

Decoupled Generative Modeling for Human-Object Interaction Synthesis

Hwanhee Jung¹ Seungwan Lee¹ Jeongyoon Yoon¹ SeungHyeon Kim¹
 Giljoo Nam² Qixing Huang³ Sangpil Kim^{1*}

¹Korea University ²Meta ³The University of Texas at Austin

Abstract

Synthesizing realistic human-object interaction (HOI) is essential for 3D computer vision and robotics, underpinning animation and embodied control. Existing approaches often require manually specified intermediate waypoints and place all optimization objectives on a single network, which increases complexity, reduces flexibility, and leads to errors such as unsynchronized human and object motion or penetration. To address these issues, we propose *Decoupled Generative Modeling for Human-Object Interaction Synthesis (DecHOI)*, which separates path planning and action synthesis. A trajectory generator first produces human and object trajectories without prescribed waypoints, and an action generator conditions on these paths to synthesize detailed motions. To further improve contact realism, we employ adversarial training with a discriminator that focuses on the dynamics of distal joints. The framework also models a moving counterpart and supports responsive, long-sequence planning in dynamic scenes, while preserving plan consistency. Across two benchmarks, *FullBody-Manipulation* and *3D-FUTURE*, *DecHOI* surpasses prior methods on most quantitative metrics and qualitative evaluations, and perceptual studies likewise prefer our results.

1. Introduction

Realistic human-object interaction synthesis (HOI) is a fundamental task with broad impact in computer vision and robotics [2, 3, 5, 26, 46]. These capabilities form the basis of modern 3D systems, enabling human motion animation and humanoid control [27, 48, 56]. Synthesizing interaction requires perceiving the pose of the object and awareness of the target goal, followed by the generation of a plausible sequence of human joint configurations that performs the required manipulation safely and intentionally [1, 47, 50, 51]. Recent approaches [24, 43] rely on interpretable natural language instructions to specify tasks, yet producing interactions that are faithful to the prompt

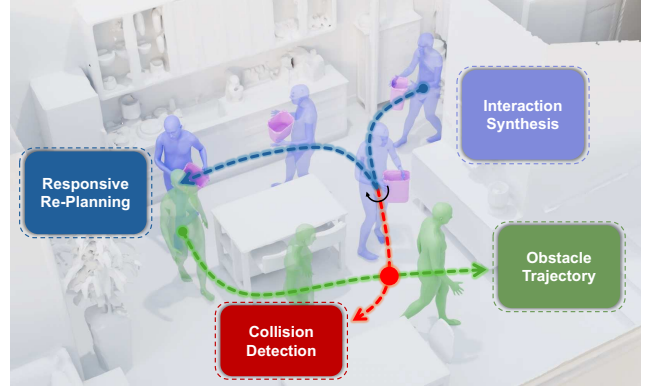


Figure 1. Overview of DecHOI for dynamic human-object interaction synthesis. The framework decouples trajectory planning and interaction synthesis, enabling collision detection and responsive re-planning for realistic, contact-consistent motion.

while keeping realism and diversity remains a difficult challenge. Prior methods [24, 43] generate 4D motion sequences of humans and objects by conditioning a diffusion model on path waypoints and text instructions. While these methods provide a strong foundation for interaction modeling that is capable of producing desired behaviors, this design introduces two challenges: i) reliance on specified intermediate waypoints that the human must follow, and ii) high optimization complexity from assigning the entire objective to a single denoising network. Specifically, for the first challenge, at inference time, the user must manually provide not only the start and goal points but also intermediate waypoints. Such reliance on externally supplied constraints narrows the model’s generative scope and reduces autonomy, while introducing procedural burden for the user. For the second challenge, HOI requires the model to simultaneously resolve the motion of both the human and the object in every frame. Assigning all objectives to a single network elevates optimization complexity and risks practical intractability. The resulting complexity often produces unsynchronized interactions, such as hovering or penetration of objects, which harms plausibility and degrades overall quality. These observations motivate us to design a new intermediate representation that is simple, flexible, informative, and still enables interactions that remain faithful to the

*Corresponding author.

instruction and plausible for HOI generation.

To this end, we propose Decoupled Generative Modeling for Human-Object Interaction Synthesis (DecHOI), a novel framework that generates trajectories and motions as separate processes. In DecHOI, a trajectory generator (TG) first produces human and object trajectories, and an action generator (AG) then synthesizes detailed actions conditioned on the generated paths. This decoupling allows the TG to produce diverse and accurate paths without reliance on specified waypoints, while the AG focuses exclusively on human actions and object motion, capturing more fine-grained details. Partitioning the model into two lightweight expert networks reduces the optimization complexity for each branch and mitigates unsynchronized interactions.

Another key factor for human-level interaction is precise hand and foot control [13, 28]. To improve the robustness of hand and foot contact and reduce undesired penetration, we introduce adversarial training applied to the object and to the distal joints. A compact discriminator focuses on hand-object interaction signals and foot dynamics, shaping the learning distribution and enabling realistic control.

For applications, the previous model [24] enables 3D scene-aware long-sequence interaction synthesis but remains restricted to static environments. To overcome this limitation, our framework models a moving counterpart and equips the human agent to either avoid the counterpart or wait, allowing it to respond to dynamic environments. This improves practicality in dynamic scenes by aligning actions with the plan and adapting to moving obstacles.

To validate the effectiveness of DecHOI, we evaluated our method on the *FullBodyManipulation* [23] and observed superior performance on most quantitative metrics and qualitative assessments. We also report results on unseen objects from the *3D-FUTURE* [9], demonstrating robustness and generalizability. For long-sequence interaction synthesis, we present scenarios in indoor environments furnished with diverse objects, where the human agent interacts with a moving counterpart and reaches various goal points by either avoiding the counterpart or waiting. The summary of our contributions is as follows.

- We propose DecHOI, a decoupled framework that separates trajectory generation from fine-grained action synthesis, reducing optimization complexity and removing the need for manual intermediate waypoints.
- We improve coordination between distal joints and an object through adversarial training with a compact discriminator, thereby reducing interpenetration.
- We enable responsive planning with a long-sequence planner that dynamically adapts to moving counterparts, supporting scene-aware interaction.
- Experiments conducted on various benchmarks demonstrate that DecHOI achieves state-of-the-art performance, surpassing previous methods in terms of realism, accu-

racy, and diversity.

2. Related Work

2.1. Human Action Generation

Human action generation aims to synthesize realistic 3D motion sequences from conditioning signals such as text descriptions, action labels, and joint-level constraints [8, 19, 30, 55, 57]. T2M [11] samples a motion length from text and then synthesizes the motion with a temporal VAE over motion snippet codes. MDM [39] adapts classifier-free diffusion [15] with a transformer backbone for high-fidelity motion, and PriorMDM [36] treats a pre-trained diffusion model as a generative prior to enable controllable composition. ACMDM [29] uses absolute global joint coordinates in a streamlined transformer diffusion model, improving fidelity and text alignment. OmniControl [44] adds spatial constraints for any joint at any time using analytic guidance with a learned realism prior, while MoMask [12] enables text-guided temporal inpainting via masked generation with hierarchical motion tokens. However, these works primarily address single human motion and neither manipulate objects nor perceive or adapt to the surrounding scene [25, 41, 42]. In contrast, we target scene-aware human-object interaction that avoids obstacles and manipulates objects in accordance with the specified intent.

2.2. Human-Object Interaction Synthesis

Human-object interaction synthesis (HOI) takes text instructions or object state as input and generates purposeful 3D human-object motion [18, 37, 49, 54]. For example, OMOMO [23] conditions on full per-frame object motion and enforces contact by integrating hand kinematics during whole body generation. CHOIS [24] and HOIFHLI [43] show that per-frame object conditioning is unnecessary because language instruction paired with sparse 3D object waypoints, together with the initial human and object states, stabilizes intended behaviors. As a complementary effort, CG-HOI [7] jointly generates human and object motion with contact cues, using contact-guided conditioning to improve realism. Despite these gains, a single network that models trajectory, pose, and contact is still hard to optimize and often produces desynchronization and contact artifacts. We instead decouple trajectory and action generation and apply contact-aware adversarial training, yielding robust synchronized HOI.

2.3. Path-conditioned HOI

Path-conditioned HOI first specifies a long-term trajectory via waypoints, object paths, or scene-level guidance, and then synthesizes it at the motion level [4, 22, 52]. CHOIS generates synchronized long-horizon human-object motion in 3D scenes by conditioning diffusion on sparse object

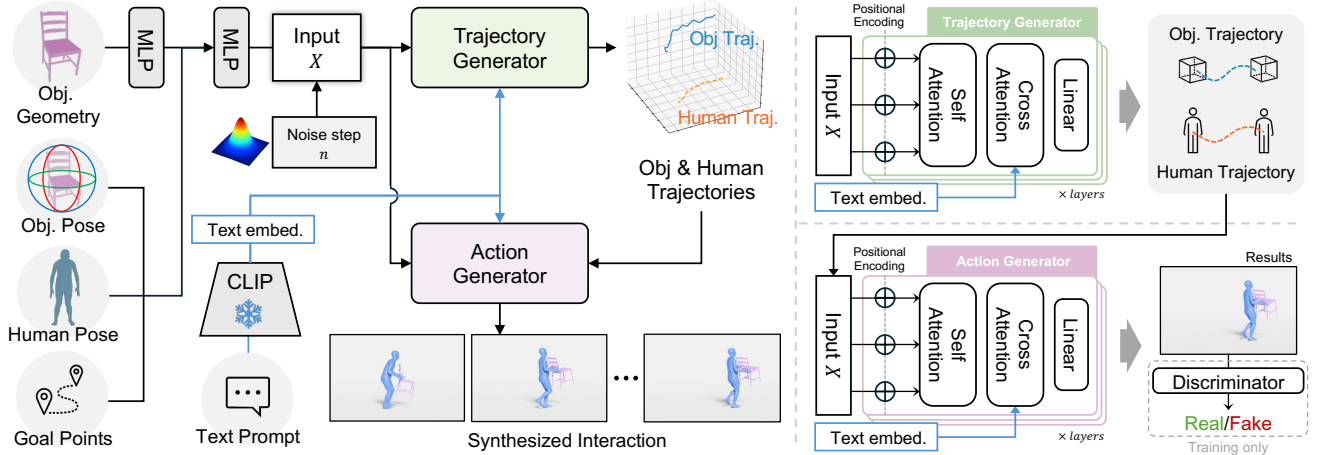


Figure 2. Architecture of DecHOI showing the decoupled trajectory and action generation process. Conditioned on the text instruction, geometry, current human and object poses, and a goal point, the trajectory generator plans paths, while the action generator produces joint motions on these paths to yield synchronized, contact-aware interactions. The right panels detail the Trajectory and Action Generators.

waypoints obtained from Habitat [35, 38] for the given start and target positions. NIFTY [21] couples an object-conditioned motion diffusion model with a learned interaction manifold that evaluates distances to a feasible interaction manifold and guides sampling toward contact-consistent behaviors. OMOMO [23] takes per-frame object motion data as input and synthesizes manipulation by denoising, incorporating hand motion to maintain contact during whole body generation. However, in inference, these formulations retain a fixed plan or guidance and therefore do not adapt to external scene changes. By contrast, our method updates the plan, enabling adaptation to external scene changes while preserving high-level intent.

3. Method

Given an instruction, current human and object poses, and a goal point, our framework synthesizes an interaction that moves the object to the goal as instructed. As illustrated in Fig. 2, we adopt a decoupled generation architecture with a trajectory generator and an action generator (Sec. 3.2) to reduce reliance on explicit waypoints and simplify the optimization objective. To improve contact realism and reduce interpenetration, we employ adversarial training with a compact discriminator that focuses on the dynamics of distal joints (Sec. 3.3). In Sec. 3.4, we present the overall optimization procedure and the inference-time guidance strategies for our method.

3.1. Background: Denoising Diffusion Model

Denoising Diffusion Probabilistic Models (DDPMs) [16, 39] are generative models that learn to approximate complex data distributions by gradually adding noise and then removing it. This paradigm has become dominant in recent work on human action generation [29, 36, 39, 44] and interaction synthesis [23, 24, 43, 45]. The forward process

progressively corrupts a clean sample \mathbf{x}_0 by adding Gaussian noise over N steps, defined as:

$$q(\mathbf{x}_n | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_n; \sqrt{\bar{\alpha}_n} \mathbf{x}_0, (1 - \bar{\alpha}_n) \mathbf{I}), \quad (1)$$

where $\alpha_n = 1 - \beta_n$ and $\bar{\alpha}_n = \prod_{s=1}^n \alpha_s$. The forward process is a Markov chain with Gaussian transitions. As n increases, \mathbf{x}_n converges to an isotropic Gaussian. The reverse process learns to invert this noising procedure with a parameterized model p_θ that progressively denoises \mathbf{x}_n to recover clean data:

$$p_\theta(\mathbf{x}_{n-1} | \mathbf{x}_n, \mathbf{c}) = \mathcal{N}(\mathbf{x}_{n-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_n, n, \mathbf{c}), \sigma_n^2 \mathbf{I}), \quad (2)$$

where \mathbf{c} denotes conditioning inputs such as text embeddings, object geometry, or start and goal states. Following prior formulations [24, 43], our denoising network predicts the clean sequence rather than the noise:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, n \sim [1, N]} \|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|_1. \quad (3)$$

Predicting $\hat{\mathbf{x}}_0$ has been found to yield better results for motion data and allows guidance losses to be applied at each denoising step.

3.2. Decoupled Generative Modeling

Jointly estimating and optimizing human and object poses is non-trivial for a single network to handle. We address this challenge with decoupled generative modeling that separates trajectory generation from action generation. At the start of the pipeline the inputs are the object pose sequence $P_o \in \mathbb{R}^{T \times 12}$ with T frames, where each frame contains the global 3D position and a 9-parameter relative rotation matrix, and the human pose sequence $P_h \in \mathbb{R}^{T \times D_h}$, where D_h comprises global joint coordinates and per-joint 6D rotation parameters. To ensure accurate interaction and instruction following, inputs also include object geometry

$B \in \mathbb{R}^{1024 \times 3}$ represented as a Basis Point Set (BPS) [33] and a text instruction.

Trajectory Generator. We formulate the trajectory generator (TG) as a conditional denoising diffusion model [16]. Given an input sequence, we apply a forward noising process for N steps following a Markov chain. For conditioning, the human and object poses of the start frame remain clean, and the object position in the end frame is also kept clean to set the goal point. Then these noisy data are concatenated with the geometry embedding $F_{\text{obj}} \in \mathbb{R}^D$, computed by a simple MLP encoder from B . We obtain the conditioning input $X \in \mathbb{R}^{T \times (12 + D_h + D)}$ for the denoising network. A transformer-based denoising network of TG generates both global object and human trajectories that are faithful to the instruction. For example, given the instruction “*Lift the chair; move the chair; and put down the chair,*” the object trajectory rises in the early part of the sequence and descends as it approaches the goal point. To learn this alignment, we incorporate a text embedding $F_{\text{text}} \in \mathbb{R}^D$ obtained from a CLIP encoder [34] into the model inputs. In contrast to prior methods [24, 43] that condition by concatenating the embedding along the sequence dimension, we use a cross-attention layer to apply the conditioning explicitly, which yields stronger alignment between text and motion features. Afterward, the TG yields continuous 3D paths for the object $\hat{\mathcal{T}}_o \in \mathbb{R}^{T \times 3}$ and the human $\hat{\mathcal{T}}_h \in \mathbb{R}^{T \times 3}$.

Action Generator. Once reliable human and object trajectories have been generated, we use them as richer conditions. The action generator (AG) receives the same input X as the trajectory generator (TG), except that the noisy global positions of the object and the human root in all T frames are replaced with the trajectories produced by TG. This dense conditioning provides stronger priors and reduces the complexity of the learning objective. AG is a lightweight diffusion model that employs a transformer-based denoising network with the same architecture as TG. Given the conditioned noisy input, it generates the full pose of the object $\hat{P}_o \in \mathbb{R}^{T \times 12}$ and the pose of the human joints $\hat{P}_h \in \mathbb{R}^{T \times D_h}$. The human pose generated is then used for parametric human modeling, and we apply SMPL-X [32] to reconstruct the human mesh and pose.

3.3. Adversarial Training for Distal Joints

Realistic human-object interaction (HOI) requires not only plausible motion but also reliable contact [6, 10, 20]. Language instructions such as “*Lift the chair*” or “*Kick the box*” rely on distal joints in the hands and feet. Existing models often cause the hands and feet to drift into empty space or intersect objects [28, 46, 53]. To address these issues, we introduce an adversarial training mechanism that regularizes contact by focusing on the distal joints.

In general, the most reliable cue to decide whether an input is real or fake is the distance between the object sur-

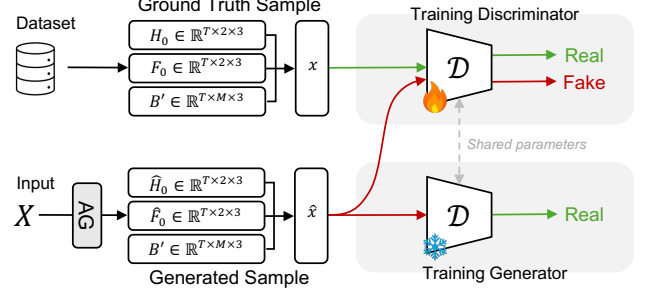


Figure 3. Adversarial module of DecHOI, where a hand and foot-focused discriminator contrasts real and generated interactions to enhance contact realism.

face and the distal joints [6, 17, 28, 40]. In ground truth, these distances are small due to complete contact, whereas generated results often exhibit larger gaps. Optimizing the generator so that the discriminator cannot distinguish real from fake therefore acts as a regularizer that drives contact toward completeness. Motivated by this observation, we design a compact discriminator \mathcal{D} that differentiates real and fake using distal joint kinematics together with object geometry. As shown in Fig. 3, \mathcal{D} receives global coordinates of the hands $H \in \mathbb{R}^{T \times 2 \times 3}$ and feet $F \in \mathbb{R}^{T \times 2 \times 3}$ over all T frames, obtained by forward kinematics from the 6D joint rotations. In addition, we sample M points from the geometry of the object B to obtain $B' \in \mathbb{R}^{T \times M \times 3}$ and apply the relative rotation and translation for each frame. For real data, we feed clean inputs x , and for fake data, we use the outputs \hat{x} produced by AG. The discriminator is trained with the following loss:

$$\mathcal{L}_{\mathcal{D}} = \frac{1}{T} \sum_{t=1}^T ([1 - s_t^{(r)}]_+ + [1 + s_t^{(f)}]_+), \quad (4)$$

$$s_t^{(r)} = \mathcal{D}(x)_t, \quad s_t^{(f)} = \mathcal{D}(\hat{x})_t,$$

where $s_t^{(r)}$ and $s_t^{(f)}$ denote the scores for real and fake data at frame t . After completing the discriminator training stage, we train AG to fool the discriminator by minimizing the following generator objective:

$$\mathcal{L}_G = -\frac{1}{T} \sum_{t=1}^T s_t^{(f)}. \quad (5)$$

Through adversarial training, the network enforces higher contact fidelity and synthesizes realistic interactions.

3.4. Training and Inference

Objectives for Training. During optimization, the trajectory generator (TG) and the action generator (AG) are trained with distinct objectives. TG focuses exclusively on planning trajectories. Given the input X , it produces $\hat{\mathcal{T}}_0 = \{\hat{\mathcal{T}}_o, \hat{\mathcal{T}}_h\}$ and is trained to reconstruct the clean trajectory representation $\mathcal{T}_0 = \{\mathcal{T}_o, \mathcal{T}_h\}$ with an L_1 objective. The loss is defined as:

Methods	Condition Matching		Human Motion Quality					Interaction Quality					GT Difference			
	$T_s \downarrow$	$T_e \downarrow$	$H_{\text{feet}} \downarrow$	$FS \downarrow$	$R_{\text{prec}} \uparrow$	$FID \downarrow$	$DIV \rightarrow$	$C_{\text{prec}} \uparrow$	$C_{\text{rec}} \uparrow$	$C_{F1} \uparrow$	$P_{\text{hand}} \downarrow$	$P_{\text{body}} \downarrow$	MPJPE \downarrow	$T_{\text{root}} \downarrow$	$T_{\text{obj}} \downarrow$	$O_{\text{obj}} \downarrow$
Lin-OMOMO [23]	-	-	10.04	0.44	0.45	16.81	6.65	0.75	0.56	0.59	0.70	0.65	17.45	29.20	75.65	-
Pred-OMOMO [23]	2.34	9.66	6.99	0.38	0.63	13.01	6.68	0.67	0.48	0.52	0.59	0.60	23.63	49.12	51.86	1.23
CHOIS [24]	1.92	8.01	6.15	0.41	0.68	1.58	8.31	0.74	0.53	0.58	0.66	0.61	18.86	44.04	51.86	1.23
HOIFHLI [43]	1.73	7.65	4.77	0.38	0.62	2.06	8.55	0.77	0.61	0.64	0.58	0.61	19.31	40.87	50.96	1.18
Ours (DecHOI)	1.59	6.91	4.42	0.38	0.72	0.33	8.86	0.80	0.64	0.67	0.53	0.54	15.27	25.47	22.96	0.86

Table 1. Quantitative comparison on the *FullBodyManipulation* [23] with CHOIS [24], HOIFHLI [43], and OMOMO [23] variants (Lin-OMOMO and Pred-OMOMO) across four categories of evaluation metrics. Arrows indicate direction: (\uparrow) means higher is better, (\downarrow) means lower is better, and (\rightarrow) means closer to the real data value is better. The real-data *DIV* reference is 9.02.

Methods	Condition Matching		Human Motion Quality				Interaction Quality	
	$T_s \downarrow$	$T_e \downarrow$	$H_{\text{feet}} \downarrow$	$FS \downarrow$	$R_{\text{prec}} \uparrow$	$FID \downarrow$	$C_{\%}$	$P_{\text{hand}} \downarrow$
Lin-OMOMO [23]	-	-	7.92	0.48	0.57	8.85	0.26	0.20
Pred-OMOMO [23]	4.72	10.92	6.61	0.45	0.59	5.16	0.40	0.17
CHOIS [24]	5.75	10.28	4.20	0.42	0.61	2.04	0.46	0.18
Ours (DecHOI)	4.27	8.43	4.06	0.41	0.69	1.01	0.48	0.15

Table 2. Quantitative results on the *3D-FUTURE* [9]. DecHOI achieves better trajectory accuracy, motion stability, and contact realism than CHOIS [24] and OMOMO [23] baselines.

$$\mathcal{L}_{\text{TG}} = \mathbb{E}_{\mathcal{T}_0, n \sim [1, N]} \|\hat{\mathcal{T}}_0 - \mathcal{T}_0\|_1. \quad (6)$$

AG receives the same input X as TG, while the trajectories are kept clean. AG is trained to reconstruct the entire motion $P_0 = \{P_o, P_h\}$ from which the generator predicts $\hat{P}_0 = \{\hat{P}_o, \hat{P}_h\}$. The reconstruction objective is:

$$\mathcal{L}_{\text{AG}} = \mathbb{E}_{P_0, n \sim [1, N]} \|\hat{P}_0 - P_0\|_1. \quad (7)$$

To further stabilize the reconstruction of the distal joints, we add a forward kinematic loss. Using the predicted relative joint rotations, we compute global hand positions \hat{H}_0 and global foot positions \hat{F}_0 from the pelvis root. These are supervised by clean distal joint positions H_0 and F_0 with:

$$\mathcal{L}_{\text{FK}} = \|\hat{H}_0 - H_0\|_1 + \|\hat{F}_0 - F_0\|_1. \quad (8)$$

The total objective for training the generators combines the reconstruction and adversarial terms:

$$\mathcal{L} = \lambda_{\text{TG}} \mathcal{L}_{\text{TG}} + \lambda_{\text{AG}} \mathcal{L}_{\text{AG}} + \lambda_{\text{FK}} \mathcal{L}_{\text{FK}} + \lambda_{\text{G}} \mathcal{L}_{\text{G}}, \quad (9)$$

where λ_{TG} , λ_{AG} , λ_{FK} , and λ_{G} are scalar weights that balance the contributions of the terms.

Inference-time Guidance. Inspired by previous work [24], our model applies reconstruction guidance to regularize the generation process. This design injects constraints at each denoising step without retraining the network for a specific purpose, which permits flexible control over the outputs. The process is formally represented as:

$$\tilde{P}_0 = \hat{P}_0 - \alpha \Sigma_n \nabla_{P_n} \mathcal{F}(\hat{P}_0), \quad (10)$$

where \mathcal{F} is a regularization objective and α controls the perturbation strength. In our implementation, \mathcal{F} encourages precise contact by penalizing distances between distal joints and the object surface, and stabilizes stance by minimizing deviations of foot joints from the ground plane to prevent hovering and penetration. Additional details are provided in the supplementary material.

4. Experiments

4.1. Datasets

Our experiments utilize two datasets designed to capture realistic human-object interactions across diverse indoor environments, following the evaluation setup in prior work [24]. i) *FullBodyManipulation* [23]: Comprising about 10 hours of synchronized human-object motion sequences, this dataset covers 15 distinct rigid objects. We consider only rigid body interactions and therefore exclude sequences with articulated objects. Motion data from 15 subjects are used for training, with data from two additional subjects reserved for testing, ensuring a consistent evaluation protocol. ii) *3D-FUTURE* [9]: Containing a broad range of 3D furniture models. This dataset offers substantial geometric diversity. To evaluate generalization to unseen appearances and shapes, we substitute the 17 test objects in *FullBodyManipulation* with unseen *3D-FUTURE* models of the same categories, enabling systematic assessment on previously unobserved instances.

4.2. Evaluation Metrics

Condition Matching. This metric evaluates how accurately the generated motion aligns with the specified object conditions. We measure the Euclidean distance between the predicted and target object start and end positions T_s and T_e in the scene, reported in centimeters (cm).

Human Motion Quality. We evaluate realism and plausibility with five metrics: foot height (H_{feet}), foot sliding (FS), R-precision (R_{prec}), Fréchet Inception Distance (FID) [14], and Diversity (DIV). H_{feet} is the mean foot-to-ground height. FS measures horizontal foot displacement during stance, following [24]. Both are reported in centimeters. R_{prec} (top-3) scores text-motion alignment. FID compares generated and ground-truth motion distributions in a learned feature space. DIV quantifies variation across samples under the same condition.

Interaction Quality. This category evaluates interaction accuracy and naturalness. We report hand-object contact precision (C_{prec}), recall (C_{rec}), F1 (C_{F1}), and percent ($C_{\%}$) as an overall contact reliability measure. We also report hand-object (P_{hand}) and body-object (P_{body}) penetration to quantify interpenetration depth (cm).

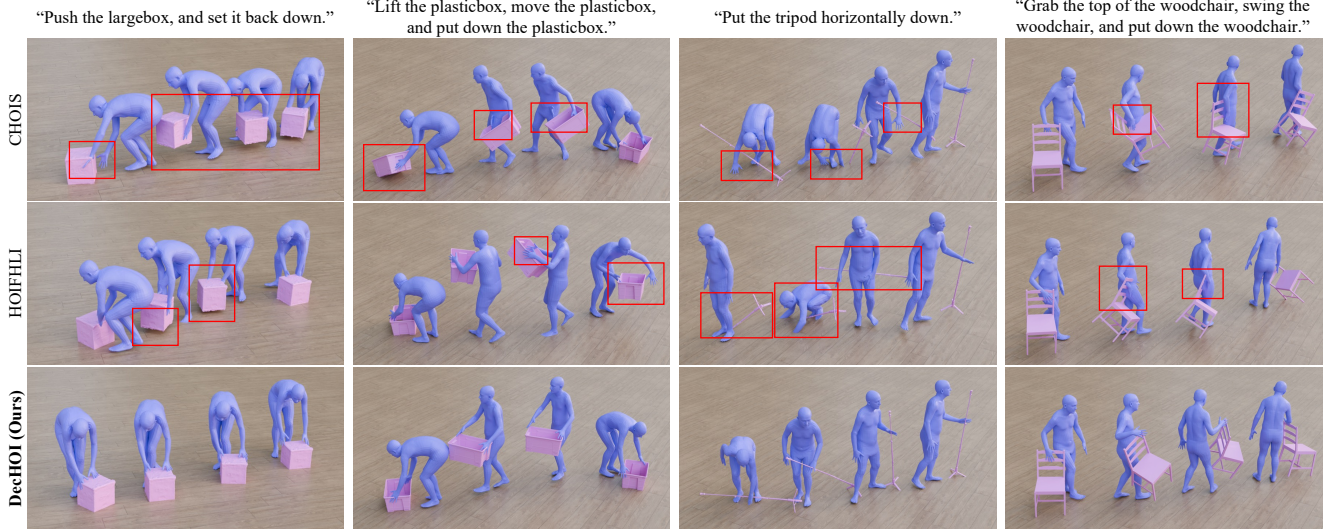


Figure 4. Qualitative comparison of DecHOI with CHOIS [24] and HOIFHLI [43] on the *FullBodyManipulation* [23]. DecHOI produces stable contacts, smooth motion, and accurate object trajectories, while prior methods show drift, penetration, or inconsistent coordination between human and object motions.

Ground Truth Difference. This metric quantifies the deviation of generated motions from ground truth motions in both spatial and rotational domains. We report mean per-joint position error (MPJPE), root translation error (T_{root}), and object translation error (T_{obj}) as Euclidean distances in centimeters. Orientation consistency is evaluated with object orientation errors (O_{obj}), computed as the Frobenius norm of rotation differences.

4.3. Quantitative Analysis

We quantitatively evaluate our approach against CHOIS [24], HOIFHLI [43], and two OMOMO [23] variants (Lin- and Pred-OMOMO) implemented following CHOIS on two datasets: *FullBodyManipulation* [23] and *3D-FUTURE* [9]. Because our method does not require specified intermediate waypoints, we evaluate all baselines without intermediate waypoint inputs to ensure a fair comparison. Note that Lin-OMOMO uses ground truth for all object-related signals, so metrics that are not applicable under this setting are reported as (-).

FullBodyManipulation. As shown in Tab. 1, our model surpasses all baselines on most metrics. OMOMO [23] variants condition on object pose at every frame, yet under our extremely limited inputs they underperform in overall human motion and interaction quality. Methods relying on sparse object conditioning [24, 43] also degrade on contact-related measures once waypoints are removed, as the reduced conditioning increases task complexity and leads to unsynchronized interactions. In contrast, DecHOI simplifies the problem through decoupled modeling, allowing the trajectory generator to produce accurate paths with low condition matching error and to attain competitive *DIV* with reduced reliance on waypoints. The resulting interactions exhibit higher realism and stronger alignment with

instructions, as reflected by improved *FID* and R_{prec} . Although contact accuracy and penetration scores often trade off against each other, our adversarial training achieves balanced performance across both and produces realistic interactions. Further analyses and ablations are provided in the supplementary material.

3D-FUTURE. To further evaluate the generalization capability of our approach, we test it on unseen *3D-FUTURE* samples that share the same object categories as those in the *FullBodyManipulation* test set. As shown in Tab. 2, our method maintains consistent trends across metrics, achieving strong trajectory accuracy and high contact reliability. Moreover, the improved *FID* indicates better alignment between generated and GT motion distributions. These results demonstrate that DecHOI effectively synthesizes realistic and coherent human-object interactions even for previously unseen objects, highlighting its robust generalization and broad applicability across diverse objects.

4.4. Qualitative Analysis

In Fig. 4, we present qualitative comparisons among CHOIS [24], HOIFHLI [43], and our DecHOI on the *FullBodyManipulation* [23]. As illustrated in the four scenes, DecHOI consistently produces more stable and realistic interactions. For the *largebox* scene, CHOIS and HOIFHLI yield unstable object motion, leading to box hovering and hand-object penetration. In contrast, DecHOI maintains a firmly grounded trajectory and cleaner contacts through stable global path modeling. In the *plasticbox* scene, baselines exhibit severe penetration between the human body and the object, along with misaligned hand contacts. In contrast, our adversarial training on distal joints mitigates these failures and enhances contact realism. For the *tripod* scene,

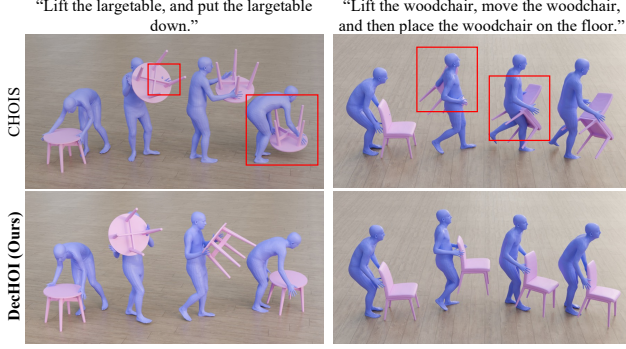


Figure 5. Qualitative comparison of DecHOI and CHOIS [24] on the *3D-FUTURE* [9], showing generalization to unseen objects.

both struggle with fine manipulation as the object drifts and rotates inconsistently, whereas our decoupled framework suppresses such unsynchronized artifacts and achieves more coherent motion. Lastly, in the *woodchair* scene, prior works show noticeable mesh intersection and incomplete contact, while DecHOI produces stable contacts and precise spatial alignment between the hand and object. As shown in Fig. 5, we further compare CHOIS and DecHOI on two unseen objects from the *3D-FUTURE* [9]: *targetable* and *woodchair*. CHOIS often fails to adapt to these novel geometries, leading to human-object intersection and motion instability, whereas our method preserves accurate contacts and smooth motion, demonstrating strong generalization to unseen shapes. More qualitative results and analysis are provided in the supplementary material.

Additionally, Fig. 6 visualizes the loss landscape to compare the training objective complexity of CHOIS, which uses a single network, and DecHOI. The surface of CHOIS appears noisy and highly rugged with many local minima, indicating sensitivity to initialization and unstable training. In contrast, DecHOI exhibits a smoother landscape with fewer local minima and more stable convergence. These observations suggest that our decoupled modeling effectively reduces overall optimization complexity.

4.5. Long-term Dynamic Planning

To evaluate long-term and dynamic human-object interactions, we introduce *DynaPlan*, which supports responsive planning and scene-aware interaction generation under multi-agent conditions. We equip both the agent and the moving counterpart with circular influence radii and detect potential collisions when their regions intersect. When a collision is detected, we re-plan with A^* , adaptively choosing either a short detour or waiting to maintain goal-directed and collision-free motion while accounting for the future path of the counterpart. To handle uncertainty in that future path, we predict it with a pre-trained trajectory prediction network [31]. We evaluate performance on about 190 long-sequence indoor scenarios. Further details are provided in the supplementary material.

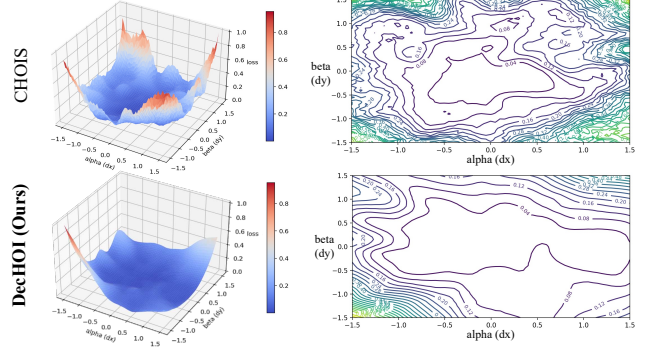


Figure 6. Visualization of training loss landscapes for DecHOI and CHOIS [24], demonstrating reduced optimization complexity.

Methods	Condition Matching		Human Motion Quality		Interaction Quality				
	$T_s \downarrow$	$T_e \downarrow$	$H_{\text{feet}} \downarrow$	$FS \downarrow$	$C\%$	$P_{\text{hand}} \downarrow$	$P_{\text{body}} \downarrow$	$P_{o \rightarrow s} \downarrow$	$P_{h \rightarrow s} \downarrow$
CHOIS [24]	2.19	8.05	3.14	0.43	0.78	0.73	0.73	1.00	0.83
Ours (DecHOI)	1.90	7.98	2.85	0.45	0.76	0.65	0.63	0.72	0.65

Table 3. Quantitative results between CHOIS [24] and DecHOI for responsive long-term interaction synthesis on the *DynaPlan*.

Quantitative Results. Tab. 3 summarizes the quantitative results for long-sequence human-object interaction synthesis on the *DynaPlan*. Compared to CHOIS [24], DecHOI achieves lower trajectory errors T_s and T_e , along with reduced instability, resulting in smoother and more stable motion over extended sequences. For interaction quality, DecHOI shows clear improvements across all penetration metrics, including P_{hand} , P_{body} , $P_{o \rightarrow s}$, and $P_{h \rightarrow s}$. Here, P_{hand} and P_{body} measure the penetration depth between the human mesh and manipulated objects, while $P_{o \rightarrow s}$ and $P_{h \rightarrow s}$ quantify collisions between the object or human and the surrounding static scene. Lower values on most metrics indicate that DecHOI effectively minimizes interpenetration and maintains collision-free trajectories even in cluttered indoor environments.

These performance gains stem from the decoupled structure of the trajectory and action generators, which separates global path planning from detailed motion synthesis, combined with contact-aware adversarial learning that enforces spatial consistency throughout long-horizon interactions. Overall, DecHOI demonstrates superior stability, plausibility, and contact accuracy in dynamic multi-agent settings, outperforming CHOIS quantitatively.

Qualitative Results. Fig. 7 visualizes DecHOI operating in two long-term dynamic indoor scenes from *DynaPlan*. In both scenes, the agent encounters a moving counterpart along its original path and initiates reactive re-planning to maintain safe, goal-directed motion. Throughout these sequences, the agent maintains a continuous and valid trajectory with the manipulated object, and the action generator smoothly adapts joint motions to the updated trajectories. These results highlight DecHOI’s ability to sustain collision-free and consistent coordination between human and object across extended dynamic environments.

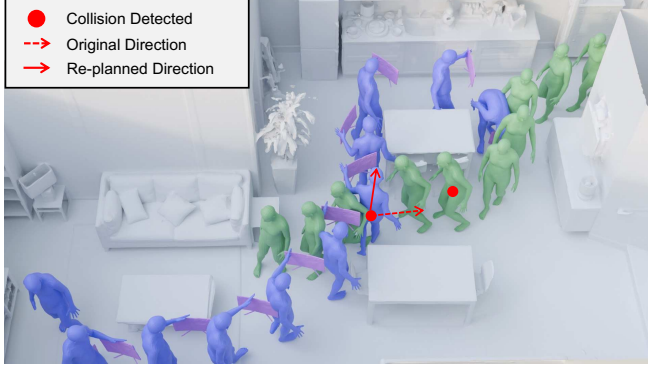


Figure 7. Visualization of DecHOI in long-sequence dynamic environments. The human agent (blue) adaptively re-plans its path when encountering a moving obstacle (green), choosing between detour and waiting behaviors to maintain goal-directed, collision-free motion.

Methods	Condition Matching		Human Motion Quality		Interaction Quality		
	$T_s \downarrow$	$T_c \downarrow$	$FS \downarrow$	$R_{prec} \uparrow$	$C_{F1} \uparrow$	$P_{hand} \downarrow$	$P_{body} \downarrow$
Baseline	1.72	7.92	0.35	0.67	0.65	0.58	0.60
w/ adversarial	1.68	7.78	0.39	0.64	0.66	0.53	0.57
w/ cross-attention	1.75	7.82	0.38	0.70	0.54	0.56	0.56
Ours (DecHOI)	1.59	6.91	0.38	0.72	0.67	0.53	0.54

Table 4. Ablation results for DecHOI on the FullBodyManipulation [23], evaluating the contribution of each component.

4.6. Ablation Study

As shown in Tab. 4, we perform an ablation study on the *FullBodyManipulation* [23] to analyze the effects of adversarial training and text conditioning based on cross-attention. The baseline, which uses only the decoupled generative modeling and concatenates text embeddings along the sequence axis following prior work [24, 43], achieves moderate performance but suffers from high penetration errors and limited contact consistency. Introducing adversarial training substantially reduces P_{hand} and P_{body} , indicating improved realism and more stable interactions, although the linguistic alignment measured by R_{prec} slightly decreases. Conversely, adding cross-attention layers for text conditioning enhances R_{prec} by effectively capturing semantic intent, but slightly degrades interaction quality due to weaker regularization. When both modules are combined in DecHOI, the model achieves the best overall performance, simultaneously improving R_{prec} , C_{F1} , and penetration metrics. These results indicate that adversarial training enhances the robustness of distal joint interactions during object manipulation, and that text conditioning based on cross-attention enables more explicit transfer of semantic information across modalities, leading to stronger alignment.

4.7. User Study

We conducted a user study to evaluate the perceptual realism and text alignment of DecHOI compared with CHOIS [24] and HOIFHLI [43]. For each comparison, a total of 60 text-scene pairs, 30 per method, were randomly sampled from our full test set, and 200 participants were recruited via Amazon Mechanical Turk (AMT). Each participant viewed two anonymized clips, each lasting about

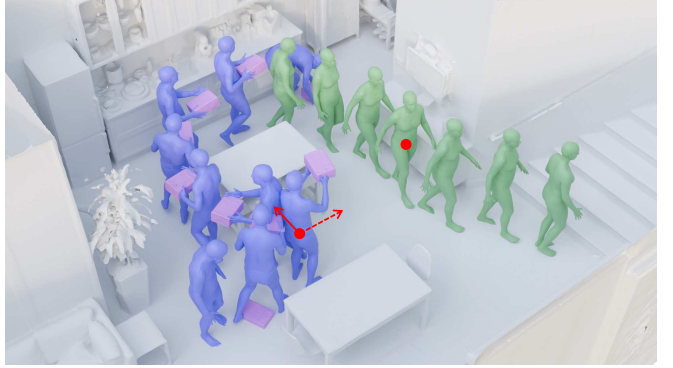


Figure 8. Stacked horizontal bars showing user preference distributions for DecHOI compared with CHOIS [24] and HOIFHLI [43] across two evaluation criteria: Text Alignment and Interaction Quality.

four seconds, in random order and answered two questions: i) which video better matches the text instruction (Text Alignment), and ii) which video shows better human-object interaction quality (Interaction Quality). Responses were collected using a two-choice scale: *Prefer Ours* or *Prefer Others*. As summarized in Fig. 8, DecHOI was consistently preferred on both criteria across both baselines, confirming that our decoupled framework yields motions that are semantically faithful to textual instructions and more realistic, stable, and natural than prior methods.

5. Conclusion

In this work, we propose a decoupled generative modeling framework for human-object interaction synthesis that separates trajectory generation from fine-grained action generation. This formulation lowers optimization complexity and enables the model to learn precise and stable motion, which reduces temporal asynchrony and penetration. It also removes the need for manually specified intermediate waypoints, improving practical applicability. Additionally, a compact adversarial discriminator focused on distal joint cues further improves interaction fidelity and strengthens contact realism. We also present a dynamic planner that delivers robust and adaptive path planning in long-horizon, multi-agent settings, as demonstrated on the *DynaPlan*. Extensive quantitative and qualitative results show state-of-the-art performance across major metrics, producing realistic and semantically aligned human-object interaction.

References

- [1] Maya Antoun and Daniel Asmar. Human object interaction detection: Design and survey. *Image and Vision Computing*, 130:104617, 2023. 1
- [2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 1
- [3] Yichao Cao, Qingfei Tang, Xiu Su, Song Chen, Shan You, Xiaobo Lu, and Chang Xu. Detecting any human-object interaction relationship: Universal hoi detector with spatial prompt learning on foundation models. *Advances in Neural Information Processing Systems*, 36:739–751, 2023. 1
- [4] Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1855–1866, 2024. 2
- [5] Yixin Chen, Sai Kumar Dwivedi, Michael J Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17100–17110, 2023. 1
- [6] Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun S Lakshmipathy, Agniv Chatterjee, Michael J Black, and Dimitrios Tzionas. Pico: Reconstructing 3d people in contact with objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1783–1794, 2025. 4
- [7] 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*, pages 19888–19901, 2024. 2
- [8] Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision*, pages 93–109. Springer, 2024. 2
- [9] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129(12):3313–3337, 2021. 2, 5, 6, 7
- [10] Dongjun Gu, Jaehyeok Shim, Jaehoon Jang, Changwoo Kang, and Kyungdon Joo. Contactgen: Contact-guided interactive 3d human generation for partners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1923–1931, 2024. 4
- [11] 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*, pages 5152–5161, 2022. 2
- [12] 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*, pages 1900–1910, 2024. 2
- [13] Yuze Hao, Jianrong Zhang, Tao Zhuo, Fuan Wen, and Hehe Fan. Hand-centric motion refinement for 3d hand-object interaction via hierarchical spatial-temporal modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2076–2084, 2024. 2
- [14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5
- [15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4
- [17] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022. 4
- [18] Kai Jia, Tengyu Liu, Mingtao Pei, Yixin Zhu, and Siyuan Huang. Primhoi: Compositional human-object interaction via reusable primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11491–11501, 2025. 2
- [19] 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*, 36:20067–20079, 2023. 2
- [20] 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*, pages 1737–1747, 2024. 4
- [21] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 947–957, 2024. 3
- [22] Jiye Lee and Hanbyul Joo. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9663–9674, 2023. 2
- [23] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 2, 3, 5, 6, 8
- [24] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pages 54–72. Springer, 2024. 1, 2, 3, 4, 5, 6, 7, 8

- [25] Lei Li and Angela Dai. Genzi: Zero-shot 3d human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20465–20474, 2024. 2
- [26] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 1
- [27] 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*, pages 1769–1782, 2025. 1
- [28] Wei Mao, Richard I Hartley, Mathieu Salzmann, et al. Contact-aware human motion forecasting. *Advances in Neural Information Processing Systems*, 35:7356–7367, 2022. 2, 4
- [29] Zichong Meng, Zeyu Han, Xiaogang Peng, Yiming Xie, and Huaizu Jiang. Absolute coordinates make motion generation easy. *arXiv preprint arXiv:2505.19377*, 2025. 2, 3
- [30] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27859–27871, 2025. 2
- [31] Abdullallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14424–14432, 2020. 7
- [32] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 4
- [33] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4332–4341, 2019. 4
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4
- [35] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 3
- [36] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3
- [37] Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao Chen, Ning Li, and Hong Qin. Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2024. 2
- [38] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems*, 34:251–266, 2021. 3
- [39] 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*, 2023. 2, 3
- [40] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8001–8013, 2023. 4
- [41] Yuan Wang, Yali Li, Xiang Li, and Shengjin Wang. Hsi-gpt: A general-purpose large scene-motion-language model for human scene interaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7147–7157, 2025. 2
- [42] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–444, 2024. 2
- [43] 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*, pages 11176–11186, 2025. 1, 2, 3, 4, 5, 6, 8
- [44] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3
- [45] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023. 3
- [46] Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, et al. Interact: Advancing large-scale versatile 3d human-object interaction generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7048–7060, 2025. 1, 4
- [47] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for

- physics-based human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12266–12277, 2025. [1](#)
- [48] Haiwei Xue, Xiangyang Luo, Zhanghao Hu, Xin Zhang, Xunzhi Xiang, Yuqin Dai, Jianzhuang Liu, Zhensong Zhang, Minglei Li, Jian Yang, et al. Human motion video generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. [1](#)
- [49] Mengqing Xue, Yifei Liu, Ling Guo, Shaoli Huang, and Changxing Ding. Guiding human-object interactions with rich geometry and relations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22714–22723, 2025. [2](#)
- [50] Jie Yang, Xuesong Niu, Nan Jiang, Ruimao Zhang, and Siyuan Huang. F-hoi: Toward fine-grained semantic-aligned 3d human-object interactions. In *European Conference on Computer Vision*, pages 91–110. Springer, 2024. [1](#)
- [51] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16284–16295, 2024. [1](#)
- [52] Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*, pages 246–263. Springer, 2024. [2](#)
- [53] 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*, pages 16010–16021, 2023. [4](#)
- [54] Jinlu Zhang, Yixin Chen, Zan Wang, Jie Yang, Yizhou Wang, and Siyuan Huang. Interactanything: Zero-shot human object interaction synthesis via llm feedback and object affordance parsing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7015–7025, 2025. [2](#)
- [55] Jianrong Zhang, Hehe Fan, and Yi Yang. Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17592–17602, 2025. [2](#)
- [56] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2430–2449, 2023. [1](#)
- [57] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. Parco: Part-coordinating text-to-motion synthesis. In *European Conference on Computer Vision*, pages 126–143. Springer, 2024. [2](#)

Decoupled Generative Modeling for Human-Object Interaction Synthesis

Supplementary Material

Overview

In this supplementary material, we provide additional details and analyses to complement the main paper. Section A summarizes the experimental setup, including key implementation details and training and inference configurations. Section B defines the evaluation protocols and metrics, while Section C introduces our reconstruction guidance for improving hand-object contact and foot grounding. Section D describes DynaPlan for collision-aware dynamic planning, and Section E reports additional experiments, including an ablation on the generator loss weight, comparisons under privileged waypoint supervision, and an oracle trajectory study. Sections F to J detail the loss landscape visualization, the implementation of the OMOMO variants, the user study design, extended qualitative results, and limitations and future directions.

A. Experimental Setup

A.1. Implementation Details

Our models are implemented in the PyTorch deep learning framework. All experiments are conducted on a single NVIDIA RTX 3090 GPU with 24 GB of memory, and training requires approximately 25 GPU hours. We optimize the generators with the Adam optimizer [5] using a learning rate of 1×10^{-4} and a batch size of 32, where each training sample is a sequence with $T = 120$ frames. During training, input sequences shorter than T are zero-padded to match the length. The transformer-based denoising network consists of 4 attention heads and 4 layers. We train the model for 580k steps and report results using the checkpoint that achieves the best validation performance. The diffusion process uses 1,000 noising steps during training, and we employ the standard DDPM [4] sampling procedure at inference time.

For adversarial training of the discriminator, we use the same optimizer configuration as for the generators, but set the learning rate to 2×10^{-5} . The discriminator is updated once for every generator update. The loss balancing weights for the trajectory generator, action generator, and forward kinematics objectives are fixed to $\lambda_{TG} = 0.1$, $\lambda_{AG} = 1.0$, and $\lambda_{FK} = 1.0$, respectively. The adversarial loss weight λ_G is adjusted within the range $[0.01, 0.05]$ during training.

In addition, to prevent error accumulation from the trajectory generator (TG) propagating into the action generator (AG), we provide clean trajectories for both the human and the object as part of the noisy input to AG during training. In other words, AG is conditioned on ground-truth trajec-

tories while the remaining components are corrupted with noise, which is inspired by the teacher forcing strategy [12] commonly used to mitigate error accumulation. Note that during training TG receives noisy inputs for all frames except the start and goal conditions. At inference time the trajectories predicted by TG are then used as inputs to AG.

A.2. Inference Runtime and Resource Consumption

Methods	Inference time (s)	Inference VRAM (GB)
CHOIS [8]	2.00	1.9
HOIFHLI [13]	162.71	5.2
Ours (DecHOI)	3.41	2.1

Table 1. Inference runtime and GPU memory usage for CHOIS [8], HOIFHLI [13], and our DecHOI.

In this section, we report the computational cost of DecHOI and the baselines used in our comparison. DecHOI employs two lightweight denoising networks as specialized experts for the decoupled generative modeling of trajectories and interactions. This design incurs a modest increase in inference time compared to CHOIS [8], while remaining significantly more efficient than HOIFHLI [13], which relies on a generation pipeline with multi-stage for grasp generation and refinement and therefore has substantially higher inference time.

We further observe that DecHOI requires comparable or even lower VRAM usage than the prior models. Considering the qualitative and quantitative improvements, the additional inference cost of DecHOI is modest relative to the overall computational budget. Moreover, when taking into account the manual effort required to annotate intermediate waypoints for the baselines, DecHOI offers a more practical solution in realistic deployment scenarios.

B. Evaluation Details

For an accurate and fair comparison, all trajectories are expressed in a common global world frame, and all protocols and thresholds used for metric computation follow the same settings as CHOIS [8].

B.1. Condition Matching

We measure T_s and T_e as the Euclidean distances between the predicted and ground-truth object centroids at the start frame and at the final target position, respectively. Since our generator uses only the start and goal points, we evaluate condition matching solely with T_s and T_e .

B.2. Human Motion Quality

Foot height. H_{feet} is the mean distance from the feet to the floor. To obtain the floor level z_{floor} , we identify frames where the toe joints move below a small speed threshold (quasi-static stance), cluster their toe heights, and take the lowest cluster as z_{floor} . All foot z -coordinates are then measured relative to this floor level when computing H_{feet} and foot sliding.

Foot sliding. For the ankles and toes, we accumulate horizontal displacement between consecutive frames, but only when each joint is near the floor, as:

$$\sum_{t=0}^{T-1} \mathbb{1}(z_{j,t} < H_j) d_{j,t} (2 - 2^{z_{j,t}/H_j}), \quad (1)$$

where $\mathbb{1}(\cdot)$ is the indicator function, $z_{j,t}$ is the height of joint j at frame t , H_j is the joint specific height threshold that defines the near floor region, and $d_{j,t}$ is the horizontal displacement between frame t and $t + 1$. The index j runs over the left and right ankles and toes. The accumulated displacement is then averaged over time and across the four joints.

R-Precision. For each text-motion pair, we compute feature vectors for the text and candidate motions, then rank the motions by cosine similarity to the text, following [2, 8]. R_{prec} (top-3) is the fraction of pairs for which the ground-truth motion appears in the top-3 ranked motions, averaged over all evaluation pairs.

FID. We compute motion features for all generated and real sequences, fit a Gaussian to each set, and use the Fréchet Inception Distance (FID) between the two Gaussians as the FID score [3]. Lower FID indicates that the distribution of generated motions is closer to that of real motions.

DIV. For each model, we collect motion features from all generated results, sample random feature pairs, and average their Euclidean distances to obtain DIV, following [2]. We report a single DIV value per model, where values closer to the DIV of real data indicate more realistic diversity.

B.3. Interaction Quality

Contact metrics. For each frame, we assign a binary contact label based on the minimum distance between the hand joints and the object surface: if the closest distance from either hand joint to any object vertex is below 5 cm, the frame is marked as in contact. Given ground-truth contact labels, $C\%$ is the fraction of frames predicted in contact, C_{prec} and C_{rec} are precision and recall for the contact class, and C_{F1} is their F1 score.

Penetration metrics. We measure mesh interpenetration using signed distance fields (SDFs). At every frame, we map human and object vertices into the corresponding SDF

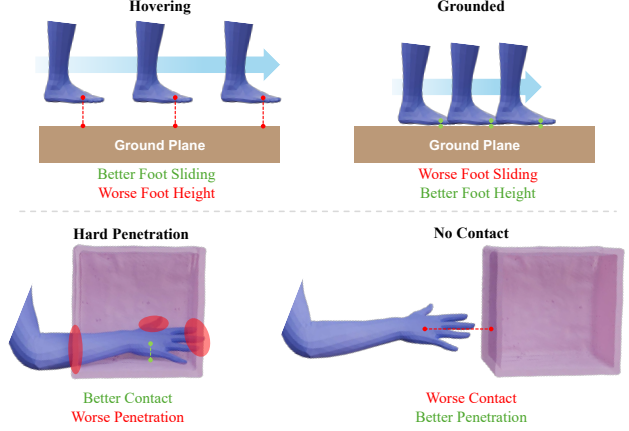


Figure 1. Visualization of trade-off relationships induced by the metric definitions. The top row illustrates the trade-off between foot related metrics, and the bottom row shows the relationship between contact and penetration metrics.

frame and sample their signed distances, with negative values indicating penetration. We average the negative signed distances into four scalars: P_{hand} and P_{body} for penetration of hand and full-body vertices into the manipulated object, and $P_{o \rightarrow s}$ and $P_{h \rightarrow s}$ for penetration of the object and the human into the static scene.

B.4. Ground Truth Difference

MPJPE. We report MPJPE as the mean Euclidean distance between predicted and ground-truth joint positions [9], averaged over all joints and frames in the same world frame.

Root and object translation error. T_{root} and T_{obj} are the mean Euclidean distances between predicted and ground-truth root joint positions and object centroids, respectively, averaged over all frames.

Object orientation error. O_{obj} measures the discrepancy between predicted and ground-truth object rotations. We compute the mean Frobenius norm of the difference between the two rotation matrices per frame and average over time.

B.5. Metrics Trade-off

Foot-Ground trade-off. Foot sliding (FS) assigns larger penalties when H_{feet} is small, as defined in Eq. B.2. Consequently, as illustrated in the top row of Fig. 1, even large horizontal displacements yield little or no foot sliding when the feet remain high above the ground (hovering). In contrast, when the feet stay very close to the floor, even small displacements are amplified in the FS score. This behavior induces a trade-off between H_{feet} and FS .

Contact-Penetration trade-off. As shown in the bottom row of Fig. 1, the contact-relative score and the penetration score also exhibit a trade-off relationship. Since con-

tact is determined based on the distance between each hand joint and the nearest object vertex, severe penetration can still yield a stable contact score. Conversely, when the hand does not approach the object, the penetration score can be favorable due to the absence of intersecting regions, while the contact score becomes poor. Therefore, metrics with such trade-off behavior should not be interpreted as a single score in isolation but evaluated jointly.

C. Inference-time Reconstruction Guidance

To obtain robust and plausible motion during inference, we apply reconstruction guidances following prior work [8]. First, to strengthen hand and foot contacts, we define a guidance term that regularizes the distance between the distal human joints and their nearest object vertices. For the hand joints this term is written as:

$$\mathcal{F}_{\text{cont}} = \|\mathbf{M}_l \odot |H_l - V_l|\|_1 + \|\mathbf{M}_r \odot |H_r - V_r|\|_1, \quad (2)$$

where H_l and H_r denote the left and right hand joint positions and V_l and V_r are the corresponding closest object vertices. The binary masks \mathbf{M}_l and \mathbf{M}_r indicate whether a hand joint is in contact with the object and are obtained by thresholding the hand to object distance with a threshold of 5 cm.

Second, to encourage stable foot grounding and a realistic stance, we introduce a guidance term that regularizes the distance between the feet and the floor plane. Given the positions of the left and right feet F_l and F_r , this term is defined as:

$$\mathcal{F}_{\text{feet}} = \|\min(F_l, F_r) - h\|_2, \quad (3)$$

where $h = 0.02$ m is a threshold for foot height estimated from the ground-truth motion. The error is computed only along the vertical coordinate.

D. DynaPlan

Inspired by [8], we project each 3D indoor scene from Replica [11] into a 2D grid layout, marking walkable regions as traversable and expanding non-walkable regions (e.g., walls and furniture) with a distance transform to enforce a human scale safety margin. *DynaPlan* then performs path planning for the agent and obstacle on this grid using A*. After the paths are optimized, the resulting trajectories are lifted back into the original 3D scene using the known metric scale [8, 11], and resampled for the agent and obstacles so they provide global plans and motion priors for full-body HOI generation in a consistent 3D coordinate frame. Fig. 2 visualizes the overall execution process of *DynaPlan*.

D.1. Obstacle Trajectory Modeling

Given obstacle start and goal points, the moving obstacle follows a deterministic A*-based trajectory, advancing by

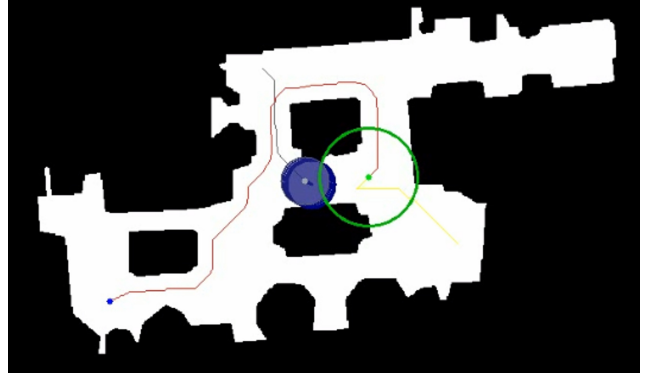


Figure 2. The green dot and circle denote the agent and R_{agent} , and the red curve shows the re-planned path. When the obstacle (gray) and its influence region (blue) enter R_{agent} , that area is given high cost in the risk map and dynamic re-planning is triggered.

one grid cell at each time step t . *DynaPlan* uses a pre-trained Social-STGCNN [10] to forecast future obstacle positions. Given the past T_{obs} obstacle positions, the model predicts T_{pred} future positions. These predicted positions do not control the obstacle. Instead, they are used only to construct predictive risk regions that *DynaPlan* uses for planning.

D.2. Collision Detection

Given an agent start and goal points, *DynaPlan* first computes an initial path using A*, minimizing the accumulated Euclidean step length ($\sum_{t=1}^T \|x_t - x_{t-1}\|_2$). In addition, we assign an influence radius to the agent and obstacle, and declare a collision whenever the obstacle’s radius along its path overlaps with that of the agent. For collision detection we use only the predicted obstacle trajectories. Once a collision is detected, the agent updates a risk map and calls A* again to re-plan its path.

D.3. Dynamic Re-planning

When a potential collision is detected, *DynaPlan* performs dynamic re-planning by updating the risk map based on the current and predicted obstacle positions. For a grid location x , the risk map $R(x)$ is:

$$R(x) = \exp\left(-\frac{\|x - o_t\|_2^2}{2\sigma^2}\right) + \sum_{k=1}^{T_{\text{pred}}} \gamma_k \exp\left(-\frac{\|x - \hat{o}_{t+k}\|_2^2}{2\sigma^2}\right). \quad (4)$$

We use a Gaussian weight with $\sigma = 1$ to assign smaller values to locations farther from the obstacle, and set $\gamma \in (0, 1)$ so that predicted future positions contribute less than the current position.

Methods	Condition Matching		Human Motion Quality		Interaction Quality		
	$T_s \downarrow$	$T_e \downarrow$	FS \downarrow	$R_{prec} \uparrow$	$C_{F1} \uparrow$	$P_{hand} \downarrow$	$P_{body} \downarrow$
$\lambda_G = 0.01$	1.69	7.60	0.45	0.70	0.69	0.55	0.56
$\lambda_G = 0.03$	1.69	7.83	0.40	0.69	0.67	0.53	0.54
$\lambda_G = 0.05$	1.98	7.87	0.42	0.69	0.66	0.53	0.56
Ours (DecHOI)	1.59	6.91	0.38	0.72	0.67	0.53	0.54

Table 2. Ablation comparing fixed and adaptive λ_G settings shows that the adaptive strategy yields a better overall balance.

Given a risk map R , the cost of a candidate path P is:

$$\mathcal{C}(P \mid R) = \sum_{t=1}^T (\|x_t - x_{t-1}\|_2 + \lambda_R R(x_t)). \quad (5)$$

We set $\lambda_R = 4.0$ to balance A* cost and risk cost.

To decide between detouring and waiting, *DynaPlan* evaluates candidate waiting time steps ranging from zero (detouring) to T_{pred} . For each waiting candidate, the agent holds its position for the corresponding duration while the obstacle progresses, and the risk map is updated accordingly. A waiting time penalty is applied to discourage long waits, by adding to the risk map cost an amount proportional to the number of waiting time steps.

DynaPlan selects the candidate with the lowest score, providing a lightweight, prediction-driven re-planning mechanism that enables collision-aware navigation.

E. Additional Experiments

E.1. Ablation Study on Generator Weight

In this section, we investigate the effect of the generator loss weight λ_G on the overall objective (Eq. 9) defined in the main paper. Tab. 2 compares fixed settings where λ_G is set to 0.01, 0.03, or 0.05 with an adaptive strategy that dynamically adjusts λ_G between 0.01 and 0.05 according to the performance of the discriminator \mathcal{D} . When \mathcal{D} becomes too strong, the adaptive scheme increases λ_G so that the generator places more emphasis on fooling the discriminator. When \mathcal{D} influence weakens, the scheme decreases λ_G to avoid over-regularization by the adversarial term.

With fixed values, training tends to drift toward an unbalanced regime where either the generator or the discriminator dominates, which often prevents the loss from converging. This lack of convergence not only degrades interaction quality but also harms the overall synthesis performance, leading to weaker condition matching and motion quality. In contrast, the adaptive weighting maintains a better balance in the adversarial training, stabilizes convergence, and yields the best performance across metrics. These results demonstrate that adaptive generator loss weighting effectively mitigates the inherent instability and bias of adversarial training and enables more reliable optimization.

E.2. Effect of Privileged Waypoint

In the main paper, we evaluate DecHOI and prior waypoint-based methods [8, 13] under a fair protocol where all models receive the same inputs. For completeness, we additionally report an experiment on the *FullBodyManipulation* [7] in which prior methods are evaluated under their original waypoint supervised configuration. In this experiment, DecHOI still operates without any intermediate waypoints, while the prior methods are given waypoints and are thus evaluated in a privileged information setting. The quantitative results of this comparison are summarized in Tab. 3. Note that for all baselines we report scores reproduced using the official released implementations, since several metrics are not reported in the CHOIS and HOIFHLI papers (e.g., DIV and P_{body}).

Both CHOIS [8] and HOIFHLI [13] benefit noticeably from waypoint supervision. In particular, human motion quality and GT difference metrics improve significantly. The GT difference metrics are highly sensitive to waypoints, and for T_{obj} we observe improvements of up to a factor of five, since waypoints directly specify object target positions. For this reason, GT difference metrics are not suitable for a direct comparison with DecHOI, which does not receive any waypoint input.

In contrast, DecHOI achieves comparable or better performance than the prior approaches on most metrics, even without privileged conditions. Considering the trade-off structure among metrics discussed in Sec. B.5, DecHOI shows slightly better overall performance in terms of H_{feet} and FS and also provides improved text alignment, realism, and diversity, enabled by our decoupled design. Moreover, given the inherent trade-off between contact score and penetration, DecHOI attains superior interaction quality than prior work, despite operating without intermediate waypoints while the baselines use them. These results suggest that DecHOI enables efficient and realistic interaction synthesis under weaker input priors and that the proposed adversarial training makes a clear contribution to improving interaction quality.

E.3. Oracle Trajectory Ablation

We evaluate an oracle setting that bypasses the trajectory generator (TG) and feeds the action generator (AG) with ground-truth object and human trajectories $\{\mathcal{T}_o^{GT}, \mathcal{T}_h^{GT}\}$ for all T frames. The AG architecture, checkpoint, and inference-time guidance are kept unchanged, and the conditioning interface follows Sec. 3.2 (see Fig. 2) in the main paper. This design disentangles the contributions of the TG and AG, allowing us to assess the standalone performance of the TG and to estimate the theoretical upper bound of AG performance within the decoupled pipeline. We conduct this experiment on the *FullBodyManipulation* [7], and report the results in Tab. 4.

Methods	Condition Matching		Human Motion Quality					Interaction Quality					GT Difference			
	$T_s \downarrow$	$T_e \downarrow$	$H_{feet} \downarrow$	$FS \downarrow$	$R_{prec} \uparrow$	$FID \downarrow$	$DIV \rightarrow$	$C_{prec} \uparrow$	$C_{rec} \uparrow$	$C_{F1} \uparrow$	$P_{hand} \downarrow$	$P_{body} \downarrow$	MPJPE \downarrow	$T_{root} \downarrow$	$T_{obj} \downarrow$	$O_{obj} \downarrow$
CHOIS [8]	1.72	6.95	4.49	0.35	0.66	0.69	8.37	0.80	0.64	0.68	0.59	0.60	15.30	24.43	13.35	0.98
HOIFHLI [13]	1.64	8.05	4.91	0.36	0.62	1.51	8.85	0.79	0.65	0.67	0.60	0.58	15.93	23.31	10.98	1.10
Ours (DecHOI)	1.59	6.91	4.42	0.38	0.72	0.33	8.86	0.80	0.64	0.67	0.53	0.54	15.27	25.47	22.96	0.86

Table 3. Quantitative comparison on the *FullBodyManipulation* [7] in the privileged waypoint supervised setting, where DecHOI operates without intermediate waypoints and CHOIS [8], HOIFHLI [13] receive sparse intermediate waypoints.

Methods	Human Motion			Interaction				
	$R_{prec} \uparrow$	$FID \downarrow$	$DIV \rightarrow$	$C_{prec} \uparrow$	$C_{rec} \uparrow$	$C_{F1} \uparrow$	$P_{hand} \downarrow$	$P_{body} \downarrow$
DecHOI	0.72	0.33	8.86	0.80	0.64	0.67	0.53	0.54
DecHOI (oracle)	0.66	0.15	8.97	0.82	0.65	0.69	0.56	0.57

Table 4. Oracle trajectory ablation study on the FullBodyManipulation [7]. DecHOI (oracle) denotes a variant in which the trajectory generator is replaced by ground truth trajectories.

Results and discussion. In this oracle configuration, metrics that quantify condition matching or deviation from the conditioning trajectories are less informative, since the AG is driven directly by ground-truth object and human paths. We therefore focus our analysis on human motion quality and interaction quality. The oracle setting generally achieves higher scores on most metrics, yet the gap relative to the standard DecHOI that uses predicted trajectories remains modest. On several metrics (R_{prec} , P_{hand} , and P_{body}), the standard DecHOI even slightly outperforms the oracle, and its overall performance is very strong when the typical trade-offs between contact and penetration are taken into account. This small performance difference indicates that the TG generates realistic trajectories that closely follow the distribution of ground-truth paths, and suggests that the AG is not tightly constrained by the TG performance but is capable of producing high quality interaction synthesis independently.

F. Loss Landscape Visualization

We visualize the loss landscape around our converged model by evaluating the training objective on a two-dimensional subspace of the parameter space, following [6]. We fix a batch \mathcal{B} drawn from the same distribution as the training data, so that variations in the loss reflect only changes in the model parameters. Let w_0 be the vector obtained by flattening and concatenating all trainable tensors of the final trained model.

To explore the neighborhood of w_0 , we construct two random perturbation directions. For each trainable tensor, we sample Gaussian noise with the same shape, normalize it to obtain a comparable scale across layers, and concatenate these tensors to form the first direction d_x . We repeat this procedure to obtain a second direction d'_y and orthogonalize it with respect to d_x using Gram-Schmidt to obtain the final direction d_y . This keeps the perturbations balanced across layers and avoids a single layer dominating the change.

We then form a grid of coefficients $\alpha, \beta \in [-r, r]$ and, for each pair (α, β) , construct a perturbed parameter vector $w(\alpha, \beta) = w_0 + \alpha d_x + \beta d_y$, reshape it back to the original tensor shapes, and evaluate the total loss on \mathcal{B} . The guidance term is omitted since it is used only at inference time and is not part of the training objective. Plotting the resulting loss values over the (α, β) grid as contour and surface plots provides a qualitative view of the local loss landscape around w_0 and indicates whether the optimization problem near the solution is relatively simple with a smooth basin or more complex with sharper curvature and narrower valleys.

G. OMOMO Implementation Details

We implement Lin-OMOMO and Pred-OMOMO within the original OMOMO framework [7], following the variant definitions introduced in CHOIS [8]. To match the conditioning used in our experiments, our OMOMO variants are provided only with the object start and goal states.

G.1. Lin-OMOMO

Lin-OMOMO uses the original OMOMO generator and training setup. The object trajectory is defined by linearly interpolating the object centroid between its start and goal positions and is used as the object translation input in the OMOMO conditioning stream. All centroid-related values are updated accordingly. Since Lin-OMOMO only interpolates the object centroids and keeps the orientation identical to the ground-truth, we omit O_{obj} for Lin-OMOMO from our reported results.

G.2. Pred-OMOMO

In Pred-OMOMO, the original OMOMO generator and training configuration are kept intact, while the object motion is obtained from CHOIS. Given the object’s start and goal conditions as input, CHOIS predicts the full trajectory of object centroids and rotations. We use this predicted object motion as the input in the OMOMO conditioning stream and update all related pose values to stay in sync with these predictions. Because both the trajectory of the object centroid and its rotations are supplied directly by CHOIS, the object-level metrics T_{obj} and O_{obj} for Pred-OMOMO are identical to those of CHOIS in our experiments.

Question. 60

Text Instruction: Lift the woodchair over your head, walk and then place the woodchair on the floor.

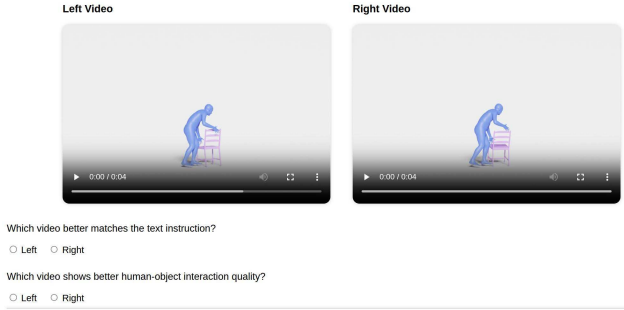


Figure 3. Example 2AFC interface in which participants read a text instruction and compare two anonymized clips to judge text alignment and interaction quality.

H. User Study Details

We conducted a two-alternative forced-choice (2AFC) perceptual study to evaluate two criteria of synthesized human-object interactions: Text Alignment (how faithfully a clip follows the natural-language instruction) and Interaction Quality (the perceived realism and plausibility of contact, including stable hand and foot contacts, few penetrations, and minimal temporal jitter).

For each comparison round (DecHOI-CHOIS [8] and DecHOI-HOIFHLI [13]), we randomly sampled 30 text-scene pairs from the full test set without scene duplication, yielding 60 clips in total. To ensure fairness, we shuffled both the order of videos and the left-right placement of each method in every trial, and fixed all video durations to 4 seconds. As shown in Fig. 3, the interface presents the Text Instruction at the top and two anonymized videos. Participants could replay videos and answer two questions. They selected which clip better matched the instruction (Text Alignment) and which clip exhibited better human-object interaction quality (Interaction Quality). To avoid presentation bias, no method identifiers or cues were provided.

Using the results in Fig. 8 of the main paper, participants preferred DecHOI in 71.5% of trials versus CHOIS and 67.5% versus HOIFHLI for Text Alignment, and in 69.0% of trials versus CHOIS and 63.5% versus HOIFHLI for Interaction Quality. These trends are consistent with our quantitative metrics and support the claim that decoupling trajectory and action generation improves both semantic faithfulness and perceptual interaction quality.

I. Additional Qualitative Results

We provide additional qualitative results that complement the examples in the main paper.

Fig. 4 extends Fig. 4 in the main paper by including additional scenes from *FullBodyManipulation* [7], comparing DecHOI against CHOIS [8] and HOIFHLI [13]. Across diverse instructions (e.g., lift-move-place, push and pull), De-

cHOI maintains stable hand-object contacts and grounded object trajectories, exhibiting less penetration and hovering. In contrast, the baselines show drift, temporal desynchronization between human and object, or incomplete contacts, especially during lift-place transitions and when objects change orientation.

Fig. 5 provides additional qualitative results that complement Fig. 5 in the main paper, showcasing *3D-FUTURE* [1] objects that are unseen during training. DecHOI maintains precise hand-object contact locations, more coherent placement motion, and consistent text to motion alignment, whereas CHOIS often produces mesh intersections or unstable object poses. The qualitative patterns align with the quantitative improvements reported for condition matching, motion stability, and contact reliability on unseen shapes.

Fig. 6 complements Fig. 7 in the main paper with additional long-horizon dynamic scenes from *DynaPlan*. We visualize collision events together with the corresponding re-planned direction. The dashed arrows indicate the original direction of motion, which would be the optimal path in free space, yet the agent re-routes around obstacles instead of strictly following this direction. These behaviors demonstrate that obstacle detection and dynamic planning are applied effectively and suggest that the framework can scale to more complex scenarios.

J. Limitations & Future Work

Our framework focuses on synthesizing human-object interactions that involve manipulating rigid objects. In real environments, however, many objects are deformable or articulated. Handling such cases requires contact regions that adapt over time to object parts whose positions change, and in some scenarios it is necessary to model underlying physical effects such as friction, inertia, and gravity. Addressing these challenges would require more sophisticated representations of articulated structure and dynamics. Incorporating recent articulation estimation networks as object priors is a promising direction, and systematically extending our approach to deformable and articulated objects constitutes an important avenue for future work.

In addition, the proposed *DynaPlan* module currently supports dynamic planning and interaction synthesis for a single manipulated object and a single moving obstacle. Real-world scenarios often involve more complex environments, such as crowds of agents and multiple objects being manipulated simultaneously. Scaling to such settings may require path planners that go beyond classical A* and can reason about many interacting agents. Moreover, supporting multiple manipulated objects would likely depend on constructing richer datasets, for example by combining motion capture with procedural composition of multi object interactions. Exploring these extensions would be highly valuable for broadening the applicability of our framework.

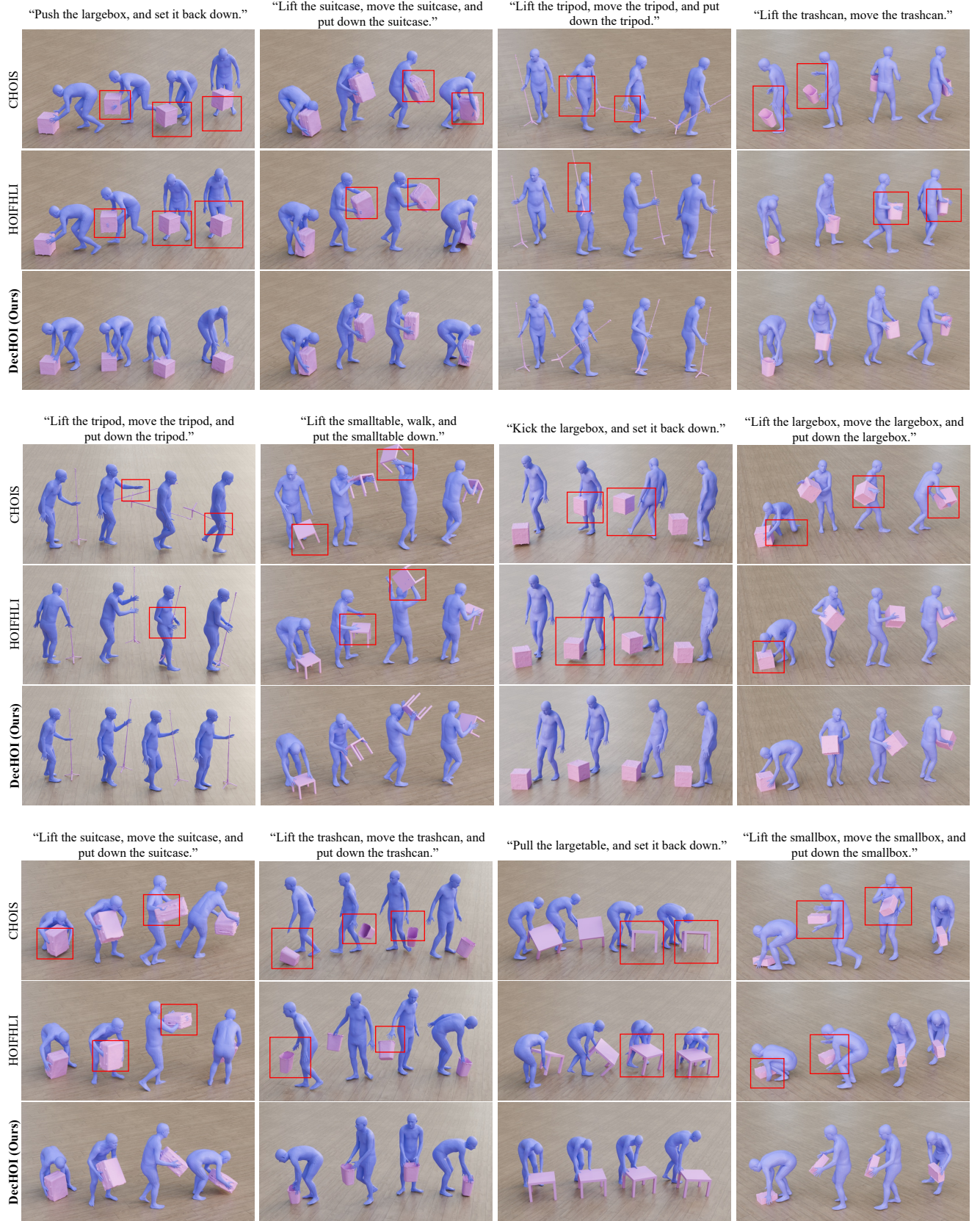


Figure 4. Additional qualitative comparison of DecHOI with CHOIS [8] and HOIFHLI [13] on the *FullBodyManipulation* [7]. Also DecHOI produces stable contacts, smooth motion, and accurate object trajectories, while prior methods show drift, penetration, or inconsistent coordination between human and object motions.

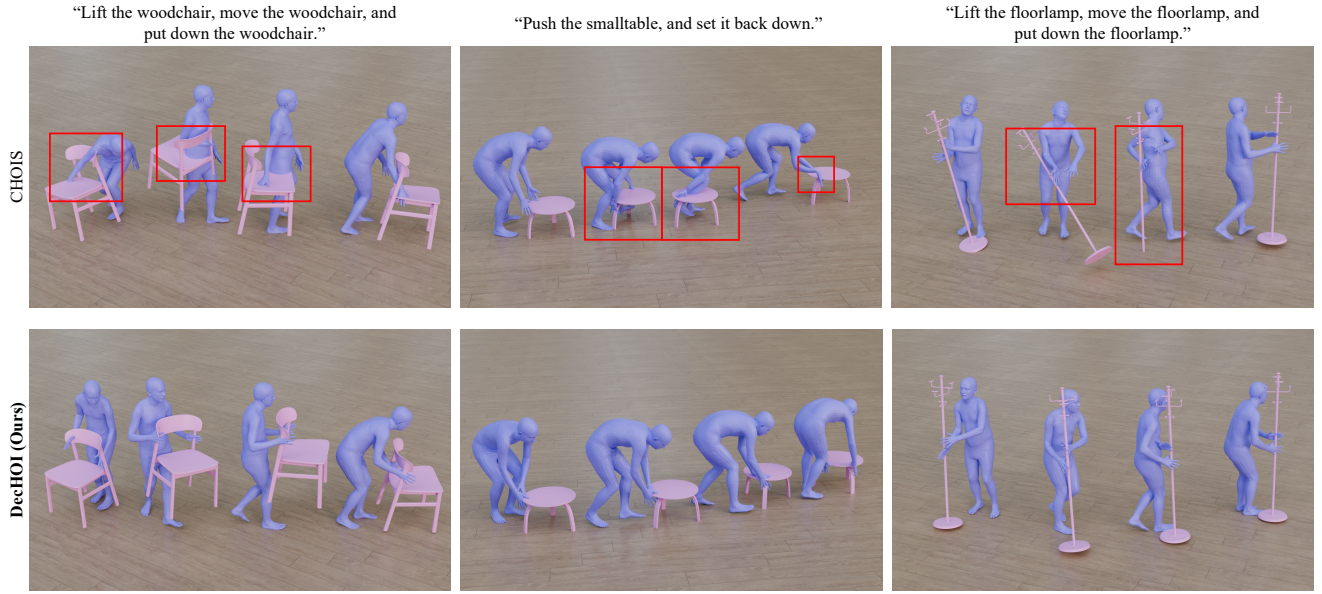


Figure 5. Additional qualitative comparison of DecHOI and CHOIS [8] on the 3D-FUTURE [1], demonstrating generalization to unseen object categories such as woodchair, smalltable, and floorlamp.

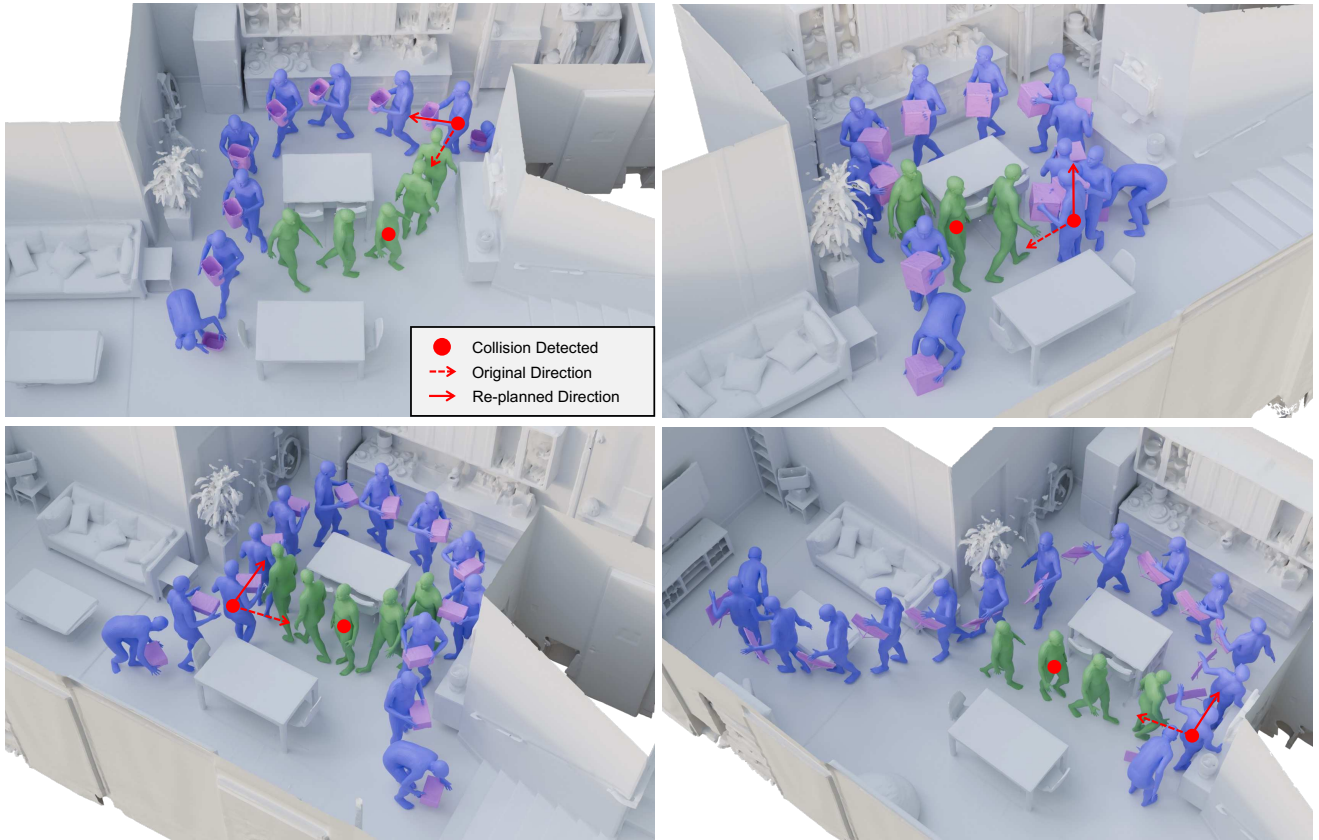


Figure 6. Visualization of DecHOI in long-sequence dynamic environments. The human agent (blue) adaptively re-plans its path when encountering a moving obstacle (green), often waiting briefly before detouring, and thereby maintaining goal-directed and collision-free motion.

References

- [1] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129(12):3313–3337, 2021. [6](#), [8](#)
- [2] 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*, pages 5152–5161, 2022. [2](#)
- [3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. [2](#)
- [4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [1](#)
- [5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. [1](#)
- [6] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018. [5](#)
- [7] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. [4](#), [5](#), [6](#), [7](#)
- [8] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pages 54–72. Springer, 2024. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [9] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. [2](#)
- [10] Abdullah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14424–14432, 2020. [3](#)
- [11] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. [3](#)
- [12] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989. [1](#)
- [13] 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*, pages 11176–11186, 2025. [1](#), [4](#), [5](#), [6](#), [7](#)