InterDreamer: Zero-Shot Text to 3D Dynamic Human-Object Interaction

Sirui Xu[†] **Ziyin Wang**[†] **Yu-Xiong Wang**[‡] **Liang-Yan Gui**[‡]
University of Illinois Urbana-Champaign
[†] Equal Contribution [‡] Equal Advising
{siruixu2, ziyin, yxw, lgui}@illinois.edu
https://sirui-xu.github.io/InterDreamer/

Abstract

Text-conditioned human motion generation has experienced significant advancements with diffusion models trained on extensive motion capture data and corresponding textual annotations. However, extending such success to 3D dynamic human-object interaction (HOI) generation faces notable challenges, primarily due to the lack of large-scale interaction data and comprehensive descriptions that align with these interactions. This paper takes the initiative and showcases the potential of generating human-object interactions *without direct training on text-interaction pair data*. Our *key insight* in achieving this is that interaction semantics and dynamics can be decoupled. Being unable to learn interaction semantics through supervised training, we instead leverage pre-trained large models, synergizing knowledge from a large language model and a text-to-motion model. While such knowledge offers high-level control over interaction semantics, it cannot grasp the intricacies of low-level interaction dynamics. To overcome this issue, we introduce a world model designed to comprehend simple physics, modeling how human actions influence object motion. By integrating these components, our novel framework, InterDreamer, is able to generate text-aligned 3D HOI sequences without relying on paired text-interaction data. We apply InterDreamer to the BEHAVE, OMOMO, and CHAIRS datasets, and our comprehensive experimental analysis demonstrates its capability to generate realistic and coherent interaction sequences that seamlessly align with the text directives.

1 Introduction

Text-guided human motion generation has made unprecedented progress through advancements in diffusion models [41, 105, 106, 131], leading to synthesis outcomes that are realistic, diverse, and controllable. This progress has ignited an increased interest in exploring expanded tasks related to text-guided human interaction generation, such as social interaction [69] and human-scene interaction [44]. However, many of these explorations are limited in that the dynamics of objects is not involved or text-guided. Aiming to bridge such a gap, this paper tackles a more challenging task – *generating versatile 3D human-object interactions (HOIs) through language guidance*, as illustrated in Figure 1.

Although a direct solution, as suggested by the concurrent work [28, 65, 91, 107, 136, 140], would be replicating the success observed in human motion generation and adopting a similar supervised approach for learning text-driven HOIs, it is not scalable. As can be observed, generating social or scene interactions is heavily dependent on extensive collections of text-interaction pair data [34, 69, 83, 130], and scaling these methods to address more complex HOIs outlined in our study could require datasets of comparable magnitude. Achieving this goal appears unattainable by merely annotating existing 3D HOI datasets [7, 30, 45, 47, 53, 66, 169, 170, 177, 180], which are relatively limited in

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

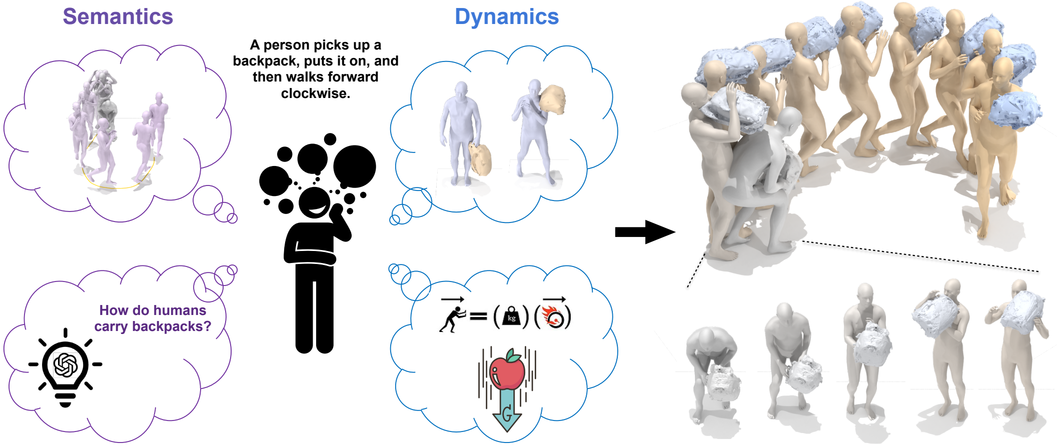

Semantics
Dynamics
A person picks up a backpack, puts it on, and then walks forward clockwise.
How do humans carry backpacks?

Figure 1: InterDreamer generates vivid 3D human-object interaction sequences guided by text descriptions, by synergizing semantics and dynamics knowledge from large-scale text-motion data (upper left), a large language model (bottom left), human-object interaction data (upper middle), and prior knowledge (bottom middle) from simple physics. We visualize the generated text-guided interaction sequence (upper right), with the beginning of the sequence unfolded (bottom right).

size. Although recent studies [28, 65, 91, 155] have annotated some of these datasets, the volume of text-interaction pairs still lags behind that available for existing text-driven motion generation efforts.

An intriguing question naturally arises: given the limited annotations of the text, *what is the potential of learning for text-conditioned HOI generation without text supervision*, which is the main focus of this paper. However, formulating the task in such a setting presents significant challenges, primarily due to the inability to directly learn the alignment between text and HOI dynamics. Our key observation is that interaction semantics and dynamics can be *decoupled*. That is, the high-level semantics of an interaction, aligned with its textual description, can be informed by *human motion* and the *initial object pose*. Meanwhile, the low-level dynamics of the interaction – specifically, the *subsequent* behavior of the object – is governed by the forces exerted by the human, within the constraints of physical laws. Motivated by these insights, we introduce InterDreamer – a novel framework that synergizes knowledge of interaction semantics and dynamics (Figure 1), both of which do not necessarily require learning from text-interaction pairs, if they are decoupled.

The semantics of interaction, although not available through direct supervised training, can be harnessed from *prior knowledge* without text-interaction pair datasets. Specifically, to acquire semantically aligned interaction, we first consult a large language model (LLM), such as GPT-4 [88] and Llama 2 [120], to provide understanding including how humans typically use specific body parts in interactions with particular objects, by exploiting its *in-context learning* capability with *few-shot prompting* [10] and *chain-of-thought prompting* [134]. The intermediate thoughts and the final thought are then used to (**i**) generate semantically aligned human motion with a pre-trained text-to-motion model; and (**ii**) identify an initial object pose that is harmonious with the generated human pose and text description, following a philosophy similar to *retrieval-augmented generation* [62].

While these large models can offer high-level motion semantic modeling, they lack crucial *low-level* dynamics knowledge. Nevertheless, by decoupling interaction dynamics from semantics, a key advantage emerges in our InterDreamer framework: interaction dynamics can be learned from motion capture data *without the necessity of text annotations*. We instantiate this idea by developing a *world model*, which predicts the subsequent state of an object affected by the interaction. The key here is to reach *generalizable representations* in different motion and objects. To do so, we exert control over the object through the motion of vertices on the human body. These vertices are solely sampled in regions where contact occurs, *agnostic* to the overall object shape and whole-body motion. Such abstraction empowers the model to learn the simple dynamics from a publicly available 3D HOI dataset BEHAVE [7], and generalize naturally to other datasets [47, 66]. The plausibility of the generated interaction is further enhanced by a subsequent optimization procedure on the synthesized human and object motion.

To summarize, our contributions are: (**i**) We address the task of synthesizing whole-body interactions with dynamic objects guided by textual commands, achieving this notably without the need for paired text-interaction data, a novel paradigm to the best of our knowledge. (**ii**) We introduce a framework that decomposes semantics and dynamics, and they can be easily integrated. Specifically, it harnesses knowledge from a large language model (LLM) and a text-to-motion model as external resources, alongside our proposed world model. Remarkably, the only component that requires additional training is the world model, which highlights the *ease of use* of our framework. (**iii**) Experimental results demonstrate that our framework, InterDreamer, is capable of producing semantically aligned and realistic human-object interactions, and generalizes *beyond existing HOI datasets*.

2 Related Work

Text-Conditioned Human Motion Generation. Significant progress has been witnessed in human motion synthesis tasks, given different kinds of external conditions, including action categories [2, 36, 61, 93], past motion [5, 17, 86, 110, 149, 150, 163], trajectories [31, 50, 51, 100, 122, 145], scene context [12, 29, 39, 44, 113, 114, 125–127, 130, 153, 175, 182, 183], and without condition [96]. Recently, human motion synthesis guided by textual descriptions [1, 6, 18, 24, 26, 34, 35, 49, 54, 57, 64, 71, 77, 81, 94, 95, 97, 104, 116, 133, 160, 162, 168, 172, 174, 178, 179, 181, 185, 187] is popular and extended to various applications, including the text-conditioned generation of multiple-person [33, 43, 63, 75, 132] and human-scene interaction [14, 21, 44, 48]. Our goal is to model human and object dynamics concurrently guided by text.

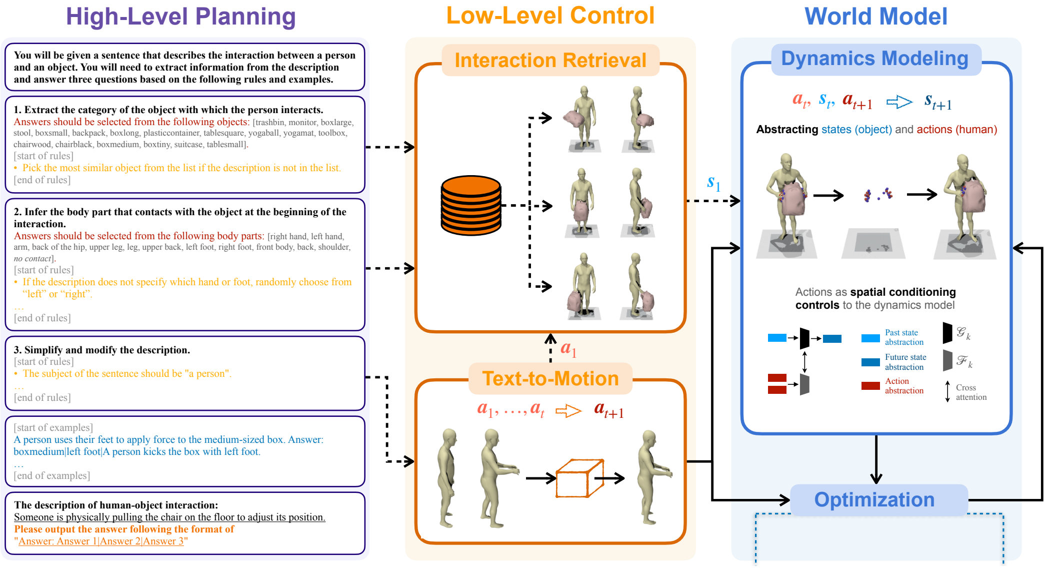
Human-Object Interaction Generation. Synthesizing hand-object interactions [11, 15, 20, 68, 73, 79, 80, 119, 137, 161, 165, 167, 184, 186] and single-frame human-object interactions [25, 42, 56, 92, 128, 143, 152, 154, 166] are popular topics and extended to zero-shot settings [52, 67, 156, 157]. Recently, researchers explore whole-body dynamic interaction generation, in kinematic-based approaches [22, 27, 32, 38, 58–60, 66, 84, 85, 99, 107–109, 111, 123, 136, 139, 140, 148, 151, 176] and physics-based approaches [4, 8, 16, 23, 40, 72, 74, 78, 87, 89, 115, 117, 124, 129, 142, 146, 147, 158]. Current methods in HOI synthesis are often restricted by a narrow scope of actions, the use of non-dynamic objects, and a lack of comprehensive whole-body motion. Our work aims to generate diverse whole-body interactions with various objects, and enables control through language input. Recent datasets [7, 30, 45, 47, 53, 66, 112, 138, 144, 155, 164, 169, 170, 180] provide the groundwork for research in this area, and concurrent efforts [28, 65, 91] demonstrate the feasibility of applying supervised learning methods via annotating datasets. However, the amount of data currently available fall short when compared to more extensive text-motion datasets [34, 70, 83]. This discrepancy in data volume limits the capability of supervised methods to capture the complexity of human-object interactions, motivating us to investigate the potential of zero-shot generation.

External Knowledge from LLMs. Large language models (LLMs) are being used for advanced visual tasks, such as editing images based on instructions [9]. In digital humans, they are used to reconstruct 3D human-object interactions [128] and generate human motion [3, 46, 159, 178] as well as human-scene interactions [141]. Our approach is inspired by [128], which uses LLMs to infer contact body parts with a given object for reconstructing 3D human-object interactions – a task different from ours. Our approach utilizes GPT-4 [88] or Llama 2 [120], to not only understand contact body parts but also narrow the distribution gap between different tasks, and provide knowledge for interaction retrieval. This is accomplished by utilizing the in-context learning capabilities of LLMs [22] and their support for retrieval-augmented generation [62].

3 Methodology

Problem Formulation. Our goal is to synthesize a sequence of 3D human-object interactions $\boldsymbol{x}$ that satisfies a descriptive text p. This sequence is a series of tuples $[(\boldsymbol{h}_1, \boldsymbol{o}_1), (\boldsymbol{h}_2, \boldsymbol{o}_2), \dots, (\boldsymbol{h}_M, \boldsymbol{o}_M)]$, where $\boldsymbol{h}_i$ represents the human pose parameters defined in the SMPL model [76], while the shape of the human is unified the same as [34]. $\boldsymbol{o}_i$ defines the pose of the rigid object in terms of its 3D spatial position and orientation. The sequence length M is variable and is dynamically determined by our text-to-motion model based on the input text p. We do *not* require text supervision for training.

Overview. Our framework, illustrated in Figure 2, can be conceptualized as a Markov decision process (MDP). We begin by dividing the motion sequence into T segments, each with m frames,



You will be given a sentence that describes the interaction between a person and an object. You will need to extract information from the description and answer three questions based on the following rules and examples.

1. Extract the category of the object with which the person interacts.
Answers should be selected from the following objects: [trashbin, monitor, boxlarge, stool, boxsmall, backpack, boxlong, plasticcontainer, tablesquare, yogaball, yogamat, toolbox, chairwood, chairblack, boxmedium, boxtiny, suitcase, tablesmall].
[start of rules]
- Pick the most similar object from the list if the description is not in the list.

[end of rules]

2. Infer the body part that contacts with the object at the beginning of the interaction.
Answers should be selected from the following body parts: [right hand, left hand, arm, back of the hip, upper leg, leg, upper back, left foot, right foot, front body, back, shoulder, *no contact*].
[start of rules]
- If the description does not specify which hand or foot, randomly choose from “left” or “right”.

…
[end of rules]

3. Simplify and modify the description.
[start of rules]
- The subject of the sentence should be "a person".

…
[end of rules]

[start of examples]
A person uses their feet to apply force to the medium-sized box. Answer: boxmedium|left foot|A person kicks the box with left foot.
…
[end of examples]

The description of human-object interaction:
Someone is physically pulling the chair on the floor to adjust its position.
Please output the answer following the format of
"Answer: Answer 1|Answer 2|Answer 3"

Figure 2: **An overview of our InterDreamer.** **(i)** Our high-level planning analyzes the description using LLMs and provides guidance to the low-level control. **(ii)** Our low-level control includes a text-to-motion model that translates text into human actions $\boldsymbol{a}_{t+1}$, and an interaction retrieval model that extracts the object's first state $\boldsymbol{s}_1$. **(iii)** Our world model executes actions to output the next state $\boldsymbol{s}_{t+1}$. We achieve this by abstracting the problem as predicting the motion of contact vertices – represented by red spheres for humans and blue spheres for objects on the top right – using human vertices as controls for the prediction of object vertices. An optimization process is coupled with the dynamics model, projecting the state and action onto valid counterparts. Solid arrows mean that the process is performed iteratively.

where $M = T \times m$. Object motion $\{\boldsymbol{o}_i\}_{i=1}^M$ can be seen as a sequence of environmental states $\{\boldsymbol{s}_t\}_{t=1}^T$, and human motion $\{\boldsymbol{h}_i\}_{i=1}^M$ is described as a sequence of actions $\{\boldsymbol{a}_t\}_{t=1}^T$ that interact with the environment. Under such an MDP setup, our framework starts with high-level planning L, which deciphers textual interaction description p by $g = L(p)$ (Sec. 3.1). Then, a text-to-motion model π translates context g into human actions, modeled as $\boldsymbol{a}_{t+1} \sim \pi(\boldsymbol{a}_{t+1} | \boldsymbol{s}_t, \{\boldsymbol{a}_i\}_{i=1}^t, g)$ (Sec. 3.2). The interaction retrieval R proposes an initial object state $\boldsymbol{s}_1 \sim R(\boldsymbol{s}_1 | \boldsymbol{a}_1, g)$, based on the initial action $\boldsymbol{a}_1$ and context g (Sec. 3.2). After that, a world model P is trained to predict future states $\boldsymbol{s}_{t+1} \sim P(\boldsymbol{s}_{t+1} | \boldsymbol{a}_t, \boldsymbol{s}_t, \boldsymbol{a}_{t+1})$ from the current action and state (Sec. 3.3). Our world model incorporates an optimization process, for both state and action refinement (Sec. 3.4). Notably, the text-to-motion and world models are executed *iteratively* until text-to-motion generates an end frame.

3.1 High-Level Planning

Leveraging LLMs' strong reasoning capabilities and inherent common sense, our high-level planning L yields interaction details $g = L(p)$ that cannot be naïvely extracted in textual descriptions p. The process undertaken by L encompasses three steps: **(i)** *Determining the object*: the LLM is employed to translate described objects into corresponding categories from a predefined list. **(ii)** *Determining initial human-object contact*: the LLM infers the body parts involved in the interaction, drawing from a list defined in the SMPL model [76]. And most importantly, **(iii)** *reducing the distribution gap*: the LLM bridges the distribution gap between the free-form textual input and the language used within the training data of the text-to-motion model [34]. This involves standardizing syntax and content according to designed guidelines. In Figure 2, we demonstrate the prompt we used with the few-shot prompting [10]. We define intermediate thoughts and the final thought, *i.e.*, answers to three questions, as detailed information $g = L(p)$, which guides the subsequent procedure, structuring the entire framework with a philosophy similar to retrieval-augmented generation [62]. Our high-level planning operates indirectly in the generation of interactions. Nonetheless, it narrows the vast range of possible interactions in the real world into a more manageable distribution within the capabilities of our framework. We incorporate GPT-4 [88] and Llama-2 [120] for evaluation.

3.2 Low-Level Control

With the information g derived from the description p, the low-level control aims to create a sequence of human actions $\{\boldsymbol{a}_t\}_{t=1}^{T}$ by a text-to-motion model, and an initial state $\boldsymbol{s}_1$ by interaction retrieval, such that they correspond to the objectives outlined by g.

Text-to-Motion. We utilize a text-to-motion model π to develop actions to be executed in the world model. At each timestep t, π receives the sequence of previous actions $\{\boldsymbol{a}_i\}_{i=1}^{t}$ and the text tokens encoded from the rewritten description in $g = L(p)$, and produces a next action $\boldsymbol{a}_{t+1}$, which later in Sec. 3.4 will be adjusted through an optimization process that intertwines actions with the object state $\boldsymbol{s}_t$. Thus, the overall process can be formally defined as $\boldsymbol{a}_{t+1} \sim \pi(\boldsymbol{a}_{t+1}|\boldsymbol{s}_t, \{\boldsymbol{a}_i\}_{i=1}^{t}, g)$, while the initial action $\boldsymbol{a}_1 \sim \pi(\boldsymbol{a}_1|g)$ is influenced merely by context g without prior actions or states, which will be used in interaction retrieval. π builds upon existing text-to-motion models, where we evaluate MDM [118], MotionDiffuse [172], ReMoDiffuse [173], and MotionGPT [46].

Interaction Retrieval. The interaction retrieval component R establishes the initial state $\boldsymbol{s}_1 \sim R(\boldsymbol{s}_1|\boldsymbol{a}_1, g)$, based on the initial action $\boldsymbol{a}_1$ generated by the text-to-motion model. We propose a user-friendly pipeline for this purpose built on handcrafted rules. First, we create databases by extracting HOI frames from the training sets of each target datasets — BEHAVE [7], OMOMO [66], and CHAIRS [47]. The indexing key for retrieval is a tuple consisting of the body part in contact and the category of the involved object. Each retrieval value is a per-frame contact map, represented by a list of K vertex pairs $\{(d_h^i, d_o^i)\}_{i=1}^{K}$. Here, d_h^i refers to the contact vertex on the human surface, while d_o^i refers to the corresponding contact vertex on the object surface. This contact map is linked to its corresponding key, creating a searchable record of interactions. During the inference stage, using the body part and object information provided by the high-level planning (Sec. 3.1), we retrieve all relevant contact maps from the database. We then sample one map $\{(d_h^i, d_o^i)\}_{i=1}^{K}$ and use it to establish the object state $\boldsymbol{s}_1 \sim R(\boldsymbol{s}_1|\boldsymbol{a}_1, g)$, thus initializing the interaction. Further details including how we ensure consistency between the sampled state and human action are provided in Sec. B.1 of the Appendix. We also discuss an alternative learning-based approach in Sec. B.1.

3.3 World Model

Our world model combines a dynamics model and the optimization process, dedicated to simulating state transitions affected by applied actions. While drawing inspiration from similar concepts utilized in robotics [103, 135] and autonomous driving systems [55], we use it here to generate HOI trajectories. This model, trained on the training set of a 3D HOI dataset such as BEHAVE [7], serves a similar role as a simulator but is much simpler – it takes the preceding object state $\boldsymbol{s}_t$ along with a pair of consecutive actions $\boldsymbol{a}_t$ and $\boldsymbol{a}_{t+1}$, and predicts the subsequent object state $\boldsymbol{s}_{t+1}$. The interplay between the low-level control and the world model ultimately produces a coherent interaction rollout.

In designing the dynamics model, a naïve method would be directly taking raw actions, states, and object geometry as input. However, this suffers from a severe generalization problem during inference: the dynamics model is likely to encounter human actions and object geometry that do not exist in the training set, since our text-to-motion model is not trained with object interaction data. To overcome this limitation, we instead focus on encoding interactions through the contact vertices on the object, which capture both the action and object geometry, as shown in Figure 2. This *locality* ensures that the dynamics model remains focused on interactions in the contact region, without being distracted by the motion of body parts and geometry details that are irrelevant to the interaction.

Input Representation. Specifically, at each timestep t, we abstract the past actions as H historical vertex trajectories $\{\{\boldsymbol{v}_i^j\}_{j=1}^{N}\}_{i=1}^{H}$, and the future actions as $F = m$ future vertex trajectories $\{\{\boldsymbol{v}_i^j\}_{j=1}^{N}\}_{i=H+1}^{H+F}$, where non-fixed variable N is the number of sampled contact vertices, and m is the length of segments as mentioned in the overview of Sec 3. Note that we train our dynamics model to forecast over a longer duration than the past motion ($F > H$), only the foremost future action will be used for autoregressive generation during the inference, as suggested in [19]. To determine these N vertices, we start with object's signed distance fields $\{\mathbf{sdf}_i\}_{i=1}^{H}$ over the past H frames, derived from the past state $\boldsymbol{s}_t$. We then sample vertices that meet the following criteria: $|\mathbf{sdf}_i(\boldsymbol{v}_i^j)| \leq \delta_1, \mathbf{sdf}_i(\boldsymbol{v}_i^k)| \leq \delta_1, \forall i = 1, \ldots, H, \forall j$, and $\|\boldsymbol{v}_i^j - \boldsymbol{v}_i^k\| \geq \delta_2, \forall j \neq k$, where δ_1 and δ_2 are two hyperparameters. The objective is to sparsely sample contact vertices while ensuring that they are sufficient to encompass the interaction. We characterize each vertex trajectory $\{\boldsymbol{v}_i^j\}_{i=1}^{H+F}$

Table 1: **Quantitative results** on evaluating the dynamics model. Our dynamics model with vertex-based action generates interactions of the best quality.

Methods	Text-to-Interaction		Interaction Prediction [148]		
	CMD ↓	Pene. (10^{-2}%) ↓	Trans. Err. (mm) ↓	Rot. Err. (10^{-3} rad) ↓	Pene. (10^{-2}%) ↓
w/o action	0.424	533	123	256	228
contact markers as action (InterDiff [148])	0.219	484	123	226	164
human motion as action	0.325	957	129	265	218
contact vertices as action (**ours**)	**0.151**	**443**	**119**	**221**	**156**

Table 2: **Quantitative results** on human motion quality given our annotation on the BEHAVE [7] dataset. We show that our high-level planning effectively adapts single human generators into human-object interaction generation. To evaluate R-Precision, a batch size of 16 is selected.

Methods	Planning (**Ours**)	R-Precision$^{\uparrow}$			FID$^{\downarrow}$	MM Dist$^{\downarrow}$	Multimodality$^{\uparrow}$	Diversity$^{\rightarrow}$
		Top 1	Top 2	Top 3				
Ground Truth	-	0.237$^{\pm0.004}$	0.392$^{\pm0.004}$	0.496$^{\pm0.005}$	0.024$^{\pm0.000}$	4.259$^{\pm0.006}$	-	6.510$^{\pm0.227}$
MDM [118]	×	0.153$^{\pm0.016}$	0.279$^{\pm0.026}$	0.398$^{\pm0.016}$	12.279$^{\pm0.217}$	5.351$^{\pm0.057}$	**7.604**$^{\pm0.190}$	7.598$^{\pm0.334}$
	✓	**0.163**$^{\pm0.010}$	**0.307**$^{\pm0.043}$	**0.402**$^{\pm0.019}$	**10.374**$^{\pm0.304}$	**5.303**$^{\pm0.117}$	7.281$^{\pm0.083}$	**7.471**$^{\pm0.427}$
MotionDiffuse [172]	×	0.205$^{\pm0.011}$	0.351$^{\pm0.002}$	0.458$^{\pm0.021}$	10.208$^{\pm0.500}$	4.837$^{\pm0.064}$	4.520$^{\pm0.163}$	7.323$^{\pm0.412}$
	✓	**0.216**$^{\pm0.032}$	**0.369**$^{\pm0.023}$	**0.472**$^{\pm0.027}$	**9.015**$^{\pm0.403}$	**4.649**$^{\pm0.029}$	**4.991**$^{\pm0.172}$	**7.295**$^{\pm0.501}$
ReMoDiffuse [173]	×	0.196$^{\pm0.009}$	0.338$^{\pm0.011}$	0.448$^{\pm0.012}$	6.385$^{\pm0.201}$	4.855$^{\pm0.029}$	5.889$^{\pm0.524}$	**7.160**$^{\pm0.306}$
	✓	**0.223**$^{\pm0.006}$	**0.368**$^{\pm0.015}$	**0.482**$^{\pm0.011}$	**5.237**$^{\pm0.174}$	**4.784**$^{\pm0.053}$	**6.350**$^{\pm0.411}$	7.201$^{\pm0.318}$
MotionGPT [46]	×	0.233$^{\pm0.003}$	0.344$^{\pm0.004}$	0.457$^{\pm0.005}$	5.497$^{\pm0.106}$	5.205$^{\pm0.027}$	1.062$^{\pm0.211}$	8.316$^{\pm0.204}$
	✓	**0.234**$^{\pm0.004}$	**0.387**$^{\pm0.003}$	**0.471**$^{\pm0.007}$	**4.751**$^{\pm0.121}$	**4.995**$^{\pm0.003}$	**1.337**$^{\pm0.193}$	**7.106**$^{\pm0.487}$

with a feature $\boldsymbol{f}^j$ to provide (i) human vertex coordinates at T-pose, providing information about the position of the human vertex on the body surface; (ii) the vertex-to-object surface vector, indicating vertex's impact on the object as well as inherently including information related to the object's shape; and (iii) the vertex's velocity relative to its nearest object vertex. Thus, the model needs to learn how the features of human action $\boldsymbol{f}^j$ affect the evolution of the state of the object.

Architecture. As demonstrated in Figure 2, the network comprises two components: $\mathcal{G}$ that operates without contact vertex conditions, applicable in scenarios where no contact occurs, and $\mathcal{F}$, which incorporates contact vertex conditions into the object trajectory when contact is present. The k-th layer of $\mathcal{G}$ can be initiated as $\mathcal{G}_k(\boldsymbol{x}_k, \boldsymbol{\Theta})$, mapping the input feature map $\boldsymbol{x}_k$ at the k-th layer to another feature map, with Θ denoting the MLP's parameters. To incorporate human vertex controls, we introduce a second network $\mathcal{F}_k(\boldsymbol{y}_k^j, \boldsymbol{\Theta}_v)$ operating on N vertex features $\{\boldsymbol{y}_k^j\}_{j=1}^N$, where $\boldsymbol{\Theta}_v$ is its parameters. With a cross-attention layer Attn, a dynamics block is formulated as: $\boldsymbol{x}_{k+1}, \{\boldsymbol{y}_{k+1}^j\}_{j=1}^N = \mathrm{Attn}(\mathcal{G}_k(\boldsymbol{x}_k, \boldsymbol{\Theta}), \{\mathcal{F}_k(\boldsymbol{y}_k^j, \boldsymbol{\Theta}_v)\}_{j=1}^N)$. We stack multiple dynamics blocks to form the model. The initial input, $\boldsymbol{x}_0$, corresponds to the previous state $\boldsymbol{s}_t$, while each $\boldsymbol{y}_0^j$ represents the feature of the vertex trajectory, containing both the trajectory $\{\boldsymbol{v}_i^j\}_{i=1}^{H+F}$ and its associated feature vector $\boldsymbol{f}^j$. The output of this model is preliminary and subject to further optimization as introduced in Sec. 3.4, which will yield the final future state. We utilize the Mean Squared Error loss to train the dynamics model. For more explanations, please refer to Sec. B.2 of the Appendix.

3.4 Optimization

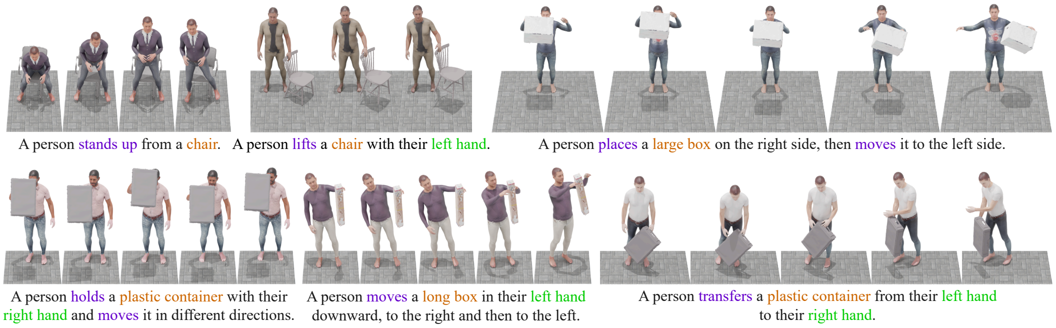
Optimization plays a role in introducing prior knowledge and avoiding the accumulation of errors. During inference, we input the action $\boldsymbol{a}_{t+1}$ and state $\boldsymbol{s}_{t+1}$ and refine them. This refinement is achieved through gradient descent on the human and object pose parameters. Our optimization includes several loss terms: a fitting loss to align optimized results with their preliminary one, a velocity loss for temporal smoothness, a contact loss to promote occurring contact, and a collision loss to reduce penetration. We provide detailed formulations in Sec. B.3 of the Appendix.

4 Experiments

Extensive comparisons evaluate the performance of our InterDreamer across two motion-relevant tasks. Details of the evaluation settings are provided in Sec. 4.1. We present both quantitative (Sec. 4.2) and qualitative (Sec. 4.3) results for our approach. Additionally, we perform ablation studies to verify the efficacy of each component within our framework. These studies also cover the

Table 3: **Quantitative results** on human motion quality on the OMOMO [66] dataset with their provided annotation. We show that our high-level planning narrows the distribution gap and adapts single human generators into human-object interaction generation. To evaluate R-Precision, a batch size of 32 is selected.

Methods	Planning (**Ours**)	R-Precision$^{\uparrow}$			FID$^{\downarrow}$	MM Dist$^{\downarrow}$	Multimodality$^{\uparrow}$	Diversity$^{\rightarrow}$
		Top 1	Top 2	Top 3				
Ground Truth	-	$0.044^{\pm0.004}$	$0.095^{\pm0.008}$	$0.151^{\pm0.009}$	$0.000^{\pm0.000}$	$6.858^{\pm0.006}$	-	$5.660^{\pm0.110}$
MDM [118]	×	$0.056^{\pm0.005}$	$0.096^{\pm0.007}$	$0.135^{\pm0.006}$	$16.638^{\pm0.631}$	$7.110^{\pm0.063}$	$2.446^{\pm0.456}$	**5.862**$^{\pm0.520}$
	✓	**0.062**$^{\pm0.006}$	**0.109**$^{\pm0.004}$	**0.155**$^{\pm0.008}$	**15.735**$^{\pm0.285}$	**6.889**$^{\pm0.082}$	**2.663**$^{\pm0.520}$	$6.461^{\pm0.841}$
MotionDiffuse [172]	×	$0.048^{\pm0.006}$	$0.094^{\pm0.008}$	$0.143^{\pm0.013}$	$15.442^{\pm0.231}$	**5.799**$^{\pm0.054}$	$1.658^{\pm0.209}$	$5.981^{\pm0.516}$
	✓	**0.075**$^{\pm0.005}$	**0.141**$^{\pm0.015}$	**0.189**$^{\pm0.009}$	**10.815**$^{\pm0.093}$	$5.916^{\pm0.094}$	**1.677**$^{\pm0.264}$	**5.718**$^{\pm0.522}$
ReMoDiffuse [173]	×	$0.062^{\pm0.003}$	$0.111^{\pm0.005}$	$0.160^{\pm0.012}$	$15.479^{\pm0.209}$	$5.690^{\pm0.049}$	$1.179^{\pm0.145}$	$6.032^{\pm0.540}$
	✓	**0.067**$^{\pm0.004}$	**0.127**$^{\pm0.006}$	**0.174**$^{\pm0.006}$	**14.560**$^{\pm0.080}$	**5.678**$^{\pm0.033}$	**1.193**$^{\pm0.202}$	**5.368**$^{\pm0.417}$
MotionGPT [46]	×	$0.061^{\pm0.005}$	$0.114^{\pm0.006}$	$0.152^{\pm0.006}$	$18.472^{\pm0.528}$	$6.358^{\pm0.076}$	**4.553**$^{\pm0.244}$	**6.726**$^{\pm0.156}$
	✓	**0.064**$^{\pm0.007}$	**0.121**$^{\pm0.007}$	**0.164**$^{\pm0.009}$	**17.512**$^{\pm0.498}$	**6.287**$^{\pm0.041}$	$4.470^{\pm0.191}$	$7.048^{\pm0.169}$


A person stands up from a chair. A person lifts a chair with their left hand. A person places a large box on the right side, then moves it to the left side.

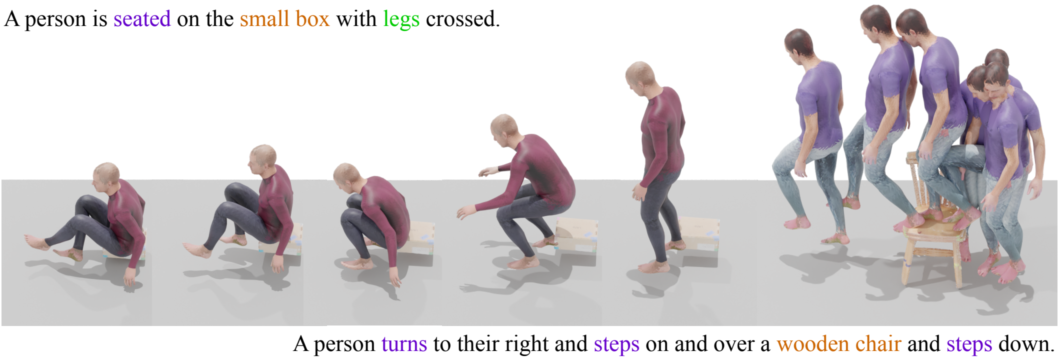
A person holds a plastic container with their right hand and moves it in different directions. A person moves a long box in their left hand downward, to the right and then to the left. A person transfers a plastic container from their left hand to their right hand.

Figure 3: **Qualitative results** on free-form text input. The interaction sequences, with textures from [13], are presented through a time-series visualization.

interaction prediction task [148] to evaluate our dynamics model. Additional details and results are presented in Sec. C and Sec. D of the Appendix. Please refer to our website for video results.

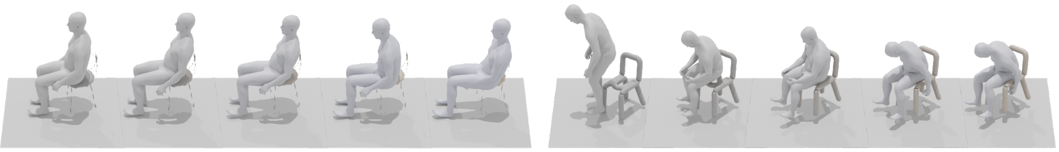
4.1 Experimental Setup

Datasets. We use BEHAVE [7], CHAIRS [47], and OMOMO [66] datasets for quantitative evaluation. The BEHAVE dataset includes recordings of 8 individuals interacting with 20 everyday objects, and our analysis focuses on 18 objects for which interaction sequences are available at 30 Hz. The human pose is modeled using SMPL-H [102], with hand poses set to an average pose *due to the absence of detailed hand pose in the dataset*. We manually segment the long interaction sequences in the test set, and annotate them with descriptions as well as their starting and ending indices, leading to 532


A person is seated on the small box with legs crossed.

A person turns to their right and steps on and over a wooden chair and steps down.

Figure 4: **Qualitative results** in more challenge scenarios with *free-form input not* from our annotations, showing the ability of our InterDreamer to fit *object sizes* and handle *complex and long sequences*. Here, our synergized models are GPT-4 [88] and MotionGPT [46].



A seated person on chair 63 adjusts the sitting position repeatedly. Someone can be seen on chair 96 and adjusts their sitting position.

Figure 5: **Qualitative results** on the CHAIRS [47] dataset. Our dynamics model trained on the BEHAVE [7] dataset generalizes well on the CHAIRS objects unseen in training. Frames are separately visualized. Here, our synergized models are GPT-4 [88] and MotionGPT [46].

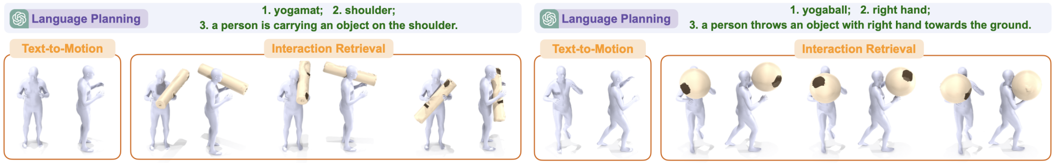
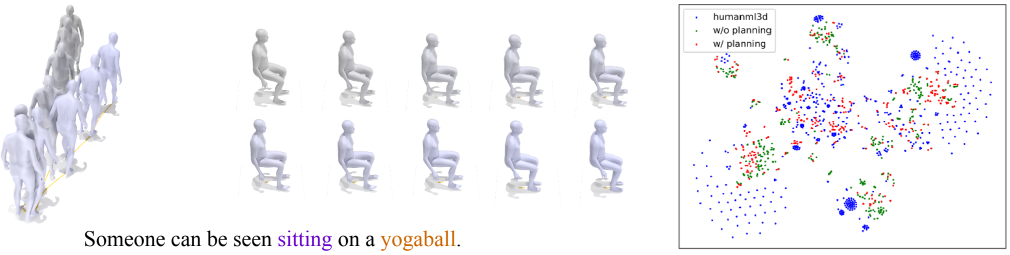

Language Planning
1. yogamat; 2. shoulder;
3. a person is carrying an object on the shoulder.
Text-to-Motion
Interaction Retrieval
Language Planning
1. yogaball; 2. right hand;
3. a person throws an object with right hand towards the ground.
Text-to-Motion
Interaction Retrieval

Figure 6: Results from the interaction retrieval. We demonstrate that our proposed retrieval approach based on handcraft rules can extract diverse and realistic interactions.

subsequences for evaluation. For qualitative evaluation, we go beyond using annotations specifically created and employ free – form text to demonstrate our model's capability on out-of-distribution inputs. To assess our model's performance with novel objects, we expand our retrieval database to include objects from the OMOMO [66] and CHAIRS [47] datasets, while we do not fine-tune the dynamics model on them–a direct qualitative evaluation without additional adaptation.

Metrics. The evaluation metrics are divided into three categories: (**i**) *Human motion quality*: The Fréchet Inception Distance (**FID**) measures the distance between the generated motion and ground truth. The MultiModality (**Multimodality**) and **Diversity** metrics assess the variance in generated human motion. **R-Precision** evaluates the consistency between the text and the generated human motion within the latent space. MultiModal distance (**MM Dist**) is the distance between the motion feature and the text feature. We follow [34] to generate motion and text features. (**ii**) *Interaction quality*: We propose **CMD** to measure the distance between contact maps of real interactions and those generated. The per-sequence contact map is defined by the percentage of time that each body part is actively in contact. The detailed formulation is provided in Sec. C of the Appendix. We also measure the collision (**Pene.** [148]), which calculates the average percentage of object vertices that have non-negative values in the human signed distance fields [90]. (**iii**) *Object motion accuracy*: The dynamics model's performance in the interaction prediction task [148] is evaluated by the accuracy of predicted object motion, including **Trans. Err.**, the average distance between predicted and ground truth, and **Rot. Err.**, the average distance between the predicted and ground truth.



Someone can be seen sitting on a yogaball.

(a) Low-level control w/o planning v.s. w/ planning

(b) Text feature w/ planning v.s. w/o planning

Figure 7: (**a**) **Ablation study** on the high-level planning. On the *left* are results from MotionGPT [46] using free-form descriptions, and on the *right* are results with our planning additionally. Without planning, the motion generative model struggles to interpret free-form HOI descriptions and generate semantically-aligned motion. (**b**) We visualize CLIP [98] features of text on HumanML3D [34] via t-SNE [82], raw HOI descriptions ("w/o planning"), and HOI descriptions processed through our high-level planning ("w/ planning"). See Table 5 for quantitative measurements.

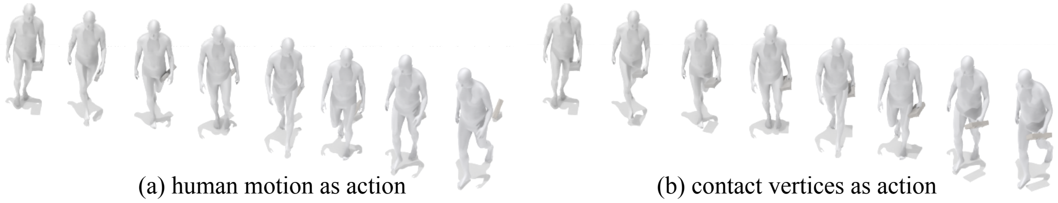


Figure 8: **Ablation study** on the dynamics model. Given the text description of "A person walks clockwise while holding a small box with left hand," our (**b**) vertex-based control can synthesize consistent contacts, which (**a**) the baseline fails to do.

Baselines. Most recent research on text-to-HOI synthesis follows a supervised learning approach [28, 91], making direct quantitative comparisons unfair. Therefore, we primarily focus on qualitative comparisons with these methods. To enable quantitative evaluation, we develop a range of baselines to assess both the overall performance of our framework and the effectiveness of its individual components. In the context of high-level planning, we utilize GPT-4 [88] and Llama-2 [120], illustrating the effectiveness of our prompts across different language models. For low-level motion generation control, our baselines include MDM [118], MotionDiffuse [172], ReMoDiffuse [173], and MotionGPT [46], which span a range of text-to-motion approaches trained on HumanML3D [34] and show the generalizability of our framework. To evaluate the dynamics model, we include different design choices: (**i**) unconditional dynamics model which operates object dynamics independently of human motion; (**ii**) using human marker features as actions to the dynamics model, similar to [148]; (**iii**) using unprocessed human motion and object pointcloud motion as input to the dynamics model; (**iv**) our proposed vertex-based actions where only the contacting vertices are used for control.

4.2 Quantitative Results

In Table 1, comparing our framework to baselines with unconditional dynamics model, InterDreamer achieves better interaction quality in terms of CMD and penetration scores, showing the importance of human influence on object motion. Against methods that utilize direct raw human motion or markers for action features, our method demonstrates enhanced performance by offering more fine-grained guidance and extracting generalizable features for dynamics modeling. Tables 2 and 3 present a comparative analysis of our approach of combining high-level planning with low-level control, where we adopt various text-to-motion models against their counterparts without high-level planning on the BEHAVE and OMOMO datasets. Our approach consistently outperforms baselines. Specifically, InterDreamer exhibits superior motion quality, reflected by a lower FID, higher R-Precision, and better diversity, highlighting the benefits of incorporating our planning to reduce the distribution gap for the motion generator to generalize in the HOI synthesis task.

4.3 Qualitative Results

Figure 3 displays several results guided by the free-form text. Our method exhibits proficiency in interpreting the textual input and synthesizing dynamic, realistic interactions, despite the absence of training with text-interaction paired data. More importantly, as illustrated in Figure 4, we selectively use more complex sequences of interactive descriptions that are *beyond the scope of the existing HOI dataset*. Figure 5 further exemplifies our method that is able to generalize effectively to the CHAIRS dataset, despite our dynamics model not being trained on it. Figure 6 depicts the retrieval procedure, resulting in a diverse set of interactions that are both high-quality and semantically aligned. More experimental results and the user study are presented in Sec. D of the Appendix.

Table 4: **Ablation study** on the high-level planning. Q1 and Q2 ask to identify the object category and the contact body part, respectively. We assess the accuracy by comparing the LLM's responses with labels we annotate. Note that the text input to LLMs may contain ambiguities; for example, the annotation is "hand" when the motion uses "right hand." We include Q1 Acc* and Q2 Acc* excluding ambiguous text.

LLM (# of parameters)	Q1 Acc ↑	Q1 Acc* ↑	Q2 Acc ↑	Q2 Acc* ↑
GPT-4 [88]	0.801	0.997	0.703	0.964
Llama-2 (7B) [120]	0.073	0.147	0.436	0.689
Llama-2 (13B) [120]	0.232	0.319	0.662	0.853
Llama-2 (70B) [120]	0.722	0.967	0.798	0.907

Table 5: **Quantitative comparison** of text similarity. The text processed by high-level planning is more similar to text in HumanML3D [34] on average, while addressing the distributional gap significantly for challenging out-of-distributional descriptions, compared to text without planning.

Sim. to [34]↑	Average	Out-of-Dist.
w/o planning	0.913	0.838
w/ planning	**0.932**	**0.927**

4.4 Ablation Study

Adaptability of High-Level Planning. Is our framework adaptable across different large language models (LLMs)? As illustrated in Table 4, our analysis contains two types of language models: GPT-4 [88], which is accessible through APIs and operates as a black box model; and Llama-2 [120], an open-source model. We demonstrate that language models with large parameters exhibit very high accuracy in responding to questions tailored to our prompts, validating the framework's adaptability.

Effectiveness of High-Level Planning with Low-Level Control. In consistency with Table 2, Figure 7 offers a qualitative comparison of text-to-motion results, contrasting results with and without LLM-revised text descriptions. The comparison shows that motion generated without LLM-enhanced descriptions often fails to correspond to the intended text, if the text is too challenging, *e.g.,* not in the distribution of HumanML3D [34], which is used to train text-to-motion models. This underscores the LLM's critical role in bridging the distribution gap. In Figure 7(b), we visualize the CLIP [98] features of descriptions from HumanML3D, our raw annotations, and those processed by high-level planning. Quantitative evidence is provided in Table 5. Text processed through high-level planning demonstrates greater similarity to the HumanML3D dataset. Additionally, we test on more challenging out-of-distribution text, selecting examples with an average cosine similarity to HumanML3D text of less than 0.85. High-level planning successfully rephrases these texts, significantly increasing their similarity. For example, in Figure 7(a), the text "Someone can be seen sitting on a yoga ball" has a cosine similarity of 0.874 to the closest in-distribution text, while the rephrased text after planning, "A person is seated on an object," achieves a similarity of 0.958.

Effectiveness of World Model. In the quantitative evaluation, we show that the performance of our framework is enhanced by the tailored design of our world model. Table 1 provides additional evidence of the effectiveness by integrating the proposed world model, as interaction correction within the InterDiff framework [148] in the interaction prediction task. This implementation demonstrates enhanced conditionality in the object dynamics modeling across two tasks, attributed to the vertex-level condition as actions. Doing so effectively removes the whole-body complexity, most of which tends to be irrelevant to the interaction. Figure 8 further indicates that our vertex-based condition can establish consistent interactions over time, while the condition by raw motion is not robust.

5 Conclusion

We tackle the task of text-guided 3D human-object interaction generation, aiming to accomplish this without relying on paired text-interaction data. To this end, we present InterDreamer that decouples interaction dynamics from semantics, formulating the task as retrieval-augmented generation and Markov decision process, where high-level planning and low-level control are introduced to generate semantically aligned human motion and initial object pose, while a world model is responsible for the object dynamics guided by the interaction. Our approach demonstrates effectiveness in this novel task, suggesting its potential for various real-world applications.

Limitations. The current utilization of dynamics modeling could be enhanced. A prospective improvement involves incorporating model-based learning techniques, which empower the agent to more effectively interact and learn a broader range of skills. The results may not be physically plausible and lead to artifacts in some cases, for example, foot skating. Hand poses are rough because they are missing from the dataset, but could be improved by integrating a physics simulator.

Acknowledgments

We thank Qiusi Zhan for supporting the implementation and evaluation of high-level planning, Jiangwei Yu and Xiyan Xu for their efforts in dataset annotation, and Xiang Li for the discussions. This work was supported in part by NSF Grant 2106825, NIFA Award 2020-67021-32799, the IBM-Illinois Discovery Accelerator Institute, the Toyota Research Institute, and the Jump ARCHES endowment through the Health Care Engineering Systems Center at Illinois and the OSF Foundation. This work used computational resources on NCSA Delta and PTI Jetstream2 through allocations CIS220014, CIS230012, CIS230013, and CIS240311 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, and on TACC Frontera through the National Artificial Intelligence Research Resource (NAIRR) Pilot.

References

[1] Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 3DV (2019) 3

[2] Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action composition for 3d humans. In: 3DV (2022) 3

[3] Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: SINC: Spatial composition of 3d human motions for simultaneous action generation. In: ICCV (2023) 3

[4] Bae, J., Won, J., Lim, D., Min, C.H., Kim, Y.M.: Pmp: Learning to physically interact with environments using part-wise motion priors. In: SIGGRAPH (2023) 3

[5] Barquero, G., Escalera, S., Palmero, C.: BeLFusion: Latent diffusion for behavior-driven human motion prediction. In: ICCV (2023) 3

[6] Barquero, G., Escalera, S., Palmero, C.: Seamless human motion composition with blended positional encodings. In: CVPR (2024) 3

[7] Bhatnagar, B.L., Xie, X., Petrov, I., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: BEHAVE: Dataset and method for tracking human object interactions. In: CVPR (2022) 1, 2, 3, 5, 6, 7, 8, 22, 25

[8] Braun, J., Christen, S., Kocabas, M., Aksan, E., Hilliges, O.: Physically plausible full-body hand-object interaction synthesis. arXiv preprint arXiv:2309.07907 (2023) 3

[9] Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023) 3

[10] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020) 2, 4

[11] Cao, J., Liu, J., Kitani, K., Zhou, Y.: Multi-modal diffusion for hand-object grasp generation. arXiv preprint arXiv:2409.04560 (2024) 3

[12] Cao, Z., Gao, H., Mangalam, K., Cai, Q.Z., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: ECCV (2020) 3

[13] Casas, D., Comino-Trinidad, M.: SMPLitex: A generative model and dataset for 3d human texture estimation from single image. In: BMVC (2023) 7

[14] Cen, Z., Pi, H., Peng, S., Shen, Z., Yang, M., Zhu, S., Bao, H., Zhou, X.: Generating human motion in 3d scenes from text descriptions. In: CVPR (2024) 3

[15] Cha, J., Kim, J., Yoon, J.S., Baek, S.: Text2HOI: Text-guided 3d motion generation for hand-object interaction. In: CVPR (2024) 3

[16] Chao, Y.W., Yang, J., Chen, W., Deng, J.: Learning to sit: Synthesizing human-chair interactions via hierarchical control. In: AAAI (2021) 3

[17] Chen, L.H., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: HumanMAC: Masked motion completion for human motion prediction. In: ICCV (2023) 3

[18] Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR (2023) 3

[19] Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. In: RSS (2023) 5

[20] Christen, S., Hampali, S., Sener, F., Remelli, E., Hodan, T., Sauser, E., Ma, S., Tekin, B.: Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. arXiv preprint arXiv:2403.17827 (2024) 3

[21] Cong, P., Dou, Z.W., Ren, Y., Yin, W., Cheng, K., Sun, Y., Long, X., Zhu, X., Ma, Y.: LaserHuman: Language-guided scene-aware human motion generation in free environment. arXiv preprint arXiv:2403.13307 (2024) 3

[22] Corona, E., Pumarola, A., Alenya, G., Moreno-Noguer, F.: Context-aware human motion prediction. In: CVPR (2020) 3

[23] Cui, J., Liu, T., Liu, N., Yang, Y., Zhu, Y., Huang, S.: AnySkill: Learning open-vocabulary physical skill for interactive agents. In: CVPR (2024) 3

[24] Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: MoFusion: A framework for denoising-diffusion-based motion synthesis. In: CVPR. pp. 9760–9770 (2023) 3

[25] Dai, S., Li, W., Sun, H., Huang, H., Ma, C., Huang, H., Xu, K., Hu, R.: InterFusion: Text-driven generation of 3d human-object interaction. In: ECCV (2024) 3

[26] Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: Motionlcm: Real-time controllable motion generation via latent consistency model. arXiv preprint arXiv:2404.19759 (2024) 3

[27] Daiya, D., Conover, D., Bera, A.: COLLAGE: Collaborative human-agent interaction generation using hierarchical latent diffusion and language models. arXiv preprint arXiv:2409.20502 (2024) 3

[28] Diller, C., Dai, A.: CG-HOI: Contact-guided 3d human-object interaction generation. In: CVPR (2024) 1, 2, 3, 9, 21

[29] Diller, C., Funkhouser, T., Dai, A.: FutureHuman3D: Forecasting complex long-term 3d human behavior from video observations. In: CVPR (2024) 3

[30] Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: CVPR (2023) 1, 3

[31] Feng, H., Ma, W., Gao, Q., Zheng, X., Xue, N., Xu, H.: Stratified avatar generation from sparse observations. In: CVPR (2024) 3

[32] Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: IMoS: Intent-driven full-body motion synthesis for human-object interactions. arXiv preprint arXiv:2212.07555 (2022) 3

[33] Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: ReMoS: Reactive 3d motion synthesis for two-person interactions. arXiv preprint arXiv:2311.17057 (2023) 3

[34] Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: CVPR (2022) 1, 3, 4, 8, 9, 10, 25

[35] Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: ECCV (2022) 3

[36] Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: ACMMM (2020) 3

[37] Han, S., Joo, H.: CHORUS: Learning canonicalized 3d human-object spatial relations from unbounded synthesized images. In: ICCV (2023) 22

[38] Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.J.: Stochastic scene-aware motion prediction. In: ICCV (2021) 3

[39] Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3d scenes by learning human-scene interaction. In: CVPR (2021) 3

[40] Hassan, M., Guo, Y., Wang, T., Black, M., Fidler, S., Peng, X.B.: Synthesizing physical character-scene interactions. In: SIGGRAPH (2023) 3

[41] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020) 1

[42] Hou, Z., Yu, B., Tao, D.: Compositional 3d human-object neural animation. arXiv preprint arXiv:2304.14070 (2023) 3

[43] Huang, B., Li, C., Xu, C., Pan, L., Wang, Y., Lee, G.H.: Closely interactive human reconstruction with proxemics and physics-guided adaption. In: CVPR (2024) 3

[44] Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: CVPR (2023) 1, 3

[45] Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: InterCap: Joint markerless 3D tracking of humans and objects in interaction. In: GCPR (2022) 1, 3

[46] Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: MotionGPT: Human motion as a foreign language. In: NeurIPS (2023) 3, 5, 6, 7, 8, 9, 26

[47] Jiang, N., Liu, T., Cao, Z., Cui, J., Chen, Y., Wang, H., Zhu, Y., Huang, S.: CHAIRS: Towards full-body articulated human-object interaction. In: ICCV (2023) 1, 2, 3, 5, 7, 8, 22

[48] Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: CVPR (2024) 3

[49] Jin, P., Li, H., Cheng, Z., Li, K., Yu, R., Liu, C., Ji, X., Yuan, L., Chen, J.: Local action-guided motion diffusion model for text-to-motion generation. arXiv preprint arXiv:2407.10528 (2024) 3

[50] Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: GMD: Controllable human motion synthesis via guided diffusion models. In: ICCV (2023) 3

[51] Kaufmann, M., Aksan, E., Song, J., Pece, F., Ziegler, R., Hilliges, O.: Convolutional autoencoders for human motion infilling. In: 3DV (2020) 3

[52] Kim, H., Han, S., Kwon, P., Joo, H.: Zero-shot learning for the primitives of 3d affordance in general objects. arXiv preprint arXiv:2401.12978 (2024) 3

[53] Kim, J., Kim, J., Na, J., Joo, H.: ParaHome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. arXiv preprint arXiv:2401.10232 (2024) 1, 3

[54] Kim, J., Kim, J., Choi, S.: Flame: Free-form language-based motion synthesis & editing. In: AAAI (2023) 3

[55] Kim, S.W., Zhou, Y., Philion, J., Torralba, A., Fidler, S.: Learning to simulate dynamic environments with gamegan. In: CVPR (2020) 5

[56] Kim, T., Saito, S., Joo, H.: NCHO: Unsupervised learning for neural 3d composition of humans and objects. In: ICCV (2023) 3

[57] Kong, H., Gong, K., Lian, D., Mi, M.B., Wang, X.: Priority-centric human motion generation in discrete latent space. In: ICCV (2023) 3

[58] Krebs, F., Meixner, A., Patzer, I., Asfour, T.: The kit bimanual manipulation dataset. In: Humanoids (2021) 3

[59] Kulkarni, N., Rempe, D., Genova, K., Kundu, A., Johnson, J., Fouhey, D., Guibas, L.: NIFTY: Neural object interaction fields for guided human motion synthesis. arXiv preprint arXiv:2307.07511 (2023)

[60] Lee, J., Joo, H.: Locomotion-Action-Manipulation: Synthesizing human-scene interactions in complex 3d environments. In: ICCV (2023) 3

[61] Lee, T., Moon, G., Lee, K.M.: Multiact: Long-term 3d human motion generation from multiple action labels. In: AAAI (2023) 3

[62] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: NeurIPS (2020) 2, 3, 4

[63] Li, B., Ho, E.S., Shum, H.P., Wang, H.: Two-person interaction augmentation with skeleton priors. In: CVPR (2024) 3

[64] Li, C., Chibane, J., He, Y., Pearl, N., Geiger, A., Pons-Moll, G.: Unimotion: Unifying 3d human motion synthesis and understanding. arXiv preprint arXiv:2409.15904 (2024) 3

[65] Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis. arXiv preprint arXiv:2312.03913 (2023) 1, 2, 3

[66] Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023) 1, 2, 3, 5, 7, 8, 22, 25, 26

[67] Li, L., Dai, A.: GenZI: Zero-shot 3d human-scene interaction generation. In: CVPR (2024) 3

[68] Li, Q., Wang, J., Loy, C.C., Dai, B.: Task-oriented human-object interactions generation with implicit neural representations. arXiv preprint arXiv:2303.13129 (2023) 3

[69] Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: InterGen: Diffusion-based multi-human motion generation under complex interactions. arXiv preprint arXiv:2304.05684 (2023) 1

[70] Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-X: A large-scale 3d expressive whole-body human motion dataset. In: NeurIPS (2023) 3

[71] Liu, H., Zhan, X., Huang, S., Mu, T.J., Shan, Y.: Programmable motion generation for open-set motion control tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1399–1408 (2024) 3

[72] Liu, L., Hodgins, J.: Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. ACM Transactions on Graphics (TOG) **37**(4), 1–14 (2018) 3

[73] Liu, S., Zhou, Y., Yang, J., Gupta, S., Wang, S.: Contactgen: Generative contact modeling for grasp generation. In: ICCV (2023) 3

[74] Liu, Y., Chen, C., Ding, C., Yi, L.: PhysReaction: Physically plausible real-time humanoid reaction synthesis via forward dynamics guided 4d imitation. arXiv preprint arXiv:2404.01081 (2024) 3

[75] Liu, Y., Chen, C., Yi, L.: Interactive humanoid: Online full-body motion reaction synthesis with social affordance canonicalization and forecasting. arXiv preprint arXiv:2312.08983 (2023) 3

[76] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM transactions on graphics (2015) 3, 4, 25, 26

[77] Lu, S., Chen, L.H., Zeng, A., Lin, J., Zhang, R., Zhang, L., Shum, H.Y.: HumanTOMATO: Text-aligned whole-body motion generation. arXiv preprint arXiv:2310.12978 (2023) 3

[78] Luo, Z., Wang, J., Liu, K., Zhang, H., Tessler, C., Wang, J., Yuan, Y., Cao, J., Lin, Z., Wang, F., et al.: SMPLOlympics: Sports environments for physically simulated humanoids. arXiv preprint arXiv:2407.00187 (2024) 3

[79] Ma, J., Chen, X., Bao, W., Xu, J., Wang, H.: MADiff: Motion-aware mamba diffusion models for hand trajectory prediction on egocentric videos. arXiv preprint arXiv:2409.02638 (2024) 3

[80] Ma, J., Xu, J., Chen, X., Wang, H.: Diff-IP2D: Diffusion-based hand-object interaction prediction on egocentric videos. arXiv preprint arXiv:2405.04370 (2024) 3

[81] Ma, S., Cao, Q., Zhang, J., Tao, D.: Contact-aware human motion generation from textual descriptions. arXiv preprint arXiv:2403.15709 (2024) 3

[82] Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 8

[83] Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: ICCV (2019) 1, 3

[84] Mandery, C., Terlemez, O., Do, M., Vahrenkamp, N., Asfour, T.: The kit whole-body human motion database. In: ICAR (2015) 3

[85] Mandery, C., Terlemez, O., Do, M., Vahrenkamp, N., Asfour, T.: Unifying representations and large-scale whole-body motion databases for studying human motion. IEEE Transactions on Robotics **32**(4), 796–809 (2016) 3

[86] Mao, W., Liu, M., Salzmann, M.: Generating smooth pose sequences for diverse human motion prediction. In: CVPR (2021) 3

[87] Merel, J., Tunyasuvunakool, S., Ahuja, A., Tassa, Y., Hasenclever, L., Pham, V., Erez, T., Wayne, G., Heess, N.: Catch & carry: reusable neural controllers for vision-guided whole-body tasks. ACM Transactions on Graphics (TOG) **39**(4), 39–1 (2020) 3

[88] OpenAI: ChatGPT. `https://chat.openai.com/` (2023) 2, 3, 4, 7, 8, 9, 10, 23, 26

[89] Pan, L., Wang, J., Huang, B., Zhang, J., Wang, H., Tang, X., Wang, Y.: Synthesizing physically plausible human motions in 3d scenes. arXiv preprint arXiv:2308.09036 (2023) 3

[90] Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: Learning continuous signed distance functions for shape representation. In: CVPR (2019) 8

[91] Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., Jiang, H.: HOI-Diff: Text-driven synthesis of 3d human-object interactions using diffusion models. arXiv preprint arXiv:2312.06553 (2023) 1, 2, 3, 9, 21

[92] Petrov, I.A., Marin, R., Chibane, J., Pons-Moll, G.: Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In: CVPR (2023) 3

[93] Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: ICCV (2021) 3

[94] Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: ECCV (2022) 3

[95] Petrovich, M., Black, M.J., Varol, G.: TMR: Text-to-motion retrieval using contrastive 3d human motion synthesis. In: ICCV (2023) 3

[96] Raab, S., Leibovitch, I., Li, P., Aberman, K., Sorkine-Hornung, O., Cohen-Or, D.: MoDi: Unconditional motion synthesis from diverse data. In: CVPR (2023) 3

[97] Raab, S., Leibovitch, I., Tevet, G., Arar, M., Bermano, A.H., Cohen-Or, D.: Single motion diffusion. arXiv preprint arXiv:2302.05905 (2023) 3

[98] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 8, 10

[99] Razali, H., Demiris, Y.: Action-conditioned generation of bimanual object manipulation sequences. In: AAAI (2023) 3

[100] Rempe, D., Luo, Z., Bin Peng, X., Yuan, Y., Kitani, K., Kreis, K., Fidler, S., Litany, O.: Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In: CVPR (2023) 3

[101] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 22

[102] Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics **36**(6) (2017) 7

[103] Seo, Y., Hafner, D., Liu, H., Liu, F., James, S., Lee, K., Abbeel, P.: Masked world models for visual control. In: CoRL (2023) 5

[104] Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023) 3

[105] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015) 1

[106] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 1

[107] Song, W., Zhang, X., Li, S., Gao, Y., Hao, A., Hou, X., Chen, C., Li, N., Qin, H.: HOIAnimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In: CVPR (2024) 1, 3

[108] Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Trans. Graph. **38**(6), 209–1 (2019)

[109] Starke, S., Zhao, Y., Komura, T., Zaman, K.: Local motion phases for learning multi-contact character movements. ACM Transactions on Graphics (TOG) **39**(4), 54–1 (2020) 3

[110] Sun, J., Chowdhary, G.: Towards globally consistent stochastic human motion prediction via motion diffusion. arXiv preprint arXiv:2305.12554 (2023) 3

[111] Taheri, O., Choutas, V., Black, M.J., Tzionas, D.: GOAL: Generating 4d whole-body motion for hand-object grasping. In: CVPR (2022) 3

[112] Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: ECCV (2020) 3

[113] Tang, J., Wang, J., Ji, K., Xu, L., Yu, J., Shi, Y.: A unified diffusion framework for scene-aware human motion estimation from sparse signals. In: CVPR (2024) 3

[114] Tendulkar, P., Surís, D., Vondrick, C.: FLEX: Full-body grasping without full-body grasps. In: CVPR (2023) 3

[115] Tessler, C., Guo, Y., Nabati, O., Chechik, G., Peng, X.B.: MaskedMimic: Unified physics-based character control through masked motion inpainting. arXiv preprint arXiv:2409.14393 (2024) 3

[116] Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: ECCV (2022) 3

[117] Tevet, G., Raab, S., Cohan, S., Reda, D., Luo, Z., Peng, X.B., Bermano, A.H., van de Panne, M.: CLoSD: Closing the loop between simulation and diffusion for multi-task character control. arXiv preprint arXiv:2410.03441 (2024) 3

[118] Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022) 5, 6, 7, 9

[119] Tian, J., Yang, L., Ji, R., Ma, Y., Xu, L., Yu, J., Shi, Y., Wang, J.: Gaze-guided hand-object interaction synthesis: Benchmark and method. arXiv preprint arXiv:2403.16169 (2024) 3

[120] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 2, 3, 4, 9, 10

[121] Turk, A.M.: Amazon mechanical turk. Retrieved August **17**, 2012 (2012) 23

[122] Wan, W., Dou, Z., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Tlcontrol: Trajectory and language control for human motion synthesis. arXiv preprint arXiv:2311.17135 (2023) 3

[123] Wan, W., Yang, L., Liu, L., Zhang, Z., Jia, R., Choi, Y.K., Pan, J., Theobalt, C., Komura, T., Wang, W.: Learn to predict how humans manipulate large-sized objects from interactive motions. IEEE Robotics and Automation Letters (2022) 3

[124] Wang, J., Hodgins, J., Won, J.: Strategy and skill learning for physics-based table tennis animation. In: SIGGRAPH (2024) 3

[125] Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3d human motion and interaction in 3d scenes. In: CVPR (2021) 3

[126] Wang, J., Rong, Y., Liu, J., Yan, S., Lin, D., Dai, B.: Towards diverse and natural scene-aware 3d human motion synthesis. In: CVPR (2022)

[127] Wang, J., Yan, S., Dai, B., Lin, D.: Scene-aware generative network for human motion synthesis. In: CVPR (2021) 3

[128] Wang, X., Li, G., Kuo, Y.L., Kocabas, M., Aksan, E., Hilliges, O.: Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In: 3DV (2022) 3

[129] Wang, Y., Lin, J., Zeng, A., Luo, Z., Zhang, J., Zhang, L.: PhysHOI: Physics-based imitation of dynamic human-object interaction. arXiv preprint arXiv:2312.04393 (2023) 3

[130] Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., Huang, S.: HUMANISE: Language-conditioned human motion generation in 3d scenes. In: NeurIPS (2022) 1, 3

[131] Wang, Z., Li, D., Jiang, R.: Diffusion models in 3d vision: A survey. arXiv preprint arXiv:2410.04738 (2024) 1

[132] Wang, Z., Wang, J., Lin, D., Dai, B.: InterControl: Generate human motion interactions by controlling every joint. arXiv preprint arXiv:2311.15864 (2023) 3

[133] Wei, D., Sun, X., Sun, H., Li, B., Hu, S., Li, W., Lu, J.: Understanding text-driven motion synthesis with keyframe collaboration via diffusion models. arXiv preprint arXiv:2305.13773 (2023) 3

[134] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS (2022) 2

[135] Wu, P., Escontrela, A., Hafner, D., Abbeel, P., Goldberg, K.: Daydreamer: World models for physical robot learning. In: CoRL (2023) 5

[136] Wu, Q., Shi, Y., Huang, X., Yu, J., Xu, L., Wang, J.: THOR: Text to human-object interaction diffusion via relation intervention. arXiv preprint arXiv:2403.11208 (2024) 1, 3

[137] Wu, Q., Dou, Z., Xu, S., Shimada, S., Wang, C., Yu, Z., Liu, Y., Lin, C., Cao, Z., Komura, T., et al.: DICE: End-to-end deformation capture of hand-face interactions from a single image. arXiv preprint arXiv:2406.17988 (2024) 3

[138] Wu, S., Liu, Y., Li, L., Bi, M., Zeng, W., Yang, X.: HIMO: A new benchmark for full-body human interacting with multiple objects. In: ECCV (2024) 3

[139] Wu, Y., Wang, J., Zhang, Y., Zhang, S., Hilliges, O., Yu, F., Tang, S.: SAGA: Stochastic whole-body grasping with contact. In: ECCV (2022) 3

[140] Wu, Z., Li, J., Liu, C.K.: Human-object interaction from human-level instructions. arXiv preprint arXiv:2406.17840 (2024) 1, 3

[141] Xiao, Z., Wang, T., Wang, J., Cao, J., Zhang, W., Dai, B., Lin, D., Pang, J.: Unified human-scene interaction via prompted chain-of-contacts. arXiv preprint arXiv:2309.07918 (2023) 3

[142] Xiao, Z., Wang, T., Wang, J., Cao, J., Zhang, W., Dai, B., Lin, D., Pang, J.: Unified human-scene interaction via prompted chain-of-contacts. In: ICLR (2024) 3

[143] Xie, X., Bhatnagar, B.L., Pons-Moll, G.: Chore: Contact, human and object reconstruction from a single rgb image. In: ECCV (2022) 3

[144] Xie, X., Lenssen, J.E., Pons-Moll, G.: InterTrack: Tracking human object interaction without object templates. arXiv preprint arXiv:2408.13953 (2024) 3

[145] Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: OmniControl: Control any joint at any time for human motion generation. arXiv preprint arXiv:2310.08580 (2023) 3

[146] Xie, Z., Starke, S., Ling, H.Y., van de Panne, M.: Learning soccer juggling skills with layer-wise mixture-of-experts. In: SIGGRAPH (2022) 3

[147] Xie, Z., Tseng, J., Starke, S., van de Panne, M., Liu, C.K.: Hierarchical planning and control for box loco-manipulation. arXiv preprint arXiv:2306.09532 (2023) 3

[148] Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In: ICCV (2023) 3, 6, 7, 8, 9, 10, 25

[149] Xu, S., Wang, Y.X., Gui, L.Y.: Diverse human motion prediction guided by multi-level spatial-temporal anchors. In: ECCV (2022) 3

[150] Xu, S., Wang, Y.X., Gui, L.: Stochastic multi-person 3d motion forecasting. In: ICLR (2023) 3

[151] Xu, X., Joo, H., Mori, G., Savva, M.: D3D-HOI: Dynamic 3d human-object interactions from videos. arXiv preprint arXiv:2108.08420 (2021) 3

[152] Xu, Z., Chen, Q., Peng, Y., Liu, Y.: Semantic-aware human object interaction image generation. In: ICML (2024) 3

[153] Xue, K., Seo, H.: Shape conditioned human motion generation with diffusion model. arXiv preprint arXiv:2405.06778 (2024) 3

[154] Yang, C., Kang, C., Kong, K., Oh, H., Kang, S.J.: Person in Place: Generating associative skeleton-guidance maps for human-object interaction image editing. In: CVPR (2024) 3

[155] Yang, J., Niu, X., Jiang, N., Zhang, R., Huang, S.: F-HOI: Toward fine-grained semantic-aligned 3d human-object interactions. arXiv preprint arXiv:2407.12435 (2024) 2, 3

[156] Yang, Y., Zhai, W., Luo, H., Cao, Y., Zha, Z.J.: LEMON: Learning 3d human-object interaction relation from 2d images. In: CVPR (2024) 3, 22

[157] Yang, Y., Zhai, W., Wang, C., Yu, C., Cao, Y., Zha, Z.J.: EgoChoir: Capturing 3d human-object interaction regions from egocentric views. arXiv preprint arXiv:2405.13659 (2024) 3

[158] Yang, Z., Yin, K., Liu, L.: Learning to use chopsticks in diverse gripping styles. ACM Transactions on Graphics (TOG) **41**(4), 1–17 (2022) 3

[159] Yao, H., Song, Z., Zhou, Y., Ao, T., Chen, B., Liu, L.: MoConVQ: Unified physics-based motion control via scalable discrete representations. arXiv preprint arXiv:2310.10198 (2023) 3

[160] Yazdian, P.J., Liu, E., Cheng, L., Lim, A.: MotionScript: Natural language descriptions for expressive 3d human motions. arXiv preprint arXiv:2312.12634 (2023) 3

[161] Ye, Y., Li, X., Gupta, A., De Mello, S., Birchfield, S., Song, J., Tulsiani, S., Liu, S.: Affordance diffusion: Synthesizing hand-object interactions. In: CVPR (2023) 3

[162] Yuan, W., Shen, W., He, Y., Dong, Y., Gu, X., Dong, Z., Bo, L., Huang, Q.: Mogents: Motion generation based on spatial-temporal joint modeling. In: ECCV (2024) 3

[163] Yuan, Y., Kitani, K.: DLow: Diversifying latent flows for diverse human motion prediction. In: ECCV (2020) 3

[164] Zhang, C., Liu, Y., Xing, R., Tang, B., Yi, L.: Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. arXiv preprint arXiv:2406.19353 (2024) 3

[165] Zhang, H., Christen, S., Fan, Z., Zheng, L., Hwangbo, J., Song, J., Hilliges, O.: ArtiGrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. arXiv preprint arXiv:2309.03891 (2023) 3

[166] Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3d human-object spatial arrangements from a single image in the wild. In: ECCV (2020) 3

[167] Zhang, J., Zhang, Y., An, L., Li, M., Zhang, H., Hu, Z., Liu, Y.: ManiDext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. arXiv preprint arXiv:2409.09300 (2024) 3

[168] Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: CVPR (2023) 3

[169] Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., Yu, J., Xu, L., Wang, J.: NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In: CVPR (2023) 1, 3

[170] Zhang, J., Zhang, J., Song, Z., Shi, Z., Zhao, C., Shi, Y., Yu, J., Xu, L., Wang, J.: Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In: CVPR (2024) 1, 3

[171] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023) 22

[172] Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: MotionDiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022) 3, 5, 6, 7, 9

[173] Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: ReMoDiffuse: Retrieval-augmented motion diffusion model. In: ICCV (2023) 5, 6, 7, 9

[174] Zhang, M., Li, H., Cai, Z., Ren, J., Yang, L., Liu, Z.: Finemogen: Fine-grained spatio-temporal motion generation and editing. In: NeurIPS (2024) 3

[175] Zhang, W., Dabral, R., Leimkühler, T., Golyanik, V., Habermann, M., Theobalt, C.: ROAM: Robust and object-aware motion generation using neural pose descriptors. arXiv preprint arXiv:2308.12969 (2023) 3

[176] Zhang, X., Bhatnagar, B.L., Starke, S., Guzov, V., Pons-Moll, G.: COUCH: Towards controllable human-chair interactions. In: ECCV (2022) 3

[177] Zhang, X., Bhatnagar, B.L., Starke, S., Petrov, I., Guzov, V., Dhamo, H., Pérez-Pellitero, E., Pons-Moll, G.: FORCE: Dataset and method for intuitive physics guided human-object interaction. arXiv preprint arXiv:2403.11237 (2024) 1

[178] Zhang, Y., Huang, D., Liu, B., Tang, S., Lu, Y., Chen, L., Bai, L., Chu, Q., Yu, N., Ouyang, W.: Motiongpt: Finetuned llms are general-purpose motion generators. arXiv preprint arXiv:2306.10900 (2023) 3

[179] Zhang, Z., Liu, R., Aberman, K., Hanocka, R.: TEDi: Temporally-entangled diffusion for long-term motion synthesis. arXiv preprint arXiv:2307.15042 (2023) 3

[180] Zhao, C., Zhang, J., Du, J., Shan, Z., Wang, J., Yu, J., Wang, J., Xu, L.: I'M HOI: Inertia-aware monocular capture of 3d human-object interactions. In: CVPR (2024) 1, 3

[181] Zhao, K., Li, G., Tang, S.: DART: A diffusion-based autoregressive motion model for real-time text-driven motion control. arXiv preprint arXiv:2410.05260 (2024) 3

[182] Zhao, K., Wang, S., Zhang, Y., Beeler, T., Tang, S.: Compositional human-scene interaction synthesis with semantic control. In: ECCV (2022) 3

[183] Zhao, K., Zhang, Y., Wang, S., Beeler, T., Tang, S.: Synthesizing diverse human motions in 3d indoor scenes. In: ICCV (2023) 3

[184] Zheng, J., Zheng, Q., Fang, L., Liu, Y., Yi, L.: CAMS: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In: CVPR (2023) 3

[185] Zhong, L., Xie, Y., Jampani, V., Sun, D., Jiang, H.: SMooDi: Stylized motion diffusion model. arXiv preprint arXiv:2407.12783 (2024) 3

[186] Zhou, K., Bhatnagar, B.L., Lenssen, J.E., Pons-Moll, G.: Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In: ECCV (2022) 3

[187] Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., Liu, Y., Komura, T., Wang, W., Liu, L.: EMDM: Efficient motion diffusion model for fast, high-quality motion generation. arXiv preprint arXiv:2312.02256 (2023) 3

In this Appendix, we include additional method details and experimental results: (**i**) We provide demo videos in the website, explained in Sec. A. (**ii**) We present additional details of interaction retrieval, world model, and optimization in Sec. B. (**iii**) We provide implementation details and additional information on the experimental setup in Sec. C. (**iv**) We provide additional qualitative experiments in Sec. D. (**v**) We provide some failure cases in Sec. E. (**vi**) We discuss the potential negative societal impact in Sec. F.

A Visualization Video

Beyond the qualitative results presented in the main paper, we include two demo videos that offer more detailed visualizations of the task, further illustrating the efficacy of our approach. These demos highlight (**i**) We conduct a qualitative comparison of our approach with existing text-to-HOI work [28, 91] within the framework of supervised learning. Note that as our setting contains no text supervision, it is unfair to compare our work with these approaches; we include the comparison here for additional reference. We evaluate our method by directly testing our trained model on the annotated data available from their websites, specifically retrieving their generated videos for direct comparison. *Remarkably, even without training on these datasets, our method generates results that demonstrate high-quality interactions.* It is even capable of synthesizing complex interactions involving *dynamically-changing* contact, such as the handover and throwing of objects.

B Additional Details of Methodology

B.1 Low-Level Control

In this section, we provide additional details on the retrieval based on handcraft rules, which is straightforward and does not require training. We also investigate a learning-based method without relying on handcrafted designs.

Handcraft Interaction Retrieval. In Sec. 3.2 of the main paper, we detail the construction of the interaction database and emphasize the use of body parts and object categories as keys to fetch semantically-aligned contact maps. Same as the main paper, we define a contact map as a list of K index pairs of vertices $\{(d_h^i, d_o^i)\}_{i=1}^K$. This section delves into the methodology for outlining an optimization process to generate the object initial pose $\boldsymbol{s}_1$ given contact maps and the initial human pose $\boldsymbol{a}_1$, and choose one pose based on a predefined metric.

Let $\boldsymbol{v}_{\boldsymbol{h}_1}[d_o]$ denote the vertex on the surface of the object, and $\boldsymbol{v}_{\boldsymbol{o}_1}[d_h]$ represent the corresponding vertex on the human body surface, where d_o and d_h are the indices of vertices. Specifically, to optimize $\boldsymbol{s}_1$, the overall optimization objective is given by,

$$E_{\text{opt}} = \lambda_{\text{fit}} E_{\text{fit}} + \lambda_{\text{cont}} E_{\text{cont}} + \lambda_{\text{pene}} E_{\text{pene}}, \tag{1}$$

where λ_{fit}, λ_{cont}, and λ_{pene} are hyperparameters.

Fitting Loss. To project a contact map to an object pose, we minimize the L2 distance between the human vertices and the object vertices indicated by the contact map,

$$E_{\text{fit}} = \|\boldsymbol{v}_{\boldsymbol{o}_1}[d_o] - \boldsymbol{v}_{\boldsymbol{h}_1}[d_h]\|_2. \tag{2}$$

Contact Loss. We leverage a contact loss to encourage the body part to contact the object surface in addition to the fitting loss,

$$E_{\text{cont}} = \sum_{\tilde{d}_h \in \mathcal{T}} \min_{\tilde{d}_o} \|\boldsymbol{v}_{\boldsymbol{o}_1}[\tilde{d}_o] - \boldsymbol{v}_{\boldsymbol{h}_1}[\tilde{d}_h]\|_2, \tag{3}$$

where $\mathcal{T} = \{\tilde{d}_h | \min_{\tilde{d}_o} \|\boldsymbol{v}_{\boldsymbol{o}_1}[\tilde{d}_o] - \boldsymbol{v}_{\boldsymbol{h}_1}[\tilde{d}_h]\|_2 \leq \epsilon\}$ includes the index of the body part that is close to the object vertex $\boldsymbol{v}_{\boldsymbol{o}_1}[\tilde{d}_o]$, where ϵ is a hyperparameter, $\tilde{d}_h$ and $\tilde{d}_o$ are vertex indices for human and object, respectively, in addition to the contact map $\{(d_h^i, d_o^i)\}_{i=1}^K$.

Penetration Loss. Given the signed-distance field of the human pose $\mathbf{sdf}_{\boldsymbol{h}_1}$, we employ a penetration loss to penalize the body-object interpenetration,

$$E_{\text{pene}} = -\sum_{d_o} \min(\mathbf{sdf}_{\boldsymbol{h}_1}(\boldsymbol{v}_{\boldsymbol{o}_1}[d_o]), 0). \tag{4}$$

The metric for determining the final pose selection is given by the expression $\mathbb{1}(E_{\text{pene}} = 0)/E_{\text{cont}}$. We sample a pose from the set generated by all contact maps, with higher metrics corresponding to higher selection probability.

Learning-Based Interaction Retrieval. Our interaction retrieval can also be achieved by integrating knowledge from several learning-based algorithms. Although being more complicated, the retrieval can be done without handcraft rules. Our framework can be divided into following. (**i**) Given the text prompt $\boldsymbol{t}$ and the initial human pose $\boldsymbol{a}_1$, we synthesize corresponding images via Stable Diffusion [101]. (**ii**) We follow [37] filter out images with low quality in interaction. (**iii**) An off-the-shelf model LEMON [156] is to employed to obtain object affordance and human contact, given the generated image paired with human pose $\boldsymbol{a}_1$ and object template. The output $\{(l_h^i)_{i=1}^K, (l_o^i)_{i=1}^K\}$ indicates the contact vertex indexes of human and object respectively, and the output $\boldsymbol{T}_1$ indicates the estimated object translation, which is used for initialization in the optimization. (**iv**) To acquire the object pose, we utilize the optimization to minimize the Chamfer distance between the human vertices and the object vertices, indicated by the contact vertices obtained in the last step.

$$E_{\text{fit}} = \sum_j \min_k \|\boldsymbol{v}_{\boldsymbol{o}_1}[l_o^k] - \boldsymbol{v}_{\boldsymbol{h}_1}[l_h^j]\|_2. \tag{5}$$

B.2 World Model

World Model for Initial States. In the particular instance where the timestep $t = 1$, the state vector $\boldsymbol{s}_1$ encapsulates a single frame. Consequently, we employ two distinct models for dynamics prediction. For predictions originating from the initial state, the history motion encompasses a single timestep $H = 1$. In contrast, for predictions for subsequent states, the historical interval covering m timesteps, where m denotes the frame count per segment.

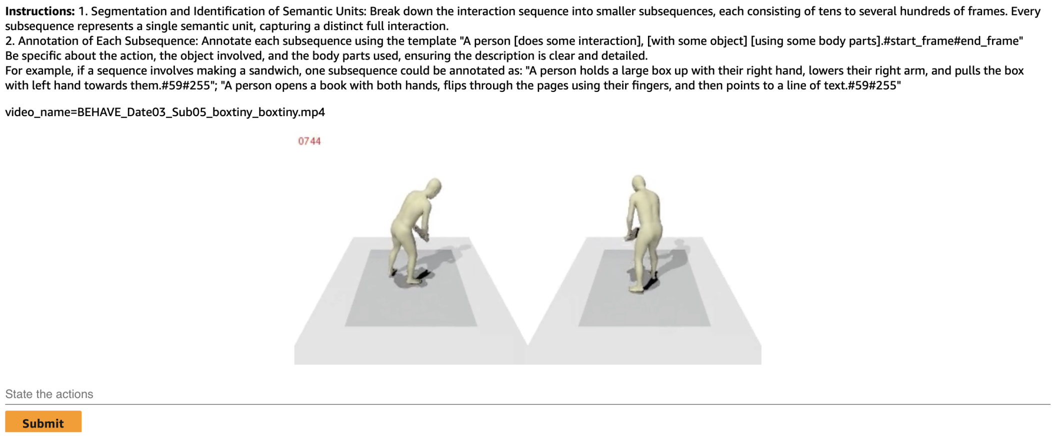
World Model for Implicit Geometry Encoding. The input to the world model includes the trajectories of the human vertices (represented by red small spheres in the top-right of Figure 2 of the main paper), along with the vertex-to-object surface vectors. By adding the vertex-to-object surface vectors to human vertices, one can easily obtain the object vertices (shown as blue small spheres in the top-right of Figure 2 of the main paper). Though the network of the world model does not receive this information directly, it can learn to combine these features to derive it when needed.

World Model for Novel Objects. The world model employs "contact vertices" as an input, which includes features derived from the object distance field. These features encompass the human vertex-to-object surface distance and the human vertex velocity relative to the nearest object vertex as introduced in Sec. 3.3 of the main paper, inherently including information related to the object's shape. This encoding is consistently applied to both training objects from the BEHAVE [7] dataset and novel objects from the CHAIRS [47] and OMOMO [66] datasets.

World Model for Non-Contact Objects. The network can process inputs without contact conditions by adopting an approach similar to ControlNet [171]. The network comprises two components: $\mathcal{G}$ that operates without contact vertex conditions, applicable in scenarios where no contact occurs, and $\mathcal{F}$, akin to the control components in ControlNet, which incorporates contact vertex conditions into the object trajectory when contact is present. When there is no contact, only the unconditional network is utilized. The model is aware of past object motion and thus needs to learn how human interaction affects the object's state. This includes understanding how objects follow contact positions or normals by $\mathcal{F}$, as well as how they move without contact by $\mathcal{G}$. With the no-contact object motion data provided by BEHAVE [7], the world model (more specifically, $\mathcal{G}$) learns to infer whether the object should free-fall based on its previous velocity or remain on the ground based on its height.

B.3 Optimization

We provide detailed formulations of optimization objectives, complementing Sec. 3.4 in the main paper. For *efficiency*, we perform optimization *sparsely* only if the loss is above a threshold to improve the efficiency. Specifically, given the reference interaction sequence $\{\boldsymbol{h}_i\}_{i=1}^L$ and $\{\boldsymbol{o}_i\}_{i=1}^L$



Instructions: 1. Segmentation and Identification of Semantic Units: Break down the interaction sequence into smaller subsequences, each consisting of tens to several hundreds of frames. Every subsequence represents a single semantic unit, capturing a distinct full interaction.
2. Annotation of Each Subsequence: Annotate each subsequence using the template "A person [does some interaction], [with some object] [using some body parts].#start_frame#end_frame"
Be specific about the action, the object involved, and the body parts used, ensuring the description is clear and detailed.
For example, if a sequence involves making a sandwich, one subsequence could be annotated as: "A person holds a large box up with their right hand, lowers their right arm, and pulls the box with left hand towards them.#59#255"; "A person opens a book with both hands, flips through the pages using their fingers, and then points to a line of text.#59#255"

video_name=BEHAVE_Date03_Sub05_boxtiny_boxtiny.mp4

Figure A: We use Amazon Mechanical Turk [121] to build an annotation platform. We provide instructions to guide the annotator to split a long sequence into several short sub-sequences with their start and end frames, and then annotate each sub-sequence. We inform annotators that our collected data are used for text-motion generation when they accept the job.

of arbitrary length L, derived from previous steps, we apply gradient descent to optimize human pose sequence $\{\boldsymbol{h}_i^*\}_{i=1}^L$ and object pose sequence $\{\boldsymbol{o}_i^*\}_{i=1}^L$, using the loss function,

$$E_{\text{opt}} = \lambda_{\text{fit}} E_{\text{fit}} + \lambda_{\text{vel}} E_{\text{vel}} + \lambda_{\text{cont}} E_{\text{cont}} + \lambda_{\text{pene}} E_{\text{pene}}, \tag{6}$$

where λ_{fit}, λ_{vel}, λ_{cont}, and λ_{pene} are hyperparameters.

Fitting Loss. We minimize the L1 distance between the input and the reference,

$$E_{\text{fit}} = \sum_{i=1}^{L} (\|\boldsymbol{h}_i^* - \boldsymbol{h}_i\|_1 + \|\boldsymbol{o}_i^* - \boldsymbol{o}_i\|_1). \tag{7}$$

Velocity Loss. We leverage a velocity loss to smooth the interaction sequence,

$$E_{\text{vel}} = \sum_{i=1}^{L-1} (\|\boldsymbol{h}_{i+1}^* - \boldsymbol{h}_i^*\|_1 + \|\boldsymbol{o}_{i+1}^* - \boldsymbol{o}_i^*\|_1). \tag{8}$$

Contact loss. We leverage a contact loss to encourage the body part to contact the object surface, if they are close to each other in the initial interaction,

$$E_{\text{cont}} = \sum_{i=1}^{L} \sum_{d_h \in \mathcal{T}_i} \min_{d_o} \|\boldsymbol{v}_{\boldsymbol{o}_i^*}[d_o] - \boldsymbol{v}_{\boldsymbol{h}_i^*}[d_h]\|_2, \tag{9}$$

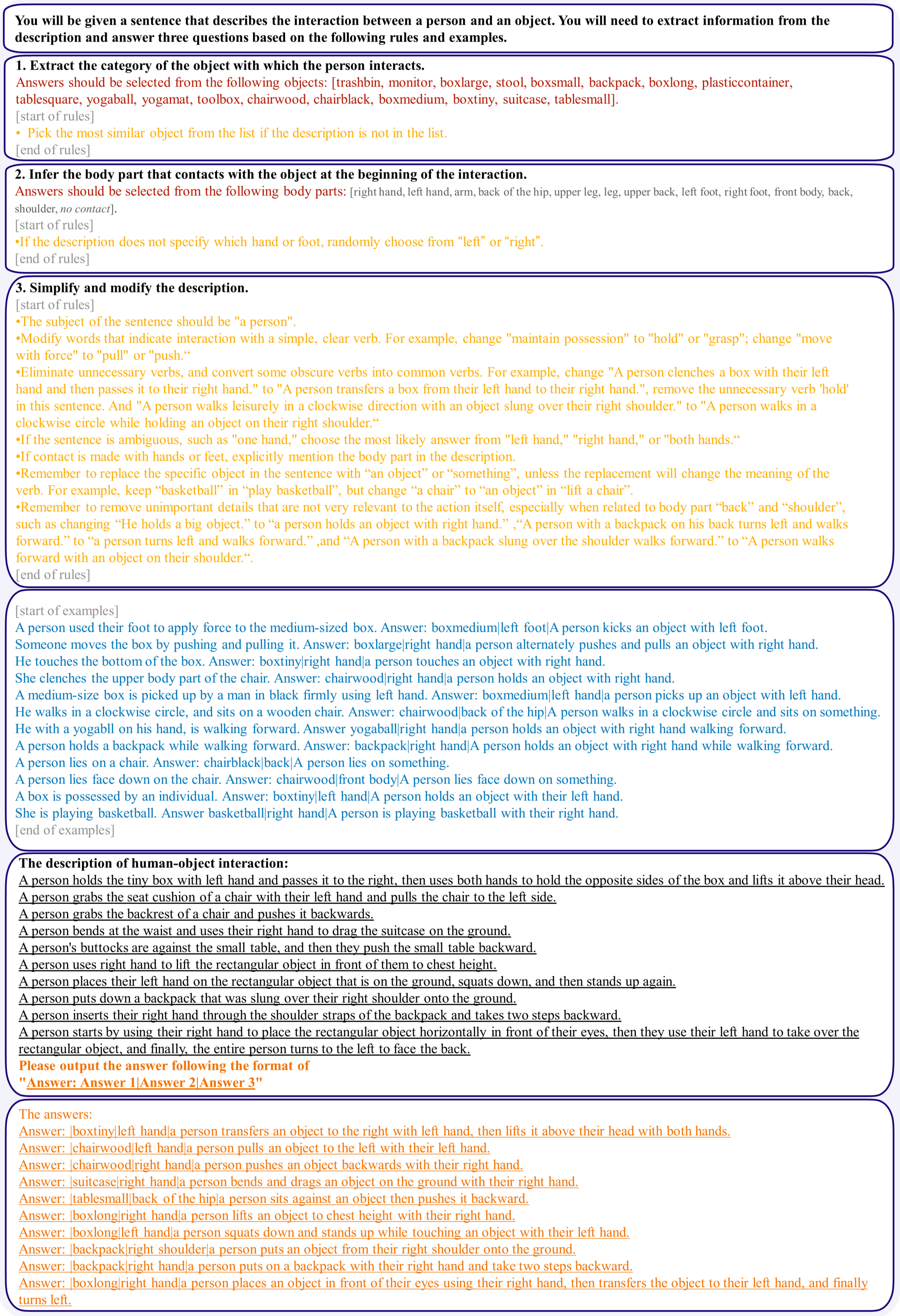
where $\boldsymbol{v}_{\boldsymbol{h}_i^*}[d_h]$ denotes the vertex on the human body surface, and $\boldsymbol{v}_{\boldsymbol{o}_i^*}[d_o]$ represents the corresponding vertex on the surface of the object. And $\mathcal{T}_i = \{d_h | \min_{d_o} \|\boldsymbol{v}_{\boldsymbol{o}_i}[d_o] - \boldsymbol{v}_{\boldsymbol{h}_i}[d_h]\|_2 \leq \epsilon\}$ includes the index of reference human vertex $\boldsymbol{v}_{\boldsymbol{h}_i}[d_h]$ that is close to the reference object vertex $\boldsymbol{v}_{\boldsymbol{o}_i}[d_o]$, where ϵ is a hyperparameter, d_h and d_o are vertex indices for human and object, respectively.

Penetration Loss. Given the signed-distance field of the human pose $\textbf{sdf}_{\boldsymbol{h}_i^*}$, we employ a penetration loss to penalize the body-object interpenetration,

$$E_{\text{pene}} = -\sum_{i=1}^{L} \sum_{d_o} \min(\textbf{sdf}_{\boldsymbol{h}_i^*}(\boldsymbol{v}_{\boldsymbol{o}_i^*}[d_o]), 0). \tag{10}$$

C Additional Details of Experimental Setup

Datasets. We include a screenshot of our annotation platform in Figure A. Our annotations are further diversified by GPT-4 [88]. The prompt used for this purpose is: `I'm going to give`



You will be given a sentence that describes the interaction between a person and an object. You will need to extract information from the description and answer three questions based on the following rules and examples.

1. Extract the category of the object with which the person interacts.

Answers should be selected from the following objects: [trashbin, monitor, boxlarge, stool, boxsmall, backpack, boxlong, plasticcontainer, tablesquare, yogaball, yogamat, toolbox, chairwood, chairblack, boxmedium, boxtiny, suitcase, tablesmall].

[start of rules]

- Pick the most similar object from the list if the description is not in the list.

[end of rules]

2. Infer the body part that contacts with the object at the beginning of the interaction.

Answers should be selected from the following body parts: [right hand, left hand, arm, back of the hip, upper leg, leg, upper back, left foot, right foot, front body, back, shoulder, *no contact*].

[start of rules]

•If the description does not specify which hand or foot, randomly choose from “left” or “right”.

[end of rules]

3. Simplify and modify the description.

[start of rules]

•The subject of the sentence should be "a person".

•Modify words that indicate interaction with a simple, clear verb. For example, change "maintain possession" to "hold" or "grasp"; change "move with force" to "pull" or "push.“

•Eliminate unnecessary verbs, and convert some obscure verbs into common verbs. For example, change "A person clenches a box with their left hand and then passes it to their right hand." to "A person transfers a box from their left hand to their right hand.", remove the unnecessary verb 'hold' in this sentence. And "A person walks leisurely in a clockwise direction with an object slung over their right shoulder." to "A person walks in a clockwise circle while holding an object on their right shoulder.“

•If the sentence is ambiguous, such as "one hand," choose the most likely answer from "left hand," "right hand," or "both hands.“

•If contact is made with hands or feet, explicitly mention the body part in the description.

•Remember to replace the specific object in the sentence with “an object” or “something”, unless the replacement will change the meaning of the verb. For example, keep “basketball” in “play basketball”, but change “a chair” to “an object” in “lift a chair”.

•Remember to remove unimportant details that are not very relevant to the action itself, especially when related to body part “back” and “shoulder”, such as changing “He holds a big object.” to “a person holds an object with right hand.” ,“A person with a backpack on his back turns left and walks forward.” to “a person turns left and walks forward.” ,and “A person with a backpack slung over the shoulder walks forward.” to “A person walks forward with an object on their shoulder.“.

[end of rules]

[start of examples]

A person used their foot to apply force to the medium-sized box. Answer: boxmedium|left foot|A person kicks an object with left foot.

Someone moves the box by pushing and pulling it. Answer: boxlarge|right hand|a person alternately pushes and pulls an object with right hand.

He touches the bottom of the box. Answer: boxtiny|right hand|a person touches an object with right hand.

She clenches the upper body part of the chair. Answer: chairwood|right hand|a person holds an object with right hand.

A medium-size box is picked up by a man in black firmly using left hand. Answer: boxmedium|left hand|a person picks up an object with left hand.

He walks in a clockwise circle, and sits on a wooden chair. Answer: chairwood|back of the hip|A person walks in a clockwise circle and sits on something.

He with a yogabll on his hand, is walking forward. Answer: yogaball|right hand|a person holds an object with right hand walking forward.

A person holds a backpack while walking forward. Answer: backpack|right hand|A person holds an object with right hand while walking forward.

A person lies on a chair. Answer: chairblack|back|A person lies on something.

A person lies face down on the chair. Answer: chairwood|front body|A person lies face down on something.

A box is possessed by an individual. Answer: boxtiny|left hand|A person holds an object with their left hand.

She is playing basketball. Answer basketball|right hand|A person is playing basketball with their right hand.

[end of examples]

The description of human-object interaction:

A person holds the tiny box with left hand and passes it to the right, then uses both hands to hold the opposite sides of the box and lifts it above their head.

A person grabs the seat cushion of a chair with their left hand and pulls the chair to the left side.

A person grabs the backrest of a chair and pushes it backwards.

A person bends at the waist and uses their right hand to drag the suitcase on the ground.

A person's buttocks are against the small table, and then they push the small table backward.

A person uses right hand to lift the rectangular object in front of them to chest height.

A person places their left hand on the rectangular object that is on the ground, squats down, and then stands up again.

A person puts down a backpack that was slung over their right shoulder onto the ground.

A person inserts their right hand through the shoulder straps of the backpack and takes two steps backward.

A person starts by using their right hand to place the rectangular object horizontally in front of their eyes, then they use their left hand to take over the rectangular object, and finally, the entire person turns to the left to face the back.

Please output the answer following the format of
"Answer: Answer 1|Answer 2|Answer 3"

The answers:

Answer: |boxtiny|left hand|a person transfers an object to the right with left hand, then lifts it above their head with both hands.

Answer: |chairwood|left hand|a person pulls an object to the left with their left hand.

Answer: |chairwood|right hand|a person pushes an object backwards with their right hand.

Answer: |suitcase|right hand|a person bends and drags an object on the ground with their right hand.

Answer: |tablesmall|back of the hip|a person sits against an object then pushes it backward.

Answer: |boxlong|right hand|a person lifts an object to chest height with their right hand.

Answer: |boxlong|left hand|a person squats down and stands up while touching an object with their left hand.

Answer: |backpack|right shoulder|a person puts an object from their right shoulder onto the ground.

Answer: |backpack|right hand|a person puts on a backpack with their right hand and take two steps backward.

Answer: |boxlong|right hand|a person places an object in front of their eyes using their right hand, then transfers the object to their left hand, and finally turns left.

Figure B: Full log of our high-level planning.

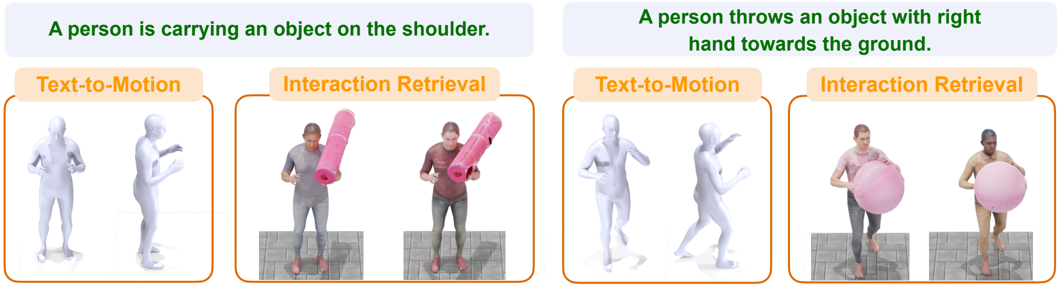

A person is carrying an object on the shoulder.
A person throws an object with right hand towards the ground.
Text-to-Motion
Interaction Retrieval
Text-to-Motion
Interaction Retrieval

Figure C: **Qualitative results** from the interaction retrieval. We demonstrate that our learning-based interaction retrieval can extract diverse and realistic interactions.

`you a description, and I would like to have three rewritten sentences with varying degrees of complexity, following the example: “...” The input text is “...” Please give me three texts that vary in complexity but keep the meaning of the sentence the same.` This results in **(i)** *less complexity*: someone holds a backpack and steps left; **(ii)** *middle complexity*: a person holds a backpack in front of them with both hands and takes a step to the left; **(iii)** *more complexity*: with both hands, a person clutches a heavy backpack firmly and brings it close to their body, then steps to the left with their left leg.

Metrics. In Sec. 4.1, we introduce the metrics employed in this paper. This section details the formula for the proposed metric CMD. The formulations for other metrics are available in the existing literature [34, 148]. CMD quantifies the discrepancy between the contact maps of ground truth interactions and those synthesized one. In this context, a contact map is characterized by the proportion of time $\{\boldsymbol{p}_i\}_{i=1}^P$ each body part maintains active contact. Here, $\boldsymbol{p}_i$ denotes the percentage of time during which the body part i is less than a threshold distance from the object. And the metric is defined as,

$$\text{CMD} = \frac{1}{P}\sum_{i=1}^{P} \|\boldsymbol{p}_i - \boldsymbol{p}_i^{\text{GT}}\|_1, \tag{11}$$

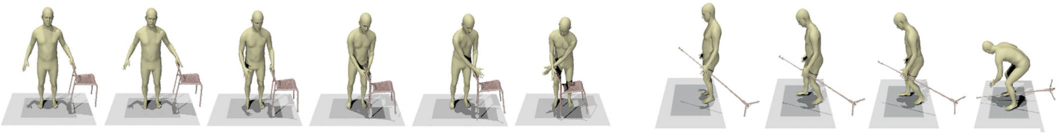
where $\boldsymbol{p}_i^{\text{GT}}$ is from the ground truth contact map, P is the number of body parts defined in SMPL [76], and we set the distance threshold as 0.03 m.

Implementation Details. The segment in the MDP contains $m = 4$ frames. The dynamics model, which includes 2 dynamics blocks as described in the main paper, is trained on the BEHAVE training set [7], with a batch size of 32, a latent dimension of 64, and for 500 epochs. For rollout after the initial step $t > 1$, our dynamics model is trained to predict over a longer timeframe ($F = 3 \times m = 12$), exceeding the past motion duration ($H = m = 4$). For the initial step $t = 1$, we train a separate dynamics model to forecast a duration of $F = 15$ given the past motion over $H = 1$ frame, consistent with Sec. B.2. The optimization process is conducted over 300 epochs, utilizing a learning rate of 0.01. The dynamic model is trained on an NVIDIA A40 GPU for a day. Our full log for high-level planning is presented in Figure B.

D Additional Qualtitative Results

Interaction Retrieval. We here visualize the intermediate retrieval results. Figure C depicts the results from learning-based retrieval, resulting in a diverse set of interactions that are both high-quality and semantically aligned.

Qualitative Experiments on OMOMO dataset. Figure D exemplifies our method that is able to generalize effectively to the OMOMO [66] dataset, despite our dynamics model not being trained on its object geometry or annotations.



Grab the top of the whitechair, swing the whitechair, and put down the whitechair.

Put the tripod horizontally down.

Figure D: **Qualitative results** on the OMOMO [66] dataset. Our method generalizes well on the OMOMO objects and annotations unseen in training. Frames are separately visualized. Here, our synergized models are GPT-4 [88] and MotionGPT [46].

E Failure Cases

We present failure cases of our method in our website, consisting of (i) inconsistency of interaction with textual description, (ii) inconsistency of human actions with textual descriptions, and (iii) wrong object category inferred by LLM.

F Potential Negative Societal Impact

Some potential negative societal impacts include: (**i**) Our approach can synthesize realistic human motion interacting with objects, which could be misused to create deceptive or harmful content, such as portraying individuals in false situations. This could contribute to the spread of misinformation. (**ii**) Our method evaluates real behavioral information, raising potential privacy concerns. Although our model utilizes a processed representation (SMPL [76]) of human motion that retains minimal identifying details – unlike raw data or images – its ability to simulate human-object interactions could still be exploited for unauthorized surveillance or behavioral analysis. For instance, with photorealistic textures, it might be used to model and generate personal habits or movements without consent, posing risks of privacy violations. However, the use of a processed representation can be positively viewed as a privacy-enhancing feature, as it minimizes the exposure of personally identifiable details.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: in Sec. 1 and 4

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: in Sec. 5

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: this paper doesn't include proof and theory.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: in Sec. 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Our code is not available at this time but will be released in the future.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: in Sec. 4

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: in Sec. 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: in Sec. C

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: in Sec. F

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: in Sec. F

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not use or release pretrained language models trained by our own

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: in Sec. 4

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: in Sec. C

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: in Sec. F

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.