DEX2HOI: DEXTEROUS BIMANUAL TWO-OBJECT INTERACTION GENERATION

Anonymous authors
Paper under double-blind review

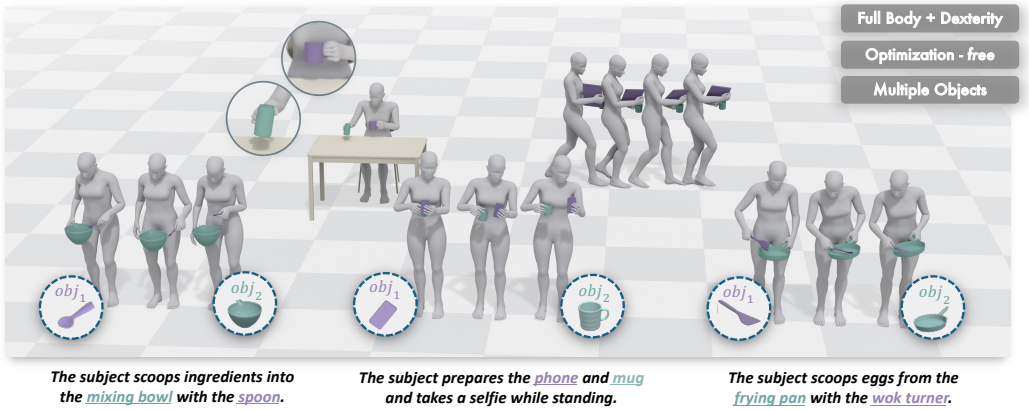

Figure 1: **Dex2HOI** is a diffusion-based model for Human-Object Interaction that generates dexterous bimanual manipulations from text, in single-shot, supporting interaction with up to two objects.

ABSTRACT

Recent advances in 4D Human-Object Interaction (HOI) generation have enabled increasingly realistic motion synthesis, particularly for single-object manipulation. Yet current research overlooks an inherent property of human behavior: people naturally coordinate both hands and manipulate multiple objects simultaneously. To address this gap, we present **Dex2HOI**, a unified diffusion model for single- and two-object HOI synthesis from text. At its core, Dex2HOI employs a Dual-Stream Diffusion approach, where each object is processed in a dedicated interaction stream and coordinated through bidirectional cross-attention. To synthesize the final motion, we introduce a Motion Fusion Network integrated with novel hand-relative object representations and contact-aware conditioning applied across the whole sequence. By sampling the diffusion process autoregressively over prefix-conditioned windows, Dex2HOI generates arbitrarily long sequences at real-time speed omitting redundant test-time optimization, achieving up to ×540 inference speed-up over prior state-of-the-art methods. Extensive evaluation on both single- and two-object benchmarks demonstrates *state-of-the-art* quantitative results, marking a step beyond conventional single-object HOI generation and toward expressive multi-object manipulation. Code and models will be released upon acceptance.

1 INTRODUCTION

Human motion is fundamentally shaped by interaction with our surroundings. We engage with objects in diverse ways, simultaneously manipulating multiple objects, or choosing to handle a single object with one or both hands depending on factors such as its weight, size, or the task at hand. This kind of coordinated bimanual multi-object manipulation, although fundamental to human-like behavior, remains severely underexplored in current human-object interaction (HOI) generation research (Sui et al., 2026). Synthesizing such motions is more than a modelling curiosity: it underpins numerous applications, from realistic interactions in AR/VR systems, to embodied agents that assist in cluttered human environments and robotic systems that rely on manipulation priors (Chen et al.; Pan et al., 2024; Wu et al., 2025).

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.

Recent progress in HOI generation has led to increasingly realistic motions on primarily single-object benchmarks, driven largely by diffusion-based models (Peng et al., 2025; Pi et al., 2025; Ron et al., 2025; Zhang et al., 2025). Despite this, most existing HOI methods face three limitations. First, nearly all state-of-the-art full-body methods assume only single-object interactions. Although a few datasets (Kim et al., 2025; Lv et al., 2024; Lu et al., 2025) include multi-object sequences, dedicated generative methods remain largely unexplored. Second, many existing methods focus only on a subset of the human body: whole-body approaches such as the OMOMO (Li et al., 2023) benchmark omit detailed hand articulation, while hand-centric methods like Text2HOI (Cha et al., 2024) and LatentHOI (Li et al., 2025b) partially address fine-grained grasping but sacrifice full-body generation. Third, current approaches in full-body HOI motion synthesis rely on auxiliary modules and optimization strategies to enforce accurate contacts. Some examples include predicting intermediate contact maps to guide generation (Diller et al., 2024), conditioning the model on full object trajectories (Zhang et al., 2025) or iterative test-time refinement such as DNO (Karunratanakul et al., 2024) with contact-aware constraints (Pi et al., 2025; Ron et al., 2025), resulting in inference overhead.

To address these limitations, we introduce a paradigm shift in both the representation of object motion and the architectural design of the generative model. We present **Dex2HOI**, a unified framework to tackle both single- and two-object full-body HOI synthesis from text. Our main objective is to approach HOI generation in a simple yet informed way: removing model redundancy, avoiding multi-stage pipelines, and producing accurate contacts in a single-shot diffusion sampling. At the representation level, we model object motion as a learnable decomposition of hand-relative branches fused through learnable weights. This decomposition encodes a strong inductive bias for hand-object contact and naturally guides the diffusion process toward smooth transitions of control between hands. At the architectural level, we propose the use of a Dual-Stream Diffusion decoder that processes each object in a dedicated interaction stream and models information across streams through bidirectional cross-attention. The body motions from each stream are then combined with our Motion Fusion Network to produce a single coherent motion. The same architecture handles *both* one- and two-object sequences without modification, supporting a unified treatment of single- and bimanual manipulation. Finally, we train an end-to-end model with contact-aware supervision, removing the need for test-time optimization or auxiliary contact-prediction modules. Together, these choices enable single-shot generation that is up to $\times 540$ **faster** than prior optimization-based *state-of-the-art* approaches. Overall, our contributions can be summarized as follows:

- We propose the first *unified* HOI generation model that synthesizes both single- and two-object interactions, advancing beyond existing single-object approaches.
- We introduce a novel hand-relative object motion representation and train end-to-end denoising models that enable efficient single-shot generation of complex sequences while maintaining physical plausibility and contact consistency across two hands and two objects.
- We evaluate **Dex2HOI** generated motions across three benchmarks, GRAB (Taheri et al., 2020) for single-object HOI, and HIMO (Lv et al., 2024) and HUMOTO (Lu et al., 2025) for two-object HOI. We achieve state-of-the-art performance in both single- and two-object settings, which we further validate through VLM-based and user preference studies.

2 Related Work

Human Motion Generation. Human motion generation is an open problem in computer vision, covering a broad set of tasks and applications. In recent years, diffusion-based models have emerged as the dominant paradigm (Chen et al., 2023; Tevet et al., 2023; Zhang et al., 2024), with the Motion Diffusion Model (MDM) (Tevet et al., 2023) establishing a simple yet effective denoising framework for text-conditioned motion synthesis. Autoregressive formulations (Shi et al., 2024; Zhao et al., 2025) extend diffusion-based motion synthesis to long-horizon generation by predicting motion segments conditioned on a temporal context window, recursively using the latter frames as the condition for the next segment. Subsequent work extends these paradigms along several axes, beyond isolated body motion. Human-Human interaction methods (Liang et al., 2024; Ruiz-Ponce et al., 2025a;b; Wang et al., 2024) coordinate motion for multiple subjects, Human-Scene interaction methods (Cong et al., 2024; Huang et al., 2023; Jiang et al., 2024; Yi et al., 2024) condition motion in static 3D environments, while Human-Object interaction (Ghosh et al., 2023; Peng et al., 2025; Pi et al., 2025; Ron et al., 2025) targets interactions between the subject and usually one object at a time.

Human-Object Interaction Generation. Progress in HOI generation has been particularly driven by the increasing presence of 4D HOI datasets, which span from small object manipulation datasets (Chao et al., 2021; Taheri et al., 2020; Yang et al., 2022), to interaction with larger everyday objects, including furniture and household devices (Jiang et al., 2023; Li et al., 2023; Bhatnagar et al., 2022) and some with even articulated objects (Fan et al., 2023). These datasets include interactions with a single object of interest at a time, while datasets that address the manipulation of more than two objects are more limited (Kim et al., 2025; Lu et al., 2025; Lv et al., 2024).

Early works on HOI generation (Ghosh et al., 2023; Taheri et al., 2022; 2024) rely on conditional variational autoencoders (CVAEs) to model the distribution of plausible whole-body interactions. Many recent works have adopted denoising diffusion to improve realism, diversity, and conditioning flexibility in HOI synthesis (Diller et al., 2024; Li et al., 2024; Peng et al., 2025; Pi et al., 2025; Ron et al., 2025; Wang et al., 2026; Xu et al., 2025a; Zeng et al., 2025; Zhang et al., 2025; Chen et al., 2026). Among these, CoDA (Pi et al., 2025) and HOIDiNi (Ron et al., 2025) formulate HOI generation as a diffusion noise optimization problem, refining generated motion through contact-aware guidance to produce more plausible whole-body interactions, while CoDA further extends to articulated object manipulation. Additionally, DiffGrasp (Zhang et al., 2025) chooses to generate whole-body grasping motions guided by whole sequence object trajectories. Other works focus on controllability and structure, including CHOIS (Li et al., 2024) for controllable interaction synthesis and Contact-Guided HOI (Diller et al., 2024) which leverages contact priors. Recent large-scale frameworks such as InterAct (Xu et al., 2025a) further push toward generalizable single-object HOI animation.

Beyond purely generative formulations, a parallel line of work formulates HOI as a control problem and learns RL policies that produce physically plausible interactions (Braun et al., 2024; Wang et al., 2023; Wu et al., 2025; Xu et al., 2025b). The aforementioned methods primarily focus on single-object interactions and do not explicitly address the generation of more complex two-object manipulations, leaving a significant gap in modeling coordinated bimanual interactions.

Hand-Object Interaction Generation. A complementary line of research targets hand-centric interaction synthesis, focusing on generating fine-grained grasps. Early approaches such as ManipNet (Zhang et al., 2021) predict dexterous finger motion using explicit hand-object spatial representations. More recent methods use denoising models to generate plausible hand-object interactions (Cha et al., 2024; Christen et al., 2024; Huang et al., 2025; Li et al., 2025b; Ye et al., 2024; Fu et al., 2026). LatentHOI (Li et al., 2025b) learns a latent hand diffusion model for generalizable grasp generation. For more than one object, ManipTrans (Li et al., 2025a) addresses bimanual two-object grasping by transferring dexterous manipulation priors across objects via residual learning. While these methods achieve high-fidelity hand-object contact, they disregard full-body motion synthesis. Dex2HOI bridges this gap by jointly modeling dexterous bimanual contact and full-body motion within a single framework.

3 METHOD

The overview of the proposed **Dex2HOI** is illustrated in Fig. 2. In Section 3.1 we introduce our *hand-relative object representation*, which disentangles object motion into wrist-local and global branches. Section 3.2 then describes our Dual-Stream HOI diffusion and Motion Fusion Network that coordinate per-object streams into a single coherent full-body motion. Finally, in Section 3.3 we introduce our geometry-informed interaction-aware training objective.

3.1 HUMAN AND OBJECT MOTION REPRESENTATIONS

We represent each motion sequence as a temporally aligned multimodal feature vector containing human rotations, human root translation and per-object trajectory representations. For a sequence of length T, the full motion feature at frame t can be defined as:

$$\mathbf{f}_t = \left[\mathbf{h}_t^{\{A,B\}},\ \mathbf{r}_t^{\{A,B\}},\ \mathbf{o}_t^{\{A,B\}}\right], \quad 0 \le t \le T \tag{1}$$

where $\mathbf{h}_t$ denotes the human 6D (Zhou et al., 2019) rotations, $\mathbf{r}_t \in \mathbb{R}^3$ denotes the human root translation, and $\mathbf{o}_t$ the *hand-relative* object representations for the two objects $\{A, B\}$.

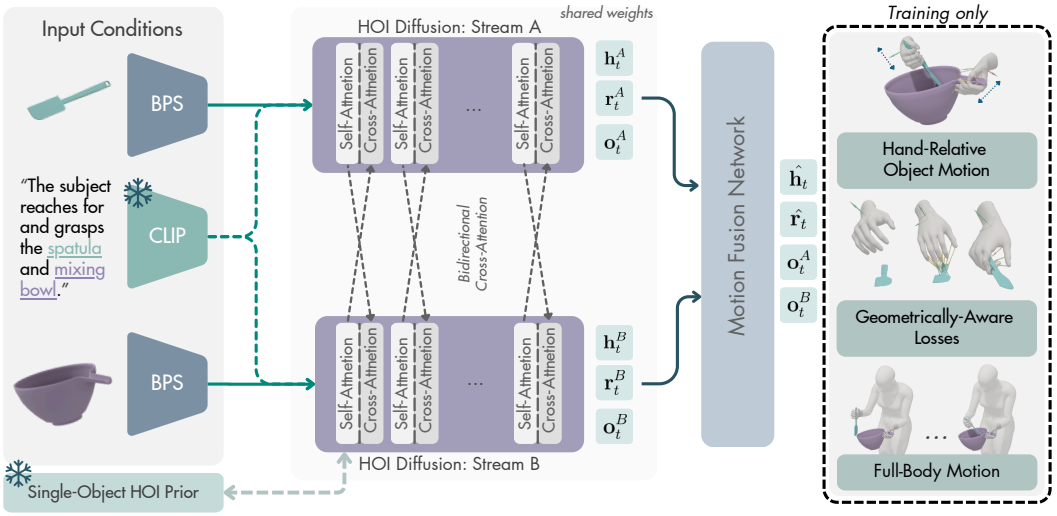

Figure 2: **Dex2HOI** generates dexterous bimanual two-object HOI sequences from text in single-shot, combining (a) Hand-Relative Object Motion Representations, (b) Dual-Stream HOI Diffusion, (c) our Motion Fusion Network, and (d) Geometrically-Aware Losses.

Human Representation. We represent human body poses at frame t as a set of 6D (Zhou et al., 2019) joint rotations $\mathbf{h}_t \in \mathbb{R}^{6J}$, where J denotes the number of joints. We set $J = 52$ for datasets that utilize the SMPL-X parametric human model (Pavlakos et al., 2019) and $J = 65$ for datasets that utilize Mixamo parametric avatars (Adobe Inc.). To facilitate training, we also parameterize the 3D root translation $\mathbf{r}_t$ as the displacement relative to the first frame position (global pelvis bone).

Hand-Relative Object Representation. Modeling object motion in global coordinates alone is insufficient for predicting interaction-dependent trajectories: prior approaches consequently resort to multi-stage pipelines or test-time optimization to recover plausible contact. Relying purely on a global object trajectory provides a weak inductive bias and forces the network to implicitly infer the controlling hand at every frame. To address this, we introduce a *hand-relative object representation* that explicitly disentangles object motion into right-hand, left-hand, and global components. Specifically, for frame t, we define the object feature vector as:

$$\mathbf{o}_t = \left[\mathbf{r}_t^{obj},\ \Delta_t^L,\ \Delta_t^R,\ \mathbf{p}_t^G,\ \boldsymbol{\alpha}_t\right] \in \mathbb{R}^{18}, \tag{2}$$

where $\mathbf{r}_t^{obj} \in \mathbb{R}^6$ denotes the global object rotation in 6D form (Zhou et al., 2019), $\Delta_t^L \in \mathbb{R}^3$ and $\Delta_t^R \in \mathbb{R}^3$ are the object translation expressed in the left and right wrist frames respectively, $\mathbf{p}_t^G \in \mathbb{R}^3$ represents the object translation in global coordinates and $\boldsymbol{\alpha}_t \in \mathbb{R}^3$ are learnable mixture weights that balance the contribution of wrist-relative displacements and global-coordinate translation.

Letting $\mathbf{p}_t^L, \mathbf{p}_t^R \in \mathbb{R}^3$ denote the global wrist positions, and $\mathbf{R}_t^L, \mathbf{R}_t^R \in SO(3)$ the corresponding global wrist rotations obtained from forward kinematics, we define the wrist-relative offsets, relative to the right (R) and left (L) hand:

$$\Delta_t^{\{R,L\}} = (\mathbf{R}_t^{\{R,L\}})^\top (\mathbf{p}_t^{obj} - \mathbf{p}_t^{\{R,L\}}) \tag{3}$$

Intuitively, these terms describe where the object lies in each wrist's local frame. To ensure accurate object translation in cases where the object is not interacting with either hand, we include $\mathbf{p}_t^G$ as a global coordinate fallback. Under this formulation, we define the global object translation through a soft mixture of the three relative representations as:

$$\hat{\mathbf{p}}_t^{obj} = w_t^L \left(\mathbf{p}_t^L + \mathbf{R}_t^L \Delta_t^L\right) + w_t^R \left(\mathbf{p}_t^R + \mathbf{R}_t^R \Delta_t^R\right) + w_t^G \mathbf{p}_t^G, \tag{4}$$

where $[w_t^L, w_t^R, w_t^G] = \mathrm{softmax}(\boldsymbol{\alpha}_t)$ are the learned mixing weights.

Using this formulation, the model learns to smoothly interpolate between left-hand control, right-hand control, and global motion through the learned weights $[w_t^L, w_t^R, w_t^G]$, which adapt at each frame depending on which hand is in contact with the object, instead of forcing all interaction regimes into a single translation branch.

3.2 MOTION MODELING

We formulate human-object motion synthesis as autoregressive conditional denoising over structured motion sequences, enabling the generation of arbitrarily long sequences (Ron et al., 2025; Tevet et al., 2025). Given an observed prefix of length T_p, the AR model D_θ predicts the future segment $\mathbf{F}_{T_p:T_p+T_f-1}$, with a prediction length of T_f frames.

Dual-Stream HOI Diffusion $\mathbf{D}_\theta$. To facilitate motion generation for multiple objects, we adopt a cooperative denoising paradigm (Liang et al., 2024) with two coordinated diffusion streams A and B with shared weights, one for each object-conditioned interaction. Stream A carries motion representation $\mathbf{f}_t^A = [\mathbf{h}_t, \mathbf{r}_t, \mathbf{o}_t^A]$ and stream B carries $\mathbf{f}_t^B = [\mathbf{h}_t, \mathbf{r}_t, \mathbf{o}_t^B]$, where both share the human representation $[\mathbf{h}_t, \mathbf{r}_t]$ but differ in their object representation. At diffusion timestep τ, the shared denoiser D_θ processes each stream conditioned on the other:

$$\hat{\mathbf{f}}_0^A = D_\theta\big(\mathbf{f}_\tau^A, \mathbf{f}_\tau^B, \mathbf{c}^A, \tau\big), \qquad \hat{\mathbf{f}}_0^B = D_\theta\big(\mathbf{f}_\tau^B, \mathbf{f}_\tau^A, \mathbf{c}^B, \tau\big), \tag{5}$$

where $\mathbf{c}^A$ and $\mathbf{c}^B$ are branch-specific conditioning signals containing: a frozen CLIP (Radford et al., 2021) text embedding; object geometry encoded as Basis Point Set (BPS) features (Prokudin et al., 2019); and a lightweight MLP projection for τ. To enable information flow between the two HOI streams, we follow (Liang et al., 2024) and apply bidirectional cross-attention $\mathrm{CA}(\cdot)$ after each layer ℓ of the shared HOI Diffusion D_θ, resulting in each stream effectively attending to the other.

Motion Fusion Network $\mathbf{D}_\mathbf{h}$. At the output of D_θ, both streams produce a complete prediction $\hat{\mathbf{f}}_0^A$ and $\hat{\mathbf{f}}_0^B$, each containing full motion features with human representation and hand-relative object representation per object present. We decompose each prediction along the feature partition introduced in Section 3.1:

$$\hat{\mathbf{f}}_0^A = \big[\,\hat{\mathbf{h}}^A,\ \hat{\mathbf{r}}^A,\ \hat{\mathbf{o}}^A\,\big], \qquad \hat{\mathbf{f}}_0^B = \big[\,\hat{\mathbf{h}}^B,\ \hat{\mathbf{r}}^B,\ \hat{\mathbf{o}}^B\,\big]. \tag{6}$$

To produce a single coherent human pose, derived from the separate HOI Diffusion streams $[\hat{\mathbf{h}}^A, \hat{\mathbf{r}}^A]$ and $[\hat{\mathbf{h}}^B, \hat{\mathbf{r}}^B]$, we introduce a dedicated *Motion Fusion Network* $D_\mathbf{h}$, implemented as a stack of self-attention layers that operates *only* on the human representations (Figure 2):

$$\big[\hat{\mathbf{h}},\ \hat{\mathbf{r}}\big] = D_\mathbf{h}\Big(\big[\hat{\mathbf{h}}^A, \hat{\mathbf{r}}^A\big],\ \big[\hat{\mathbf{h}}^B, \hat{\mathbf{r}}^B\big]\Big). \tag{7}$$

The fused body prediction $[\hat{\mathbf{h}}, \hat{\mathbf{r}}]$ replaces the per-stream body outputs, while the object predictions $\hat{\mathbf{o}}^A$ and $\hat{\mathbf{o}}^B$ are taken directly from their originating streams without modification. This design enforces a clean factorization of the generation process: each object's dynamics are denoised under its own object-specific conditioning, while the body is reconciled in a shared latent that accesses both interactions. As we demonstrate through ablations in Section 4.3, the Motion Fusion Network outperforms simpler alternatives such as averaging or selecting one of the two human predictions, which often lead to inconsistent motions.

3.3 TRAINING

Since our HOI diffusion model solely predicts rotation parameters for the human body, we apply a differentiable forward kinematics layer $\mathrm{FK}(\cdot)$ to map predicted poses to world-space joint positions, thus enabling velocity and contact losses that are otherwise inaccessible in the rotation domain. Along with the main diffusion objective $\mathcal{L}_{diff}$ and the well-established geometry regularization and SDF losses (Liang et al., 2024; Pi et al., 2025; Ron et al., 2025), which cover joint and position velocity (supervision against target), and interpenetrations (body-object and inter-object SDFs), we also enhance our training objective by introducing the following contact-aware conditioning:

Hand-relative object representation supervision. Our hand-relative object representation $\mathbf{o}_t$ induces a structured decomposition of object motion into wrist-relative and global branches, according to Eq. 4. We explicitly supervise the different components. First, we supervise the mixture weights by comparing the predicted weights $[\hat{w}_t^L, \hat{w}_t^R, \hat{w}_t^G] = \mathrm{softmax}(\hat{\boldsymbol{\alpha}}_t)$, extracted from the predicted representation $\hat{\mathbf{o}}t$, against hand proximity-derived ground-truth targets $[w_t^L, w_t^R, w_t^G]$:

$$\mathcal{L}_{mix} = \|\hat{w}_t^s - w_t^s\|_2^2, \quad s \in \{L, R, G\}. \tag{8}$$

Second, we supervise the decoded object translation $\hat{\mathbf{p}}_t^{obj}$, obtained via the mixture decoder of Eq. 4, against the ground-truth world-space position $\mathbf{p}_t^{obj}$:

$$\mathcal{L}_{trans} = \left\| \hat{\mathbf{p}}_t^{obj} - \mathbf{p}_t^{obj} \right\|_2^2, \tag{9}$$

enforcing correct trajectory reconstruction while ensuring gradients propagate through all branches and mixture weights. Finally, we regularize the predicted wrist-relative offsets $\hat{\Delta}_t^L$ and $\hat{\Delta}_t^R$, as defined in Eq. 3, against the ground-truth Δ_t^L and Δ_t^R, further encouraging geometric consistency:

$$\mathcal{L}_{off} = \frac{1}{2}\left(\left\| \hat{\Delta}_t^L - \Delta_t^L \right\|_2^2 + \left\| \hat{\Delta}_t^R - \Delta_t^R \right\|_2^2 \right). \tag{10}$$

Hand-Object contact regression. We enforce fine-grained interaction geometry via a unified finger-object distance loss. Letting $d_{t,j}$ denote the ground-truth signed nearest-surface distance from hand joint j at frame t to the object mesh, and $\hat{d}_{t,j}$ the predicted counterpart, we regress nearest-surface distances collectively via Huber loss ρ_δ:

$$\mathcal{L}_{dist} = \frac{1}{|\mathcal{A}|} \sum_{(t,j)\in\mathcal{A}} \rho_\delta \left(\hat{d}_{t,j} - d_{t,j} \right), \tag{11}$$

where $\mathcal{A} = (t, j) : d_{t,j} < d_{min}$ selects frames where hand joint j lies within approach radius $d_{min} = 0.10$ of the object surface. The full training objective is additionally discussed in the Appendix.

4 Evaluation

4.1 Datasets

To train and evaluate Dex2HOI we use the following datasets:

GRAB. A dataset (Taheri et al., 2020) with 10 subjects interacting with 51 everyday objects. We follow the typical split (Ron et al., 2025; Pi et al., 2025) using subjects 1 through 9 for training and subject 10 (145 sequences) as the test set.

HUMOTO. A mo-cap dataset (Lu et al., 2025) that contains both single and multi-object manipulation tasks in Mixamo representations (Adobe Inc.). We retain clips that contain up to two moving objects, splitting according to the provided annotations at a frame level. The resulting benchmark yields 337 train and 45 test sequences, mostly from tabletop manipulation scenes; the broader 1/2-object subset used for one-object training and evaluator training contains 1828 train and 322 test sequences.

HIMO. A multi-object interaction dataset (Lv et al., 2024) covering simultaneous manipulation of 2 or 3 objects. The 2-object split contains 2578 train and 522 test sequences, and the 3-object split contains 1045 train and 176 test sequences, following the official splits. In this work, we use the 2-object split to evaluate our work. More about the datasets can be found in the Appendix.

4.2 Motion Evaluators and Metrics

To evaluate the generated motions on the HUMOTO interactions, we trained a dedicated transformer-based motion evaluator on ground-truth motions of the human joints and of the two objects, following the protocol in (Ghosh et al., 2023). The motion encoder is used to compute FID↓, Diversity→, and MultiModality (MM ↑ / MModality ↑) based on extracted motion embeddings and assess distributional quality, while for text-alignment metrics MultiModal Distance (MMdist ↓) and R-Precision (R@3 ↑), we train an additional motion-text aligner using a contrastive objective. Finally, the penetration depth metric measures the mean hand and body–object interpenetration distance across frames.

To evaluate on GRAB we follow the existing protocol (Ghosh et al., 2023; Ron et al., 2025). In contrast to the HUMOTO evaluator that is trained using a contrastive learning objective, the GRAB evaluator serves as an action-recognition network, since GRAB comes with discrete action labels (Taheri et al., 2020). We report the same six metrics across all GRAB experiments, with the difference that MultiModality is calculated between generations of the same *action* labels instead of instruction prompts, following (Ghosh et al., 2023).

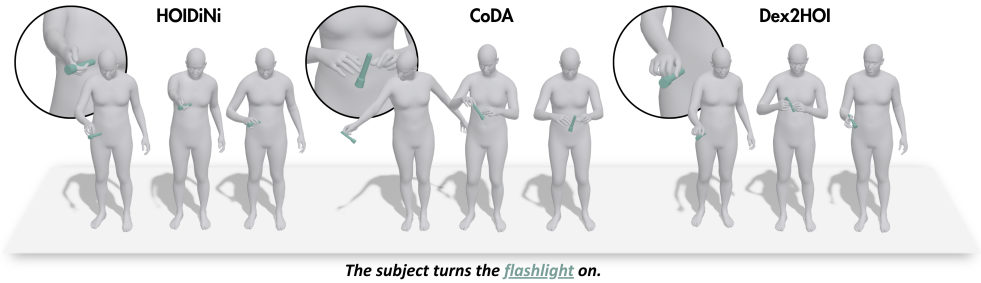

Figure 3: **Dex2HOI** comparative qualitative results on GRAB (single-object HOI synthesis).

	Method	FID ↓	Diversity →	MM →	Pen. (mm) ↓	R@3 ↑	MMdist ↓	Runtime (s) ↓
	GT	—	0.706	0.194	—	0.617	—	—
1 object	MDM	0.514	0.674	0.186	5.85	0.570	1.137	0.56±0.03
	IMoS (w/o obj optim)	0.521	0.608	0.207	6.23	0.557	1.165	0.75±0.05
	IMoS	0.494	0.632	0.189	7.02	0.557	1.165	59.22±2.15
	HOIDiNi (w/o DNO)	0.503	0.715	0.208	10.88	0.587	**1.126**	0.43±0.02
	HOIDiNi	0.584	0.783	0.183	7.34	0.528	1.179	122.70±3.41
	CoDA (w/o DNO)	0.697	0.483	**0.189**	7.39	0.567	1.369	2.79±0.08
	CoDA	0.558	0.694	0.249	**5.36**	0.592	1.446	135.26±12.7
	Dex2HOI	**0.479**	**0.702**	0.161	5.70	**0.593**	1.222	**0.25**±0.01

Table 1: Single-object motion evaluation on the GRAB dataset (Taheri et al., 2020): Dex2HOI provides a *favorable* trade-off across realism, multimodality and action retrieval metrics, while maintaining quality hand-object contacts and without utilizing test-time optimization (Karunratanakul et al., 2024). In the last column we report the average time it takes to infer a 60-frame sample. We highlight **best** and second best results and (→) indicates results closer to the ground-truth values are better. Results across all tables are averaged over 10 independent seeded runs.

4.3 Experiments

Baselines. On the two-object benchmarks HUMOTO and HIMO, we compare **Dex2HOI** against the HIMO-Gen generative baseline (Lv et al., 2024) and an MDM baseline (Tevet et al., 2023), adapted to jointly predict object and human motion. Given the limited availability of multi-object evaluation benchmarks and to also evaluate our separate HOI Diffusion Models D_θ, we perform additional evaluation on the single-object against previous state-of-the-art full-body HOI methods, namely HOIDiNi (Ron et al., 2025), CoDA (Pi et al., 2025) and IMoS (Ghosh et al., 2023) and MDM (Tevet et al., 2023). Apart from MDM, these methods employ some form of hand-object contact informed optimization at inference, producing plausible results at an additional computational cost. Adapting these optimization frameworks to two-object sequences is a non-trivial task; Their contact objectives are formulated around a single object–hand pair, with no mechanism to resolve which hand should contact which object at each timestep. Naively running per-object optimization in parallel without coordination signals risks yielding inter-object penetrations and incoherent motion.

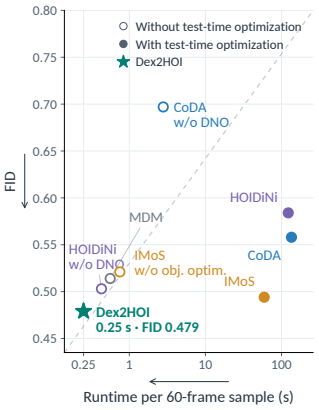

Figure 4: Optimization Pareto Diagram - lower FID and faster runtime is better.

Quantitative Results. As shown in Table 1, **Dex2HOI** achieves the best FID, Diversity and R@3 on GRAB while running up to ×540 times faster on average than CoDA and ×490 times than HOIDiNi, depending on the selected duration of the iterative noise optimization (we use 200 opt. steps on both methods over 60-frame predictions). Notably, our single-shot generation surpasses optimization-based methods across standardized metrics, without the exhaustive runtime requirements - which is also summarized in the Pareto analysis in Figure 4. Similarly, on the HUMOTO dataset (Table 2), **Dex2HOI** achieves the best FID, penetration, and MModality across both single- and two-object settings, with a great gain over prior methods particularly on single-object instances.

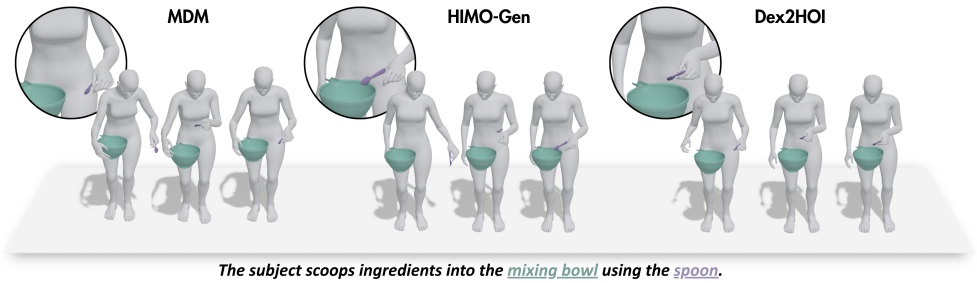

Figure 5: Qualitative contact comparison against HIMO-Gen and MDM on HUMOTO.

	Method	FID ↓	Diversity →	MModality ↑	Pen. (mm) ↓	R-Pre@3 ↑	MMdist ↓
2 objects	GT	—	1.061	—	—	0.520	1.210
	MDM (Tevet et al., 2023)	0.896	**0.797**	0.009	12.8	0.118	1.420
	HIMO-Gen (Lv et al., 2024)	0.886	0.606	0.019	11.7	**0.295**	1.281
	Dex2HOI	**0.655**	0.780	**0.056**	**9.8**	0.275	**1.260**
1 object	GT	—	1.192	—	—	0.665	1.250
	MDM (Tevet et al., 2023)	1.394	0.464	0.053	18.2	0.086	1.445
	HIMO-Gen (Lv et al., 2024)	1.418	0.413	0.008	20.7	0.188	1.410
	Dex2HOI	**0.880**	**0.848**	**0.103**	**13.1**	**0.250**	**1.323**

Table 2: **Quantitative evaluation on the HUMOTO dataset (Lu et al., 2025)**, reporting performance on both two-object (top) and single-object (bottom) settings. We compare **Dex2HOI** against HIMO-Gen (Lv et al., 2024) and MDM (Tevet et al., 2023) and highlight **best** and second-best results.

Ablation Study. We design our ablation study around the core components of Dex2HOI and we provide the following variants for two-object evaluation: **(i) w/o the Motion Fusion Network**, where we replace the learned $D_{\mathbf{h}}(\cdot)$ model with an element-wise average of the two stream body predictions. **(ii) w/o Dual-stream HOI**, where we use a single denoising stream that jointly predicts human motion together with both objects. **(iii) w/o hand-relative object representations**, where we replace the proposed representation with global object trajectory prediction. **(iv) w/o mixture supervision**, where we remove the $\mathcal{L}_{\text{mix}}$, $\mathcal{L}_{\text{trans}}$, and $\mathcal{L}_{\text{off}}$ terms, leaving only the diffusion MSE loss on the hand-relative object representation. This tests whether the hand-relative representation can be learned implicitly, or whether weight supervision is required. **(v) w/o finger-contact supervision $\mathcal{L}_{\text{dist}}$**, where we remove our contact regression loss from the learning objective. Finally, **(vi) GRAB single prior**, where the diffusion layers are initialized from frozen single-object GRAB training, and only cross-attention and D_h are learned from two-object data. On GRAB, we further ablate the following: (i) hand-relative representations, (ii) mixture supervision, and (iii) L_{dist}.

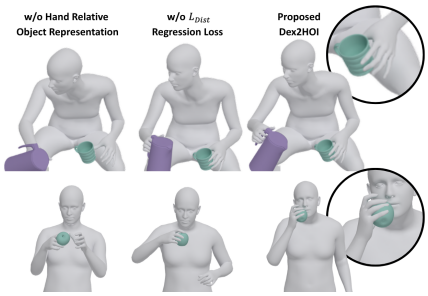

Figure 6: **Representation and contact ablations.** Hand-relative representations couple object motion to the controlling wrists, while L_{dist} supervises fine-grained hand-object contact.

Ablation Results. Table 3 shows that the proposed gains do not arise from simply scaling a motion diffusion model to additional output dimensions. Replacing the hand-relative representation with global object trajectories causes the largest degradation, while removing the dual-stream formulation or learned Motion Fusion consistently reduces generation quality, supporting explicit object-wise modeling and learned coordination. Mixture supervision further improves both motion and interaction quality, indicating that explicitly modeling transitions between left-hand, right-hand, and global object control is beneficial. Notably, removing L_{dist} leaves FID nearly unchanged ($0.655 \rightarrow 0.656$) but increases penetration from 9.8 to 11.6 mm, showing that it specifically improves local interaction geometry rather than global motion statistics. Qualitative results are also shown in Figures 3 and 5 for GRAB and HUMOTO respectively, while Figure 6 qualitatively illustrates the effect of the proposed representation and contact supervision. Finally, although **Dex2HOI** is designed for single-shot generation, we additionally provide an **iterative refinement** ablation for applications that prioritize

contact fidelity over inference speed. We use a deterministic sequence-level training-free optimizer which solves for a gated per-frame correction to the elbow, wrist and finger rotations of each hand under a 5 mm contact gate. The refinement results in minor finger-distance corrections while leaving the body, root trajectory, and object trajectory unchanged. This procedure is strictly post-hoc and not used in **any visual result in this paper**, unless otherwise specified in the Appendix.

	Method	FID ↓	Diversity →	MModality ↑	Pen. (mm) ↓	R-Pre@3 ↑	MMdist ↓
2 objects	GT	—	1.061	—	—	0.520	1.210
	w/o Fusion (avg layer)	0.758	0.761	0.055	10.9	0.271	1.255
	w/o Dual-stream HOI	0.885	0.638	0.025	11.4	0.243	**1.255**
	w/o hand-relative obj reps	0.950	0.514	0.026	13.8	0.262	1.281
	w/o mixture supervision	0.831	0.586	0.033	11.8	0.167	1.294
	w/o contact regression L_{Dist}	0.656	0.749	0.063	11.6	0.274	1.260
	w/ GRAB single prior	**0.576**	0.766	**0.109**	11.4	0.275	1.259
	Proposed	0.655	**0.780**	0.056	**9.8**	**0.275**	1.260
	+ contact refinement (Δ)	*-0.001*	*+0.012*	*+0.000*	*-0.3*	*+0.000*	*-0.000*
1 object	GT	—	0.706	0.194	—	0.617	1.175
	w/o hand-relative obj reps	0.621	0.651	0.149	10.9	0.437	1.346
	w/o mixture supervision	0.596	0.677	0.159	9.9	0.466	1.261
	w/o contact regression L_{Dist}	0.570	0.679	0.154	6.6	0.421	1.374
	Proposed	**0.479**	**0.702**	**0.161**	**5.7**	**0.593**	**1.222**
	+ contact refinement (Δ)	*-0.000*	*+0.011*	*+0.000*	*-1.3*	*+0.000*	*+0.005*

Table 3: **Ablation study on HUMOTO (2 objects) and GRAB (1 object).** Removing the Hand-Relative object representations and other key model parts consistently degrades performance, confirming that both components are essential. The GRAB-pretrained variant remains competitive, indicating that grasping priors transfer effectively from SMPL-X (GRAB) to Mixamo characters.

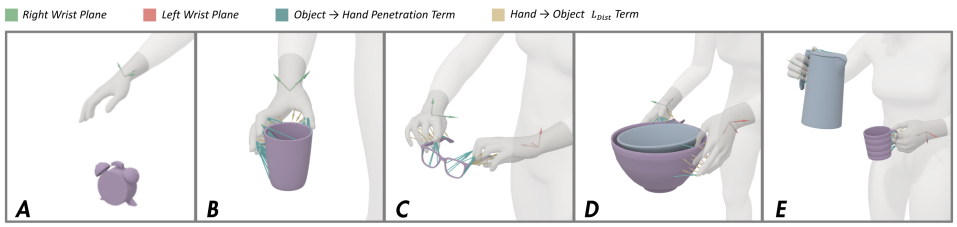

Figure 7: Dex2HOI's hand-relative object representation models object translation relative to the wrists while retaining a global object component, allowing our unified model to represent diverse hand–object interactions. **(A) No-contact**: the object remains spatially separated from both hands, using global object motion fallback. **(B) Single-hand contact:** the object is primarily constrained relative to a single wrist, supporting one-handed grasping. **(C) Bimanual handling** of a single object and **(D)** of multiple objects, and **(E) Independent two-object manipulation**.

Perceptual Assessment. We complement metric-based evaluation with VLM and human perceptual studies assessing three properties central to dexterous HOI: motion naturalness, hand–object contact quality and text prompt alignment. For the VLM evaluation, we query InternVL3-8B to evaluate rendered while, separately, participants compare anonymized side-by-side renderings of competing methods. Full protocols, prompts, and results are provided in the Appendix.

5 Conclusion

We presented **Dex2HOI**, a unified framework for human-object interaction generation that extends beyond conventional single-object HOI to coordinated bimanual dexterous manipulation of up to two objects. We model multi-object interactions through dual denoising streams, coupled via bidirectional cross-attention and fused into a single coherent full-body motion. By modeling object motion through our hand-relative representation, we unify single-hand and bimanual manipulation within the same formulation, producing plausible hand-object interactions in single-shot inference, with no additional contact optimization. Our extensive evaluation shows that Dex2HOI generalizes across both single-object and multi-object settings, achieving *SOTA* results and improved motion realism.

REFERENCES

Adobe Inc. Adobe mixamo. https://www.mixamo.com/.

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

Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *International Conference on 3D Vision (3DV)*, 2024.

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

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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)*, 2023.

Yuanpei Chen, Chen Wang, Yaodong Yang, and Karen Liu. Object-centric dexterous manipulation from human motion data. In *Conference on Robot Learning (CoRL)*.

Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, Jiankang Deng, Cordelia Schmid, and Stefanos Zafeiriou. Ho-flow: Generalizable hand-object interaction generation with latent flow matching. *arXiv preprint arXiv:2604.10836*, 2026.

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 Conference Papers*, 2024.

Peishan Cong, Ziyi Wang, Zhiyang Dou, Yiming Ren, Wei Yin, Kai Cheng, Yujing Sun, Xiaoxiao Long, Xinge Zhu, and Yuexin Ma. Laserhuman: language-guided scene-aware human motion generation in free environment. *arXiv preprint arXiv:2403.13307*, 2024.

Christian Diller et al. Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jinguo Zhu et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

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

Hongming Fu, Wenjia Wang, Xiaozhen Qiao, Rolandos Alexandros Potamias, Taku Komura, Shuo Yang, Zheng Liu, and Bo Zhao. Egograsp: World-space hand-object interaction estimation from egocentric videos. *arXiv preprint arXiv:2601.01050*, 2026.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Eurographics*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, Kris Kitani, Matt Feiszli, and Hao Tang. Hoigpt: Learning long sequence hand-object interaction with language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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. *arXiv preprint arXiv:2212.10621*, 2023.

Nan Jiang et al. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia Conference Papers*, 2024.

Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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

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

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision (ECCV)*, 2024.

Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 2024.

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

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

Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. In *International Conference on 3D Vision (3DV)*, 2024.

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

Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. HOI-Diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025.

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

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Roey Ron, Guy Tevet, Haim Sawdayee, and Amit H. Bermano. Hoidini: Human-object interaction through diffusion noise optimization. *arXiv preprint arXiv:2506.15625*, 2025.

Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. Mixermdm: Learnable composition of human motion diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Pablo Ruiz-Ponce, Sergio Escalera, José García-Rodríguez, Jiankang Deng, and Rolandos Alexandros Potamias. Interact2ar: Full-body human-human interaction generation via autoregressive diffusion models. *arXiv preprint arXiv:2512.19692*, 2025b.

Yi Shi, Jingbo Wang, Xuekun Jiang, Bingkun Lin, Bo Dai, and Xue Bin Peng. Interactive character control with auto-regressive motion diffusion models. *ACM Transactions on Graphics*, 2024.

Kewei Sui, Anindita Ghosh, Inwoo Hwang, Bing Zhou, Jian Wang, and Chuan Guo. A survey on human interaction motion generation. *International Journal of Computer Vision*, 2026.

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

Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J. Black. Grip: Generating interaction poses using spatial cues and latent consistency. In *International Conference on 3D Vision (3DV)*, 2024.

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

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit Haim Bermano, and Michiel van de Panne. CLoSD: Closing the loop between simulation and diffusion for multi-task character control. In *International Conference on Learning Representations (ICLR)*, 2025.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023.

Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. Intercontrol: Zero-shot human interaction generation by controlling every joint. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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 *International Conference on Learning Representations (ICLR)*, 2026.

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

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liangyan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics*, 2021.

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

Yonghao Zhang, Qiang He, Yanguang Wan, Yinda Zhang, Xiaoming Deng, Cuixia Ma, and Hongan Wang. Diffgrasp: Whole-body grasping synthesis guided by object motion using a diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *International Conference on Learning Representations (ICLR)*, 2025.

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

APPENDIX

In the **supplementary video** provided, we include a brief presentation of our method, contributions and evaluation, along with detailed **videos** of the figures shown in the paper. **Generated motions are best viewed in video form.** Training and evaluation code along with checkpoints will get released upon publication.

In this Appendix, we provide additional details to complement the main paper: Section 7 reports a full implementation breakdown, including architectural and training objective details. Section 8 presents *further evaluation*, with extended results on GRAB (Taheri et al., 2020) under the LatentHOI (Li et al., 2025b) physics protocol and additional quantitative and qualitative comparisons on HIMO (Lv et al., 2024) and HUMOTO (Lu et al., 2025) datasets. We also show examples of unseen object geometry adaptations and hand dexterity evaluation along with examples of multi- object HOI generation. Section 10 details our *perceptual study analysis*, VLM and User based. We close with a discussion of limitations and future directions in Section 11.

6 MOTION GENERATION WITH DIFFUSION MODELS - BACKGROUND

Diffusion-based Motion Generation. Denoising diffusion models formulate motion generation as the iterative reversal of a Gaussian corruption process. Given a clean motion sequence $x_0 \in \mathbb{R}^{T\times D}$, the forward diffusion process progressively perturbs the motion with Gaussian noise according to

$$q(x_\tau \mid x_0) = \mathcal{N}\left(x_\tau; \sqrt{\bar{\alpha}_\tau} x_0, (1-\bar{\alpha}_\tau) I\right), \tag{12}$$

or equivalently,

$$x_\tau = \sqrt{\bar{\alpha}_\tau}\, x_0 + \sqrt{1-\bar{\alpha}_\tau}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{13}$$

where $\tau \in \{1, \ldots, N\}$ denotes the diffusion timestep, $\alpha_\tau = 1 - \beta_\tau$, and $\bar{\alpha}_\tau = \prod_{s=1}^{\tau} \alpha_s$. As τ increases, the structured motion x_0 is gradually transformed toward Gaussian noise.

The reverse process learns a conditional denoiser D_θ that reconstructs the clean motion,

$$\hat{x}_0 = D_\theta(x_\tau, c, \tau), \tag{14}$$

where c denotes the generation conditions - in our case text and object geometry. Following common motion diffusion formulations, the model is trained by sampling a diffusion timestep τ and noise ϵ, constructing x_τ according to Eq. 13, and minimizing the reconstruction objective

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0,\tau,\epsilon}\left[\|x_0 - D_\theta(x_\tau, c, \tau)\|_2^2\right]. \tag{15}$$

Conceptually, the diffusion process can be summarized as:

$$\boxed{x_0} \xrightarrow[\text{forward diffusion}]{+\epsilon} \boxed{x_\tau} \xrightarrow[\text{text + BPS}]{D_\theta} \boxed{\hat{x}_0}. \tag{16}$$

In **Dex2HOI**, this formulation is extended to two coordinated object-specific diffusion streams. Each stream denoises the interaction between the human and one object, while bidirectional cross-attention layers enable information exchange between the streams:

$$\boxed{x_\tau^A} \xleftrightarrow[\text{cross-attention}]{} \boxed{x_\tau^B} \xrightarrow{D_\theta} \left\{ \begin{array}{c} [\hat{h}^A, \hat{r}^A, \hat{o}^A] \\ [\hat{h}^B, \hat{r}^B, \hat{o}^B] \end{array} \right\}. \xrightarrow{D_h} \boxed{[\hat{h}, \hat{r}, \hat{o}^A, \hat{o}^B]}. \tag{17}$$

The two streams independently model object-specific interactions while sharing information throughout denoising. The Motion Fusion Network D_h unifies the two predicted human representations into a single body motion, while retaining the object motion predicted by each corresponding stream (only the body pose gets fused). To further support our unified one and two object training regime, when only one object is present the remaining stream and Motion Fusion Network D_h are masked and not used for prediction.

7 IMPLEMENTATION DETAILS

Implementation details. Our two-object model uses a shared Transformer decoder with 8 layers, together with 8 cross-attention layers. All decoder and cross-attention blocks use a latent dimension of 512, 4 attention heads, and a feed-forward dimension of 1024. Object conditioning is provided by a shared BPS encoder (Prokudin et al., 2019). For the Motion Fusion Network, the two body streams h_A, h_B are represented as a 2-token sequence at each frame and fused with a lightweight self-attention head using 4 attention heads; For HOI diffusion D_θ, the model is trained with 10 diffusion steps, a cosine noise schedule, and fixed-small variance, and inference uses DDIM-10 sampling. Training uses EMA and AdamW with a learning rate of 1e-4 and 50-100k training steps, depending on the dataset. All training and evaluation experiments were executed between two NVIDIA GeForce RTX 4090 GPUs.

Classifier-Free Guidance. We apply classifier-free guidance (CFG) (Ho & Salimans, 2022) only to the text input conditions. During training, the encoded text tokens are randomly zeroed with probability p_{drop}, while all geometric and temporal inputs remain unchanged. At sampling time, we run the denoiser twice: once with the full text prompt and once with the text embedding masked out. The two predictions are combined using the guidance scale w, which amplifies the direction induced by the prompt. Importantly, the unconditional branch keeps the same object point-cloud embeddings, object masks, prefix frames, initial transforms, and optional root/object x_0 position conditions as the conditional branch. Thus CFG improves semantic prompt alignment while preserving the scene geometry and first-frame constraints, without adding extra networks or training objectives.

Training Objective Analysis. In Section 3.3, we define our full training objective as:

$$\mathcal{L} = \mathcal{L}_{diff} + \mathcal{L}_{pos} + \mathcal{L}_{vel} + \underbrace{\lambda_w \mathcal{L}_{mix} + \lambda_t \mathcal{L}_{trans} + \lambda_o \mathcal{L}_{off}}_{\textit{object representation supervision}} + \lambda_p \mathcal{L}_{dist} + \lambda_{pen} \mathcal{L}_{pen} \tag{18}$$

Velocity and Position Supervision. Beyond the denoising objective $\mathcal{L}_{diff}$ we add first-order velocity matching losses to suppress temporal jitter and enforce smooth dynamics across the generated motion. We compute the following geometric losses through forward kinematics, which converts the predicted body poses (6D rotations) to 3D joint positions.

Joint Position Loss. The global joint position loss $\mathcal{L}_{pos}$ penalises discrepancies in the predicted 3D locations of body joints:

$$\mathcal{L}_{pos}(\mathbf{p}, \hat{\mathbf{p}}) = \|\mathbf{p} - \hat{\mathbf{p}}\|_2^2, \tag{19}$$

where $\mathbf{p} \in \mathbb{R}^{T \times J \times 3}$ are the ground-truth global joint positions with J joints, and $\hat{\mathbf{p}}$ are the corresponding predicted positions. This loss enforces spatial accuracy in the generated motions.

Joint Velocity Loss. To promote temporal smoothness and physical plausibility, we apply a velocity loss between the ground truth $\mathbf{v}$ and predicted $\hat{\mathbf{v}}$ joint positions:

$$\mathcal{L}_{vel}^{J}(\mathbf{v}, \hat{\mathbf{v}}) = \|\mathbf{v} - \hat{\mathbf{v}}\|_2^2, \tag{20}$$

where $\mathbf{v}_t = \mathbf{p}_t - \mathbf{p}_{t-1}$ represents the joint velocities computed as the difference between consecutive frames. This loss discourages unnatural jittering and encourages smooth, realistic motion trajectories.

Hand-Object Penetration Loss. To penalize fingertip interpenetration with the object surface, we apply a penalty on the signed distance field evaluated at predicted fingertip positions $\hat{\mathbf{h}}$:

$$\mathcal{L}_{pen}^{HO}(\hat{\mathbf{h}}) = \max\big(0,\ -\mathrm{SDF}(\hat{\mathbf{h}})\big), \tag{21}$$

where $\mathrm{SDF}(\cdot)$ returns positive values in free space and negative values inside the object. This loss fires only when a fingertip penetrates the object surface, pushing predicted joints back to the exterior.

Body-Object Penetration Loss. An analogous penalty is applied to the full body mesh vertices $\mathbf{v}$ to prevent torso and limb collisions with the object:

$$\mathcal{L}_{pen}^{BO}(\mathbf{v}) = \max\big(0,\ -\mathrm{SDF}(\mathbf{v})\big). \tag{22}$$

Object-Object Penetration Loss. To address the possibility of penetration between objects - when simultaneous manipulation is involved - the same criterion is applied between the objects themselves. Let $\mathcal{P}_k$ denote a set of points sampled on the surface of object k, expressed in world coordinates under its predicted pose. Then the penetration penalty is symmetric in the two objects:

$$\mathcal{L}_{pen}^{OO} = \frac{1}{|\mathcal{P}_b|} \sum_{\mathbf{p}\in\mathcal{P}_b} \max\big(0,\, -\mathrm{SDF}_a(\mathbf{p})\big) \;+\; \frac{1}{|\mathcal{P}_a|} \sum_{\mathbf{p}\in\mathcal{P}_a} \max\big(0,\, -\mathrm{SDF}_b(\mathbf{p})\big), \tag{23}$$

where $\mathrm{SDF}_k(\cdot)$ is the signed distance field of object k.

Empirically, in our experiments we set the object-representation supervision lambdas as $\lambda_w = \lambda_o = 0.5$, the object translation weight $\lambda_t = 1.0$, while for the position and velocity losses we use $\lambda_{pos} = 1.0$ and $\lambda_{vel} = 1.0$ respectively. Finally, we use a penetration weight of $\lambda_{pen} = 1.0$ across penetration terms, and $\lambda_p = 2.0$ for our hand-object contact regression loss.

Ground-truth mixture weights. The ground-truth targets $[w_t^L, w_t^R, w_t^G]$ are derived from hand proximity to the object at each frame. For each hand $s \in \{L, R\}$ we measure the Euclidean distance d_t^s from the wrist center to the object center, then compute gated contact score:

$$\phi_t^s = \sigma(r_c - d_t^s)\,, \qquad s \in \{L, R\}, \tag{24}$$

where $r_c = 0.15\,\mathrm{m}$ is the contact radius ($\phi_t^s \approx 1$ when the hand is within r_c, $\phi_t^s \approx 0$ otherwise). The global score captures frames where neither hand controls the object: $\phi_t^G = (1 - \phi_t^L)(1 - \phi_t^R)$. The three scores are ℓ^1-normalised to obtain the final weights:

$$\big[w_t^L,\, w_t^R,\, w_t^G\big] = \frac{1}{\phi_t^L + \phi_t^R + \phi_t^G} \big[\phi_t^L,\, \phi_t^R,\, \phi_t^G\big]. \tag{25}$$

When only the left hand is near the object $w^L \approx 1$; when both hands are near $w^L \approx w^R \approx 0.5$; and when neither hand is close $w^G \approx 1$, delegating object position to the free global branch. Figure 9 shows the **mean** weight regression across training on GRAB (top) and HUMOTO (bottom).

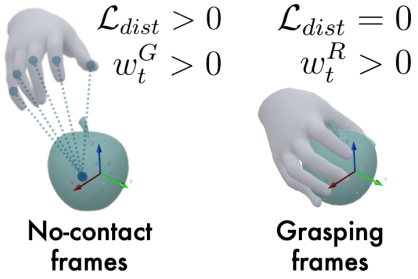

Figure 8: Object Motion Paradigm; No-contact frames have positive $\mathcal{L}_{Dist}$ and use w_t^G. Contact frames have $w_t^G \approx 0$.

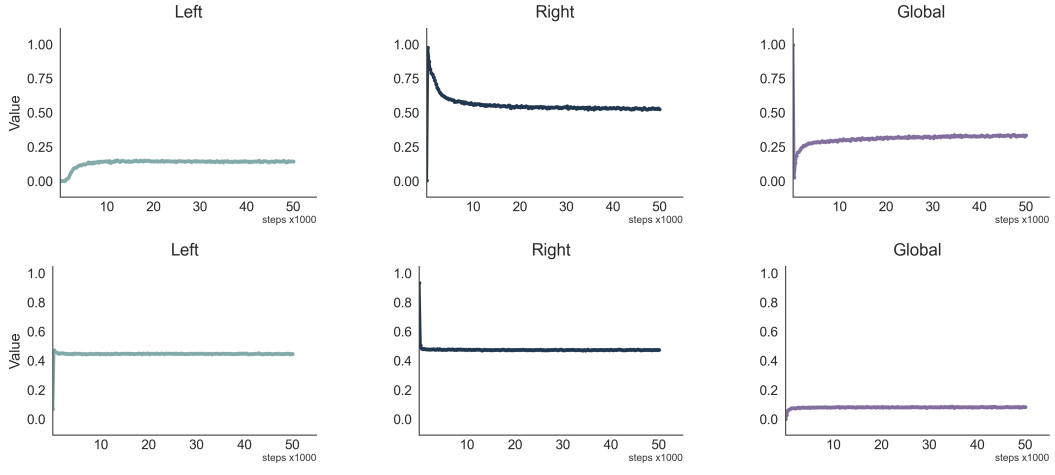

Figure 9: Average predicted mixture weights (w) over training for the Hand-Relative Object Representation. Left/right denote the hand-conditioned object branches, and global denotes the global branch for GRAB (top) and HUMOTO (bottom) datasets.

8 Further Evaluation

Metrics Analysis. As in the main paper, we evaluate generated motions along three axes: distributional realism, text–motion alignment, and physical plausibility. Namely, FID measures the distance

between the distribution of generated and real motion embeddings. Diversity measures the spread of generated motions in embedding space. MultiModality (MM) quantifies how varied the model's outputs are for the same text prompt. R-Precision@3 measures top-3 motion-to-text retrieval accuracy from a pool of candidates (we use size 32). MMdist measures the alignment between each generated motion and its conditioning text in the shared embedding space. Finally, penetration depth quantifies physical plausibility by measuring how object vertices intersect the body mesh.

8.1 More Evaluation on the GRAB Dataset

To further assess the quality of the contacts produced by **Dex2HOI**, we compare against hand-centric grasping method LatentHOI (Li et al., 2025b), which focuses exclusively on dexterous hand–object interaction without modeling the full body. This method is not evaluated on motion representation metrics (FID, R-Precision, etc.), and instead uses a physics-inspired protocol, explained below.

To enable a fair comparison and also test our method's ability to generalize beyond known objects, we retrain Dex2HOI on the GRAB dataset following the unseen-object protocol of LatentHOI: the four held-out object categories (*apple*, *mug*, *train*, *elephant*) are excluded from training and used exclusively for evaluation. This setting probes whether our hand-relative representation produces grasps that transfer to novel geometries, rather than overfitting to learned object instances.

We adopt the LatentHOI (Li et al., 2025b) physics protocol: interpenetration volume (IV, cm^3) and depth (ID, cm) measure hand–object mesh overlap; contact ratio (CR) measures the fraction of hand vertices within contact distance of the object; IVU is the normalized in-volume fraction over off-ground frames; Phy is the percentage of off-ground frames with positive hand–object contact; and SD and OD measure per-prompt and overall motion diversity. For the hand-only comparison, we report mesh-based hand–object penetration computed bidirectionally between the hand mesh and the object mesh, so that both hand-inside-object and object-inside-hand failure modes are captured. Diversity is computed across subjects performing the same action (inter-subject variation). Table 4 reports the results, with **bold** and underlined entries indicating the best and second-best generative models respectively. Finally, Figure 10 compares generated grasps on the 4 unseen object geometries, while Figure 11 provides additional qualitative results on GRAB, against CoDA and HOIDiNi.

	IV [cm^3] ↓		ID [cm] ↓		CR [%] ↑					
Method	Right	Left	Right	Left	Right	Left	IVU ↓	Phy ↑	SD ↑	OD →
GT	4.97	1.20	0.92	0.18	8.21	1.56	0.12	98.68	–	0.40
IMoS (Ghosh et al., 2023)	10.38	–	1.25	–	4.61	–	0.53	84.88	0.00	**0.46**
MDM (Tevet et al., 2023)	9.12	2.61	1.24	0.51	8.21	1.29	0.19	89.81	0.18	0.29
MLD (Chen et al., 2023)	9.62	3.14	1.06	0.49	10.23	0.87	0.24	85.68	0.20	0.31
LatentHOI (Li et al., 2025b)	**6.38**	**1.66**	**0.77**	**0.29**	11.94	1.11	**0.10**	**96.16**	0.13	0.27
Dex2HOI	7.39	2.59	1.94	0.75	**15.02**	**7.35**	0.26	81.98	**0.20**	0.47

Table 4: GRAB hand–object interaction evaluation against LatentHOI (Li et al., 2025b).

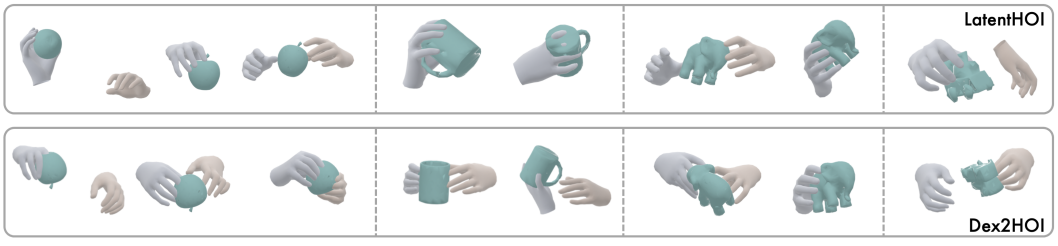

Figure 10: **Dex2HOI** vs LatentHOI (Li et al., 2025b) grasping generations on 4 unseen object geometries.

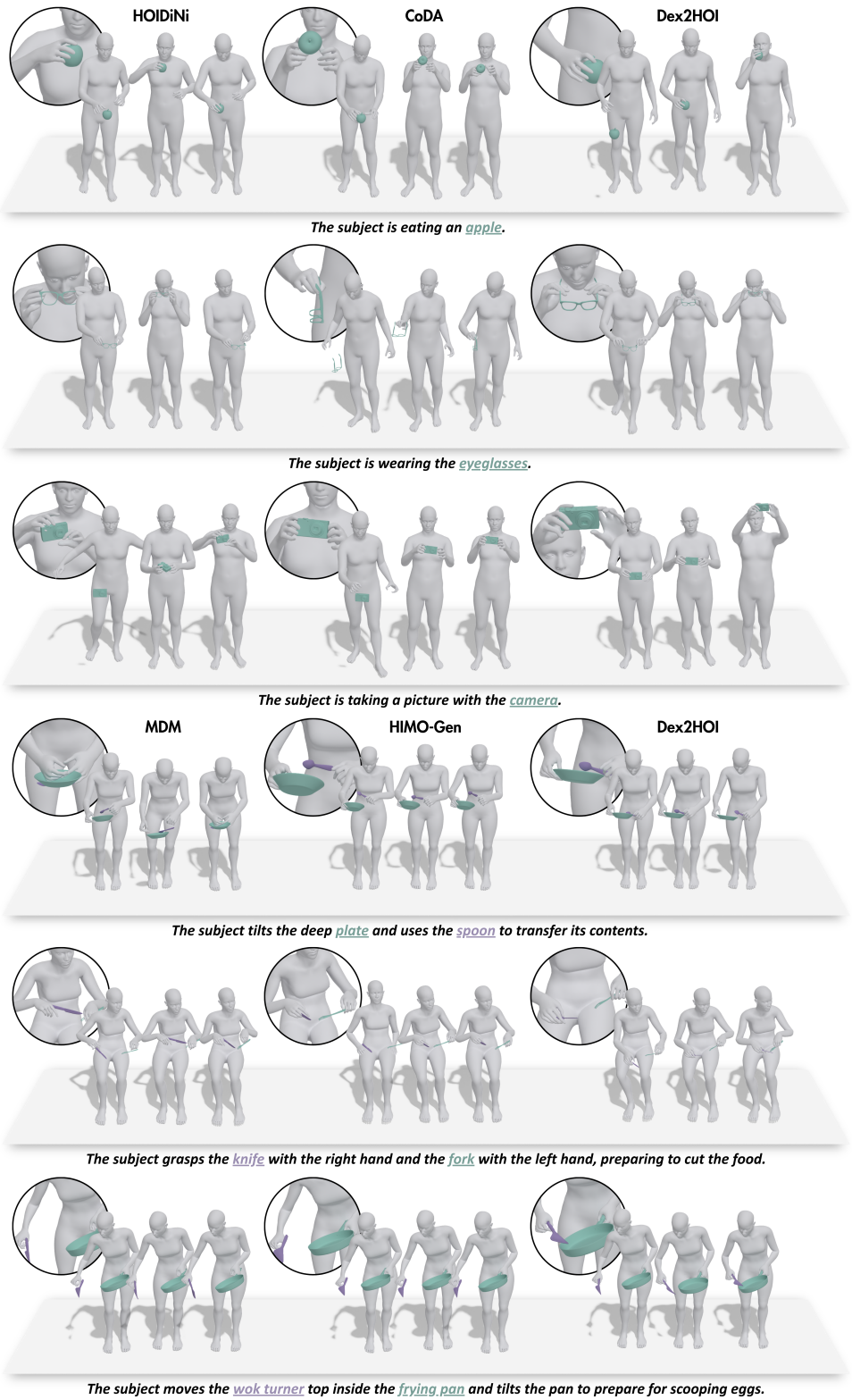

Figure 11: Additional qualitative comparisons of **Dex2HOI (right)** on the GRAB dataset, against HOIDiNi (left) and CoDA (middle), and on HUMOTO (right), against HOIDiNi (left) and CoDA (middle). Dex2HOI generates higher-quality interactions in single-shot.

8.2 HUMOTO Unseen Geometries and Dexterity Study

HUMOTO Two-Object Test Split. Throughout our experiments we use a split that contains 45 two-object sequences spanning seven scene domains (since the original HUMOTO dataset does not come with a pre-made split). Kitchen and cooking interactions constitute the largest subset (27/45, 60%), followed by office (6/45, 13.3%), dining-room (4/45, 8.9%), living-room (3/45, 6.7%), bedroom (3/45, 6.7%), garage (1/45, 2.2%), and outdoor (1/45, 2.2%) scenarios. The interactions are manipulation-centric: grasping is the most frequent dominant action, followed by placement and transfer operations including moving, placing, putting, transferring, and pouring. Across the 90 moving-object instances (45 sequences $\times$ two objects), the most frequent objects are mug (12), spoon (9), wok turner (9), deep plate (8), and frying pan (7), while 15 additional object categories occur only once. The full train/test split will be released upon publication to facilitate future research on motion-generation tasks involving diverse HUMOTO objects and actions.

To assess whether Dex2HOI learns transferable interaction priors rather than memorizing the object geometries observed during training, we perform an additional zero-shot object-switching experiment. Specifically, we replace the original HUMOTO object geometries at inference with semantically related but geometrically unseen objects sourced from HIMO, while preserving the interaction prompt and generating the motion without retraining or test-time optimization. As illustrated in Fig. 12, Dex2HOI adapts the generated interaction to substantial geometric changes, including replacing a vacuum flask and mug with unseen container geometries, and replacing a mixing bowl and spatula with a frying pan and whisk.

Diversity	**GT**	**MDM**	**HIMO-Gen**	**Dex2HOI**
Left Hand World	16.037	15.120 $\pm$ 0.082	14.228 $\pm$ 0.103	**15.552** $\pm$ **0.041**
Left Hand Wrist-relative	3.371	3.125 $\pm$ 0.020	2.951 $\pm$ 0.024	**3.368** $\pm$ **0.012**
Left Thumb	1.118	1.054 $\pm$ 0.006	1.011 $\pm$ 0.009	**1.115** $\pm$ **0.004**
Left Index	1.667	1.567 $\pm$ 0.008	1.491 $\pm$ 0.014	**1.669** $\pm$ **0.005**
Left Middle	1.677	1.568 $\pm$ 0.009	1.466 $\pm$ 0.013	**1.673** $\pm$ **0.005**
Left Ring	1.534	1.424 $\pm$ 0.011	1.320 $\pm$ 0.012	**1.531** $\pm$ **0.006**
Left Pinky	1.424	1.316 $\pm$ 0.012	1.206 $\pm$ 0.010	**1.421** $\pm$ **0.005**
Right Hand World	16.965	15.851 $\pm$ 0.076	14.975 $\pm$ 0.104	**16.762** $\pm$ **0.038**
Right Hand Wrist-relative	3.287	3.067 $\pm$ 0.012	2.876 $\pm$ 0.029	**3.284** $\pm$ **0.010**
Right Thumb	1.235	1.124 $\pm$ 0.006	1.044 $\pm$ 0.014	**1.238** $\pm$ **0.004**
Right Index	1.711	1.590 $\pm$ 0.006	1.505 $\pm$ 0.017	**1.708** $\pm$ **0.005**
Right Middle	1.571	1.486 $\pm$ 0.006	1.407 $\pm$ 0.014	**1.573** $\pm$ **0.004**
Right Ring	1.419	1.320 $\pm$ 0.005	1.247 $\pm$ 0.011	**1.416** $\pm$ **0.004**
Right Pinky	1.306	1.211 $\pm$ 0.004	1.129 $\pm$ 0.011	**1.308** $\pm$ **0.003**

Table 5: **Hand-motion diversity on the HUMOTO test set.** World-space hand diversity captures both global hand motion and articulation, whereas wrist-relative diversity removes global wrist displacement and therefore more directly measures local hand articulation. Per-finger results further isolate individual finger motion. Dex2HOI consistently remains closest to the ground-truth diversity.

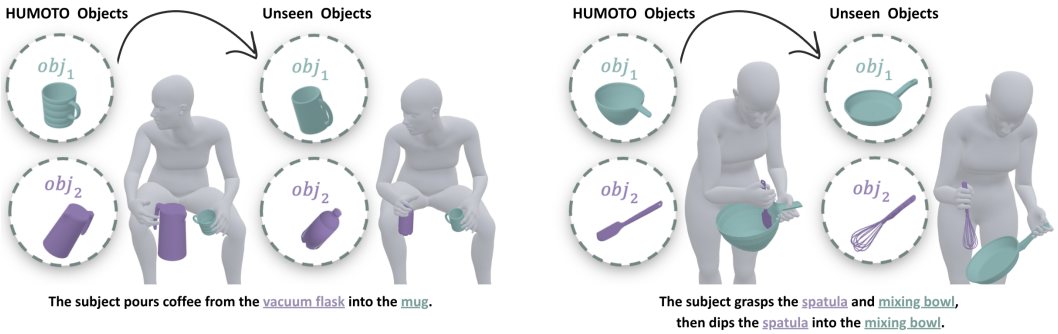

Figure 12: **Zero-shot object switching on HUMOTO.**

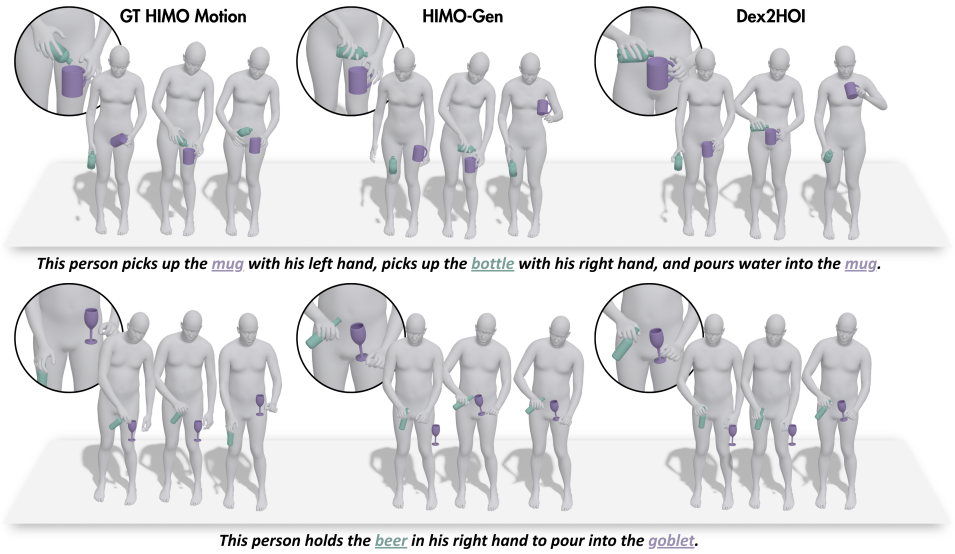

Figure 13: Additional qualitative results on the HIMO dataset (Lv et al., 2024); We show Dex2HOI generations (right) against HIMO ground-truth motion (left).

8.3 More Evaluation on HIMO Dataset

HIMO. For our next experiments, we use the 2-object split (Lv et al., 2024), which contains 2,578 training and 522 test sequences, to assess dexterous bimanual two-object motion generation.

HIMO Motion Evaluator. Following our previous protocol, we train a dedicated HIMO motion evaluator with a contrastive text–motion objective on the 2-object split, enabling the computation of FID, R-Precision, MMdist, Diversity, and Multimodality maintaining consistency with prior text-to-motion benchmarks. This evaluator is used to compute all distributional metrics reported in Table 6, where Dex2HOI again surpasses baselines HIMO-Gen and MDM.

HIMO vs. HUMOTO Ground-Truth Quality. While HIMO provides a valuable benchmark for two-object manipulation, ground-truth motions exhibit noticeable contact-displacement artifacts — objects occasionally drift away from the hand during manipulation (as seen in the GT Motion section of Figure 13). For this reason, we refrain from showing generated videos on HIMO in the supplementary video, as these artifacts in the training data propagate to the generated outputs and are difficult to interpret in isolation. Quantitatively, however, Dex2HOI still produces faithful hand-object following: as reported in Table 6, our method achieves the best FID (0.46), R-Precision@3 (0.369), and penetration (9.15mm), indicating that the model captures the underlying interaction structure despite the noisier supervision signal. Furthermore, as shown in the qualitative generated examples provided in Figure 13, Dex2HOI produces coordinated hand-object contacts that align well with the instruction prompt.

	Method	FID ↓	Diversity →	MModality ↑	Pen. (mm) ↓	R-Pre@3 ↑	MMdist ↓
2 objects	GT	—	1.184	—	—	0.791	1.029
	MDM (Tevet et al., 2023)	0.54	**1.097**	**0.628**	13.22	0.295	**1.083**
	HIMO-Gen (Lv et al., 2024)	1.09	0.445	0.323	9.95	0.140	1.393
	Dex2HOI	**0.46**	0.865	0.470	**9.15**	**0.369**	1.280

Table 6: Evaluation of **HIMO** *two-object* dataset. We highlight **best** and second best results.

9 Expanding to More than Two Objects

Although Dex2HOI explicitly models up to two objects within each denoising window, its autoregressive formulation enables longer interactions involving a larger set of objects over time. We leverage

this capability to experiment with longer horizon motion prediction and sequential interaction with more than two objects. In the case of HUMOTO (and HIMO) we can clearly distinguish between two types of multi-object interactions - both of which are included in the supplementary video provided:

(1) sequential single-hand hand-offs, where the subject is never in contact with all three objects at once - e.g. they set an object down before picking up the third - which is exactly a chained sequence of two-object interactions and can be represented as two consecutive two-object windows with a hand-off event, requiring no new architecture; and

(2) compositional carrying, where multiple secondary objects are moved rigidly together with a base object (e.g. carrying a tray with glasses); here the secondary objects' motion can be determined by the base object's motion and can be captured by a single unified BPS representation rather than as independent streams.

Figure 14 demonstrates this capability by switching the active object conditions throughout autoregressive generation, particularly in the HUMOTO.kitchen table setting.

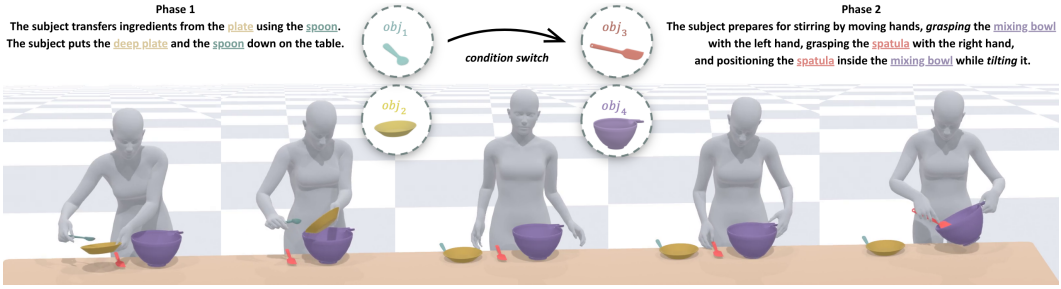

Figure 14: **Autoregressive generation beyond two objects.** While each Dex2HOI denoising window models at most two active objects, the object conditions can be switched between autoregressive windows. Previously generated motion is retained as the temporal prefix while new object geometries and interaction conditions are introduced, allowing a continuous sequence to compose interactions with more than two objects without modifying the model architecture or retraining.

10 Perceptual Studies Analysis

To complement the quantitative metrics reported in Tab. 1, we conduct two perceptual evaluation studies: an automated VLM-based evaluation and a human preference study. Both are designed to probe dimensions of motion quality that FID/diversity scores do not capture, namely perceptual naturalness, semantic alignment with the text prompt, and hand-contact plausibility.

Automated VLM Evaluation. We evaluate generated motions using a video-language model (VLM) to obtain **scalable** assessment across all test sequences. For each generated motion we render a mesh-body video and present it to **InternVL3-8B** (et al., 2025) alongside three structured prompts:

1. *Naturalness*: "Rate how natural and human-like the body movement looks on a scale of 1 to 5, where 1 = completely unnatural/robotic and 5 = perfectly natural human motion."
2. *Semantic alignment*: "Rate how well the motion matches the description `{prompt}` on a scale of 1 to 5."
3. *Contact quality*: "Rate how well the hands make physical contact with the object on a scale of 1 to 5, where 1 = hands never touch the object and 5 = hands grasp the object convincingly."

Each question is answered independently and the model is constrained to reply with a single integer from 1 to 5, making the evaluation deterministic. We run this protocol on two benchmarks; On the **GRAB** test set we evaluate 145 test set sequences for each of CoDA, HOIDiNi, and Dex2HOI. On the **HUMOTO** test set we evaluate 45 sequences for MDM, HIMO-Gen, and Dex2HOI. To further

validate our results, we run a *pairwise* choice evaluation: for every pair of methods (A, B), the model is shown both videos and chooses which exhibits better naturalness, contact, or overall quality.

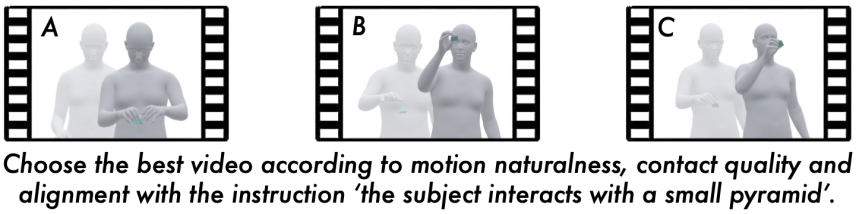

Figure 15: **User study**. A sample of the user study where 30 participants evaluated Dex2HOI generated motions against baseline methods.

As shown in Figure 16, **Dex2HOI** consistently achieves superior interaction scores, in both single- and two-object benchmarks, demonstrating more natural motions and better contacts than competing baselines, further validated by our user study.

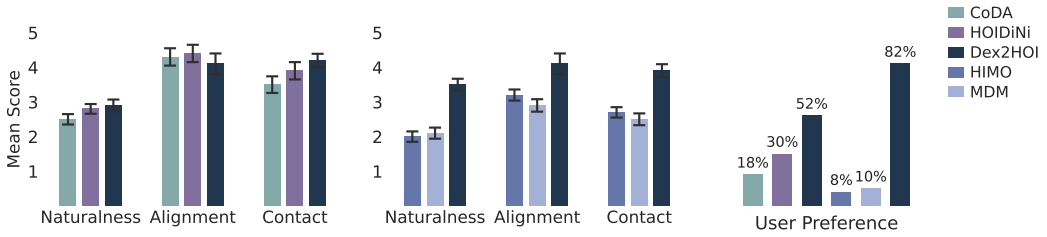

Figure 16: **Perceptual Studies**: *Left*: VLM-eval for 1-object GRAB dataset. *Middle*: VLM-eval for 2-object HUMOTO dataset. *Right*: Collective User Preference. **Dex2HOI** achieves superior scores in realism and contact quality and is preferred by the majority of participants over competing approaches.

11 Discussion and Broader Impact

Limitations. Despite **Dex2HOI**'s promising results, certain challenges remain. While our architecture supports long sequence synthesis through autoregressive sampling, the diversity and complexity of the motions remain bounded by the bimanual interactions present in current 4D HOI datasets, which are limited. Future work should address long-horizon action planning and the development of standardized benchmarks for scaling multi-object bimanual manipulation.

Broader Impact. Dex2HOI advances the synthesis of dexterous bimanual single- and two-object interactions, with potential to support a range of beneficial applications including immersive AR/VR experiences, animation pipelines, and the training of embodied agents and assistive robotic systems that operate in human-populated environments, while adhering to established ethical guidelines regarding synthetic data generation and use.