SURFACE-SYNC HUMAN-OBJECT INTERACTION MOTION GENERATION VIA CONTACT MODELING AND JOINT HAND-BODY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human–object interaction (HOI) motion generation remains challenging because realistic interactions require modeling not only human and object motions, but also their evolving physical relationship at the interaction interface. Existing approaches mainly represent interactions through joints, trajectories, or object poses, leaving surface correspondence, contact quality, and mechanical stability implicit. In this paper, we propose Surface-Sync, a unified framework for HOI motion generation that explicitly models interaction from representation and physical reasoning perspectives. We introduce a surface-centric interaction representation based on sparse surface markers, which captures dynamic human–object surface correspondences through contact activation and canonical object-surface anchors. This representation provides a unified basis for incorporating geometric compatibility and wrench-space physical priors, enabling the generation of physically stable and coherent interactions. Furthermore, we propose an interaction-driven whole-body coordination mechanism that propagates local interaction constraints from end-effectors to the surrounding body structure, improving the consistency between manipulation and global motion. Extensive experiments demonstrate that our approach generates more realistic, physically plausible, and coordinated human–object interaction motions compared with existing methods.

1 INTRODUCTION

The synthesis of realistic human motion has achieved remarkable progress with the development of deep generative models. However, generating realistic human-object interactions (HOI) remains a challenging problem. Unlike isolated human motion, where the plausibility of a sequence can often be evaluated from body trajectories alone, human-object interaction is fundamentally governed by the relationship between two continuously interacting surfaces. Whether grasping an object, pushing a tool, or supporting a load, successful interaction requires the human and object surfaces to maintain appropriate spatial alignment, contact consistency, and physical compatibility throughout the motion.

Despite recent advances, existing HOI generation methods mainly focus on modeling human and object motions using joints, sparse points, or mesh vertices. These representations are effective for describing motion trajectories, but they fail to explicitly capture the interaction interface where the actual manipulation occurs. As a result, generated motions may achieve visually plausible poses while still exhibiting unrealistic interaction patterns, such as floating contacts, loose grasps, and surface penetration as shown in Fig.1. This limitation reveals a fundamental gap in current HOI modeling: interaction is not merely a relation between motion trajectories, but a structured relationship between continuously evolving human and object surfaces.

First, existing joint-based representations are insufficient to describe fine-grained surface interaction. A valid interaction is not determined by whether a human point is close to an object point, but by how local surface regions align and evolve over time. For example, a stable grasp requires the hand surface to conform to the object surface, rather than simply reaching a nearby position as shown in Fig.1(a). Motivated by this observation, we introduce a surface-centric interaction representation, which explicitly models the evolving relationship between human and object surfaces.

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.



(a) (b) (c)

Figure 1: (a)When grasping an object, the hand and the object should form a continuous and coherent contact surface like the left half instead of the right half. (b)To ensure grasp stability, the grasping requires force closure. The right half is a simple counterexample, where the contact forces do not satisfy it. (c)During human motion, whole-body coordination is required to support hand interactions; the person shown on the right half excessively stretches the arm to move the object to a farther location, whereas the left side shows a more natural motion with the body leaning forward.

Specifically, we represent both human and object geometry using sparse surface markers that preserve local geometric structures. For each interaction-capable human marker, we model its contact state and corresponding location on the object surface in the object's canonical coordinate system. This formulation transforms implicit geometric proximity into an explicit interaction structure, describing not only whether interaction occurs, but also where and how the interaction evolves on the object surface. Furthermore, the object-centric formulation removes the ambiguity caused by global object motion and enables consistent modeling of interaction patterns across different poses and sequences.

Second, surface correspondence alone is insufficient, since realistic interaction requires not only contact existence but also contact quality and physical compatibility. Motivated by this, we propose physics-aware interaction optimization. Based on the proposed surface-centric interaction representation, local geometric and dynamic properties can be directly derived from the interaction relationship. For example, surface alignment characterizes whether interacting regions are geometrically compatible, while relative motion describes the temporal consistency of the contact. Moreover, the local contact geometry and friction constraints provide the basis for characterizing the mechanical capability of the interaction through wrench space as shown in Fig.1(b). Therefore, our representation naturally supports physics-aware optimization including contact coherence and force closure, enabling the model to generate interactions that are not only geometrically close but also physically stable.

Third, realistic HOI generation requires coordination beyond local interaction regions. Existing methods often model hand-object manipulation and whole-body motion separately, ignoring the fact that interaction and body movement are strongly coupled. In natural behaviors, hand-object contact constrains body configuration, while body posture provides necessary support for maintaining stable interaction as shown in Fig.1(c). For example, handing an object to the distance requires coordinated adjustments of the wrist, arm, shoulder, and torso rather than isolated hand movement. To address this issue, we propose an interaction-driven whole-body coordination mechanism. By treating the surface-level interaction as the anchor of the kinematic chain, local contact requirements are propagated through the human structure to guide global body configuration, allowing fine-grained manipulation and whole-body motion to be jointly coordinated.

Our contributions are summarized as follows:

- We introduce a surface-centric interaction representation that explicitly captures evolving human-object surface relationships in an object-centric manner, providing a unified foundation for modeling geometric and physical interaction constraints.
- Based on the proposed representation, we incorporate contact quality and physical compatibility modeling, including geometric coherence and mechanical stability, enabling the generation of physically plausible interactions.
- We propose an interaction-driven whole-body coordination mechanism that propagates local interaction constraints through the human kinematic structure, producing globally coordinated motion.
- Extensive experiments demonstrate that our method generates more realistic, stable, and coordinated human-object interactions compared with existing approaches.

2 Related Works

2.1 Hand-Object Interaction Generation

The synthesis of realistic hand-object interactions has evolved from static grasp generation toward dynamic and semantically conditioned interaction synthesis. Early works such as GrabNet(Taheri et al., 2020) and ContactGen(Liu et al., 2023) focus on generating object-conforming hand configurations, while contact-aware methods (Liu et al., 2023; Cha et al., 2024) exploit object-surface contact information to improve geometric consistency. With the introduction of large-scale datasets such as GRAB(Taheri et al., 2020), subsequent studies have advanced toward dynamic interaction generation. Text2HOI(Cha et al., 2024) leverages diffusion models to synthesize temporally coherent hand-object manipulation sequences from text, and recent works further improve interaction modeling through contact guidance, interaction dynamics, and specialized architectures (Wu et al., 2026; Xu et al., 2026).

Despite these advances, existing hand-object interaction methods mainly model the hand and object states, while the evolving physical relationship between their surfaces remains implicit or is introduced only through contact-related conditioning and auxiliary objectives. Such formulations describe where contact may occur, but do not explicitly represent the temporal evolution of individual contacts or whether their spatial arrangement forms a physically valid configuration with sufficient force and torque support. Consequently, contact correspondence, temporal contact evolution, and mechanical properties of the interaction are rarely unified within a single representation.

In contrast, our method treats the interaction interface itself as an explicit generative state. Instead of using contact only as a spatial prior or refinement signal, we model the evolving correspondence between human and object surfaces throughout the interaction. These correspondences further enable the derivation of local geometric properties and admissible wrench spaces, allowing geometric consistency and mechanical stability to be jointly modeled.

2.2 Whole-Body and Object Interaction

The study of human motion generation has evolved from general-purpose motion synthesis (Tevet et al., 2023; Zhang et al., 2022a; Jin et al., 2023; Zhang et al., 2023a; Chen et al., 2023; Jiang et al., 2023a; Tevet et al., 2022; Wang et al., 2023; Pinyoanuntapong et al., 2024; Athanasiou et al., 2022; 2023; Shafir et al., 2023) toward human interactions with objects and environments (Ye et al., 2023; Tian et al., 2026; Ma et al., 2025; Liu et al., 2023; Li et al., 2023b; Christen et al., 2024; Wang et al., 2022; Petrov et al., 2023; Kim et al., 2023; Hou et al., 2023). Compared with isolated motion synthesis, generating coordinated whole-body motion during object manipulation introduces an additional challenge, as local hand-object interaction must remain consistent with global posture and body dynamics. Early scene-level methods such as SAMP(Hassan et al., 2021) and PLACE(Zhang et al., 2020), as well as subsequent works (Hassan et al., 2021; Zhang et al., 2020; Jiang et al., 2023b; Kulkarni et al., 2023; Pi et al., 2023; Zhang et al., 2022b; Starke et al., 2019), mainly focus on human interaction with static environments. Object-centric approaches such as GOAL(Taheri et al., 2022) and OMOMO(Li et al., 2023a) further introduce object-conditioned human motion generation, but assume predefined object trajectories. More recent methods, including InterDiff(Xu et al., 2023) and HOI-Diff(Peng et al., 2023), jointly synthesize human and object motions, while subsequent works improve interaction modeling through kinematic structures, interaction dynamics, and physical constraints (Zeng et al., 2025; Wu et al., 2026; Wang et al., 2025).

Despite these advances, existing approaches typically treat human motion and object motion as the primary state variables and model their interaction through feature exchange, kinematic structures, or auxiliary contact/physical supervision. Consequently, interaction is mainly represented at the pose, joint, trajectory, or dynamics level, while the evolving correspondence between human and object surfaces and the mechanical properties of contact configurations remain implicit.

In contrast, we focus on representing the *interaction interface itself*. Building upon surface markers, we introduce a dynamic interaction state that explicitly describes the evolving correspondence between human and object surface regions. Surface markers have previously been explored for human motion representation, e.g., MOJO(Zhang et al., 2021b) and InterAct(Xu et al., 2025), but mainly serve to describe body states and geometry rather than interaction relations. Our method instead leverages the surface domain to model dynamic human-object correspondences and derive their

geometric and wrench-space properties, providing a unified representation for surface compatibility, contact coherence, and force closure. Furthermore, an end-effector-driven structural constraint propagates interaction-induced adjustments from the hands to the whole body.

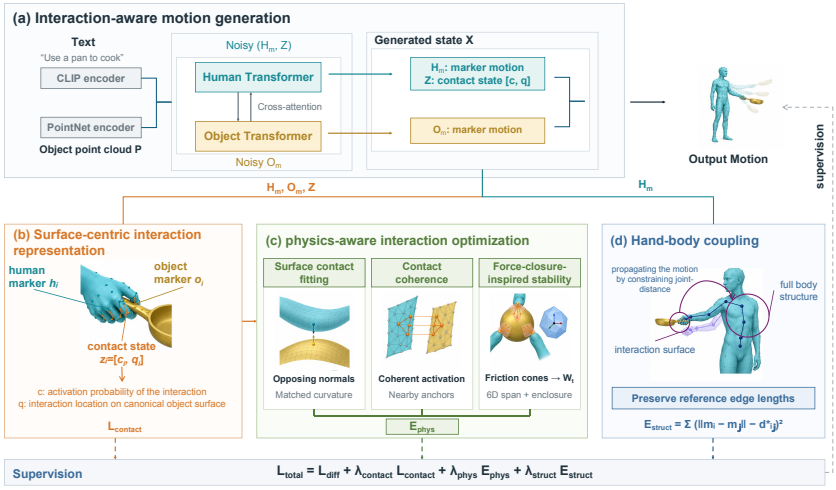

Figure 2: Overview of Surface-Sync. (a) The interaction-aware motion generation model synchronously generates human and object motions. (b) The surface-centric interaction representation augments the synchronized motion with explicit contact activation and canonical object anchors. (c) Physics-aware interaction optimization imposes surface fitting, local contact coherence, and force-closure stability. (d) Joint hand-body optimization propagates interaction-induced adjustments from the end-effectors to the arms and torso through structural constraints.

3 METHOD

3.1 BASELINE REVISITING

We build upon a conditional dual-branch diffusion model following HOI-Diff (Peng et al., 2023) as our generation backbone. The model adopts two branches to separately denoise human and object motions, while exchanging information through cross-attention to capture their mutual dependency. Given an object geometry condition $\mathbf{P}$ and a textual description $\mathbf{s}$, the baseline aims to generate synchronized human and object motions: $(\hat{\mathbf{H}}_p, \hat{\mathbf{O}}_p) = \mathcal{F}(\mathbf{P}, \mathbf{s})$, where $\mathbf{H}_p$ denotes the human motion represented by joint trajectories, and $\mathbf{O}_p$ denotes the object motion represented by rigid transformations. Although this formulation enables synchronization between human and object, the interaction remains implicit. Specifically, joint trajectories and object poses do not explicitly describe which surface regions participate in interaction, how contacts evolve during manipulation, or whether the resulting interaction satisfies physical constraints. Our method retains this diffusion backbone while replacing the implicit interaction modeling with an explicit surface-centric formulation.

3.2 OVERVIEW

Realistic human-object interaction generation requires modeling not only individual human and object motions, but also the physical relationship established between their surfaces. For example, when a person picks up a cup, the hand motion, the contacted region on the cup, and the resulting body configuration are mutually dependent. A realistic interaction therefore requires explicit modeling of contact formation, physical validity, and whole-body coordination.

To address these challenges, we propose Surface-Sync, which consists of four stages as shown in Fig.2. We use an interaction-aware generation model for motion generation(Fig.2(a) and Sec.3.4). First, we introduce a surface-centric interaction representation based on sparse surface markers(Fig.2(b) and Sec.3.3), which explicitly models the geometric entities involved in interaction. Second, we derive physical objectives from the predicted contacts to encourage geometrical fitness and mechanically stable interactions(Fig.2(c) and Sec.3.5). Finally, we propagate local interaction constraints to the surrounding body to maintain whole-body consistency(Fig.2(d) and Sec.3.6).

Given text and object geometry conditions, our model jointly generates $\hat{\mathbf{X}} = (\hat{\mathbf{H}}_m, \hat{\mathbf{O}}_m, \hat{\mathbf{Z}})$, where $\mathbf{H}_m$ and $\mathbf{O}_m$ denote human and object surface motions, respectively, and $\mathbf{Z}$ represents the dynamic contact state. The surface motions provide geometric information, while the contact state serves as the bridge between motion generation and physical reasoning.

3.3 Surface-Centric Interaction Representation

Existing HOI representations mainly rely on human joints and object poses. However, physical interaction occurs at the surface level: the feasibility of a grasp or support depends on local surface geometry rather than only the relative pose of skeleton joints. Therefore, a joint-based representation introduces a mismatch between the representation space and the actual interaction space.

To address this issue, we introduce a surface-centric representation that models both human and object motions using surface markers. The representation explicitly describes the geometric entities involved in interaction and provides a unified space for modeling dynamic correspondences and physical constraints.

Human/Object Surface Parameterization. For each frame t, the human body is represented by N_h surface markers sampled from the body mesh: $\mathbf{H}_{m,t} \in \mathbb{R}^{N_h \times 6}$, where each marker contains its 3D position and velocity. Similarly, the object surface is represented by N_o markers sampled from the canonical object surface: $\mathbf{O}_{m,t} \in \mathbb{R}^{N_o \times 6}$. The complete human and object surface sequences are denoted as $\mathbf{H}_m$ and $\mathbf{O}_m$. The number of sampled markers is determined in practice by balancing computational efficiency and modeling accuracy, providing our method with greater flexibility in adjustment. Compared with joint-based representations, surface markers directly describe the regions involved in physical interaction and provide the geometric basis for modeling contact relationships.

Dynamic Surface Correspondence Modeling. Surface motions specify where human and object regions are located, but they do not indicate whether two regions actually interact. Therefore, we introduce an explicit dynamic surface correspondence modeling to describe evolving human-object correspondences. For each interaction-capable human marker i, we define its contact state as $\mathbf{z}_{i,t} = [c_{i,t}, \mathbf{q}_{i,t}]$, where $c_{i,t} \in [0, 1]$ denotes the contact activation probability and $\mathbf{q}_{i,t}$ denotes the corresponding location on the canonical object surface. The contact state is formulated in continuous surface coordinates instead of discrete object indices. By representing $\mathbf{q}_{i,t}$ in the object's canonical space, our representation explicitly captures where the interaction occurs relative to the object geometry, which cannot be directly inferred from global positions that vary with object poses. This design allows the interaction region to move continuously over the object surface. For example, when a hand slides along a cup or a tool, the model can represent the changing contact location while preserving its semantic correspondence on the object surface rather than assuming a fixed correspondence. The complete interaction representation is $\mathbf{Z}_t = [\mathbf{z}_{i,t}]_{i=1}^{N_c}$, where N_c denotes the number of interaction-capable human markers. As interactions usually occur on human hands, we select these markers from human hands and N_c is smaller than N_h. The predicted contact state provides explicit surface correspondences between human and object regions, which are subsequently used to evaluate geometric and physical constraints.

3.4 Interaction-Aware Motion Generation

The interaction representation defines how human and object surfaces interact, but these relationships should be generated consistently with the motion itself. Therefore, we incorporate the contact state into the diffusion process and jointly predict motion and interaction evolution.

We augment the human branch state with the interaction representation: $\mathbf{X}_H^{(0)} = (\mathbf{H}_m, \mathbf{Z})$, while the object branch generates object surface motions: $\mathbf{X}_O^{(0)} = \mathbf{O}_m$. During diffusion denoising, the human and object branches exchange information through bidirectional cross-attention: $(\hat{\mathbf{H}}_m, \hat{\mathbf{Z}}) = \mathcal{M}_H(\mathbf{X}_H^{(\ell)}, \mathbf{X}_O^{(\ell)}, \ell, \mathbf{z}_{\text{text}}, \mathbf{z}_{\text{obj}})$, $\hat{\mathbf{O}}_m = \mathcal{M}_O(\mathbf{X}_O^{(\ell)}, \mathbf{X}_H^{(\ell)}, \ell, \mathbf{z}_{\text{text}}, \mathbf{z}_{\text{obj}})$. Here, $\mathbf{z}_{\text{text}}$ and $\mathbf{z}_{\text{obj}}$ denote text and object geometry conditions.

By jointly predicting contact and motion, the model learns contact evolution as an intrinsic part of the generation process rather than estimating contacts after motion generation. This formulation ensures that generated trajectories and interaction states remain mutually consistent.

3.5 PHYSICS-AWARE INTERACTION OPTIMIZATION

The predicted contact states provide explicit interaction correspondences, but physically plausible interaction requires more than contact existence. For example, when a hand grasps an object, the contacting surfaces should exhibit compatible orientations and sufficient mechanical stability; otherwise, the generated interaction may appear plausible but violate physical constraints. Therefore, we introduce physics-aware objectives from three perspectives: surface fitness, contact coherence, and force-closure-inspired stability.

Surface Contact Fitting. A valid contact requires fitting and compatible local geometry between interacting surfaces. For example, when a palm presses against a table, the palm and table surfaces should exhibit consistent orientations; similarly, fingers wrapping around a curved object should match the local shape of the object surface.

For each predicted contact, we evaluate the geometric relationship between the corresponding human and object surface regions. Specifically, we consider the fitness of surface orientations and local geometric structures. Let $\mathbf{n}^h_{i,t}$ denote the normal of human surface marker i, and let $\mathbf{n}^o(\mathbf{q}_{i,t})$ denote the object surface normal evaluated at the predicted canonical correspondence $\mathbf{q}_{i,t}$. The surface normal of each marker is estimated by fitting a local plane to the neighboring markers on the corresponding mesh. Since two contacting surfaces should face each other, we define the normal fitness term as

$$E_{\text{normal}} = \sum_{t,i} c_{i,t} \left| (\mathbf{n}^h_{i,t})^\top \mathbf{n}^o(\mathbf{q}_{i,t}) + 1 \right|^2, \tag{1}$$

which ensures the energy is zero when the normals are antiparallel. We further measure local geometric fitness using surface curvature. Let $\kappa^h_{i,t}$ and $\kappa^o(\mathbf{q}_{i,t})$ denote the local curvature of the human and object surfaces, respectively. The curvature values are computed from the local shape of the fitting corresponding mesh surfaces. The curvature consistency term is defined as

$$E_{\text{geom}} = \sum_{t,i} c_{i,t} \left| \kappa^h_{i,t} - \kappa^o(\mathbf{q}_{i,t}) \right|^2. \tag{2}$$

The overall surface fitness objective is $E_{\text{surf}} = E_{\text{normal}} + \lambda_\kappa E_{\text{geom}}$.

Contact Coherence. Physical interactions usually involve continuous contact regions rather than isolated contact points. Therefore, neighboring contact regions should evolve consistently in both activation states and spatial correspondences. We define a local consistency objective over neighboring human surface markers. Let $\mathcal{N}_H$ denote the local neighborhood pairs on the human surface, where the neighborhood is constructed according to the spatial adjacency of surface markers. The coherence objective is

$$E_{\text{coh}} = \sum_t \sum_{(i,j)\in\mathcal{N}_H} \left[(c_{i,t} - c_{j,t})^2 + c_{i,t} c_{j,t} d_{\mathcal{S}}(\mathbf{q}_{i,t}, \mathbf{q}_{j,t}) \right], \tag{3}$$

where $d_{\mathcal{S}}(\cdot,\cdot)$ denotes the geodesic distance on the object surface, which is computed along the canonical space of the object mesh represented by $\mathbf{q}$. The first term encourages neighboring human regions to have consistent contact activation $c \in [0,1]$, while the second term prevents discontinuous object correspondences when neighboring regions are simultaneously in contact.

Force-Closure-Inspired Stability. Geometric fitness alone does not guarantee stable interaction. For example, a grasp requires multiple contacts to jointly resist external disturbances.

For each contact region, we consider feasible contact forces sampled from the local friction cone determined by the surface normal. The friction cone describes the set of physically feasible force directions under the local contact geometry. Each contact force produces both translational force and rotational torque, forming a 6D wrench: $\mathbf{w}^r_{i,t} = \begin{bmatrix} \mathbf{f}^r_{i,t} \\ \mathbf{r}_{i,t} \times \mathbf{f}^r_{i,t} \end{bmatrix}$, where $\mathbf{f}^r_{i,t}$ denotes a sampled feasible contact force and $\mathbf{r}_{i,t}$ denotes the contact position relative to the object center. We collect all contact wrenches into a matrix: $\mathbf{W}_t = [c_{i,t}\mathbf{w}^r_{i,t}]_{i,r}$. To encourage the generated contacts to provide sufficient control authority, we maximize the volume spanned by the wrench basis:

$$E_{\text{span}} = -\frac{1}{T} \sum_t \log\det(\mathbf{W}_t \mathbf{W}_t^\top + \delta \mathbf{I}). \tag{4}$$

Furthermore, force closure requires the origin to lie inside the convex hull of feasible wrenches. We approximate this property with an enclosure objective:

$$E_{\text{enc}} = \frac{1}{T} \sum_{t} \text{softplus}(\epsilon - s(\mathbf{W}_t)), \tag{5}$$

where $s(\mathbf{W}_t)$ measures the weakest support direction of the wrench set. The final force-closure-inspired stability objective is $E_{\text{fc}} = E_{\text{span}} + \lambda_{\text{enc}} E_{\text{enc}}$. The overall physics objective is $E_{\text{phys}} = \lambda_{\text{surf}} E_{\text{surf}} + \lambda_{\text{coh}} E_{\text{coh}} + \lambda_{\text{fc}} E_{\text{fc}}$. These objectives transform the interaction representation from a geometric correspondence into a physically meaningful interaction configuration.

3.6 Whole-Body Interaction Coordination

Physical constraints mainly affect local contact regions, especially the hands. However, human motion is globally coordinated: modifying hand-object interaction without considering surrounding body parts may result in unnatural arm or torso configurations.

Therefore, we introduce a structural coordination objective that propagates local physics-aware interaction requirements from contact regions to the whole body, enabling the global pose configuration to adapt coherently to the constraints imposed by hand-object interactions. We connect interaction-related body markers with neighboring body regions through structural constraints:

$$E_{\text{struct}} = \sum_{t} \sum_{(i,j)} (\|\mathbf{m}_{i,t} - \mathbf{m}_{j,t}\|_2 - d_{ij}^*)^2, \tag{6}$$

where (i, j) denotes a pair of structurally connected body markers, $\mathbf{m}_{i,t}$ denotes the position of body marker i, and d_{ij}^* denotes the reference structural distance computed from the original motion. This objective allows local interaction requirements to influence surrounding body configurations while preserving human motion structure. As a result, the generated motion achieves both physically plausible contacts and globally coordinated body movements.

3.7 Training Objectives

The overall training objective combines diffusion reconstruction, contact supervision, physical constraints, and whole-body coordination: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{contact}} \mathcal{L}_{\text{contact}} + \lambda_{\text{phys}} E_{\text{phys}} + \lambda_{\text{struct}} E_{\text{struct}}$. The diffusion objective reconstructs human and object surface motions together with the dynamic contact state. The contact objective supervises contact activation and surface correspondence. The physics objective enforces geometric fitness and mechanical stability, while the structural objective maintains whole-body consistency. By jointly optimizing these objectives, Surface-Sync establishes a unified framework that connects surface representation, interaction generation, physical reasoning, and whole-body coordination.

4 Experiments

4.1 Settings

Datasets and Methods. We evaluate our method on InterAct(Xu et al., 2025) dataset, which is a large-scale 3D human-object interaction dataset containing approximately 30.7 hours of motion data and 34.1K sequence-level textual descriptions. We conduct crossdataset experiments on NeuralDome(Zhang et al., 2023b), BEHAVE(Bhatnagar et al., 2022), and IMHD^2(Zhao et al., 2024), which are parts of InterAct. We compare our method with several baselines, which cover general motion diffusion, interaction-aware generation, kinematic-chain modeling, and modality-specific denoising guidance.

These datasets cover a broad range of motion durations, object categories, and interaction complexities, enabling comprehensive evaluation of physically plausible interaction generation across different capture modalities and scales. To unify the motion representation and the textual annotation, we use the processed version by InterAct(Xu et al., 2025) for a fair comparison.

Metrics. We follow HOI-Diff(Peng et al., 2023) to employ FID / R / Diversity(Div.) / Contact Distance(CDis) as the metrics for a fair comparison. To better evaluate the correctness of interaction at the contact interface, we employ 3 additional metrics to quantify the physical plausibility and contact quality of the synthesized motions:**1) Penetration(pene)** measures human–object interpenetration by computing the signed distance from human surface vertices to the object surface. We report the maximum penetration depth averaged over the sequence. **2) Contact Ratio(CR)** is defined as

Table 1: Quantitative results on InterAct. ↑/↓ indicate preferred directions, and → denotes comparison with the reference distribution.

Method	FID↓	R@3↑	Div.→	CDis↓	Pene.↓	CR↑	Realism↑
Ground Truth	0.000	0.909	7.781	0.232	0.076	0.092	0.86
HOI-Diff (Peng et al., 2023)	0.689	0.740	7.620	0.337	0.109	0.041	0.48
CHOIS (Li et al., 2024)	0.572	0.766	**7.717**	0.325	0.131	0.037	0.46
InterDiff (Xu et al., 2023)	0.215	0.824	**7.717**	0.329	0.116	0.038	0.59
Text2HOI (Cha et al., 2024)	0.331	0.745	7.631	0.342	**0.105**	0.042	0.62
LIGHT w/o guidance (Wang et al., 2026)	0.196	0.715	7.708	0.321	0.121	0.042	0.69
LIGHT w/ guidance (Wang et al., 2026)	0.148	0.754	7.712	0.304	0.132	0.046	0.72
Ours	**0.107**	**0.838**	7.703	**0.274**	0.106	**0.051**	**0.76**

the average ratio of human surface vertices whose distances to the object surface fall below a fixed threshold, reflecting the extent and consistency of hand–object contact. **3) Realism**(Zhang et al., 2021a; Cha et al., 2024) measures the proportion of frames in which the object motion is consistent with the hand–object contact configuration, evaluating whether object movement can be reasonably explained by the implied contact forces.

Table 2: Cross-dataset evaluation. Models trained on one dataset are evaluated on each of the other two. Bold denotes the best reported value for each transfer direction and metric. Lower FID and penetration and higher R@3, CR, and Realism are preferred.

Train	Test	Method	FID↓	R@3↑	Pene.↓	CR↑	Realism↑
NeuralDome	BEHAVE	InterDiff	15.63	0.16	5.01	0.019	0.25
		HOI-Diff	6.87	0.19	4.92	0.017	0.33
		Ours	**6.32**	**0.28**	**4.77**	**0.035**	**0.50**
	IMHD^2	InterDiff	18.12	0.19	7.16	0.011	0.19
		HOI-Diff	16.24	0.26	7.02	0.018	0.22
		Ours	**14.53**	**0.56**	**5.25**	**0.031**	**0.31**
BEHAVE	NeuralDome	InterDiff	17.52	0.22	8.10	0.009	0.18
		HOI-Diff	17.79	0.28	7.92	0.012	0.21
		Ours	**16.24**	**0.57**	**7.23**	**0.017**	**0.26**
	IMHD^2	InterDiff	17.43	0.21	6.82	0.014	0.23
		HOI-Diff	18.10	0.27	6.23	0.017	0.25
		Ours	**16.79**	**0.54**	**5.28**	**0.028**	**0.32**
IMHD^2	NeuralDome	InterDiff	18.13	0.14	8.24	0.010	0.20
		HOI-Diff	17.02	0.19	7.87	0.010	0.21
		Ours	**15.39**	**0.41**	**7.15**	**0.019**	**0.25**
	BEHAVE	InterDiff	16.73	0.15	5.23	0.015	0.23
		HOI-Diff	7.27	0.23	5.48	0.021	0.23
		Ours	**6.81**	**0.29**	**4.62**	**0.031**	**0.41**

4.2 Quantitative Results

As shown in Tab. 1, our method achieves strong overall performance across both motion quality and interaction-related metrics. **1)** Our method achieves the best FID and R@3, indicating improved motion quality and stronger semantic alignment with interaction intents. Meanwhile, the comparable diversity score demonstrates that the introduced interaction constraints improve generation quality without significantly reducing motion variability. **2)** For interaction-specific metrics, our method achieves the lowest CDis and the highest CR while maintaining a competitive penetration score, suggesting that the generated hand–object contacts are more accurate and physically plausible rather than merely increasing contact through excessive interpenetration. This improvement is attributed to the explicit modeling of surface correspondence and physical constraints, which encourages coherent and stable interaction patterns. **3)** Our method obtains the highest realism score, indicating that the generated object motions are better aligned with the corresponding hand contacts. These results demonstrate that explicitly modeling interaction geometry and physical compatibility leads to more realistic and consistent human–object interactions.

We also conduct a cross-dataset experiment in Tab. 2, with a particular focus on the generalization of surface-level interaction modeling. As the results show, our approach consistently outperforms the evaluated baselines when trained on one dataset and tested on the others. These results suggest that our framework preserves its relative advantages across datasets with different motion styles and object configurations. We attribute this robustness in part to the emphasis on contact geometry and hand–body coordination, which provide useful constraints across interaction scenarios. Overall, the cross-dataset results support the transferability of our surface-based interaction framework.

Table 3: Ablation of the major components.

Surface Rep.	Physics Opt.	Body Coord.	FID↓	R@3↑	CDis↓	Realism↑
×	×	×	0.689	0.740	0.337	0.48
✓	×	×	0.342	0.737	0.299	0.63
✓	✓	×	0.137	0.801	0.275	0.72
✓	✓	✓	0.107	0.838	0.274	0.76

Table 4: Ablation of the interaction representation design.

Variant	FID↓	R@3↑	CDis↓	Realism↑
Geometry-only Contact	0.268	0.725	0.281	0.71
Contact Activation Only	0.195	0.761	0.279	0.72
Full Ours	0.107	0.838	0.274	0.76

Table 5: Ablation of the physics-aware objectives.

E_{surf}	E_{coh}	E_{fc}	FID↓	R@3↑	CDis↓	Realism↑
×	✓	✓	0.154	0.805	0.313	0.65
✓	×	✓	0.113	0.822	0.325	0.68
✓	✓	×	0.142	0.810	0.284	0.62
✓	✓	✓	0.107	0.838	0.274	0.76

4.3 Ablation Studies

Effect of Major Components. We first evaluate the contribution of the major components of Surface-Sync, including the surface-centric interaction representation, physics-aware optimization, and whole-body coordination. As shown in Table 3, each component brings consistent improvements, demonstrating that explicit interaction modeling, physical constraints, and structural coordination provide complementary benefits for HOI motion generation.

Effect of Surface-Centric Interaction Representation. We further analyze the design of the proposed interaction representation. As shown in Table 4, explicitly modeling interaction states improves generation quality over geometry-only contact estimation, while introducing dynamic surface correspondence further enhances interaction consistency. These results verify the importance of representing both contact occurrence and interaction location for HOI generation.

Effect of Physics-Aware Interaction Optimization. We investigate the contribution of different terms in the physics-aware objective. As shown in Table 5, removing any physical constraint leads to degraded performance, indicating that surface fitting, contact coherence, and force-closure-inspired stability capture complementary aspects of realistic interaction. Combining all terms achieves the best overall results.



Figure 3: The left half shows a qualitative comparison. The text condition is *A person picks up a bucket from the ground and swings it*. The right half gives more visualizations of our method

4.4 Qualitative Results

Here we present qualitative visualizations. More results in video format can be found in the supplementary materials. The left half of fig. 3 visualizes text-conditioned generation results for the prompt *A person picks up a bucket from the ground and swings it*. HOI-Diff fails to establish a meaningful interaction: the bucket starts moving even when the hand is far from its surface, indicating a lack of explicit contact modeling. Removing physical priors from our method improves spatial alignment but still results in unstable and intermittent contact during manipulation. In contrast, the full model maintains consistent hand–bucket contact throughout the interaction, demonstrating that physical priors are essential for grounding text-driven motion in physically plausible human–object interactions. The right half of fig. 3 shows that our method generalizes across objects with diverse geometries, including pillows, buckets, and large furniture such as tables. It supports both single-hand and two-hand interactions, while consistently producing more coordinated full-body motions, such as bending to pick up a bucket or leaning forward to pull a table, highlighting its robustness and physical plausibility across varied HOI scenarios.

5 Conclusion

In this work, we proposed Surface-Sync, a surface-aware HOI generation framework that jointly optimizes full-body motion with contact geometry. By leveraging surface markers and physical priors, our method produces more stable and realistic interactions with reduced penetration and sliding. In addition, the proposed hand–body coordination mechanism improves kinematic consistency. Extensive experiments validate the effectiveness and generality of our approach.

AI USE STATEMENT

In this work, we used generative AI tools for [assist with translation]. We have not used generative AI tools for [propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, and interpret results], and [generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, and clean and reformat dataset] are not applicable to this work. Additionally, we used generative AI tools for [summarize or analyse existing literature]. We have reviewed all AI-assisted work. [LLM-helped translations have been checked by authors]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies human-object interaction motion generation using existing motion datasets and includes a user study to evaluate the perceived quality of generated motions. The user study was conducted with voluntary participants. Participants were informed about the purpose of the study before participation. No personally identifiable information was collected, and all responses were anonymized and used solely for research purposes. All experiments are conducted on publicly available datasets or datasets used under their corresponding licenses and agreements. We follow the original data usage policies and respect the privacy considerations associated with these resources. The proposed method aims to improve the physical plausibility and semantic consistency of generated human-object interactions for research purposes, such as motion synthesis, digital animation, and embodied intelligence. While generative models may potentially be misused to create misleading or inappropriate content, our work focuses on controllable 3D motion generation and does not target applications involving impersonation, surveillance, or other harmful uses. We acknowledge that motion generation models may inherit biases present in training data, including biases related to human appearance, behavior, and interaction patterns. Since our method operates on abstract motion and geometry representations rather than personal identities or demographic attributes, we do not intentionally model or infer sensitive human characteristics. Nevertheless, careful evaluation and responsible deployment are necessary when applying such models to real-world scenarios. We declare that there are no conflicts of interest or external sponsorships influencing the design, evaluation, or presentation of this research. All experimental results are reported following standard research practices, and implementation details are provided to facilitate reproducibility.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed Surface-Sync framework, including the surface-centric interaction representation, interaction-aware motion generation, physics-aware interaction optimization, and whole-body coordination mechanism in Sec. 3. The overall training objective and experimental protocols are described in Sec. 3.7 and Sec. 4, respectively. Additional experimental results and qualitative visualizations are provided in the appendix. We will release the source code and trained models upon publication to facilitate further research and independent verification.

REFERENCES

Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pp. 414–423. IEEE, 2022.

Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Sinc: Spatial composition of 3d human motions for simultaneous action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9984–9995, 2023.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *CVPR*, 2022.

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*, pp. 1577–1585, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18000–18010, June 2023.

Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proceedings of the International Conference on Computer Vision 2021*, October 2021.

Zhi Hou, Baosheng Yu, and Dacheng Tao. Compositional 3d human-object neural animation, 2023. URL https://arxiv.org/abs/2304.14070.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023a.

Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *ICCV*, 2023b.

Peng Jin, Yang Wu, Yanbo Fan, Zhongqian Sun, Yang Wei, and Li Yuan. Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. *arXiv preprint arXiv:2311.01015*, 2023.

Taeksoo Kim, Shunsuke Saito, and Hanbyul Joo. Ncho: Unsupervised learning for neural 3d composition of humans and objects, 2023. URL https://arxiv.org/abs/2305.14345.

Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis, 2023.

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 42(6), 2023a.

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

Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations, 2023b. URL https://arxiv.org/abs/2303.13129.

Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

Junyi Ma, Jingyi Xu, Xieyuanli Chen, and Hesheng Wang. Diff-ip2d: Diffusion-based hand-object interaction prediction on egocentric videos, 2025. URL https://arxiv.org/abs/2405.04370.

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. *arXiv preprint arXiv:2312.06553*, 2023.

Ilya A. Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone?, 2023. URL https://arxiv.org/abs/2306.00777.

Huaijin Pi, Sida Peng, Minghui Yang, Xiaowei Zhou, and Hujun Bao. Hierarchical generation of human-object interactions with diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15061–15073, October 2023.

Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: Bidirectional autoregressive motion model. In *Computer Vision – ECCV 2024*, 2024.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023.

Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6), November 2019. ISSN 0730-0301. doi: 10.1145/3355089.3356505. URL https://doi.org/10.1145/3355089.3356505.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. URL https://grab.is.tue.mpg.de.

Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://goal.is.tue.mpg.de.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pp. 358–374. Springer, 2022.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Jie Tian, Ran Ji, Lingxiao Yang, Suting Ni, Yuexin Ma, Lan Xu, Jingyi Yu, Ye Shi, and Jingya Wang. Gaze-guided hand-object interaction synthesis: Dataset and method, 2026. URL https://arxiv.org/abs/2403.16169.

Ruiyan Wang, Lin Zuo, Zonghao Lin, Qiang Wang, Zhengxue Cheng, Rong Xie, Jun Ling, and Li Song. Pa-hoi: A physics-aware human and object interaction dataset. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 12769–12775, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3758217. URL https://doi.org/10.1145/3746027.3758217.

Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors, 2022. URL https://arxiv.org/abs/2209.02485.

Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22035–22044, 2023.

Ziyin Wang, Sirui Xu, Chuan Guo, Bing Zhou, Jiangshan Gong, Jian Wang, Yu-Xiong Wang, and Liang-Yan Gui. Unleashing guidance without classifiers for human-object interaction animation. In *ICLR*, 2026.

Lin Wu, Zhixiang Chen, and Jianglin Lan. Hoi-dyn: Learning interaction dynamics for human-object motion diffusion. *Advances in Neural Information Processing Systems*, 38:90795–90825, 2026.

Hang Xu, Yang Xiao, Changlong Jiang, Haohong Kuang, Kaidi Zhang, Min Du, and Ran Wang. Moeg-hoi: mixture of expert groups for one-stage hand-object interaction motion generation with hand-finger-joint semantic guidance. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i13.38102. URL https://doi.org/10.1609/aaai.v40i13.38102.

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

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. InterAct: Advancing large-scale versatile 3d human-object interaction generation. In *CVPR*, 2025.

Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mellon, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22479–22489, 2023. doi: 10.1109/CVPR52729.2023.02153.

Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng, and Wei-Shi Zheng. Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12358–12369, 2025. doi: 10.1109/CVPR52734.2025.01153.

He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Trans. Graph.*, 40(4), July 2021a. ISSN 0730-0301. doi: 10.1145/3450626.3459830. URL https://doi.org/10.1145/3450626.3459830.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023a.

Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023b.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022a.

Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision (3DV)*, November 2020.

Xiaohan Zhang, Bharat Lal Bhatnagar, Vladimir Guzov, Sebastian Starke, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*. Springer, October 2022b.

Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3372–3382, 2021b.

Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'M HOI: Inertia-aware monocular capture of 3d human-object interactions. In *CVPR*, 2024.

A Appendix

Here we will present extra details about our method and experiments, and dive deeply into our approach. In Sec.A.1, we dive deeper and give more additional experiments. In Sec.A.2, we present more visualizations and qualitative comparisons. In Sec.A.3, we provide implementation details on the evaluation metrics we used. We provide more details about the marker representation implementation in Sec.A.4.

Table 6: Evaluation of whole-body motion quality and structural consistency. Structural Error is only reported for our surface-based representation, where the corresponding structural graph is explicitly defined.

Method	Body FID↓	Body Accel. Error↓	Body Jerk↓	Structural Error↓	Foot Sliding↓
InterDiff	1.09	1.92	67.4	–	0.09
HOI-Diff	1.13	1.77	93.2	–	0.12
Ours w/o Structural Constraint	1.07	1.63	45.2	13.47	0.07
Ours	**1.02**	**1.24**	**27.3**	**9.26**	**0.05**

Table 7: Analysis of the predicted interaction representation. Bold denotes the best result.

Method	Contact F1↑	Contact Temporal Jitter↓	Force-Closure Rate↑
GT	–	0.35	0.91
Geometry-only Contact	0.68	2.82	0.57
Contact Activation Only	0.72	2.57	0.59
Full Ours	**0.78**	**1.44**	**0.84**

A.1 Additional Experiments

A.1.1 Effect of Whole-Body Structural Coupling

Table 6 evaluates distributional body-motion quality, motion derivatives, structural consistency, and foot sliding. Even without the structural constraint, our interaction-aware formulation produces competitive whole-body motion quality and consistently improves motion smoothness and foot stability over the reported baselines. This suggests that explicitly modeling human-object interaction already provides useful guidance for coordinating whole-body motion. Adding the structural constraint consistently improves all evaluated aspects of whole-body motion, including distributional fidelity, acceleration and jerk, structural consistency, and foot stability. These results support the role of structural coupling in propagating interaction constraints from the hand-object interface to the full kinematic chain, thereby producing smoother and more structurally coherent whole-body motion. Structural Error is compared only between our variants, since it is defined on their shared structural graph.

A.1.2 Effect of Explicit Interaction Representation

As shown in Table 7, explicitly modeling dynamic interaction correspondences progressively improves the quality of the predicted interaction representation. Compared with geometry-only contact estimation or contact activation prediction alone, introducing contact anchors $\mathbf{q}$ provides more precise surface-level correspondence modeling, leading to improved contact accuracy, temporal consistency, and physical plausibility. Further incorporating the full interaction representation with geometry- and mechanics-aware constraints achieves the best performance across all evaluated metrics. These results demonstrate that explicitly capturing evolving human-object surface relationships, together with physical interaction modeling, provides a more reliable representation for generating realistic and physically consistent HOI motions.

A.1.3 Marker Set Choice

Table 8: Ablations on the marker set placement choice. An up-arrow↑ indicates the performance is better if the value is higher. We use **bold** to represent the best result.

marker set choice	pene↓	CR↑	realism↑	FID↓
CMU	0.293	0.039	0.62	0.293
SSM2	0.278	0.041	0.59	0.212
Ours(InterAct)	**0.106**	**0.051**	**0.76**	**0.107**

We further investigate the impact of different marker set placement strategies on interaction modeling. As shown in Table 8, marker sets defined on existing human motion datasets, such as CMU

Table 9: User study results on human-object interaction quality. Higher scores indicate better perceptual quality.

Method	Interaction Realism ↑	Contact Quality ↑	Motion Naturalness ↑	Overall Preference ↑
InterDiff	2.41	2.10	3.42	0.16
HOI-Diff	2.75	2.13	3.38	0.21
Ours	**3.61**	**3.17**	**3.69**	**0.63**

and SSM2, lead to inferior interaction quality compared with our InterAct-based marker selection. Although these marker sets provide reasonable coverage of human motion dynamics, they are not specifically designed to capture interaction-critical surface regions, resulting in less accurate contact modeling and weaker physical consistency. In contrast, our marker set placement focuses on interaction-relevant surface areas, enabling more precise characterization of human-object correspondence and contact states. This leads to improved contact realism, reduced penetration, and better overall motion quality, demonstrating the importance of designing surface markers according to the requirements of HOI generation.

A.1.4 User Study

To further evaluate the perceptual quality of generated human-object interactions, we conduct a user study with 20 participants from diverse backgrounds. Each participant is asked to evaluate 10 sets of generated interaction sequences, where each set contains results from InterDiff, HOI-Diff, and our method. The order of the presented videos is randomly shuffled to reduce potential bias. Participants are asked to assess each result from three perspectives: interaction realism, contact quality, and motion naturalness, using a five-point Likert scale, where higher scores indicate better perceptual quality. In addition, participants select their preferred result among the three methods for each sequence. The evaluation focuses on whether the generated motions exhibit realistic human-object coordination, physically plausible contacts, and natural full-body movements.

As shown in Table 9, our method consistently receives higher perceptual scores across all evaluation criteria. Compared with existing approaches, our generated motions achieve more realistic human-object interactions and more plausible contacts, demonstrating the effectiveness of explicitly modeling interaction geometry and physical constraints. Moreover, our method achieves better motion naturalness, indicating that the proposed interaction-aware formulation improves not only local hand-object coordination but also the overall quality of full-body motion. The preference results further show that participants favor our method by a substantial margin, suggesting that the generated interactions are more convincing and better aligned with human perception.

A.2 Additional Visualization

In the multimedia supplementary materials, we present 5 more sets of video comparisons between HOI-diff(Peng et al., 2023), our method without physics-aware interaction optimizations, and our full implementation.

For results of *A person holds a pillow overhead and lowers it*, HOI-Diff produces motions where the pillow and the hand move asynchronously, indicating that object motion is weakly grounded to hand contact. Without physics-aware interaction optimizations, our model suffers from noticeable interpenetration between the hand and the pillow during lowering. In contrast, the full model maintains consistent contact and synchronized motion throughout the interaction.

For results of *A person picks up a bucket from the ground and swings it*, both HOI-Diff and our variant without physics-aware interaction optimizations cause the bucket to rise before the hand reaches it, and the subsequent swinging motion is poorly aligned with hand movement. Our full method establishes contact before lifting and preserves coherent hand–object coupling during the swing.

For results of *A person pulls a table backwards*, HOI-Diff generates an incorrect object trajectory where the table does not slide along the ground. Both baseline results also lack proper coordination

between the hand and the body. Our method instead produces a more realistic interaction, where the person bends forward and engages the whole body to pull the table, demonstrating improved hand–body coordination and physically plausible object motion.

As for results of *A person waves a badminton racket around*, both HOI-Diff and the variant without physics-aware interaction optimizations generate an interaction without stable grasp. Our result guarantees that the racket is firmly grasped by the human.

For the results of *A person lift a table high*, HOI-Diff produces an incorrect interaction where the hands do not properly grasp the table legs, resulting in an implausible lifting motion. It also does not lift the table high enough as the condition requires. The version without physical prior similarly generates inaccurate hand placement and weak contact. Our model generates a more realistic interaction, where the person correctly grips the table legs and lifts with coordinated body motion, demonstrating improved grasp accuracy and physical plausibility.

A.3 METRICS

We follow (Peng et al., 2023) to employ FID, R-precision, Diversity, and Contact Distance. We adopt FID to quantify the distributional discrepancy between real and generated motions based on a pretrained motion encoder. R-Precision measures the alignment between generated motions and their associated text descriptions. Diversity reflects the variability among generated motions. We also follow (Xu et al., 2025) to leverage 1)Penetration that refers to the depth of intersection, defined as the maximum negative distance from human vertices to the object's surface, averaged over the entire sequence, and 2)Contact Ratio that indicates the average proportion of human vertices whose distances to the object fall below a specified threshold. Furthermore, we compute the Foot Skating Ratio, which represents the fraction of frames where foot displacement exceeds 2.5 cm during ground contact (foot height is lower than 5 cm).

A.4 MARKER REPRESENTATION

We represent the human body using a sparse set of 77 markers, each of which is defined by a fixed correspondence to a vertex on the SMPL-X(Loper et al., 2015) mesh. This vertex-based definition ensures a deterministic and reproducible mapping between the marker representation and the underlying parametric body model.

Let $\mathbf{V} \in \mathbb{R}^{N\times 3}$ denote the SMPL-X mesh vertices. The marker positions $\mathbf{M} \in \mathbb{R}^{77\times 3}$ are obtained by directly indexing the mesh vertices as

$$\mathbf{M}_i = \mathbf{V}_{k_i}, \quad i = 1, \ldots, 77, \tag{7}$$

where k_i denotes the SMPL-X vertex index corresponding to the i-th marker.

The complete list of vertex indices used for the 77-marker set, following the SMPL-X vertex indexing convention, is provided below:

```
[5920, 5621, 5882, 3486, 4430, 3258, 2029, 5694, 5645,
9117, 4302, 4319,  707, 4788, 4198, 3998, 8846, 4726,
3682, 3688, 5890, 5901, 3961, 4722, 5698, 5868, 3892,
3712, 4479, 4088, 3477, 4839, 5787, 4291, 8576, 6248,
7166, 6021, 3067, 8250, 8388, 8339, 3116, 7040, 7099,
2198, 7524, 7115, 6265, 6746, 8634, 7462, 6443, 6449,
8584, 8595, 6709, 7458, 8392, 8527, 7215, 6832, 7575,
6503, 8481, 5945, 5947, 8079, 7669, 7794, 7905, 8022,
5361, 4933, 5058, 5169, 5286]
```

This fixed mapping is used consistently throughout all experiments to convert between the marker-based motion representation and SMPL-X mesh geometry.