

# HOI-Diff: Text-Driven Synthesis of 3D Human-Object Interactions using Diffusion Models

## Supplementary Material

### A. Additional Details of Methodology

In Sec. 3 of our main paper, we presented the foundational design of each key component in our HOI-Diff pipeline. Here, we delve into an elaborate explanation of model architecture, learning objectives and additional details associated with each crucial component.

#### A.1. Dual-branch diffusion model (DBDM)

The Communication Module (CM) in DBDM is based on the cross attention mechanism. Formally,

$$\tilde{f}^h = \text{MLP}(\text{Attn}(f^h \mathbf{W}_Q, f^o \mathbf{W}_K, f^o \mathbf{W}_V)), \quad (8)$$

$$\tilde{f}^o = \text{MLP}(\text{Attn}(f^o \mathbf{W}_Q, f^h \mathbf{W}_K, f^h \mathbf{W}_V)), \quad (9)$$

where  $\text{MLP}(\cdot)$  denotes fully-connected layers,  $\text{Attn}(\cdot)$  is the attention block [54], and  $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$  are learned projection matrices for query, key, and value, respectively.

The training objective of this full model is based on reconstruction loss

$$\mathcal{L}_{hoi} = \mathbb{E}_{t \sim [1, T]} \|M_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{x}_0\|_2^2, \quad (10)$$

where  $\mathbf{x}_0$  is the ground truth of the HOI sequence.

#### A.2. Affordance prediction diffusion model (APDM).

**Model architecture.** The affordance prediction diffusion model comprises eight Transformer layers for the encoder with a PointNet++ [38] to encode the object’s point clouds. The training objective of this diffusion model is also based on reconstruction loss

$$\mathcal{L}_{aff} = \mathbb{E}_{t \sim [1, T]} \|A_\theta(\mathbf{y}_t, t, \mathbf{p}, \mathbf{d}) - \mathbf{y}_0\|_2^2, \quad (11)$$

where  $\mathbf{y}_0$  is the ground-truth affordance data.  $\mathbf{p}$  and  $\mathbf{d}$  denote object point cloud and text description (prompt), respectively.  $A_\theta$  represents the affordance prediction diffusion model.

**Inferring object state with GPT-3.5-turbo in APDM.** To infer the state of an object, we directly leverage the strong prior knowledge of large language models to derive the result. Specifically, we utilize the GPT-3.5-turbo [34] API by inputting specific instructions, allowing it to infer the result directly based on the input HOI text description. The prompt template for instruction is shown in Figure 6.

#### A.3. Affordance-guided interaction correction.

During the inference stage, it’s found that the predicted object contact positions may occasionally be inaccurately positioned, residing either inside or outside the object. To rectify this, we implement post-processing steps that replace these predicted contact points, denoted as  $\mathbf{y}_0^o$ , with their nearest neighbors from the object’s point clouds. This adjustment aims to enhance the accuracy of the updated contact points, aligning them more closely with their actual positions on the object’s surface. However, employing these updated contact points directly for contact constraints, particularly in the absence of detailed human shape information, introduces a new challenge. It can potentially lead to penetration issues within the contact area while reconstructing the human mesh in the final stage. To mitigate contact penetration, we adopt a method that recalculates points at a specified distance outward, perpendicular to the normal, originating from the object’s contact points. This process can be formulated as:  $\tilde{\mathbf{y}}_0^o = \hat{\mathbf{y}}_0^o + \mathbf{v}_n^i * d$ , where  $i \in \{1, 2\}$  indicates the  $i^{th}$  object contact points,  $\mathbf{v}_n^i$  denotes the normal vector at that point and  $d = 0.05$  is a contact distance threshold.

As for smoothness term, we formulate it as

$$G_{smo} = \sum_{l=1}^{L-1} \|\mathbf{x}_0^o(l+1) - \mathbf{x}_0^o(l)\|^2, \quad (12)$$

where  $\mathbf{x}_0^o(l)$  is the predicted 6DoF pose of the object in the  $l$ -th frame.

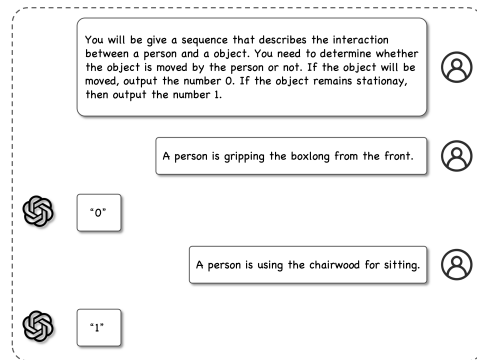


Figure 6. Prompt template for inferring object state.

**Algorithm 1** Affordance-guided Interaction Correction

---

**Require:** Input  $c = (d, p)$  consisting of a textual description  $d$  and object point cloud  $p$ , HOI-Diff model  $M_\theta$ , objective function  $G(\mu_t^h, \mu_t^o, y_0)$ , and estimated affordance  $y_0 = (y_0^h, y_0^o, y_0^o)$ .

- 1:  $x_T^h, x_T^o \leftarrow \text{sample from } \mathcal{N}(0, I)$
- 2:  $K = 1$
- 3: **for all**  $t$  from  $T$  to 1 **do**
- 4:  $x_t^h, x_t^o \leftarrow M_\theta(x_t^h, x_t^o, t, c)$  # Get  $\mu_t^h, \mu_t^o$  according to Eq.(2) with  $\Sigma_t$
- 5: **if**  $t = 1$  **then**
- 6:  $K = 100$
- 7: **end if**
- 8: **for all**  $k$  from  $K$  to 1 **do** # Separately perturb
- 9:  $\mu_t^h \leftarrow \mu_t^h - \tau_1 \Sigma_t \nabla_{\mu_t^h} G(\mu_t^h, \mu_t^o, y_0)$ ,  $\mu_t^o \leftarrow \mu_t^o - \tau_2 \Sigma_t \nabla_{\mu_t^o} G(\mu_t^h, \mu_t^o, y_0)$
- 10: **end for**
- 11:  $x_{t-1}^h \sim \mathcal{N}(\mu_t^h, \Sigma_t)$ ,  $x_{t-1}^o \sim \mathcal{N}(\mu_t^o, \Sigma_t)$
- 12: **end for**
- 13: **return**  $x_0^h, x_0^o$

---

**B. Implementation Details**

Both our DBDM and APDM are built on the Transformer [54] architecture. Similar to MDM [51], we employ the CLIP model to encode text prompts, adhering to a classifier-free generation process. Our models are trained using PyTorch [35] on 1 NVIDIA A5000 GPU. We set control strength of guidance as  $\tau_1 = 1$ ,  $\tau_2 = 100$ , and  $\Sigma_t = \min(\Sigma_t, 0.01)$ . Both the DBDM and APDM are trained on the same data for 20k steps.

Both the DBDM and APDM architectures of HOI-Diff are based on Transformers with 4 attention heads, a latent dimension of 512, a dropout of 0.1, a feed-forward size of 1024, and the GeLU activation [15]. The number of learned parameters for each model is stated in Table 4.

Our training setting involves 20k iterations for the DBDM and 10k iterations for the APDM model. These iterations utilize a batch size of 32 and employ the AdamW optimizer [30] with a learning rate set at  $10^{-4}$ . We use  $T=1000$  and  $N=500$  diffusion steps in DBDM and APDM, respectively.

**C. Additional Details of Baselines**

- $\text{MDM}^{\text{finetuned}}$ : We finetune MDM [51] on BEHAVE dataset without considering the object motion.
- $\text{MDM}^*$ : We extend the original feature dimensions of the input and output processing in MDM [51] from  $D^h$  to  $D^h + D^o$ , enabling support for HOIs sequences. The model is trained from scratch on BEHAVE dataset [4].
- $\text{PriorMDM}^*$ : The proposed approach for dual-person motion generation employs paired fixed MDMs [51] per individual to ensure uniformity within generated human motion distributions. This design leverages a singular ComMDM to coordinate between the two branches of fixed MDM instances, streamlining training and maintaining consistency across generated motions. Given that both branches are based on MDM that pretrained on human motion datasets, direct utilization of them for human-object in-

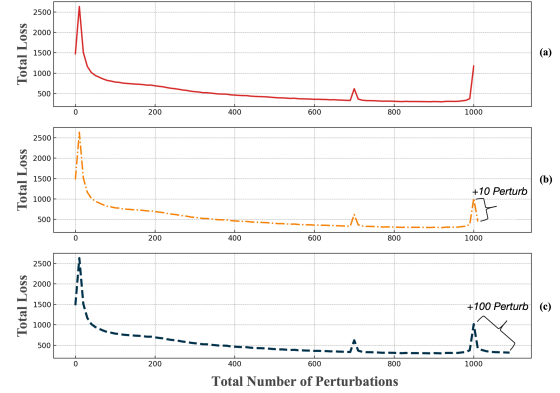


Figure 7. Effect of different total numbers of perturbations in the whole denoising process. (a) Perturb one time in each denoising step (in total  $T = 1000$ ). (b) Perturb one time in first  $T - 1$  denoising steps, and repeatedly perturb 10 times in the final denoising step. (c) Perturb one time in first  $T - 1$  denoising steps, and repeatedly perturb 100 times in the final denoising step.

teractions in our task is infeasible. We maintain one branch dedicated to humans, leveraging pre-trained weights, while adapting the input and output processing of another branch specifically for generating object motion. Following this, we fine-tune the human MDM branch while initiating the learning of object motion from scratch within the object branch. Eventually, we integrate ComMDM to facilitate communication and coordination between these distinct branches handling human and object interactions.

- **InterDiff**: InterDiff [61] is originally designed for a prediction task rather than text-driven HOIs generation. To tailor it to our task, we replace its Transformer encoder with a CLIP encoder and modify its feature dimensions of the input and output layers.

To ensure fair comparisons, all the above baselines as well as our own models are all trained on BEHAVE and OMOMO datasets for 20k steps.

**D. Additional Details of Evaluation Metrics**

For detailed information regarding metrics employed in human motion generation, including *FID*, *R-Precision*, and *Diversity*, we refer readers to [13, 51] for comprehensive understanding.

**Contact Distance.** Expanding on the concept of *Contact Distance*, we utilize the *chamfer distance* metric to quantify the closeness between human body joints and the object surface. This computation leverages ground-truth affordance data that includes human contact labels and object contact

points,

$$ContactDistance = \frac{1}{L} \sum_l^L CD(\hat{x}_l^h, \hat{p}_l), \quad (13)$$

where  $\hat{x}_l^h$  represents two human contact joints at the  $l$ -th frame, indexed according to ground-truth contact labels. Additionally,  $\hat{p}_l$  denotes two object contact points derived from the object motion  $x_l^o$  at frame  $l$ , also indexed based on ground-truth information.  $CD$  denotes the *chamfer distance*.

**Penetration Score.** We followed the Li et al. [26] to compute the penetration score (Pene), each vertex of the body ( $V_i$ ) is queried against the precomputed Signed Distance Field (SDF) of the object. This process yields a corresponding distance value for each vertex. The penetration score is then formalized as:

$$Pene = \frac{1}{n} \sum_{i=1}^n |\min(d_i, 0)|, \quad (14)$$

measured in centimeters (cm).

## E. Inference Time

In Table 5, we provide the inference times for both baselines and our full method, including its key components. All measurements were conducted using an NVIDIA A5000 GPU. Training an additional model for affordance information and using classifier guidance for interaction correction do contribute to increased inference costs. However, despite the longer inference time, our complete method notably enhances the accuracy of 3D HOIs generation.

	Params (M)	FID ↓	R-precision (Top-3) ↑
MDM*	49.85	6.98	0.36
Ours (Full)	47.74	1.62	0.46

Table 6. With comparable model size, the performance results of MDM\* and Ours (Full).

## F. Additional Ablation Studies

**Different perturbing times in classifier guidance.** As discussed in Sec. 3.4, in the later stage of classifier guidance, diffusion models tend to strongly attenuate the introduced signals. Therefore, we iteratively perturb the predicted mean of motion for  $K$  times at the final denoising step. In Figure 7, we present the ablation results, illustrating the impact of different numbers of perturbations. Notably, we observe that employing 100 perturbations leads to re-convergence and yields the desired results.

**Different guidance strength.** As detailed in Sec. 3.4, we employ distinct control strengths for classifier guidance, considering the varying feature densities in predicted human

and object motion. Rather than employing equal control strengths, we opt to assign a higher control strength to object motion, allowing it to closely align with human contact joints, as illustrated in Figure 8.

**Different model with comparable model size.** Although our method involves a slightly larger number of model parameters, our model is specifically designed for HOI generation. As seen in the Table 6, if we attempt to scale MDM\* to the same model size, its performance remains subpar.

## G. User Study

For each method, we select 15 prompts from the BEHAVE dataset and 10 prompts from the OMOMO dataset, covering various interaction types and object items. We sample twice with each prompt to gather a total of 50 results. 40 participants are asked to choose their most preferred generation results from these samples. This user study requires pairwise comparisons of our method with other baseline on generated interaction quality, as shown in Figure 15. The results in Figure 9 indicate strong preference for our method: it is favored over the baselines in 89.6% (Ours vs. MDM\*), 73.8% (Ours vs. PriorMDM\*) and 95.3% (Ours vs. Interdiff).

## H. Additional Qualitative Results

In this section, we present additional qualitative results showcasing the model’s performance evaluated on the OMOMO dataset, and the effectiveness of APDM.

**Qualitative results on OMOMO dataset.** We present additional qualitative results on the OMOMO dataset, rendered with SMPL [29] shapes, as shown in Figure 11. It is evident that our method can generalize effectively to unseen objects and produce realistic 3D human-object interactions.

**Qualitative results of APDM.** To verify the accuracy of estimated contact points on object surface, we provide additional visual results in Figure 13. It can be seen that our method can predict realistic and practical contact points based on text descriptions. With APDM, we even can generate different interactions with the same object based on the input description, as shown in the Figure 14.

**Generalization capability.** To verify the model’s generalization capability, except of unseen object test on OMOMO dataset, we also downloaded several objects from Sketchfab<sup>3</sup>, adjusted them to a reasonable scale, and used them as inputs. As shown in Figure 12, our model successfully establishes reasonable HOI contact with these previously unseen objects.

<sup>3</sup><https://sketchfab.com/>

Model	DBDM	APDM
Parameters ( $\cdot 10^6$ )	8.82	38.92

Table 4. **Model Parameters.** The number of learned parameters of our two core architectures.

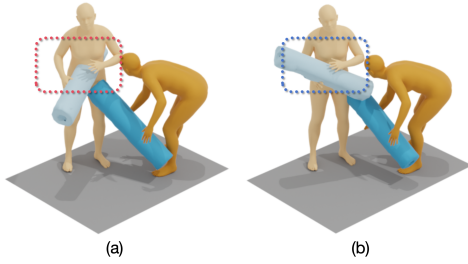


Figure 8. **Effect of different control strengths for classifier guidance.** (a) We use equal strengths of  $\tau_1 = 1, \tau_2 = 1$  to perturb the predicted mean of human motion and object motion, respectively. (b) We use different strengths of  $\tau_1 = 1, \tau_2 = 100$  for the perturbation. We can see that different strengths work better.

## I. Annotation for BEHAVE Dataset

**Text Annotating Process.** Initially, we manually annotate the interaction types and the specific human body parts involved, delineating actions like “lift” associated with the “left hand” or “hold” involving “two hands”. Subsequently, to generate complete sentences, we leverage the capabilities of GPT-3.5 to assist in formulating the entirety of the description.

**Examples of Annotated Textual Descriptions.** In Table 7, we showcase a selection of our annotated textual descriptions for the BEHAVE dataset [4].

**Analysis of Annotated Textual Descriptions.** All text descriptions encompass 36 distinct interaction verbs associated with 20 different objects. Figure 16 illustrates the frequency of each verb, indicating their respective occurrences.

**Affordance Data.** Our affordance data includes 8-dimensional human contact labels and object contact points. We employ *chamfer distance* to measure the distance between all human body joints and object surface points. Following a predefined distance threshold  $\gamma = 0.12$ , we identify the 8 contact points on the object surface corresponding to the 8 primary human body joints. Subsequently, we derive the human contact labels by encoding the indexes of contact joints into an 8-dimensional vector represented by binary values.

Method	MDM*	PriorMDM*	Ours (Full)
Time (s)	32.3	38.6	118.0
Component	APDM	DBDM	Interaction Correction
Time (s)	24.2	46.4	47.4

Table 5. **Inference Time (on NVIDIA A5000 GPU).** We report the inference time for baselines, our full method, and its key components.

Object	Textual Descriptions
<i>backpack</i>	A person is carrying the backpack in front.
	The person is raising a backpack with his right hand.
	The person at the front presently has control over the backpack.
<i>chairwood</i> (wooden chair)	A person is using the chairwood for sitting.
	The person is propelling the chairwood on the ground.
	Someone is hoisting a chairwood by his left hand.
<i>tablesquare</i> (square table)	A person is lifting the tablesquare, utilizing his left hand.
	Someone is clutching onto a tablesquare from the front.
	An individual is moving the tablesquare back and forth.
<i>boxlong</i> (long box)	A person is gripping the boxlong from the front.
	A person is raising the boxlong using his left hand.
	Someone hoists the boxlong with his left hand.
<i>toolbox</i>	Someone is grasping the toolbox upfront.
	The person has a firm hold on the toolbox with his right hand.
	A person is gripping the toolbox with his left hand.
<i>yogaball</i>	A person is shifting a yogaball back and forth on the floor using his hands.
	The person is occupying a yogaball.
	A person is employing an yogaball to engage in an upper body game.

Table 7. **Examples of our annotated textual descriptions for the BEHAVE dataset rephrased by GPT-3.5 [34].**

## J. Additional Details of OMOMO Dataset

The OMOMO dataset comprises data captured for a total of 15 objects. Adhering to their official split strategy depicted in [27](Figure 5), we allocate 10 objects for training and 5 objects for testing. This split allows us to further evaluate the model’s generalization ability to new objects. Notably, the OMOMO dataset itself provides text annotation, and we use GPT-3.5 to add subjects to it and embellish it appropriately. For affordance data, we preprocess it the same way we handle BEHAVE.



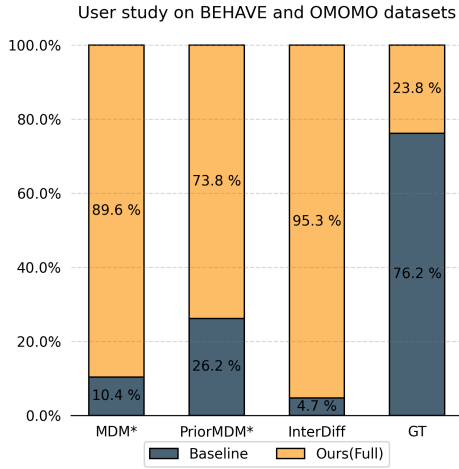


Figure 9. **Perceptual User Study.** Most participants prefer our method over the baselines.

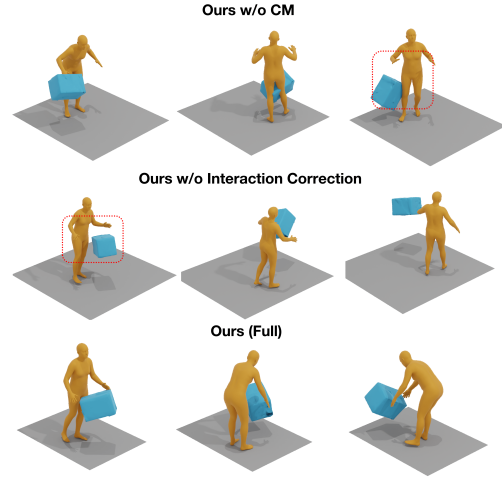


Figure 10. **Visual results of different variants of our model in ablation studies.**

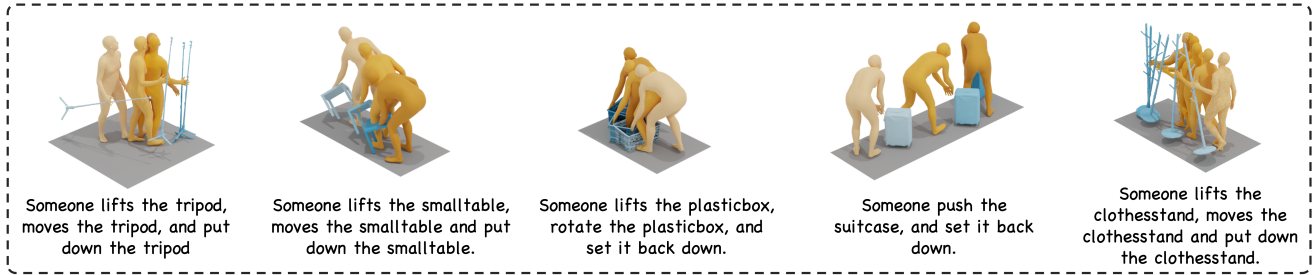


Figure 11. Additional qualitative evaluation on OMOMO dataset. Given object geometry and text description, our method can generate high-quality human-object interactions even for the unseen objects (tripod, smalltable, suitcase).

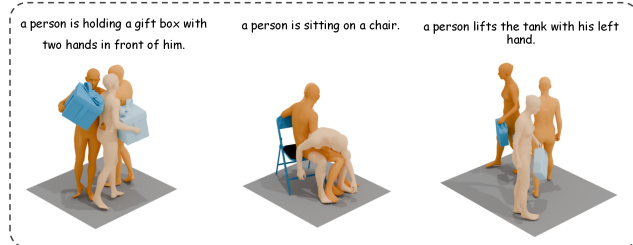


Figure 12. Additional qualitative evaluation on unseen objects.

## K. Common Questions

**Why use Skeletal Pose Representation rather than SMPL parameters?** Most state-of-the-art text-to-motion methods adopt the skeletal pose representation proposed by Guo et al. [13], demonstrating excellent performance and stability. While some works [2] argue that SMPL parameters [29] contains shape and global information, it does not generate as smooth motions as skeletal-based approaches. Consequently, we adopt the skeletal pose representation and aim to leverage strong pose priors from the pretrained text-to-motion

model [51] to ensure the authenticity of generated human motion.

**Can we handle multi-phase interactions between humans and objects?** Due to the lack of fine-grained textural descriptions in the current 3D HOI dataset, we primarily consider only one interaction phase. However, we have found that an LLM can still reason well for multiple phases given a template such as: *You will be given a sentence that describes an interaction between a person and an object across multiple phases. Your task is to divide the interaction into phases based on the state of the object and determine the state for each phase. If the object is being moved by the person during a phase, output the number 0. If the object remains stationary during a phase, output the number 1.*

For example, given the text description: *The box is on the ground. A person is picking up the box and holding it forward, then putting the box towards the table. The box is on the table* The result from GPT-3.5-turbo: *"Phase 1: The box is on the ground - State: 1 (stationary); Phase 2: The person is picking up the box and holding it forward - State: 0 (moved); Phase 3: The person is putting the box towards the table - State: 0 (moved); Phase 4: The box is on the*

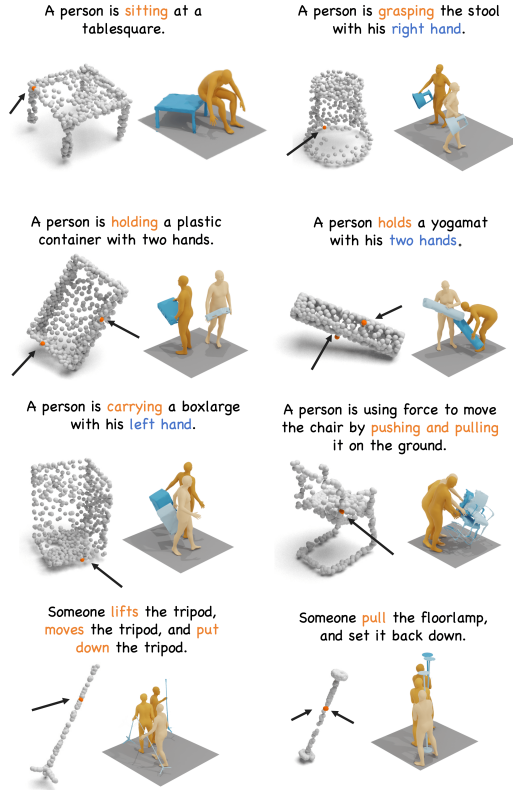


Figure 13. Visual results of estimated contact points. Our APDM, trained on the BEHAVE dataset, can accurately estimating contact positions for objects based on textual descriptions. Furthermore, it showcases the capability to generalize to unseen objects in the OMOMO dataset, as demonstrated in the last row.

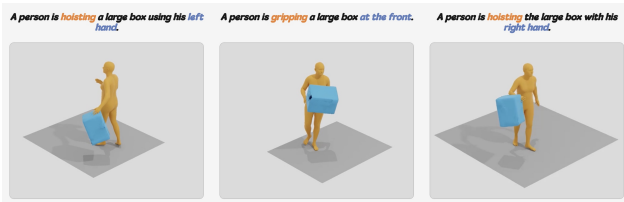


Figure 14. Leveraging the power of the APDM module, our method can generate diverse HOIs for the same object using different contacting body parts and contact points.

table - State: 1 (stationary). We will address the generation of multiple phases of 3D HOI in future work.

#### Can we generate hand motion with articulated fingers?

The BEHAVE and OMOMO datasets do not capture and provide raw hand parameters, despite utilizing SMPLH and SMPLX models to fit human body meshes for rendering. Consequently, in this paper, we focus solely on whole-body human motion, excluding articulated hand and finger movements.

#### Which one looks more realistic and is coherent with text description?

Prompt: The person is gripping with suitcase with his right hand.



Figure 15. An example question for our text-to-hoi user study.

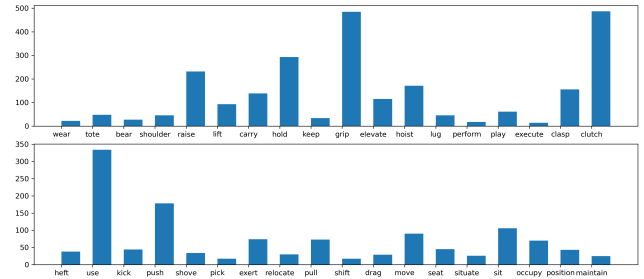


Figure 16. Analysis of word frequency We count the occurrences of each interaction verb from all text descriptions to illustrate their respective frequencies.

	Acc (%) ↑	Time (s) ↓
GPT-3.5	95.6	0.518
Gemini-1.5-Pro-Exp-0801	99.4	0.259
Gemma-2-27B	98.6	0.522
LLaMA-2-13B	99.4	0.259
APDM + MLP	79.5	2.420

Table 8. LLMs’ inference accuracy (Acc) and average inference time (Time) on object state prediction.

**Why do we use large language models (LLMs) to predict object state based on the input description?** We aim to leverage LLMs for inferring object states, and our results demonstrate that they perform efficiently and effectively. As shown in the Table 8, we evaluated the performance of object state prediction with GPT-3.5-turbo [34] and obtained an average precision of 95.6% on the validation set, with an average response time of 0.518 seconds. The results suggest that GPT-3.5-turbo is sufficiently accurate without adding significant overhead. We also evaluated the prediction performance using other LLMs, including Gemini-1.5-Pro-Exp-0801 [41] (99.4%, 0.259s), Gemma-2-27B [50] (98.6%, 0.522s), and LLaMA-2-13B [53] (94.4%, 0.521s), the latter two being publicly available.

To further validate the effectiveness of the LLM module, we modified the APDM module by adding an MLP head to predict the object status. The newly added MLP takes

in the features consisting of object geometry information and CLIP embeddings. We used an MSE loss. We got average precision 79.5% and average time 2.42s for this design on the validation set, which is significantly worse than the results of GPT-3.5-turbo (95.6%, 0.518s), Gemma-2-27b (98.6%, 0.522s), Gemini-1.5-Pro-Exp-0801 (99.4%, 0.259s) and LLaMA-2-13B (4.4%, 0.521s).

In future work, we believe the LLM can play a more important role in 3D HOI, e.g. providing high-level instruction for more complex human-object interactions, and our initial use of the LLM offers insights into its potential applications and how it can be effectively utilized.

## L. Supplementary Video

Beyond the qualitative results presented in the main paper, our supplementary materials offer comprehensive demos that provide an in-depth visualization of our task, further showcasing the effectiveness of our approach.

In these demonstrations, we highlight the better performance of our method, HOI-Diff, in producing diverse and realistic 3D HOIs while maintaining adherence to physical validity. Notably, the visualizations show that HOI-Diff consistently generates smooth, vivid interactions, accurately capturing human-object contacts.

Additionally, we present the visual ablation results and emphasize the significance and effectiveness of our affordance-guided interaction correction, underscoring its substantial impact on improving the overall performance and quality of the generated 3D HOIs.

## M. Limitations

The existing datasets for 3D HOIs are limited in terms of action and motion diversity, posing a challenge for synthesizing long-term interactions in our task. Furthermore, the effectiveness of our model’s interaction correction component is contingent on the precision of affordance estimation. Despite simplifying this task, achieving accurate affordance estimation remains a significant challenge, impacting the overall performance of our model. A promising direction for future research involves integrating a sophisticated affordance model pre-trained on an extensive 3D object dataset, along with text prompts. Such an advancement could significantly enhance the realism and accuracy of human-object contact in our model, leading to more natural and precise HOIs synthesis.

## N. Social Impacts

On the positive side, it may offers the research community valuable insights into understanding human behaviors. On the negative side, it remains uncertain whether individuals can be identified solely based on their poses and movements. However, compared to traditional input images of people, this method poses a lower risk of invading personal privacy.