UNLEASHING GUIDANCE WITHOUT CLASSIFIERS FOR HUMAN-OBJECT INTERACTION ANIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating realistic human-object interaction (HOI) animations remains challenging because it requires jointly modeling dynamic human actions and diverse object geometries. Prior diffusion-based approaches often rely on handcrafted contact priors or human-imposed kinematic constraints to improve contact quality. We propose a data-driven alternative in which guidance emerges from the denoising pace itself, reducing dependence on manually designed priors. Building on diffusion forcing, we factor the representation into modality-specific components and assign individualized noise levels with asynchronous denoising schedules. In this paradigm, cleaner components guide noisier ones through cross-attention, yielding guidance without auxiliary classifiers. We find that this data-driven guidance is inherently contact-aware, and can be further enhanced when training is augmented with a broad spectrum of synthetic object geometries, encouraging invariance of contact semantics to geometric diversity. Extensive experiments show that pace-induced guidance more effectively mirrors the benefits of contact priors than conventional classifier-free guidance, while achieving higher contact fidelity, more realistic HOI generation, and stronger generalization to unseen objects and tasks.

1 INTRODUCTION

Human-object interaction (HOI) animation, which involves generating dynamic motion sequences of a person interacting with objects, has become an important problem in computer vision and graphics. Realistic HOI animation underpins applications ranging from virtual reality, gaming, and robotics simulation, where agents must manipulate or use objects in a human-like way. Recent advances in generative model have opened new possibilities for synthesizing such animations directly from high-level specifications, *e.g.,* text descriptions. In particular, diffusion models (Peng et al., 2023; Diller & Dai, 2024; Li et al., 2023a) have emerged as a powerful generative framework capable of modeling high-dimensional distribution of human motions and contacts, by iteratively denoising noise into plausible HOI sequences.

Despite this promise, a persistent challenge is ensuring high-quality realistic interactions between humans and objects. A plain diffusion model, without physics simulation, often produces artifacts: hands miss their targets, objects drift or penetrate the body, or contact is unstable over time. Prior work has sought to mitigate this through *guidance*. One line of methods employs external classifiers trained for *human-specified tasks*, *e.g.*, contact or affordance regression (Peng et al., 2023; Diller & Dai, 2024) to steer the diffusion process, but such classifier guidance is cumbersome to design, and may overfit to specific priors. Another line introduces *handcrafted rules* with kinematic constraints, such as forcing hands to align with objects via inverse kinematics (Xu et al., 2024), but these approaches sacrifice generality and are computationally inefficient. Together, these efforts highlight: existing diffusion-based HOI animation still relies heavily on external priors not directly from the data.

Extending classifier-free guidance (CFG) (Ho & Salimans, 2022) to HOI animation is a natural way to reduce the reliance on external priors, but CFG, especially CFG based on text dropout for text-conditioned generation, chiefly improves global distributional alignment and offers limited control over the fine-grained, persistent contact central to HOI. We therefore introduce LIGHT, *i.e.,* *L*earning *I*mplicit *G*uidance for *H*uman-object in*T*eraction, a complementary data-driven framework where guidance arises from the relative denoising pace of separate components. Concretely, we factor the representation into modality-specific components (*e.g.,* human and object) and assign each its own

noise level with asynchronous schedules. And we formulate two paths: (**I**) a *staged schedule*, where one modality (*e.g.*, human) is kept cleaner while the other (*e.g.*, object) follows its prescribed noise levels, approximating a conditional trajectory; and (**II**) a *uniform schedule* that weakly conditions by assigning all with prescribed noise levels, here the "unconditional" is realized as noisier conditioning as a *soft* form of CFG. The contrast between these two paths produces a guidance effect analogous to CFG. As the lag between schedules approaches zero, LIGHT collapses to joint denoising with no guidance; as the lag grows large, it approximates conditioning dropout in CFG. Our experiment shows that hard dropout on text improves global distributional alignment, whereas the soft guidance from LIGHT adjusts more on low-level contact details, indicating that LIGHT learns, purely from data, to reduce contact errors without hand-crafted priors. Note that assigning different noise levels to diffusion models naturally corresponds to the diffusion forcing mechanism (Chen et al., 2024), and LIGHT provides a principled extension of diffusion forcing into a guidance framework.

LIGHT relies solely on data priors; however, this does not constrain the model's flexibility to incorporate additional world knowledge like physics provided by humans, as such human-induced priors can also be reflected through data manipulation. We find that LIGHT achieves further improvement when training data is augmented with prior via *contact-aware shape-spectrum augmentation*. Specifically, objects are augmented with geometrically diverse alternatives from large repositories (Chang et al., 2015; Deitke et al., 2023), and motions are retargeted to preserve contact semantics and relative pose. These synthetic pairs teach the model that contact should remain invariant to irrelevant shape changes. This results in a stronger prior that the asynchronous schedules can exploit, improving generalizability to, *e.g.,* unseen objects from training.

In summary, our contribution lies in three folds: First, we introduce LIGHT, a novel guidance mechanism in which asynchronous denoising induces guidance without reliance on external classifiers. In contrast to CFG's reliance on hard dropout, our formulation provides a softer and more flexible alternative, naturally extending diffusion forcing into the guidance framework and potentially *inspiring future applications* in other domains. Second, we propose contact-aware shape-spectrum augmentation, a strategy that preserves contact semantics while varying object geometry, thereby enabling the model to acquire more robust generative capabilities and improve generalization. Third, we conduct extensive experimental evaluations demonstrating that the proposed approach consistently outperforms existing baselines and facilitates the generation of vivid and realistic interactions.

2 Related Work

Denoising Diffusion Models for Human Animation. Denoising diffusion models (Sohl-Dickstein et al., 2015; Song et al., 2020; Ho et al., 2020) synthesize data by gradually denoising samples drawn from a noise distribution, effectively reversing a stochastic diffusion process. Recently, these models have achieved notable success in human motion generation, producing realistic and diverse animations (Barquero et al., 2023; Tevet et al., 2023; Zhang et al., 2022a; Raab et al., 2023; Zhang et al., 2023b; Shafir et al., 2023; Zhang et al., 2023e). A prominent example, MDM (Tevet et al., 2023), leverages transformer architectures to predict clean human motion trajectories from noisy inputs during the denoising stages. To generate conditional motion sequences, such diffusion methods typically incorporate external conditions into the diffusion process, including textual descriptions (Petrovich et al., 2023; Guo et al., 2022b; Petrovich et al., 2022; Zhang et al., 2022a; 2023d; Tevet et al., 2022a; Barquero et al., 2024). Further advancements have extended diffusion-based approaches to more conditioning scenarios, including guiding animations along specific trajectories (Karunratanakul et al., 2023; Rempe et al., 2023; Xie et al., 2023).

Denoising Diffusion Models for Interaction Animation. Recent advancements in interaction animation have shown significant progress in diverse scenarios, including human-human interactions (Liang et al., 2023; CMU; Mehta et al., 2018; Xu et al., 2023a) and human interactions within static scenes (Hassan et al., 2021; Cao et al., 2020; Hassan et al., 2019). Early human-object interaction (HOI) animation methods typically leveraged kinematic models combined with conventional generative frameworks. These earlier approaches primarily addressed simplified scenarios, either involving static or small objects (Xie et al., 2022; Zhang et al., 2020; Wang et al., 2022; Petrov et al., 2023; Taheri et al., 2022; Wu et al., 2022; Kulkarni et al., 2023; Zhang et al., 2022b), or focusing exclusively on hand-object interactions (Li et al., 2023c; Ye et al., 2023; Zheng et al., 2023; Zhou et al., 2022a; Zhang et al., 2024a; 2023a). Consequently, they struggled to effectively capture

complex, dynamic, whole-body interactions. Recently, denoising diffusion models have emerged as a powerful paradigm capable of modeling sophisticated interactions involving dynamically moving objects (Peng et al., 2023; Diller & Dai, 2024; Li et al., 2023a; Wu et al., 2024a;b; Song et al., 2024; Xu et al., 2024; Zhang et al., 2024b). Concurrently, new datasets have expanded the range of possible interactions – from low-dynamic manipulations (Taheri et al., 2020) to highly dynamic scenarios engaging multiple body parts simultaneously (Bhatnagar et al., 2022; Jiang et al., 2023; Huang et al., 2022; Zhang et al., 2023c; Fan et al., 2023; Li et al., 2023b; Zhao et al., 2024; Kim et al., 2024b; Jiang et al., 2024; Yang et al., 2024).

Guidance with Denoising Diffusion Models. Diffusion models can be steered by additional guidance signals to better satisfy conditioning constraints. In image and text generation, for instance, classifier guidance (Dhariwal & Nichol, 2021) uses an external classifier's gradients to direct the denoising process toward desired outcomes, whereas classifier-free guidance (Ho & Salimans, 2022) foregoes a separate classifier by leveraging the model's own conditional and unconditional predictions for guidance. Inspired by such strategies, recent works in 3D human–object interaction (HOI) generation have explored explicit contact-aware or auxiliary-guided approaches to ensure realism. Without explicit constraints, diffusion-based HOI models often produce unrealistic artifacts, e.g., floating objects or missing contact points. To mitigate this, InterDiff (Xu et al., 2023b), for example, employs a kinematics-informed predictor that iteratively refines the diffusion output with corrections, improving human-object prediction accuracy. Similarly, HOI-Diff (Peng et al., 2023) integrates an auxiliary affordance module to steer the model toward consistent contact and affordance cues throughout generation. These approaches demonstrably boost realism but introduce additional complexity by relying on extra networks or hand-crafted constraints, which can hinder generalization. CG-HOI (Diller & Dai, 2024) cast interaction synthesis as a multi-task learning problem: an auxiliary contact prediction task is learned jointly to guide the motion generation, thereby avoiding training separate guidance models. However, it still imposes predefined relationships between predicted contact and object movements, using weighted combinations that embed human-designed assumptions. In contrast to these externally guided methods, we propose a guidance strategy without any assumptions on human-object contact, where the proposed approach gradually applies guidance according to a temporal schedule for different components during denoising, promoting a new principled manner and and potentially inspiring future applications in other domains.

3 METHODOLOGY

Overview. We formalize the HOI animation task as the synthesis of a 3D motion sequence in which a human interacts with an object over a time horizon of T frames. The input is a text description $\boldsymbol{d}$, a canonical geometry of the object represented as a point cloud $\boldsymbol{P}$, and a body shape parameter β from SMPL-H (Romero et al., 2017), and the output of LIGHT is a sequence comprising both human motion and object motion trajectories. The human state is represented by joint positions $\boldsymbol{j}^p \in \mathbb{R}^{T\times 52\times 3}$, where 30 of total 52 joints are for both hands. Object trajectories are in translations $\boldsymbol{o}^t \in \mathbb{R}^{T\times 3}$ and rotations in a 6D representation $\boldsymbol{o}^r \in \mathbb{R}^{T\times 6}$. Thus, each frame $\boldsymbol{x}_t$ is fully characterized by the tuple $\boldsymbol{x}_t = (\boldsymbol{j}_t^p, \boldsymbol{o}_t^t, \boldsymbol{o}_t^r)$. We omit the time step t for simplicity.

Preliminaries. Diffusion forcing (Chen et al., 2024) is a recent advancement in diffusion modeling that generalizes the conventional diffusion process by allowing independent noise schedules for each token within a sequence. Unlike standard diffusion models, which uniformly apply a single noise schedule across all tokens at each diffusion step, diffusion forcing treats each token independently, enabling flexible and staged denoising processes tailored to sequential generation tasks. Formally, consider a sequence of tokens $\boldsymbol{x}(\mathbf{0})$, representing the unperturbed data. Diffusion forcing introduces token-specific noise levels $\boldsymbol{\lambda}$ with each element drawn independently from $\{0, 1, 2, \ldots, K\}$, where K is the total number of denoising steps. Each token of $\boldsymbol{x}(\mathbf{0})$ is then corrupted individually, $\boldsymbol{x}(\boldsymbol{\lambda}) = \langle\sqrt{\bar{\alpha}(\boldsymbol{\lambda})}, \boldsymbol{x}(0)\rangle + \langle\sqrt{1-\bar{\alpha}(\boldsymbol{\lambda})}, \boldsymbol{\epsilon}\rangle$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ denotes Gaussian noise, and $\bar{\alpha}(\boldsymbol{\lambda})$ defines the cumulative noise schedule for handling tokens corrupted by varying degrees of noise. We use dot product $\langle, \rangle$ because the noise level for different tokens can be varied.

We slightly adjust the original diffusion forcing framework to directly predict the clean data $\tilde{\boldsymbol{x}}(\mathbf{0}) = \mathcal{G}_\theta(\boldsymbol{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \boldsymbol{d})$ from its noised counterpart $\boldsymbol{x}(\boldsymbol{\lambda})$, a typical solution for the motion generation task (Tevet et al., 2023), rather than predicting the noise. The training objective is formulated as

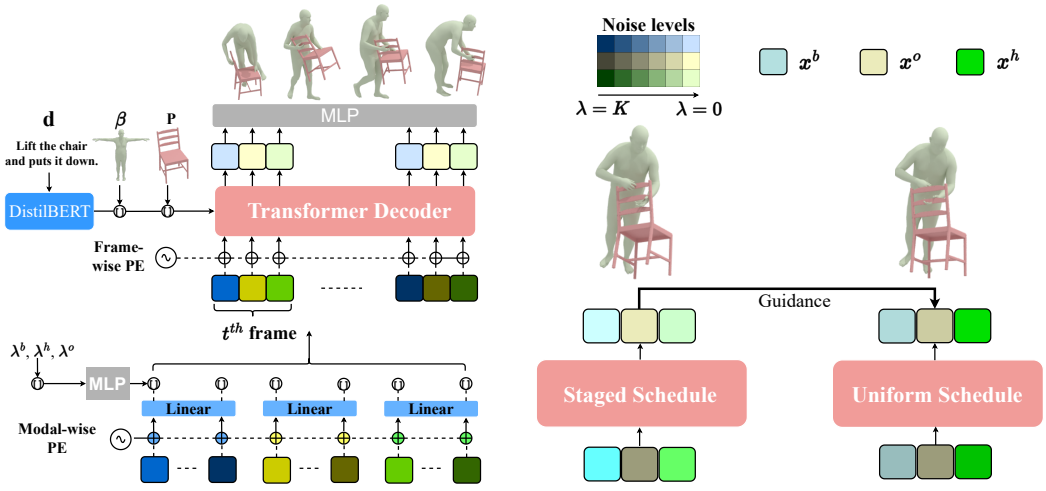

Figure 1: **Overview of LIGHT .** *Left: Training.* We form three modalities – body, hand, and object – each diffused with its own noise level. After adding modal-wise and frame-wise positional encodings, the tokens are processed by a shared Transformer decoder and an MLP head to predict clean motion. *Right: Inference.* We compare a *uniform* schedule that denoises all modalities synchronously with a *staged* schedule that keeps one modality cleaner from the uniform run.

minimizing the reconstruction error,

$$\mathcal{L}_{\text{DF}} = \mathbb{E}_{\boldsymbol{x}(0),\boldsymbol{\lambda}} \|\hat{\boldsymbol{x}}(\mathbf{0}) - \mathcal{G}_\theta(\boldsymbol{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \boldsymbol{d})\|^2, \tag{1}$$

where $\hat{\boldsymbol{x}}(\mathbf{0})$ denotes the ground truth data. $\mathcal{G}_\theta(\boldsymbol{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \boldsymbol{d})$ denotes the model prediction of clean data from the generative model $\mathcal{G}$, parameterized by θ, conditioned on the partially noised sequence $\boldsymbol{x}(\boldsymbol{\lambda})$, the noise level $\boldsymbol{\lambda}$ for each token, and the textual description $\boldsymbol{d}$. For notational simplicity, we omit other inputs including human shape parameter $\boldsymbol{\beta}$ and object geometry $\boldsymbol{P}$ here; we describe how these additional conditioning variables are incorporated into our generative model below.

Token Separation. Following Cha et al. (2024), we explicitly decompose the representation $\boldsymbol{x}$ into distinct modalities for the human body, hands, and objects, denoted as $\boldsymbol{x}^b$, $\boldsymbol{x}^h$, and $\boldsymbol{x}^o$, respectively, which formulates these three components as separate token groups, yielding a total of $3 \times T$ tokens and noise levels $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}^b, \boldsymbol{\lambda}^h, \boldsymbol{\lambda}^o\} \in \mathbb{R}^{T \times 3}$. As common practice for whole-body motion generation (Lu et al., 2023), this separation is beneficial for differentiating body and hand representations due to their distinct characteristics: the body encompasses 22 joints and typically exhibits larger-scale spatial movements, while the hands contain 30 joints with fine-grained and intricate finger motions. Empirically, we observe that separating the hand component improves motion generation quality in tasks with frequent hand interactions such as on the GRAB dataset (Taheri et al., 2020) (Figure 3).

Overall Inference Procedure. As shown in Figure 1, at test time, we perform two coupled denoising passes that share the same denoiser $\mathcal{G}_\theta$, as outlined in Algorithm 1. First, a *uniform* schedule denoises all modalities in sync under a same schedule, enhanced with text classifier-free guidance. Second, a *staged* schedule applies modality-specific schedules. The staged schedule incorporates the cleaner generation from the prior uniform schedule and introduces pace-induced guidance, combined with text CFG, as incremental guidance. The final sample is obtained from the staged schedule. Design choices and hyperparameters are provided in Sec. 4.

Inference with Uniform Schedule. The reverse process of diffusion forcing gradually denoises a noisy latent sequence $\boldsymbol{x}_U(\boldsymbol{\lambda})$ into a clean sequence $\boldsymbol{x}_U(\mathbf{0})$, defined as:

$$\begin{aligned} \tilde{\boldsymbol{x}}_U &= \mathcal{G}_\theta(\boldsymbol{x}_U(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \boldsymbol{d}) \\ &\quad + \omega_1 \big(\mathcal{G}_\theta(\boldsymbol{x}_U(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \boldsymbol{d}) - \mathcal{G}_\theta(\boldsymbol{x}_U(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \emptyset)\big), \end{aligned} \tag{2}$$

$$\boldsymbol{x}_U(\boldsymbol{\lambda} - \mathbf{1}) = \langle \sqrt{\bar{\alpha}(\boldsymbol{\lambda} - \mathbf{1})}, \tilde{\boldsymbol{x}}_U \rangle + \langle \sqrt{\mathbf{1} - \bar{\alpha}(\boldsymbol{\lambda} - \mathbf{1})}, \boldsymbol{\epsilon} \rangle, \tag{3}$$

where our generative model predicts the denoised sequence $\tilde{\boldsymbol{x}}_U$ given the partially noised input sequence $\boldsymbol{x}_U(\boldsymbol{\lambda})$, with classifier-free guidance (CFG) with text $\boldsymbol{d}$ dropout. ω_1 is a hyperparameter to control the text CFG. With $\tilde{\boldsymbol{x}}_U$ further noised back to $\boldsymbol{x}_U(\boldsymbol{\lambda} - \mathbf{1})$, this iterative denoising step

Algorithm 1 Inference with pace-induced guidance

Require: Total number of denoising steps K; lagged modality index m_2; cleaner modality index m_1; $\boldsymbol{x}_U(\boldsymbol{K})$ and $\boldsymbol{x}_S(\boldsymbol{K})$ initialized with same Gaussian noise; lag offset vector $\boldsymbol{\delta}$

for $k = K, K-1, \ldots, 1$ **do**
  Assign $\boldsymbol{\lambda}$ as an all-k vector
  Obtain denoised $\tilde{\boldsymbol{x}}_U$ from $\boldsymbol{x}_U(\boldsymbol{\lambda})$ and $\boldsymbol{\lambda}$ with text CFG ▷ Equation 2
  Add noise back to $\boldsymbol{x}_U(\boldsymbol{\lambda} - \mathbf{1})$ and record ▷ Equation 3
end for
for $k = K, K-1, \ldots, 1$ **do**
  Assign $\boldsymbol{\lambda}$ as an all-k vector
  if $\boldsymbol{\lambda} - \boldsymbol{\delta} \geq \mathbf{0}$ for all elements **then**
    Obtain the input of conditional branch $\boldsymbol{x}'_S(\boldsymbol{\lambda})$ by merging $\boldsymbol{x}_U(\boldsymbol{\lambda} - \boldsymbol{\delta})$ ▷ Equation 4
    Obtain denoised $\tilde{\boldsymbol{x}}_S$ from $\boldsymbol{x}_S(\boldsymbol{\lambda})$ and $\boldsymbol{\lambda}$ with text CFG ▷ Equation 5
    Update $\tilde{\boldsymbol{x}}_S$ with $\boldsymbol{x}'_S(\boldsymbol{\lambda})$ by LIGHT ▷ Equation 6, incremental guidance
  else
    Obtain denoised $\tilde{\boldsymbol{x}}_S$ from $\boldsymbol{x}_S(\boldsymbol{\lambda})$ and $\boldsymbol{\lambda}$ with text CFG
  end if
  Add noise back to $\boldsymbol{x}_S(\boldsymbol{\lambda} - \mathbf{1})$ ▷ Equation 7
end for
return $\boldsymbol{x}_S(\mathbf{0})$

yields clearer HOI sequences until the inference advances to step $\mathbf{0}$. In this scenario, all frames and modalities are denoised simultaneously as plain diffusion models, with $\boldsymbol{\lambda}$ set to be the same and start at K. The resulting trajectory $\boldsymbol{x}_U(\cdot)$ are incorporated into the staged schedule to formulate LIGHT.

Inference with Staged Schedule. Our *staged inference* strategy assigns asynchronized denoising schedules, enabling distinct denoising paces across modalities. Specifically, for specified modalities m_1, m_2 satisfying $m_1 \cap m_2 = \varnothing, m_1 \cup m_2 = \{b, h, o\}$. *i.e.,* together m_1 and m_2 cover all modalities without overlap (*e.g.,* $m_1 = \{h\}, m_2 = \{b, o\}$). This partitioning is flexible. In Table D, we verify that all possible combinations yield improvements to varying degrees. Then, we set

$$\boldsymbol{x}'_S = \left(\boldsymbol{x}_U^{m_1}(\boldsymbol{\lambda}^{m_1} - \boldsymbol{\delta});\, \boldsymbol{x}_S^{m_2}(\boldsymbol{\lambda}^{m_2})\right), \quad \boldsymbol{\lambda}' = \left((\boldsymbol{\lambda}^{m_1} - \boldsymbol{\delta});\, \boldsymbol{\lambda}^{m_2}\right), \tag{4}$$

where the vectors $\boldsymbol{\lambda}^{m_1}$ and $\boldsymbol{\lambda}^{m_2}$ are all-k vectors at denoising step k, and $(\cdot\,;\,\cdot)$ denotes the concatenation operation. The offset vector $\boldsymbol{\delta}$ determines how much earlier m_1 is denoised compared to m_2, thereby creating a pacing discrepancy across modalities. This discrepancy allows m_1 to reach a cleaner state from the previous generation $\boldsymbol{x}_U^{m_1}$, and when it is combined with the current output of $\boldsymbol{x}_S^{m_2}$ to formulate $\boldsymbol{x}'_S$, it can lead to what we refer to as pace-induced guidance. Formally, the guided update is given by

$$\begin{aligned}
\tilde{\boldsymbol{x}}_S = \mathcal{G}_\theta(\boldsymbol{x}_S(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \boldsymbol{d}) & \\
+ \omega_1(\mathcal{G}_\theta(\boldsymbol{x}_S(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \boldsymbol{d}) - \mathcal{G}_\theta(\boldsymbol{x}_S(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \varnothing)) & \quad (5)\\
+ \omega_2(\mathcal{G}_\theta(\boldsymbol{x}'_S, \boldsymbol{\lambda}', \boldsymbol{d}) - \mathcal{G}_\theta(\boldsymbol{x}_S(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \boldsymbol{d}))\,, & \quad (6)\\
\boldsymbol{x}_S(\boldsymbol{\lambda} - \mathbf{1}) = \langle \sqrt{\bar{\alpha}(\boldsymbol{\lambda} - \mathbf{1})}, \tilde{\boldsymbol{x}}_S \rangle + \langle \sqrt{\mathbf{1} - \bar{\alpha}(\boldsymbol{\lambda} - \mathbf{1})}, \boldsymbol{\epsilon} \rangle, & \quad (7)
\end{aligned}$$

where ω_1, ω_2 are the scalar weights controlling the strength of the conditional influence. The resulting guided prediction for the m_2 component is then re-noised according to the schedule and propagated to the next denoising step, while m_1 continues to follow the trajectory of $\boldsymbol{x}_U(\boldsymbol{k}), k \in \{0, 1, 2, \ldots, K\}$.

Architecture. The left figure in Figure 1 illustrates the full architecture $\mathcal{G}_\theta$. (i) *Motion tokens.* At each frame we form three motion tokens, body $\boldsymbol{x}_t^b$, hand $\boldsymbol{x}_t^h$ and object $\boldsymbol{x}_t^o$, and attach a *modal-wise positional encoding* that distinguishes these modalities. A small linear projection is applied to each modality so that the three streams share a common hidden dimensionality before fusion. (ii) *Noise-level and temporal encodings.* Each motion token receives an additive noise-level embedding together that encodes the diffusion step with a *frame-wise timestep embedding*. (iii) *Object-geometry token.* Object shape is encoded once per sequence. From the point cloud $\boldsymbol{P}$, we concatenate an un-normalised Basis Point Set (BPS) (Prokudin et al., 2019) descriptor with a normalized BPS (Zhang et al., 2024c) whose maximal point-to-centroid distance is normalized to 0.95; the original object

Table 1: **Quantitative comparisons** on the InterAct dataset (Xu et al., 2025) between our method and baseline approaches. We report R-Precision with batch sizes 64 and 256.

Method	R-Precision↑						FID↓	MM Dist↓	Diversity→	FSR↓	Pene↓	Contact→	Interaction↑		
	Batch Size = 64			Batch Size = 256											
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
Ground Truth	$0.808^{\pm 0.002}$	$0.950^{\pm 0.000}$	$0.988^{\pm 0.000}$	$0.574^{\pm 0.000}$	$0.773^{\pm 0.005}$	$0.879^{\pm 0.000}$	$0.000^{\pm 0.000}$	$1.439^{\pm 0.001}$	$7.509^{\pm 0.049}$	$0.030^{\pm 0.001}$	$0.036^{\pm 0.000}$	$0.261^{\pm 0.000}$	$0.891^{\pm 0.000}$	$0.906^{\pm 0.000}$	$0.898^{\pm 0.000}$
HOI-Diff (Peng et al., 2023)	$0.703^{\pm 0.026}$	$0.908^{\pm 0.006}$	$0.953^{\pm 0.000}$	$0.402^{\pm 0.027}$	$0.602^{\pm 0.005}$	$0.730^{\pm 0.000}$	$0.909^{\pm 0.034}$	$2.640^{\pm 0.012}$	$7.449^{\pm 0.022}$	$0.032^{\pm 0.004}$	$0.056^{\pm 0.002}$	$0.096^{\pm 0.003}$	$0.695^{\pm 0.023}$	$0.451^{\pm 0.013}$	$0.495^{\pm 0.006}$
CHOIS (Li et al., 2023a)	$0.744^{\pm 0.013}$	$0.920^{\pm 0.019}$	$0.972^{\pm 0.017}$	$0.445^{\pm 0.011}$	$0.693^{\pm 0.008}$	$0.818^{\pm 0.003}$	$0.493^{\pm 0.028}$	$2.155^{\pm 0.056}$	$7.654^{\pm 0.102}$	$0.060^{\pm 0.008}$	$0.068^{\pm 0.000}$	$0.158^{\pm 0.004}$	$0.776^{\pm 0.027}$	$0.664^{\pm 0.017}$	$0.675^{\pm 0.026}$
InterDiff (Xu et al., 2023b)	$0.777^{\pm 0.002}$	$\mathbf{0.941}^{\pm 0.004}$	$\mathbf{0.984}^{\pm 0.000}$	$0.484^{\pm 0.005}$	$0.721^{\pm 0.008}$	$0.848^{\pm 0.011}$	$0.194^{\pm 0.008}$	$1.901^{\pm 0.001}$	$7.648^{\pm 0.024}$	$0.048^{\pm 0.007}$	$0.064^{\pm 0.018}$	$0.171^{\pm 0.002}$	$0.788^{\pm 0.009}$	$0.684^{\pm 0.000}$	$0.699^{\pm 0.002}$
Text2HOI (Cha et al., 2024)	$0.559^{\pm 0.004}$	$0.762^{\pm 0.004}$	$0.859^{\pm 0.000}$	$0.430^{\pm 0.011}$	$0.637^{\pm 0.016}$	$0.764^{\pm 0.014}$	$0.434^{\pm 0.026}$	$2.795^{\pm 0.025}$	$7.434^{\pm 0.018}$	$0.031^{\pm 0.001}$	$\mathbf{0.048}^{\pm 0.003}$	$0.124^{\pm 0.002}$	$0.755^{\pm 0.007}$	$0.548^{\pm 0.007}$	$0.590^{\pm 0.004}$
LIGHT (**Ours**) w/o guidance	$0.764^{\pm 0.015}$	$0.927^{\pm 0.002}$	$0.975^{\pm 0.004}$	$0.475^{\pm 0.003}$	$0.729^{\pm 0.003}$	$0.850^{\pm 0.003}$	$0.202^{\pm 0.013}$	$1.921^{\pm 0.026}$	$7.615^{\pm 0.003}$	$\mathbf{0.029}^{\pm 0.008}$	$0.054^{\pm 0.005}$	$0.186^{\pm 0.000}$	$\mathbf{0.792}^{\pm 0.004}$	$0.731^{\pm 0.034}$	$0.733^{\pm 0.025}$
LIGHT (**Ours**) w/ guidance	$\mathbf{0.781}^{\pm 0.004}$	$0.927^{\pm 0.006}$	$0.975^{\pm 0.004}$	$\mathbf{0.498}^{\pm 0.003}$	$\mathbf{0.730}^{\pm 0.005}$	$\mathbf{0.855}^{\pm 0.011}$	$\mathbf{0.188}^{\pm 0.010}$	$\mathbf{1.863}^{\pm 0.036}$	$\mathbf{7.547}^{\pm 0.022}$	$0.031^{\pm 0.006}$	$0.051^{\pm 0.008}$	$0.178^{\pm 0.003}$	$0.791^{\pm 0.000}$	$\mathbf{0.753}^{\pm 0.006}$	$\mathbf{0.755}^{\pm 0.001}$

scale (largest radius) is appended as an extra scalar. This geometry vector is processed by an MLP and then concatenated to text tokens. (iv) *Text token.* The input prompt $\boldsymbol{d}$ is encoded by a frozen DistilBert (Sanh et al., 2019) text encoder, producing a sequence of language tokens. These tokens are injected into each of the transformer decoder's layers through their cross-attention blocks. (v) *Transformer decoder and prediction head.* The concatenated sequence, including body, hand and object tokens with all positional, temporal embeddings, is passed through a Transformer decoder, with text and object-geometry tokens injected by cross-attention. A lightweight MLP head maps the denoised motion latents to final output, which are trained to match the clean ground-truth targets.

Training Schedule. Following Kim et al. (2024a); Xiu et al. (2025); Zhang et al. (2024d); Chen et al. (2024), we assign independent noise levels across modalities, formally defined as $\boldsymbol{\lambda}^b, \boldsymbol{\lambda}^h, \boldsymbol{\lambda}^o \sim U\{\mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{K}\}$ independently. Training with independent noise levels encourages the model to capture a broad range of asynchronized conditional distributions, enabling diverse conditionalities at inference. Unlike diffusion forcing (Chen et al., 2024), we do not vary noise levels across frames; instead, all frames within the same modality share the same noise level.

Training Loss. The model is trained with a composite loss consisting of a diffusion supervised term L_{DF} and a regularization term L_{reg}. The total loss is defined as $\mathcal{L} = \mathcal{L}_{\text{DF}} + \mathcal{L}_{\text{reg}}$. The regularization term L_{reg} promotes plausible human-object interactions and comprises three components: (i) a *bone-length loss* that penalizes deviations in limb lengths from ground truth, (ii) a *contact loss* that aligns designated human joints with expected contact regions on the object, and (iii) a *velocity loss* that matches predicted object and human motion velocity with ground truth. Full details of these components are provided in Sec. B.1 of the Appendix.

Contact-Aware Shape-Spectrum Augmentation. We employ an optimization-based augmentation strategy to enhance our dataset by transferring original HOI sequences onto novel objects from the same category. Specifically, we first train a correspondence network following Xie et al. (2024) that maps points from the source object's surface to corresponding points on new object surface selected from ShapeNet (Chang et al., 2015) and Objaverse (Deitke et al., 2023) within the same object category. With the learned correspondence, we replace the original object with the novel object, optimizing the placement so that the original human-object contact points are preserved – the new object's corresponding points remain consistently matched to the same human contacts. The optimization objectives are detailed in Sec. B.2 of the Appendix.

4 Experiments

Dataset. We conduct experiments mainly on the InterAct dataset (Xu et al., 2025), and ablate on its major subsets, BEHAVE (Bhatnagar et al., 2022) and OMOMO (Li et al., 2023a). InterAct includes fine-grained textual annotations accompanying these sequences, and we adopt its official training-testing split for all evaluations. As the data originally annotates human motion using SMPL-H (Romero et al., 2017) and SMPL-X (Pavlakos et al., 2019), we standardize all representations to SMPL-H and utilize SMPL-H joints consistently across our experiments, using official SMPL conversion. We manually filter out implausible motion sequences from the original MoCap data; for instance, we exclude cases from OMOMO in which hand orientations were incorrectly inverted, and cases from IMHD where the human is distorted given their shape and pose. We apply object augmentation to enrich the dataset from 217 objects to 1121 objects, with examples in Figure A.

Metrics. Following the standard practice (Guo et al., 2022a), we measure realism and diversity of the generated HOI sequence, alignment with textual descriptions, and physical plausibility of interactions. To evaluate realism and diversity, we utilize the Fréchet Inception Distance (**FID**), which

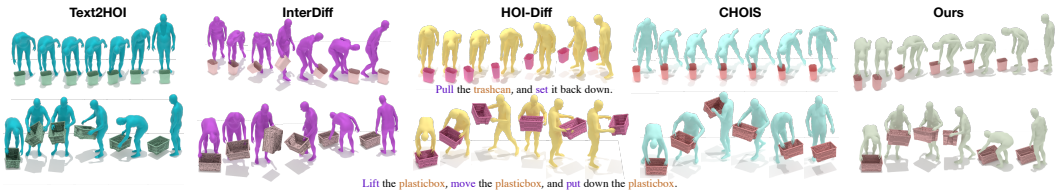

Figure 2: **Qualitative comparison** with baselines. Our method yields more realistic human-object interactions, fewer contact/penetration artifacts, more accurate finger positioning, and better text-motion alignment.

Table 2: **Ablation study** of token-separation strategies on the InterAct dataset (Xu et al., 2025). We report R-Precision with batch sizes 64 and 256.

hand–body separation	human–object separation	R-Precision↑ Batch Size = 64 Top 1	Top 2	Top 3	Batch Size = 256 Top 1	Top 2	Top 3	FID↓	MM Dist↓	Diversity→	FSR↓	Pene↓	Contact→	Interaction↑ C_{prec}	C_{rec}	C_{F1}
✓	✓	**0.781**$^{\pm 0.004}$	**0.927**$^{\pm 0.006}$	**0.975**$^{\pm 0.004}$	**0.498**$^{\pm 0.003}$	**0.730**$^{\pm 0.005}$	**0.855**$^{\pm 0.011}$	**0.188**$^{\pm 0.010}$	**1.863**$^{\pm 0.036}$	**7.547**$^{\pm 0.022}$	**0.031**$^{\pm 0.006}$	**0.051**$^{\pm 0.008}$	0.178$^{\pm 0.003}$	0.791$^{\pm 0.000}$	**0.755**$^{\pm 0.001}$	**0.753**$^{\pm 0.006}$
–	✓	0.755$^{\pm 0.006}$	0.926$^{\pm 0.004}$	0.973$^{\pm 0.006}$	0.492$^{\pm 0.022}$	0.729$^{\pm 0.008}$	0.852$^{\pm 0.011}$	0.197$^{\pm 0.011}$	1.947$^{\pm 0.012}$	7.561$^{\pm 0.165}$	0.048$^{\pm 0.004}$	0.057$^{\pm 0.004}$	0.192$^{\pm 0.003}$	0.789$^{\pm 0.017}$	0.714$^{\pm 0.008}$	0.728$^{\pm 0.008}$
✓	–	0.741$^{\pm 0.004}$	0.917$^{\pm 0.006}$	0.970$^{\pm 0.002}$	0.473$^{\pm 0.011}$	0.713$^{\pm 0.003}$	0.842$^{\pm 0.014}$	0.203$^{\pm 0.002}$	2.004$^{\pm 0.013}$	7.574$^{\pm 0.119}$	0.036$^{\pm 0.002}$	0.064$^{\pm 0.011}$	0.172$^{\pm 0.005}$	**0.811**$^{\pm 0.011}$	0.725$^{\pm 0.020}$	0.736$^{\pm 0.015}$

quantifies feature distribution similarity between generated sequences and ground-truth samples, and a **Diversity** metric, measuring variability across generated HOIs. To assess textual alignment, we adopt **R-Precision**, and Multimodal Distance (**MM Dist**), quantifying the feature-level distance between generated HOIs and corresponding text embeddings. We introduce three metrics tailored explicitly for assessing the plausibility and quality of generated HOI sequences. The Foot Skating Ratio (**FSR**) measures the proportion of frames exhibiting unrealistic foot sliding. The Penetration Ratio (**Pene**) calculates the average fraction of object vertices intersecting the human mesh across the sequence. The Contact Ratio (**Contact**) assesses the frequency with which the human and object maintain consistent contact throughout the motion. More details can be found in Sec. C.2 of the Appendix. Following Li et al. (2023b), we also report contact precision (C_{prec}), recall (C_{rec}), and F1 score (C_{F1}) to measure frame-wise contact accuracy, though these may not fully capture plausibility due to the diversity of text-to-HOI generation (see Sec. C.2 of the Appendix).

Existing methods (Peng et al., 2023; Diller & Dai, 2024; Wu et al., 2024a; Song et al., 2024) typically rely on feature extractors trained using relatively small-scale HOI datasets or their evaluator only measure human motion quality, which can negatively impact the robustness and reliability of evaluations. To address this limitation, we retrain our feature extraction models using the larger-scale InterAct dataset, leveraging regressed joint representations in conjunction with object BPS Prokudin et al. (2019) features. Additional details can be found in Sec. C.1 of the Appendix.

Implementation Details. In the main paper, we fix the modiality pairs by setting $m_1 = \{b, h\}, m_2 = \{o\}$. Additional experimensts on all of the component combinations are presented in the Appendix D. Unless otherwise specified, we set the guidance weights as $\omega_1 = 0.5$, $\omega_2 = 3.0$ with a noise-level offset $\delta = 250$ while the total denoising steps $K = 500$. During training, object geometry is represented using a BPS (Prokudin et al., 2019) comprising 1024 points. Our models are trained on NVIDIA A100 GPUs, leveraging mixed-precision training with FP16 precision and flash attention. Training converges within approximately 24 hours using a single GPU. For the transformer backbone, we adopt an 8-layer transformer decoder structure with a latent dimension of 512 and a feed-forward dimension of 1024. Additional implementation details are in the Sec. B.

Baselines. We compare against four recent HOI generation baselines. HOI-Diff (Peng et al., 2023), a diffusion-based framework for text-driven HOI, employs a transformer backbone (Tevet et al., 2022b) with classifier guidance via an affordance predictor (Dhariwal & Nichol, 2021). CHOIS (Li et al., 2023a), originally requiring inputs beyond text, is adapted here for text-only prompts. InterDiff (Xu et al., 2023b) is modified by replacing its historical motion encoder with a text encoder. Text2HOI (Cha et al., 2024) is incorporated for full-body HOI generation, following their protocol to train a static contact-map estimator. To assess our guidance strategy, we further evaluate two variants: (**i**) our method without guidance and (**ii**) our full model with LIGHT.

Quantitative Evaluation. As shown in Table 1, LIGHT surpasses prior diffusion-based text-to-HOI methods on most key metrics: it attains higher R-precision for tighter text-animation alignment and lower FID and MM Dist scores for better generation quality. Tables E and F demonstrate that our method surpass other baselines, when trained on small-scale datasets, mirroring baseline's setting. Sec. D demonstrate that our framework is able to generalize on two new tasks without retraining, with guidance showing improvement over counterpart without guidance, even on these unseen tasks.

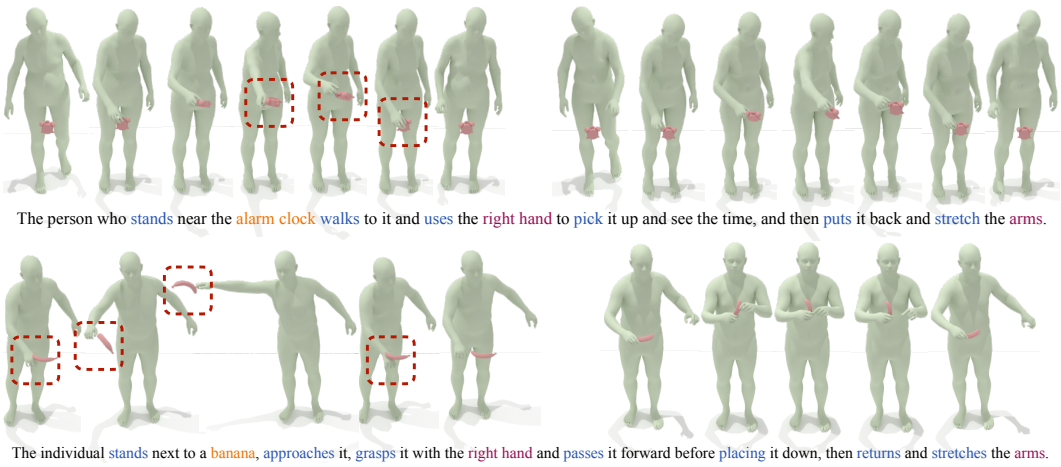

Figure 3: **Qualitative comparison** between our method using body and hand merged into a single token (**left**) versus separating body and hand into distinct tokens (**right**). Unrealistic grasping artifacts produced by the single-token approach are highlighted in red dashed boxes. Our separate-token strategy yields better results.

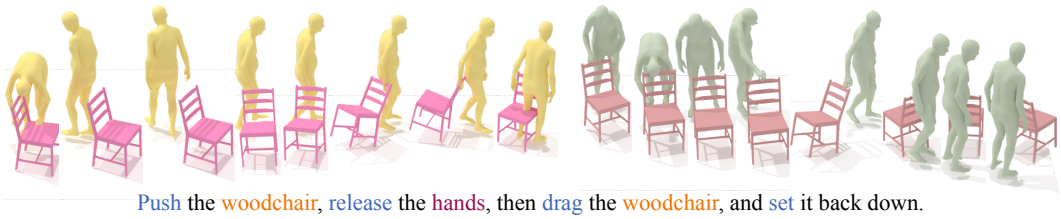

Figure 4: **Qualitative comparison. Left**: our LIGHT *without* guidance. **Right**: our full method *with* guidance, which markedly enhances generation quality.

Qualitative Evaluation. In the baseline comparison (Figure 2), our method attains higher contact accuracy, whereas the baseline exhibits noticeable penetration and floating artifacts, and their finger articulation is often incorrect. We also conduct a user study to complement the qualitative validation, which is discussed in Sec. E, demonstrating that ours produces animation with higher quality.

Impact of Pace-Induced Guidance. We evaluate the impact of our proposed guidance. Figure 4 shows examples comparing HOI sequences generated with and without pace-induced guidance. With guidance enabled, generated animations clearly exhibit more precise contact dynamics and fewer unrealistic motion artifacts. This highlights that our approach effectively enhances the plausibility and overall visual quality of generated results. Table 1 corroborates these observations: combining guidance with our pipeline achieves the strongest performance.

Impact of Separating Tokens. In Figure 3, we show that separate hand token modeling yields noticeably more precise grasping motions and object placements. By comparison, the unified-token baseline often produces unnatural grasps and misaligned objects, underscoring the benefits of our separation strategy. Quantitative results in Table 2 reinforce this trend: token separation for three modalities with the our framework achieves superior performance across all metrics. We also conduct ablation on subset evaluation (Tables H and G), to mirror the setting of existing baselines (Li et al., 2023a; Xu et al., 2023b; Peng et al., 2023) trained on smaller dataset.

Impact of Augmented Data. Table A and Figure A illustrates that our augmentation framework generates realistic human-object interaction animations well suited for training. We evaluate models trained with augmented data under three distinct train-test splits: (i) Partitioned by sequence, consistent with the main experimental setup. Here, the test set contains no novel objects – all objects also appear in training. (ii) Partitioned by object category, where the test set consists of in-category objects unseen during training. (iii) Partitioned by object category, where the test set consists of cross-category objects entirely unseen during training. In all cases, the model is trained on a combination of the original training data and its in-category augmented data, then evaluated on the corresponding test set using our pretrained evaluator without fine-tuning. Table 3 shows that augmentation consistently improves performance across metrics, with especially large gains on the unseen-object evaluation.

Table 3: **Ablation study.** We compare models trained *with* and *without* data augmentation on the InterAct dataset (Xu et al., 2025). Experiments on unseen objects include in-category and cross-category objects never observed during training. We report R-Precision with batch sizes 64 and 256.

Augmented	Unseen	R-Precision↑						FID↓	MM Dist↓	Diversity→	FSR↓	Pene↓	Contact→	Interaction↑		
		Batch Size = 64			Batch Size = 256											
		Top 1	Top 2	Top 3	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
–	Ground Truth	$0.808^{\pm0.002}$	$0.950^{\pm0.000}$	$0.988^{\pm0.000}$	$0.574^{\pm0.000}$	$0.773^{\pm0.005}$	$0.879^{\pm0.000}$	$0.000^{\pm0.000}$	$1.439^{\pm0.001}$	$7.509^{\pm0.049}$	$0.030^{\pm0.009}$	$0.036^{\pm0.000}$	$0.261^{\pm0.000}$	$0.891^{\pm0.000}$	$0.906^{\pm0.000}$	$0.898^{\pm0.000}$
×	×	**0.781**$^{\pm0.004}$	$0.927^{\pm0.006}$	$0.975^{\pm0.004}$	**0.498**$^{\pm0.003}$	$0.730^{\pm0.005}$	$0.855^{\pm0.011}$	$0.188^{\pm0.010}$	$1.863^{\pm0.036}$	**7.547**$^{\pm0.022}$	$0.031^{\pm0.006}$	**0.051**$^{\pm0.008}$	$0.178^{\pm0.003}$	$0.791^{\pm0.000}$	$0.755^{\pm0.001}$	$0.753^{\pm0.006}$
✓	×	$0.770^{\pm0.011}$	**0.930**$^{\pm0.006}$	**0.981**$^{\pm0.004}$	$0.491^{\pm0.014}$	**0.743**$^{\pm0.005}$	**0.860**$^{\pm0.005}$	**0.183**$^{\pm0.023}$	**1.817**$^{\pm0.037}$	$7.576^{\pm0.029}$	**0.030**$^{\pm0.004}$	$0.052^{\pm0.001}$	$0.193^{\pm0.001}$	**0.809**$^{\pm0.004}$	**0.765**$^{\pm0.002}$	**0.772**$^{\pm0.001}$
×	In-category	$0.253^{\pm0.026}$	$0.350^{\pm0.026}$	$0.423^{\pm0.015}$	$0.152^{\pm0.027}$	$0.207^{\pm0.027}$	$0.258^{\pm0.032}$	$0.472^{\pm0.098}$	$3.097^{\pm0.022}$	$5.978^{\pm0.065}$	$0.035^{\pm0.009}$	$0.275^{\pm0.043}$	$0.192^{\pm0.003}$	**0.778**$^{\pm0.034}$	$0.674^{\pm0.005}$	$0.697^{\pm0.014}$
✓	In-category	**0.256**$^{\pm0.026}$	**0.372**$^{\pm0.026}$	**0.439**$^{\pm0.006}$	**0.162**$^{\pm0.008}$	**0.219**$^{\pm0.016}$	**0.264**$^{\pm0.008}$	**0.429**$^{\pm0.067}$	**3.077**$^{\pm0.015}$	**6.036**$^{\pm0.064}$	**0.032**$^{\pm0.009}$	**0.264**$^{\pm0.023}$	$0.209^{\pm0.004}$	$0.770^{\pm0.005}$	**0.710**$^{\pm0.009}$	**0.713**$^{\pm0.005}$
×	Cross-category	**0.066**$^{\pm0.013}$	**0.122**$^{\pm0.004}$	**0.159**$^{\pm0.009}$	**0.029**$^{\pm0.014}$	**0.045**$^{\pm0.019}$	**0.053**$^{\pm0.019}$	$4.455^{\pm0.304}$	**4.938**$^{\pm0.006}$	$5.299^{\pm0.137}$	$0.133^{\pm0.047}$	$0.076^{\pm0.026}$	$0.080^{\pm0.005}$	**0.744**$^{\pm0.025}$	$0.312^{\pm0.024}$	$0.403^{\pm0.020}$
✓	Cross-category	$0.062^{\pm0.004}$	$0.111^{\pm0.015}$	$0.156^{\pm0.009}$	$0.020^{\pm0.005}$	$0.037^{\pm0.008}$	$0.047^{\pm0.000}$	**3.956**$^{\pm0.050}$	$4.981^{\pm0.013}$	**5.619**$^{\pm0.127}$	**0.102**$^{\pm0.058}$	**0.063**$^{\pm0.016}$	$0.078^{\pm0.015}$	$0.742^{\pm0.033}$	**0.331**$^{\pm0.065}$	**0.424**$^{\pm0.080}$

This demonstrates that object-level augmentation substantially enhances generalizability, benefiting both in-category and cross-category unseen objects.

Impact of Guidance Intensity and Denoising lagging. As shown in Figure 5, varying the guidance weight ω_2 and the inter-branch denoising offset δ reveals a clear optimum. Weak guidance (small ω_2) degrades FID; $\omega_2 = 4$ injects just enough prior for the best quality; larger ω_2 overconstrains, reducing diversity and motion fluidity. Likewise, small offsets ($\delta < 100$) keep the staged and uniform branches too similar for effective correction, $\delta = 200$ strikes the best balance, and larger offsets might cause excessive divergence and abrupt fusion.

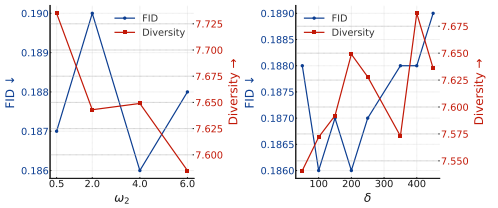

Figure 5: Ablation study on the schedule-based guidance weight ω_2 and denoising preceding offset δ. Left: δ fixed at 200 while varying ω_2. Right: ω_2 fixed at 4.0 while varying δ.

Analysis of Guidance Direction. We analyze how guidance directions – LIGHT and text-conditioned CFG – relate to two reference directions: (**i**) the displacement toward the ground-truth (GT) $\boldsymbol{g}_{\text{GT}} = \hat{\boldsymbol{x}}(\mathbf{0}) - \boldsymbol{x}_S(\boldsymbol{\lambda})$. where $\hat{\boldsymbol{x}}(\mathbf{0})$ is defined in Sec. 3 as ground truth conterpart, and (**ii**) the penetration-reducing gradient ∇L_{pen}. We provides comparisons between two produced guidance directions, where (**i**) the guidance direction of our LIGHT is defined as $\boldsymbol{g}_{\text{LIGHT}} = \mathcal{G}_\theta(\boldsymbol{x}'_S, \boldsymbol{\lambda}', \boldsymbol{d}) - \mathcal{G}_\theta(\boldsymbol{x}_S(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \boldsymbol{d})$, as defined in Equation 6. (**ii**) the text-cfg direction is $\boldsymbol{g}_{\text{CFG}} = \mathcal{G}_\theta(\boldsymbol{x}_U(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \boldsymbol{d}) - \mathcal{G}_\theta(\boldsymbol{x}_U(\boldsymbol{\lambda}), \boldsymbol{\lambda}, \emptyset)\big)$. as defined in Equation 2. We report the mean cosine similarity between two guidance directions and two reference directions. Empirically, our pace-induced guidance consistently steers motion toward the desired distribution, achieving GT-aligned similarity on par with text-conditioned CFG. To probe effects on low-level contact, we also measure the cosine similarity between two directions with and the penetration-reducing gradient ∇L_{pen}. As shown in Table 4, our LIGHT shows more correlation with this gradient compared to pure text CFG. We hypothesize that the hard-dropout form of CFG biases samples toward the marginal data distribution – enhancing global plausibility but diminishing sensitivity to contact-specific cues. In LIGHT, the "unconditional" path retains weak conditioning via noisier, lagged components (*e.g.*, slight human-object surface misalignment). Then, the contrast between clean and noise branch will focus on surface-alignment signals and explicitly promotes depenetration and alignment.

Table 4: Comparison of the mean gradient similarity with penetration-descending direction, direction towards GT distribution. $\delta = 250$ is used and mean gradient similarity is calculated on all the guiding steps. we define $\langle \mathbf{a}, \mathbf{b} \rangle = \frac{\mathbf{a}}{\|\mathbf{a}\|} \cdot \frac{\mathbf{b}}{\|\mathbf{b}\|}$.

Dataset	$<\boldsymbol{g}_{\text{CFG}}, \boldsymbol{g}_{\text{GT}}>$	$<\boldsymbol{g}_{\text{LIGHT}}, \boldsymbol{g}_{\text{GT}}>$	$<\boldsymbol{g}_{\text{CFG}}, \nabla L_{\text{pen}}>$	$<\boldsymbol{g}_{\text{LIGHT}}, \nabla L_{\text{pen}}>$
InterAct (Xu et al., 2025)	0.401	0.401 + 0.002	0.217	0.217 + **0.035**
OMOMO (Li et al., 2023a)	0.395	0.395 + 0.029	0.239	0.239 + **0.032**

5 Conclusion

We introduced LIGHT, a framework for human–object interaction animation that jointly models dynamic human motion and diverse object geometries without relying on handcrafted priors or kinematic constraints. By decoupling modality-specific components with individualized noise schedules, the model enables cleaner elements to guide noisier ones, yielding inherently contact-aware generation. Generalization is further enhanced through synthetic object augmentation, which serves as a data prior to promote invariance of contact semantics across geometric variations. Collectively,

these contributions establish a new form of guidance for HOI animation, providing both a scalable path toward more generalizable modeling and a natural extension of guidance within the diffusion forcing paradigm.

Ethics statement. The realism and flexibility of animations produced by our approach could facilitate misuse, such as generating convincing yet fictitious scenarios involving human interactions with objects, which might lead to misinformation. To proactively mitigate these risks, we explicitly adopt abstract human-body representations—specifically, the SMPL model—which inherently lacks identifiable facial or biometric features. Thus, our approach substantially decreases the likelihood of synthesized content being exploited for identity misrepresentation or other privacy-invasive applications, thereby ensuring responsible and ethical usage.

Reproducibility statement. We provide a detailed description of the model architecture in Sec. 3. Details of adapted baselines are introduced in Sec. 4. For augmentation, we include all the details in Sec. B.2. Comprehensive implementation details, including hyperparameters, evaluation metrics, and the training procedure of the evaluator, are reported in the Sec. 3 and further expanded in Sec. B, to facilitate reproducibility.

REFERENCES

CMU graphics lab motion capture database. http://mocap.cs.cmu.edu/. 2

German Barquero, Sergio Escalera, and Cristina Palmero. BeLFusion: Latent diffusion for behavior-driven human motion prediction. In *ICCV*, 2023. 2

German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *CVPR*, 2024. 2

Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *CVPR*, 2022. 3, 6, 18, 19, 20

Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020. 2

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *CVPR*, 2024. 4, 6, 7, 15, 20

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 6, 16, 19

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *NeurIPS*, 2024. 2, 3, 6

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 2, 6, 16, 19

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3, 7

Christian Diller and Angela Dai. CG-HOI: Contact-guided 3d human-object interaction generation. In *CVPR*, 2024. 1, 3, 7

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 *CVPR*, 2023. 3

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022a. 6, 15, 17

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*, 2022b. 2

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, 2019. 2

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021. 2

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 3

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *GCPR*, 2022. 3

Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. CHAIRS: Towards full-body articulated human-object interaction. In *ICCV*, 2023. 3

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *CVPR*, 2024. 3

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. GMD: Controllable human motion synthesis via guided diffusion models. In *ICCV*, 2023. 2, 18

Gwanghyun Kim, Alonso Martinez, Yu-Chuan Su, Brendan Jou, José Lezama, Agrim Gupta, Lijun Yu, Lu Jiang, Aren Jansen, Jacob Walker, et al. A versatile diffusion transformer with mixture of noise levels for audiovisual generation. *Advances in Neural Information Processing Systems*, 37: 11837–11865, 2024a. 6

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. ParaHome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. *arXiv preprint arXiv:2401.10232*, 2024b. 3

Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. NIFTY: Neural object interaction fields for guided human motion synthesis. *arXiv preprint arXiv:2307.07511*, 2023. 2

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. *arXiv preprint arXiv:2312.03913*, 2023a. 1, 3, 6, 7, 8, 9, 17, 18, 19, 20, 21

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023b. 3, 7, 15, 17

Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. *arXiv preprint arXiv:2303.13129*, 2023c. 2

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. InterGen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 2

Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arxiv:2310.12978*, 2023. 4, 17

Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *3DV*, 2018. 2

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 17

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 6

Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. HOI-Diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 1, 3, 6, 7, 8, 15, 20

Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *CVPR*, 2023. 2

Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, 2022. 2

Mathis Petrovich, Michael J Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *ICCV*, 2023. 2, 17

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *ICCV*, 2019. 5, 7

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 16

Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. Single motion diffusion. *arXiv preprint arXiv:2302.05905*, 2023. 2

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 17

Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *CVPR*, 2023. 2

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. 3, 6

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 6

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao Chen, Ning Li, and Hong Qin. Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *CVPR*, 2024. 3, 7

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 3, 4, 17, 20

Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. GOAL: Generating 4d whole-body motion for hand-object grasping. In *CVPR*, 2022. 2

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *ECCV*, 2022a. 2

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022b. 7

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 2, 3

Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *3DV*, 2022. 2

Qianyang Wu, Ye Shi, Xiaoshui Huang, Jingyi Yu, Lan Xu, and Jingya Wang. THOR: Text to human-object interaction diffusion via relation intervention. *arXiv preprint arXiv:2403.11208*, 2024a. 3, 7

Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. SAGA: Stochastic whole-body grasping with contact. In *ECCV*, 2022. 2

Zhen Wu, Jiaman Li, and C Karen Liu. Human-object interaction from human-level instructions. *arXiv preprint arXiv:2406.17840*, 2024b. 3

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *ECCV*, 2022. 2

Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. InterTrack: Tracking human object interaction without object templates. *arXiv preprint arXiv:2408.13953*, 2024. 6, 16

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*, 2023. 2

Jingqiao Xiu, Fangzhou Hong, Yicong Li, Mengze Li, Wentao Wang, Sirui Han, Liang Pan, and Ziwei Liu. Egotwin: Dreaming body and view in first person. *arXiv preprint arXiv:2508.13013*, 2025. 6

Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. *arXiv preprint arXiv:2312.16051*, 2023a. 2

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023b. 3, 6, 7, 8, 15, 20

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *arXiv preprint arXiv:2403.19652*, 2024. 1, 3

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. InterAct: Advancing large-scale versatile 3d human-object interaction generation. In *CVPR*, 2025. 6, 7, 9, 17, 18, 19, 21

Jie Yang, Xuesong Niu, Nan Jiang, Ruimao Zhang, and Siyuan Huang. F-HOI: Toward fine-grained semantic-aligned 3d human-object interactions. In *ECCV*, 2024. 3

Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023. 2

Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. ArtiGrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. *arXiv preprint arXiv:2309.03891*, 2023a. 2

Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020. 2

Jiajun Zhang, Yuxiang Zhang, Liang An, Mengcheng Li, Hongwen Zhang, Zonghai Hu, and Yebin Liu. Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. *arXiv preprint arXiv:2409.09300*, 2024a. 2

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023b. 2

Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023c. 3

Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *CVPR*, 2024b. 3

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022a. 2

Wanyue Zhang, Rishabh Dabral, Vladislav Golyanik, Vasileios Choutas, Eduardo Alvarado, Thabo Beeler, Marc Habermann, and Christian Theobalt. Bimart: A unified approach for the synthesis of 3d bimanual interaction with articulated objects. *arXiv preprint arXiv:2412.05066*, 2, 2024c. 5

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. COUCH: Towards controllable human-chair interactions. In *ECCV*, 2022b. 2

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. *arXiv preprint arXiv:2306.10900*, 2023d. 2

Zihan Zhang, Richard Liu, Kfir Aberman, and Rana Hanocka. TEDi: Temporally-entangled diffusion for long-term motion synthesis. *arXiv preprint arXiv:2307.15042*, 2023e. 2

Zihan Zhang, Richard Liu, Rana Hanocka, and Kfir Aberman. Tedi: Temporally-entangled diffusion for long-term motion synthesis. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024d. 6

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

Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. CAMS: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *CVPR*, 2023. 2

Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *ECCV*. Springer, 2022a. 2

Keyang Zhou, Bharat Lal Bhatnagar, Bernt Schiele, and Gerard Pons-Moll. Adjoint rigid transform network: Task-conditioned alignment of 3d shapes. In *2022 international conference on 3D vision (3DV)*, pp. 1–11. IEEE, 2022b. 16

6 APPENDIX

In the Appendix, we provide additional methodological details and experimental results: (i) a demo video, described in Sec. A; (ii) an expanded explanation of our loss formulation and object-augmentation procedure in Sec. B; (iii) definitions of the proposed metrics in Sec. C.2; and (iv) extra experimental results, omitted from the main paper for space, in Sec. E.

LLM Usage. We employ large language models (LLMs), such as ChatGPT, to assist in polishing our paper. Specifically, LLMs are used to correct grammatical errors, refine word choice, and improve overall fluency. We do not use LLM in formulating our methodology or running experiments

A VISUALIZATION VIDEO

Beyond the qualitative results in the main paper, we provide an accompanying video with richer visualizations. These demos highlight (**i**) qualitative pairs of *augmented inputs vs. ground-truth* trajectories to illustrate the diversity introduced by our augmentation pipeline; (**ii**) a qualitative comparison between our method and the existing text-to-HOI baselines (Peng et al., 2023; Li et al., 2023b; Xu et al., 2023b; Cha et al., 2024); (**iii**) comparisons *with vs. without augmentation* to demonstrate how augmentation improves generalization and visual plausibility; (**iv**) an ablation study of our guidance mechanism, showing that our proposed full guidance is able to further improve the quality of the generation.

B ADDITIONAL DETAILS OF METHODOLOGY

B.1 TRAINING LOSS

In training, we regularize our LIGHT with three loss terms, foot-skating loss (L_{fs}), velocity loss (L_v), and contact loss (L_{cont}). The total regularization objective is

$$L_{\text{reg}} = \lambda_{\text{fs}} L_{\text{fs}} + \lambda_{\text{v}} L_{\text{v}} + \lambda_{\text{cont}} L_{\text{cont}}, \tag{8}$$

where we set $\lambda_{\text{fs}} = 1, \lambda_{\text{v}} = 0.02, \lambda_{\text{cont}} = 0.1$.

Foot-skating loss. To constrain foot sliding during ground contact, we penalize the deviation between predicted and ground-truth foot velocities:

$$L_{\text{fs}} = \sum_{t=1}^{T} \sum_{f=1}^{4} c_t^f \left\| \left(\boldsymbol{j}_{t,f}^p - \boldsymbol{j}_{t-1,f}^p \right) - \left(\hat{\boldsymbol{j}}_{t,f}^p - \hat{\boldsymbol{j}}_{t-1,f}^p \right) \right\|_2^2, \tag{9}$$

where $c_t^f \in \{0, 1\}$ is the ground-truth foot-contact label for joint f at time t, and $\hat{\boldsymbol{j}}_{t,f}^p$ (*resp.* $\hat{\boldsymbol{j}}_{t,f}^p$) denotes the predicted (*resp.* ground-truth) 3-D position of that joint. We follow (Guo et al., 2022a) for the foot-joint definitions and contact annotation.

Velocity loss. To encourage smooth temporal evolution of both human and object trajectories, we match first-order differences between predictions and ground truth:

$$\begin{aligned} L_{\text{v}} \; = \lambda_{\text{pv}} & \sum_{t=1}^{T} \sum_{j=1}^{J} \left\| \left(\boldsymbol{j}_{t,j}^p - \boldsymbol{j}_{t-1,j}^p \right) - \left(\hat{\boldsymbol{j}}_{t,j}^p - \hat{\boldsymbol{j}}_{t-1,j}^p \right) \right\|_2^2 \\ & + \lambda_{\text{otv}} \sum_{t=1}^{T} \left\| \left(\boldsymbol{o}_t^t - \boldsymbol{o}_{t-1}^t \right) - \left(\hat{\boldsymbol{o}}_t^t - \hat{\boldsymbol{o}}_{t-1}^t \right) \right\|_2^2 \\ & + \lambda_{\text{orv}} \sum_{t=1}^{T} \left\| \left(\boldsymbol{o}_t^r - \boldsymbol{o}_{t-1}^r \right) - \left(\hat{\boldsymbol{o}}_t^r - \hat{\boldsymbol{o}}_{t-1}^r \right) \right\|_2^2, \end{aligned} \tag{10}$$

where $\boldsymbol{j}_{t,j}^p$ is the position of human joint j, while $\boldsymbol{o}_t^t$ and $\boldsymbol{o}_t^r$ are the object translation and rotation at time t. We set $\lambda_{\text{pv}} = 1, \lambda_{\text{otv}} = 1, \lambda_{\text{orv}} = 1$.

Contact loss. To promote accurate human–object interactions, we minimize joint-to-surface distances whenever contact is expected:

$$L_{\text{cont}} = \sum_{t=1}^{T} \sum_{j=1}^{J} \left(d\big(\hat{\boldsymbol{j}}_{t,j}^{p},\, \hat{\boldsymbol{V}}_{t}^{o}\big)\, \hat{c}_{t}^{j} \right)^{2}, \tag{11}$$

where $d(\cdot, \cdot)$ returns the minimum Euclidean distance between joint $\hat{\boldsymbol{j}}_{t,j}^{p}$ and the set of ground truth object vertices $\boldsymbol{V}_{t}^{o}$, and c_{t}^{j} is the ground-truth binary contact label for joint j at time t, where we extract this information from ground truth data whether any joint to mesh distance is less than 0.03 m.

B.2 Additional Details on Object Augmentation

We employ an optimization-based strategy to enhance our dataset by transferring each original human-object interaction (HOI) sequence onto a novel object instance of the same category. Specifically, given an original HOI sequence with human motion and an object trajectory, we first train a dense correspondence network following Xie et al. (2024); Zhou et al. (2022b). Specifically, first, we extract the AABB (Axis-Aligned Bounding Box) of the object, and normalize to make the diagram of AABB to be 1, and the center of AABB to be at the origin. Then we train the corresponding network following Zhou et al. (2022b). The model architecture consists of a PointNet (Qi et al., 2017) and multilayer perceptron (MLP), which maps an unordered point cloud to the orientation. We follow Zhou et al. (2022b) to supervise the model with a chamfer distance loss. This network learns to map points on the source object's surface to corresponding points on a new object instance selected from a 3D shape repository in the same category, *e.g.*, ShapeNet (Chang et al., 2015) or Objaverse (Deitke et al., 2023). Using the learned correspondence, we replace the original object in the HOI sequence with the novel object, obtaining an initial transformed trajectory for the new object.

The goal is to position and move the new object such that the human's contact interactions remain consistent with the original sequence. In other words, for each contact constraint observed in the original sequence, *i.e.,* a specific human joint maintaining contact at frame with the original object, the corresponding surface point on the new object, identified via the correspondence network, should likewise remain in contact with that same human joint. We achieve this by optimizing the new object's pose parameters and the original human parameters over time to preserve the spatial alignment of these contact points while maintaining the physical plausibility of the interaction. To refine the new object's trajectory, we formulate an objective function that is a weighted sum of multiple alignment terms:

$$\mathcal{L} = \lambda_{\text{con}} L_{\text{con}} + \lambda_{\text{normal}} L_{\text{normal}} + \lambda_{\text{colli}} L_{\text{colli}} + \lambda_{\text{init}} L_{\text{init}} + \lambda_{\text{acc}} L_{\text{acc}}, \tag{12}$$

where λ_{con}, λ_{normal}, λ_{colli}, λ_{init}, and λ_{acc} are weighting coefficients. Each loss term L_{*} is designed to enforce a particular aspect of alignment between the human and the new object:

- L_{con}: penalizes any deviation in the distance between the human's contact points and the new object's corresponding surface points.
- L_{normal}: encourages the surface normals of the new object at contact regions to align with those of the source object, ensuring that the orientation of the object relative to the human's contacting limb (*e.g.*, a hand grasp) remains consistent with the original interaction.
- L_{colli}: penalizes interpenetration or unintended collisions between the human body and the new object, enforcing that the object remains outside the human geometry except at the intended contact regions.
- L_{init}: regularizes the solution towards the original motion.
- L_{acc}: promotes smooth motion by penalizing abrupt changes in velocity and acceleration.

By minimizing $\mathcal{L}$, we refine the human and novel object's trajectory so that the human's interaction with the object remains natural and consistent with the original sequence. Thus, the new object is placed and moved in such a way that its correspondence-defined contact points remain matched to the human's grasp or touch points throughout the sequence.

Table A: Quantitative evaluation of augmented data quality relative to ground truth annotations.

Dataset	Data source	Pene$^{\downarrow}$	Floating$^{\downarrow}$	$CH^{\uparrow}_{prec}$	$CH^{\uparrow}_{rec}$	$CH^{\uparrow}_{F1}$
InterAct (Xu et al., 2025)	Ground Truth	0.008	0.005	0.947	0.911	0.920
	Augmentation	0.010	0.011	0.980	0.980	0.980
GRAB (Taheri et al., 2020)	Ground Truth	0.003	0.000	1.000	1.000	1.000
	Augmentation	0.004	0.000	0.994	0.998	0.996

C Additional Details of Experimental Setup

C.1 Evaluator representation and evaluator training

Our evaluator takes as input the global human joint locations, object rotation, object translation, and object dynamic BPS. The dynamic BPS is computed using a static basis point set, following Li et al. (2023b); Xu et al. (2025). Consequently, the evaluator assesses the quality of the entire HOI sequence rather than evaluating human or object motion in isolation. Following Lu et al. (2023); Petrovich et al. (2023), we jointly train the text and HOI motion encoders of our evaluator. Departing from traditional training on classification tasks (Guo et al., 2022a), we utilize a sequence-level contrastive learning objective based on the InfoNCE loss (Oord et al., 2018), following recent implementations (Lu et al., 2023; Petrovich et al., 2023). Our text encoding module incorporates Sentence-BERT (Reimers & Gurevych, 2019).

C.2 Evaluation Metrics

We outline our metric definitions below. Standard text-to-motion metrics are not restated; we simply extend them with the object and contact representations introduced in the main paper. For their original formulations, see Guo et al. (2022a).

Contact Percentage. This metric measures how consistently the human's body joints stay in contact with the object, indicating the quality of physical interaction. Extending Li et al. (2023b) from hand contact to whole-body contact, we define

$$\text{Contact} = \frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}\mathbf{1}\Big(d\big(\boldsymbol{j}^{p}_{t,j},\, \boldsymbol{V}^{o}_{t}\big) < \gamma\Big), \tag{13}$$

where $\boldsymbol{j}^{p}_{t,j} \in \mathbb{R}^3$ is the 3-D position of the j-th joint at time t, $\boldsymbol{V}^{o}_{t}$ is the set of object vertices at time t, and $d(\boldsymbol{j}^{p}_{t,j}, \boldsymbol{V}^{o}_{t})$ is the minimum distance from the joint to the object surface. We use the contact threshold $\gamma = 0.05$ following Li et al. (2023a). The more alignment of values of Equation equation 13 with the ground truth indicates better quality.

Frame-Wise Contact Matching. Following Li et al. (2023b), we report frame-wise contact matching, where C_{F1}, C_{prec}, and C_{rec} denote the F1 score, precision, and recall of detected contacts between hands and objects. We use a contact threshold of $\gamma = 0.05$, consistent with Li et al. (2023a). However, this metric is not entirely suitable for the text-to-HOI task, since frame-level alignment is not strictly enforced. In generative tasks conditioned only on text, it is standard practice to evaluate outputs with distributional metrics (*e.g.,* FID in text-to-motion or text-to-image), rather than element- or frame-wise measures such as MPJPE or MSE. Nevertheless, we include frame-wise contact scores as they provide a partial indication of contact quality.

Penetration. Penetration quantifies geometric violations between the human mesh and the object. After reconstructing the full human mesh via forward kinematics, we compute

$$\text{Penetration} = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\Big|\min\big(\text{sdf}^{h_t}(\boldsymbol{v}_{o_i}),\, 0\big)\Big|, \tag{14}$$

where $\text{sdf}^{h_t}(\boldsymbol{v}_{o_i})$ is the signed distance from object vertex $\boldsymbol{v}_{o_i}$ to the human mesh at time t. Negative SDF values indicate penetration depth; clamping positive values to zero ensures that only inter-penetrations are accumulated. Lower scores in Eq. equation 14 correspond to fewer or shallower penetrations.

Table B: **Quantitative evaluation** of unconditional generation on the OMOMO dataset (Li et al., 2023a). For R-Precision we adopt a batch size of 64. We don't retrain our model for unconditional generation.

m_1	m_2	FID↓	Diversity→	FSR↓	Pene↓	Contact→	Interaction↑		
							C_{prec}	C_{rec}	C_{F1}
–	–	$0.400^{\pm 0.003}$	$\mathbf{7.364}^{\pm 0.021}$	$\mathbf{0.014}^{\pm 0.004}$	$0.086^{\pm 0.003}$	$0.180^{\pm 0.004}$	$0.803^{\pm 0.009}$	$0.636^{\pm 0.015}$	$0.683^{\pm 0.013}$
o, h	b	$0.348^{\pm 0.014}$	$7.407^{\pm 0.064}$	$0.015^{\pm 0.007}$	$0.084^{\pm 0.010}$	$0.185^{\pm 0.003}$	$0.803^{\pm 0.009}$	$0.636^{\pm 0.029}$	$0.685^{\pm 0.027}$
b, h	o	$0.373^{\pm 0.020}$	$7.216^{\pm 0.087}$	$\mathbf{0.014}^{\pm 0.004}$	$\mathbf{0.075}^{\pm 0.004}$	$0.185^{\pm 0.010}$	$\mathbf{0.812}^{\pm 0.002}$	$0.651^{\pm 0.031}$	$\mathbf{0.699}^{\pm 0.027}$
b, o	h	$0.354^{\pm 0.011}$	$7.243^{\pm 0.074}$	$0.016^{\pm 0.048}$	$0.082^{\pm 0.014}$	$0.183^{\pm 0.004}$	$0.801^{\pm 0.010}$	$0.638^{\pm 0.030}$	$0.686^{\pm 0.024}$
b	o, h	$\mathbf{0.345}^{\pm 0.018}$	$7.279^{\pm 0.172}$	$0.016^{\pm 0.007}$	$0.078^{\pm 0.020}$	$0.181^{\pm 0.010}$	$0.807^{\pm 0.006}$	$0.637^{\pm 0.035}$	$0.685^{\pm 0.028}$
o	b, h	$0.380^{\pm 0.024}$	$7.432^{\pm 0.164}$	$0.015^{\pm 0.002}$	$0.089^{\pm 0.002}$	$0.191^{\pm 0.011}$	$0.810^{\pm 0.000}$	$\mathbf{0.653}^{\pm 0.034}$	$\mathbf{0.699}^{\pm 0.025}$
h	b, o	$0.348^{\pm 0.005}$	$7.285^{\pm 0.030}$	$0.015^{\pm 0.007}$	$0.083^{\pm 0.022}$	$0.186^{\pm 0.007}$	$0.809^{\pm 0.000}$	$0.645^{\pm 0.022}$	$0.692^{\pm 0.018}$

Table C: **Quantitative comparisons** of controllable human motion generation conditioned on object motion and text. We report R-Precision with batch sizes 64 and 256.

Dataset	m_1	m_2	R-Precision↑ Batch Size = 64			R-Precision↑ Batch Size = 256			FID↓	MPJPE↓	HJPE↓	FSR↓	Pene↓	Contact→	Interaction↑		
			Top 1	Top 2	Top 3	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
OMOMO (Li et al., 2023a)	–	–	$\mathbf{0.672}^{\pm 0.000}$	$\mathbf{0.906}^{\pm 0.000}$	$\mathbf{0.961}^{\pm 0.000}$	$\mathbf{0.322}^{\pm 0.003}$	$\mathbf{0.549}^{\pm 0.003}$	$\mathbf{0.711}^{\pm 0.000}$	$0.105^{\pm 0.003}$	$227.862^{\pm 0.853}$	$419.108^{\pm 2.486}$	$0.014^{\pm 0.008}$	$0.782^{\pm 0.009}$	$0.065^{\pm 0.001}$	$0.742^{\pm 0.004}$	$0.296^{\pm 0.004}$	$0.396^{\pm 0.002}$
	h	b	$0.668^{\pm 0.005}$	$0.898^{\pm 0.000}$	$\mathbf{0.961}^{\pm 0.000}$	$0.316^{\pm 0.000}$	$0.547^{\pm 0.000}$	$0.707^{\pm 0.005}$	$0.105^{\pm 0.003}$	$\mathbf{222.927}^{\pm 1.018}$	$\mathbf{406.182}^{\pm 0.193}$	$\mathbf{0.013}^{\pm 0.005}$	$\mathbf{0.761}^{\pm 0.040}$	$0.066^{\pm 0.001}$	$\mathbf{0.801}^{\pm 0.010}$	$0.300^{\pm 0.003}$	$0.400^{\pm 0.008}$
	b	h	$0.668^{\pm 0.005}$	$0.902^{\pm 0.005}$	$\mathbf{0.961}^{\pm 0.000}$	$0.318^{\pm 0.003}$	$0.547^{\pm 0.000}$	$0.707^{\pm 0.000}$	$\mathbf{0.104}^{\pm 0.002}$	$223.435^{\pm 0.681}$	$407.231^{\pm 0.222}$	$\mathbf{0.013}^{\pm 0.006}$	$0.778^{\pm 0.039}$	$0.066^{\pm 0.001}$	$0.788^{\pm 0.014}$	$\mathbf{0.302}^{\pm 0.006}$	$\mathbf{0.402}^{\pm 0.010}$
InterAct (Xu et al., 2025)	–	–	$\mathbf{0.797}^{\pm 0.000}$	$\mathbf{0.949}^{\pm 0.005}$	$\mathbf{0.992}^{\pm 0.000}$	$\mathbf{0.551}^{\pm 0.005}$	$\mathbf{0.766}^{\pm 0.005}$	$\mathbf{0.873}^{\pm 0.003}$	$0.228^{\pm 0.010}$	$214.467^{\pm 0.456}$	$352.036^{\pm 1.491}$	$0.015^{\pm 0.006}$	$\mathbf{0.585}^{\pm 0.034}$	$0.084^{\pm 0.003}$	$0.691^{\pm 0.011}$	$0.364^{\pm 0.001}$	$0.415^{\pm 0.009}$
	h	b	$0.785^{\pm 0.005}$	$\mathbf{0.949}^{\pm 0.005}$	$0.984^{\pm 0.000}$	$0.547^{\pm 0.001}$	$\mathbf{0.873}^{\pm 0.003}$	$\mathbf{0.873}^{\pm 0.003}$	$0.224^{\pm 0.005}$	$\mathbf{209.672}^{\pm 2.364}$	$346.499^{\pm 3.214}$	$\mathbf{0.013}^{\pm 0.004}$	$0.605^{\pm 0.064}$	$0.086^{\pm 0.005}$	$\mathbf{0.695}^{\pm 0.015}$	$\mathbf{0.380}^{\pm 0.003}$	$\mathbf{0.425}^{\pm 0.009}$
	b	h	$0.785^{\pm 0.005}$	$\mathbf{0.949}^{\pm 0.005}$	$0.984^{\pm 0.000}$	$\mathbf{0.547}^{\pm 0.001}$	$\mathbf{0.766}^{\pm 0.005}$	$\mathbf{0.873}^{\pm 0.003}$	$\mathbf{0.220}^{\pm 0.005}$	$211.505^{\pm 0.775}$	$\mathbf{346.246}^{\pm 3.554}$	$0.015^{\pm 0.005}$	$0.640^{\pm 0.102}$	$0.085^{\pm 0.005}$	$0.687^{\pm 0.013}$	$0.378^{\pm 0.003}$	$0.419^{\pm 0.010}$
BEHAVE (Bhatnagar et al., 2022)	–	–	$\mathbf{0.672}^{\pm 0.000}$	$\mathbf{0.906}^{\pm 0.000}$	$\mathbf{0.961}^{\pm 0.000}$	$0.381^{\pm 0.019}$	$0.623^{\pm 0.024}$	$\mathbf{0.736}^{\pm 0.003}$	$0.105^{\pm 0.003}$	$227.862^{\pm 0.853}$	$419.108^{\pm 2.486}$	$0.014^{\pm 0.008}$	$0.782^{\pm 0.009}$	$0.065^{\pm 0.001}$	$0.742^{\pm 0.004}$	$0.296^{\pm 0.004}$	$0.396^{\pm 0.002}$
	h	b	$0.668^{\pm 0.005}$	$0.898^{\pm 0.000}$	$0.961^{\pm 0.000}$	$\mathbf{0.393}^{\pm 0.014}$	$\mathbf{0.631}^{\pm 0.019}$	$0.732^{\pm 0.019}$	$0.105^{\pm 0.003}$	$\mathbf{222.927}^{\pm 1.018}$	$\mathbf{406.182}^{\pm 0.193}$	$\mathbf{0.013}^{\pm 0.005}$	$\mathbf{0.761}^{\pm 0.040}$	$0.066^{\pm 0.001}$	$\mathbf{0.801}^{\pm 0.010}$	$0.300^{\pm 0.003}$	$0.400^{\pm 0.008}$
	b	h	$0.668^{\pm 0.005}$	$0.902^{\pm 0.005}$	$\mathbf{0.961}^{\pm 0.000}$	$0.383^{\pm 0.005}$	$0.627^{\pm 0.014}$	$0.729^{\pm 0.011}$	$\mathbf{0.104}^{\pm 0.002}$	$223.435^{\pm 0.681}$	$407.231^{\pm 0.222}$	$\mathbf{0.013}^{\pm 0.006}$	$0.778^{\pm 0.039}$	$0.066^{\pm 0.001}$	$0.788^{\pm 0.014}$	$\mathbf{0.302}^{\pm 0.006}$	$\mathbf{0.402}^{\pm 0.010}$

Foot-Skating Ratio (FSR). FSR gauges how well the character's feet remain stationary during ground-contact phases, a key indicator of motion realism. Following Karunratanakul et al. (2023), we compute

$$\text{FSR} = \frac{1}{T}\sum_{t=1}^{T} c_t^F \, \mathbf{1}\Big(\big\|(\boldsymbol{j}_{t,f}^p - \boldsymbol{j}_{t-1,f}^p)\cdot \text{fps}\big\| < \epsilon\Big), \tag{15}$$

where $c_t^F \in \{0,1\}$ is the ground-contact indicator for a foot joint at time t, $\boldsymbol{j}_{t,f}^p$ is that foot's position, and fps converts frame-to-frame displacement to velocity. We adopt $\epsilon = 0.025$ and a ground-contact height threshold of 0.05 as in Karunratanakul et al. (2023). FSR ranges from 0 to 1; higher values signify less foot sliding and more realistic foot-ground interaction.

D Additional Analysis of Guidance

Quantitative Evaluation of Our Guidance Across Tasks. In addition to the Text-to-HOI task presented in the main paper, we evaluate our guidance under different task settings. Specifically, we consider (i) *unconditional HOI generation*, where the model takes as input an object mesh and a human shape to synthesize a complete HOI sequence without extra conditioning, and (ii) *controllable HOI generation*, where the model is conditioned on a richer set of inputs—including the full object motion sequence, object mesh, human shape, and a textual description—to generate the corresponding HOI sequence. For both settings, we reuse the model from the main paper without retraining. This is made possible by our independent noise scheduling, which allows the model to noised out unused conditioning signals or overlap one denoising modality with input condition. Tables B and C demonstrate that our guidance remains effective across multiple tasks.

Analysis of Different Component Combinations. In this section, we experiment with multiple component combinations. Table D demonstrates that our guidance is effective on flexible component combinations.

E Other Additional Experimental Results

User Study. We conducted a user study with 15 participants. Among them, 86.7% reported that our augmented data exhibits fewer artifacts than the ground truth (GT), while 100% agreed that the interaction consistency between GT and augmented sequences is preserved. Additional statistics are provided in Figure B, which further demonstrate that (**i**) our method qualitatively outperforms the baseline, and (**ii**) the proposed guidance mechanism leads to improved performance.

Table D: **Ablation** of modality combinations m_1, m_2 on InterAct (Xu et al., 2025), BEHAVE (Bhatnagar et al., 2022), and OMOMO (Li et al., 2023a) datasets. We report R-Precision with batch sizes 64 and 256.

Dataset	ω_1	m_1	m_2	R-Precision↑ Batch Size = 64 Top 1	Top 2	Top 3	R-Precision↑ Batch Size = 256 Top 1	Top 2	Top 3	FID↓	MM Dist↓	Diversity→	FSR↓	Pene↓	Contact→	Interaction↑ C_{prec}	C_{rec}	C_{F1}
BEHAVE (Bhatnagar et al., 2022)	0.0	–	–	0.504±0.005	0.723±0.005	0.812±0.011	0.252±0.003	0.436±0.024	0.529±0.014	0.714±0.039	3.199±0.060	6.727±0.286	0.259±0.047	0.107±0.013	0.137±0.005	0.615±0.003	0.567±0.007	0.552±0.003
		o, h	***b***	0.512±0.005	**0.715**±0.005	0.816±0.016	0.256±0.003	0.428±0.024	0.525±0.008	0.703±0.033	3.198±0.056	6.954±0.338	0.258±0.047	0.110±0.013	0.186±0.004	0.616±0.004	0.572±0.001	0.555±0.001
		b, h	***o***	0.520±0.005	0.711±0.000	**0.824**±0.016	0.256±0.014	**0.441**±0.032	**0.533**±0.014	0.706±0.040	3.191±0.066	6.935±0.003	0.259±0.048	0.109±0.014	0.187±0.002	0.614±0.003	0.571±0.005	0.554±0.002
		b, o	***h***	0.516±0.022	0.711±0.011	0.820±0.011	0.252±0.008	0.436±0.030	0.527±0.016	0.706±0.023	3.199±0.061	6.864±0.008	**0.257**±0.047	**0.108**±0.011	0.187±0.002	**0.618**±0.004	0.575±0.002	0.557±0.000
		b	***o, h***	0.516±0.032	0.703±0.011	**0.824**±0.005	0.254±0.016	0.430±0.005	0.536±0.005	0.697±0.029	**3.189**±0.068	**6.998**±0.394	**0.257**±0.047	0.109±0.016	0.187±0.001	0.616±0.005	0.571±0.004	0.554±0.002
		o	***b, h***	0.508±0.022	0.703±0.022	0.816±0.016	0.252±0.008	0.430±0.005	0.526±0.003	**0.696**±0.038	3.208±0.050	6.657±0.240	0.260±0.049	0.111±0.016	0.189±0.001	0.617±0.004	**0.584**±0.013	**0.564**±0.004
		h	***b, o***	**0.523**±0.022	0.711±0.000	0.820±0.022	**0.261**±0.018	0.436±0.001	0.531±0.010	0.699±0.024	3.195±0.053	6.845±0.258	0.258±0.047	0.111±0.015	0.187±0.005	0.615±0.004	0.572±0.004	0.554±0.004
	0.5	–	–	0.488±0.027	0.691±0.016	0.793±0.027	0.234±0.011	**0.426**±0.011	**0.516**±0.022	0.764±0.053	3.233±0.023	**6.907**±0.116	0.252±0.043	0.114±0.018	0.187±0.003	0.617±0.003	0.569±0.007	0.557±0.003
		o, h	***b***	0.496±0.027	0.676±0.027	0.793±0.027	0.242±0.011	0.420±0.011	0.506±0.005	0.752±0.056	3.226±0.019	6.763±0.571	0.253±0.041	0.116±0.017	0.187±0.003	0.617±0.003	0.575±0.004	0.559±0.002
		b, h	***o***	0.500±0.011	**0.699**±0.027	0.797±0.022	0.242±0.005	0.420±0.032	0.510±0.024	0.746±0.063	**3.221**±0.034	6.569±0.201	0.252±0.042	**0.112**±0.011	0.188±0.002	0.618±0.004	0.581±0.002	0.563±0.001
		b, o	***h***	**0.508**±0.011	0.684±0.016	0.801±0.016	0.240±0.003	0.429±0.005	0.508±0.014	0.736±0.037	3.223±0.020	6.793±0.058	0.252±0.041	0.118±0.019	0.191±0.000	0.615±0.000	**0.583**±0.009	**0.564**±0.006
		b	***o, h***	0.504±0.016	0.691±0.016	**0.805**±0.011	0.242±0.000	0.416±0.011	0.514±0.011	**0.726**±0.037	**3.221**±0.018	6.646±0.038	**0.251**±0.041	0.118±0.018	0.190±0.000	0.614±0.000	0.580±0.006	0.562±0.004
		o	***b, h***	0.500±0.032	0.680±0.011	0.789±0.011	**0.246**±0.005	0.422±0.003	0.504±0.003	0.744±0.048	3.234±0.044	6.875±0.462	0.260±0.046	**0.112**±0.012	0.189±0.000	**0.623**±0.011	**0.583**±0.011	0.563±0.009
		h	***b, o***	0.500±0.022	0.684±0.038	0.789±0.022	0.242±0.000	0.414±0.008	0.506±0.000	0.747±0.062	3.224±0.023	6.812±0.124	0.252±0.040	0.114±0.012	0.188±0.002	0.617±0.004	0.582±0.001	0.563±0.000
InterAct (Xu et al., 2025)	0.0	–	–	0.775±0.009	0.928±0.004	0.975±0.000	0.496±0.005	0.732±0.003	0.855±0.011	0.190±0.011	1.860±0.044	7.587±0.069	**0.030**±0.012	0.055±0.005	0.178±0.002	0.790±0.001	0.724±0.011	0.729±0.005
		o, h	***b***	0.775±0.009	0.928±0.000	**0.975**±0.000	0.498±0.003	0.734±0.005	0.857±0.014	0.187±0.010	1.862±0.049	7.584±0.061	**0.030**±0.013	0.056±0.007	0.178±0.003	0.792±0.001	0.727±0.012	0.733±0.008
		b, h	***o***	**0.781**±0.004	0.927±0.006	**0.975**±0.004	0.498±0.003	0.730±0.005	0.855±0.011	0.188±0.010	1.863±0.036	7.547±0.022	0.031±0.006	**0.051**±0.008	0.178±0.003	0.791±0.000	0.753±0.006	**0.755**±0.001
		b, o	***h***	0.773±0.006	0.928±0.000	0.973±0.002	0.496±0.005	0.738±0.005	0.857±0.008	0.187±0.009	**1.859**±0.050	7.649±0.051	**0.030**±0.083	0.055±0.008	0.177±0.003	0.791±0.001	0.726±0.016	0.731±0.010
		b	***o, h***	0.773±0.006	0.928±0.000	0.973±0.002	0.496±0.005	**0.738**±0.005	0.857±0.008	**0.186**±0.011	**1.859**±0.050	7.636±0.047	**0.030**±0.014	0.055±0.007	0.178±0.003	0.792±0.000	0.730±0.015	0.733±0.008
		o	***b, h***	0.780±0.006	**0.930**±0.002	**0.975**±0.000	0.486±0.008	0.734±0.000	**0.861**±0.008	**0.186**±0.006	1.866±0.043	7.606±0.197	**0.030**±0.013	0.059±0.005	0.184±0.002	**0.800**±0.012	**0.755**±0.016	**0.755**±0.011
		h	***b, o***	0.773±0.011	0.928±0.004	0.973±0.002	**0.502**±0.008	0.736±0.003	0.857±0.014	0.187±0.010	1.861±0.052	**7.697**±0.053	**0.030**±0.002	0.056±0.007	0.180±0.002	0.791±0.001	0.735±0.012	0.737±0.006
	0.5	–	–	0.764±0.015	0.927±0.002	0.975±0.004	**0.475**±0.003	**0.729**±0.003	0.850±0.003	0.202±0.013	1.922±0.027	7.577±0.165	0.029±0.018	0.054±0.004	0.186±0.000	0.792±0.004	0.730±0.035	0.733±0.026
		o, h	***b***	0.758±0.015	0.925±0.004	0.975±0.004	0.471±0.003	0.723±0.000	0.850±0.003	0.191±0.009	1.931±0.019	7.523±0.092	**0.029**±0.019	0.054±0.002	0.186±0.002	0.792±0.004	0.736±0.021	0.742±0.017
		b, h	***o***	0.762±0.013	**0.927**±0.002	0.975±0.004	**0.475**±0.003	0.727±0.005	**0.852**±0.005	0.195±0.012	**1.916**±0.022	7.569±0.018	**0.029**±0.018	0.054±0.003	0.186±0.001	0.793±0.004	0.738±0.023	0.742±0.015
		b, o	***h***	0.758±0.015	0.925±0.004	0.975±0.004	0.471±0.003	0.723±0.000	0.848±0.005	0.191±0.009	1.931±0.019	7.523±0.092	**0.029**±0.019	0.054±0.002	0.186±0.002	0.792±0.004	0.736±0.021	0.742±0.017
		b	***o, h***	0.756±0.013	0.923±0.002	0.969±0.000	0.471±0.003	0.725±0.008	**0.852**±0.005	0.195±0.010	1.928±0.024	7.567±0.028	**0.029**±0.011	**0.053**±0.005	0.184±0.002	0.791±0.001	0.731±0.027	0.736±0.020
		o	***b, h***	**0.766**±0.013	**0.927**±0.002	**0.980**±0.002	0.469±0.000	**0.729**±0.003	**0.852**±0.000	**0.190**±0.012	1.919±0.025	**7.634**±0.055	**0.029**±0.012	0.056±0.003	0.192±0.001	**0.801**±0.010	**0.761**±0.020	**0.761**±0.014
		h	***b, o***	0.756±0.013	0.925±0.004	0.975±0.004	0.467±0.003	0.723±0.000	0.846±0.003	**0.190**±0.010	1.929±0.014	7.617±0.002	**0.029**±0.009	0.054±0.004	0.187±0.001	0.792±0.004	0.738±0.019	0.743±0.013
OMOMO (Li et al., 2023a)	0.0	–	–	0.681±0.000	**0.898**±0.002	0.970±0.002	**0.324**±0.001	0.551±0.001	0.711±0.003	0.090±0.009	1.489±0.000	7.269±0.080	0.014±0.009	0.061±0.004	0.214±0.005	0.842±0.000	0.755±0.005	0.786±0.004
		o, h	***b***	**0.681**±0.000	0.897±0.004	**0.972**±0.000	**0.324**±0.001	0.551±0.001	**0.713**±0.000	0.088±0.009	1.483±0.001	7.371±0.247	0.013±0.008	0.061±0.003	0.215±0.006	**0.843**±0.001	0.755±0.000	0.787±0.000
		b, h	***o***	**0.681**±0.000	0.897±0.004	0.969±0.004	**0.324**±0.001	0.551±0.001	0.711±0.001	**0.087**±0.011	1.488±0.002	**7.541**±0.202	0.013±0.008	0.061±0.003	0.216±0.002	**0.843**±0.002	0.763±0.017	0.794±0.011
		b, o	***h***	0.672±0.004	**0.898**±0.002	0.967±0.002	**0.324**±0.000	**0.553**±0.003	0.711±0.003	**0.087**±0.007	1.520±0.003	7.351±0.114	**0.011**±0.003	0.061±0.012	0.227±0.001	**0.843**±0.001	0.765±0.001	0.794±0.001
		b	***o, h***	**0.681**±0.000	**0.898**±0.006	0.970±0.002	**0.324**±0.001	**0.553**±0.001	**0.713**±0.003	**0.087**±0.010	1.490±0.005	7.308±0.043	0.014±0.004	0.061±0.007	0.219±0.002	**0.843**±0.000	0.763±0.010	0.793±0.008
		o	***b, h***	0.678±0.000	0.897±0.004	0.969±0.004	**0.324**±0.001	0.551±0.001	0.711±0.001	**0.087**±0.009	1.511±0.007	7.429±0.088	**0.011**±0.003	**0.058**±0.001	0.231±0.000	0.842±0.000	**0.773**±0.020	**0.798**±0.013
		h	***b, o***	**0.681**±0.000	0.897±0.004	**0.972**±0.000	**0.324**±0.001	**0.553**±0.003	**0.713**±0.003	0.088±0.011	**1.477**±0.006	7.513±0.034	0.013±0.008	0.061±0.003	0.217±0.007	**0.843**±0.001	0.763±0.006	0.791±0.004
	0.5	–	–	0.675±0.000	**0.898**±0.002	0.966±0.004	0.322±0.003	**0.551**±0.005	0.711±0.011	0.090±0.007	1.521±0.002	7.442±0.067	0.011±0.003	**0.058**±0.005	0.226±0.002	0.843±0.000	0.762±0.004	0.792±0.003
		o, h	***b***	0.677±0.002	**0.898**±0.002	**0.969**±0.004	**0.324**±0.000	0.549±0.008	**0.713**±0.011	**0.087**±0.006	**1.518**±0.004	7.384±0.109	0.011±0.001	**0.058**±0.005	0.224±0.002	**0.844**±0.001	0.761±0.000	0.793±0.001
		b, h	***o***	0.675±0.004	**0.898**±0.002	**0.969**±0.004	**0.324**±0.003	**0.551**±0.005	**0.713**±0.011	0.088±0.006	1.521±0.002	7.396±0.089	0.011±0.002	**0.058**±0.005	0.227±0.003	0.843±0.001	0.763±0.007	0.793±0.005
		b, o	***h***	0.677±0.006	**0.898**±0.006	0.967±0.002	0.323±0.003	0.549±0.008	0.709±0.008	0.088±0.007	1.528±0.005	7.365±0.125	0.011±0.002	0.059±0.005	0.227±0.002	**0.844**±0.001	0.764±0.014	0.793±0.010
		b	***o, h***	0.677±0.006	**0.898**±0.002	0.967±0.002	**0.324**±0.000	**0.551**±0.005	0.711±0.011	0.089±0.007	1.523±0.003	7.420±0.022	0.011±0.002	0.059±0.006	0.229±0.003	**0.844**±0.000	0.765±0.008	0.794±0.005
		o	***b, h***	**0.681**±0.000	0.897±0.004	0.966±0.004	**0.324**±0.001	**0.551**±0.008	0.711±0.011	0.088±0.008	1.526±0.002	**7.543**±0.113	**0.010**±0.002	0.059±0.006	0.227±0.005	0.840±0.000	**0.767**±0.017	**0.795**±0.011
		h	***b, o***	0.675±0.004	**0.898**±0.002	0.967±0.002	**0.324**±0.000	**0.551**±0.005	0.711±0.011	0.089±0.006	1.524±0.005	7.420±0.145	0.012±0.003	**0.058**±0.007	0.226±0.000	**0.844**±0.001	0.764±0.011	0.793±0.006

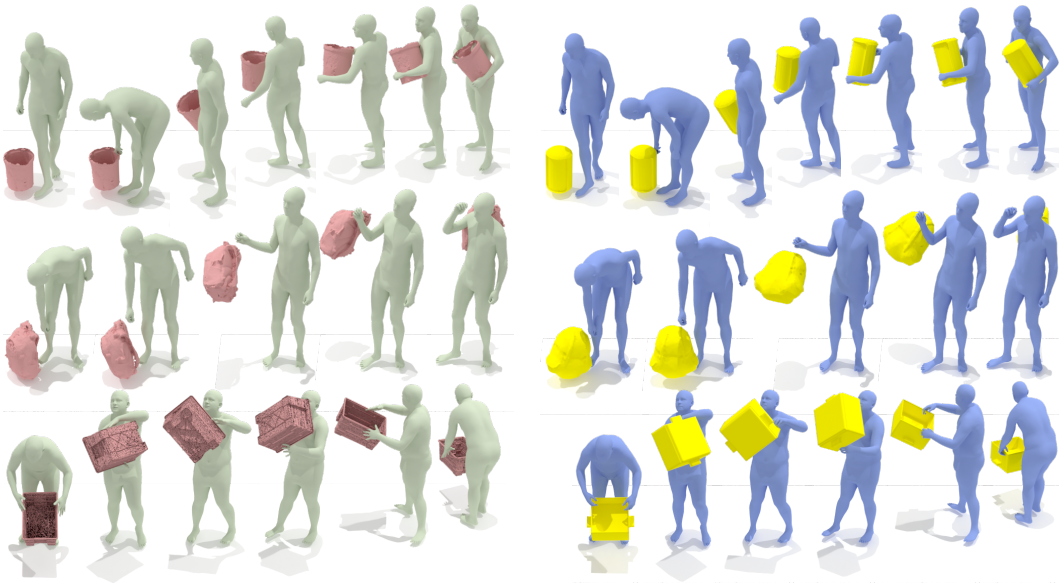

Figure A: We visualize our augmentation results, where original MoCap sequences (**left**) are enriched by introducing novel objects (**right**) sourced from ShapeNet (Chang et al., 2015) and Objaverse (Deitke et al., 2023). Through optimization, human and object motions are aligned at corresponding contact points.

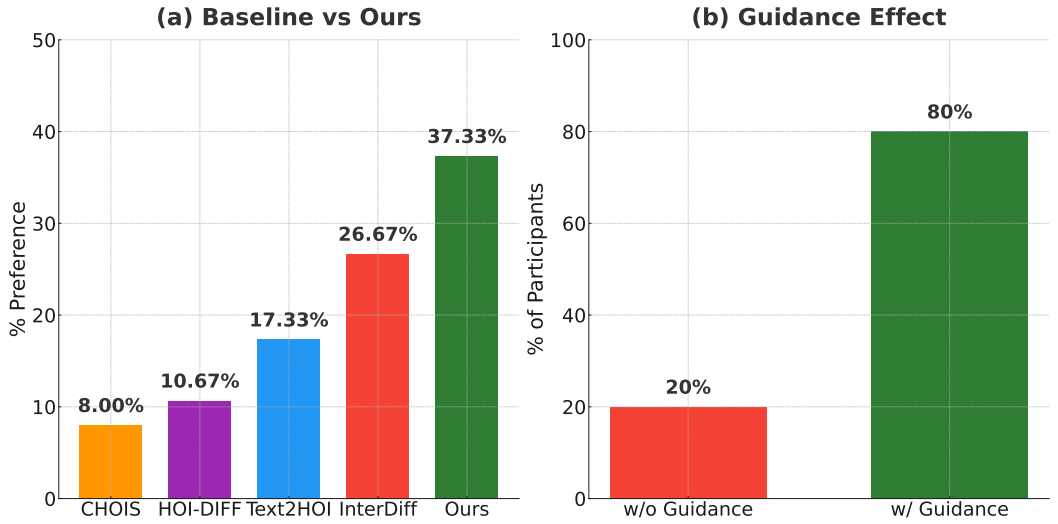

Figure B: **User study results.** 15 participants evaluates our method comparing baselines, and our guidance mechanism. Percentages indicate participant preference.

Table E: **Quantitative comparisons** on the OMOMO dataset (Li et al., 2023a) between our method and baseline approaches. We report R-Precision with batch sizes 64 and 256.

Method	R-Precision↑						FID↓	MM Dist↓	Diversity→	FSR↓	Pene↓	Contact→	Interaction↑		
	Batch Size = 64			Batch Size = 256											
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
Ground Truth	$0.683^{\pm0.002}$	$0.909^{\pm0.000}$	$0.981^{\pm0.000}$	$0.320^{\pm0.000}$	$0.562^{\pm0.000}$	$0.723^{\pm0.000}$	$-0.000^{\pm0.000}$	$1.447^{\pm0.000}$	$7.356^{\pm0.184}$	$0.012^{\pm0.008}$	$0.066^{\pm0.000}$	$0.401^{\pm0.000}$	$0.872^{\pm0.000}$	$0.875^{\pm0.000}$	$0.874^{\pm0.000}$
HOI-Diff Peng et al. (2023)	$0.480^{\pm0.015}$	$0.719^{\pm0.017}$	$0.838^{\pm0.022}$	$0.233^{\pm0.015}$	$0.402^{\pm0.019}$	$0.519^{\pm0.017}$	$1.448^{\pm0.066}$	$2.683^{\pm0.079}$	$7.119^{\pm0.064}$	$0.020^{\pm0.006}$	$0.106^{\pm0.001}$	$0.149^{\pm0.005}$	$0.758^{\pm0.003}$	$0.586^{\pm0.002}$	$0.631^{\pm0.004}$
CHOIS Li et al. (2023a)	$0.662^{\pm0.000}$	$0.887^{\pm0.013}$	$0.967^{\pm0.006}$	$0.301^{\pm0.001}$	$0.525^{\pm0.003}$	$0.688^{\pm0.002}$	$0.148^{\pm0.000}$	$1.644^{\pm0.022}$	$\mathbf{7.358}^{\pm0.201}$	$0.018^{\pm0.033}$	$0.055^{\pm0.002}$	$0.257^{\pm0.000}$	$0.840^{\pm0.003}$	$\mathbf{0.789}^{\pm0.004}$	$\mathbf{0.809}^{\pm0.005}$
InterDiff Xu et al. (2023b)	$\mathbf{0.689}^{\pm0.002}$	$\mathbf{0.903}^{\pm0.004}$	$\mathbf{0.978}^{\pm0.000}$	$0.316^{\pm0.005}$	$0.547^{\pm0.002}$	$\mathbf{0.711}^{\pm0.005}$	$0.110^{\pm0.004}$	$1.514^{\pm0.003}$	$7.390^{\pm0.111}$	$0.017^{\pm0.005}$	$\mathbf{0.054}^{\pm0.005}$	$0.244^{\pm0.010}$	$\mathbf{0.843}^{\pm0.005}$	$0.775^{\pm0.004}$	$0.799^{\pm0.002}$
Text2HOI Cha et al. (2024)	$0.653^{\pm0.004}$	$0.867^{\pm0.006}$	$0.955^{\pm0.002}$	$0.301^{\pm0.002}$	$0.523^{\pm0.005}$	$0.676^{\pm0.005}$	$0.155^{\pm0.004}$	$1.710^{\pm0.007}$	$7.339^{\pm0.074}$	$0.012^{\pm0.028}$	$0.093^{\pm0.001}$	$0.189^{\pm0.000}$	$0.837^{\pm0.002}$	$0.695^{\pm0.001}$	$0.742^{\pm0.001}$
LIGHT (**Ours**) w/o guidance	$0.675^{\pm0.000}$	$0.898^{\pm0.002}$	$0.966^{\pm0.004}$	$0.322^{\pm0.003}$	$\mathbf{0.551}^{\pm0.005}$	$\mathbf{0.711}^{\pm0.011}$	$0.090^{\pm0.007}$	$1.521^{\pm0.002}$	$7.340^{\pm0.009}$	$\mathbf{0.011}^{\pm0.003}$	$0.058^{\pm0.001}$	$0.226^{\pm0.002}$	$\mathbf{0.843}^{\pm0.000}$	$0.762^{\pm0.004}$	$0.792^{\pm0.003}$
LIGHT (**Ours**) w/ guidance	$0.678^{\pm0.000}$	$0.897^{\pm0.004}$	$0.969^{\pm0.004}$	$\mathbf{0.324}^{\pm0.001}$	$\mathbf{0.551}^{\pm0.001}$	$\mathbf{0.711}^{\pm0.001}$	$\mathbf{0.087}^{\pm0.009}$	$\mathbf{1.511}^{\pm0.007}$	$7.429^{\pm0.088}$	$\mathbf{0.011}^{\pm0.003}$	$0.058^{\pm0.001}$	$0.231^{\pm0.000}$	$0.842^{\pm0.000}$	$0.773^{\pm0.020}$	$0.798^{\pm0.013}$

Quantitative Evaluation of the Quality of Augmented Data. Table A demonstrates that our augmented data can match the quality of the ground-truth set, showing minimal contact errors and demonstrating its suitability as synthetic training data. And our augmentation of small objects on GRAB (Taheri et al., 2020) could also greatly align the contact of the ground truth dataset

Quantitative Evaluation on additional datasets. The Table E, F and ablations in Table G, H, demonstrates that our guidance method is effective across multiple datasets, rather than only on a large-scale dataset. And also, token separation and independent noise schedule are also effective.

Disscussion of Runtime efficiency. Our guidance requires three model forward passes per step, and a complete inference procedure is needed before applying guidance. Specifically, for a batch of 64 samples, each sample consisting of 300 frames, our guided inference completes in approximately 72 seconds on one A100 gpu. In contrast, HOI-Diff, InterDiff both takes about 15 seconds without guidance.

Limitations. Despite the encouraging results, our approach currently faces several limitations. First, our framework is designed specifically for single-object scenarios, and extending it to multi-object interactions remains an open challenge. Furthermore, the current model does not explicitly incorporate static environmental contexts or detailed scene geometry, potentially limiting the contextual realism of generated motions. Additionally, while our method improves interaction coherence and physical

Table F: **Quantitative comparisons** on the BEHAVE dataset Bhatnagar et al. (2022) between our method and baseline approaches. We report R-Precision with batch sizes 64 and 256.

Method	R-Precision↑						FID↓	MM Dist↓	Diversity→	FSR↓	Pene↓	Contact→	Interaction↑		
	Batch Size = 64			Batch Size = 256											
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
Ground Truth	$0.914^{\pm0.000}$	$0.941^{\pm0.005}$	$0.953^{\pm0.000}$	$0.680^{\pm0.000}$	$0.872^{\pm0.003}$	$0.914^{\pm0.001}$	$-0.000^{\pm0.000}$	$1.523^{\pm0.006}$	$7.046^{\pm0.088}$	$0.252^{\pm0.025}$	$0.107^{\pm0.000}$	$0.392^{\pm0.000}$	$0.827^{\pm0.000}$	$0.875^{\pm0.000}$	$0.848^{\pm0.000}$
HOI-Diff Peng et al. (2023)	$0.422^{\pm0.011}$	$0.629^{\pm0.016}$	$0.727^{\pm0.043}$	$0.205^{\pm0.024}$	$0.377^{\pm0.024}$	$0.471^{\pm0.008}$	$0.879^{\pm0.094}$	$3.526^{\pm0.006}$	$6.636^{\pm0.209}$	$0.277^{\pm0.054}$	$0.097^{\pm0.007}$	$0.221^{\pm0.001}$	$\mathbf{0.619}^{\pm0.012}$	$\mathbf{0.620}^{\pm0.011}$	$\mathbf{0.575}^{\pm0.010}$
CHOIS Li et al. (2023a)	$0.453^{\pm0.000}$	$0.672^{\pm0.011}$	$0.750^{\pm0.011}$	$0.221^{\pm0.008}$	$0.395^{\pm0.005}$	$0.490^{\pm0.014}$	$0.850^{\pm0.058}$	$3.480^{\pm0.010}$	$6.829^{\pm0.084}$	$0.260^{\pm0.060}$	$0.096^{\pm0.035}$	$0.205^{\pm0.008}$	$0.594^{\pm0.001}$	$0.582^{\pm0.015}$	$0.552^{\pm0.007}$
InterDiff Xu et al. (2023b)	$0.500^{\pm0.043}$	$\mathbf{0.715}^{\pm0.005}$	$0.809^{\pm0.016}$	$0.275^{\pm0.030}$	$0.467^{\pm0.046}$	$0.559^{\pm0.027}$	$0.770^{\pm0.020}$	$3.259^{\pm0.009}$	$6.844^{\pm0.154}$	$0.290^{\pm0.070}$	$\mathbf{0.084}^{\pm0.006}$	$0.210^{\pm0.001}$	$0.646^{\pm0.022}$	$0.631^{\pm0.018}$	$0.596^{\pm0.006}$
Text2HOI Cha et al. (2024)	$0.336^{\pm0.011}$	$0.496^{\pm0.005}$	$0.602^{\pm0.011}$	$0.135^{\pm0.035}$	$0.246^{\pm0.022}$	$0.334^{\pm0.008}$	$1.317^{\pm0.081}$	$3.761^{\pm0.041}$	$6.600^{\pm0.119}$	$0.255^{\pm0.053}$	$0.140^{\pm0.004}$	$0.163^{\pm0.004}$	$0.607^{\pm0.001}$	$0.549^{\pm0.015}$	$0.514^{\pm0.003}$
LIGHT (**Ours**) w/o guidance	$0.488^{\pm0.027}$	$0.691^{\pm0.016}$	$0.793^{\pm0.027}$	$0.234^{\pm0.011}$	$0.426^{\pm0.011}$	$0.516^{\pm0.022}$	$0.764^{\pm0.053}$	$3.233^{\pm0.023}$	$\mathbf{6.907}^{\pm0.116}$	$\mathbf{0.252}^{\pm0.043}$	$0.114^{\pm0.018}$	$0.187^{\pm0.003}$	$0.617^{\pm0.003}$	$0.569^{\pm0.007}$	$0.557^{\pm0.003}$
LIGHT (**Ours**) w/ guidance	$\mathbf{0.508}^{\pm0.022}$	$0.703^{\pm0.022}$	$\mathbf{0.816}^{\pm0.016}$	$\mathbf{0.252}^{\pm0.008}$	$\mathbf{0.430}^{\pm0.005}$	$\mathbf{0.526}^{\pm0.003}$	$\mathbf{0.696}^{\pm0.038}$	$\mathbf{3.208}^{\pm0.050}$	$6.657^{\pm0.240}$	$0.260^{\pm0.049}$	$0.111^{\pm0.016}$	$0.189^{\pm0.001}$	$0.617^{\pm0.004}$	$0.584^{\pm0.013}$	$0.564^{\pm0.004}$

Table G: **Ablation study** of token-separation strategies on the BEHAVE dataset Xu et al. (2025).

hand-body separation	human-object separation	R-Precision↑ Batch Size = 64 Top 1	Top 2	Top 3	Batch Size = 256 Top 1	Top 2	Top 3	FID↓	MM Dist↓	Diversity→	FSR↓	Pene↓	Contact→	Interaction↑ C_{prec}	C_{rec}	C_{F1}
✓	✓	$0.508^{\pm0.022}$	$0.703^{\pm0.022}$	**0.816**$^{\pm0.016}$	**0.252**$^{\pm0.008}$	$0.430^{\pm0.005}$	$0.526^{\pm0.003}$	**0.696**$^{\pm0.038}$	$3.208^{\pm0.050}$	**6.657**$^{\pm0.240}$	**0.260**$^{\pm0.049}$	$0.111^{\pm0.016}$	$0.189^{\pm0.001}$	$0.617^{\pm0.004}$	$0.584^{\pm0.013}$	$0.564^{\pm0.004}$
–	✓	$0.539^{\pm0.000}$	$0.703^{\pm0.065}$	$0.801^{\pm0.049}$	$0.223^{\pm0.022}$	$0.430^{\pm0.016}$	$0.514^{\pm0.019}$	$0.923^{\pm0.043}$	$3.186^{\pm0.026}$	$6.845^{\pm0.420}$	$0.274^{\pm0.049}$	**0.100**$^{\pm0.000}$	$0.173^{\pm0.002}$	$0.626^{\pm0.003}$	$0.569^{\pm0.003}$	$0.564^{\pm0.006}$
✓	–	**0.559**$^{\pm0.005}$	**0.746**$^{\pm0.027}$	$0.812^{\pm0.000}$	$0.248^{\pm0.014}$	**0.453**$^{\pm0.016}$	**0.535**$^{\pm0.032}$	$0.808^{\pm0.003}$	**3.135**$^{\pm0.035}$	$6.756^{\pm0.073}$	$0.289^{\pm0.075}$	**0.100**$^{\pm0.000}$	$0.204^{\pm0.002}$	**0.635**$^{\pm0.019}$	**0.627**$^{\pm0.011}$	**0.588**$^{\pm0.008}$

Table H: **Ablation study** of token-separation strategies on the OMOMO dataset (Li et al., 2023a). We report R-Precision with batch sizes 64 and 256.

hand–body separation	human–object separation	R-Precision↑ Batch Size = 64 Top 1	Top 2	Top 3	Batch Size = 256 Top 1	Top 2	Top 3	FID↓	MM Dist↓	Diversity→	FSR↓	Pene↓	Contact→	Interaction↑ C_{prec}	C_{rec}	C_{F1}
✓	✓	$0.678^{\pm0.000}$	$0.897^{\pm0.004}$	$0.969^{\pm0.004}$	**0.324**$^{\pm0.001}$	$0.551^{\pm0.001}$	$0.711^{\pm0.001}$	**0.087**$^{\pm0.009}$	$1.511^{\pm0.007}$	$7.429^{\pm0.088}$	**0.011**$^{\pm0.003}$	**0.021**$^{\pm0.001}$	$0.231^{\pm0.000}$	**0.842**$^{\pm0.000}$	$0.773^{\pm0.020}$	$0.798^{\pm0.013}$
–	✓	$0.678^{\pm0.004}$	$0.902^{\pm0.002}$	**0.973**$^{\pm0.002}$	$0.320^{\pm0.000}$	$0.551^{\pm0.005}$	$0.711^{\pm0.005}$	$0.093^{\pm0.008}$	**1.483**$^{\pm0.006}$	**7.366**$^{\pm0.071}$	$0.017^{\pm0.004}$	$0.062^{\pm0.006}$	$0.223^{\pm0.005}$	$0.840^{\pm0.003}$	$0.775^{\pm0.005}$	$0.798^{\pm0.005}$
✓	–	**0.688**$^{\pm0.000}$	**0.906**$^{\pm0.004}$	**0.973**$^{\pm0.002}$	$0.322^{\pm0.003}$	**0.553**$^{\pm0.003}$	**0.715**$^{\pm0.000}$	$0.106^{\pm0.003}$	$1.499^{\pm0.013}$	$7.378^{\pm0.193}$	$0.031^{\pm0.004}$	$0.079^{\pm0.002}$	$0.236^{\pm0.003}$	$0.835^{\pm0.003}$	**0.790**$^{\pm0.000}$	**0.804**$^{\pm0.003}$

plausibility, it may still struggle with highly dynamic or physically complex interactions requiring precise and rapid manipulations. Finally, due to the computational complexity introduced by staged conditioning and diffusion forcing, real-time or interactive applications might be challenging.