AW3D: ACTION PROTOTYPE LEARNING WITH WORLD MODELLING FOR TEXT-DRIVEN 3D HUMAN-OBJECT INTERACTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Realistic three-dimensional human–object interaction (HOI) animation is important for character animation, embodied simulation and humanoid control. Although existing text-conditioned HOI generators can produce physically plausible interactions, they remain limited in generating motions that faithfully follow textual instructions, often omitting required actions or executing them with incorrect timing or ordering. Here we present AW3D, an action-prototype framework with interaction world modelling, whose central idea is to explicitly organize high-level interaction intent into an ordered sequence of variable-duration action stages before dense motion generation. AW3D incorporates three key designs: *Joint-Embedding Action Prototype Learning*, which constructs a data-driven vocabulary of semantically meaningful interaction stages; *Distribution-Guided Action Sequence Decomposition*, which decomposes high-level intent into an ordered sequence of action prototypes and estimates their durations; and *Interaction World Modelling for Role-Guided Dense Motion Generation*, which adapts the resulting sequence to coordinate body, hand and object motions according to their roles in the interaction. Extensive experiments demonstrate that AW3D achieves state-of-the-art semantic alignment and motion quality, while remaining competitive in physical plausibility and interaction consistency. Code will be released upon acceptance.

1 INTRODUCTION

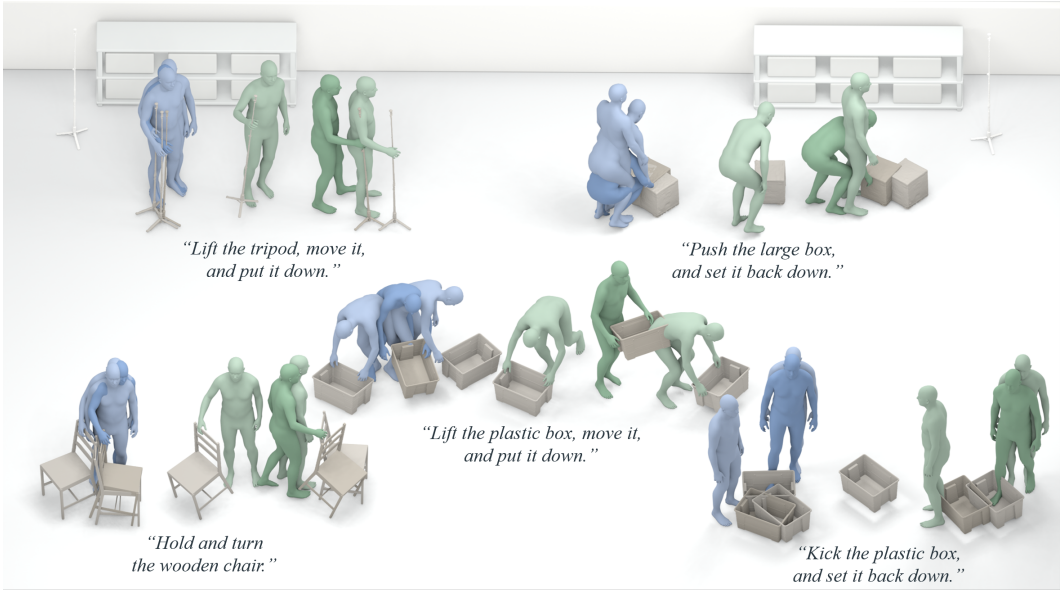

Figure 1: Text-to-motion alignment in HOI generation. Direct generation (blue) may omit or weakly realize action stages, whereas AW3D (green) better preserves their content and temporal progression.

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.

3D human–object interaction (HOI) synthesis seeks to generate coordinated human and object motions that accomplish a specified interaction. This capability is central to computer graphics, computer vision, and robotics, where virtual characters and embodied agents must not only move naturally but also manipulate objects in purposeful ways. It underpins a broad range of applications, including character animation, embodied simulation, and humanoid control (Li et al., 2023; 2025). Compared with isolated human-motion synthesis, HOI synthesis poses an additional challenge: the movements of the body, hands, and object must remain spatially and temporally coupled throughout the interaction. Recent advances in generative modelling, particularly diffusion models (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022), have made it possible to synthesize such interactions from natural-language descriptions, offering an intuitive and scalable interface for specifying complex interaction intent (Sahili et al., 2025; Zhu et al., 2023; Sui et al., 2026).

Existing methods have substantially improved the controllability and physical plausibility of generated interactions. Motion- or trajectory-conditioned approaches use full object motion sequences or sparse object waypoints to provide explicit spatial control (Li et al., 2023; 2025); contact- and guidance-aware methods exploit predicted contacts, affordances, or modality-wise denoising to encourage physically plausible interactions (Diller & Dai, 2024; Cha et al., 2024; Peng et al., 2025; Wang et al., 2026); and relation- or kinematics-aware models strengthen spatial coordination between human and object motion (Wu et al., 2024; Zeng et al., 2025). Hierarchical generators further separate coarse motion planning from fine-grained interaction synthesis (Pi et al., 2023; Jung et al., 2026). Although these methods can produce physically plausible motions with realistic contacts and spatial relationships, we find that they remain limited in faithfully following textual instructions. Consequently, required actions may be omitted, performed in the wrong order, or executed with incorrect timing.

We attribute these failures to the mismatch between the temporal abstraction of language and that of dense motion. A textual instruction typically describes an interaction as a short sequence of semantic stages—for example, reaching for an object, grasping it, lifting it, and placing it elsewhere. Dense motion generators, however, directly predict body, hand, and object states over hundreds of frames. Without an explicit intermediate representation of action stages, the same generation process must simultaneously determine *what* should happen, *when* each action should occur, and *how* it should be physically executed. As illustrated in Fig. 1, this entanglement makes it difficult to maintain the intended action order and timing over long sequences, particularly when different stages have substantially different durations.

To address this limitation, we propose AW3D, an action-prototype framework with interaction world modelling that explicitly organizes high-level intent into an ordered sequence of variable-duration action stages before dense motion generation. AW3D comprises three key designs. First, *Joint-Embedding Action Prototype Learning* discovers recurring interaction stages from continuous HOI sequences and constructs a data-driven vocabulary of action prototypes. Second, *Distribution-Guided Action Sequence Decomposition* translates high-level intent into an ordered prototype sequence with adaptive stage durations while retaining the uncertainty of plausible action transitions. Third, *Interaction World Modelling for Role-Guided Dense Motion Generation* adapts the planned prototypes to the evolving interaction state and provides role-specific guidance for coordinated body, hand, and object motion. Together, these designs decouple high-level action organization from low-level motion realization, enabling dense HOI generation to follow the intended action content, order, and timing.

We evaluate AW3D on InterAct, a unified benchmark containing diverse 3D HOI sequences with detailed textual annotations (Xu et al., 2025). Extensive experiments show that AW3D achieves state-of-the-art semantic alignment and motion-distribution quality while remaining competitive in physical plausibility and interaction consistency. Its action-prototype pathway also consistently improves multiple HOI generation backbones, particularly for long and compositionally complex interactions.

The main contributions of this work are as follows:

- We introduce AW3D, an action-prototype framework with interaction world modelling that bridges high-level language intent and continuous 3D HOI motion by explicitly organizing variable-duration action stages before dense motion generation.

- We develop a unified approach that learns a data-driven vocabulary of action prototypes, decomposes interaction intent into an ordered sequence with adaptive durations, and converts this sequence into role-specific guidance for coordinated body, hand, and object motion.
- Extensive experiments demonstrate state-of-the-art performance in semantic alignment and motion-distribution quality, competitive physical and interaction quality, and consistent improvements across different HOI generation backbones.

2 Related Work

3D Human–Object Interaction Synthesis. 3D human–object interaction (HOI) synthesis generates coordinated body, hand, and object motions from language, object geometry, or motion controls. Existing methods reduce motion ambiguity through object trajectories, sparse waypoints, or decoupled path planning (Li et al., 2023; 2025; Pi et al., 2023; Jung et al., 2026). Diffusion- and relation-based approaches further improve physical plausibility by modelling contacts, affordances, and geometric or kinematic dependencies (Peng et al., 2025; Diller & Dai, 2024; Xu et al., 2023; Wu et al., 2024; Xue et al., 2025; Zeng et al., 2025; Wang et al., 2026), while recent methods enhance long-horizon generation and compositionality through symbolic plans, reusable interaction primitives, or visual priors (Jiang et al., 2024; Wu et al., 2025; Jia et al., 2025; Geng et al., 2025; Cai et al., 2026). Despite increasingly realistic interactions, existing methods often struggle to faithfully translate textual instructions into motion, resulting in omitted, mistimed, or misordered actions. To address this, AW3D learns action prototypes from HOI sequences and explicitly organizes their temporal progression before dense motion generation, thereby improving text–motion alignment without requiring prescribed trajectories.

World Modelling Methods. World models capture dynamics in compact latent spaces by representing scenes as entities and relations (Kipf et al., 2020; Wu et al., 2023; Lei et al., 2025; Zhou et al., 2025). Masked representation learning predicts hidden image or video regions (He et al., 2022; Tong et al., 2022), whereas Joint-Embedding Predictive Architectures (JEPA) replace pixel reconstruction with latent prediction, progressing from visual representation learning to action-conditioned prediction and planning (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025). Recent work further integrates joint-embedding prediction with object-centric dynamics (Nam et al., 2026). Despite their success in modelling visual dynamics, these approaches focus primarily on fixed-rate observations or tracked entities, leaving the variable-duration interaction stages in HOI unexplored. We introduce the world-modelling paradigm into HOI synthesis and propose AW3D, which learns action prototypes from object-centric relations, predicts their condition-dependent temporal progression, and uses the resulting interaction states to guide dense HOI generation.

3 Method

3.1 Overview

Given a text instruction c, a canonical object point cloud $\mathbf{P}$, an SMPL-H body shape $\boldsymbol{\beta}$ (Loper et al., 2023; Romero et al., 2022), and a target length L, we synthesize a sequence of coordinated human and object motion. The human motion comprises joint positions $\mathbf{j}^p \in \mathbb{R}^{L\times 52\times 3}$ and hand rotations $\mathbf{j}^{r_h} \in \mathbb{R}^{L\times 30}$; the object trajectory comprises translations $\mathbf{o}^t \in \mathbb{R}^{L\times 3}$ and continuous 6D rotations $\mathbf{o}^r \in \mathbb{R}^{L\times 6}$ (Zhou et al., 2019). We represent each frame and the full sequence as

$$\mathbf{x}_t = \left[\mathbf{j}^p_{t,1}; \ldots; \mathbf{j}^p_{t,52}; \mathbf{j}^{r_h}_t; \mathbf{o}^r_t; \mathbf{o}^t_t\right] \in \mathbb{R}^{195}, \qquad \mathbf{X} = \mathbf{x}_{1:L} \in \mathbb{R}^{L\times 195}. \tag{1}$$

Generating $\mathbf{X}$ directly from text requires a model to infer both the high-level interaction sequence and its frame-level realization. This is particularly challenging for compositional instructions: a generated motion may be locally plausible yet omit, reorder, or prematurely terminate a required action stage. We address this problem by introducing an intermediate action-prototype space that separates *what should happen* from *how it should move*.

Our framework comprises three stages (Fig. 2). First, we discover reusable action prototypes from object-centric human–object relation states. Second, a text-conditioned planner predicts a temporally ordered distribution over these prototypes. Third, an interaction world model adapts the plan to the

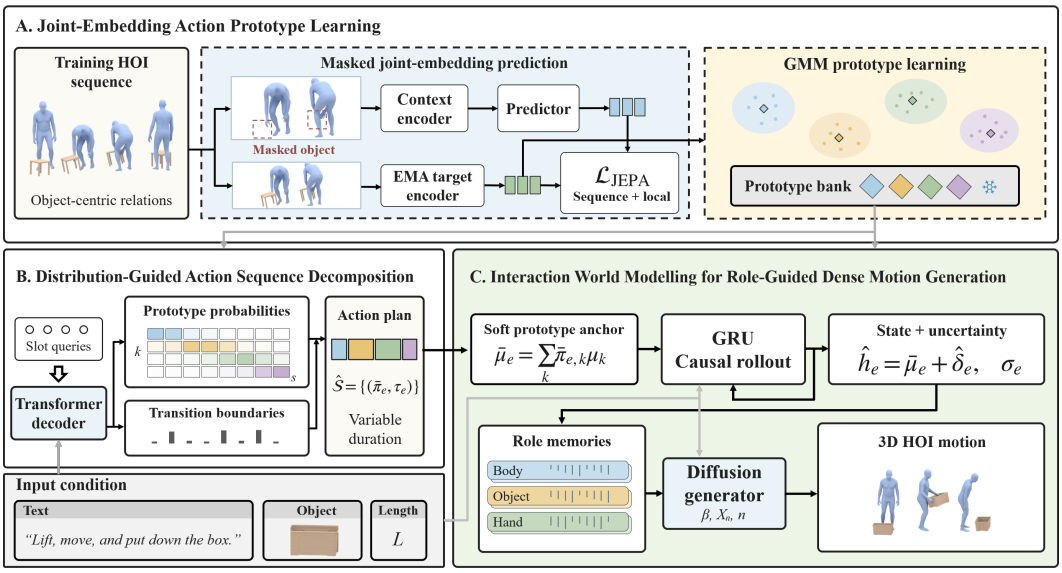

Figure 2: Overview of AW3D. (A) Joint-Embedding Action Prototype Learning discovers recurring object-centric interaction stages and organizes them into a shared prototype bank. (B) Distribution-Guided Action Sequence Decomposition predicts soft prototype distributions and transition boundaries, determining which stages occur, in what order, and for how long. (C) Interaction World Modelling predicts how the planned stages evolve under the current HOI condition and transforms their states into role-specific guidance for dense motion generation.

current object and interaction history, producing role-specific guidance for dense motion generation. Let $\mathbf{q} = \Phi_{\text{txt}}(c)$ and $\mathbf{g} = \Phi_{\text{obj}}(\mathbf{P})$ denote the text and object representations. The complete process is

$$\begin{aligned} \boldsymbol{\Pi} &= \mathcal{P}_\psi(\mathbf{q}, \mathbf{g}, L), \qquad \widehat{\mathcal{S}} = \mathcal{D}(\boldsymbol{\Pi}, L), \\ \widehat{\mathcal{H}} &= \mathcal{W}_\omega(\widehat{\mathcal{S}}, \mathbf{q}, \mathbf{g}, L), \qquad \mathbf{M} = \mathcal{U}(\boldsymbol{\Pi}, \widehat{\mathcal{H}}, L), \\ \mathbf{X} &\sim p_\theta(\cdot \mid c, \mathbf{P}, \boldsymbol{\beta}, L, \mathbf{M}) \,. \end{aligned} \tag{2}$$

Here, $\boldsymbol{\Pi}$ contains prototype and boundary predictions, $\widehat{\mathcal{S}}$ is the decoded soft action-stage plan, $\widehat{\mathcal{H}}$ contains its condition-specific interaction states, and $\mathbf{M}$ provides frame-level guidance to the motion generator. This hierarchy makes the temporal structure of the instruction explicit before synthesizing detailed kinematics. Next, we introduce each key design in detail.

3.2 Joint-Embedding Action Prototype Learning

Our first objective is to obtain intermediate states that are compact enough for sequence planning while remaining sufficiently informative to constrain physical interaction. To this end, we learn action prototypes from object-centric relational dynamics rather than raw global poses.

We first transform the human joints into the object coordinate system to remove variations caused by global object placement. The joints are then grouped into functional body regions, with the object treated as an additional entity. For entity n at frame t, we construct a descriptor $\mathbf{r}_t^{(n)} \in \mathbb{R}^D$ that encodes its object-relative position and velocity, distance to the object surface, soft contact, and object motion. Coordinated changes in these signals define stage boundaries, partitioning each sequence into intervals $\mathcal{E}_e = [a_e, b_e)$ with coherent interaction dynamics. Averaging the descriptors within each interval yields $\mathbf{F} \in \mathbb{R}^{E \times N \times D}$, where E and N denote the numbers of stages and entities, respectively. The coordinate transformation, contact function, and boundary extraction procedure are detailed in Appendix A.1.

Directly clustering these descriptors tends to emphasize appearance-level similarity while overlooking the dependencies between body regions and the object. We therefore learn stage representations through masked latent prediction. Our joint-embedding predictive architecture (Assran et al., 2023;

Bardes et al., 2024; Nam et al., 2026) consists of a context encoder f_ξ, an exponential-moving-average target encoder $f_{\bar{\xi}}$, and a predictor g_ϕ. During training, we mask one active entity across all interaction stages and predict its latent trajectory from the remaining entities. This prediction task requires the model to infer how the masked entity participates in the interaction, rather than merely reconstructing its observed coordinates.

To capture relational dynamics at complementary temporal scales, the architecture comprises two prediction branches. The sequence branch models dependencies across stages and thereby captures the progression of the interaction. In parallel, the shared local branch predicts the masked entity independently within each stage, emphasizing the contact and relational patterns that define stage identity. We optimize

$$\mathcal{L}_{\mathrm{JEPA}} = \mathrm{SL1}\Big(\widehat{\mathbf{h}}_{\mathrm{seq}}, \mathrm{sg}(\mathbf{h}_{\mathrm{seq}})\Big) + \lambda_{\mathrm{loc}}\,\mathrm{SL1}\Big(\widehat{\mathbf{h}}_{\mathrm{loc}}, \mathrm{sg}(\mathbf{h}_{\mathrm{loc}})\Big)\,, \tag{3}$$

where sg stops gradients through the target encoder.

After training, we freeze the target encoder and aggregate its local entity features to obtain a stage embedding $\mathbf{h}_e^*$. We then organize the learned embedding space using a Gaussian mixture model:

$$p(\mathbf{h}^*) = \sum_{k=1}^{K} \kappa_k\, \mathcal{N}(\mathbf{h}^*; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,. \tag{4}$$

The component means $\{\boldsymbol{\mu}_k\}_{k=1}^K$ form the action-prototype bank. Without imposing a fixed action vocabulary, each prototype represents a recurring human–object relational state, such as approaching, establishing contact, manipulating, or releasing. Together, these prototypes provide the discrete structure required to plan a complete interaction.

3.3 Distribution-Guided Action Sequence Decomposition

The prototype bank defines a vocabulary of interaction states, but it does not specify which states are needed to fulfill a particular instruction or how they should evolve over time. We therefore learn a planner that translates the text instruction, object geometry, and target motion length into an ordered sequence of action prototypes.

A text encoder, an object encoder, and a length embedding first integrate the three conditioning signals into a shared memory:

$$\mathbf{C} = \Phi_c(\mathbf{q}, \mathbf{g}, L). \tag{5}$$

Human–object interactions may contain different numbers of stages, each with a variable duration. To represent this temporal structure without assuming a fixed number of events, we rasterize the prototype labels of stages discovered from training motions onto a normalized low-rate temporal grid. A long stage therefore spans multiple adjacent slots, preserving its coarse duration, while slots beyond the requested motion length are masked. Let y_s denote the prototype label assigned to slot s, and let

$$y_s^b = \mathbf{1}[y_s \neq y_{s+1}] \tag{6}$$

indicate whether a stage transition occurs after that slot. Further implementation details are provided in Appendix A.3.

Given the condition memory $\mathbf{C}$, a set of learned slot queries predicts both what interaction state should occur at each temporal location and when the interaction should transition to the next stage. Specifically, each query attends to $\mathbf{C}$ and produces prototype logits $\mathbf{u}_s \in \mathbb{R}^K$ together with a boundary logit b_s. The planner is trained over all valid slots using

$$\mathcal{L}_{\mathrm{plan}} = \mathcal{L}_{\mathrm{CE}}^{\mathrm{smooth}}(\mathbf{u}, \mathbf{y}) + \lambda_b \mathcal{L}_{\mathrm{BCE}}(\mathbf{b}, \mathbf{y}^b), \tag{7}$$

where the first term supervises the interaction state at each slot and the second encourages accurate localization of stage transitions.

Although the most likely prototype provides a discrete stage label, committing to a hard assignment would discard uncertainty among semantically related interaction states. We therefore retain the calibrated prototype posterior

$$\pi_{s,k} = \mathrm{softmax}\left(\frac{1}{T|\mathcal{I}_{e(s)}|}\sum_{s' \in \mathcal{I}_{e(s)}} \mathbf{u}_{s'}\right)_k, \tag{8}$$

where T is a temperature parameter and $\mathcal{I}_{e(s)}$ is the boundary-decoded segment containing slot s. This distribution preserves plausible alternatives when the intended interaction state is ambiguous and provides smoother conditioning for the subsequent world model.

The predicted boundaries group the valid slots into $\widehat{E}$ contiguous segments, denoted by $\{\mathcal{I}_e\}_{e=1}^{\widehat{E}}$. Each segment shares one prototype posterior across its slots; we record this distribution and its normalized temporal extent:

$$\overline{\boldsymbol{\pi}}_e = \boldsymbol{\pi}_s \quad (s \in \mathcal{I}_e), \qquad \boldsymbol{\tau}_e = \left(\tau_e^{\text{start}}, \tau_e^{\text{end}}, \tau_e^{\text{dur}}\right). \tag{9}$$

The resulting plan is expressed as

$$\widehat{\mathcal{S}} = \{(\overline{\boldsymbol{\pi}}_e, \boldsymbol{\tau}_e)\}_{e=1}^{\widehat{E}}. \tag{10}$$

In this way, the planner decomposes a global textual instruction into an explicit sequence of interaction stages, specifying not only their identities and ordering but also their durations and associated uncertainty. This structured plan establishes the temporal progression of the interaction; the next step is to adapt these prototype distributions to the geometry and motion context of the specific object.

3.4 Interaction World Modelling for Role-Guided Dense Motion Generation

The planned prototypes describe the semantic progression of an interaction, but they remain shared abstractions: the same manipulation prototype may require different hand configurations and object trajectories depending on the object geometry, linguistic context, and preceding interaction states. To bridge this gap, we introduce an interaction world model $\mathcal{W}_\omega$ that evolves the planned sequence into condition-specific latent states, which are subsequently transformed into role-specific guidance for dense motion generation.

For each segment e, the prototype posterior defines a soft anchor, while a predicted innovation adapts this anchor to the current interaction:

$$\overline{\boldsymbol{\mu}}_e = \sum_{k=1}^{K} \overline{\pi}_{e,k} \boldsymbol{\mu}_k, \qquad \widehat{\mathbf{h}}_e = \overline{\boldsymbol{\mu}}_e + \widehat{\boldsymbol{\delta}}_e. \tag{11}$$

Here, $\overline{\boldsymbol{\mu}}_e$ retains the shared semantics of the planned prototype distribution, whereas $\widehat{\boldsymbol{\delta}}_e$ captures the object- and context-dependent realization. Let $\mathbf{c}_0$ fuse the text, object, and target-length conditions. The world model performs a causal rollout over the planned segments:

$$\begin{aligned} \mathbf{z}_e &= \mathrm{GRU}\Big(W_\mu \overline{\boldsymbol{\mu}}_e + W_\delta \widehat{\boldsymbol{\delta}}^{\mathrm{n}}_{e-1} + W_\tau \boldsymbol{\tau}_e + \mathbf{c}_0, \mathbf{z}_{e-1}\Big), \\ \left[\widehat{\boldsymbol{\delta}}^{\mathrm{n}}_e, \log \boldsymbol{\sigma}^2_e\right] &= g_\delta(\mathbf{z}_e), \qquad \boldsymbol{\ell}_e = g_p(\mathbf{z}_{e-1}). \end{aligned} \tag{12}$$

The normalized innovation is converted to $\widehat{\boldsymbol{\delta}}_e$ using fixed training-set statistics, as detailed in Appendix A.4. Feeding each normalized prediction into the next segment allows the rollout to depend on the interaction realized so far, rather than treating individual stages independently. Meanwhile, the prior logits $\boldsymbol{\ell}_e$ predict the current planned prototype from the preceding latent state, and $\boldsymbol{\sigma}^2_e$ captures uncertainty in the continuous realization.

During training, each predicted segment is temporally matched to a frozen target embedding $\mathbf{h}^*_e$, from which we derive the corresponding target innovation. The world model is optimized with

$$\mathcal{L}_{\text{world}} = \mathcal{L}_{\text{NLL}} + \lambda_r \mathcal{L}_{\text{res}} + \lambda_t \mathcal{L}_{\text{temp}} + \lambda_p \mathcal{L}_{\text{proto}} + \lambda_s \mathcal{L}_{\text{state}}. \tag{13}$$

The objective jointly models heteroscedastic innovation uncertainty, reconstructs the target innovation, matches its segment-to-segment changes, predicts the planned prototype from the preceding latent state, and aligns the realized states with their targets. Complete loss definitions and the temporal matching procedure are provided in Appendix A.4.

The resulting rollout captures interaction dynamics at the segment level, whereas motion synthesis requires frame-level guidance. Moreover, the body, hands, and object realize complementary aspects of the same interaction state. We therefore upsample the rollout and project it into role-specific memories. For each planning slot s, we take its soft prototype anchor $\overline{\boldsymbol{\mu}}_s = \overline{\boldsymbol{\mu}}_{e(s)}$ from the decoded segment and estimate the confidence of the plan using its normalized posterior entropy:

$$\alpha_s = 1 - \frac{H(\boldsymbol{\pi}_s)}{\log K}. \tag{14}$$

A concentrated posterior thus provides stronger guidance, whereas an ambiguous plan is introduced more conservatively.

Let $s(t)$ map frame t to its planning slot and $e(s)$ map each slot to its decoded segment. For each role $r \in \{\mathrm{body}, \mathrm{object}, \mathrm{hand}\}$, we construct the frame-level memory as

$$\begin{aligned}\mathbf{u}_{t,r} = A_\mu \overline{\boldsymbol{\mu}}_{s(t)} + A_{\delta,r}\widehat{\boldsymbol{\delta}}_{e(s(t))} + A_{q,r}\left[\rho_{e(s(t))}, \eta_{e(s(t))}\right] \\ + \mathbf{e}^{\mathrm{phase}}_{s(t)} + \mathbf{e}^{\mathrm{role}}_{r}, \qquad \mathbf{m}_{t,r} = \alpha_{s(t)} A_o \,\mathrm{LN}(\mathrm{GELU}(\mathbf{u}_{t,r}))\,.\end{aligned} \tag{15}$$

Here, ρ_e denotes the prior probability of the dominant planned prototype, and η_e is the mean log inverse standard deviation of the predicted innovation. The shared projection A_μ maintains a common interaction anchor across roles, while $A_{\delta,r}$ and $A_{q,r}$ allow the body, object, and hands to respond differently to the realized state and its reliability. The phase and role embeddings further encode temporal position and stream identity.

These memories condition the corresponding body, object, and hand token streams in the dense diffusion generator. Given a noisy motion sequence $\mathbf{X}_n$ at diffusion step n, the generator predicts

$$\widehat{\mathbf{X}}_0 = \mathcal{G}_\theta(\mathbf{X}_n, n, c, \mathbf{P}, \boldsymbol{\beta}, L, \mathbf{M})\,. \tag{16}$$

This design establishes a continuous path from language-level intent to detailed motion: the planner first identifies the ordered interaction structure, the world model adapts that structure to the current object and interaction history, and the dense generator realizes it through coordinated body, hand, and object trajectories. By resolving semantic progression before frame-level synthesis, AW3D preserves the structure of complex instructions while retaining the kinematic detail required for physically plausible interaction.

4 Experiments

4.1 Experimental Setup

Dataset. We conduct experiments on InterAct (Xu et al., 2025), a unified benchmark with fine-grained text annotations that includes BEHAVE (Bhatnagar et al., 2022), OMOMO (Li et al., 2023), and other HOI datasets. We follow its official split, SMPL-H conversion (Romero et al., 2022), and sequence filtering. Every learned component of AW3D is trained on the official training split. Appendix B.1 provides the complete data and evaluation protocols.

Baselines. We compare with four representative HOI generators under a unified text-conditioned setting. HOI-Diff (Peng et al., 2025) uses affordance-guided diffusion, while LIGHT (Wang et al., 2026) is a text-driven whole-body HOI model evaluated with and without sampling guidance. CHOIS (Li et al., 2025) originally requires additional object-motion controls; we remove these controls and condition it on the text prompt, object geometry, and target length. Text2HOI (Cha et al., 2024) is designed for hand–object motion; we adapt its output representation to full-body HOI and train its static contact-map estimator on InterAct. Thus, all methods receive the same task-level conditions and use the same processed split and evaluator.

Metrics. Following standard text-to-motion evaluation (Guo et al., 2022), we assess text alignment with R-Precision and Multimodal Distance (MM Dist), motion distribution with FID and Diversity, and physical interaction with FSR, Pene, Contact, and frame-wise contact precision, recall, and F1. R-Precision uses a batch size of 256. Appendix B.3 defines all metrics and their directions.

Implementation. We optimize the three modules sequentially. To match the conditions available at inference, the dense generator is trained with five-fold out-of-fold plans predicted by the action sequence model. At evaluation, the action sequence model is refitted to the complete training split. Architectures, hyperparameters, and optimization schedules are provided in Appendix B.2.

4.2 Main Results

We first compare AW3D with representative text-conditioned HOI generators. Table 1 shows that AW3D delivers the strongest overall semantic fidelity and distribution quality. It consistently outperforms all baselines across the three R-Precision measures while achieving the lowest FID

Table 1: Quantitative comparison on InterAct. Bold and underline denote the best and second-best generated results. For Diversity and Contact, values closer to Ground Truth are better.

Method	R-Precision↑			FID↓	MM Dist↓	Div.→	FSR↓	Pene↓	Contact→	Interaction↑		
	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
Ground Truth	.598±.001	.831±.001	.906±.002	.000±.000	1.575±.001	7.843±.015	.100±.237	.0884±.0002	.202±.000	1.000	1.000	1.000
HOI-Diff	.253±.007	.406±.003	.514±.006	.953±.086	3.128±.037	7.464±.003	.072±.001	**.043±.012**	.056±.007	**.772±.029**	.332±.022	.403±.017
Text2HOI	.258±.014	.422±.007	.526±.008	.744±.060	3.109±.003	7.751±.049	**.042±.015**	.441±.155	**.140±.010**	.764±.040	**.642±.029**	**.652±.024**
CHOIS	.305±.001	.478±.007	.597±.001	.679±.006	2.866±.014	7.695±.015	.044±.002	.109±.043	.086±.010	.744±.002	.451±.033	.507±.033
LIGHT w/o guidance	.404±.007	.607±.001	.717±.007	.234±.034	2.901±.029	7.686±.041	.085±.200	.148±.035	.119±.004	.719±.005	.544±.009	.571±.006
LIGHT w/ guidance	.428±.000	.643±.001	.748±.007	.171±.044	2.789±.028	7.752±.053	.094±.230	.175±.026	.132±.005	.700±.003	.598±.009	.605±.001
AW3D (ours)	**.485±.003**	**.723±.003**	**.826±.001**	**.101±.014**	**2.401±.015**	**7.852±.030**	.088±.224	.149±.009	.126±.003	.756±.026	.603±.008	.630±.003

Table 2: Transfer of AW3D across three HOI generators on InterAct. Bold denotes the better result within each pair. For Diversity and Contact, values closer to 7.843 and 0.202 are better.

Method	R-Precision↑			FID↓	MM Dist↓	Div.→	FSR↓	Pene↓	Contact→	Interaction↑		
	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
CHOIS	.305±.001	.478±.007	.597±.001	.679±.006	2.866±.014	7.695±.015	**.044±.002**	.109±.043	.086±.010	.744±.002	.451±.033	.507±.033
CHOIS + AW3D	**.333±.003**	**.522±.009**	**.640±.009**	**.559±.007**	**2.645±.009**	**7.726±.005**	.069±.007	**.088±.014**	**.113±.005**	**.800±.004**	**.535±.042**	**.590±.041**
Text2HOI	.258±.014	.422±.007	.526±.008	.744±.060	3.109±.003	**7.751±.049**	.042±.015	.441±.155	**.140±.010**	**.764±.040**	**.642±.029**	**.652±.024**
Text2HOI + AW3D	**.291±.001**	**.479±.005**	**.601±.012**	**.684±.088**	**2.755±.025**	7.697±.051	**.037±.013**	**.238±.026**	.136±.008	.756±.022	.601±.066	.633±.041
LIGHT	**.404±.007**	.607±.001	.717±.007	.234±.034	2.901±.029	7.686±.041	**.085±.200**	**.148±.035**	.119±.004	.719±.005	.544±.009	.571±.006
LIGHT + AW3D	.403±.012	**.615±.003**	**.726±.016**	**.131±.014**	**2.854±.014**	**7.728±.137**	.096±.230	.154±.005	**.130±.007**	**.737±.014**	**.578±.027**	**.606±.026**

and MM Dist, with Diversity closest to that of the ground-truth motions. AW3D also improves contact recall and contact F1 over LIGHT with guidance while remaining comparable on the other physical metrics. We attribute this result to the interaction world model, which transforms high-level prototypes into context-dependent interaction states, and to the role-specific guidance that coordinates human and object motion during dense generation. Together, sequence organization and role-specific guidance improve semantic consistency while preserving physical plausibility.

To assess transfer across dense generators, we integrate the action-prototype pathway into CHOIS, Text2HOI, and the original LIGHT architecture while preserving each backbone's training and inference protocol. As shown in Table 2, the proposed pathway improves most semantic-alignment and motion-distribution metrics across all three generators, with particularly consistent gains in retrieval accuracy, FID, and MM Dist. It also improves or maintains comparable interaction quality across different motion representations, conditioning mechanisms, and contact models. These consistent gains show that action prototypes provide a transferable temporal prior for organizing complex HOI sequences and supplying context-aware guidance during dense generation.

Figures 3 and 4 further show how this advantage appears in generated motion. In the cross-method comparison, AW3D more consistently completes the instructed progressions of sitting, crossing the lower legs, and standing up, as well as lifting, moving, and replacing a tripod. The paired comparison holds the dense generator fixed and isolates the action-prototype pathway. With this pathway, the model produces a clear table displacement and completes the ordered kick–lift–move–place interaction with the trashcan. The corresponding baseline weakly realizes or omits these transitions. These examples show that action-prototype decomposition and interaction world modelling improve text–motion alignment by preserving the action stages and their temporal dependencies throughout dense HOI generation.

Table 3: Long-sequence evaluation on 413 InterAct test motions ($200 < L \leq 300$).

Method	R@3 ↑	FID↓	MM Dist↓	Div.→	FSR↓	Pene↓	Contact→	Interaction↑		
								C_{prec}	C_{rec}	C_{F1}
Ground Truth	.934±.000	.000±.000	1.662±.000	7.868±.187	.083±.196	.079±.000	.213±.000	1.000±.000	1.000±.000	1.000±.000
CHOIS	.663±.009	.950±.069	3.425±.029	7.918±.080	.114±.007	.112±.026	.098±.002	.765±.002	.559±.011	.599±.008
CHOIS + AW3D	**.744±.008**	**.551±.031**	**3.122±.027**	**7.827±.197**	**.111±.008**	**.109±.011**	**.119±.004**	**.797±.006**	**.634±.013**	**.668±.016**
Text2HOI	.595±.007	.795±.011	3.582±.050	**7.749±.119**	.092±.004	**.281±.018**	.143±.003	.767±.003	.695±.009	.690±.010
Text2HOI + AW3D	**.717±.003**	**.763±.009**	**3.160±.017**	7.713±.085	**.074±.002**	.395±.026	**.163±.005**	**.812±.004**	**.739±.004**	**.737±.002**
LIGHT	.728±.009	.404±.041	3.115±.045	7.722±.173	**.070±.166**	.124±.006	**.136±.002**	**.823±.012**	**.686±.007**	**.711±.004**
LIGHT + AW3D	**.753±.009**	**.266±.019**	**3.082±.038**	**7.796±.188**	.084±.209	**.111±.004**	.135±.004	.821±.004	.680±.012	.709±.015

Table 4: Ablation study under the same experimental conditions. Bold and underline denote the best and second-best results. For Diversity and Contact, values closer to 7.843 and 0.202 are better.

Variant	R@3 ↑	FID↓	MM Dist↓	Div.→	FSR↓	Pene↓	Contact→	Interaction↑		
								C_{prec}	C_{rec}	C_{F1}
No prototype pathway	.823±.010	.124±.014	2.402±.005	**7.848±.174**	.085±.212	.160±.001	.125±.005	.755±.018	.597±.009	.627±.004
Reconstruction-based prototype learning	.820±.007	.113±.012	2.413±.002	7.778±.022	.085±.226	.153±.016	**.127±.002**	.750±.001	.592±.013	.621±.011
Shuffled prototype sequence	**.833±.005**	.119±.005	**2.395±.005**	7.827±.004	.083±.219	.159±.014	.125±.001	.743±.010	.579±.002	.605±.001
Static prototype memory	.826±.000	.109±.017	2.400±.015	7.745±.040	**.081±.211**	.161±.003	.125±.005	.754±.002	**.603±.011**	.622±.007
AW3D (full)	.826±.001	**.101±.014**	2.401±.015	7.852±.030	.088±.224	**.149±.009**	.126±.003	**.756±.026**	**.603±.008**	**.630±.003**

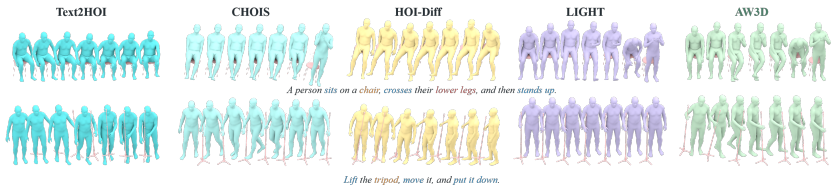

Figure 3: Qualitative comparison with existing HOI generators on InterAct. AW3D more consistently realizes the complete ordered action sequence.

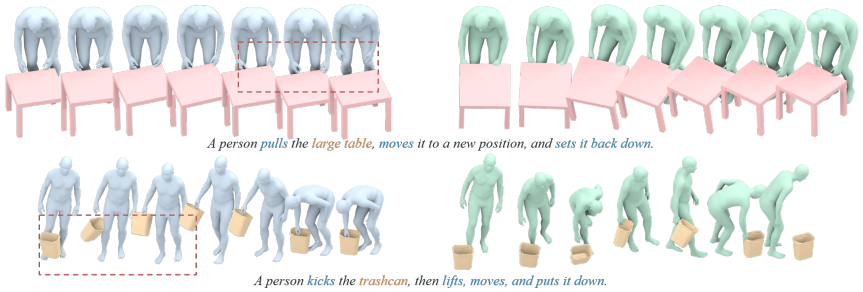

Figure 4: Controlled comparison of the same generator without and with the action-prototype pathway. Left: baseline. Right: AW3D. Red boxes mark incomplete stages.

Because AW3D explicitly models the correspondence between an instruction and its ordered action progression, we expect its advantages to become more pronounced for long-horizon generation, where temporal dependencies are harder to preserve. We therefore evaluate three generators and their AW3D-enhanced variants on the 413 sequences longer than 200 frames in the official test set. As shown in Table 3, AW3D consistently improves the overall balance among semantic alignment, motion-distribution quality, and interaction fidelity across different architectures. These results confirm that action-prototype sequence modelling provides an effective temporal prior for maintaining coherent, instruction-aligned interactions over extended horizons.

4.3 Ablation Study

To assess the contribution of each component, we conduct controlled ablations of the action-prototype pathway. Table 4 shows that replacing relation-state prediction with reconstruction degrades motion and interaction quality; shuffling prototypes while preserving their content and model capacity confirms the importance of temporal order; and using fixed prototype centers demonstrates the benefit of condition-specific state updates. Together, these results validate the complementary roles of relation-aware prototype learning, sequence decomposition, and interaction world modelling.

To examine the learned prototype space, Fig. 5 shows that the inferred boundaries align with contact transitions, while prototype assignments correlate more strongly with contact states than with temporal phases. The sequence model also outperforms a phase-only baseline in PlanScore, confirming that the learned space captures inter-action structure and ordered action progression. Further analyses are provided in Appendix C.

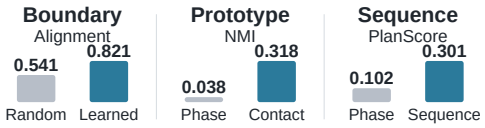

Figure 5: Diagnostics of the learned action-prototype pathway. NMI is normalized mutual information; PlanScore is the harmonic mean of valid-slot prototype accuracy and transition F1.

5 Conclusion

We presented AW3D, a framework that organizes text-driven 3D HOI at the action-stage level before dense motion synthesis. It learns recurring object-centric relation states as action prototypes, decomposes each instruction into a variable-duration action sequence, and adapts this sequence through interaction world modelling to guide body, hand, and object motion. Experiments show improved text–motion alignment, motion distribution quality, and interaction consistency, with consistent gains across HOI generators and long interaction sequences. Overall, AW3D provides a general approach to generating realistic and semantically aligned 3D human–object interactions.

AI USE STATEMENT

Generative AI tools were used to refine the wording of author-proposed research hypotheses and assist with translation. They also assisted with routine code editing and improving the readability of the manuscript. Generative AI tools were not used to design the research methodology or experiments, implement the proposed method, conduct qualitative or thematic analysis, or interpret experimental results. The remaining required-disclosure tasks were not applicable to this work. The authors verified the cited literature and supporting sources, tested AI-assisted code edits, checked all reported results against the original experiment records, and reviewed all AI-assisted text. The authors developed the method, designed and conducted the experiments, implemented the core approach, performed the analyses, and take full responsibility for the complete submission.

REPRODUCIBILITY STATEMENT

Section 3 defines AW3D and its training objectives. Appendix A provides the complete details of interaction-stage discovery, action-prototype learning, action sequence decomposition, interaction world modelling, and dense motion generation. Appendix B documents the data splits, preprocessing, training configurations, evaluation metrics, and model size. Appendices C and D provide the diagnostic and generalization protocols.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, June 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting Feature Prediction for Learning Visual Representations from Video. *Transactions on Machine Learning Research*, 2024. URL `https://mlanthology.org/tmlr/2024/bardes2024tmlr-revisiting/`.

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 (CVPR)*, pp. 15935–15946, June 2022.

Songjin Cai, Linjie Zhong, Ling Guo, and Changxing Ding. Vihoi: Human-object interaction synthesis with visual priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30686–30695, June 2026.

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1577–1585, June 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)*, pp. 19888–19901, June 2024.

Zichen Geng, Zeeshan Hayder, Wei Liu, and Ajmal Saeed Mian. Auto-regressive diffusion for generating 3d human-object interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3131–3139, 2025.

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)*, pp. 5152–5161, June 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

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 (ICCV)*, pp. 11491–11501, October 2025.

Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Hwanhee Jung, Seunggwan Lee, Jeongyoon Yoon, SeungHyeon Kim, Giljoo Nam, Qixing Huang, and Sangpil Kim. Decoupled generative modeling for human-object interaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2253–2263, June 2026.

Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1gax6VtDB.

Anson Lei, Bernhard Schölkopf, and Ingmar Posner. SPARTAN: A sparse transformer world model attending to what matters. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=uS5ch7GjZ4.

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

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 54–72, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72940-9.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-JEPA: Learning World Models through Object-Level Latent Masking. *arXiv e-prints*, art. arXiv:2602.11389, February 2026. doi: 10.48550/arXiv.2602.11389.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2903–2913, June 2025.

Huaijin Pi, Sida Peng, Minghui Yang, Xiaowei Zhou, and Hujun Bao. Hierarchical generation of human-object interactions with diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15061–15073, October 2023.

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 (ICCV)*, October 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

Ali Rida Sahili, Najett Neji, and Hedi Tabia. Text-driven motion generation: Overview, challenges and directions. *arXiv preprint arXiv:2505.09379*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Kewei Sui, Anindita Ghosh, Inwoo Hwang, Bing Zhou, Jian Wang, and Chuan Guo. A survey on human interaction motion generation. *Int. J. Comput. Vis.*, 134(3):113, 2026. doi: 10.1007/S11263-025-02582-5. URL https://doi.org/10.1007/s11263-025-02582-5.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093. Curran Associates, Inc., 2022. doi: 10.52202/068431-0732. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf.

Ziyin Wang, Sirui Xu, chuan guo, Bing Zhou, Jiangshan Gong, Jian Wang, Yu-Xiong Wang, and Liang-Yan Gui. Unleashing guidance without classifiers for human-object interaction animation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 46121–46141, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4c8ce3c63f6b66d6811c6d67c68e487b-Paper-Conference.pdf.

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

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)*, pp. 11176–11186, October 2025.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TFbwV6I0VLg.

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 (ICCV)*, pp. 14928–14940, October 2023.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. Interact: Advancing large-scale versatile 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7048–7060, June 2025.

Mengqing Xue, Yifei Liu, Ling Guo, Shaoli Huang, and Changxing Ding. Guiding human-object interactions with rich geometry and relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22714–22723, June 2025.

Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng, and Wei-Shi Zheng. Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12358–12369, June 2025.

Gaoyue Zhou, Hengkai Pan, Yann Lecun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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.

A Additional Method Details

This section provides the implementation details of the three-stage AW3D pipeline described in Section 3. We first discover variable-duration action stages from object-centric relation changes and organize recurring stages into an action-prototype space. The action sequence model then decomposes the test-time conditions into an ordered soft prototype sequence. Finally, the interaction world model adapts this sequence to the current interaction and forms role-specific guidance for dense HOI generation.

A.1 Object-Centric Interaction Stages

We partition the 52 SMPL-H joints into five non-overlapping human entities: torso, left leg, right leg, left arm and hand, and right arm and hand. The manipulated object forms the sixth entity. For each human entity, a 12-dimensional descriptor summarizes its object-local position and velocity, soft contact, surface distance, entity speed, and object motion. The object entity uses its local velocity, whole-body contact, and motion statistics. The resulting object-centric representation captures relation changes consistently across global scene placements.

Let $\mathbf{p}_{t,j}$ be the world-space position of joint j, and let $\mathbf{R}_t^o$ and $\mathbf{o}_t^t$ be the object rotation and translation. Under the row-vector convention, the object-local joint position is

$$\widetilde{\mathbf{p}}_{t,j} = (\mathbf{p}_{t,j} - \mathbf{o}_t^t)\mathbf{R}_t^o. \tag{17}$$

For the nearest distance $d_{t,j}$ from the joint to the canonical object surface, the continuous contact score is

$$a_{t,j} = \exp\left[-\frac{1}{2}\left(\frac{d_{t,j}}{\sigma}\right)^2\right]. \tag{18}$$

Let Δa_t, Δv_t, and Δo_t denote frame-wise changes in contact, relative motion, and object motion. With robust scale normalization ρ, transition saliency is

$$\chi_t = \rho(\Delta a_t) + \lambda_v \rho(\Delta v_t) + \lambda_o \rho(\Delta o_t). \tag{19}$$

After temporal smoothing, well-separated local maxima above a sequence-adaptive threshold define the stage boundaries. Table 5 summarizes the configuration used to construct these stages.

Table 5: Configuration for object-centric interaction-stage construction.

Parameter	Value	Parameter	Value
Entities	6	Features per entity	12
Soft-contact scale σ	0.05	Smoothing window	5 frames
Peak quantile	0.65	Minimum duration	10 frames
Stage-count cap	24	Saliency weights (λ_v, λ_o)	$(0.5, 0.5)$

Before saliency aggregation, each nonzero transition signal is divided by its median magnitude and clipped to $[0, 5]$. Local maxima above a sequence-specific quantile define candidate boundaries, subject to the minimum-duration constraint. The observed maximum is 21 stages, below the cap of 24. For the boundary diagnostic in Table 10, a boundary is aligned when it lies within three frames of an upper-quartile nonzero soft-contact change. The random control samples the same number of eligible positions from each sequence.

A.2 Joint-Embedding Action Prototype Learning

The joint-embedding predictor receives the object-centric descriptors with separate stage and entity axes. During training, one active entity tube is masked across all valid stages. The online context encoder predicts latent targets produced by an exponential-moving-average target encoder. Masked entities are sampled in proportion to their activity, with positive smoothing so that each example has a valid target. Sequence-level and stage-local losses are evaluated on masked valid tokens. Predicted and target embeddings are independently normalized before the smooth L_1 objective.

The stage-local branch shares the encoder weights and processes every stage independently. It predicts the masked entity from the remaining entity descriptors within the same stage, centering prototype

identity on relation content. We fit a 64-component diagonal Gaussian mixture model (GMM) to these local representations from all 74,225 training stages. Table 6 summarizes the joint-embedding and prototype-learning configuration.

Table 6: Joint-embedding action prototype configuration.

Parameter	Value	Parameter	Value
Input dimension	12	Latent width	256
Context layers	4	Predictor layers	2
Attention heads	8	Dropout	0
EMA momentum	$0.996 \rightarrow 1.0$	Local-loss weight	0.5
GMM components K	64	Planning slots S	30
GMM iterations	20	GMM batch size	8,192
Variance floor	10^{-4}	Training stages	74,225

A.3 Distribution-Guided Action Sequence Decomposition

At inference, the action sequence model uses only the conditions available to the dense HOI generator: T5 language features (Raffel et al., 2020), a BPS object descriptor (Prokudin et al., 2019), and normalized sequence length. These features form the condition memory of a Transformer decoder with 30 learned queries. Separate heads predict a 64-way action-prototype distribution and the transition probability between adjacent slots.

The 30 slots form a normalized low-rate grid over the maximum 300-frame horizon. Each slot spans ten frames on this grid and serves as a temporal support unit. Predicted boundaries merge adjacent units into variable-duration action stages. Each stage carries a soft prototype distribution, a confidence value, and a duration determined by the predicted transitions.

More generally, for S planning slots and maximum supported length $L_{\max}$, a sequence of length L uses

$$S_L = \min\left(S, \left\lceil \frac{LS}{L_{\max}} \right\rceil\right) \tag{20}$$

valid slots. The objective is evaluated over these slots, and the remaining slots are masked.

Prototype classification uses label-smoothed cross-entropy, and boundary prediction uses binary cross-entropy. We select the sequence model using PlanScore, the harmonic mean of valid-slot prototype accuracy and transition F1. A scalar temperature and the boundary threshold are estimated from held-out validation predictions. Five-fold cross-fitting generates out-of-fold plans for the training sequences, and a model fitted to the complete training split produces evaluation plans. The resulting soft sequence retains uncertainty among related prototypes while specifying the identity, order, and duration of the action stages supplied to the interaction world model.

A.4 Interaction World Modelling for Role-Guided Dense Motion Generation

The interaction world model aligns each decoded action stage with frozen joint-embedding states according to temporal overlap. It predicts the continuous innovation that adapts the shared prototype center to the current language, object geometry, and motion length. A GRU cell performs this prediction as a closed-loop rollout at the stage rate. Its hidden state is initialized from the fused condition, and the previous normalized innovation is zero for the first stage. At stage e, the model first predicts the prototype prior from the preceding hidden state. It then receives the current prototype center, stage timing, condition context, and previous innovation, and predicts a normalized innovation mean and diagonal log variance. The predicted innovation is passed to the next stage during both training and inference.

Let $\mathbf{h}_e^*$ be the overlap-weighted joint-embedding target for segment e. Its target innovation and normalized target innovation are

$$\boldsymbol{\delta}_e^* = \mathbf{h}_e^* - \overline{\boldsymbol{\mu}}_e, \qquad \boldsymbol{\delta}_e^{*,\mathrm{n}} = (\boldsymbol{\delta}_e^* - \mathbf{m}_\delta) \oslash \mathbf{s}_\delta, \tag{21}$$

where $\mathbf{m}_\delta$ and $\mathbf{s}_\delta$ are fixed training-set statistics and $\oslash$ denotes element-wise division. The residual head predicts the normalized mean $\widehat{\boldsymbol{\delta}}_e^{\mathrm{n}}$ and log variance $\ell_{e,d}^{\sigma} = \log \sigma_{e,d}^2$. The innovation used to

reconstruct the interaction state is

$$\widehat{\boldsymbol{\delta}}_e = \mathbf{m}_\delta + \mathbf{s}_\delta \odot \widehat{\boldsymbol{\delta}}_e^{\mathrm{n}}. \tag{22}$$

With valid-segment set $\mathcal{V}$ and interaction-state width D_h, the heteroscedastic Gaussian objective in normalized innovation space is

$$\mathcal{L}_{\mathrm{NLL}} = \frac{1}{2|\mathcal{V}|D_h} \sum_{e\in\mathcal{V}} \sum_{d=1}^{D_h} \Big[\exp(-\ell_{e,d}^{\sigma})(\widehat{\delta}_{e,d}^{\mathrm{n}} - \delta_{e,d}^{*,\mathrm{n}})^2 + \ell_{e,d}^{\sigma} + \log(2\pi) \Big]. \tag{23}$$

The auxiliary terms supervise the normalized innovation, its segment-to-segment change, the current prototype predicted from the preceding latent state, and the reconstructed state:

$$\mathcal{L}_{\mathrm{res}} = \mathrm{SL1}(\widehat{\boldsymbol{\delta}}^{\mathrm{n}}, \boldsymbol{\delta}^{*,\mathrm{n}}), \qquad \mathcal{L}_{\mathrm{temp}} = \mathrm{SL1}(\Delta\widehat{\boldsymbol{\delta}}^{\mathrm{n}}, \Delta\boldsymbol{\delta}^{*,\mathrm{n}}), \tag{24}$$

$$\mathcal{L}_{\mathrm{proto}} = \mathrm{CE}(\boldsymbol{\ell}, \mathbf{y}), \qquad \mathcal{L}_{\mathrm{state}} = 1 - \cos(\overline{\boldsymbol{\mu}} + \widehat{\boldsymbol{\delta}}, \mathbf{h}^*). \tag{25}$$

All auxiliary terms are averaged over valid segments, and $\mathcal{L}_{\mathrm{temp}}$ is evaluated only over valid adjacent segment pairs. For decoded segment e, let $k_e = \arg\max_k \overline{\pi}_{e,k}$ be its dominant planned prototype. The prior probability and mean log inverse standard deviation used by the role-specific memory are

$$\rho_e = \mathrm{softmax}(\boldsymbol{\ell}_e)_{k_e}, \qquad \eta_e = -\frac{1}{2D_h} \sum_d \log \sigma_{e,d}^2. \tag{26}$$

The stage-rate outputs are expanded to frames according to the decoded durations. A shared prototype anchor carries the action-stage identity. Role-specific projections combine the predicted innovation, reliability statistics, and learned role embeddings to form body, object, and hand memories. These memories guide the dense denoiser through a zero-initialized residual projection. During training, the complete prototype residual is removed independently for 20% of samples. This whole-sequence dropout keeps the original conditioning pathway active and improves robustness to uncertain sequence predictions.

For completeness, the forward diffusion process for role r at timestep n is

$$\mathbf{x}_n^r = \sqrt{\overline{\alpha}_n}\mathbf{x}_0^r + \sqrt{1-\overline{\alpha}_n}\boldsymbol{\epsilon}^r, \qquad \boldsymbol{\epsilon}^r \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{27}$$

B Experimental Details

B.1 Datasets and Evaluation Protocols

Primary InterAct protocol. The main evaluation follows the official InterAct split, sequence filtering, and SMPL-H conversion. Stage discovery and all learned components use only the official training motions. Every compared method is evaluated on the same filtered test set with identical preprocessing and evaluator settings.

Object-disjoint protocol. The unseen-object evaluation partitions complete object-instance groups. Every object and all its associated sequences belong to one partition, yielding a test set composed exclusively of unseen object instances. The split contains 7,465 training sequences over 119 objects and 2,349 test sequences over 22 unseen objects. Motion normalization and all learned components are estimated from the training partition.

Dataset-specific protocols. OMOMO and BEHAVE use their official training and test splits and the preprocessing protocol used by LIGHT (Wang et al., 2026). We learn a separate action-prototype space, action sequence model, interaction world model, normalization statistics, and dense generator for each dataset. All methods within a dataset use the same preprocessing and evaluation protocol.

B.2 Training Configuration

Table 7 lists the configuration of the final InterAct model. The joint-embedding prototype encoder, action sequence model, and interaction world model use AdamW with linear warm-up and cosine

decay. The dense generator follows the same optimizer and learning-rate schedule. All controlled runs use the same random seed.

Table 7: Training configuration for the final InterAct model.

Stage	Steps	Batch	Learning rate	Weight decay	Warm-up
Joint-embedding prototype encoder	25k	128	3×10^{-4}	0.05	5k
Diagonal GMM	20 EM	8,192	–	–	–
Action sequence model	60k max.	64	1×10^{-4}	0.05	2k
Interaction world model	50k max.	256	2×10^{-4}	0.02	2k
Dense HOI generator	300k max.	64	1×10^{-4}	0	2.5k

The action sequence model has width 256, four Transformer decoder layers, eight attention heads, dropout 0.1, and label smoothing 0.1. The boundary loss has unit weight. The interaction world model has width 384 and dropout 0.1. Its weights for residual reconstruction, temporal consistency, prototype prediction, and state alignment are 0.25, 0.15, 0.20, and 0.10. The world model used by the dense generator is selected on the validation split.

The dense denoiser has width 512, eight Transformer layers, and a maximum length of 300 frames. It predicts clean motion using a 500-step cosine diffusion process. The role-specific motion losses, condition dropout, and exponential moving average settings remain fixed across all controlled runs.

B.3 Evaluation Metrics

We use the InterAct evaluator and follow the text-to-motion protocol of Guo et al. (2022) and the physical metrics adopted by LIGHT (Wang et al., 2026). Let $\mathbf{e}_i^t$ and $\mathbf{e}_i^m$ denote the evaluator embeddings of the text and motion in the i-th paired sample. Let $\mathcal{X}^r$ and $\mathcal{X}^g$ denote real and generated motions. For a sequence with recorded length L_i, the evaluator uses the first $T_i = L_i - 4$ frames for contact and penetration. Foot skating follows the evaluator's masked fixed-length protocol defined below.

R-Precision. For a batch of B paired samples, we first construct the cross-modal distance matrix

$$D_{ij} = \left\| \mathbf{e}_i^t - \mathbf{e}_j^m \right\|_2 .$$

Each text query ranks all motion candidates in ascending order of D_{ij}. R-Precision at rank k is

$$R@k = \frac{1}{B} \sum_{i=1}^{B} \mathbb{I}\big[i \in \mathrm{TopK}_j(-D_{ij}, k) \big] .$$

It is the fraction of queries for which the paired motion appears among the k nearest candidates. We use $B = 256$ and report $R@1$, $R@2$, and $R@3$; higher values indicate stronger text–motion alignment.

Multimodal Distance. MM Dist measures the distance between each text and its paired generated motion in the shared evaluator space:

$$\mathrm{MMDist} = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{e}_i^t - \mathbf{e}_i^m \right\|_2 .$$

MM Dist directly measures the pairwise semantic discrepancy between each instruction and its generated motion. Lower values indicate closer agreement with the instruction and complement the batch-level ranking measured by R-Precision.

Fréchet Inception Distance. Let $(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ and $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ be the empirical mean and covariance of evaluator motion embeddings for $\mathcal{X}^r$ and $\mathcal{X}^g$. We compute

$$\mathrm{FID} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \mathrm{Tr}\Big(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_g)^{1/2} \Big) .$$

FID evaluates the discrepancy between the generated and real motion distributions in feature space. Lower values indicate that generated interactions better match the real distribution.

Diversity. From the generated motion embeddings, we independently draw K_d pairs of sample indices (a_l, b_l) and compute

$$\text{Diversity} = \frac{1}{K_d} \sum_{l=1}^{K_d} \left\| \mathbf{e}_{a_l}^m - \mathbf{e}_{b_l}^m \right\|_2 .$$

The evaluator uses $K_d = 300$ when more than 300 motions are available and $K_d = 50$ otherwise. The real-motion score is the reference: values below it can indicate mode contraction, and substantially higher values can reflect implausible variation.

Foot-Skating Ratio. For each foot f, let $\mathbf{p}_{t,f}$ be its 3D position and let $\mathbf{p}_{t,f}^{xz}$ denote its projection onto the ground plane. At 30 fps, the instantaneous horizontal speed is

$$v_{t,f} = 30 \left\| \mathbf{p}_{t+1,f}^{xz} - \mathbf{p}_{t,f}^{xz} \right\|_2 ,$$

and $\bar{v}_{t,f}$ is its average over a five-frame window. A foot is considered in ground contact across frames t and $t+1$ when both heights are below $h = 0.05$ m. A transition is marked as skating when at least one contacting foot has both $v_{t,f} > 0.5$ m/s and $\bar{v}_{t,f} > 0.5$ m/s. Let m_t be the evaluator mask for valid transitions in a sequence padded to $T_{\max}$ frames. The official evaluator computes

$$\text{FSR} = \frac{1}{T_{\max} - 1} \sum_{t=1}^{T_{\max}-1} m_t \mathbb{I}[\exists f :\ c_{t,f} = 1,\ v_{t,f} > 0.5,\ \bar{v}_{t,f} > 0.5] ,$$

where $c_{t,f}$ is the ground-contact indicator. Lower FSR indicates less foot sliding.

Penetration. Let $\mathbf{q}_{t,j}$ be human joint j expressed in the object rest frame at time t, and let $s(\mathbf{q})$ be the object signed-distance field, with negative values inside the object. For $J = 52$ joints, penetration is

$$\text{Pene} = \frac{1}{TJ} \sum_{t=1}^{T} \sum_{j=1}^{J} \left| \min(s(\mathbf{q}_{t,j}), 0) \right| .$$

The score averages the penetration depth of joints inside the object. Lower values indicate fewer or shallower interpenetrations. Following the reporting convention of the main tables, we multiply the mean value by 100.

Contact Rate. Let $\mathcal{O}_t$ be the posed object surface and let $\mathbf{j}_{t,j}$ be human joint j at frame t. With the common threshold $\tau_c = 0.05$ m, the joint-contact indicator is

$$c_{t,j} = \mathbb{I}\left[\min_{\mathbf{o} \in \mathcal{O}_t} \left\| \mathbf{j}_{t,j} - \mathbf{o} \right\|_2 < \tau_c \right] .$$

For a sequence of length T with $J = 52$ joints, we compute

$$\text{Contact} = \frac{1}{TJ} \sum_{t=1}^{T} \sum_{j=1}^{J} c_{t,j},$$

and average the result over test sequences. This metric measures how frequently the generated human approaches the object. The real-data contact rate provides the target because contact duration varies across actions.

Frame-Wise Hand–Object Contact. To assess whether contact occurs at the correct time, we reduce the 30 hand joints to one binary state per frame:

$$y_t = \mathbb{I}\left[\min_{j \in \mathcal{H}} \min_{\mathbf{o} \in \mathcal{O}_t} \left\| \mathbf{j}_{t,j} - \mathbf{o} \right\|_2 < \tau_c \right] ,$$

where $\mathcal{H}$ is the set of hand joints. We obtain y_t^r from the reference interaction and y_t^g from the generated interaction. For each sequence, true positives, false positives, and false negatives are accumulated over its evaluator horizon, after which

$$C_{\text{prec}} = \frac{TP}{TP + FP}, \qquad C_{\text{rec}} = \frac{TP}{TP + FN}, \qquad C_{F1} = \frac{2 C_{\text{prec}} C_{\text{rec}}}{C_{\text{prec}} + C_{\text{rec}}}.$$

The sequence-level scores are averaged over the test set. When neither the reference nor the generated sequence contains hand contact, the evaluator assigns one to all three scores; other zero-denominator cases receive zero. Precision penalizes spurious contact, recall penalizes missed contact, and C_{F1} summarizes their balance. Higher values indicate better temporal agreement with the reference hand–object interaction.

B.4 Model Size and Inference Efficiency

Table 8 reports the trainable parameter count of each component. The joint-embedding encoder is used only to construct the action-prototype space. At inference, the 6.419M-parameter action sequence model and the 5.656M-parameter interaction world model run once at the low stage rate. The role-guidance adapter adds 0.603M trainable parameters to the 59.514M-parameter dense generator.

Table 8: Trainable parameter count and inference role of each component.

Component	Trainable parameters	Used at inference
Joint-embedding prototype encoder	4.817M	No
Action prototype bank	0	Yes, frozen
Action sequence model	6.419M	Yes
Interaction world model	5.656M	Yes
Base dense generator	59.514M	Yes
Dense generator + role-guidance adapter	60.117M	Yes

Table 9 reports end-to-end sampling efficiency for a 300-frame interaction with batch size one on a single NVIDIA RTX 4090. Each pair follows the sampling schedule of its host generator. AW3D latency includes action sequence decomposition and interaction world modelling. Results are averaged over ten timed runs after two warm-up runs.

Table 9: Inference efficiency on a single RTX 4090. Time is reported as mean $\pm$ standard deviation. Peak memory excludes dataset-level caches.

Method	Time (s)↓	Seq./s↑	Peak mem. (GB)↓	Relative time↓
Text2HOI	0.435±0.001	2.301	0.505	1.00×
Text2HOI + AW3D	0.508±0.002	1.969	0.509	1.17×
CHOIS	0.428±0.004	2.335	0.422	1.00×
CHOIS + AW3D	0.509±0.002	1.966	0.425	1.19×
LIGHT	9.986±0.350	0.100	0.538	1.00×
LIGHT + AW3D	10.686±0.416	0.094	0.542	1.07×

Across the three host generators, AW3D increases end-to-end latency by 7.0–18.8% and peak memory by 0.003–0.004 GB. The action sequence model and interaction world model require 11.5 ms per sequence when measured separately. Dense motion sampling therefore remains the main inference cost, while the explicit action-stage pathway adds a modest computational overhead.

C Analysis of the Learned Action-Prototype Space

The component ablation in Section 4.3 measures the contribution of each module to generated motion. Here, we examine whether the learned intermediate variables support the intended action-stage decomposition. Let Z denote discrete prototype assignments and Y denote either temporal phase or contact state. We compute normalized mutual information (NMI) as

$$\mathrm{NMI}(Z, Y) = \frac{I(Z; Y)}{\sqrt{H(Z)H(Y)}},$$

where $I(\cdot;\cdot)$ is mutual information and $H(\cdot)$ is entropy. For action sequence prediction, let A_{proto} be prototype accuracy over valid slots and let F_{tr} be the F1 score of predicted stage transitions over valid adjacent slots. PlanScore is

$$\mathrm{PlanScore} = \frac{2A_{\mathrm{proto}}F_{\mathrm{tr}}}{A_{\mathrm{proto}} + F_{\mathrm{tr}}}.$$

Table 10: Diagnostics of the learned action-prototype space on InterAct. Bold denotes the relation- or sequence-aware result.

Diagnostic	Result	
Stage discovery		
Stages / sequence, median [P10, P90]	8 [5, 12]	
Duration (frames), median [P10, P90]	19 [11, 47]	
Boundary alignment↑	random: .541	learned: **.821**
Prototype bank		
Occupied / effective prototypes	64 / 56.91	
NMI	phase: .0385	contact: **.3182**
Sequence decomposition		
PlanScore↑	phase: .1015	sequence: **.3008**

Table 10 follows the action-prototype pathway from stage discovery to action sequence decomposition. The discovery process yields a median of eight stages per sequence and 19 frames per stage. Its boundaries align with relation changes more closely than matched random positions. All 64 prototypes are occupied, with 56.91 effective prototypes, and contact-state NMI is higher than temporal-phase NMI. The sequence model also raises PlanScore from 0.1015 to 0.3008 over a phase baseline. These results support the central design of AW3D: the intermediate space organizes relation-aware action stages and their temporal progression before dense motion synthesis.

D Additional Generalization and Dataset-Specific Evaluations

The main experiments evaluate AW3D on the unified InterAct benchmark. We extend this evaluation in two directions. The object-disjoint split tests whether action-stage organization transfers to unseen object geometry within InterAct. The OMOMO and BEHAVE experiments test whether the complete pathway can be learned independently from datasets with different motion and object distributions. Together, these settings examine whether organizing action content, order, and duration before dense motion generation remains effective across object instances and dataset-specific HOI distributions.

D.1 Generalization to Unseen Object Instances

Table 11 evaluates two matched generator pairs on the split described in Section B.1. Every component, including the action-prototype space, is learned from the object-disjoint training partition. The paired comparisons isolate the effect of action-stage organization when the test interactions involve unseen object instances.

Table 11: Object-disjoint evaluation on InterAct. Models in each pair use the same training and evaluation protocol.

Method	R@3↑	FID↓	MM Dist↓	Div.→	FSR↓	Pene↓	Contact→	Interaction↑		
								C_{prec}	C_{rec}	C_{F1}
Ground Truth	.593±.001	.000±.000	1.791±.004	7.558±.073	.030±.057	.0420±.0001	.255±.001	1.000±.000	1.000±.000	1.000±.000
LIGHT	.270±.006	4.437±.081	4.465±.003	**7.020±.036**	**.028±.063**	**.0718±.0060**	.108±.004	.877±.009	.548±.011	.627±.007
LIGHT + AW3D	**.283±.010**	**4.232±.065**	**4.403±.024**	6.992±.041	.031±.067	.0813±.0021	**.119±.004**	**.901±.010**	**.591±.014**	**.672±.007**
No prototype pathway	**.358±.006**	3.804±.037	**4.040±.012**	6.989±.052	.029±.064	**.0565±.0048**	.086±.003	.864±.021	.481±.010	.564±.003
AW3D (full)	.344±.004	**3.685±.004**	4.064±.010	**7.094±.010**	**.028±.063**	.0786±.0067	**.091±.003**	**.885±.022**	**.488±.011**	**.575±.010**

Across both matched pairs, AW3D reduces FID and improves all three interaction metrics on unseen objects. In the LIGHT pair, it further improves R@3, MM Dist, and Contact. These results show that the learned action-stage structure transfers to unseen object geometry and provides effective temporal guidance for coordinating human and object motion.

D.2 Dataset-Specific Evaluation on OMOMO

Table 12 evaluates all methods on the official OMOMO split with shared preprocessing and evaluation. The action-prototype space and every AW3D module are learned from the OMOMO training split. This setting tests whether variable-duration action stages can be recovered from OMOMO and used to organize its object-manipulation sequences before dense motion generation.

Table 12: Dataset-specific evaluation on OMOMO under the official split. Bold and underline denote the best and second-best generated results. For Diversity and Contact, values closer to Ground Truth are better. R-Precision uses batch size 256.

Method	R-Precision↑			FID↓	MM Dist↓	Div.→	FSR↓	Pene↓	Contact→	Interaction↑		
	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
Ground Truth	.313±.004	.560±.001	.733±.001	.000±.000	1.498±.000	7.464±.044	.012±.039	.0667±.0000	.263±.000	1.000	1.000	1.000
HOI-Diff	.280±.009	.479±.007	.646±.005	.844±.016	2.073±.009	7.430±.047	**.008±.002**	.1185±.0065	.164±.004	.914±.009	.771±.006	.814±.004
Text2HOI	.300±.001	.531±.005	.687±.001	.498±.027	1.840±.026	**7.464±.024**	.021±.002	.2753±.0102	.176±.002	.895±.012	.793±.015	.816±.012
CHOIS	.311±.003	.552±.012	.724±.004	.277±.014	1.677±.017	7.453±.078	.019±.000	.1473±.0079	.168±.003	.918±.002	.781±.015	.817±.013
LIGHT	.290±.009	.520±.022	.679±.012	.125±.017	1.784±.032	7.401±.002	.013±.044	**.1057±.0024**	.195±.002	.939±.009	.841±.002	.873±.006
LIGHT + AW3D	.299±.005	.521±.004	.689±.005	.153±.012	1.746±.027	7.419±.079	.014±.046	.1148±.0019	**.207±.001**	.941±.002	**.871±.002**	**.894±.002**
AW3D (full)	**.318±.000**	**.558±.004**	**.727±.005**	**.074±.001**	**1.531±.002**	7.483±.063	.014±.045	.1119±.0000	.191±.003	**.950±.002**	.853±.003	.886±.004

On OMOMO, the complete model achieves the strongest R-Precision, FID, and MM Dist among the evaluated generators. Adding AW3D to LIGHT also improves semantic retrieval and strengthens contact and interaction consistency. The matched and full-model results jointly show that organizing the requested action progression improves text–motion alignment and coordinated object manipulation under the OMOMO distribution.

D.3 Dataset-Specific Evaluation on BEHAVE

Table 13 applies the same protocol to BEHAVE. Its training split is used to learn an independent action-prototype space and the complete AW3D pipeline. The different interaction and object distribution provides a complementary test of whether explicit action-stage organization remains effective when learned from a dataset-specific HOI corpus.

Table 13: Dataset-specific evaluation on BEHAVE under the official split. Bold and underline denote the best and second-best generated results. For Diversity and Contact, values closer to Ground Truth are better. R-Precision uses batch size 256.

Method	R-Precision↑			FID↓	MM Dist↓	Div.→	FSR↓	Pene↓	Contact→	Interaction↑		
	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
Ground Truth	.566±.005	.916±.003	.930±.000	.000±.000	1.502±.016	6.798±.230	.178±.235	.0906±.0003	.207±.000	1.000	1.000	1.000
HOI-Diff	.105±.000	.221±.003	.289±.005	.967±.026	4.013±.105	7.077±.192	**.134±.003**	.1635±.0040	.134±.002	.670±.019	.554±.012	.562±.015
Text2HOI	.129±.032	.250±.043	.324±.000	2.799±.416	4.232±.014	7.264±.019	.146±.025	.1956±.0391	.123±.003	.701±.004	.652±.017	.635±.004
CHOIS	.139±.003	.268±.003	.346±.014	1.484±.138	4.058±.047	7.520±.237	.173±.010	**.1252±.0073**	.095±.002	.653±.022	.499±.033	.508±.035
LIGHT	.240±.008	.412±.003	.496±.016	.542±.021	3.415±.062	7.053±.466	.163±.228	.1365±.0010	.156±.004	.722±.015	.692±.003	.675±.000
LIGHT + AW3D	.238±.005	.416±.035	.504±.022	**.502±.013**	3.378±.032	6.892±.272	.160±.246	.1378±.0176	**.173±.009**	.727±.003	**.724±.017**	**.698±.011**
AW3D (full)	**.318±.008**	**.535±.005**	**.592±.003**	.727±.001	**2.950±.014**	**6.823±.290**	.166±.257	.1539±.0076	.143±.004	**.767±.011**	.661±.006	.672±.000

On BEHAVE, adding AW3D to LIGHT improves the overall balance of semantic retrieval, motion-distribution quality, and interaction consistency. The complete model achieves the strongest R-Precision and MM Dist among the evaluated generators, with Diversity closest to the reference distribution. Together with the OMOMO results, these comparisons show that the action-stage pathway can be learned from distinct HOI distributions and consistently improves instruction-aligned dense motion generation.

E Additional Qualitative Results

Aggregate metrics measure semantic alignment and interaction quality, but they do not directly show whether every instructed action stage appears in the generated sequence. Figures 7 and 8 therefore compare paired generations at matched time points. Each pair uses the same instruction, object condition, sequence length, and initial noise. The examples cover object transport and placement, reorientation, foot-guided manipulation, seated actions, ball interaction, bat swinging, and overhead lifting. Across these settings, AW3D more consistently completes the requested progression and coordinates human motion with the corresponding object response.

E.1 Perceptual User Study

The qualitative examples make action-stage progression visible. We further ask whether viewers perceive better instruction adherence and human–object coordination in complete animations. We conducted a blind, two-alternative forced-choice study on Credamo. Participants compared videos without and with AW3D. For each pair, they selected the animation that better followed the requested actions and their order, and the one with more natural human–object coordination.

Figure 6 shows a preference for AW3D in both text alignment (59.9%) and interaction quality (57.4%). Viewers more often judged its animations to preserve the requested actions and their order. They also judged the human–object coordination more natural. These findings are consistent with AW3D's goal of organizing action progression and guiding motion generation with the evolving interaction state. Together with the quantitative and visual results, the study supports the quality of its text-aligned HOI animations.

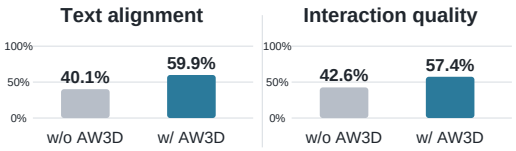

Figure 6: Perceptual preferences for HOI animations without and with AW3D.

F Limitations

AW3D currently focuses on one person interacting with one rigid object. Articulated objects, deformable objects, multiple manipulated objects, and scene-level interactions require richer relation states. Prototype quality also follows the motion and contact observations available during training, and rare prototype compositions can carry higher sequence uncertainty. The confidence gate attenuates their contribution. Exact collision-free and dynamically valid motion additionally requires explicit physical simulation. Integrating action-prototype world modelling with physical dynamics is a useful direction for future work.

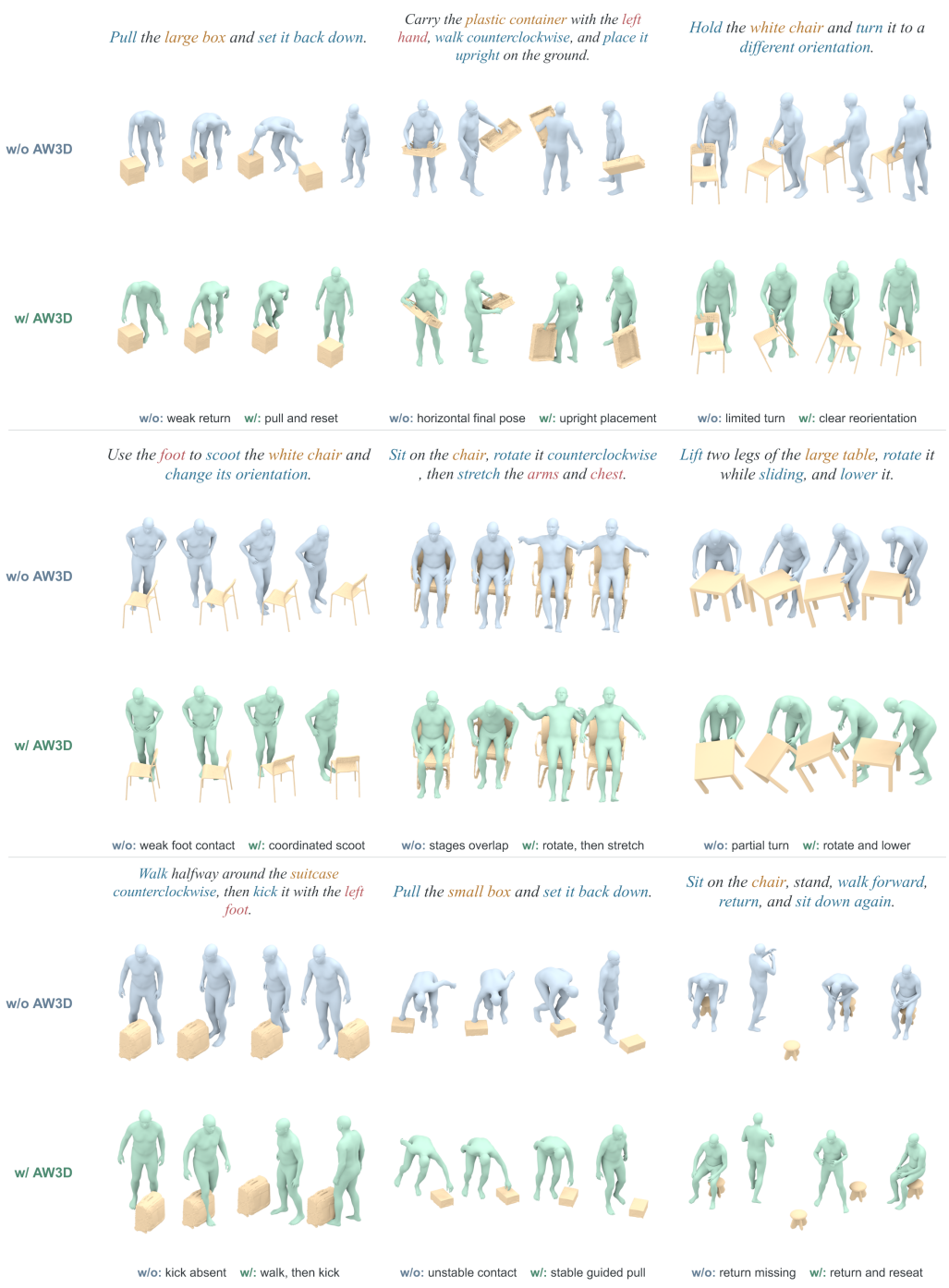

Figure 7: Additional paired comparisons at matched time points. Top: w/o AW3D. Bottom: w/ AW3D. The short annotations identify the main stage-level difference in each pair.

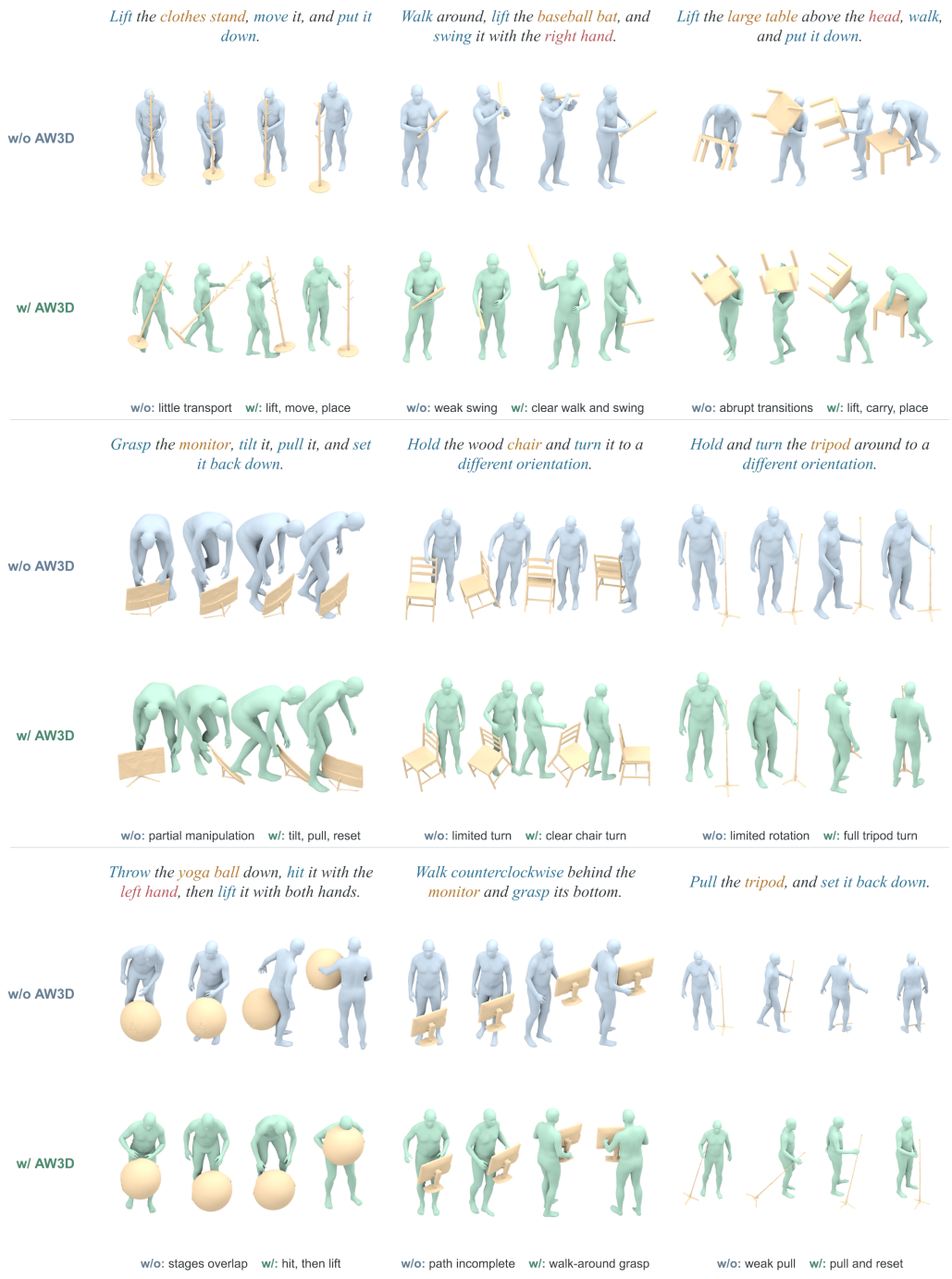

Figure 8: Additional paired comparisons at matched time points. Top: w/o AW3D. Bottom: w/ AW3D. The short annotations identify the main stage-level difference in each pair.