ULIFLOW: UNIFIED LATENT HUMAN-OBJECT INTERACTION FLOW

Anonymous authors
Paper under double-blind review

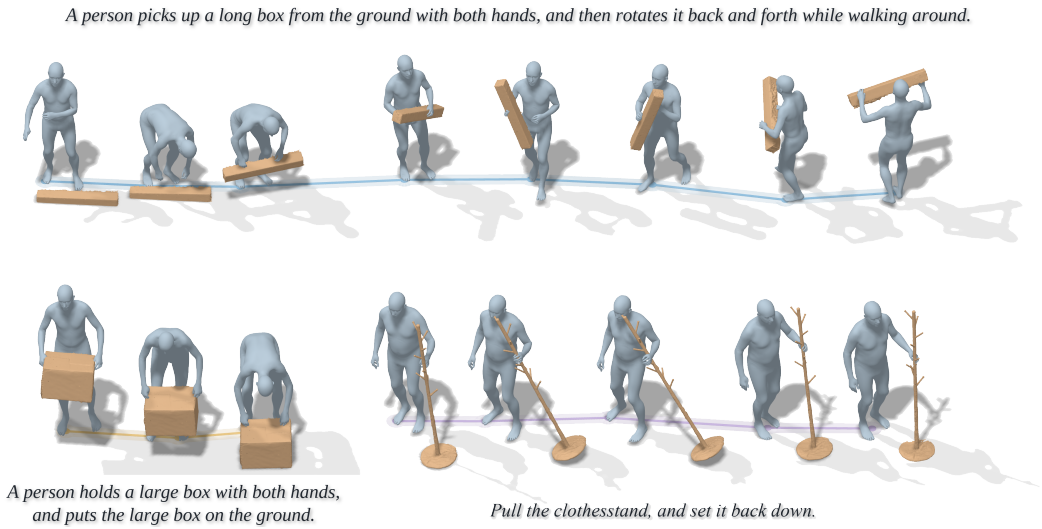

Figure 1: Given a textual prompt and a requested duration, ULIFlow jointly generates coherent human and object motions as a variable-length human–object interaction (HOI) sequence.

ABSTRACT

Text-conditioned 3D human–object interaction (HOI) generation remains challenging because individually plausible human and object motions do not necessarily form a coherent interaction. We argue that a key limitation lies in conventional skeleton-based representations, which compactly describe body kinematics while leaving surface-level interaction geometry largely implicit, often requiring additional contact-specific regularizers. We propose ULIFlow, a framework that explicitly models continuous human–object relations within a unified latent space. To this end, we derive continuous marker–object relations from sparse body-surface markers and treat them as an explicit interaction modality alongside human motion and object dynamics. A Mixture-of-Transformers (MoT) autoencoder is then designed to jointly model these heterogeneous streams while preserving their modality-specific structures. Importantly, we let the relation stream directly participate in human and object reconstruction, encouraging the shared latent representation to preserve both individual dynamics and the interaction geometry that coordinates them. To further achieve variable-length generation, we introduce a blockwise Flow Forcing scheme with global sequence-progress conditioning, enabling progressive continuation from fixed-length training windows while maintaining consistent long-horizon dynamics. Extensive experiments demonstrate that ULIFlow consistently outperforms prior work in overall generation quality and human–object coordination, while preserving coherent interactions over requested durations beyond the training horizon.

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.

1 INTRODUCTION

Text-conditioned 3D human–object interaction (HOI) generation aims to synthesize human motion and object dynamics from a language description and has attracted increasing attention (Li et al., 2024; Cha et al., 2024; Peng et al., 2025; Xu et al., 2024; Wang et al., 2026b). In particular, generating one interaction counterpart, as demonstrated by OMOMO (Li et al., 2023) and DAViD (Kim et al., 2025), has shown promising results. However, joint generation of plausible human and object motions remains challenging and relatively underexplored. Since individual plausibility does not ensure interaction coherence, simply combining the generated object and human motion trajectories may exhibit spatial misalignment or frequent interpenetration, which we define as a lack of *coordination*.

One straightforward way to improve generation coordination is to impose hand-crafted rules or kinematic constraints (Diller & Dai, 2024; Li et al., 2024; Peng et al., 2025). Although such regularization promotes physical consistency, such formulations often depend on task-specific assumptions and require substantial manual effort. To reduce this reliance on heuristics, recent methods instead learn human motion and object dynamics jointly, through interaction tokenization (Geng et al., 2025) or modality-aware fusion (Wang et al., 2026a). Despite these advances, the synthesized sequences can still exhibit physical artifacts such as interpenetration, suggesting that jointly modeling the two trajectories with prior techniques is insufficient to ensure coherent interaction. In particular, while this strategy has proved effective for human–human motion generation (Shafir et al., 2024), extending its success to HOI remains challenging: the distinct kinematic structures of humans and objects complicate their geometric coupling. Consequently, existing formulations may still require interaction-specific priors to regularize the generated interactions (Zeng et al., 2025).

We argue that the key reason joint modeling for HOI fails lies in the long-standing practice of adopting only skeleton-based representations for encoding human dynamics. As skeletal representations provide a compact description that primarily captures internal body structure, the required contact cues, which by contrast occur on the human surface, remain implicitly represented. Consequently, contact information between humans and objects is weakly encoded in the resulting latent space, which therefore necessitates additional contact modeling (e.g., regularization (Xu et al., 2023)) to reduce interaction artifacts.

In this paper, we ask: *can we jointly model human motion, object dynamics, and their continuous geometric relations in a shared latent space, such that complex contact patterns can be learned directly from data without contact-specific regularization?* Our core insight is that a powerful latent space for HOI requires explicit modeling of interaction geometry to handle diverse and evolving contact patterns. However, as the widely adopted skeletal representation is naturally limited in expressiveness, it cannot sufficiently characterize the rich geometric relations required for modeling diverse contact patterns.

To this end, we propose **U**nified **L**atent Human–Object **I**nteraction **Flow** (ULIFlow), which models continuous human–object dynamics in a unified latent space under the auto-encoding paradigm. Specifically, we introduce a complementary representation, body-surface markers (Zhang et al., 2021), to compensate for the limited geometric expressiveness of skeletal representations. Based on these markers, we explicitly derive continuous marker–object relations, including relative positions and velocities, and treat them as a distinct interaction modality alongside human and object dynamics. To accommodate the heterogeneous structures of human, object, and relation modalities, we design our model as a Mixture-of-Transformers (MoT) autoencoder (Liang et al., 2024) that combines modality-specific transformations with shared attention, preserving the characteristics of each stream while learning their cross-modal dependencies. Notably, we couple human and object decoding with the relation stream to dynamically guide the reconstruction of both trajectories using explicit spatiotemporal cues. As the resulting latent space is jointly informed by individual dynamics and explicit interaction geometry, it enables coordinated human–object motion to be modeled without additional contact-specific regularization.

Another challenge is to generate interactions of a requested duration from fixed-length training windows. To enable variable-length synthesis, we adapt the Diffusion Forcing mechanism (Chen et al., 2024) to our subsequent Flow Matching stage. Concretely, we divide each training window into temporal blocks with different noise levels, enabling the model to progressively continue an interaction from partially resolved motion. As shown in Fig. 1, our design allows the generated HOI sequences to have arbitrary length while maintaining coherent long-term dynamics.

Extensive experiments demonstrate that ULIFlow consistently outperforms prior methods in overall generation quality and human–object coordination. These advantages remain evident in long-horizon generation, where ULIFlow preserves coherent interactions as the requested duration extends beyond the fixed training window. Our contributions are as follows:

- We propose ULIFlow, a unified latent framework for text-conditioned HOI generation at any requested duration. By explicitly modeling continuous human–object geometry, ULIFlow learns coordinated dynamics directly from data without contact-specific regularization or heuristic interaction designs.
- We introduce a Mixture-of-Transformers (MoT) autoencoder that jointly encodes human, object, and relation modalities via modality-specific transformations and relation-aware decoding.
- We design a blockwise Flow Forcing scheme to align finite-window training with progressive rollout, enabling stable HOI generation beyond the training horizon.

2 Related Work

Generative modeling for HOI. Recent HOI generation methods have increasingly adopted diffusion models for the conditional synthesis of human and object motion (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020; Xu et al., 2023; Li et al., 2024). Despite their strong generative capability, diffusion-based HOI modeling remains susceptible to interaction artifacts. (Xu et al., 2023; Li et al., 2024; Cha et al., 2024). To improve interaction fidelity, existing approaches prevalently introduce contact- and affordance-aware guidance (Kulkarni et al., 2024; Diller & Dai, 2024) or explicitly model kinematic and relational dependencies (Zeng et al., 2025; Xue et al., 2025). While these advances substantially improve human–object consistency, they generally rely on dedicated interaction modeling or guidance mechanisms in addition to the underlying motion generator, which can be time-consuming to explicitly create the rules. Along this direction, HOI-Diff (Peng et al., 2025) learns an auxiliary affordance model to predict contact regions as guidance, while LIGHT (Wang et al., 2026b) derives cross-modal guidance from asynchronous denoising. These approaches reduce manual specification of guidance, but do not necessarily eliminate contact-specific supervision. In particular, (Wang et al., 2026b) still employs a contact loss based on binary labels derived from geometric thresholds. Our method instead learns human–object coordination from continuous geometric supervision without contact-specific regularization.

Human motion representations. Prior methods for text-conditioned motion generation (Guo et al., 2022; Tevet et al., 2023) and HOI modeling (Xu et al., 2023; Zeng et al., 2025) commonly adopt joint-based representations to compactly describe body kinematics. However, joints primarily characterize body kinematics and provide limited information about the body surface where physical interaction occurs. A common way to recover richer body geometry is to use parametric human models, such as SMPL (Loper et al., 2015), which reconstruct a posed body surface from compact pose and shape parameters (Petrovich et al., 2021; 2022). For example, GOAL (Taheri et al., 2022) samples body-surface vertices to characterize human–object proximity during whole-body grasping. To retain surface information more compactly, marker-based representations describe human motion using sparse points distributed over the body surface. MOJO (Zhang et al., 2021) demonstrates that such markers capture motion information unavailable from skeletal joints, while subsequent work extends them to scene-aware motion generation (Zhang & Tang, 2022; Zhao et al., 2023) and HOI modeling (Wu et al., 2022; Xu et al., 2025). These studies establish sparse body-surface markers as a compact representation for retaining interaction-relevant surface information.

Unified multimodal modeling. Recent work has demonstrated the effectiveness of unified multimodal modeling across diverse tasks. For example, Unified-IO 2 (Lu et al., 2023) unifies multimodal perception and action prediction, while Chameleon (Team, 2025) jointly supports text–image understanding and generation. However, extending this success to HOI remains challenging, as heterogeneous human and object dynamics must maintain precise geometric coordination. Prior efforts explore joint interaction tokenization (Geng et al., 2025) and modality-aware fusion (Wang et al., 2026a), yet joint modeling alone does not ensure consistent contact, leaving misalignment and interpenetration as unresolved concerns. Our method builds on the insight that explicitly modeling continuous

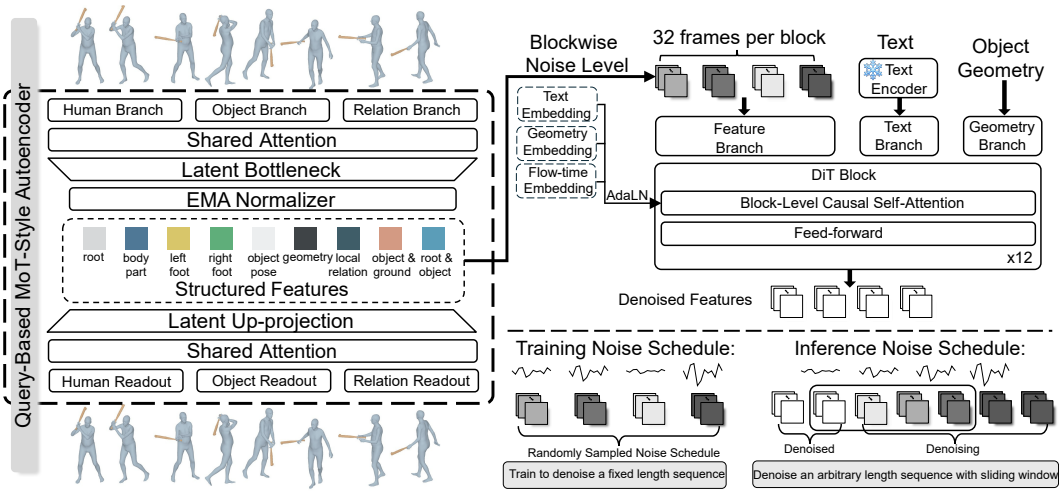

Figure 2: **Overview of ULIFlow**. **Left:** Unified latent modeling. The query-based MoT autoencoder jointly encodes and reconstructs human, object, and relation streams in a structured latent space. **Right:** Conditional generation. A conditional latent DiT uses independently sampled blockwise flow times and sliding-window inference to support generation at any requested duration.

human–object relations in the latent space enables joint learning of individual dynamics and geometric coordination, which addresses interaction inconsistencies by using relational cues to guide the reconstruction of both trajectories.

Variable-length generation. Prior work has explored variable-length synthesis in human motion (Barquero et al., 2024; Han et al., 2024; Xiao et al., 2025; Zhuo et al., 2025) and video generation (Song et al., 2025; Cai et al., 2026). Such flexibility is also important for HOI, since even the same interaction can be performed over different durations. However, extending an interaction requires human and object motion to remain coordinated throughout the generated sequence. To meet this requirement, we adapt Diffusion Forcing (Chen et al., 2024) to blockwise flow matching within our unified HOI latent space. By learning continuation from partially resolved histories and conditioning local windows on global sequence progress, our method generates coordinated interactions of the requested duration from fixed-length training windows.

3 METHOD

Given a text prompt c, object geometry G, and a requested duration T, we jointly generate human motion H and a rigid-object trajectory O over T frames to realize the described interaction. We represent human motion as a sequence of body-surface marker positions,

$$H = \{h_t\}_{t=1}^{T}, \qquad h_t = \{\boldsymbol{p}_{t,j}^{H}\}_{j=1}^{N},$$

where $\boldsymbol{p}_{t,j}^{H} \in \mathbb{R}^3$ is the position of marker j at frame t, for $t \in \{1, \dots, T\}$ and $j \in \{1, \dots, N\}$. The object trajectory $O = (\boldsymbol{p}^{O}, \boldsymbol{r}^{O})$ comprises translations $\boldsymbol{p}^{O} \in \mathbb{R}^{T\times 3}$ and continuous 6D rotations $\boldsymbol{r}^{O} \in \mathbb{R}^{T\times 6}$ (Zhou et al., 2018). Therefore, The joint state at frame t is $x_t = (\boldsymbol{p}_t^{H}, \boldsymbol{p}_t^{O}, \boldsymbol{r}_t^{O})$.

Continuous relation representation. To make geometric coordination an explicit learning target, we derive a continuous relation representation from the evolving spatial configuration between H and O. At each frame t, we identify the nearest object-surface point $\boldsymbol{s}_{t,j}^{O}$ for every body-surface marker $\boldsymbol{p}_{t,j}^{H}$ and define their relative offset as

$$\boldsymbol{\delta}_{t,j} = \boldsymbol{s}_{t,j}^{O} - \boldsymbol{p}_{t,j}^{H}. \tag{1}$$

This offset describes the relative position of the object surface with respect to the marker, while $d_{t,j} = \|\boldsymbol{\delta}_{t,j}\|_2$ measures their spatial separation.

Since instantaneous proximity alone cannot describe how an interaction evolves, we further characterize the first- and second-order temporal variations of the offset by $\nabla\boldsymbol{\delta}_{t,j}$ and $\nabla^2\boldsymbol{\delta}_{t,j}$, respectively. To retain information about where the interaction occurs on the object, we also express the corresponding nearest surface point in the canonical object frame as $\bar{\boldsymbol{s}}^O_{t,j}$. Together, these quantities define the local relation feature:

$$\boldsymbol{r}_{t,j} = \left(\boldsymbol{\delta}_{t,j}, d_{t,j}, \nabla\boldsymbol{\delta}_{t,j}, \nabla^2\boldsymbol{\delta}_{t,j}, \bar{\boldsymbol{s}}^O_{t,j}\right). \tag{2}$$

To also capture the object's relation to the supporting surface, we apply the same construction to the object–ground pair. Here, $\boldsymbol{q}_t^{\mathrm{gnd}} = \Pi_{\mathrm{gnd}}(\boldsymbol{p}_t^O)$ denotes the projection of the object center onto the ground plane, while $\boldsymbol{r}_t^{\mathrm{gnd}}$ follows the relation definition in Eq. 2. Collecting the marker–object and object–ground relations across frames yields the local relation field $R^{\mathrm{local}} = \{(\{\boldsymbol{r}_{t,j}\}_{j=1}^N, \boldsymbol{r}_t^{\mathrm{gnd}})\}_{t=1}^T$.

Beyond these local relations, the overall human–object arrangement provides complementary global context. To describe this context, we use $\boldsymbol{p}_t^{\mathrm{root}} = N^{-1}\sum_{j=1}^N \boldsymbol{p}_{t,j}^H$ as the human root proxy and define $\boldsymbol{\gamma}_t = \boldsymbol{p}_t^O - \boldsymbol{p}_t^{\mathrm{root}}$ and $\rho_t = \|\boldsymbol{\gamma}_t\|_2$. Together with its temporal variations, these quantities form $\boldsymbol{r}_t^{\mathrm{global}} = (\boldsymbol{\gamma}_t, \rho_t, \nabla\boldsymbol{\gamma}_t, \nabla^2\boldsymbol{\gamma}_t)$ and the global relation field $R^{\mathrm{global}} = \{\boldsymbol{r}_t^{\mathrm{global}}\}_{t=1}^T$. Combining the two fields yields the complete relation representation $R = (R^{\mathrm{local}}, R^{\mathrm{global}})$. Because both components are derived directly from H and O, R provides direct supervision for spatial coordination without contact annotations or interaction-specific pseudo-labels.

Tokenization. To jointly encode human motion, object dynamics, and their geometric relations, we organize H, O, and R into structured token streams. Given that human motion and relation features are distributed over multiple body-surface markers, learned body-region queries aggregate marker-level observations into compact region tokens (Jaegle et al., 2021), while dedicated tokens retain root and foot information as well as root–object and object–ground relations. For the object stream, each rigid pose is represented by a temporal token, whereas the static geometry G is summarized using a fixed set of geometry queries (Prokudin et al., 2019). As a final step, all tokens are projected into a common feature space while retaining their modality identities and temporal positions, yielding the human, object, and relation streams X^H, X^O, and X^R, respectively.

Unified latent space modeling. Since human motion, object dynamics, and human–object relations exhibit different structures and statistical distributions, directly mixing their tokens can obscure modality-specific information. Following recent unified multimodal architectures (Esser et al., 2024; Niu et al., 2026; Agarwal et al., 2026), we therefore build the autoencoder with Mixture-of-Transformers (MoT) blocks (Liang et al., 2024), which retain modality-specific transformations while exchanging information through shared attention. As shown in the left part of Fig. 2, the encoder maps the three token streams into a unified latent field as

$$Z = \mathcal{N}_{\mathrm{EMA}}\left(\mathcal{E}_{\mathrm{MoT}}(X^H, X^O, X^R)\right), \tag{3}$$

where $\mathcal{E}_{\mathrm{MoT}}$ denotes the composite MoT-based encoder, which jointly fuses and compresses the three modality streams into a compact latent field while preserving their modality identities and temporal positions. The normalizer $\mathcal{N}_{\mathrm{EMA}}$ stabilizes the scale of the resulting latent features.

For reconstruction, the MoT decoder $\mathcal{D}_{\mathrm{MoT}}$ expands the compact latent field back to full temporal resolution and reconstructs the human, object, and relation sequences:

$$(\hat{H}, \hat{O}, \hat{R}) = \mathcal{D}_{\mathrm{MoT}}(Z). \tag{4}$$

Within the decoder, information is propagated across the three modality streams before time-indexed queries recover their full temporal sequences. Importantly, relation features remain coupled with both human and object decoding, allowing explicit relational cues to guide the reconstruction of $\hat{H}$ and $\hat{O}$ rather than serving only as an auxiliary reconstruction target. By requiring the same latent field to reconstruct all three streams, the autoencoder encourages Z to retain both individual dynamics and the interaction geometry that coordinates them. Further details of the autoencoder are provided in Sec. A.1 of the Appendix.

Conditional latent flow model. Within the learned latent space, we model interaction generation using Flow Matching (Lipman et al., 2022; Liu et al., 2022) and parameterize the conditional velocity

field with a DiT (Peebles & Xie, 2023). As shown in the right part of Fig. 2, the model predicts the latent velocity as

$$v_\psi\left(Z_{\boldsymbol{\tau}}, \boldsymbol{\tau} \mid c, G, p\right) = \mathcal{F}_\psi\left(Z_{\boldsymbol{\tau}}; c, G, \boldsymbol{\tau}, p\right), \tag{5}$$

where $\boldsymbol{\tau}$ specifies the flow-time assignment of each latent token. During long-horizon rollout, fixed-length training windows may correspond to different stages of the complete interaction, creating inconsistent temporal context between training and inference. Accordingly, we introduce a normalized sequence-progress condition $p \in [0, 1]$ to specify the global position of the current window within the requested duration. By using the same progress definition during training and inference, the model retains consistent sequence-level temporal context throughout generation.

Training schedule. To learn variable-length continuation from fixed-length training windows, we introduce a blockwise forcing schedule for Flow Matching. Unlike Diffusion Forcing (Chen et al., 2024), which assigns noise levels at the token level, our schedule assigns each temporal block an independently sampled shared flow time $\tau_b \sim \mathcal{U}(0, 1)$, where 0 and 1 denote the noise and data endpoints, respectively. This exposes the model to partially resolved causal histories. Under block-causal attention, each block uses only its available causal context, while the sequence-progress condition p maintains a global temporal reference. Together, these designs align fixed-window training with long-horizon rollout.

Training loss. We here give the training losses for each module of our model:

Autoencoder loss. We train the autoencoder with a composite reconstruction objective

$$\mathcal{L}_{\mathrm{AE}} = \lambda_H \mathcal{L}_H + \lambda_O \mathcal{L}_O + \lambda_R \mathcal{L}_R + \lambda_{\mathrm{world}} \mathcal{L}_{\mathrm{world}} + \lambda_{\mathrm{cov}} \mathcal{L}_{\mathrm{cov}}. \tag{6}$$

Here, $\mathcal{L}_H$ and $\mathcal{L}_O$ apply robust reconstruction losses to the human and object states and their first- and second-order temporal changes. The relation term $\mathcal{L}_R$ supervises both local and global relation features, while $\mathcal{L}_{\mathrm{world}}$ constrains the composed marker positions in world coordinates. Finally, $\mathcal{L}_{\mathrm{cov}}$ regularizes the aggregate first- and second-order statistics of the latent field. All reconstruction terms are masked according to their valid temporal and entity supports.

Flow Matching loss. For each valid latent token z_i, we sample Gaussian noise $\epsilon_i \sim \mathcal{N}(0, I)$ and use the flow time τ_i assigned by its block to define $z_i(\tau_i) = (1 - \tau_i)\epsilon_i + \tau_i z_i$ and the target velocity $u_i = z_i - \epsilon_i$. The conditional flow model is optimized with

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{Z, \epsilon, \boldsymbol{\tau}} \left[\frac{1}{|\Omega| d_z} \sum_{i \in \Omega} \left\| v_{\psi, i}(Z_{\boldsymbol{\tau}}, \boldsymbol{\tau} \mid c, G, p) - u_i \right\|_2^2 \right], \tag{7}$$

where $\boldsymbol{\tau} = \{\tau_i\}_{i \in \Omega}$ collects the token-wise flow times, Ω denotes the valid latent tokens, and d_z is the latent channel dimension. The generative model is therefore supervised entirely in the structured latent space, while the autoencoder provides interaction-aware latent targets and decoding. Implementation details are provided in Sec. A.3 of the Appendix.

Inference. During inference, we initialize an ordered sequence of latent blocks from Gaussian noise according to the requested duration T. We integrate the predicted velocities using a staggered schedule, allowing earlier blocks to reach the clean endpoint before later blocks. To extend generation while keeping the model context fixed, we progressively shift the window forward, retaining recent completed blocks as history and introducing new noisy blocks. As the window advances, we update the progress condition p according to the window's global position within the requested sequence. We then decode the completed latents in overlapping windows, assemble the outputs in temporal order, and crop the sequence to exactly match T frames. Further details of conditional generation are provided in Sec. A.2 of the Appendix.

4 Experiments

Dataset. We conduct our experiments on InterAct (Xu et al., 2025), a large-scale benchmark for text-conditioned 3D human–object interaction generation. InterAct consolidates seven existing datasets into 21.81 hours of annotated 3D whole-body interactions with fine-grained textual descriptions and provides a marker-based representation of human motion using body-surface markers (Zhang et al.,

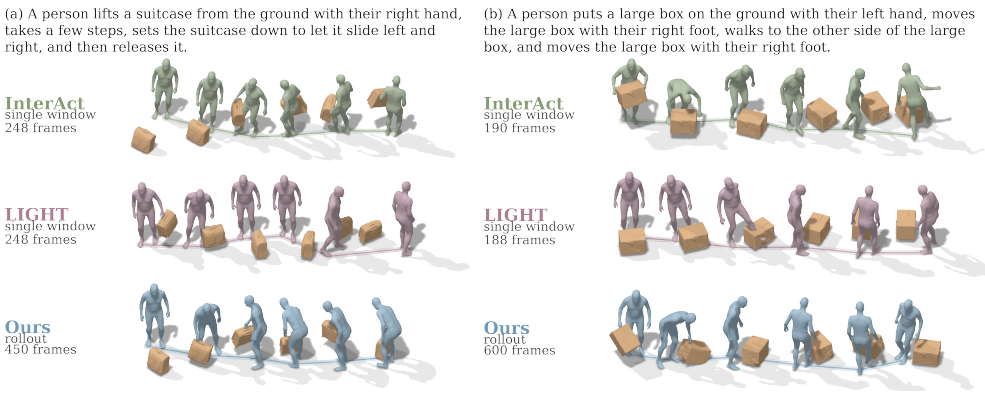

Figure 3: **Qualitative comparison** of variable-length generation with prior works. ULIFlow preserves coherent interaction beyond fixed generation windows.

Table 1: **Quantitative comparison** of generation quality on the InterAct dataset (Xu et al., 2025).

Method	FID ↓	R-Precision ↑			MM-Dist ↓	Diversity →	MModality ↑
		Top 1	Top 2	Top 3			
Ground Truth	0.000 ±0.000	0.869 ±0.003	0.952 ±0.001	0.971 ±0.002	4.012 ±0.000	11.939 ±0.000	–
Text2HOI (Cha et al., 2024)	24.521 ±0.035	0.425 ±0.001	0.606 ±0.008	0.697 ±0.004	6.268 ±0.012	6.108 ±0.023	2.544 ±0.119
CHOIS (Li et al., 2024)	6.914 ±0.147	0.644 ±0.005	0.870 ±0.007	0.958 ±0.010	5.278 ±0.006	9.378 ±0.028	2.758 ±0.017
HOI-Diff (Peng et al., 2025)	14.445 ±0.056	0.134 ±0.007	0.263 ±0.016	0.362 ±0.008	10.230 ±0.010	6.252 ±0.044	0.059 ±0.005
InterAct (Xu et al., 2025)	1.768 ±0.019	0.749 ±0.014	0.933 ±0.011	0.968 ±0.005	**4.507 ±0.003**	11.416 ±0.038	3.253 ±0.031
LIGHT (Wang et al., 2026b)	8.054 ±0.058	0.584 ±0.011	0.747 ±0.007	0.807 ±0.015	7.280 ±0.005	11.042 ±0.068	3.019 ±0.356
ULIFlow w/o MoT structure	3.113 ±0.033	0.841 ±0.009	0.941 ±0.001	0.971 ±0.005	5.457 ±0.016	11.458 ±0.014	3.361 ±0.208
ULIFlow w/o relation branch	1.244 ±0.021	**0.861 ±0.008**	**0.947 ±0.006**	**0.976 ±0.001**	4.890 ±0.020	11.677 ±0.015	3.253 ±0.198
ULIFlow (Ours)	**0.812 ±0.014**	0.852 ±0.008	0.945 ±0.008	0.972 ±0.005	4.852 ±0.013	**11.833 ±0.048**	**3.617 ±0.219**

2021). We adopt this marker representation to construct the continuous human–object relations used throughout our model. We follow the official data split for training and evaluation and sample masked 128-frame windows from the interaction sequences during training.

Metrics. Following standard practice in text-to-motion generation (Guo et al., 2022), we evaluate generated HOI sequences in terms of realism, text–interaction alignment, and diversity. We report Fréchet Inception Distance (**FID**) to measure feature-distribution similarity, **R-Precision** and Multimodal Distance **(MM-Dist)** to assess text–interaction alignment, and **Diversity** and Multimodality (**MModality**) to characterize variation across generated interactions.

To further evaluate human–object coordination, we measure body–object penetration using the penetration ratio together with the mean, 95th-percentile, and maximum penetration depths. For long-horizon generation, we evaluate the standard generation metrics on fixed-length subsequences extracted from extended rollouts and assess temporal stability around block boundaries using Peak Jerk and Area Under the Jerk (AUJ) (Cai et al., 2026). We additionally report human and object jerk statistics over the generated sequences to characterize motion stability. For metrics marked with →, values closer to the ground-truth statistics indicate better agreement with real motion dynamics. Full metric definitions and evaluation protocols are provided in Sec. B of the Appendix.

Baselines. We compare ULIFlow with five state-of-the-art text-conditioned HOI generation methods: Text2HOI (Cha et al., 2024), CHOIS (Li et al., 2024), HOI-Diff (Peng et al., 2025), InterAct (Xu et al., 2025), and LIGHT (Wang et al., 2026b). Since these methods adopt different motion representations, we first standardize their outputs for evaluation using the body-surface marker representation of InterAct (Xu et al., 2025). InterAct natively produces this representation, while we convert the outputs of the remaining baselines following prior adaptation protocols, without modifying the underlying generated motions. As the released checkpoints and evaluation settings of the adapted baselines cover different subsets of InterAct, we evaluate each method only on the subset supported by its original release.

Table 2: **Quantitative comparison** of long-horizon generation on the InterAct dataset (Xu et al., 2025). The two column groups are computed over non-overlapping 128-frame subsequences and 64-frame windows centered at continuation boundaries, respectively.

Method	128-frame subsequences						Boundary-aligned 64-frame windows			
	FID ↓	R-Precision ↑			Div →	MM-Dist ↓	FID ↓	Div →	Peak jerk →	AUJ →
		Top 1	Top 2	Top 3						
Ground Truth	$0.000_{\pm 0.000}$	$0.920_{\pm 0.000}$	$0.969_{\pm 0.000}$	$0.987_{\pm 0.000}$	$11.104_{\pm 0.000}$	$3.946_{\pm 0.000}$	$0.000_{\pm 0.000}$	$10.860_{\pm 0.000}$	$1.530_{\pm 0.000}$	$32.841_{\pm 0.000}$
InterAct (Xu et al., 2025)	$3.097_{\pm 0.180}$	$0.698_{\pm 0.014}$	$0.856_{\pm 0.026}$	$0.918_{\pm 0.022}$	$\mathbf{10.928_{\pm 0.036}}$	$6.296_{\pm 0.049}$	$3.713_{\pm 0.160}$	$\mathbf{10.670_{\pm 0.036}}$	$1.734_{\pm 0.025}$	$10.416_{\pm 0.143}$
LIGHT (Wang et al., 2026b)	$5.658_{\pm 0.128}$	$0.583_{\pm 0.009}$	$0.752_{\pm 0.013}$	$0.818_{\pm 0.008}$	$10.795_{\pm 0.125}$	$7.276_{\pm 0.028}$	$4.914_{\pm 0.069}$	$10.664_{\pm 0.039}$	$\mathbf{1.445_{\pm 0.077}}$	$18.012_{\pm 0.192}$
ULIFlow	$\mathbf{2.686_{\pm 0.060}}$	$\mathbf{0.753_{\pm 0.018}}$	$\mathbf{0.909_{\pm 0.007}}$	$\mathbf{0.955_{\pm 0.009}}$	$10.782_{\pm 0.051}$	$\mathbf{6.023_{\pm 0.031}}$	$\mathbf{3.415_{\pm 0.086}}$	$10.592_{\pm 0.027}$	$1.203_{\pm 0.024}$	$\mathbf{39.300_{\pm 0.356}}$

Table 3: **Quantitative comparison** of motion stability on the InterAct dataset (Xu et al., 2025). Jerk is reported in cm/frame^3.

Method	Human jerk			Object jerk		
	Mean →	Peak →	AUJ →	Mean →	Peak →	AUJ →
Ground Truth	$0.546_{\pm 0.000}$	$2.202_{\pm 0.000}$	$98.692_{\pm 0.000}$	$0.796_{\pm 0.000}$	$5.736_{\pm 0.000}$	$144.689_{\pm 0.000}$
InterAct (Xu et al., 2025)	$0.216_{\pm 0.003}$	$3.196_{\pm 0.042}$	$38.742_{\pm 0.340}$	$0.548_{\pm 0.034}$	$9.094_{\pm 0.500}$	$91.701_{\pm 3.920}$
LIGHT (Wang et al., 2026b)	$0.342_{\pm 0.001}$	$\mathbf{2.722_{\pm 0.080}}$	$66.963_{\pm 0.306}$	$0.536_{\pm 0.005}$	$4.089_{\pm 0.078}$	$105.587_{\pm 1.291}$
ULIFlow	$\mathbf{0.608_{\pm 0.003}}$	$1.442_{\pm 0.027}$	$\mathbf{113.770_{\pm 0.418}}$	$\mathbf{0.846_{\pm 0.007}}$	$\mathbf{4.466_{\pm 0.252}}$	$\mathbf{156.460_{\pm 1.902}}$

Table 4: **Quantitative comparison** of physical plausibility on the InterAct dataset (Xu et al., 2025). Body–object penetration depths are reported in cm.

Method	Ratio ↓	Mean ↓	P95 ↓	Max ↓
Ground Truth	$0.003_{\pm 0.000}$	$0.622_{\pm 0.000}$	$1.366_{\pm 0.000}$	$1.585_{\pm 0.000}$
InterAct (Xu et al., 2025)	$0.006_{\pm 0.001}$	$0.905_{\pm 0.145}$	$2.002_{\pm 0.227}$	$2.421_{\pm 0.182}$
LIGHT (Wang et al., 2026b)	$0.006_{\pm 0.001}$	$0.905_{\pm 0.068}$	$2.037_{\pm 0.140}$	$2.513_{\pm 0.180}$
ULIFlow	$\mathbf{0.003_{\pm 0.001}}$	$\mathbf{0.845_{\pm 0.088}}$	$\mathbf{1.805_{\pm 0.214}}$	$\mathbf{2.163_{\pm 0.215}}$

Generation quality. We compare ULIFlow against prior methods to evaluate the overall quality of generated interactions. Table 1 shows that ULIFlow surpasses the baseline approaches across most metrics while retaining diversity close to real interactions, which suggests that our unified modeling jointly harvests motion realism, text–interaction alignment, and generative variation. Using the same model and protocol without subset-specific fine-tuning, results on BEHAVE (Bhatnagar et al., 2022) and OMOMO (Li et al., 2023) in Table B exhibit the same trend.

Physical coordination. We then evaluate in Table 3 physical coordination through motion stability, body–object penetration, and visual contact quality. Compared with prior work, we can see that ULIFlow consistently generates human and object dynamics whose smoothness is closer to that of real sequences across the evaluated statistics. The penetration comparison shown in Table 4 indicates that ULIFlow consistently produces fewer and less severe geometric conflicts than prior work, demonstrating stronger spatial coordination. In Figure 4, we show that ULIFlow achieves better human–object coordination with less floating and object drift in both the slow-moving (top) and fast-moving (bottom) regimes. We further conduct a relation-pathway ablation in Tables G and H, which show that explicit relation modeling consistently improves performance across all human–object coordination metrics.

Variable-length generation. As shown in Table 2, ULIFlow preserves local motion quality, text consistency, and continuation stability when extending interactions beyond the fixed training horizon. Figure 3 further visualizes this improvement by showing that, under the same text condition, ULIFlow follows a requested duration suited to the described interaction and produces a more coherent sequence than fixed-window baselines.

Architecture Design Ablation. We evaluate the impact of different architectural designs in the autoencoder. Table 1 shows that replacing the MoT architecture with a conventional flat Transformer autoencoder causes the most persistent degradation, while removing the relation pathway also weakens overall performance. The autoencoder ablation and subset results in Table 5 and Table F show the same tendency without compromising human or object reconstruction. Together, these results confirm that modality-aware processing and explicit relation modeling provide a more effective interface for flow-based generation. See Sec. C.2 for further details.

Temporal Rollout Design Ablation. We ablate blockwise forcing and progress conditioning under long-horizon generation. As shown in Figure 5, either ablation preserves plausible early motion but

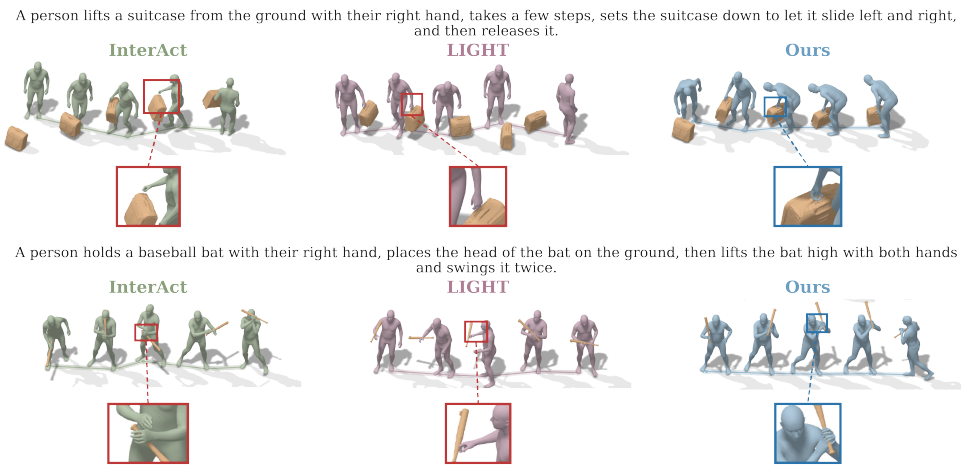

Figure 4: **Qualitative comparison** of contact and motion coupling. ULIFlow achieves more physically plausible hand–object contact.

Table 5: **Ablation study** of latent autoencoder designs on the InterAct dataset (Xu et al., 2025).

MoT structure	Latent regularization	Relation branch	Parameters (M)	Marker (cm) ↓	Object translation (cm) ↓	Object rotation (deg) ↓	Effective rank (%) ↑
×	✓	✓	60.66	3.805	3.174	10.393	6.96
✓	✓	×	60.41	**1.297**	1.757	**1.512**	72.73
✓	×	✓	60.40	1.367	**1.148**	1.516	47.80
✓	✓	✓	60.40	1.348	1.263	1.520	**82.15**

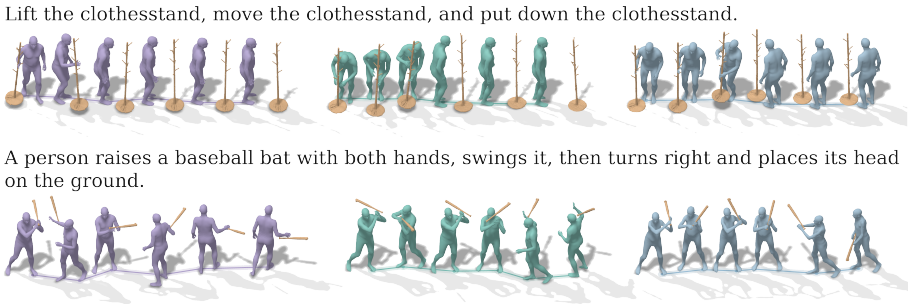

Figure 5: **Qualitative comparison** of long-horizon rollout designs. **Left**: ULIFlow without progress conditioning, **Mid**: ULIFlow without blockwise forcing, and **Right**: the complete model. Blockwise forcing and progress conditioning jointly preserve coordinated motion during late rollout.

gradually loses human–object coordination, whereas their combination maintains a stable late rollout. The quantitative comparisons in Tables C and D further evidence their complementary roles across requested durations and continuation boundaries.

5 Conclusion

In this work, we introduce ULIFlow, a framework for text-conditioned HOI generation that jointly models human motion, object dynamics, and continuous interaction geometry in a unified latent space. Derived from body-surface markers, these continuous relations allow ULIFlow to represent interaction geometry as an explicit modality alongside human and object dynamics. Through modality-specific processing and relation-aware decoding, the MoT autoencoder supports coordinated reconstruction without contact-specific regularization, while blockwise Flow Forcing and sequence-progress conditioning enable variable-length rollout beyond the training horizon. Across extensive experiments on InterAct, ULIFlow achieves state-of-the-art performance in motion realism and interaction coordination.

AI USE STATEMENT

We use ChatGPT and OpenAI Codex to assist with the refinement of some internal module designs, parts of the implementation, literature search and summarization, and language polishing. The core research idea, overall architecture, experimental design and evaluation protocols, result interpretation, dataset processing, qualitative analysis, and mathematical formulation are developed and conducted by the authors without generative AI assistance. All AI-assisted design suggestions, code, and literature references are independently reviewed and verified by the authors, who take full responsibility for the final content of the paper and associated artifacts.

ETHICS STATEMENT

This work uses existing HOI datasets and does not involve the collection of new human-subject data. The generated outputs are represented using abstract body-surface markers and rigid-object geometry rather than identifiable visual appearance or biometric information. Nevertheless, realistic HOI synthesis could potentially be used to produce misleading synthetic behavior, and the model may inherit biases and coverage limitations from the underlying datasets. We therefore encourage clear disclosure of synthetically generated content and discourage its use in deceptive, impersonation-related, or safety-critical applications.

REPRODUCIBILITY STATEMENT

We describe the model architecture, learning objectives, and inference procedure in Sec. 3. Complete training configurations and checkpoint-selection protocols are provided in Sec. A.3, while baseline adaptation and evaluation details are documented in Sec. B. Codes will be released.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 457–469. IEEE, 2024.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15914–15925. IEEE, 2022.

Yiyi Cai, Yuhan Wu, Kunhang Li, You Zhou, Bo Zheng, and Haiyang Liu. Flooddiffusion: Tailored diffusion forcing for streaming motion generation. pp. 2295–2304, 2026.

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

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. volume 37, pp. 24081–24125, 2024.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 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*, pp. 5152–5161, 2022.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. volume 33, pp. 6840–6851, 2020.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Hyeonwoo Kim, Sangwon Baik, and Hanbyul Joo. David: Modeling dynamic affordance of 3d objects using pre-trained video diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10330–10341. IEEE, 2025.

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*, pp. 947–957, 2024.

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 *European Conference on Computer Vision*, pp. 54–72. Springer, 2024.

Qifan Li, Xingyu Zhou, Jinhua Zhang, Weiyi You, and Shuhang Gu. Taming sampling perturbations with variance expansion loss for latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43643–43652, 2026.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. 2022.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. 2023.

Dantong Niu, Zhuoyang Liu, Zekai Wang, Boning Shao, Zhao-Heng Yin, Anirudh Pai, Yuvan Sharma, Stefano Saravalle, Ruijie Zheng, Jing Wang, et al. T-rex: Tactile-reactive dexterous manipulation. *arXiv preprint arXiv:2606.17055*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

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

Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer VAE. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10985–10995, 2021.

Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4332–4341, 2019.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations*, volume 2024, pp. 8717–8733, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. 2020.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. 2025.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13253–13263. IEEE, 2022.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL https://arxiv.org/abs/2405.09818.

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. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Yin Wang, Ziyao Zhang, Zhiying Leng, Haitian Liu, Frederick WB Li, Mu Li, and Xiaohui Liang. Multimodal prior-augmented text-driven 3d human-object interaction generation. *Science China Information Sciences*, 69(5):150105, 2026a.

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 *ICLR*, 2026b.

Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision*, pp. 257–274. Springer, 2022.

Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10086–10096. IEEE, 2025.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14882–14894. IEEE, 2023.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. volume 37, pp. 52858–52890, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7048–7060. IEEE, 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 Computer Vision and Pattern Recognition Conference*, pp. 22714–22723, 2025.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

Zhengrong Yue, Taihang Hu, Mengting Chen, Haiyu Zhang, Zihao Pan, Tao Liu, Zikang Wang, Jinsong Lan, Xiaoyong Zhu, Bo Zheng, et al. What matters for diffusion-friendly latent manifold? prior-aligned autoencoders for latent diffusion. *arXiv preprint arXiv:2605.07915*, 2026.

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*, pp. 12358–12369, 2025.

Le Zhang, Ning Mang, and Aishwarya Agrawal. Rit: Vanilla diffusion transformers suffice in representation space. *arXiv preprint arXiv:2605.21981*, 2026.

Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20481–20491, 2022.

Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3372–3382, 2021.

Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, , and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *International conference on computer vision (ICCV)*, 2023.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5738–5746, 2018. URL `https://api.semanticscholar.org/CorpusID:56178817`.

Wenjie Zhuo, Fan Ma, and Hehe Fan. Infinidreamer: Arbitrarily long human motion generation via segment score distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14688–14698. IEEE, 2025.

A Supplementary Method Details

A.1 Structured Autoencoder

Token construction. Considering that the marker-level human and relation fields are substantially denser than the rigid-object state, we first reduce them to comparable token sets without removing their spatial organization. At each frame, eight learned body-region queries aggregate the 77 surface markers, while separate root and foot tokens retain global translation and support information. This construction yields 11 human tokens per frame. Given the rigid structure of the object, we use one dynamic pose token per frame and four static geometry tokens obtained from its Basis Point Set (BPS) descriptor (Prokudin et al., 2019). To preserve the correspondence between motion and relation features, eight relation queries aggregate marker-level relations and two additional tokens retain the object–ground and root–object relations, yielding 10 relation tokens per frame. Finally, learned identity encodings distinguish marker regions, relation types, entities, and modalities, while marker-to-part assignments and marker–object proximity guide the corresponding query aggregation.

Continuous relation fields. To instantiate the continuous relation representation defined in Sec. 3, the first 77 relation entries preserve a one-to-one correspondence with the human markers. An additional local entry applies the same construction to the object–ground pair, while a separate global entry records the root–object offset, its norm, and their first- and second-order temporal differences. Since these local, ground, and global quantities follow different numerical scales, we normalize the three groups independently before tokenization. This keeps the relation pathway aligned with the body surface while preventing any one field from dominating the query aggregation.

Modality-aware encoding. Given that dynamic object poses and static object geometry require different transformations despite belonging to the same semantic modality, the three input streams are implemented with four computational routes comprising human dynamics, object pose, static object geometry, and relation dynamics. To permit cross-modal exchange without forcing identical feature transformations, we share attention across routes and retain route-specific normalization and feed-forward layers. For route s, one block is expressed as

$$\widetilde{X}^s = X^s + \mathrm{Attn}\Big(\mathrm{LN}_s(X^s), \{\mathrm{LN}_{s'}(X^{s'})\}_{s'}\Big),$$
$$Y^s = \widetilde{X}^s + \mathrm{FFN}_s\Big(\mathrm{LN}'_s(\widetilde{X}^s)\Big).$$

Here, X^s and Y^s denote the input and output features of route s, respectively. To exchange information across the complete interaction, the attention operator uses route s as its query source and the concatenation of all routes as keys and values. Meanwhile, the route-specific transformations preserve modality-dependent structure. Since only dynamic tokens carry physical time, temporal rotary position embeddings (RoPE) (Su et al., 2023) are applied along their common frame axis, while static geometry tokens retain only their learned identities.

Temporal compression. Since preserving every frame would make the latent field grow linearly with the input duration, we compress all dynamic routes onto a shared set of temporal anchors. To maintain both global coverage and interaction-critical changes, the hybrid selection criterion allocates proportions of 0.5, 0.3, and 0.2 to uniform coverage, relation saliency, and motion energy, respectively. For a sequence of T frames, the number of anchors is $K = \min(32, \lceil T/4 \rceil)$. At anchor a_k, the retained token is augmented with information from the complete temporal stream through gated attention:

$$\bar{x}_{k,s} = x_{a_k,s} + g_{k,s} \odot \sum_{t=1}^{T} \alpha_{k,s,t} V_s x_{t,s}, \qquad \sum_t \alpha_{k,s,t} = 1.$$

Here, $\alpha_{k,s,t}$ is the attention weight from anchor k to frame t on route s, V_s is a learned value projection, and $g_{k,s}$ controls the amount of temporal context added to the anchor. Given a 128-frame window, the compressed topology contains 32 copies of the 11 human, one object-pose, and 10 relation tokens, together with four static geometry tokens, for a total of 708 tokens. To obtain the final latent field without changing these identities, a shared nonlinear projection reduces each token from the model width to $d_z = 16$.

Latent normalization and whitening. Since modality-specific routes can produce latent channels with different scales, valid tokens update channel-wise exponential moving statistics dur-

ing autoencoder optimization. With momentum 0.99, the normalized reconstruction latent is $z^{\mathrm{rec}}_{i,c} = (z^{\mathrm{raw}}_{i,c} - m_c)/\sqrt{v_c + \epsilon}$, where m_c and v_c are the running mean and variance of channel c and are updated without gradient propagation. The covariance term in Eq. 6 further penalizes nonzero aggregate means, off-diagonal covariance, and deviation of diagonal covariance from one. To provide the flow model with a fixed, well-conditioned target space, we then estimate a zero-phase component analysis (ZCA) transform from valid training latents after autoencoder optimization. If $C = U\,\mathrm{diag}(\lambda)U^\top$ denotes their covariance and $\widetilde{\lambda}_i = \max(\lambda_i, \epsilon)$, then

$$W = U\,\mathrm{diag}(\widetilde{\lambda}^{-1/2})U^\top, \qquad z^{\mathrm{white}} = (z^{\mathrm{rec}} - \mu)W^\top.$$

Accordingly, the conditional flow model is optimized in z^{white} space. Before decoding, the fixed inverse transform recovers $z^{\mathrm{rec}} = z^{\mathrm{white}}(W^{-1})^\top + \mu$. Invalid tokens remain zero under both transformations.

Relation-coupled decoding. To prevent relation reconstruction from becoming an auxiliary prediction that is detached from the generated motion, the decoder injects relation features directly into both human and object readout. It first maps each latent token back to the shared feature dimension and applies eight MoT blocks followed by temporal densification. At each decoded frame, relation queries gather human, object-pose, geometry, and relation features. Subsequently, the human-marker and object-pose queries read the decoded relation entries together with their own modality features. A final frame-wise coupling block updates all three sets of entries before the prediction heads recover $(\hat{H}, \hat{O}, \hat{R})$. In this way, the relation pathway contributes to trajectory reconstruction rather than serving only as an additional supervision target.

Reconstruction objectives. Given the heterogeneous fields represented by each stream, we apply masked Smooth L1 supervision field by field before combining their objectives. For each stream $q \in \{H, O, R\}$, the reconstruction term is

$$\mathcal{L}_q = \sum_{f \in \mathcal{F}_q} w_f\,\mathrm{SmoothL1}(\hat{f}, f).$$

Here, $\mathcal{F}_q$ denotes the supervised fields associated with stream q, and w_f is the corresponding field weight. To preserve human dynamics, the human objective supervises root-relative marker positions, the global root, foot heights, and their temporal differences. To retain rigid-object motion, the object objective supervises 6D rotation, translation, velocity, and acceleration. To maintain geometric coordination, the relation objective supervises marker–surface offsets, distances, canonical nearest points, their temporal differences, and the global root–object fields. In addition, the world-space marker term composes the predicted root-relative markers with the predicted root and constrains the resulting global positions. We use $\lambda_H = \lambda_O = \lambda_{\mathrm{world}} = 1$, $\lambda_R = 2$, and a final covariance weight of 10^{-2}. Position, rotation, distance, and root fields use unit weight. Canonical surface points and foot heights use weight 0.5, while first- and second-order differences use weights 0.25 and 0.10, respectively.

Noise-robust decoder fine-tuning. Since latents produced by the conditional flow model need not lie exactly on the encoded training manifold, we further adapt the decoder to small latent errors. After estimating the whitening transform, we freeze the tokenizer and encoder and optimize only the decoder. Gaussian perturbations are applied in the whitened latent space and mapped back through the fixed inverse transform, while 25% of examples retain unperturbed latents to preserve reconstruction at the clean endpoint. To introduce this perturbation gradually, we keep $\sigma_{\max} = 0$ for the first 2,000 optimization steps and then increase it linearly to 0.125 by step 20,000, sampling each perturbation scale uniformly from $[0, \sigma_{\max}]$. Finally, finite-difference velocity and acceleration losses on decoded human markers and object translations are added with coefficients 0.5 and 0.25, respectively. This procedure improves decoding under predicted latents without modifying the encoder or its target latent distribution.

A.2 Conditional Flow and Variable-Length Inference

Generative target and conditioning streams. To separate the state being generated from information available at inference, we use the fixed autoencoder only to construct the Flow Matching target. Its reconstruction latent Z^{rec} is mapped to Z^{white} by the whitening transform above, and the complete whitened field, including its dynamic tokens and four static latent-geometry tokens, is treated as the

generated state. Importantly, these latent-geometry tokens are part of the target and remain distinct from the object-geometry condition computed directly from the input G. Given the different structures of language and shape, text prompts are encoded by a frozen FLAN-T5-XL encoder (Chung et al., 2022), while sixteen learned queries aggregate the object BPS descriptor into a separate geometry-condition stream. The noisy latent, text, and geometry streams are then projected to a common model width before entering the conditional transformer.

Conditional modulation. Considering that latent state, language, and geometry play different roles, the conditional DiT retains route-specific adaptive normalization and feed-forward layers across its 12 blocks while using shared attention for cross-stream exchange. Shared query, key, value, and output projections provide the common exchange path, whereas zero-initialized route-specific residual projections allow each stream to specialize during optimization. To inject global semantic and geometric context, pooled text and geometry features modulate the latent route together with its token-wise flow time. Since flow time does not identify where a window belongs in the requested sequence, a separate progress condition is computed from the normalized window center and injected through the text and geometry routes. Shared attention then communicates this global reference to the latent route. The same definition is used for sampled training windows and sliding inference windows. Finally, temporal RoPE is restricted to latent tokens that carry physical frame indices, leaving text and static geometry conditions without temporal rotation.

Blockwise training schedule. To expose the model to causal histories containing both clean and noisy latent regions, we group dynamic tokens by temporal anchor, with eight consecutive anchors in each block. Every block independently samples $\tau_b \sim \mathcal{U}(0, 1)$, while the static latent-geometry tokens receive a separate flow time and the text and geometry conditions remain unperturbed. To prevent information from later blocks entering earlier predictions, we use the block-causal attention mask:

$$M_{ij} = 1 \quad \Longleftrightarrow \quad j \in \mathcal{C} \;\text{ or }\; b(j) \le b(i),$$

where $b(i)$ denotes the temporal block containing latent token i and $\mathcal{C}$ contains all text and geometry-condition tokens. Given that tokens within one anchor range represent the same local interval, attention remains bidirectional inside each block. To support classifier-free guidance without changing the target latent field, both conditions are retained with probability 0.7, while text-only, geometry-only, and unconditional examples are each sampled with probability 0.1.

Staggered rollout. Given that training uses a finite context while requested sequences may be much longer, inference advances a fixed-size window over the ordered latent blocks. Each window contains w blocks, of which h completed blocks are retained as fully denoised causal context. To stagger the denoising trajectories, let S be the number of integration steps assigned to each block and let $\Delta = \lceil S/(w-h) \rceil$ be the offset between consecutive block schedules. For $b \in \{1, \dots, w\}$, the flow time at solver step k is

$$\tau_b(k) = \mathrm{clip}\left(\frac{k - (b-1)\Delta}{S}, 0, 1 \right).$$

Only active blocks with $0 \le \tau_b < 1$ are updated. Using Euler integration, their update is

$$Z_b^{(k+1)} = Z_b^{(k)} + \frac{1}{S} v_{\psi,b}(Z^{(k)}, \boldsymbol{\tau}^{(k)} \mid c, G, p).$$

Once the earliest active block reaches the data endpoint, it becomes fixed causal context and the window advances. Since object shape is constant across the rollout, static latent-geometry tokens are integrated during the first block schedule and remain fixed thereafter. To recover motion from the extended latent sequence, the decoder processes overlapping latent windows, after which the decoded blocks are assembled in temporal order and only the final boundary is cropped to the requested duration.

A.3 Architecture and Optimization Settings

Autoencoder configuration. For reproducibility, the autoencoder uses a model width of 256 and four attention heads. Its shared latent field has 708 tokens of dimension 16. The MoT encoder has six layers, the decoder has eight MoT layers, and the temporal densifier has four layers. To optimize this model, we use 100,000 steps of AdamW $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 10^{-4}, a peak learning rate of 10^{-4}, 2,000 warm-up steps, and cosine decay to 10^{-5}. The covariance regularizer is activated after 10,000 steps and linearly reaches its full weight at step 30,000. We estimate the ZCA

transform with $\epsilon = 10^{-6}$. The noise-robust decoder is then fine-tuned for an additional 30,000 steps, with 500 warm-up steps and cosine decay from 2×10^{-5} to 5×10^{-6}. Both phases use a per-GPU batch size of 6 without gradient accumulation on three NVIDIA GeForce RTX 5080 GPUs, giving an effective batch size of 18. The ZCA transform and decoder fine-tuning are initialized from the final autoencoder checkpoint. We retain the decoder checkpoint with the lowest validation reconstruction drift under latent perturbations of $\sigma = 0.125$.

Conditional-flow configuration. The conditional flow model uses width 384, 12 layers, and six attention heads. To learn the velocity field, we optimize it for 110,000 steps with AdamW $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 10^{-2}, and learning rate 10^{-4}. The learning rate increases linearly during the first 2,000 steps, remains constant through step 100,000, and follows a cosine decay to 10^{-5} over the final 10,000 steps. We maintain an exponential moving average of the trainable model with decay 0.9999. To reduce training memory and computation, all models use bfloat16 mixed precision. Unless stated otherwise, inference uses Euler integration with $S = 32$, $w = 4$, and $h = 1$, corresponding to a 128-frame temporal context with one fully denoised history block retained as causal context. Training uses a per-GPU batch size of 16 with two-step gradient accumulation on three NVIDIA GeForce RTX 5080 GPUs, yielding an effective batch size of 96. The conditional flow model is trained with the selected noise-robust decoder checkpoint and the fixed ZCA transform. All reported generation results use the EMA weights from the final 110,000-step checkpoint.

B Evaluation Details

B.1 Baseline Adaptation Protocol

To evaluate methods with different human representations in a common space, we convert their released outputs to the official InterAct 77-marker layout. For CHOIS, the generated SMPL-X parameters are passed through its body-model forward process. For HOI-Diff and LIGHT, their generated 22- and 52-joint trajectories are fitted to SMPL-H using the corresponding released fitting procedures. Since Text2HOI generates only articulated hands, its generated wrist states and hand poses drive a neutral SMPL-X carrier whose remaining body degrees of freedom are fixed. In every case, the 77 markers are selected using the official InterAct marker vertex indices.

We decode each generated object trajectory with the corresponding baseline's released transformation convention and use the InterAct object samples when constructing evaluator inputs. When a baseline and InterAct use different world frames, a single rigid transformation estimated from their initial human and object states maps the generated sequence to the InterAct coordinate system. No future ground-truth frames are used by this alignment, and the generated temporal evolution is otherwise unchanged. All converted records are stored at 30 fps and evaluated with the same frozen evaluator described below.

B.2 InterAct Feature-Space Protocol

Following the official InterAct evaluation protocol (Xu et al., 2025), we use its frozen marker–BPS evaluator, training-set normalization statistics, and BPS basis for all feature-space metrics. At each frame, the evaluator input concatenates the 3D coordinates of 77 body-surface markers with 128 BPS displacement vectors computed from the translated object basis to its nearest surface points, yielding a 615-dimensional representation. The official motion and text encoders map a generated interaction and its caption to 256-dimensional features.

We retain the standard definitions of FID, R-Precision, MM-Dist, Diversity, and MModality used by InterAct and text-to-motion evaluation (Guo et al., 2022). FID compares the Gaussian statistics of generated and reference motion features. R-Precision ranks motion features by their Euclidean distance to a text feature within retrieval pools of 32 samples and reports Top-1, Top-2, and Top-3 accuracy. MM-Dist is the mean distance between matched text and motion features, while Diversity averages the distance between 300 randomly sampled pairs of generated motion features. For MModality, we generate 30 samples for each evaluated condition and average 10 randomly drawn within-condition pairs. All other feature metrics are reported as the mean and sample standard deviation over three independent generation runs with seeds 20260622, 1, and 2. MModality is evaluated from one repeated-sampling run, with uncertainty estimated by 1,000 condition-level bootstrap resamples. Outputs from adapted baselines are first converted to the common 77-marker

representation and then evaluated by the same frozen InterAct evaluator. Each adapted method is scored on all conditions supported by its released model and data subset.

B.3 Physical and Temporal Metrics

Body–object penetration. For each valid frame t and body marker j, we query its signed distance to the posed object mesh. Let $q_{t,j} = \max(\mathrm{sdf}_{O_t}(\boldsymbol{p}^H_{t,j}), 0)$, where the signed distance is positive inside the object. The penetration ratio is the fraction of valid marker queries for which $q_{t,j} > 0$. The mean, 95th percentile, and maximum penetration depths are computed over the penetrating queries and converted from meters to centimeters. We use the available decimated object meshes and uniformly sample approximately 48 frames per sequence for signed-distance queries. Sequences without a valid object mesh are excluded from this metric.

Relation coordination. To measure contact without using the reference object trajectory for the generated motion, we compute marker–surface distances independently for each generated human–object pair and each ground-truth pair. Let $d_t = \min_{j,k} \|\boldsymbol{p}^H_{t,j} - \boldsymbol{s}^O_{t,k}\|_2$ and $c_t = \mathbf{1}[d_t < \epsilon]$, where $\epsilon = 2$ cm. Contact recall is the fraction of ground-truth contact frames for which the generated pair also satisfies $c_t = 1$, and contact distance is the generated d_t averaged over those reference contact frames. Hand-contact metrics use the same definitions with the marker set restricted to the hands. To distinguish reduced penetration from simply keeping the object away from the body, we additionally measure the fraction of object travel occurring across consecutive frames for which neither endpoint is in contact. Finally, we summarize relation dynamics for each sequence by the temporal mean of $|d_t - d_{t-1}|$ and report the 1-Wasserstein distance between the generated and ground-truth distributions.

Motion stability. For a temporally aligned point trajectory $\boldsymbol{x}_{t,j}$, we compute the third-order finite difference

$$\Delta^3 \boldsymbol{x}_{t,j} = \boldsymbol{x}_{t,j} - 3\boldsymbol{x}_{t-1,j} + 3\boldsymbol{x}_{t-2,j} - \boldsymbol{x}_{t-3,j}.$$

The per-frame jerk is the mean Euclidean norm of this quantity over points, reported in cm/frame^3. Human jerk is evaluated on the 77 body-surface markers, while object jerk is evaluated on temporally corresponding object surface points. We report the temporal mean, the maximum as Peak Jerk, and the sum over valid frames as AUJ. Since realistic motions need not minimize jerk, these statistics are compared with their ground-truth counterparts rather than ranked monotonically.

B.4 Long-Horizon Protocol

To assess local quality within extended rollouts, we divide each valid sequence into complete, non-overlapping 128-frame subsequences. Each subsequence inherits the caption of its parent interaction and is evaluated with the same InterAct feature-space protocol. We additionally extract 64-frame windows centered at every 128-frame continuation boundary, using 32 frames on each side. Incomplete subsequences and boundary windows are discarded. FID and Diversity are computed over the boundary windows, while Peak Jerk and AUJ are computed from their human marker trajectories. The values in the main paper are averaged over the same three generation seeds used for the standard feature metrics.

C Supplementary Experiments

C.1 Additional Evaluations

Inference efficiency. With longer requested sequences, an effective rollout should extend computation without proportionally expanding memory. Table A shows that runtime follows the increasing number of velocity evaluations, whereas peak memory remains nearly unchanged. This scaling follows from the fixed-size context described in Sec. A.2, which allows the generated duration to increase without a corresponding growth in memory.

Subset-level generation quality. Beyond aggregate benchmark performance, the BEHAVE and OMOMO subsets test whether the observed improvement persists across distinct source domains. Under the protocol of Sec. B, Table B shows that ULIFlow retains the strongest overall distributional

Table A: **Quantitative comparison** of inference efficiency across requested sequence lengths on the InterAct dataset (Xu et al., 2025). Runtime and peak memory are measured on one NVIDIA GeForce RTX 5080.

Frames	Blocks	Velocity evaluations	Time (s)	FPS	Peak VRAM (GiB)
128	4	65	2.817 ± 0.004	45.43	3.267
256	8	109	4.827 ± 0.007	53.04	3.270
384	12	153	6.927 ± 0.038	55.43	3.274
512	16	197	9.367 ± 0.312	54.66	3.277

Table B: **Quantitative comparison** of generation quality on the BEHAVE (Bhatnagar et al., 2022) and OMOMO (Li et al., 2023) subsets of the InterAct dataset with prior works.

BEHAVE subset

Method	FID ↓	R-Precision ↑			MM-Dist ↓	Diversity →
		Top 1	Top 2	Top 3		
Ground Truth	0.000 ± 0.000	0.887 ± 0.000	0.956 ± 0.000	0.969 ± 0.000	3.242 ± 0.000	10.805 ± 0.000
HOI-Diff (Peng et al., 2025)	14.440 ± 0.008	0.138 ± 0.000	0.256 ± 0.006	0.358 ± 0.004	10.230 ± 0.000	6.253 ± 0.002
InterAct (Xu et al., 2025)	5.080 ± 0.079	0.698 ± 0.010	0.865 ± 0.010	0.917 ± 0.018	5.805 ± 0.069	10.679 ± 0.036
LIGHT (Wang et al., 2026b)	4.607 ± 0.151	0.556 ± 0.017	0.769 ± 0.027	0.881 ± 0.033	6.252 ± 0.043	10.518 ± 0.038
ULIFlow	$\mathbf{3.396 \pm 0.122}$	$\mathbf{0.748 \pm 0.024}$	$\mathbf{0.892 \pm 0.007}$	$\mathbf{0.948 \pm 0.014}$	$\mathbf{5.308 \pm 0.028}$	$\mathbf{10.807 \pm 0.038}$

OMOMO subset

Method	FID ↓	R-Precision ↑			MM-Dist ↓	Diversity →
		Top 1	Top 2	Top 3		
Ground Truth	0.000 ± 0.000	0.812 ± 0.000	0.956 ± 0.000	0.990 ± 0.000	3.275 ± 0.000	10.891 ± 0.000
CHOIS (Li et al., 2024)	6.914 ± 0.147	0.644 ± 0.005	0.870 ± 0.007	0.958 ± 0.010	5.278 ± 0.006	9.378 ± 0.028
InterAct (Xu et al., 2025)	2.356 ± 0.079	$\mathbf{0.838 \pm 0.007}$	$\mathbf{0.975 \pm 0.007}$	$\mathbf{0.996 \pm 0.004}$	3.612 ± 0.022	10.744 ± 0.006
LIGHT (Wang et al., 2026b)	29.684 ± 0.231	0.337 ± 0.009	0.534 ± 0.016	0.646 ± 0.010	8.063 ± 0.022	9.729 ± 0.013
ULIFlow	$\mathbf{1.367 \pm 0.042}$	0.796 ± 0.004	0.958 ± 0.001	0.989 ± 0.001	$\mathbf{3.577 \pm 0.015}$	$\mathbf{10.990 \pm 0.043}$

quality while remaining competitive in text–interaction alignment and diversity. Since both subsets use the same trained model and evaluator, the resulting pattern is not tied to the aggregate composition of the full benchmark.

C.2 Ablation Studies

Reconstruction fidelity. To test whether relation modeling consumes capacity needed for human and object reconstruction, we compare autoencoders with nearly matched parameter counts in Table 5. The complete model retains marker and object errors comparable to the variant without the relation branch, indicating that explicit relation modeling does not materially compromise reconstruction. In contrast, replacing MoT with a flat Transformer substantially worsens both, despite a similar parameter budget.

Latent utilization. Prior work indicates that latent organization and utilization, beyond reconstruction fidelity, matter for downstream generative learning (Zhang et al., 2026; Yue et al., 2026; Yao et al., 2025; Li et al., 2026). To assess this property, we report the entropy-based effective rank of the latent covariance spectrum, normalized by channel dimension (Roy & Vetterli, 2007), in Table 5. The complete model distributes variance more evenly across latent directions, whereas removing the relation branch or latent regularization reduces rank and the flat Transformer exhibits severe concentration. This broader latent utilization accompanies the preserved reconstruction quality, while the downstream comparison in Table 1 evaluates its generative impact.

Generation across requested lengths. To examine how the rollout mechanisms behave as the requested duration increases, we evaluate progress conditioning and blockwise forcing across three sequence lengths in Table C. Although all variants remain competitive within the training horizon, the complete design becomes increasingly reliable beyond it across distributional quality and text alignment. Coupling the global temporal reference described in Sec. A.2 with blockwise conditioning therefore preserves coherent continuation as more latent blocks are generated.

Continuation-boundary quality. At continuation boundaries, local artifacts directly reveal whether separately generated blocks form a smooth trajectory. Table D reports feature-space and temporal statistics in windows centered on these boundaries. Across both extended durations, the complete

Table C: **Quantitative comparison** of long-horizon generation across requested sequence lengths on the InterAct dataset (Xu et al., 2025). Metrics are computed over non-overlapping 128-frame subsequences.

Frames	Progress condition	Blockwise forcing	FID ↓	R-Precision ↑			MM-Dist ↓	Diversity →
				Top 1	Top 2	Top 3		
128	Ground Truth		0.000 ±0.000	0.635 ±0.000	0.781 ±0.000	0.875 ±0.000	6.738 ±0.000	11.171 ±0.000
128	×	✓	7.319 ±0.219	0.691 ±0.024	**0.830 ±0.012**	**0.892 ±0.026**	6.172 ±0.032	11.419 ±0.066
128	✓	×	**6.837 ±0.281**	0.688 ±0.028	0.806 ±0.033	0.875 ±0.021	6.099 ±0.058	**11.229 ±0.029**
128	✓	✓	7.712 ±0.166	**0.694 ±0.006**	**0.830 ±0.012**	0.878 ±0.016	**6.093 ±0.025**	11.236 ±0.035
256	Ground Truth		0.000 ±0.000	0.563 ±0.000	0.719 ±0.000	0.811 ±0.000	6.849 ±0.000	11.076 ±0.000
256	×	✓	4.888 ±0.227	0.575 ±0.005	0.740 ±0.011	0.827 ±0.002	6.466 ±0.020	11.271 ±0.033
256	✓	×	3.862 ±0.171	0.612 ±0.008	0.763 ±0.010	0.842 ±0.000	6.249 ±0.027	11.073 ±0.029
256	✓	✓	**3.789 ±0.029**	**0.615 ±0.005**	**0.782 ±0.009**	**0.850 ±0.008**	**6.190 ±0.034**	**11.118 ±0.032**
384	Ground Truth		0.000 ±0.000	0.524 ±0.000	0.706 ±0.000	0.788 ±0.000	6.961 ±0.000	10.981 ±0.000
384	×	✓	4.383 ±0.102	0.547 ±0.006	0.753 ±0.004	0.824 ±0.011	6.566 ±0.027	11.223 ±0.013
384	✓	×	3.639 ±0.014	0.576 ±0.011	0.770 ±0.004	0.838 ±0.006	**6.361 ±0.013**	11.059 ±0.044
384	✓	✓	**3.367 ±0.011**	**0.591 ±0.003**	**0.776 ±0.001**	**0.841 ±0.003**	6.394 ±0.012	**11.062 ±0.015**

Table D: **Quantitative comparison** of continuation-boundary quality across requested sequence lengths on the InterAct dataset (Xu et al., 2025). Metrics are computed over 64-frame windows centered at continuation boundaries.

Frames	Progress condition	Blockwise forcing	FID ↓	Diversity →	Peak jerk →	AUJ →
256	Ground Truth		0.000 ±0.000	11.066 ±0.000	0.852 ±0.000	17.320 ±0.000
256	×	✓	5.938 ±0.136	11.203 ±0.070	1.275 ±0.044	40.060 ±0.491
256	✓	×	**4.853 ±0.169**	11.056 ±0.015	1.322 ±0.023	38.877 ±0.076
256	✓	✓	4.917 ±0.146	**11.027 ±0.054**	**1.192 ±0.004**	**36.580 ±0.165**
384	Ground Truth		0.000 ±0.000	11.073 ±0.000	0.861 ±0.000	17.204 ±0.000
384	×	✓	4.487 ±0.117	11.210 ±0.030	1.263 ±0.033	40.310 ±0.419
384	✓	×	**3.554 ±0.104**	11.072 ±0.026	1.299 ±0.006	39.502 ±0.102
384	✓	✓	3.800 ±0.079	**10.977 ±0.021**	**1.176 ±0.017**	**36.815 ±0.241**

Table E: **Quantitative comparison** of latent-noise robustness on the InterAct dataset (Xu et al., 2025). Latent perturbations are applied in the whitened space before decoding.

Decoder	σ	Marker (cm) ↓	Object trans. (cm) ↓	Object rot. (deg) ↓	Acceleration	Jerk
Original	0.00	1.348	**1.263**	**1.520**	0.786	1.328
Noise-robust		**1.344**	1.457	1.606	**0.462**	**0.682**
Original	0.05	2.153	2.459	2.285	1.002	1.700
Noise-robust		**1.805**	**2.039**	**2.245**	**0.496**	**0.732**
Original	0.10	3.530	4.339	3.567	1.388	2.370
Noise-robust		**2.688**	**3.149**	**3.350**	**0.578**	**0.856**
Original	0.20	6.550	8.271	6.357	2.332	4.017
Noise-robust		**4.753**	**5.705**	**5.831**	**0.824**	**1.234**

schedule most consistently suppresses abrupt motion changes while retaining competitive distributional quality. Together with the sequence-level comparison in Table C, these results show that progress conditioning and blockwise forcing address complementary aspects of long-horizon generation.

Decoder robustness to latent perturbations. Because generated latents need not lie exactly on the encoded manifold, we perturb the whitened representation before applying the decoder in Eq. (4). Table E shows that the noise-robust decoder degrades more gracefully as the perturbation grows and substantially suppresses acceleration and jerk while retaining comparable clean reconstruction. Exposure to a neighborhood around the encoded manifold allows the post-training procedure in Sec. A.1 to broaden the region that can be decoded into plausible human–object dynamics.

Subset-level architecture ablation. Across the same source domains, Table F isolates whether the latent-modeling contributions remain consistent. Replacing MoT with a conventional flat architecture yields the most persistent degradation, while relation modeling and latent regularization affect complementary aspects of generation. In line with the modality-aware encoding and decoding in Eqs. (3) and (4), the complete design maintains the strongest overall balance across the evaluated criteria.

Table F: **Quantitative comparison** of latent autoencoder variants on the BEHAVE (Bhatnagar et al., 2022) and OMOMO (Li et al., 2023) subsets of the InterAct dataset.

BEHAVE subset

Method	FID ↓	R-Precision ↑			MM-Dist ↓	Diversity →
		Top 1	Top 2	Top 3		
Ground Truth	0.000 ± 0.000	0.887 ± 0.000	0.956 ± 0.000	0.969 ± 0.000	3.242 ± 0.000	10.805 ± 0.000
ULIFlow w/o MoT structure	5.251 ± 0.093	0.660 ± 0.026	0.827 ± 0.014	0.908 ± 0.007	5.973 ± 0.008	10.454 ± 0.021
ULIFlow w/o relation branch	$\mathbf{3.025 \pm 0.140}$	0.717 ± 0.007	0.865 ± 0.010	0.923 ± 0.004	5.442 ± 0.057	10.719 ± 0.039
ULIFlow w/o latent regularization	3.347 ± 0.035	0.740 ± 0.010	0.890 ± 0.004	$\mathbf{0.950 \pm 0.006}$	5.348 ± 0.050	10.728 ± 0.048
ULIFlow	3.396 ± 0.122	$\mathbf{0.748 \pm 0.024}$	$\mathbf{0.892 \pm 0.007}$	0.948 ± 0.014	$\mathbf{5.308 \pm 0.028}$	$\mathbf{10.807 \pm 0.038}$

OMOMO subset

Method	FID ↓	R-Precision ↑			MM-Dist ↓	Diversity →
		Top 1	Top 2	Top 3		
Ground Truth	0.000 ± 0.000	0.812 ± 0.000	0.956 ± 0.000	0.990 ± 0.000	3.275 ± 0.000	10.891 ± 0.000
ULIFlow w/o MoT structure	4.728 ± 0.084	0.730 ± 0.008	0.921 ± 0.010	0.972 ± 0.008	4.565 ± 0.053	10.812 ± 0.047
ULIFlow w/o relation branch	1.866 ± 0.161	0.782 ± 0.008	0.949 ± 0.006	0.984 ± 0.003	3.680 ± 0.026	$\mathbf{10.901 \pm 0.021}$
ULIFlow w/o latent regularization	1.387 ± 0.029	0.791 ± 0.003	0.953 ± 0.004	0.988 ± 0.002	3.610 ± 0.004	10.981 ± 0.035
ULIFlow	$\mathbf{1.367 \pm 0.042}$	$\mathbf{0.796 \pm 0.004}$	$\mathbf{0.958 \pm 0.001}$	$\mathbf{0.989 \pm 0.001}$	$\mathbf{3.577 \pm 0.015}$	10.990 ± 0.043

Table G: **Quantitative comparison** of physical coordination for the relation-pathway ablation on the InterAct dataset (Xu et al., 2025). Contact metrics use a 2-cm threshold, and unsupported travel is the fraction of object displacement occurring without body contact at either endpoint.

Method	Pen. ratio (%) ↓	Mean depth (cm) ↓	P95 depth (cm) ↓	Max depth (cm) ↓
Ground Truth	0.295 ± 0.000	0.977 ± 0.000	2.082 ± 0.000	2.339 ± 0.000
ULIFlow w/o relation branch	0.583 ± 0.043	1.364 ± 0.474	2.650 ± 0.526	3.116 ± 0.541
ULIFlow	$\mathbf{0.379 \pm 0.042}$	$\mathbf{1.312 \pm 0.572}$	$\mathbf{2.186 \pm 0.688}$	$\mathbf{2.516 \pm 0.713}$

Method	Contact recall (%) ↑	Hand recall (%) ↑	Contact dist. (cm) ↓	Unsupported travel (%) ↓	Relation-velocity gap (cm/frame) ↓
Ground Truth	100.000 ± 0.000	100.000 ± 0.000	1.130 ± 0.000	27.764 ± 0.000	0.000 ± 0.000
ULIFlow w/o relation branch	51.128 ± 0.275	46.176 ± 0.145	4.042 ± 0.056	50.839 ± 0.475	0.144 ± 0.003
ULIFlow	$\mathbf{69.680 \pm 0.122}$	$\mathbf{68.850 \pm 0.436}$	$\mathbf{2.697 \pm 0.048}$	$\mathbf{28.624 \pm 0.330}$	$\mathbf{0.030 \pm 0.000}$

Effect of the relation pathway on physical coordination. To directly assess whether relation modeling improves coordination rather than only feature-space generation quality, we compare the complete model with the existing ablation that removes relation tokenization, relation reconstruction, and relation-coupled readout together. Table G uses 1,189 conditions shared by both variants for contact-based metrics and a fixed common subset of 125 sequences with valid object meshes for penetration. The complete relation pathway improves contact and hand-contact recall while reducing both contact distance and unsupported object travel. Its lower penetration therefore does not arise from separating the body and object. Moreover, the smaller relation-velocity gap shows that the benefit extends from static proximity to the temporal evolution of their relative geometry.

Contact-threshold sensitivity. Since binary contact measurements depend on the selected proximity threshold, Table H repeats the comparison at 1, 2, and 5 cm. The complete model improves precision, recall, and F1 at every threshold, with the same trend for hand contact. The coordination advantage is therefore consistent across both strict and permissive contact definitions.

Table H: **Quantitative comparison** of contact-threshold sensitivity on the InterAct dataset (Xu et al., 2025). All contact decisions compare the generated human with its generated object; ground truth is used only to define the reference contact frames.

Threshold (cm)	Method	Precision (%) ↑	Recall (%) ↑	F1 (%) ↑	Hand F1 (%) ↑
1	ULIFlow w/o relation branch	38.106 ± 0.673	28.172 ± 0.496	32.395 ± 0.568	33.304 ± 0.555
1	ULIFlow	$\mathbf{39.658 \pm 0.015}$	$\mathbf{44.461 \pm 0.393}$	$\mathbf{41.922 \pm 0.182}$	$\mathbf{43.098 \pm 0.094}$
2	ULIFlow w/o relation branch	61.607 ± 0.276	51.128 ± 0.275	55.880 ± 0.266	54.041 ± 0.233
2	ULIFlow	$\mathbf{64.903 \pm 0.101}$	$\mathbf{69.680 \pm 0.122}$	$\mathbf{67.207 \pm 0.054}$	$\mathbf{67.464 \pm 0.272}$
5	ULIFlow w/o relation branch	82.087 ± 0.056	78.855 ± 0.455	80.438 ± 0.263	75.167 ± 0.298
5	ULIFlow	$\mathbf{84.974 \pm 0.081}$	$\mathbf{88.498 \pm 0.333}$	$\mathbf{86.700 \pm 0.194}$	$\mathbf{84.930 \pm 0.122}$