human-only pretraining data distinguish captured hand articulation from neutral placeholders on a per-window basis. Motion-X++ and BEAT2 provide the hand-labeled windows in the selected cache; SeqZero, AIST++, OBP, and LAFAN1 use neutral placeholder hands. For windows whose local hand rotations are all neutral, the hand-rotation reconstruction and finger forward-kinematics losses are masked, while body supervision remains active.

During HOI adaptation, source-aware reconstruction weights are used for TACO: root position and body rotations receive weight 1, while the observed hand rotations receive weight 2. The other HOI sources, including OakInk2, do not receive this TACO weighting merely because part of the body motion was fitted. In particular, the selected 315-D training configuration does not apply an additional OakInk2-specific lower-body down-weight. Other sequence-tag-specific masks are independent of the TACO and OakInk2 fitting categories.

B Human Generator Details

B.1 Architecture and Conditioning

The human generator operates on 120-frame sequences sampled at 30 fps. Each frame follows the 315-D representation defined in the main paper.

The scene contains at most four object-conditioning slots. Each object token consists of a 256-D encoded BPS geometry feature, a 3-D initial translation, and a 9-D initial rotation representation, resulting in a 268-D token. Unused slots are masked.

B.2 Human-Only Pretraining

During human-only pretraining, all object-condition slots are masked. We train the diffusion model using the human-only datasets described in Sec. A, with the first frame supplied as the initial human condition.

Motion channels unavailable in a source sequence are excluded through the dataset-specific masks described above.

B.3 Interaction Adaptation

Interaction adaptation is initialized from the human-only pretrained checkpoint. We enable object conditioning and train using the captured full-body HOI and fitted manipulation sequences described in Sec. A. Human-only batches are interleaved with interaction batches.

For a human-only batch, a frozen copy of the pretrained generator supplies a motion reference for the root and body channels. The prior term is not applied to the hand channels, allowing the hand motion to adapt to the interaction supervision.

B.4 Human Motion Sampling

At inference time, we condition the adapted model on the text description, initial human state, and object conditions and sample the remaining human trajectory.

C Driver Generator Details

C.1 Hand–Driver Contact Prediction

Let F_H denote the final temporal hidden features of the frozen human generator. The contact head combines these features with queries associated with the two generated driver slots. For each hand and frame, it predicts a categorical distribution over three classes: no contact, driver 1, and driver 2.

We optimize

$$\mathcal{L}_C = \mathcal{L}_{\mathrm{CE}} + 0.01\,\mathcal{L}_{\mathrm{temp}}, \tag{43}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the categorical cross-entropy loss. The temporal term is applied when the ground-truth driver assignment remains unchanged across consecutive frames and penalizes changes in the corresponding predicted probabilities.

C.2 Human Conditioning

Before driver generation, the generated 315-D human representation is converted to joint positions through forward kinematics. The driver model receives these joint positions together with wrist orientations as its human-motion condition.

C.3 Driver Motion Representation and Decoding

For each driver i, the diffusion model predicts

$$\widetilde{D}_t^i = \left[\Delta\mathbf{p}_{L,t}^i, \Delta\mathbf{p}_{R,t}^i, \mathbf{R}_t^i\right] \in \mathbb{R}^{15}. \tag{44}$$

The first two components are 3-D translation offsets relative to the left- and right-hand anchors. The final component $\mathbf{R}_t^i \in \mathbb{R}^9$ is interpreted as an unconstrained 3×3 rotation prediction. With two driver slots, the model predicts 30 dimensions per frame. Unused driver slots are masked.

For hand $h \in \{L, R\}$, the hand anchor is defined as the mean position of the corresponding 15 finger joints,

$$\mathbf{a}_{h,t} = \frac{1}{15} \sum_{k \in \mathcal{J}_h} \mathbf{x}_{k,t}. \tag{45}$$

The two hand-relative predictions yield candidate object translations

$$\tilde{\mathbf{p}}_{h,t}^{\,i} = \mathbf{a}_{h,t} + \Delta\mathbf{p}_{h,t}^i. \tag{46}$$

When $\hat{c}_{L,t}^i + \hat{c}_{R,t}^i > 0$, the two candidates are combined using the normalized soft contact probabilities:

$$\mathbf{p}_t^i = \frac{\hat{c}_{L,t}^i \tilde{\mathbf{p}}_{L,t}^{\,i} + \hat{c}_{R,t}^i \tilde{\mathbf{p}}_{R,t}^{\,i}}{\hat{c}_{L,t}^i + \hat{c}_{R,t}^i}. \tag{47}$$

The fallback is invoked only when both driver-contact probabilities are exactly zero. A high probability for the no-contact class alone does not trigger the fallback as long as either driver-contact probability is nonzero. When both are zero, we use the position hypothesis associated with the sequence-level preferred hand.

To obtain a valid orientation, we apply Gram–Schmidt orthogonalization to the first two rows of the predicted 3×3 matrix and construct the third row from their cross product. The resulting matrix lies in $SO(3)$ and is used together with $\mathbf{p}_t^i$ as the final driver pose.

C.4 Driver Diffusion Training

The driver model is first trained with the base diffusion objective using paired human trajectories and ground-truth hand–driver contacts. After this stage, we further fine-tune the model with an additional cubic B-spline residual loss to encourage smooth driver trajectories.

The B-spline loss used during driver-model fine-tuning is distinct from the second-difference trajectory stabilization applied before physics replay in Sec. D.1.

During training, the model receives ground-truth human motion and ground-truth contacts. At inference time, these conditions are replaced by the generated human trajectory and the predicted soft contacts.

D Physics Simulation Details

D.1 Driver Trajectory Stabilization

Before every physics rollout, we stabilize each generated driver trajectory using a second-difference smoothing objective. For each trajectory component x, we solve

$$\min_{\bar{x}} \ \|\bar{x} - x\|_2^2 + \lambda \|D_2 \bar{x}\|_2^2, \qquad \bar{x}_0 = x_0, \quad \bar{x}_{T-1} = x_{T-1}, \tag{48}$$

where each row of D_2 contains the discrete second-difference stencil $[1, -2, 1]$. The endpoint constraints preserve the first and last generated poses exactly.

The user-facing smoothing strength $s \in [0, 1]$ is converted to

$$\lambda = \frac{s}{\max(1 - s, 10^{-6})}. \tag{49}$$

The same quadratic solve is applied independently to the translation and hemisphere-aligned quaternion components. Quaternion trajectories are renormalized after smoothing. This is a component-wise Euclidean smoothing operation and is not an optimization directly on $SO(3)$.

D.2 Scene and Collision Construction

For each interaction, we construct a MuJoCo scene from the object geometries and their initial poses. The source collision mesh for object i is its scaled `{object_name}_cleaned_simplified.obj` mesh.

We perform local-space convex decomposition using CoACD. Each resulting convex component is stored as a separate mesh and attached as a separate MuJoCo `geom type="mesh"` to the same object body. Thus, an object is represented by multiple convex collision components rather than being replaced by a single global convex hull.

Unless overridden by an experiment-specific configuration, CoACD uses threshold 0.08, at most 16 convex hulls, resolution 1000, merging enabled, and decimation disabled. Existing decomposition caches are reused when available.

Driver objects follow the prescribed generated trajectories, static objects remain fixed, and responder objects are instantiated as dynamic rigid bodies.

D.3 Prescribed Driver Replay

Driver trajectories are kinematically prescribed throughout physics replay. Depending on the simulation backend, this is implemented either with a MuJoCo mocap body or with a free-joint body whose state is overwritten at every simulation substep. In the latter case, the driver pose together with its linear and angular velocity is written directly to the corresponding state variables. In both cases, contacts affect responder objects without changing the prescribed driver trajectory.

The generated motions are defined at 30 fps. Each interval between two consecutive motion frames is divided into 96 MuJoCo substeps, giving

$$\Delta t_{\text{sim}} = \frac{1}{30 \times 96} = \frac{1}{2880}\,\text{s}. \tag{50}$$

With the default replay mode, driver translation is linearly interpolated between consecutive generated frames and driver orientation is interpolated using quaternion SLERP. The interpolated pose is supplied at every substep before calling the MuJoCo simulation step. Responder states are recorded at the end of each 30-fps frame interval. Frame zero is stored as the initialized state without an additional physics step.

D.4 Human Collision Proxies

The object-only simulation backend does not instantiate a human collision model. For human-aware simulation, the generated human trajectory drives collision proxies constructed from the 24-joint skeleton. Capsules are attached along parent–child body segments, and hand spheres are included by default.

The human proxies follow the prescribed human trajectory and are not dynamically modified by MuJoCo. Human–responder collision is selected when the simulation scene is constructed. When this collision mode is disabled, the proxy bodies may remain present for visualization and trajectory updates, but their collision masks are set to zero. Human collision is therefore fixed for a rollout rather than being activated or deactivated at individual contact events.

D.5 Simulation Parameters

Responder collision geometries use a uniform density of 450 unless an experiment-specific configuration overrides this value. We do not assign measured per-object masses or explicit per-object inertia tensors. Instead, the collision geometry and uniform density are provided to MuJoCo, which constructs the corresponding rigid-body properties during model compilation.

The default friction coefficients are

$$\boldsymbol{\mu} = (2.0,\ 0.1,\ 0.02), \tag{51}$$

corresponding to sliding, torsional, and rolling friction. The default contact configuration uses

$$\mathrm{condim} = 6, \qquad \mathrm{margin} = 0.001, \tag{52}$$

together with

$$\mathrm{solref} = (0.012, 1.0), \qquad \mathrm{solimp} = (0.97, 0.995, 0.001). \tag{53}$$

The human-aware free-joint kinematic driver backend uses a separate driver density of 50000 for its prescribed driver body. This value does not apply to responder objects.

E Outcome-Conditioned Editing Details

E.1 Outcome Objectives and Event Selection

Given a desired responder outcome, we first evaluate the discrepancy between the current simulated responder motion and the requested goal. The editor then identifies contact events that can causally influence this error. When an appropriate physical contact is absent, nearby geometric interactions are considered as candidate edit locations.

E.2 Local Driver Solver

For a selected event, the responder outcome error is converted into a desired change in responder velocity. The corresponding impulse is constrained by the local contact normal and friction cone and converted into a target change in driver contact-point velocity.

Let θ denote the local translational and rotational driver controls. We linearize their effect on the contact-point position and velocity as

$$\Delta\mathbf{p}_c \simeq J_p\theta, \qquad \Delta\mathbf{v}_c \simeq J_v\theta. \tag{54}$$

The local correction is obtained from

$$\theta^* = \arg\min_{\theta} \left\| \begin{bmatrix} J_p \\ J_v \end{bmatrix} \theta - \begin{bmatrix} 0 \\ \Delta\mathbf{v}_c \end{bmatrix} \right\|_2^2 + \|\Lambda\theta\|_2^2. \tag{55}$$

The first block approximately preserves the original contact position, whereas the second encourages the requested contact velocity. Λ penalizes large trajectory modifications.

For support interactions, we additionally consider a correction that aligns the supporting surface normal against gravity.

E.3 Candidate Evaluation and Iteration

For each local correction, we evaluate several scaled candidate edits through full MuJoCo replay. A candidate is accepted only when the resulting responder behavior improves the target objective. After an accepted edit, contact events are recomputed on the new rollout and the procedure may be repeated.

Editing terminates when the requested outcome is reached, when no candidate improves the objective, or when the iteration budget is exhausted.

E.4 Human Retargeting

After a driver edit is accepted, the human can be retargeted to remain consistent with the modified driver trajectory. We preserve the wrist motion relative to the driver and optimize smooth arm pose corrections while holding the edited driver fixed. Root motion, torso motion, lower-body motion, and finger articulation are retained from the original generated human trajectory.

A final physics replay using the edited driver and retargeted human produces the final edited interaction.