HOHI4D: SCALABLE SYNTHESIS OF 3D HUMAN-OBJECT-HUMAN INTERACTION DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

3D human-object-human interaction (HOHI) modeling remains underexplored due to the lack of scalable datasets and the difficulty of acquiring accurate multi-agent interaction data. Existing approaches rely on controlled capture setups or strong assumptions on object pose and contact, limiting their applicability in complex scenarios. In this work, we propose a fully automated and controllable pipeline for large-scale HOHI data synthesis, combining prompt-driven video generation with SAM3D-based reconstruction. Our framework enables diverse interaction generation, reduces object drift via geometry-aware refinement, and removes reliance on predefined object templates. Based on this pipeline, we construct HOHI4D, a large-scale dataset with over 5,000 sequences and 13 hours of motion. Furthermore, we introduce a structured generative model for HOHI. We design HOHI-VAE to learn disentangled latent representations of individual motion, object dynamics, and global interaction, and build HOHI-Diff, a latent diffusion model operating in this structured space. Our approach enables stable and coherent generation of complex multi-agent interactions, providing a scalable benchmark for HOHI modeling.

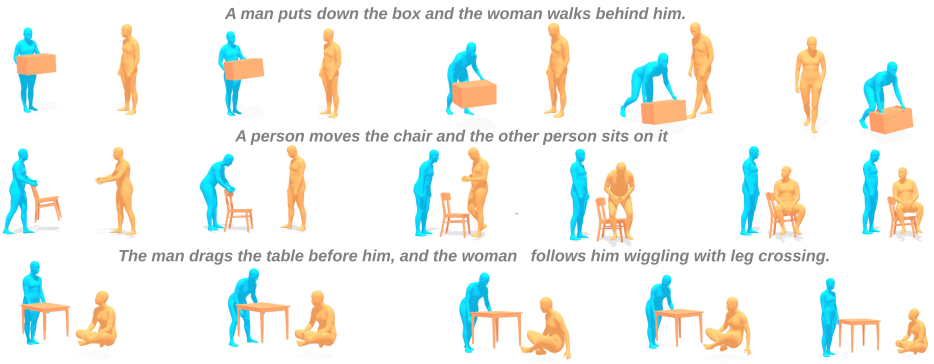


Figure 1: **HOHI4D dataset overview.** Our pipeline generates diverse, controllable and robust human-object-human interactions.

1 INTRODUCTION

Recent advances in generative modeling have significantly expanded the scope and realism of human motion synthesis. Early efforts primarily focused on text-to-motion generation Guo et al. (2022); Chen et al. (2023); Tevet et al. (2023); Zhang et al. (2023); Jiang et al. (2024a); Zhang et al. (2024); Guo et al. (2020); Yuan et al. (2023); Zhong et al. (2023); Wang et al. (2023a); Kong et al. (2023); Gao et al. (2024); Guo et al. (2024); Han et al. (2024); Li et al. (2024b), while more recent works have extended this setting to human-object interaction (HOI) Diller and Dai (2024); Peng et al. (2025); Li et al. (2024a); Xu et al. (2023; 2024b); Wang et al. (2023b); Wu et al. (2025b), human-human interaction (HHI) Liang et al. (2024); Javed et al. (2025); Wang et al. (2025); Cen et al. (2025); Ruiz-Ponce et al. (2024); Chopin et al. (2024), and human-scene interaction (HSI) Wang et al. (2022); Jiang et al. (2024b); Hassan et al. (2019; 2021); Yi et al. (2023). However, many real-world embodied scenarios involve a more general form of interaction, where multiple humans coordinate through a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

shared object. We refer to this setting as 3D Human-Object-Human Interaction (HOHI). Compared with HHI, HOHI introduces object-mediated coordination; compared with HOI, object motion is no longer induced by a single actor but jointly shaped by multiple humans. This tri-partite coupling makes HOHI fundamentally different from existing pairwise interaction settings and substantially more challenging to model.

Despite its importance for embodied intelligence and multi-agent interaction Zhang et al. (2021); Hasan et al. (2026); Wu et al. (2025a); Feng et al. (2026), HOHI remains largely underexplored. The main bottleneck lies in data acquisition. Capturing accurate 3D HOHI data requires synchronized human motion, object geometry, object pose, and their temporal coordination, which is expensive and difficult to obtain even in controlled capture environments. Existing datasets are therefore limited in scale, object diversity, and interaction coverage. More importantly, the object in HOHI often undergoes transient manipulation, exchange, occlusion, and joint control by multiple agents, making precise contact annotation and object pose tracking particularly brittle. As a result, current benchmarks are insufficient for learning generalizable HOHI representations.

A promising direction is to leverage the recent foundation-model ecosystem for scalable data construction. Modern video generative models provide strong visual and motion priors for synthesizing diverse interaction videos, while foundation-level 3D reconstruction models such as SAM3D-Body (SAM3DB) Yang et al. (2026) and SAM3D-Object (SAM3DO) Team et al. (2025) make it possible to recover human geometry and object geometry from monocular videos. This enables a new video-to-3D paradigm for constructing HOHI data at scale, avoiding the heavy dependence on predefined CAD templates, calibrated multi-view systems, or ground-truth depth required by conventional 6D pose estimation and tracking pipelines Wang et al. (2019); Wen et al. (2024).

However, directly applying this paradigm to HOHI is non-trivial. Generated videos may contain ambiguous contact, strong occlusion, and fast object manipulation, while SAM3D-Object outputs are often sparse and temporally inconsistent across frames. Naive reconstruction can therefore suffer from scale inconsistency, object drift, jittering trajectories, and unstable interaction dynamics. These issues are particularly severe in HOHI, where the shared object must remain globally consistent with the motions of both humans. Thus, scalable HOHI construction requires not only strong generative and reconstruction priors, but also controllable interaction synthesis and temporally stable object motion refinement.

To address these challenges, we propose a controllable and robust data synthesis pipeline for large-scale HOHI generation. Instead of relying on brittle explicit contact supervision, our pipeline focuses on globally consistent and temporally stable interaction reconstruction. We first use structured prompts to control interaction types, spatial layouts, and object categories, producing semantically coherent videos with diverse object-mediated interactions. We then reconstruct human motion and object geometry using SAM3D-based models, followed by geometry-aware alignment and temporal refinement to mitigate object drift and jitter. This design allows us to construct HOHI data at scale while preserving interaction consistency across humans and objects. Based on this pipeline, we construct **HOHI4D**, a large-scale HOHI dataset containing over 5,000 interaction sequences and more than 13 hours of motion data. Compared with existing HOHI datasets such as CORE4D Liu et al. (2025), HOHI4D provides substantially larger scale, richer object variation, text annotations, and more diverse interaction patterns. It therefore serves as a scalable benchmark for studying object-mediated multi-agent interaction and generative HOHI modeling.

Beyond data construction, we further introduce a structured latent generative framework for HOHI. Directly modeling HOHI sequences in the original motion space is inefficient due to their high dimensionality and strong temporal redundancy. More importantly, existing flat latent representations are insufficient for capturing tri-partite coordination among two humans and a shared object. To this end, we propose **HOHI-VAE**, which factorizes individual human motion, object dynamics, and global interaction into dedicated latent variables. Built upon this structured latent space, we further propose **HOHI-Diff**, a latent diffusion model that performs generation directly in the compact interaction space. This enables the model to focus on high-level coordination dynamics while delegating low-level motion reconstruction to the VAE decoder, resulting in coherent human-object-human generation.

In summary, our contributions are threefold:

- We propose a controllable and robust data synthesis pipeline that combines prompt-driven video generation, SAM3D-based 4D reconstruction, and temporal pose refinement, enabling scalable HOHI data generation without requiring explicit contact supervision.
- We introduce **HOHI4D**, a large-scale dataset with over 5,000 sequences and more than 13 hours of motion data, providing a new benchmark for human-object-human interaction generation.
- We present **HOHI-VAE** and **HOHI-Diff**, a structured latent diffusion framework tailored for HOHI generation, which explicitly models individual human motion, object dynamics, and global tri-partite interaction.

2 Related Work

Human Motion Generation Human motion generation has advanced rapidly with large-scale datasets and generative models. Text-driven synthesis on *HumanML3D* Guo et al. (2022) includes diffusion-based Sohl-Dickstein et al. (2015); Ho et al. (2020); Song et al. (2021) methods (MDM Tevet et al. (2023), MotionDiffuse Zhang et al. (2024), MLD Chen et al. (2023)) and language-model-based approaches (T2M-GPT Zhang et al. (2023), MotionGPT Jiang et al. (2024a)), with recent works (e.g., MoMask, SALAD) incorporating structural priors.

Beyond single-person motion, HOI generation (e.g., HOI-Diff Peng et al. (2025), CHOIS Li et al. (2024a)) leverages datasets such as BEHAVE and Full-Body-Manip Li et al. (2023), while HHI generation (e.g., InterGen Liang et al. (2024), InterMask Javed et al. (2025), TIMotion Wang et al. (2025), Ready-to-React Cen et al. (2025)) is explored on InterHuman Liang et al. (2024) and InterX Xu et al. (2024a). However, most works focus on single-modality or pairwise interactions, leaving the more complex *human-object-human* (HOHI) setting underexplored due to limited data and unified modeling.

Human & Object Interaction Reconstruction Reconstructing human-object interactions typically relies on RGB or RGB-D inputs to jointly estimate human motion and object pose. Methods such as CARI4D Xie et al. (2026) combine monocular video with estimated depth (e.g., UniDepth Piccinelli et al. (2024)) and models like FoundationPose for object pose estimation, while others (e.g., VisTracker, InterTrack) focus on tracking under predefined interaction patterns. These approaches often assume accurate depth, known camera intrinsics, or object CAD models, which limits robustness. Errors in depth, calibration, or pose estimation—especially for unseen objects—lead to instability, and reliance on predefined affordances further restricts generalization.

Video-based 3D Reconstruction with Foundation Models Foundation models significantly improve video-based 3D reconstruction. SAM3D provides a unified framework, where SAM3DB reconstructs human geometry and SAM3DO recovers object geometry and initial pose. By directly reconstructing 3D from video, SAM3D reduces dependence on depth and calibration, making it suitable for in-the-wild scenarios. However, SAM3D is not designed for interaction modeling: SAM3DO produces sparse, temporally inconsistent object meshes without motion constraints, while SAM3DB lacks contact and interaction awareness. Applying SAM3D to HOHI leads to noisy trajectories, scale inconsistency, and unstable contact, motivating a dedicated interaction-aware framework.

Human-Object-Human Interaction HOHI datasets and methods remain scarce. *CORE4D* is the only representative dataset, but is limited in scale and diversity, with synthetic augmentation mainly varying object geometry and often suffering from object drift and inaccurate contact. These limitations arise from the high cost of data acquisition, which requires multi-view setups, calibration, and tracking systems. Consequently, existing datasets lack scalability and robustness. In summary, current research faces two key gaps: (1) lack of scalable, high-quality HOHI datasets, and (2) absence of robust pipelines for jointly reconstructing human and object dynamics without strong assumptions (e.g., CAD models or precise sensors), motivating our framework.

3 Data Synthesis

We propose a scalable and controllable pipeline for synthesizing human-object-human (HOHI) interaction data from text prompts. As illustrated in Fig. 2, our pipeline consists of three stages:

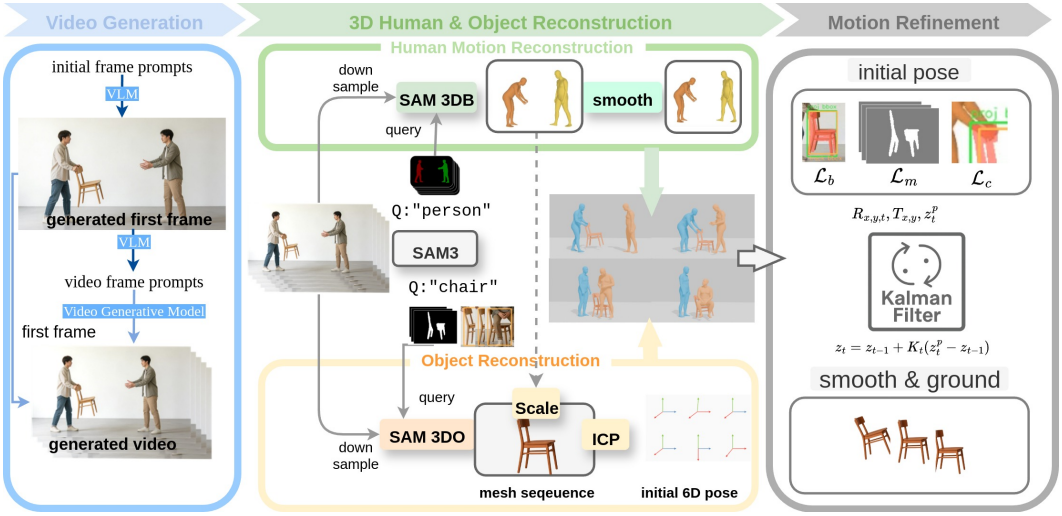


Figure 2: Diagram for data synthesis. The pipeline contains video generation, 3D human and object reconstruction, and object motion refinement

(1) controllable video generation (Sec. 3.1), (2) 4D human-object reconstruction (Sec. 3.2), and (3) object pose refinement (Sec. 3.3).

Given a set of predefined interaction prompts, we first generate diverse multi-agent interaction videos using a video generative model. We then reconstruct 3D human motion and object geometry from the generated videos using SAM3D-based models. Finally, we estimate and refine object poses over time through geometric alignment and temporal filtering, resulting in physically plausible and temporally consistent HOHI sequences.

3.1 Controllable Video Generation

The quality and diversity of synthesized interactions largely depend on the controllability of video generation. Prior datasets such as CORE4D focus on limited interaction patterns, resulting in restricted motion diversity.

To address this, we design a structured prompt pipeline to control interaction types, initial poses, and spatial configurations. We first use a vision-language model (VLM) to generate initial frame prompts that specify the relative positions and interaction intent of humans and objects. These prompts are used to synthesize initial frames via *Seedream 5.0 Lite* Seedream et al. (2025), providing spatially consistent layouts for subsequent video generation. We consider 15 object categories covering common household items, each with multiple variations in shape and scale. We further design category-specific prompts to encode functional priors (e.g., sitting for chairs, carrying for boxes), promoting semantically and physically plausible interactions.

Based on the initial frames, we derive temporally consistent prompts and feed them into *Seedance 1.5* Seedance et al. (2025) to generate full interaction sequences. This hierarchical prompting strategy ensures both semantic consistency and motion diversity.

3.2 4D Human-Object Reconstruction

Given the synthesized videos, we reconstruct human motion and object geometry over time. Human meshes are recovered using SAM3DB Yang et al. (2026) and SAM-4D-Body Gao et al. (2025), with temporal downsampling and smoothing applied to improve efficiency and stability.

For objects, we use SAM3DO to reconstruct per-frame meshes, which are often sparse and temporally inconsistent. We therefore aggregate these meshes and normalize their scale based on human reconstruction, yielding a canonical object mesh with stable geometry.

3.3 Object Pose Refinement

While SAM3DO provides coarse object geometry, it does not yield reliable temporally consistent object poses. To address this, we estimate and refine the object pose sequence in two stages.

First, we align each reconstructed object mesh to the canonical object mesh using Iterative Closest Point (ICP), obtaining an initial estimate of per-frame poses. This step converts the sparse mesh sequence into a unified pose representation. However, the resulting pose sequence is often noisy and temporally inconsistent. To enforce smoothness and physical plausibility, we introduce a Kalman filter-based refinement module. Specifically, we model the object pose as a latent state and iteratively update it based on predicted motion and observed alignment results. This filtering process effectively reduces jitter and mitigates drift caused by reconstruction errors.

In addition, we incorporate multi-objective optimization constraints, including bounding box consistency, mask overlapping, and human-object contact regularization. These constraints further improve the stability of object motion and ensure realistic interaction dynamics. Through the above refinement process, we obtain temporally coherent and physically plausible object pose sequences, which, together with reconstructed human motion, form high-quality HOHI data.

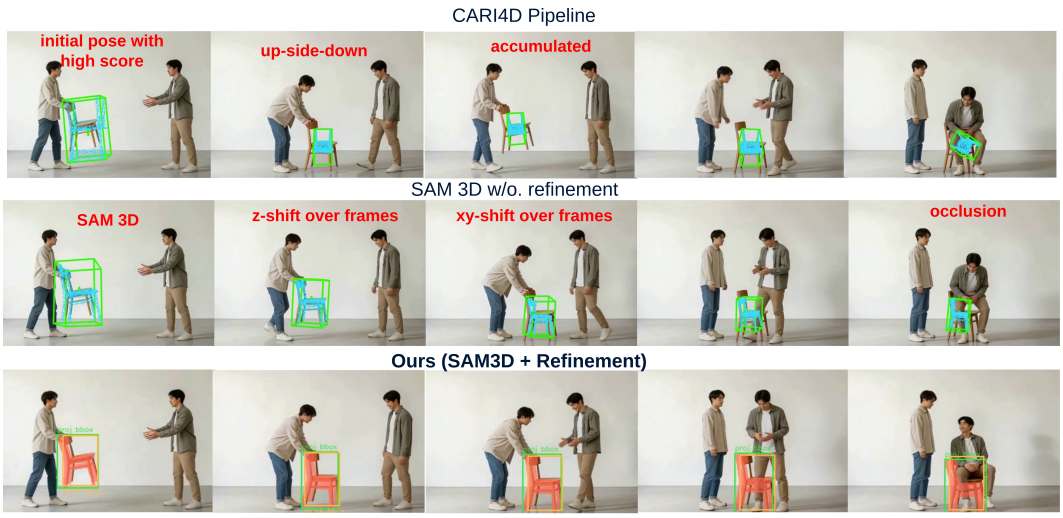


Figure 3: Comparison with the CARI4D Xie et al. (2026) pipeline and SAM3D only without refinement. The estimated posed object is projected back to the video.

An alternative data synthesis pipeline, CARI4D Xie et al. (2026), relies heavily on FoundationPose for object pose estimation. However, in the absence of accurate depth information and priors for unseen objects, the method may assign high confidence to incorrect initial pose hypotheses, leading to early-stage errors that propagate through the sequence. In contrast to both CARI4D and the direct use of SAM3DO, our approach significantly improves the stability of object pose estimation by incorporating geometry-aware refinement and temporal consistency, as illustrated in Fig. 3. Additionally, we evaluate tracking performance using bounding box mIoU and collision volume on our test dataset. The results in Tab. 2 show that our method consistently outperforms prior approaches across all metrics.

Our dataset significantly improves over CORE4D in terms of scale, diversity, and representation quality. With substantially more sequences and longer duration, HOHI4D captures richer multi-agent interaction patterns. More importantly, by leveraging MHR Ferguson et al. (2025) instead of traditional SMPL/SMPLX Loper et al. (2015); Pavlakos et al. (2019) meshes, our dataset provides a more robust and expressive motion representation that better handles complex interactions and reconstruction noise. Together, these advantages make HOHI4D a more suitable benchmark for large-scale HOHI modelling and generative learning.

Algorithm 1 Ray-based Pose Refinement with Projection-Supervised Kalman Depth Filtering

Input: camera intrinsics and image $\{K_t, \mathbf{u}_t\}_{t=1}^{T}$, sparse pose set $\{(\mathbf{R}_i^s, \mathbf{t}_i^s)\}$, canonical points $\mathcal{P}$
Output: refined pose sequence $\{(\mathbf{R}_t, \mathbf{t}_t)\}_{t=1}^{T}$

Initialize depth state $(\hat{z}_0, P_0)$
for $t = 1$ to T **do**
  Retrieve nearest sparse pose to obtain coarse rotation $\mathbf{R}_t^c$ and depth prior z_t^{prior}.
  Optimize object pose $(\mathbf{R}_t, \mathbf{t}_t)$ by minimizing a joint objective:
    $(\mathbf{R}_t, \mathbf{t}_t) = \arg\min_{\mathbf{R},\mathbf{t}} \ \mathcal{L}_b + \mathcal{L}_c + \mathcal{L}_m$.
  Compute viewing ray $\mathbf{r}_t \leftarrow \pi^{-1}(K_t, \mathbf{u}_t)$.
  Construct baseline depth estimate z_t^{base} via blending of hand-contact cues and sparse prior.
  if projection supervision is enabled **then**
    Solve $z_t^{\text{proj}} = \arg\max_{z \in \mathcal{Z}_t} S_t(z)$.
    Compute confidence weight λ_t based on projection consistency.
  Fuse depth observation:
  $z_t^{\text{obs}} = (1 - \lambda_t) z_t^{\text{base}} + \lambda_t z_t^{\text{proj}}$.
  Perform Kalman update:
  $(\hat{z}_t, P_t) = \text{KalmanUpdate}(\hat{z}_{t-1}, P_{t-1}, z_t^{\text{obs}})$.
  Recover translation $\mathbf{t}_t = \hat{z}_t \mathbf{r}_t$, and set $\mathbf{R}_t \leftarrow \mathbf{R}_t^c$.
Apply sequence-level rotation refinement on $\{\mathbf{R}_t\}$ with contact-aware smoothing.

Table 1: Comparison between CORE4D and our HOHI4D dataset. HOHI4D provides larger scale, longer duration, more object diversity, and includes text annotations with higher frame rates.

Method	Number	Duration (h)	Objects	Text	FPS
CORE4D	920	2.8	32	×	15
Ours	5000	13.2	5000	✓	30

Table 2: Quantitative comparison of object reconstruction quality.

Method	mIoU$_{bbox}$	mIoU$_{mask}$	Collision ↓
CARI4D	0.382	0.203	0.388
SAM3DO	0.479	0.157	0.645
Ours	**0.741**	**0.482**	**0.370**

4 HOHI MODELLING

4.1 PROBLEM SETUP

Given a Human-Object-Human Interaction (HOHI) sequence of length T, we denote the two human motions as $\mathbf{H}_A, \mathbf{H}_B \in \mathbb{R}^{T \times D_h}$, the object motion as $\mathbf{O} = \{(\mathbf{R}_t, \mathbf{t}_t)\}_{t=1}^{T}$, and the canonical object point cloud as $\mathbf{P} \in \mathbb{R}^{K \times 3}$. Here, $\mathbf{R}_t \in SO(3)$ and $\mathbf{t}_t \in \mathbb{R}^3$ represent the object rotation and translation at frame t.

In addition, each sequence is associated with a textual description $\mathbf{c}_{\text{text}}$ and an object representation derived from the point cloud $\mathbf{P}$, which together serve as conditioning signals for HOHI generation. Our goal is to learn a conditional generative model that produces coherent human-object-human interactions given both semantic (text) and geometric (object) inputs.

Directly modeling HOHI sequences in the original motion space is inefficient due to high dimensionality and temporal redundancy. Moreover, HOHI data involves three tightly coupled factors: individual human motion, cross-human interaction, and object dynamics, making it challenging to learn a compact representation. To address this, we first learn a structured latent space using a variational autoencoder (HOHI-VAE), which compresses a full HOHI sequence into disentangled latent variables capturing individual motion, global interaction, and object dynamics.

This structured latent space serves as the modeling space for our subsequent conditional diffusion model.

4.2 HOHI-VAE

Different from HHI or HOI settings that involve only pairwise interactions, HOHI requires modeling a global interaction pattern jointly induced by two humans and a shared object. Following prior work such as DHVAE Geng et al. (2026), we adopt a structured latent design to explicitly disentangle

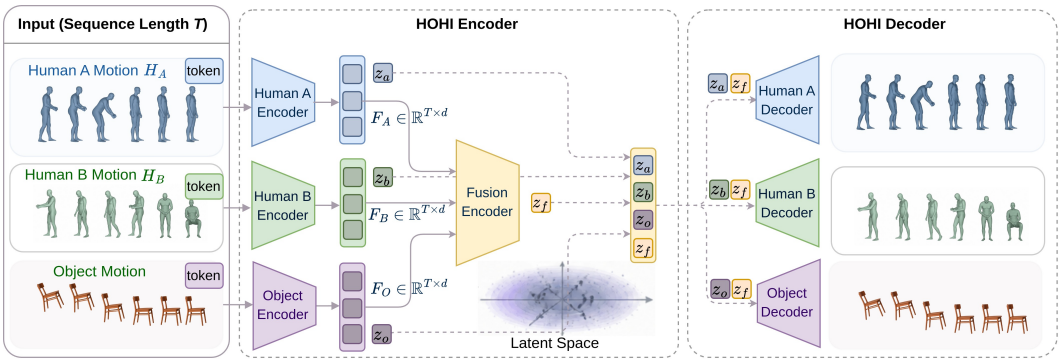


Figure 4: HOHI-VAE contains three individual encoders and one fusion encoder, capturing individual and global features.

individual motion, object dynamics, and global interaction. As shown in Fig. 4, HOHI-VAE consists of three modality-specific encoders, namely human encoders $\mathcal{E}_A$, $\mathcal{E}_B$, an object encoder $\mathcal{E}_O$, and a fusion encoder $\mathcal{E}_F$ that aggregates all streams. For the human branches, we follow MLD Chen et al. (2023) by prepending a learnable latent token to each motion sequence, whose output parameterizes the individual latents z_a and z_b, while the remaining features $\mathbf{F}_A$ and $\mathbf{F}_B$ preserve temporal information.

For the object branch, we disentangle geometry and motion. Specifically, we extract deterministic shape tokens from the canonical point cloud as a geometry prior, and encode motion-dependent features by transforming sampled points with the ground-truth trajectory, producing an object latent z_o and temporal features $\mathbf{F}_O$. The features from all streams are then fused via a Transformer-based encoder:

$$z_f = \mathcal{E}_F(\mathbf{F}_A, \mathbf{F}_B, \mathbf{F}_O). \tag{1}$$

The global latent z_f captures tri-partite coordination across both humans and the object, including role assignment and temporal synchronization. Compared to flat latent formulations, this structured decomposition introduces a stronger inductive bias for modeling coordinated multi-agent interactions.

During decoding, z_f is combined with z_a, z_b, and z_o to condition all branches, ensuring consistent reconstruction of human motions and object dynamics. Overall, HOHI-VAE compresses high-dimensional interaction sequences into a compact and disentangled latent space, facilitating stable generative modeling. The model is trained with a combination of reconstruction and regularization objectives:

$$\mathcal{L}_{\text{VAE}} = \lambda_h \mathcal{L}_{\text{human}} + \lambda_o \mathcal{L}_{\text{obj}} + \lambda_v \mathcal{L}_{\text{vel}} + \beta \mathcal{L}_{\text{KL}}.$$

The human reconstruction loss $\mathcal{L}_{\text{human}}$ is computed for both motion streams, enforcing accurate recovery of individual human motions. The object reconstruction loss $\mathcal{L}_{\text{obj}}$ is defined on the transformed object points, which better captures rigid object motion than directly supervising rotation and translation parameters. To encourage temporal consistency, we further introduce a velocity loss $\mathcal{L}_{\text{vel}}$, which penalizes discrepancies between predicted and ground-truth motion velocities for both humans and the object. This term helps reduce jittering artifacts and improves motion smoothness.

The KL divergence term regularizes all latent variables towards a standard Gaussian prior:

$$\mathcal{L}_{\text{KL}} = \text{KL}(q(z_a)||p(z)) + \text{KL}(q(z_b)||p(z)) + \text{KL}(q(z_o)||p(z)) + \text{KL}(q(z_f)||p(z)).$$

Together, these objectives ensure that the learned latent space is both compact and expressive, while preserving physically plausible and interaction-consistent HOHI dynamics.

4.3 HOHI-DIFF

Based on the structured latent space learned by HOHI-VAE, we perform generative modelling via diffusion directly in the compressed space. Instead of operating on raw motion sequences, we model the joint latent: $z = [z_a, z_b, z_o, z_f]$, which compactly represents individual human motion, object dynamics, and global interaction.

We adopt a standard forward diffusion that progressively perturbs the latent with Gaussian noise:

$$q(z_t|z_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}z_0, (1-\bar{\alpha}_t)\mathbf{I}),$$

where $\bar{\alpha}_t$ follows a predefined noise schedule.

The denoising network ϵ_θ is trained to predict the injected noise: $\epsilon_\theta(z_t, t, c)$, where c denotes conditioning signals such as text or interaction prompts. In our setting, the conditioning is applied to all latent components, allowing the model to jointly guide human motion, object behaviour, and interaction patterns.

Different from conventional latent diffusion models, our formulation explicitly preserves the structured decomposition of HOHI. In particular, the global interaction latent z_f provides a high-level coordination signal that guides the denoising of z_a, z_b, and z_o. This enables the model to maintain consistent interaction semantics during generation, rather than relying on implicit correlations. The model is optimized using the standard noise prediction objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|^2\right].$$

The denoising network ϵ_θ is instantiated as a Diffusion Transformer (DiT) Peebles and Xie (2023), which operates on the structured latent tokens using self-attention. This architecture naturally captures long-range dependencies and cross-modal interactions among z_a, z_b, z_o, and z_f, making it well-suited for HOHI generation. For conditioning, we adopt a simple yet effective additive formulation. Specifically, the condition embedding c is constructed by summing three components: a text embedding obtained from a pretrained CLIP encoder, a point-based feature extracted via a PointMLP to encode object-related information, and a sinusoidal time-step embedding. The combined embedding is then injected into the DiT through adaptive normalization layers, allowing the model to jointly modulate human motion, object dynamics, and interaction patterns during denoising. After denoising, the predicted clean latent $\hat{z}_0$ is decoded using the pretrained HOHI-VAE decoder:

$$(\hat{\mathbf{H}}_A, \hat{\mathbf{H}}_B, \hat{\mathbf{O}}) = D_{\text{VAE}}(\hat{z}_a, \hat{z}_b, \hat{z}_o, \hat{z}_f).$$

This design allows the diffusion model to focus on high-level interaction structure in a compact latent space, while delegating low-level motion reconstruction to the VAE decoder, resulting in more stable training and more coherent HOHI generation.

4.4 Experiments

Evaluation. To comprehensively evaluate HOHI generation, we follow prior HHI works Liang et al. (2024); Xu et al. (2024a) and design a contrastive learning-based CLIP-like framework, termed HOHI-CLIP. The proposed framework consists of two complementary evaluation modules. The first module evaluates text-to-motion alignment and motion realism. Specifically, we adopt a CLIP-style dual-encoder that aligns textual descriptions with human motion features. Based on this embedding space, we compute standard text-to-motion metrics, including Fréchet Inception Distance (FID_t), recall precision(RP_t), and Multi-Modality Distance (MMD_t), and Diversity (Div_t)to measure the quality and semantic consistency of generated human motions. The second module evaluates human-object interaction consistency. Following a similar contrastive formulation, we construct a joint embedding space between human motion and object motion. This enables us to measure whether the generated object dynamics are compatible with the corresponding human interaction. In this setting, (FID_o) is used to evaluate the distributional similarity between generated and real object motions, while recall precision (RP_o) and Multi-Modality Distance (MMD_o) measures the alignment between generated human motion and object motion pairs. Together, these two modules provide a comprehensive evaluation of HOHI generation, covering motion realism, semantic alignment, and interaction consistency.

Baseline. We compare our method with representative baselines, including MDM, MLD, and CHOIS. As our task requires end-to-end HOHI generation, we adapt these models to accommodate joint human-object representations while preserving their original designs as much as possible. For MDM, we concatenate human motion and object motion along the feature dimension to form a unified sequence representation. The model is conditioned on both text and object information, where textual features are extracted using a CLIP encoder and object features are encoded via a point-based

Table 3: Quantitative comparison on HOHI generation. Each result is repeated 20 times with a confidence interval reported, and the **bold** font denotes the best results.

Method	Text-Human Alignment						Human-Object Interaction		
	$\text{FID}_t(\times 10^{-2}) \downarrow$	RP_t@1↑	RP_t@2↑	RP_t@3↑	$\text{MMD}_t \downarrow$	$\text{Div}_t \rightarrow$	$\text{FID}_o(\times 10^{-2}) \downarrow$	$\text{MMD}_o \downarrow$	$\text{RP}_o \uparrow$
Ground Truth	-	$0.367^{\pm.006}$	$0.551^{\pm.006}$	$0.655^{\pm.005}$	$0.749^{\pm.001}$	$1.361^{\pm.023}$	-	$0.783^{\pm.001}$	$0.402^{\pm.004}$
MDM Tevet et al. (2023)	$4.126^{\pm.005}$	$0.357^{\pm.004}$	$0.520^{\pm.004}$	$0.629^{\pm.004}$	$0.763^{\pm.000}$	$1.312^{\pm.020}$	$8.734^{\pm.010}$	$0.854^{\pm.010}$	$0.323^{\pm.005}$
MLD Chen et al. (2023)	$4.577^{\pm.006}$	$0.384^{\pm.004}$	$0.547^{\pm.005}$	$0.648^{\pm.004}$	$0.751^{\pm.001}$	$1.320^{\pm.015}$	$6.174^{\pm.007}$	$0.825^{\pm.007}$	$0.352^{\pm.006}$
CHOIS Li et al. (2024a)	$2.202^{\pm.005}$	$0.406^{\pm.005}$	$0.596^{\pm.005}$	$0.722^{\pm.004}$	$0.671^{\pm.000}$	$\mathbf{1.344}^{\pm.027}$	$3.842^{\pm.004}$	$0.797^{\pm.007}$	$0.387^{\pm.005}$
Ours	$\mathbf{1.408}^{\pm.004}$	$\mathbf{0.460}^{\pm.004}$	$\mathbf{0.662}^{\pm.004}$	$\mathbf{0.773}^{\pm.003}$	$\mathbf{0.610}^{\pm.000}$	$1.341^{\pm.019}$	$\mathbf{2.865}^{\pm.003}$	$\mathbf{0.788}^{\pm.005}$	$\mathbf{0.396}^{\pm.005}$

Table 4: Ablation study of the proposed framework. We analyze the efficacy of the fusion encoder, object shape token, and DiT backbone on both evaluation settings.

Fusion Encoder	Shape Token	DiT	$\text{FID}_t \downarrow$	RP_t @1 ↑	$\text{MMD}_t \downarrow$	$\text{FID}_o \downarrow$	$\text{RP}_o \uparrow$
×	✓	✓	$3.875^{\pm.005}$	$0.407^{\pm.004}$	$0.680^{\pm.000}$	$4.498^{\pm.004}$	$0.375^{\pm.005}$
✓	×	✓	$1.701^{\pm.004}$	$0.455^{\pm.004}$	$0.616^{\pm.000}$	$5.740^{\pm.003}$	$0.307^{\pm.005}$
✓	✓	×	$\mathbf{1.343}^{\pm.004}$	$0.451^{\pm.003}$	$0.623^{\pm.000}$	$4.242^{\pm.003}$	$0.366^{\pm.005}$
✓	✓	✓	$1.408^{\pm.004}$	$\mathbf{0.460}^{\pm.004}$	$\mathbf{0.610}^{\pm.000}$	$\mathbf{2.865}^{\pm.003}$	$\mathbf{0.396}^{\pm.005}$

MLP. These condition embeddings are injected into the Transformer denoiser to guide the generation process. In addition, we adopt the DDIM sampling strategy to accelerate inference. For MLD, we follow the same conditioning scheme, incorporating both text and object features, while keeping its original architecture and hyperparameters unchanged to ensure a fair comparison. For CHOIS, we adapt the input feature dimensionality to match our unified HOHI representation, while maintaining its original model structure. These modifications allow all baselines to operate under the same HOHI setting, ensuring a consistent and fair evaluation.

Qualitative Results. The main results on the HOHI4D dataset are reported in Tab. 3. Compared with MDM, MLD, and CHOIS, our method achieves significant improvements across all major metrics, particularly on both FID scores, where it obtains over 30% relative improvement. These results demonstrate the superiority of our framework in generating more realistic human motions and object interactions.

Ablation Study. We conduct ablations to validate the effectiveness of our structured latent design and key components. **Structured Latent vs. Flat Latent.** We compare our model with a flat latent variant that encodes all modalities into a single latent space by removing the fusion encoder. As shown in Tab. 4, the flat latent formulation leads to a clear performance drop across all metrics, indicating that it fails to capture the tri-partite coordination in HOHI. In contrast, our structured latent design provides a more suitable inductive bias for modeling multi-agent interactions. **Object Shape Tokens.** Removing shape tokens has little effect on text-human alignment, but significantly degrades object-related performance. This shows that explicit geometry priors are important for modeling object dynamics and interaction consistency. **Diffusion Backbone.** Replacing the DiT backbone results in slightly worse performance, but the model remains competitive, suggesting that the main gain comes from the structured latent representation rather than backbone capacity.

5 Conclusion

In this work, we present a scalable framework for 3D Human-Object-Human Interaction (HOHI) generation, which addresses the long-standing lack of large-scale datasets and unified modeling methods in this domain. We first introduce a fully automated and controllable data synthesis pipeline that combines prompt-driven video generation, SAM3D-based reconstruction, and temporal pose refinement to generate realistic multi-agent interaction data at scale. Based on this pipeline, we construct HOHI4D, a large-scale benchmark containing over 5,000 interaction sequences and more than 13 hours of motion data. Furthermore, we propose HOHI-VAE and HOHI-Diff, a structured latent diffusion framework that disentangles individual human motion, object dynamics, and global interaction patterns for coherent HOHI generation. Extensive experiments demonstrate that our method significantly outperforms existing baselines in both motion realism and interaction consistency.

REFERENCES

Zhi Cen, Huaijin Pi, Sida Peng, Qing Shuai, Yujun Shen, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Ready-to-react: Online reaction policy for two-character interaction generation. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Xin Chen, Wen Jiang, Biao Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Baptiste Chopin, Hao Tang, and Mohamed Daoudi. Bipartite graph diffusion model for human interaction generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5333–5342, 2024.

Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19888–19901, 2024.

Zhaohan Feng, Ruiqi Xue, Lei Yuan, Yang Yu, Ning Ding, Meiqin Liu, Bingzhao Gao, Jian Sun, Xinhu Zheng, and Gang Wang. Multi-agent embodied ai: Advances and future directions. *Science China Information Sciences*, 69(5):151202, 2026.

Aaron Ferguson, Ahmed A. A. Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Fabian Prada, Federica Bogo, Igor Santesteban, Javier Romero, Jenna Zarate, Jeongseok Lee, Jinhyung Park, Jinlong Yang, John Doublestein, Kishore Venkateshan, Kris Kitani, Ladislav Kavan, Marco Dal Farra, Matthew Hu, Matthew Cioffi, Michael Fabris, Michael Ranieri, Mohammad Modarres, Petr Kadlecek, Rawal Khirodkar, Rinat Abdrashitov, Romain Prévost, Roman Rajbhandari, Ronald Mallet, Russell Pearsall, Sandy Kao, Sanjeev Kumar, Scott Parrish, Shoou-I Yu, Shunsuke Saito, Takaaki Shiratori, Te-Li Wang, Tony Tung, Yichen Xu, Yuan Dong, Yuhua Chen, Yuanlu Xu, Yuting Ye, and Zhongshi Jiang. Mhr: Momentum human rig, 2025.

Mingqi Gao, Yunqi Miao, and Jungong Han. Sam-body4d: Training-free 4d human body mesh recovery from videos. *arXiv preprint arXiv:2512.08406*, 2025.

Xuehao Gao, Yang Yang, Zhenyu Xie, Shaoyi Du, Zhongqian Sun, and Yang Wu. Guess: Gradually enriching synthesis for text-driven human motion generation. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

Zichen Geng, Zeeshan Hayder, Bo Miao, Jian Liu, Wei Liu, and Ajmal Saeed Mian. Disentangled hierarchical vae for 3d human-human interaction generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, 2020.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1900–1910, 2024.

Bo Han, Hao Peng, Minjing Dong, Yi Ren, Yixuan Shen, and Chang Xu. Amd: Autoregressive motion diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2022–2030, 2024.

Shaid Hasan, Breenice Lee, Sujan Sarker, and Tariq Iqbal. A multimodal framework for human-multi-agent interaction. *arXiv preprint arXiv:2603.23271*, 2026.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, October 2019. URL https://prox.is.tue.mpg.de.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. Intermask: 3d human interaction generation via collaborative masked modeling. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024a.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1737–1747, 2024b.

Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14806–14816, 2023.

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 42(6), 2023.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024a.

Tianyu Li, Calvin Qiao, Guanqiao Ren, KangKang Yin, and Sehoon Ha. Aamdm: Accelerated auto-regressive motion diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1813–1823, 2024b.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision (IJCV)*, pages 1–21, 2024.

Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 1769–1782, 2025.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6), October 2015.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2023.

Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2869–2879. IEEE Computer Society, 2025.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10106–10116, 2024.

Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. in2in: Leveraging individual information to generate human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1941–1951, June 2024.

Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15. JMLR.org, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (DDIM)*, 2021.

SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

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

Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3343–3352, 2019.

Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. Timotion: Temporal and interactive framework for efficient human-human motion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 7169–7178, 2025.

Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22035–22044, 2023a.

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

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Di Wu, Xian Wei, Guang Chen, Hao Shen, Xiangfeng Wang, Wenhao Li, and Bo Jin. Generative multi-agent collaboration in embodied ai: A systematic review. *arXiv preprint arXiv:2502.11518*, 2025a.

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

Xianghui Xie, Bowen Wen, Yan Chang, Hesam Rabeti, Jiefeng Li, Ye Yuan, Gerard Pons-Moll, and Stan Birchfield. Cari4d: Category agnostic 4d reconstruction of human-object interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026.

Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *CVPR*, pages 22260–22271, 2024a.

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

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *Advances in Neural Information Processing Systems*, 37: 52858–52890, 2024b.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026.

Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.

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

Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 509–519, 2023.

APPENDIX

CONTENTS

1 Introduction **1**

2 Related Work **3**

3 Data Synthesis **3**

3.1 Controllable Video Generation 4

3.2 4D Human-Object Reconstruction 4

3.3 Object Pose Refinement 5

4 HOHI Modelling **6**

4.1 Problem Setup 6

4.2 HOHI-VAE 6

4.3 HOHI-Diff 7

4.4 Experiments 8

5 Conclusion **9**

A Failures of CORE4D **14**

B Refinement Algorithm **15**

C Implementation Details **16**

D Prompt Design Details **16**

E Dataset Distribution **18**

F Dataset Construction Pipeline and Efficiency. **18**

G Generated Video Sample **19**

H Failure Cases **19**

I Limitation **20**

J Training logs **20**

K Video **20**

A FAILURES OF CORE4D

In this section, we highlight several prevalent failure modes observed in CORE4D, including severe object sliding, object disappearance, distorted human poses, and highly unnatural motion patterns.

Notably, these issues occur frequently throughout the dataset, which significantly limits its suitability for modeling both human–object interactions (HOI) and human–human interactions (HHI). The results are shown in Fig. 5.

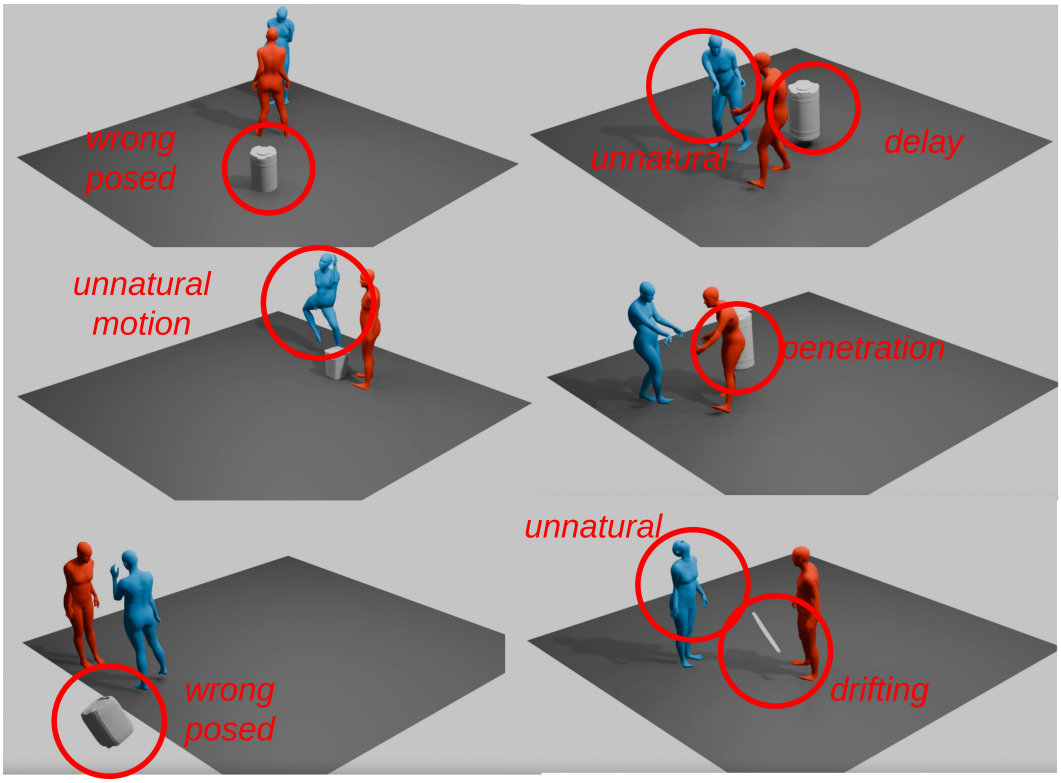


Figure 5: Failure cases in CORE4D dataset

B Refinement Algorithm

Per-frame depth score (projection supervision). For frame t, let pixel center be $\mathbf{u}_t = (u_t, v_t)$, ray be

$$\mathbf{r}_t = \frac{\mathbf{s}_{cv\to body} \odot \left(K_t^{-1}[u_t, v_t, 1]^\top\right)}{\left\|\mathbf{s}_{cv\to body} \odot \left(K_t^{-1}[u_t, v_t, 1]^\top\right)\right\|_2},$$

and coarse rotation $\mathbf{R}_t^c$ from nearest sparse pose.

For candidate depth z, translation is $\mathbf{t}(z) = z\mathbf{r}_t$, and projected geometry gives: mask score $s_m(z)$, bbox IoU $s_b(z)$, center score $s_c(z)$, size score $s_s(z)$. The depth score is

$$S_t(z) = w_m s_m(z) + w_b s_b(z)^\gamma + w_c s_c(z) + w_s s_s(z) - w_r \left|\log \frac{z}{z_t^{prior}}\right|.$$

Then

$$z_t^{proj} = \arg\max_{z\in\mathcal{Z}_t} S_t(z).$$

Depth observation fusion. Base observation (from hand/contact/sparse prior) is z_t^{base}. Projection confidence gives blend weight $\lambda_t \in [0, \lambda_{\max}]$, so

$$z_t^{obs} = (1-\lambda_t) z_t^{base} + \lambda_t z_t^{proj},$$

and measurement noise is reduced by confidence:

$$R_t = R_t^{base} \cdot \max(0.05,\ 1 - \rho\lambda_t)\,.$$

Algorithm 2 Ray-based pose refinement with projection-supervised Kalman depth

Input: $\{K_t, \mathbf{u}_t\}_{t=1}^{T}$, sparse poses $\{(\mathbf{R}_i^s, \mathbf{t}_i^s)\}$, canonical points $\mathcal{P}$
Output: refined poses $\{(\mathbf{R}_t, \mathbf{t}_t)\}_{t=1}^{T}$

1. Initialize Kalman state $(\hat{z}_0, P_0)$.
2. For $t = 1, \ldots, T$:
   (a) Pick nearest sparse pose $\rightarrow$ coarse rotation $\mathbf{R}_t^c$, sparse depth prior z_t^{prior}.
   (b) Compute ray $\mathbf{r}_t$ from $(K_t, \mathbf{u}_t)$.
   (c) Build z_t^{base} from hand/contact/sparse blending.
   (d) If projection supervision enabled: search $z \in \mathcal{Z}_t$, maximize $S_t(z)$, get z_t^{proj} and λ_t.
   (e) Fuse observation: $z_t^{obs} = (1 - \lambda_t) z_t^{base} + \lambda_t z_t^{proj}$.
   (f) Kalman update on depth: $(\hat{z}_t, P_t)$.
   (g) Set translation $\mathbf{t}_t = \hat{z}_t \mathbf{r}_t$, keep $\mathbf{R}_t = \mathbf{R}_t^c$.
3. Apply sequence-level rotation micro-refinement to $\{\mathbf{R}_t\}$ (contact-aware smoothing).

1D Kalman update (depth only).

$$\hat{z}_t^- = \hat{z}_{t-1}, \quad P_t^- = P_{t-1} + Q,$$
$$K_t = \frac{P_t^-}{P_t^- + R_t},$$
$$\hat{z}_t = \hat{z}_t^- + K_t \left(z_t^{obs} - \hat{z}_t^-\right), \quad P_t = (1 - K_t) P_t^-.$$

Final translation:

$$\mathbf{t}_t = \hat{z}_t \, \mathbf{r}_t.$$

C Implementation Details

Implementation Details. We implement both **HOI-VAE** and **HOI-Diff** in PyTorch/Lightning and train on the processed HOHI dataset (`data/hohi`) with variable-length full sequences and frame masks. For **HOI-VAE**, we use a Transformer-based hierarchical latent design with model width 512, 8 attention heads, FFN size 2048, and dropout 0.1. Each human stream encoder uses 4 Transformer layers, the interaction/top encoder uses 4 layers, and the object branch uses a point-based encoder with 1024 sampled points, point feature width 256, 2 motion layers, and 2 shape layers. The latent token setting is 2 tokens for each human stream and 2 tokens for the object stream. The decoder mirrors this setup with 4-layer Transformer decoders for two human streams and object motion. The VAE objective is a weighted sum of human reconstruction, object reconstruction (including rotation and translation terms), velocity loss, and KL regularization, with weights $\beta_{\mathrm{KL}} = 10^{-3}$, $w_{\mathrm{rec,h}} = 1.0$, $w_{\mathrm{rec,obj}} = 1.0$, $w_{\mathrm{rot}} = 0.5$, $w_{\mathrm{trans}} = 0.5$, and $w_{\mathrm{vel}} = 0.1$. We train HOI-VAE for 2000 epochs with AdamW (lr $= 10^{-4}$, weight decay $= 10^{-2}$), cosine LR decay (minimum lr $= 10^{-6}$), batch size 8, and gradient clipping 1.0.

For **HOI-Diff**, we freeze the trained HOI-VAE and denoise VAE latent tokens using an AdaLN-style DiT denoiser (8 blocks, width 512, 8 heads, FFN 2048, dropout 0.1). The condition encoder fuses text (CLIP ViT-B/32 + 4-layer text head), timestep embedding (dim 256), and object point-cloud condition (1024 sampled points, feature dim 256) into a 512-d condition vector. Diffusion training uses a DDPM scheduler with 1000 timesteps, scaled-linear β schedule ($\beta_{\text{start}} = 8.5 \times 10^{-4}$, $\beta_{\text{end}} = 1.2 \times 10^{-2}$), and ϵ-prediction loss (MSE). Sampling uses DDIM with 50 inference steps and $\eta = 0$. We train HOI-Diff for 1000 epochs using AdamW (lr $= 10^{-4}$, weight decay $= 10^{-2}$) with exponential LR decay (half-life 500 epochs), batch size 16, gradient clipping 1.0, and validation every 10 epochs.

D Prompt Design Details

In this section, we showcase some prompts that were used to generate the first frame and the whole video. In designing the prompt templates for data synthesis, we consider several practical constraints to ensure stable reconstruction and high-quality interaction generation.

First, we reduce appearance ambiguity between the two humans to avoid identity confusion during segmentation. Specifically, we enforce distinct visual attributes (e.g., different genders or clothing styles) to prevent the model from merging two individuals into a single instance.

Second, we ensure that all human body parts remain fully visible within the frame, which is critical for accurate 3D human reconstruction.

Third, we constrain the scene to be clean and minimally cluttered to avoid irrelevant objects that may interfere with video generation and reconstruction. At the same time, we avoid using overly simplistic backgrounds (e.g., solid-color or green-screen settings), as these may lead to failure cases in SAM3D-based models.

Fourth, we enforce a static camera setup without zooming or viewpoint changes, which improves temporal consistency and stabilizes both human and object reconstruction.

Finally, we minimize severe occlusions whenever possible, as heavy occlusion can significantly degrade both reconstruction quality and interaction consistency.

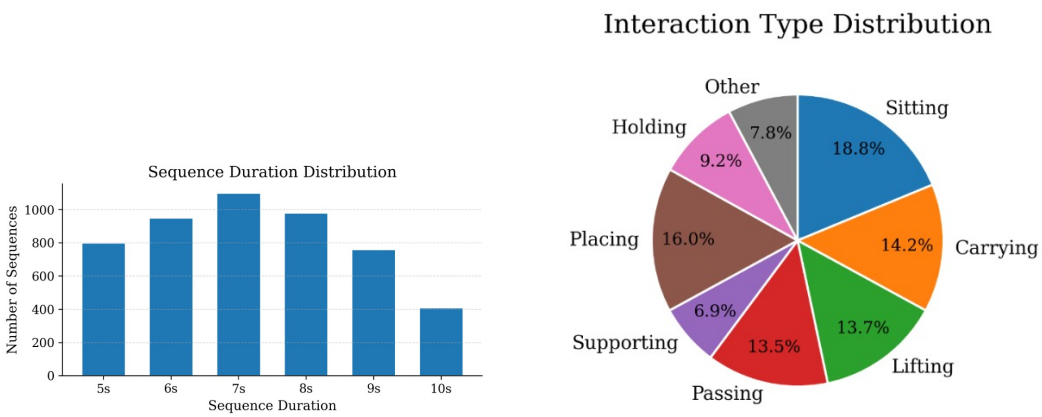
Based on these considerations, we design structured prompt templates to guide controllable and stable HOHI data generation.

```
 PREFIX: "A long-shot photo, with a wide-angle lens facing
straight ahead, there are exactly two people:  one man and one
woman,"
```

```
 SURFIX: ", with a clean indoor background, showing full bodies
of the man and the woman, and no other objects"
```

```
 Pattern:
"C1 one-holds one-observes",
"C2 two-people opposite positions",
"C3 hold-support-drag static",
"C4 one seated one standing-or-floor",
"C5 both hold from two sides",
"C6 receiving posture",
"C7 standing-plus-floor contact",
"C8 no-contact placement",
"C9 seated-and-guarding",
"C10 diagonal mirrored contact",
```

```
 STANDING POSES
"standing upright and facing the chair",
"standing with a relaxed posture",
"standing half a step from the chair and looking at it",
FLOOR POSES
"sitting on the floor with legs folded",
"sitting on the floor with knees bent",
"sitting on the floor in a relaxed pose",
OPPOSITE POSITION PAIRS
("on the left side of the chair", "on the right side of the
chair")
("in front of the chair", "behind the chair"),
("at the front-left corner of the chair", "at the back-right
corner of the chair"),
("at the front-right corner of the chair", "at the back-left
corner of the chair"),
("half a step to the left of the chair", "half a step to the
right of the chair")
```



(a) Sequence length distribution.

(b) Interaction type distribution.

Figure 6: **Dataset statistics of HOHI4D.** (a) Distribution of sequence durations, showing a balanced coverage from 5 to 10 seconds, which supports both short-term and long-horizon interactions. (b) Distribution of interaction types across different functional categories, indicating diverse object-mediated human–human interactions without severe category bias.

E Dataset Distribution

To better understand the properties of the proposed HOHI4D dataset, we provide a statistical analysis of the generated interactions in terms of diversity, temporal characteristics, and motion dynamics.

Temporal Characteristics. We analyze the distribution of sequence lengths in HOHI4D. As shown in Fig. 6a, most sequences fall within a moderate duration range from 5 to 10 seconds in 30 FPS, while longer sequences are also included to capture complex multi-stage interactions. This diversity in temporal length allows the dataset to model both short-term reactions and long-horizon coordination.

Interaction Diversity. HOHI4D covers a wide range of object categories and interaction patterns. We categorize interactions based on object types and functional roles (e.g., sitting, carrying, lifting, passing), and observe a balanced distribution across categories. This indicates that the dataset is not biased toward a small set of interaction modes, but instead captures diverse human–object–human coordination behaviors as shown in Fig. 6b.

Human-Object Motion Dynamics. We further examine the motion statistics of both humans and objects. Specifically, we compute the average translation magnitude and rotation changes of the object across time, as well as the relative distance between humans and the object. The results show that the dataset contains a wide spectrum of motion patterns, including both subtle interactions (e.g., holding or stabilizing) and large-scale movements (e.g., carrying or repositioning).

Discussion. Overall, these statistics demonstrate that HOHI4D provides diverse and dynamically rich interaction data, which is essential for learning generalizable HOHI models.

F Dataset Construction Pipeline and Efficiency.

We provide a detailed breakdown of the data synthesis pipeline to demonstrate its scalability and computational cost.

Our pipeline consists of five sequential stages. First, we generate structured initial images from text descriptions using a text-to-image model, which takes approximately 20 seconds per sample. Second, we derive temporally consistent text prompts from the generated images via a vision-language model, requiring around 10 seconds. Third, these prompts are fed into a video generation model to produce full interaction sequences, which is the most time-consuming step and takes approximately 100 seconds. Fourth, we reconstruct human motion from the generated videos using SAM3D-based models, which requires around 50 seconds per sequence. Finally, object reconstruction and pose estimation are performed using SAM3D-Object, taking approximately 3 minutes per sequence.

Overall, the entire pipeline takes roughly 5 minutes per sample. Importantly, all stages are fully automated and can be parallelized across multiple GPUs, enabling scalable large-scale data generation without human intervention.

G GENERATED VIDEO SAMPLE

In this section, we demonstrate some generated videos as shown in Fig. 7.

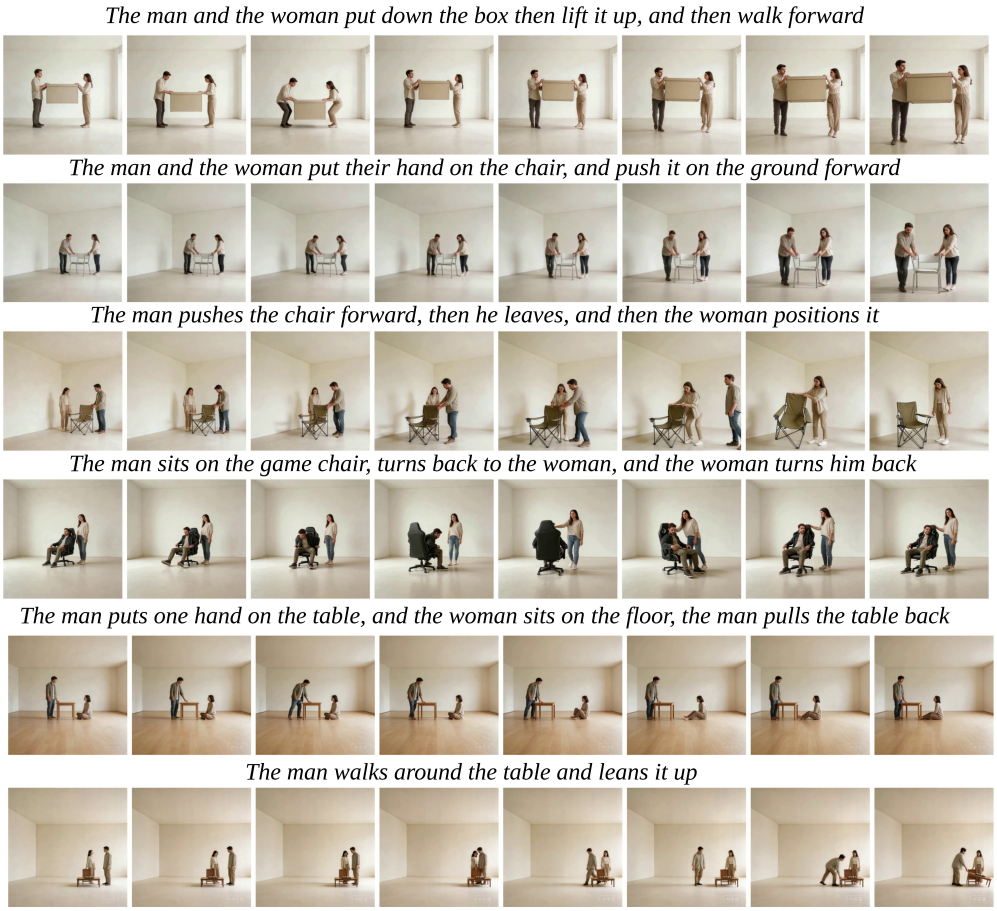


Figure 7: Samples of generated video

H FAILURE CASES

Failure Cases. We observe two primary failure modes in our pipeline.

First, when objects undergo significant deformation, the reconstruction becomes inaccurate since our method assumes rigid-body transformations, leading to incorrect contact estimation.

Second, when a human temporarily disappears from the scene (e.g., due to occlusion or generation artifacts), the reconstructed motion may become unstable or incomplete. Example cases are shown in Fig. 8.

To mitigate these issues, we plan to incorporate rigidity-aware constraints during video generation to filter out non-rigid object behaviors. In addition, we will explore more robust temporal memory mechanisms to maintain human identity and continuity, reducing failures caused by missing observations.

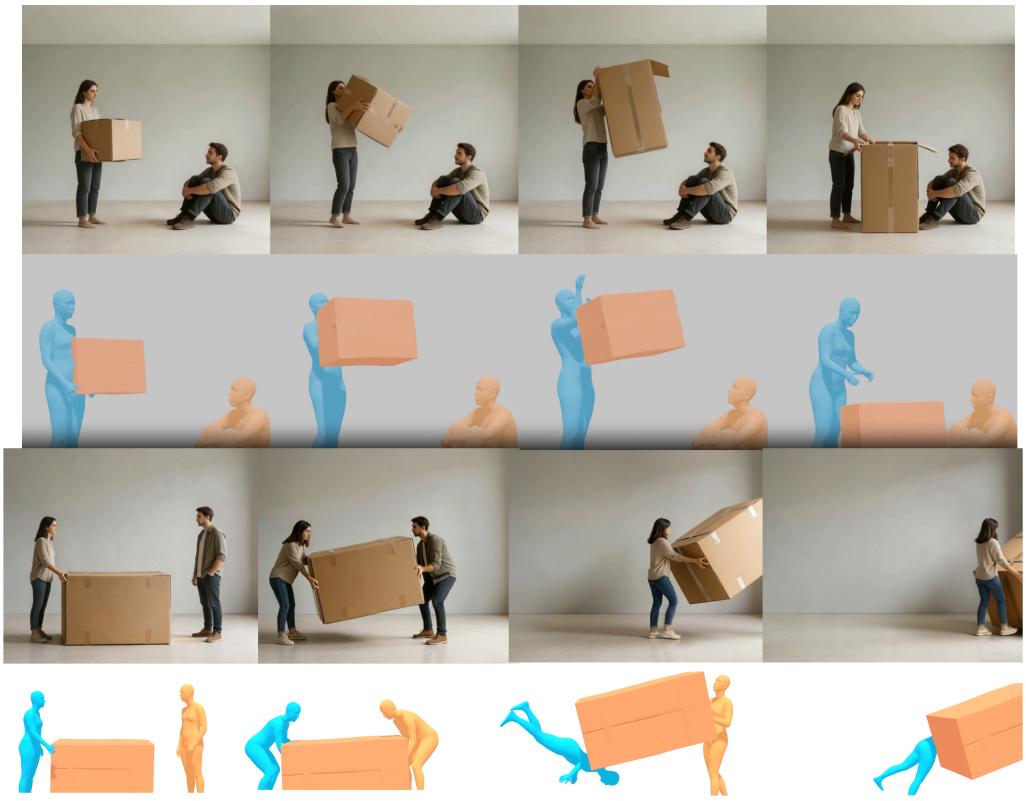


Figure 8: Two main failure cases

I LIMITATION

Despite promising results, our work still has several limitations. First, although our reconstruction and refinement pipeline substantially improves temporal consistency, accurate contact modeling in complex multi-agent interactions remains challenging, especially under severe occlusion or ambiguous object geometry. Second, the current framework mainly focuses on rigid objects and may not generalize well to highly deformable or articulated objects. In addition, while HOHI4D significantly expands the scale and diversity of existing HOHI datasets, the synthesized data may still contain reconstruction artifacts inherited from video generation and 3D reconstruction models. Finally, our diffusion framework operates in a compressed latent space and currently emphasizes global interaction dynamics rather than fine-grained physical constraints, leaving explicit physics-aware modeling as future work.

J TRAINING LOGS

The training logs are attached in the Supplementary.zip; please use `tensorboard` command to view them.

K VIDEO

The video `supplementary.mp4` with visualizations is attached in the Supplementary.zip.