JointHOI: Jointly Generating Contact Maps Enhances Hand–Object Interaction Generation

Mingyeong Song[1], Jungbin Cho[2,3], Jisoo Kim[2], Ananya Bal[3], Kartik Sharma[3], Youngjae Yu[4], Laszlo A. Jeni[3], and Junhyug Noh[1⋆]

[1] Ewha Womans University, Seoul, Korea
[2] Yonsei University, Seoul, Korea
[3] Carnegie Mellon University, Pittsburgh PA, USA
[4] Seoul National University, Seoul, Korea

Project Page: `https://smk11602.github.io/JointHOI/`

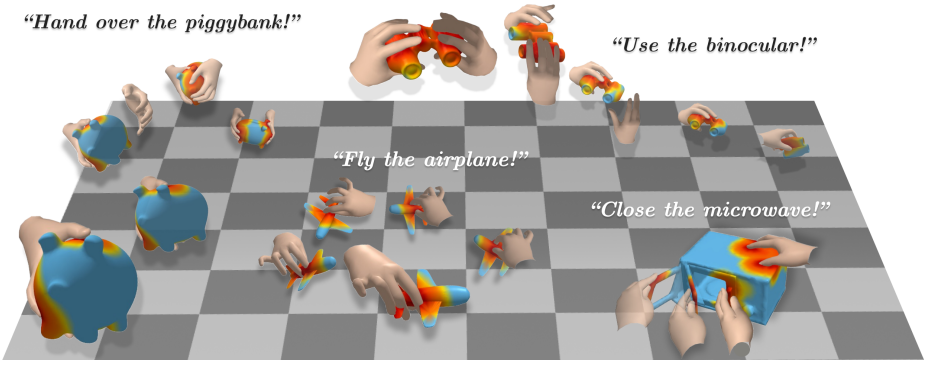

Fig. 1: JointHOI generates bimanual hand–object interaction sequences from text prompts, along with dynamic contact maps that capture time-varying hand–object proximity and help reduce artifacts such as floating and interpenetration.

Abstract. Text-driven hand–object interaction (HOI) generation is gaining attention for immersive applications and robotics, yet producing *physically plausible* interactions remains challenging. Even when individual motions appear natural, small contact errors can cause conspicuous artifacts such as floating and interpenetration. Prior methods mitigate these issues using explicit contact cues or implicit grasp priors, but typically rely on multi-stage pipelines and fail to model *temporally evolving* contact. We present JointHOI, a single-stage diffusion framework that jointly generates 3D hand–object motion and dynamic, distance-based contact maps from text. By treating contact as an auxiliary *inner modality*, joint generation enables the model to learn contact–motion coupling during training. At inference, contact-guided sampling enforces consistency between generated contact maps and motion-implied geometry,

⋆ Corresponding author: J. Noh (junhyug@ewha.ac.kr).

improving temporal stability and reducing penetration and floating. Experiments on GRAB and ARCTIC demonstrate consistent improvements in text adherence and physical plausibility over prior methods.

1 Introduction

Hand–object interactions are defined by contact: where the hand touches the object, when contact forms or breaks, and how that contact evolves as the motion unfolds. This makes contact fidelity central to HOI generation for robotics and AR/VR – robots require precise, stable contacts for reliable manipulation, while AR/VR demands visually convincing interactions that withstand close human scrutiny. Generating HOI from text is therefore not merely a motion synthesis problem; it requires producing *contact-consistent* kinematics with high geometric fidelity. Otherwise, even small errors manifest as interpenetration, floating fingers, or unstable grasps, artifacts that humans are sensitive to [2, 27].

Recent text-conditioned methods such as Text2HOI [5] take an important step toward semantic control, but realistic HOI remains challenging because interaction dynamics are still captured only indirectly. The core difficulty is that contact in HOI is *spatiotemporally structured*: it varies across the object surface and evolves continuously over time as hands approach, grasp, manipulate, and release. Representations that collapse this structure – either by removing temporal variation or by hiding geometry in a latent code – make it difficult to maintain physical plausibility throughout a generated sequence.

Accordingly, recent works [5, 7, 13, 20] augment motion generation with additional interaction modeling. A common approach is to pretrain an autoencoder that implicitly captures hand–object relationships in a latent space. While effective, this introduces a multi-stage training pipeline and lacks an explicit, physically grounded contact signal that can be inspected or enforced. Alternatively, Text2HOI explicitly predicts contact maps and uses them as a conditioning signal for motion generation; however, its contact representation is *static* and *binary*, which cannot capture time-varying contact evolution and provides only a coarse notion of proximity. Moreover, it relies on a multi-stage pipeline and an additional refinement step, further increasing complexity. We summarize these design choices across recent text-driven HOI methods in Tab. 1.

In this paper, we propose JointHOI, a single-stage joint diffusion framework that concurrently generates HOI motion and *dynamic, distance-based* contact maps directly from text. Our key idea is to treat contact as an *inner modality* of hand–object motion: we co-generate 3D trajectories together with contact evolution within a unified diffusion process so the model can learn motion–contact dependencies explicitly rather than inferring them only implicitly from kinematics. Concretely, we parameterize contact as a distance-to-surface field over a fixed set of object-surface anchors, yielding a dense, continuous signal that tracks contact evolution frame-by-frame. Jointly modeling motion and contact mitigates error propagation in sequential pipelines and lets the model internalize physically grounded co-variation patterns (*e.g.*, approach, slide, and release) directly through transformer self-attention.

Method	No post-ref.†	Single-stage‡	Implicit interaction§	Explicit contact¶	Dynamic contact	Distance contact	Inner guidance
Text2HOI [5]	✗	✗	✗	✓	✗	✗	✗
DiffH2O [7]	✓	✗	✗	✗	✗	✗	✗
LatentHOI [20]	✓	✗	✓	✗	✗	✗	✗
HOIGPT [13]	✓	✗	✓	✗	✗	✗	✗
JointHOI (Ours)	✓	✓	✗	✓	✓	✓	✓

Table 1: Design matrix for text-driven HOI generation. JointHOI combines (i) *single-stage* generation without post-refinement and (ii) *explicit* interaction modeling via co-generated *dynamic, distance-based* contact maps, which further enables inference-time *inner guidance* for physically plausible hand–object motion. †No post-refinement: no extra optimization/post-processing after generation. ‡Single-stage: no separate grasp/trajectory stage. §Implicit interaction: interaction encouraged only via latent/objective terms (no explicit contact signal). ¶Explicit contact: an explicit contact representation is predicted/used for generation.

Joint generation, however, does not guarantee that the synthesized motion will faithfully satisfy the predicted contacts at inference. We therefore introduce Contact Inner Guidance (CIG), an inference-time sampling strategy that steers the denoising process toward motions that better respect hand–object proximity. Concretely, CIG enforces consistency between the predicted contact maps and the contact inferred from the synthesized geometry, and uses classifier-based guidance to steer hand motion, reducing penetration and floating artifacts. Crucially, CIG operates strictly as a training-free guidance mechanism during inference without introducing external neural networks, separate grasp planning stages, or post-processing refinement steps, thereby fully preserving the unified single-stage nature of our pipeline. Extensive experiments show that JointHOI synthesizes more natural motions with better text adherence, and that contact guidance further improves physical plausibility by reducing penetration and floating. We also conduct targeted ablations to isolate each component and validate the overall design.

Our primary contributions are summarized as follows:

- We introduce **JointHOI**, a single-stage joint diffusion framework that generates 3D hand–object motion together with *dynamic, distance-based* contact maps directly from text.
- We present **Contact Inner Guidance (CIG)**, an inference-time guided sampling strategy with proximity-aware weighting that leverages co-generated contact maps to steer denoising and reduce penetration and floating artifacts without requiring secondary training stages.
- We provide extensive evaluations and targeted ablations demonstrating improved text adherence, motion naturalness, and physical plausibility, validating the effectiveness of each component and the overall pipeline.

2 Related Works

Text to Hand–Object Interaction Generation. Recent methods generate HOI sequences conditioned on text and object geometry to enable semantic control without explicit object trajectories, but realistic HOI requires fine-grained, temporally evolving contact that is hard to infer from motion alone. Text2HOI [5] and DiffH2O [7] factorize generation into multi-stage designs (*e.g.*, grasp or contact map prediction followed by motion refinement), while LatentHOI [20] and HOI-GPT [13] employ latent priors or autoregressive modeling. While similar contact-aware or geometry-guided representations have recently emerged in the broader human–object interaction domain [21, 34], existing hand–object generation frameworks still heavily rely on complex multi-stage pipelines or static binary cues, which hinder modeling long-horizon, time-varying contact. In contrast, JointHOI performs single-stage generation and explicitly models interaction with *distance-based, dynamic* contact maps, providing a spatiotemporal contact signal that can be enforced during sampling.

Contact in Human Motion Modeling. Contact is widely used to represent interactions between 3D human surfaces and environments [8,33], but these full-body methods lack the fine-grained hand articulation required for realistic HOI, which is emphasized in grasp generation [14, 22, 38]. In hand–object reconstruction, frameworks like ContactOpt [11] and Cao *et al.* [4] utilize contact optimization to resolve physical violations from vision inputs. Similarly, for motion generation, HOIDINI [30] and CODA [26] incorporate contact constraints via inference-time optimization (*e.g.*, DNO [15]), often requiring hundreds of steps. In contrast, we learn bimanual interactions within a single diffusion model and leverage co-generated contact for highly efficient inference-time guidance. Closest to our setting, BimArt [37] predicts HOI contact maps but assumes fixed object trajectories; we target *text-to-HOI* where object motion is synthesized jointly with contact and hand motion.

Joint Generative Modeling and Inner Modalities. Joint generation of multiple modalities has been increasingly explored both for multimodal synthesis and for improving the fidelity of a primary output. Prior work co-generates appearance with geometry cues such as depth or 3D structure [3, 12, 18, 36] and jointly synthesizes audio with motion/visual signals [19]. General any-to-any frameworks [23,24,35] and masked multimodal modeling (*e.g.*, 4M) [1,25] further suggest the benefit of shared representations across modalities. More recently, auxiliary *inner* modalities – signals directly derivable from the primary output – have been used as additional generation targets to strengthen generation; for example, VideoJAM [6] co-generates optical flow to enhance temporal coherence, and Redi [17] co-generates DINOv2 features to improve image synthesis. Motivated by the central role of contact in HOI, we treat dynamic contact maps as an inner modality of hand–object motion and leverage the co-generated contact signals to guide diffusion sampling toward physically plausible interactions.

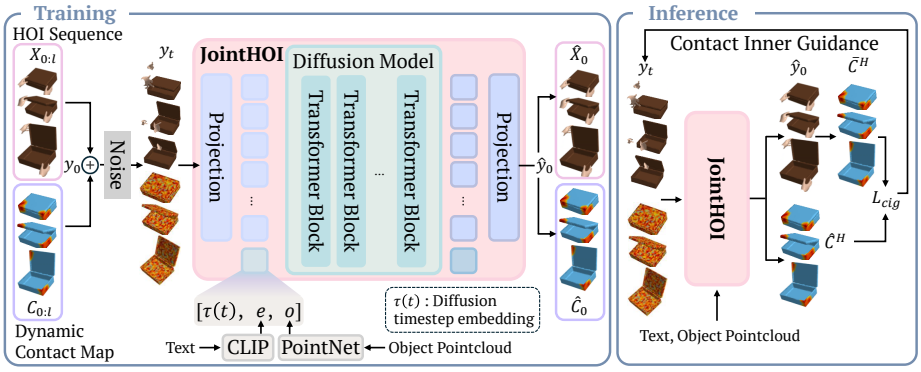

Fig. 2: Overview of JointHOI. Training: We jointly diffuse an HOI sequence y comprising bimanual motion, object motion, and per-hand dynamic contact maps, and train a Transformer diffusion model to predict $\hat{y}_0$ from noisy inputs y_t conditioned on text (CLIP) and object geometry (PointNet). **Inference:** We apply Contact Inner Guidance (CIG) using a contact-consistency energy $\mathcal{L}_{\text{cig}}$ that encourages agreement between the generated contact maps and the contact implied by the synthesized geometry, improving physical plausibility without post-processing.

3 Method

We study text-driven generation of bimanual 3D hand–object interactions conditioned on object geometry. The key challenge is *contact-consistent* motion: small errors in contact timing or proximity readily produce artifacts such as penetration, jitter, and floating. Prior text-to-HOI approaches often rely on multi-stage pipelines with static/binary contact cues or implicit latent priors, increasing complexity and limiting their ability to model *time-varying* contact.

We address these limitations with **JointHOI**, a *single-stage joint diffusion* framework that jointly synthesizes HOI motion and a *distance-based dynamic contact map* in one generative process (Fig. 2).

Our approach consists of (i) a continuous, temporally dense contact representation that explicitly tracks frame-wise hand–object proximity (Sec. 3.2), (ii) a joint diffusion model that learns contact–motion coupling in a shared latent space by generating motion and contact simultaneously (Sec. 3.3), and (iii) Contact Inner Guidance (CIG), an inference-time guided sampling strategy that enforces agreement between explicitly generated contacts and motion-implied contacts (Sec. 3.4).

3.1 Preliminaries

Notation. We use $t \in \{1, \ldots, T\}$ to index diffusion steps of a *noised* variable (*e.g.*, x_t, y_t), and $l \in \{1, \ldots, L_{\max}\}$ to index temporal frames (*e.g.*, x_l, y_l). We

omit the diffusion index when referring to clean motion variables and deterministic geometry mappings, and write $(\cdot)_t$ only for noised diffusion variables.

Diffusion models. Diffusion models learn to invert a fixed Markov noising process. Given clean data x_0, the forward process produces $\{x_t\}_{t=1}^{T}$ via

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{1-\beta_t}\, x_{t-1},\, \beta_t I\big), \tag{1}$$

which yields the marginal

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s). \tag{2}$$

The reverse process is parameterized by a denoiser D_θ conditioned on an external signal c (*e.g.*, text/object) and the timestep. We adopt *clean-sample prediction* and estimate

$$\hat{x}_0 = D_\theta(x_t, c, t), \tag{3}$$

trained with the reconstruction objective

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{x_0,t,\epsilon}\left[\|x_0 - D_\theta(x_t, c, t)\|_2^2\right]. \tag{4}$$

At inference time, we sample $x_T \sim \mathcal{N}(0, I)$ and iteratively apply the learned reverse transitions to obtain a final sample.

3D hand–object motion representation. We represent a bimanual interaction as three time-aligned sequences

$$x = \big(x^L, x^R, x^O\big) \quad \text{of length } L_{\max}, \tag{5}$$

where $x^L = \{x_l^L\}_{l=1}^{L_{\max}}$, $x^R = \{x_l^R\}_{l=1}^{L_{\max}}$, and $x^O = \{x_l^O\}_{l=1}^{L_{\max}}$ share the same frame index l. Each hand state is $x_l^H = [t_l^H, \theta_l^H] \in \mathbb{R}^{99}$ for $H \in \{L, R\}$, where $t_l^H \in \mathbb{R}^3$ is the global translation and $\theta_l^H \in \mathbb{R}^{16\times 6}$ are MANO pose parameters in 6D rotation form (flattened). The object state is $x_l^O = [t_l^O, \varphi_l^O, \delta_l] \in \mathbb{R}^{10}$, where $t_l^O \in \mathbb{R}^3$ is translation, $\varphi_l^O \in \mathbb{R}^6$ is 6D rotation, and $\delta_l \in \mathbb{R}$ is an articulation scalar. For rigid-object datasets (*e.g.*, GRAB [31]), we omit articulation and use $x_l^O = [t_l^O, \varphi_l^O] \in \mathbb{R}^9$. Our model outputs predicted trajectories $\hat{x} = \big(\hat{x}^L, \hat{x}^R, \hat{x}^O\big)$; quantities derived from predictions inherit the hat notation.

Differentiable geometry. We use MANO [29] as a differentiable parametric hand model to obtain per-frame meshes and joints:

$$(V_H, J_H) = \text{MANO}(x^H), \tag{6}$$

where $V_H \in \mathbb{R}^{L_{\max}\times V\times 3}$ and $J_H \in \mathbb{R}^{L_{\max}\times J\times 3}$ (V=778, J=21). For the object, given a canonical point cloud $P \in \mathbb{R}^{N\times 3}$, we map it to the world frame at each time l using the object state:

$$P_{\text{def}}^l = \mathcal{T}(P; t_l^O, \varphi_l^O, \delta_l), \tag{7}$$

where $\mathcal{T}$ applies per-frame translation/rotation (and articulation δ_l when available; omitted for rigid objects), yielding $P_{\text{def}} \in \mathbb{R}^{L_{\max}\times N\times 3}$ (and analogously $\hat{P}_{\text{def}}$ from $\hat{x}^O$).

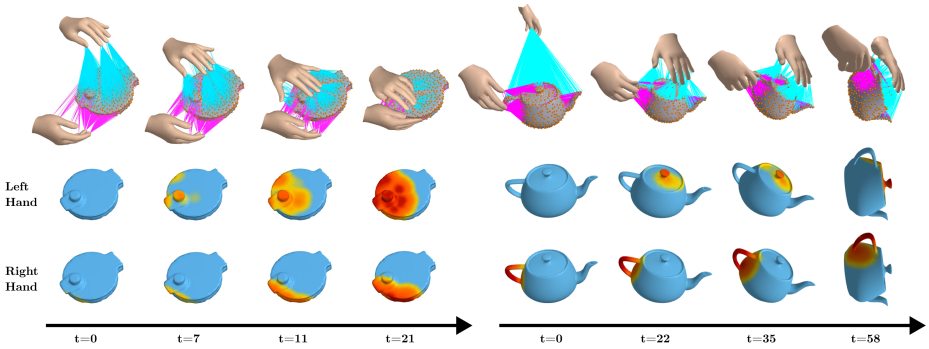

Fig. 3: Dynamic contact map visualization. Top: For each object-surface anchor, we connect it to its nearest point on the left (cyan) and right (magenta) hand, illustrating the per-hand proximity field over time. **Bottom:** We render the corresponding anchor-to-hand distances as heatmaps on the object surface (shown separately for left/right hands), highlighting how contact regions emerge, move, and disappear throughout the interaction.

3.2 Dynamic Contact Map

Prior text-to-HOI methods either use *static* contact cues (*e.g.*, a binary map) as a conditioning signal or learn an *implicit* interaction prior in a latent space (*e.g.*, via a VAE [16]). In contrast, we introduce a *dynamic contact map* that is (i) *temporally dense*, capturing frame-wise contact evolution, and (ii) *explicit*, providing an interpretable geometric signal that can be directly leveraged at inference time. We represent contact as *continuous distances* rather than binary labels, which avoids threshold sensitivity and yields a smooth signal near contact boundaries – particularly beneficial for gradient-based guidance during sampling. As illustrated in Fig. 3, this distance field yields a temporally evolving proximity signal over the object surface for each hand.

Anchor-based distance representation. For each object, we predefine a fixed set of N_c=1024 anchor points

$$\mathcal{A} = \{a_n\}_{n=1}^{N_c} \tag{8}$$

in the object canonical frame. We employ farthest point sampling (FPS) to ensure that $\mathcal{A}$ uniformly covers the object surface. These anchors define an object-centric discretization that is shared across time and invariant to global motion, allowing contacts to be tracked consistently as the object moves. Given an object motion sequence $x^O = \{x_l^O\}_{l=1}^{L_{\max}}$, we transform anchors into the global frame at each time l, yielding

$$\mathcal{A}_l = \{a_{l,n}\}_{n=1}^{N_c}. \tag{9}$$

We define bimanual contact at frame l as $C_l = (C_l^L, C_l^R)$ with $C_l^H \in \mathbb{R}^{N_c}$ for $H \in \{L, R\}$, where each entry stores the minimum Euclidean distance from an anchor point to the corresponding hand surface. In practice, we compute this distance

by taking the minimum over hand mesh vertices, which provides an efficient and differentiable approximation of surface proximity. Let $\mathcal{U}_l^H = \{V_H[l, m, :]\}_{m=1}^V$ denote the hand mesh vertices at frame l; we compute

$$C_l^H[n] = \min_{u \in \mathcal{U}_l^H} \|u - a_{l,n}\|_2, \qquad n = 1, \ldots, N_c, \quad H \in \{L, R\}. \tag{10}$$

Design choices. As shown in Fig. 3, contact patterns are asymmetric across hands and evolve over time with complex spatial patterns.

Therefore, we use per-hand bimanual contact maps [37] to reduce ambiguity in bimanual interactions, and adopt dynamic, distance-based maps to represent complex spatiotemporal contact structures.

3.3 Joint Diffusion Model

Rather than treating contact as a separate, precomputed condition, we *jointly* generate HOI motion and contact dynamics with a single diffusion model. Given an object geometry descriptor and a text prompt c, we directly model the joint distribution of bimanual hand motion, object trajectory, and dynamic contact. This joint formulation reduces error propagation inherent to sequential pipelines and enables the denoiser to learn *contact–motion coupling* in a shared latent space. Intuitively, motion and contact must co-vary under physical constraints: as the hand approaches and reorients around a surface patch, contact distances should decrease smoothly and evolve consistently over time.

Unified sequence representation. Let $x_l^L \in \mathbb{R}^{99}$, $x_l^R \in \mathbb{R}^{99}$, and $x_l^O \in \mathbb{R}^{10}$ denote per-frame motion states (Sec. 3.1), and let $C_l^L, C_l^R \in \mathbb{R}^{N_c}$ denote per-frame dynamic contact maps (Sec. 3.2). We concatenate them into a single per-frame token,

$$y_l = \left[x_l^L,\ x_l^R,\ x_l^O,\ C_l^L,\ C_l^R\right] \in \mathbb{R}^{d_y}, \qquad d_y = 99 + 99 + 10 + 2N_c, \tag{11}$$

and stack over time to obtain the joint sequence

$$y = \{y_l\}_{l=1}^{L_{\max}} \in \mathbb{R}^{L_{\max} \times d_y}. \tag{12}$$

We define diffusion directly on y, so each transformer layer can model both (i) *within-frame* dependencies between kinematics and contact and (ii) *across-frame* temporal evolution via self-attention.

Transformer denoiser and conditioning. We use a transformer denoiser that takes the noisy sequence y_t and predicts the corresponding clean sequence $\hat{y}_0$. The model is conditioned on (i) an object feature vector $\mathbf{o} \in \mathbb{R}^{1088}$ (global/local PointNet [28] features, scale, centroid), (ii) a CLIP text embedding $\mathbf{e} \in \mathbb{R}^{d_e}$, and (iii) a timestep embedding $\tau(t)$. We inject these conditions using *prefix conditioning*: we append a small set of learnable condition tokens to the beginning of the token sequence, $z = [z^{\text{text}}, z^{\text{obj}}, z^t]$, initialized from $\mathbf{e}$, $\mathbf{o}$, and $\tau(t)$, respectively. The denoiser then applies standard *full self-attention* over the concatenated sequence $[z;\ \phi(y_t)]$, where $\phi(\cdot)$ linearly projects y_t to the model dimension. As a

result, each frame token can attend to the text token to capture action semantics and to the object token to adapt contact evolution to object-specific geometry (*e.g.*, different grasp regions for a mug handle vs. a cup body), without introducing a separate cross-attention module. We write the denoiser compactly as

$$\hat{y}_0 \;=\; D_\theta\big(y_t,\ \mathbf{o},\ \mathbf{e},\ \tau(t)\big). \tag{13}$$

Training objective. We adopt clean-sample prediction and train the denoiser to reconstruct the clean joint sequence:

$$\mathcal{L}_{\text{joint}} = \mathbb{E}_{y,t,\epsilon}\Big[\big\|y - D_\theta(y_t,\ \mathbf{o},\ \mathbf{e},\ \tau(t))\big\|_2^2\Big]. \tag{14}$$

Because y contains both kinematic variables and contact distances, optimizing $\mathcal{L}_{\text{joint}}$ encourages the transformer to learn *cross-modal dependencies* through self-attention: motion dimensions can inform contact dimensions and vice versa, allowing the model to internalize physically grounded correlations such as decreasing distances during approach, increasing distances during release, and coordinated contact changes during sliding or rolling.

Inference and parsing. At inference time, we sample $y_T \sim \mathcal{N}(0, I)$ and iteratively apply the learned reverse transitions conditioned on $(\mathbf{o}, \mathbf{e})$ to obtain $\hat{y}_0$. We parse $\hat{y}_0$ into $\hat{x}^L, \hat{x}^R, \hat{x}^O, \hat{C}^L, \hat{C}^R$ by slicing the corresponding dimensions, and compute hand meshes and motion-implied contact geometry using the operators in Sec. 3.1 and Sec. 3.2.

3.4 Contact Inner Guidance

Although our model is trained to jointly predict motion and contact, long and highly dynamic interactions can still accumulate geometric inconsistencies during iterative denoising. We therefore improve physical plausibility at inference time by enforcing consistency between (i) the contact maps explicitly generated by the diffusion model and (ii) the contact maps implied by the generated hand–object geometry. We achieve this via a classifier-based *guided sampling* scheme, where a differentiable contact-consistency energy provides gradients that steer the denoising trajectory. We refer to this procedure as *Contact Inner Guidance (CIG)*, since it leverages contact – a co-generated inner modality – to guide hand–object motion generation at inference time.

Contact consistency energy. At a reverse step, let $\hat{y}_0$ be the current clean estimate predicted by the denoiser. We parse $\hat{y}_0$ into $\hat{x}^L, \hat{x}^R, \hat{x}^O, \hat{C}^L, \hat{C}^R$. From the predicted motion, we compute a *motion-implied* dynamic contact map $\bar{C}^H = \{\bar{C}_l^H\}_{l=1}^{L_{\max}}$ using the same construction as Sec. 3.2: we transform anchors with $\hat{x}^O$ to obtain $\hat{\mathcal{A}}_l$, extract hand vertices $\hat{\mathcal{U}}_l^H$ from $\text{MANO}(\hat{x}^H)$, and set $\bar{C}_l^H[n] = \min_{u \in \hat{\mathcal{U}}_l^H} \|u - \hat{a}_{l,n}\|_2$. We measure agreement between the explicitly generated contacts and the motion-implied contacts with the log-ratio energy

$$\mathcal{L}_{\text{cig}} = \sum_{H \in \{L,R\}} \Big\| \log(\hat{C}^H + \varepsilon) - \log(\bar{C}^H + \varepsilon) \Big\|_1, \tag{15}$$

where $\hat{C}^H, \bar{C}^H \in \mathbb{R}^{L_{\max} \times N_c}$ stack per-frame contact vectors. We use the log scale to emphasize fine-grained, near-contact regions (small hand–object distances), motivated by the observation that these regions carry the most critical cues for realistic interaction, whereas far-field distances are less informative. The constant ε stabilizes the computation near zero.

Classifier-based guidance. We guide sampling by nudging the noisy variable toward samples that reduce $\mathcal{L}_{\text{cig}}$. Concretely, at each reverse step we first obtain the model prediction $\hat{y}_0 = D_\theta(y_t, \mathbf{o}, \mathbf{e}, \tau(t))$, compute $\mathcal{L}_{\text{cig}}(\hat{y}_0)$, and backpropagate through the differentiable geometry operators (MANO and the distance computation) to obtain a guidance gradient with respect to the current noisy variable, $\nabla_{y_t}\mathcal{L}_{\text{cig}}$. We then apply a guidance update with scale w,

$$y_t \leftarrow y_t - w\,\nabla_{y_t}\mathcal{L}_{\text{cig}}, \tag{16}$$

followed by the standard diffusion reverse transition. Intuitively, this steers denoising toward joint samples whose explicitly generated contact distances match those physically induced by the generated motion, reducing artifacts such as contact jitter, floating, and interpenetration.

4 Experiments

4.1 Datasets

We evaluate JointHOI on two widely used bimanual hand–object interaction datasets, GRAB [31] and ARCTIC [9]. For both datasets, we use the text prompts released by Text2HOI [5] as language conditioning. Since no official train/test split is publicly provided for these datasets in the text-to-HOI setting, we construct the splits using a shared protocol: we form a balanced test set by uniformly sampling object–action pairs, and use the remaining sequences for training. All methods are trained and evaluated on the same splits.

GRAB [31]. It provides bimanual interactions with 51 rigid objects across 29 action categories. All sequences are downsampled to 30 fps, and interaction clips are padded or truncated to a maximum length of 196 frames.

ARCTIC [9]. It contains bimanual interactions with 11 articulated objects spanning 10 action categories. In addition to motion, it provides object articulation angles, enabling evaluation on interactions involving moving parts. All sequences are downsampled to 30 fps and padded or truncated to a maximum length of 64 frames.

4.2 Implementation Details

All experiments are conducted on a single NVIDIA RTX A6000 GPU (48GB). We optimize the network using Adam (lr=10^{-4}) and a linear noise schedule with T=1,000 diffusion steps for training, and use 100 sampling steps at inference. We train with batch size 128 on GRAB and 64 on ARCTIC. Training is run for up to

380K iterations on GRAB and 60K iterations on ARCTIC, with early stopping when the training loss plateaus. At inference time, we generate sequences with the same number of frames as the corresponding ground-truth clip to ensure consistent comparison across metrics.

Baselines. We compare JointHOI to state-of-the-art text-to-HOI methods, including MDM [32], DiffH2O [7], LatentHOI [20], and Text2HOI [5]. All baselines are trained using their official implementations under the same data splits and evaluation protocol. MDM serves as a single-stage diffusion baseline without interaction-specific guidance. For the two-stage DiffH2O, we report its standard setting and DiffH2O*, where the grasp stage is conditioned on ground-truth grasp signals to isolate the quality of interaction synthesis. Similarly, we report Text2HOI and Text2HOI*, where Text2HOI* removes the post-hoc refinement module to evaluate the raw motion generated from contact-map conditioning.

4.3 Evaluation Metrics

We evaluate JointHOI from two complementary perspectives: (i) semantic motion quality and (ii) physical plausibility of the hand–object interaction.

Motion Metrics. Following the protocol of IMOS [10], we train an RNN-based action classifier for each dataset and use its hidden features as motion embeddings. We report:

- **Accuracy (Acc):** Top-1 and Top-3 classification accuracy of generated motions under the action classifier, measuring semantic alignment with the target action.
- **FID:** Fréchet distance between feature distributions of real and generated sequences, quantifying overall motion realism.

Physical Plausibility Metrics. Since motion metrics may miss contact artifacts, we further evaluate physical consistency by averaging the following metrics over both hands:

- **Interpenetration Volume (IV) and Depth (ID):** Following LatentHOI [20], we measure the intersection volume (cm^3) and maximum penetration depth (cm) between hand and object meshes. Lower values indicate fewer physical violations.
- **Contact Ratio (CR):** Following Bimart [37], CR is the fraction of frames (where the object translates or articulates) that maintain valid contact (distance $< 5\,mm$). Since CR can be trivially increased by penetration, we consider the best interaction to have *CR close to GT* while keeping IV/ID low.

4.4 Experimental Results

Main results on ARCTIC and GRAB. Tab. 2 summarizes quantitative results on ARCTIC and GRAB.

On **ARCTIC**, JointHOI achieves the highest semantic fidelity and the lowest FID among synthesized methods, while substantially reducing penetration and keeping CR close to GT. It consistently outperforms multi-stage baselines

Table 2: Quantitative comparison on ARCTIC and GRAB. Bold indicates the best performance among synthesized methods (excluding GT). * denotes the oracle-assisted variant where the grasp stage is conditioned on ground-truth grasp signals.

Dataset	Method	Acc ↑ Top-1	Acc ↑ Top-3	FID ↓	Interpenetration ↓ IV (cm^3)	Interpenetration ↓ ID (cm)	CR (%) →
ARCTIC	GT	0.966	0.992	–	4.419	0.288	95.56
	MDM [32]	0.739	0.839	0.365	9.542	0.581	61.37
	DiffH2O* [7]	0.636	0.862	0.424	6.491	0.475	94.28
	DiffH2O [7]	0.560	0.823	0.547	6.412	0.527	**94.91**
	LatentHOI [20]	0.549	0.846	0.301	8.597	0.496	78.81
	Text2HOI* [5]	0.815	0.923	0.152	8.930	0.587	93.28
	Text2HOI [5]	0.816	0.927	0.148	8.281	0.524	97.33
	JointHOI (Ours)	**0.948**	**0.983**	**0.033**	**4.406**	**0.426**	93.71
GRAB	GT	0.779	0.895	–	2.790	0.526	95.02
	MDM [32]	0.395	0.500	1.203	4.915	0.650	81.22
	DiffH2O* [7]	0.595	0.737	0.410	**3.282**	0.610	87.24
	DiffH2O [7]	0.458	0.526	0.533	3.607	0.846	87.82
	LatentHOI [20]	0.584	0.721	0.214	4.356	0.570	88.39
	Text2HOI* [5]	**0.716**	0.805	0.118	8.114	0.810	97.85
	Text2HOI [5]	0.711	0.831	0.116	7.790	0.778	98.24
	JointHOI (Ours)	0.663	**0.847**	**0.031**	3.419	**0.525**	**93.00**

Table 3: Incremental ablation from the baseline by progressively adding our modules. Each row adds one component on top of the previous setting.

Setting	Acc ↑ Top-1	Acc ↑ Top-3	FID ↓	Interpenetration ↓ IV (cm^3)	Interpenetration ↓ ID (cm)	CR (%) →
Baseline (two-stage)	0.520	0.738	0.291	8.096	0.528	73.40
+ One-stage (joint)	0.795	0.919	0.274	7.057	0.548	81.72
+ Dynamic contact map (unified)	0.894	0.948	0.145	7.146	0.432	85.28
+ Dynamic contact map (L/R)	0.935	0.981	0.038	6.988	0.439	92.21
+ Contact Inner Guidance	**0.948**	**0.983**	**0.033**	**4.406**	**0.426**	**93.71**

and remains highly competitive even against the oracle-assisted DiffH2O*. On **GRAB**, JointHOI delivers the best Top-3 accuracy and lowest FID, demonstrating superior motion realism. While Text2HOI shows a slightly higher Top-1 accuracy, our significant Top-3 gain indicates that JointHOI more reliably captures ambiguous action semantics (*e.g.*, *use* vs. *inspect*). Notably, JointHOI achieves this while halving the penetration volume (IV) compared to Text2HOI. These results highlight that jointly modeling motion and interaction constraints leads to higher-fidelity, physically grounded interactions.

Effect of our design choices.

Tab. 3 reports an incremental ablation on ARCTIC. Switching from a two-stage baseline to one-stage joint training already yields a large improvement in semantic fidelity (Top-1 Acc 0.520→0.795), supporting our motivation that joint

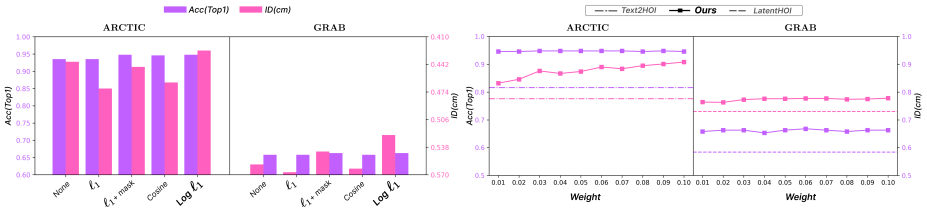

Fig. 4: Ablation of Contact Inner Guidance (CIG). We report Top-1 action accuracy (Acc, left y-axis) and penetration depth (ID, right y-axis) on ARCTIC and GRAB; the ID axis is inverted, so higher is better for both metrics. **Left:** Effect of contact-consistency loss design (None, ℓ_1, ℓ_1+mask, cosine, and Log-ℓ_1). Log-ℓ_1 provides the best overall trade-off, consistently improving physical plausibility without sacrificing semantic accuracy. **Right:** Effect of the guidance weight w. Dashed horizontal lines denote the best prior (SOTA) results; CIG remains robust across a wide range of w and consistently outperforms these baselines.

Table 4: Inference time for a 196-frame sequence on an RTX A6000.

Metric	MDM [32]	DiffH2O [7]	LatentHOI [20]	Text2HOI [5]	JointHOI
Time (s) ↓	10.61	124.44	15.95	16.79	10.98
FPS ↑	18.47	1.57	12.29	11.67	17.85

optimization mitigates stage-wise error accumulation. Introducing a distance-based dynamic contact map further improves realism (FID 0.274→0.145), consistent with the intuition that contact should be modeled as a time-varying signal rather than a static cue. Notably, separating the dynamic contact map by hand (L/R) provides a substantial additional gain in motion realism (FID 0.145 → 0.038), suggesting that bimanual interactions exhibit asymmetric, loosely coupled contact dynamics. Finally, applying CIG achieves the best overall performance, markedly reducing penetration (IV 6.988 → 4.406) while maintaining strong motion quality.

Effect of CIG loss design and guidance weight. Fig. 4 evaluates (left) different CIG consistency losses and (right) the guidance weight w, using Top-1 Acc and penetration depth (ID). Across both datasets, the results support our proximity-based intuition: losses that emphasize near-contact discrepancies improve physical plausibility without hurting semantics. In particular, Log-ℓ_1 yields the best overall Acc–ID trade-off by focusing on relative errors near contact, but other proximity-aware variants (*e.g.*, ℓ_1+mask) also perform strongly, achieving competitive accuracy with reduced penetration. The weight sweep further shows that CIG is not sensitive to w: ARCTIC and GRAB exhibit a broad stable range where Acc remains nearly unchanged while ID is consistently reduced. We therefore use w=0.05 in all subsequent experiments.

Inference efficiency. Table 4 compares the inference time for a 196-frame sequence on an NVIDIA RTX A6000 GPU. JointHOI is substantially more ef-

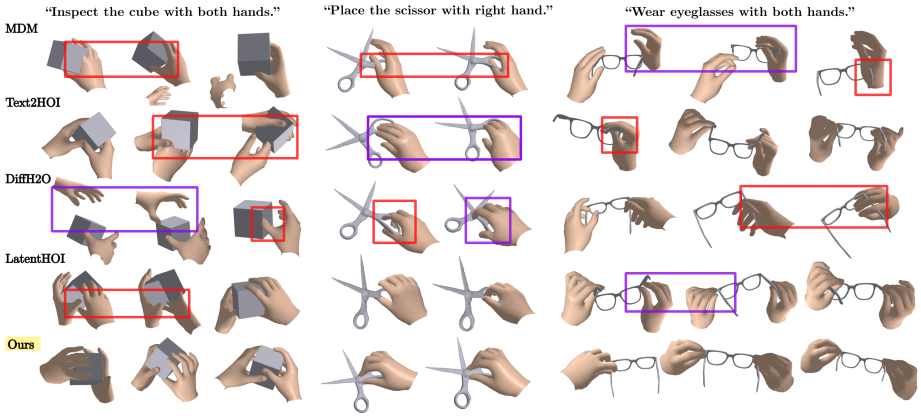

Fig. 5: Qualitative comparison on text-to-HOI generation. We show representative frames for three prompts (columns) and compare MDM [32], Text2HOI [5], DiffH2O [7], and JointHOI (rows). JointHOI produces more coherent bimanual coordination and more physically plausible hand–object interactions, with fewer artifacts such as floating (in purple boxes), unstable grasps, and penetration (in red boxes).

ficient than multi-stage or latent-based methods (DiffH2O [7], LatentHOI [20], and Text2HOI [5]), while remaining close to the single-stage baseline MDM [32]. The overhead of Contact Inner Guidance (CIG) is modest because it operates directly on the joint motion–contact representation using fixed anchors, preserving inference speed while improving physical plausibility.

Qualitative results. Fig. 5 compares text-to-HOI generations from MDM, Text2HOI, DiffH2O, LatentHOI, and JointHOI. Across prompts, JointHOI produces more coherent bimanual coordination and more stable hand–object proximity. For *"Inspect the cube with both hands"*, baselines often show weak two-hand engagement, while JointHOI maintains coordinated bimanual contact. For *"Place the scissor with right hand"*, competing methods frequently exhibit floating or unstable grasps during placement, whereas JointHOI preserves plausible right-hand control with fewer contact artifacts. For *"Wear an eyeglasses with both hands"*, a challenging case due to thin structures, JointHOI better aligns both hands to the frame and maintains consistent contact. Overall, these results indicate that jointly modeling motion with dynamic contact and enforcing contact consistency during sampling yields more physically plausible interactions with fewer floating and interpenetration artifacts.

Visualization of dynamic contact maps.

Fig. 6 compares the explicitly generated contact maps with those analytically derived from the synthesized hand–object motions. The high consistency between the two demonstrates that our model effectively learns motion–contact coupling. This supports our design choice of treating contact as an inner modality, showing that the predicted contact maps offer reliable signals for inference-time guidance.

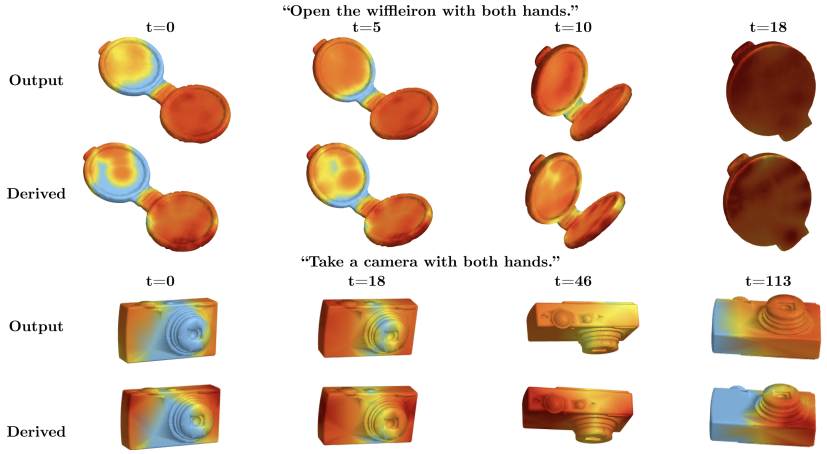

Fig. 6: Visualization of dynamic contact maps. For each timestep, the first row (**Output**) shows the explicitly generated contact maps by JointHOI, and the second row (**Derived**) shows those analytically derived from the motions. The close agreement between them demonstrates that the predicted contact faithfully reflects the synthesized geometry.

5 Conclusion

We presented JointHOI, a single-stage joint diffusion framework for text-driven hand–object interaction generation that co-generates 3D motion with dynamic, distance-based contact maps. By treating contact as an inner modality, JointHOI explicitly models the spatiotemporal evolution of hand–object proximity and learns contact–motion coupling within a unified generative process. To further improve physical plausibility at inference time, we introduced contact-guided sampling via classifier-based guidance, which enforces consistency between the generated contact maps and the contact implied by the synthesized geometry, reducing common artifacts such as penetration, floating, and contact jitter. Extensive experiments on GRAB and ARCTIC demonstrate consistent improvements in text adherence, motion quality, and physical plausibility over prior approaches, and ablations validate the contribution of each component.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2026-25499022), and Global – Learning & Academic research institution for Master's · PhD students, and Postdocs (G-LAMP) Program of the NRF funded by the Ministry of Education (RS-2025-25442252).

References

1. Bachmann, R., Kar, O.F., Mizrahi, D., Garjani, A., Gao, M., Griffiths, D., Hu, J., Dehghan, A., Zamir, A.: 4m-21: An any-to-any vision model for tens of tasks and modalities. Advances in Neural Information Processing Systems **37**, 61872–61911 (2024)
2. Burns, E., Razzaque, S., Panter, A.T., Whitton, M.C., McCallus, M.R., Brooks Jr, F.P.: The hand is more easily fooled than the eye: Users are more sensitive to visual interpenetration than to visual-proprioceptive discrepancy. Presence: teleoperators & virtual environments **15**(1), 1–15 (2006)
3. Byung-Ki, K., Dai, Q., Hyoseok, L., Luo, C., Oh, T.H.: Jointdit: Enhancing rgb-depth joint modeling with diffusion transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25261–25271 (2025)
4. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12417–12426 (2021)
5. Cha, J., Kim, J., Yoon, J.S., Baek, S.: 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)
6. Chefer, H., Singer, U., Zohar, A., Kirstain, Y., Polyak, A., Taigman, Y., Wolf, L., Sheynin, S.: Videojam: Joint appearance-motion representations for enhanced motion generation in video models. In: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 7595–7616. PMLR (2025)
7. Christen, S., Hampali, S., Sener, F., Remelli, E., Hodan, T., Sauser, E., Ma, S., Tekin, B.: Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
8. Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19888–19901 (2024)
9. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: Arctic: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12943–12954 (2023)
10. Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: Imos: intent-driven full-body motion synthesis for human-object interactions. In: Computer Graphics Forum. vol. 42, pp. 1–12. Wiley Online Library (2023)
11. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: Contactopt: Optimizing contact to improve grasps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1471–1481 (2021)
12. Hassan, M., Stapf, S., Rahimi, A., Rezende, P., Haghighi, Y., Brüggemann, D., Katircioglu, I., Zhang, L., Chen, X., Saha, S., et al.: Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22404–22415 (2025)
13. Huang, M., Chu, F.J., Tekin, B., Liang, K.J., Ma, H., Wang, W., Chen, X., Gleize, P., Xue, H., Lyu, S., et al.: Hoigpt: Learning long-sequence hand-object interaction with language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7136–7146 (2025)

14. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11107–11116 (2021)
15. Karunratanakul, K., Preechakul, K., Aksan, E., Beeler, T., Suwajanakorn, S., Tang, S.: Optimizing diffusion noise can serve as universal motion priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1334–1345 (2024)
16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
17. Kouzelis, T., Karypidis, E., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Boosting generative image modeling via joint image-feature synthesis. Advances in Neural Information Processing Systems (2025)
18. Krishnan, A., Yan, X., Casser, V., Kundu, A.: Orchid: Image latent diffusion for joint appearance and geometry generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28217–28227 (2025)
19. Kwon, M., Shin, J., Jung, J., Park, J., Uh, Y.: Jam-flow: Joint audio-motion synthesis with flow matching (2025), https://arxiv.org/abs/2506.23552
20. Li, M., Christen, S., Wan, C., Cai, Y., Liao, R., Sigal, L., Ma, S.: Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17416–17425 (2025)
21. Li, Z., Li, J., Gao, M., Yang, J., Wang, W., Zheng, F.: Conta-hoi: Towards physically plausible human-object interaction generation via contact-aware modeling. In: 2026 International Conference on 3D Vision (3DV). pp. 946–955. IEEE (2026)
22. Liu, S., Zhou, Y., Yang, J., Gupta, S., Wang, S.: Contactgen: Generative contact modeling for grasp generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20609–20620 (2023)
23. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26439–26455 (2024)
24. Lu, J., Clark, C., Zellers, R., Mottaghi, R., Kembhavi, A.: Unified-io: A unified model for vision, language, and multi-modal tasks. arXiv preprint arXiv:2206.08916 (2022)
25. Mizrahi, D., Bachmann, R., Kar, O., Yeo, T., Gao, M., Dehghan, A., Zamir, A.: 4m: Massively multimodal masked modeling. Advances in Neural Information Processing Systems **36**, 58363–58408 (2023)
26. Pi, H., Cen, Z., Dou, Z., Komura, T.: Coda: Coordinated diffusion noise optimization for whole-body manipulation of articulated objects. Advances in Neural Information Processing Systems **38**, 31758–31780 (2026)
27. Prachyabrued, M., Borst, C.W.: Visual interpenetration tradeoffs in whole-hand virtual grasping. In: 2012 IEEE Symposium on 3D User Interfaces (3DUI). pp. 39–42. IEEE (2012)
28. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
29. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)
30. Ron, R., Tevet, G., Sawdayee, H., Bermano, A.: Hoidini: Human-object interaction through diffusion noise optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5739–5749 (2026)

31. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: European conference on computer vision. pp. 581–600. Springer (2020)
32. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=SJ1kSyO2jwu
33. Wang, Z., Chen, Y., Jia, B., Li, P., Zhang, J., Zhang, J., Liu, T., Zhu, Y., Liang, W., Huang, S.: Move as you say interact as you can: Language-guided human motion generation with scene affordance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 433–444 (2024)
34. Xue, M., Liu, Y., Guo, L., Huang, S., Ding, C.: Guiding human-object interactions with rich geometry and relations. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22714–22723 (2025)
35. Yu, L., Shi, B., ..., Aghajanyan, A.: Scaling autoregressive multi-modal models: Pretraining and instruction tuning (2023)
36. Zhang, Q., Zhai, S., Martin, M.A.B., Miao, K., Toshev, A., Susskind, J., Gu, J.: World-consistent video diffusion with explicit 3d modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21685–21695 (2025)
37. Zhang, W., Dabral, R., Golyanik, V., Choutas, V., Alvarado, E., Beeler, T., Habermann, M., Theobalt, C.: Bimart: A unified approach for the synthesis of 3d bimanual interaction with articulated objects. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27694–27705 (2025)
38. Zhang, Z., Wang, H., Yu, Z., Cheng, Y., Yao, A., Chang, H.J.: Nl2contact: Natural language guided 3d hand-object contact modeling with diffusion model. In: European Conference on Computer Vision. pp. 284–300. Springer (2024)