TRIFACTORHOI: STRUCTURED DIFFUSION FOR CONTROLLABLE AND CONTACT–CONSISTENT HUMAN–OBJECT INTERACTION SYNTHESIS

Anonymous authors
Paper under double-blind review

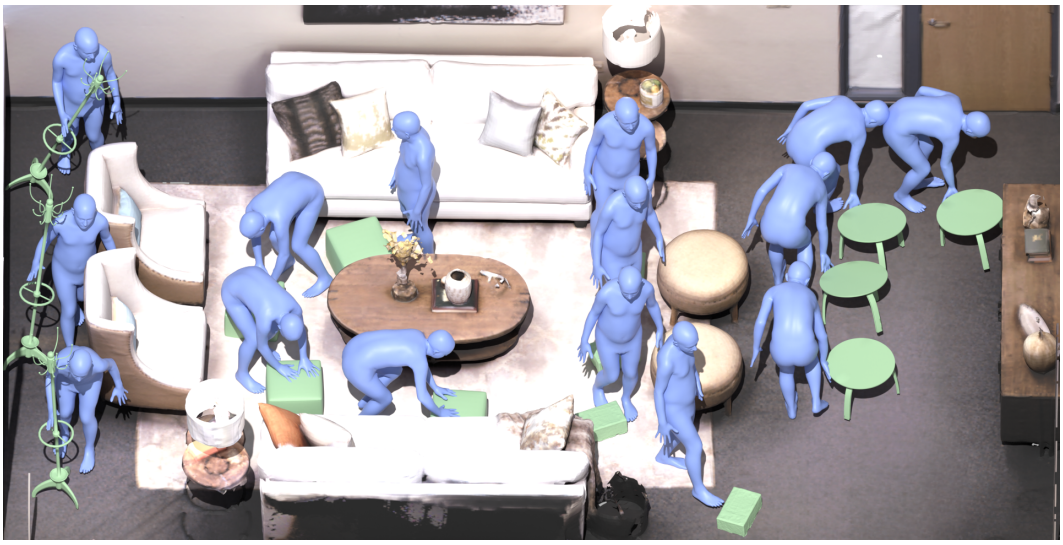

Figure 1: Diverse human–object interactions generated by TriFactorHOI. Blue human poses and green manipulated objects illustrate whole-body coordination during manipulation.

ABSTRACT

Recent diffusion models have advanced human–object interaction (HOI) generation. However, generating coordinated human–object motion from sparse waypoints while following user-specified routes and maintaining stable contacts remains challenging. We propose TriFactorHOI, a structured diffusion framework that organizes HOI generation through object path, part, and contact factors. For trajectory control, a reference path interpolated from sparse waypoints serves both as a generation condition and as the baseline for planar object translation, with learned residuals accommodating interaction-related deviations. A part-based spatiotemporal denoiser preserves separate temporal and part axes to model within-part dynamics and cross-part human–object coordination. Hand–object and foot–ground guidance refine contacts during denoising. After sampling, a full-sequence foot-refinement network predicts support probabilities and correction confidence for world-space foot anchoring and temporal inverse kinematics. Experiments on FullBodyManipulation demonstrate improved waypoint adherence, motion-distribution quality, and contact stability. With the reference-path condition held fixed, residual prediction reduces waypoint error and FID relative to absolute planar translation. Tests on unseen 3D-FUTURE objects without retraining further support generalization to new object geometries. The code and processed datasets will be made publicly available.

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

Generating realistic 3D human motion from natural-language instructions is a long-standing goal in computer vision, graphics, and embodied intelligence (Guo et al., 2022). Recent diffusion models have substantially improved the diversity and semantic fidelity of synthesized motion (Zhang et al., 2022; Tevet et al., 2023). Advances in 3D asset (Xiang et al., 2025) and language-conditioned scene generation (Lin & Mu, 2024; Yang et al., 2024b) further motivate animating humans within diverse virtual environments. Human–object interaction (HOI) additionally requires synchronized body and object motion, geometry awareness, and plausible hand–object and foot–ground contacts (Li et al., 2023). Sparse object waypoints specify the intended route without prescribing intermediate poses. CHOIS (Li et al., 2024) conditions joint motion diffusion on language, object geometry, initial states, and these waypoints. Waypoint matching alone does not establish realistic HOI. The Lin-OMOMO baseline reported by CHOIS achieves zero waypoint error through interpolation but has higher FID and lower contact F1 than CHOIS itself. This motivates using the interpolated route as a reference for joint motion generation rather than fixing it as the final object trajectory.

This joint generation setting poses several linked modeling demands. Articulated body parts and the rigid object must evolve together despite their different kinematics: arms manipulate the object while root and leg movements support transport. Persistent foot support also requires world-space consistency across frames (Kovar et al., 2002). These demands motivate a route-based object representation, coupled part modeling, and contact-aware refinement.

Prior work separates interaction semantics from dynamics (Xu et al., 2024) or trajectory planning from action synthesis (Jung et al., 2026). We pursue *structured decoupling* across three interrelated factors—**object paths**, **parts**, and **contacts**—using representations aligned with their roles. Here, factors denote distinct modeling and inference roles within one coupled generator. Attention, interaction guidance, and kinematic refinement coordinate them.

Based on this principle, we introduce **TriFactorHOI**, a structured diffusion framework for controllable and contact-consistent HOI generation. The *path factor* constructs a dense reference route from user-specified sparse waypoints and uses it both to condition generation and as the baseline for planar object translation. This places the requested path directly in the motion representation, while learned residuals allow local interaction-related adjustments. The *part factor* partitions the human and object state into seven semantic components—root, torso, left and right arms, left and right legs, and object. A part-based spatiotemporal denoiser models within-part dynamics along time and cross-part coordination across the body and object. The *contact factor* addresses both primary physical interfaces in full-body HOI. Geometry-aware hand guidance encourages hand–object proximity and palm alignment while penalizing penetration. Foot–ground guidance acts during denoising; after sampling, full-sequence support and correction-confidence prediction, persistent world-space planar anchors, and temporal inverse kinematics stabilize foot–ground contact. Together, the Path and Part Factors generate waypoint-aligned HOI motion; the Contact Factor refines hand contact and foot height during denoising, then stabilizes foot support after sampling.

We train the TriFactorHOI generator on FullBodyManipulation (Li et al., 2023) and evaluate it on unseen objects in this benchmark and unseen 3D-FUTURE objects (Fu et al., 2021) for cross-dataset testing. TriFactorHOI shows clear advantages in trajectory control, contact quality, semantic alignment, and motion-distribution quality. In a comparison using the same sparse waypoints, the waypoint-relative path design lowers both waypoint error and FID. A matched-token ablation finds lower foot sliding for semantic parts than free tokens under the same full inference pipeline (Appendix A.3.1); contact ablations assess refinement.

Our contributions are summarized as follows:

- We introduce TriFactorHOI, a structured diffusion framework that organizes controllable human–object interaction generation around object paths, semantic parts, and physical contacts. This design assigns distinct roles to trajectory control, coordinated motion modeling, and contact refinement within one coupled generator.
- We combine waypoint-relative object-motion generation, part-based spatiotemporal denoising, and contact guidance with support-aware foot refinement for controllable, coordinated, and contact-consistent human–object motion.

- Experiments on FullBodyManipulation demonstrate improved waypoint adherence, motion-distribution quality, and contact stability. Ablations evaluate the component designs, while tests on unseen 3D-FUTURE objects without retraining support generalization to new object geometries.

2 Related Work

2.1 Text-Driven Human Motion Generation

Text-driven human motion generation has been explored with variational models (Guo et al., 2022), autoregressive generation (Zhang et al., 2023a), masked Transformers (Guo et al., 2024), and diffusion-based approaches (Zhang et al., 2022; Tevet et al., 2023; Chen et al., 2023). Guided diffusion supports flexible spatial and temporal constraints (Karunratanakul et al., 2023; Xie et al., 2024), while continuous motion–language models separate modality-specific streams and share attention for controlled fusion (Zhu et al., 2026). Long sequences are addressed through motion composition (Barquero et al., 2024) and autoregressive streaming (Zhao et al., 2025; Xiao et al., 2025). Structured modeling includes fine-grained spatiotemporal features (Zhang et al., 2023b), separate temporal and spatial state-space blocks (Zhang et al., 2024c), and part-level interaction semantics (Fan et al., 2024). SALAD (Hong et al., 2025) retains seven skeletal units in a human-motion latent and applies temporal followed by skeletal attention. Body-part tokenization (Cao et al., 2026) and part-structured codebook-space generation (Fang et al., 2026) also preserve local body dynamics.

We build on this axis-wise modeling principle for HOI, jointly updating six human-part tokens and one rigid-object motion token during denoising.

2.2 Human–Object Interaction Generation and Trajectory Control

HOI generation spans distinct conditioning protocols. Some methods jointly synthesize human and object motion from text and geometry (Diller & Dai, 2024; Peng et al., 2025). ChainHOI (Zeng et al., 2025) uses joint–object nodes and semantic kinematic-chain queries to aggregate structural information before fusing it back into the node representation. Other methods condition human motion on complete object trajectories or dynamic target geometry (Li et al., 2023; Cong et al., 2025). CHOIS instead uses sparse object waypoints (Li et al., 2024). InterDreamer decouples semantics and dynamics (Xu et al., 2024), while DecHOI separates trajectory planning from action synthesis without prescribed intermediate waypoints (Jung et al., 2026). These inputs and tasks are not directly interchangeable evaluation protocols. At the pose level, TriDi jointly models human, object, and interaction modalities (Petrov et al., 2025). Sequence generation also benefits from part-affordance graphs (Li & Dai, 2026), contact planning followed by noise-space optimization (Ron et al., 2025), and staged object-motion, contact-field, and body-motion generation for multi-object and articulated interactions (Peng et al., 2026). MaMi-HOI (Wang et al., 2026) conditions a parallel kinematic branch on sparse waypoints and applies geometry-aware feature corrections. Other models exploit interaction dynamics (Wu et al., 2025a), visual priors (Cai et al., 2026), and discrete contact-aware motion tokens (Ji et al., 2026).

Together, these studies show the need to coordinate object routes, human–object motion, and contact. Our time-by-part tokens carry noisy human and object motion throughout denoising and are decoded by semantic part. Hand–object and foot–ground contacts receive dedicated guidance and refinement.

2.3 Contact-Aware and Physically Plausible Motion Synthesis

Physical plausibility has been addressed through geometric/contact training losses (Tevet et al., 2023; Diller & Dai, 2024), sampling guidance (Li et al., 2024; Peng et al., 2025), foot correction during denoising (Huang et al., 2024), and optimization-based IK (Kovar et al., 2002; Mourot et al., 2022). Hand–object models exploit local geometry (Zhou et al., 2024), contact priors (Cha et al., 2024; Zhang et al., 2025), and correspondence-guided refinement (Zhang et al., 2024a), complementing whole-body grasp synthesis (Taheri et al., 2022). Learned correctors also address noisy hand–object trajectories (Liu & Yi, 2024) and HOI forecasting (Xu et al., 2023). Beyond geometry, FORCE models physical object attributes and human force–resistance relations (Zhang et al.,

2024b), while simulation-based projection (Yuan et al., 2023) and learned controllers (Wang et al., 2023; Xu et al., 2025; 2026) address interaction dynamics. Collision and reachability guidance also supports physically interactable scene synthesis (Yang et al., 2024a).

Stable foot support additionally requires temporally persistent contact, which per-frame proximity alone does not encode. This motivates us to combine geometric hand guidance with temporally coherent foot refinement, improving contact while preserving the intended interaction.

3 METHOD

3.1 PRELIMINARIES

Given object geometry $\mathcal{G}$, a text instruction $\mathcal{I}$, an initial human–object state $\mathcal{M}_1$, and sparse object waypoints $\mathcal{W}$, we generate a T-frame sequence $\mathcal{M} = \{\mathcal{M}_t\}_{t=1}^{T}$ of synchronized full-body motion and rigid object poses. The sequence should follow the text and waypoints while maintaining plausible human–object and foot–ground contact. We adopt conditional motion diffusion: letting $\mathbf{x}_0$ denote the clean motion and $\mathbf{x}_n$ its noisy state at diffusion step n, the denoiser f_θ predicts

$$\hat{\mathbf{x}}_0 = f_\theta(\mathbf{x}_n, n, \mathcal{C}), \qquad \mathcal{C} = \{\mathcal{I}, \mathcal{G}, \mathcal{M}_1, \mathcal{W}\}. \tag{1}$$

3.2 METHOD OVERVIEW

As shown in Fig. 2, TriFactorHOI integrates three complementary factors with distinct roles in one coupled generator. The Path Factor constructs a dense reference route from user-specified sparse waypoints and represents planar object translation relative to it. The Part Factor jointly denoises semantic body parts and rigid object motion, conditioned on the text, object geometry, initial state, and reference path. The Contact Factor combines interaction-aware guidance during reverse diffusion with support-aware foot refinement after sampling, improving hand–object interaction and foot–ground stability.

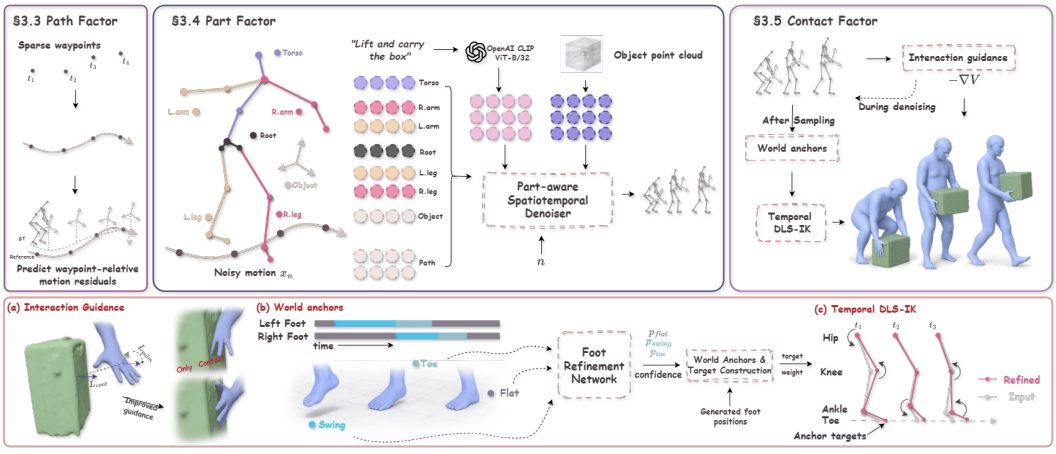

Figure 2: Overview of TriFactorHOI. Top: the Path Factor constructs a reference path for waypoint-relative object-motion generation; the Part Factor models coordinated human–object motion with a part-aware spatiotemporal denoiser; and the Contact Factor applies interaction guidance during denoising and foot refinement after sampling. Bottom: (1) geometry-aware hand–object guidance; (2) support-phase and confidence prediction followed by world-space anchor and target construction from generated foot positions; and (3) temporal DLS-IK using the resulting targets and weights.

3.2.1 DATA PREPROCESSING

Given an interaction object, we first normalize its geometry into a canonical coordinate system and sample a fixed-size point cloud from the object surface as the geometric condition. We interpolate

sparse object waypoints into a frame-wise reference path. This path is used both as an explicit condition for motion generation and as the baseline of the waypoint-relative object-motion representation introduced in Sec. 3.3.

3.2.2 INFERENCE PIPELINE

At inference time, we encode the text, canonical object point cloud, and interpolated reference path into condition tokens. The initial human–object state and observed sparse waypoints are supplied as masked motion conditions. Starting from Gaussian noise, the part-based denoiser iteratively predicts a joint human–object motion in the waypoint-relative space. Its prediction can be differentiably decoded into an absolute motion state at every reverse step, allowing interaction-aware guidance to encourage foot–ground contact, hand–object proximity, and palm alignment while penalizing penetration. Object motion stays fixed at each guided update to preserve waypoint control.

After reverse diffusion, the predicted object residual is combined with the reference path to recover the absolute object trajectory. The generated human motion is then passed to the full-sequence foot refinement network, which predicts support-phase probabilities and correction confidence. These predictions and the generated world-space foot positions are used to construct persistent contact anchors, target positions, and constraint weights. AnchorIK, our temporal damped least-squares inverse-kinematics (DLS-IK) stage, then updates the lower-body joint rotations to stabilize support. The synchronized human and object sequences form the final output.

3.3 PATH FACTOR: WAYPOINT-RELATIVE MOTION GENERATION

Sparse object waypoints express the user's intended object-transport route, while the text instruction specifies the desired interaction. The Path Factor incorporates this spatial intent into joint human–object generation, aiming to carry out the instructed interaction along the requested route. We interpolate the waypoints into a frame-wise reference path $\mathbf{R}_{1:T} = \mathrm{Interp}(\mathcal{W})$. This reference has two roles: its encoded tokens condition motion generation, and its planar coordinates define the baseline for object translation. Instead of predicting absolute planar positions, the denoiser predicts their residuals relative to this baseline:

$$\mathbf{z}^{o}_{xy,t} = \frac{\mathbf{o}_{xy,t} - \mathbf{r}_{xy,t}}{\mathbf{s}_{xy}}, \qquad \hat{\mathbf{o}}_{xy,t} = \mathbf{r}_{xy,t} + \mathbf{s}_{xy}\hat{\mathbf{z}}^{o}_{xy,t}, \tag{2}$$

Here, $\mathbf{s}_{xy}$ is a data-dependent scale. The reference provides an explicit route condition, while the residual allows local deviations associated with the generated interaction. This reparameterization applies only to planar object translation; human motion, object height, and object rotation retain their original representations and are generated jointly. Decoding adds the predicted residual back to the reference to recover the absolute object trajectory. This conversion is differentiable, so geometric losses, inference-time guidance, downstream refinement, and evaluation operate on the same absolute HOI state. The reference path conveys the user's route instruction, while learned residuals allow local adjustments as human and object motion are generated jointly. Appendix A.2 specifies the interpolant.

3.4 PART FACTOR: PART-BASED SPATIOTEMPORAL DIFFUSION

Human articulation, rigid-object motion, and semantic contact cues form one coupled HOI state despite their distinct structures. We partition this joint noisy motion into seven semantic parts rather than treating an entire frame as one token: the root, torso, left and right arms, left and right legs, and object. The channel grouping follows CHOIS (Appendix A.2).

Part-structured motion patches. Each temporal patch is encoded into one motion token per part, incorporating the noisy motion and observed motion conditions. The resulting representation $\mathbf{Z} \in \mathbb{R}^{K\times 7\times D}$ explicitly retains the time and part axes, where K is the number of temporal patches and D is the feature dimension. Part-specific input projections read the assigned noisy motion and observed-condition channels; part-specific output projections predict only the corresponding motion channels. The seven tokens can still exchange information through part attention within each temporal patch.

Part-aware denoising block. The ST blocks in Fig. 3 separate within-part temporal dynamics from cross-part coordination. Temporal attention captures how each part evolves, while part attention

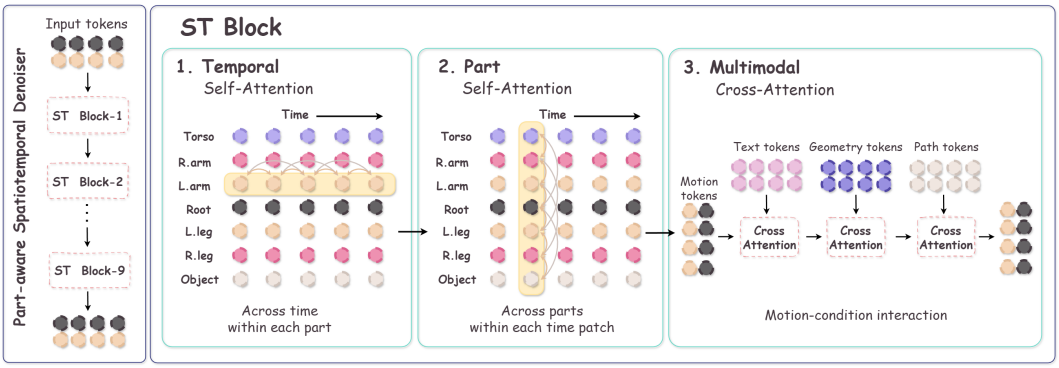

Figure 3: Part-aware Spatiotemporal Denoiser. Left: nine stacked ST blocks update the motion tokens. Right: each block applies temporal self-attention across time within each part, part self-attention across human and object parts within each temporal patch, and sequential cross-attention to text, geometry, and path tokens.

connects the articulated body and moving object. This factorization preserves local motion structure without treating the parts as independently generated sequences.

Multimodal conditioning and reconstruction. Text, object geometry, and the dense reference path are encoded as separate condition tokens. Using motion tokens as queries, cross-attention incorporates semantic, geometric, and route information while updating only the motion features. The conditions remain external to the $K \times 7$ motion-token structure; in particular, the path is not an additional motion part. Finally, the updated tokens are decoded part-wise and assembled into the clean human–object motion prediction $\hat{\mathbf{x}}_0$.

3.5 Contact Factor: Contact-Aware Physical Refinement

The Contact Factor, detailed in the lower row of Fig. 2, targets the two primary physical interfaces in full-body HOI generation: hand–object interaction and foot–ground contact. We use geometry-aware hand guidance and foot–ground height guidance during reverse diffusion. After sampling, a learned full-sequence foot refinement network (FR) predicts support probabilities and correction confidence. These predictions guide anchor construction and temporal IK for stable foot contact.

3.5.1 Hand–Object Interaction Guidance

At late reverse-diffusion steps, the predicted absolute HOI state is optimized with a hand-contact energy

$$\mathcal{E}_{\text{hand}} = \lambda_{\text{cont}}\mathcal{E}_{\text{contact}} + \lambda_{\text{palm}}\mathcal{E}_{\text{palm}} + \lambda_{\text{pen}}\mathcal{E}_{\text{penetration}}. \tag{3}$$

The contact term attracts predicted hand-contact points toward the world-space object surface at its current pose. Beyond point-distance attraction, our palm term encourages the palm plane to be parallel to a locally fitted object surface plane, discouraging edge-on configurations that can cause penetration. We penalize the normal misalignment $1 - (\mathbf{n}_{\text{palm}}^{\top}\mathbf{n}_{\text{surf}})^2$, weighted by local surface planarity and activated for predicted contacts near the surface. During the final 25 reverse steps, object-state variables remain fixed within each guided update; the next step resumes joint denoising.

The same guided steps steer foot height before full-sequence support prediction, anchor construction, and temporal IK after sampling.

3.5.2 Foot–Ground Contact Refinement

The refinement pipeline involves coupled data, learning, and geometric steps. Appendix A.1 details the auxiliary data-generation network, FR training, and inference algorithm; here we focus on the core ideas. Beyond frame-wise foot–ground proximity, we stabilize support over time using learned support phases, persistent contact anchors, and temporal inverse kinematics.

Training pairs and support-phase learning. We train FR with paired clean and low-noise human motions. Low-noise perturbations help preserve the stepping rhythm and temporal correspondence with the clean targets, focusing learning on local contact artifacts rather than changes in action semantics or step planning. A full-sequence temporal Transformer predicts a correction confidence and probabilities over three support states for each foot,

$$y_{t,f} \in \{\text{swing}, \text{toe-support}, \text{flat-support}\}. \tag{4}$$

The network therefore predicts when and how strongly a foot should be anchored; it does not directly regress the final joint positions or contact anchors.

Persistent anchors and temporal inverse kinematics. At inference time, predicted support probabilities are converted into continuous support intervals with hysteresis. World-space anchors are constructed from the generated toe and ankle positions in the horizontal plane. A toe anchor is initialized when support begins and remains fixed throughout the support interval; an ankle anchor is initialized when flat-support begins and retained while that state persists. Swing imposes no anchor constraints, toe-support activates the toe constraint, and flat-support additionally activates the ankle constraint.

We construct target positions by blending the generated foot positions toward their active anchors. The correction strength increases gradually after touchdown and is modulated by the predicted support probabilities and confidence. For the active ankle and toe constraints, a damped least-squares update of the hip, knee, and ankle rotations takes the form

$$\Delta \mathbf{q} = (\mathbf{J}_{\text{kin}}^{\top} \mathbf{W} \mathbf{J}_{\text{kin}} + \lambda \mathbf{I})^{-1} \mathbf{J}_{\text{kin}}^{\top} \mathbf{W} \mathbf{e}, \tag{5}$$

where $\mathbf{e}$ contains ankle and toe target errors, $\mathbf{J}_{\text{kin}}$ is their kinematic Jacobian, $\mathbf{W}$ encodes the support and correction-confidence weights, and λ is the damping coefficient. The temporal solver additionally penalizes a pairwise horizontal displacement residual across adjacent support frames to reduce sliding; swing frames do not contribute to this term. This rotation-based projection preserves bone lengths, root and upper-body motion, object motion, and semantic contact channels.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on FullBodyManipulation (Li et al., 2023) and conduct cross-dataset testing on unseen 3D-FUTURE objects (Fu et al., 2021). The subject split, object selection and category pairing, and sparse-waypoint construction follow the CHOIS protocol (Li et al., 2024). The latter benchmark tests transfer to new object geometries under the same motion-conditioning protocol.

Evaluation metrics. Following CHOIS (Li et al., 2024), we report start, end, and planar waypoint errors (T_s, T_e, T_{xy}); text alignment and motion-distribution quality using Top-3 R_{prec} and FID; contact precision, recall, F1, and ratio using C_{prec}, C_{rec}, C_{F1}, and $C_{\%}$; and hand penetration using P_{hand}. Foot height H_{feet} and foot sliding (FS) assess foot–ground stability. MPJPE, T_{root}, T_{obj}, and R_{obj} measure joint-position, root-translation, object-translation, and object-rotation errors, respectively. Distances are in centimeters; ↓/↑ indicate lower/higher is better.

Implementation details. All local experiments report averages over 20 independent diffusion-sampling runs (seeds 1–20). Architecture, training, inference, and computational details are provided in Appendix A.2.

4.2 Comparative Results

Appendix A.5 reports every run, together with sample standard deviations and means. Results for prior methods are cited from the original papers. We compare reported metrics where the input and evaluation settings are compatible; metrics absent from these tables were either unreported or obtained under a protocol that is not directly comparable. We verified that our evaluation and MaMi-HOI use the same CHOIS evaluation cases: the same 482 FullBodyManipulation test cases (480 for R-precision and FID) and the same 17 unseen 3D-FUTURE mesh IDs paired with the same held-out

test motions; planar object waypoints are sampled every 30 frames. Both evaluations use the same CHOIS-released feature-extractor checkpoint for R-precision and FID. The MaMi-HOI entries in Tables 1 and 2 are taken from Wang et al. (2026).

Table 1: Quantitative comparison on the FullBodyManipulation sparse-waypoint protocol. Best and second-best results are shown in bold and underlined, respectively.

	Condition Matching			Human Motion				Interaction					GT Difference			
Method	$T_s \downarrow$	$T_e \downarrow$	$T_{xy} \downarrow$	$H_{\text{feet}} \downarrow$	FS $\downarrow$	$R_{\text{prec}} \uparrow$	FID $\downarrow$	$C_{\text{prec}} \uparrow$	$C_{\text{rec}} \uparrow$	$C_{\text{F1}} \uparrow$	$C_{\%} \uparrow$	$P_{\text{hand}} \downarrow$	MPJPE $\downarrow$	$T_{\text{root}} \downarrow$	$T_{\text{obj}} \downarrow$	$R_{\text{obj}} \downarrow$
CHOIS (ECCV 2024)	1.71	6.31	2.87	4.20	0.35	0.64	0.69	0.80	0.64	0.67	0.54	0.59	15.30	24.43	12.53	0.99
HOI-Dyn (NeurIPS 2025)	1.75	5.58	3.26	3.07	0.37	–	0.48	–	–	0.71	0.60	0.64	15.60	23.90	12.47	0.90
ARFM (arXiv 2025.12)	1.55	6.58	2.64	4.08	0.35	–	–	0.80	0.70	0.72	0.60	0.65	15.41	22.89	11.90	0.94
ViHOI (CVPR 2026)	–	–	–	–	0.26	–	–	**0.83**	0.75	0.75	0.64	0.58	14.97	–	–	–
MaMi-HOI (ICML 2026)	1.53	4.51	2.14	4.10	0.37	0.71	0.39	0.80	0.72	0.73	0.62	0.57	13.50	19.65	9.98	0.78
TriFactorHOI	**0.35**	**1.45**	**0.13**	**2.99**	**0.23**	**0.74**	**0.34**	0.81	**0.83**	**0.80**	**0.72**	**0.54**	**13.28**	**18.93**	**8.45**	**0.73**

Table 2: Cross-dataset testing on unseen 3D-FUTURE objects.

	Condition Matching			Human Motion				Interaction	
Method	$T_s \downarrow$	$T_e \downarrow$	$T_{xy} \downarrow$	$H_{\text{feet}} \downarrow$	FS $\downarrow$	$R_{\text{prec}} \uparrow$	FID $\downarrow$	$C_{\%} \uparrow$	$P_{\text{hand}} \downarrow$
CHOIS (ECCV 2024)	4.12	7.35	2.92	3.75	0.38	0.62	1.60	0.48	0.15
HOI-Dyn (NeurIPS 2025)	4.60	6.17	2.95	2.56	0.37	–	1.62	0.54	0.26
MaMi-HOI (ICML 2026)	3.43	5.87	2.21	3.35	0.37	**0.69**	1.57	0.54	0.23
TriFactorHOI	**3.18**	**5.53**	**0.99**	**2.50**	**0.25**	**0.69**	**1.46**	**0.73**	**0.14**

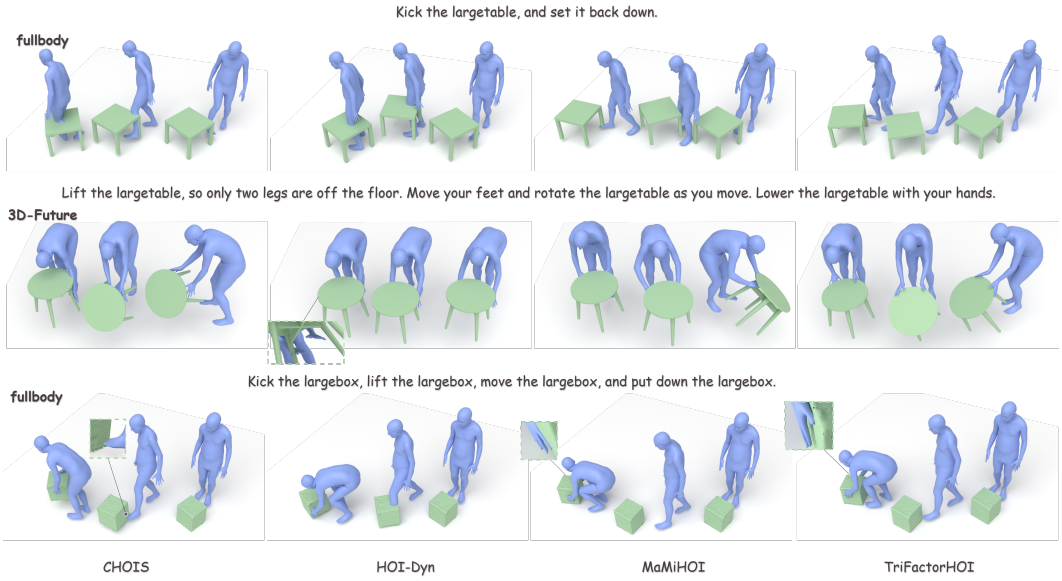

Figure 4: Qualitative comparison for three text-conditioned interaction cases. Each row shows one instruction, and each column shows representative motion frames from one method. Humans and manipulated objects are shown in blue and green, respectively.

Table 1 compares TriFactorHOI with methods evaluated under the FullBodyManipulation sparse-waypoint protocol. TriFactorHOI reduces start, end, and planar waypoint errors relative to the reported baselines. Because the interpolated reference path passes through the supplied planar waypoints, T_{xy} measures adherence at these frames and does not, by itself, establish motion or interaction quality. In the matched path-design ablation (Table 5), with the sparse waypoints, reference-path condition, interaction guidance, and foot refinement held fixed, waypoint-relative residual prediction reduces T_{xy} from 1.01 to 0.13 cm and FID from 0.37 to 0.34 relative to absolute planar prediction, while contact F1 and foot sliding remain at 0.80 and 0.23 at the reported precision. The full method's Top-3 R_{prec} of 0.74 also indicates strong text alignment. Lower foot sliding and hand

penetration, together with improved contact F1, further show that accurate object transport is compatible with stable foot support and plausible hand–object interaction. R_{prec} and FID are omitted for ARFM (Xie et al., 2025) and ViHOI (Cai et al., 2026) because their reported evaluators are not directly compatible with this protocol.

Table 2 shows that TriFactorHOI leads or ties the compared methods on all nine reported 3D-FUTURE metrics without retraining. ARFM and ViHOI report no quantitative results for this 3D-FUTURE benchmark.

Qualitative analysis. Figure 4 compares table and box manipulation across CHOIS, HOI-Dyn, MaMi-HOI, and TriFactorHOI. In the table-tilting example, TriFactorHOI couples object reorientation with coordinated bending and reaching, while some baselines show little change in object orientation. Additional qualitative comparisons are provided in Appendix A.4 (Fig. 5).

4.3 Ablation Study

Unless otherwise noted, ablations average 20 independent diffusion-sampling runs (seeds 1–20) and report arithmetic means. Palm-orientation guidance, foot refinement, and the remaining component studies are reported in Appendix A.3.

4.4 User Study

We complement the automatic metrics with preference rankings on 23 text-conditioned cases. One case is sampled at **random** from each of the 13 FullBodyManipulation object categories, and 10 cases are sampled at **random** from the unseen 3D-FUTURE objects. Each case shows four videos, generated by CHOIS, HOI-Dyn, MaMi-HOI, and TriFactorHOI. Method names are hidden, and the assignment of methods to the four options is shuffled across cases. Participants rank all four videos. The first-ranked video receives 2 points, the second receives 1 point, and the third and fourth receive 0. Each method is included in 410 rankings.

Table 3: User-study scores from 82 participants, five questions each on average. Each ranking awards 2, 1, 0, and 0 points. Each method is included in 410 rankings, 246 on FullBodyManipulation and 164 on 3D-FUTURE. The dataset columns sum to the total.

Method	Points ↑	FullBody ↑	3D-FUTURE ↑	First ↑	Second ↑	3rd/4th ↓
TriFactorHOI	**559**	**295**	**264**	230	99	81
MaMi-HOI	315	201	114	73	169	168
CHOIS	253	165	88	79	95	236
HOI-Dyn	103	77	26	28	47	335

Table 3 sums these points on both datasets. TriFactorHOI scores 295 on FullBodyManipulation and 264 on 3D-FUTURE, for 559 overall, ahead of MaMi-HOI (315), CHOIS (253), and HOI-Dyn (103). It is ranked first 230 times and third or fourth only 81 times. CHOIS receives 79 first places, six more than MaMi-HOI, but MaMi-HOI has 169 second places against CHOIS's 95 and the higher total.

5 Conclusion

TriFactorHOI combines waypoint-relative control, part-aware motion modeling, and contact refinement for full-body human–object interaction. On FullBodyManipulation, its 20-seed means reduce planar waypoint error, FID, and foot sliding by approximately 93.9%, 11.9%, and 10.3%, respectively, against the strongest prior result for each metric in Table 1. With the same sparse-waypoint inputs, its path design also lowers waypoint error and FID versus absolute translation. Testing on unseen 3D-FUTURE objects without retraining supports transfer to new object geometries.

Limitations and future work. Our kinematic contact refinement assumes planar ground and does not guarantee dynamic feasibility. Guided inference adds cost, and few-step acceleration has only been evaluated without guidance. Future work will address dynamics, non-planar support, and guidance-aware acceleration.

AI USE STATEMENT

Generative AI assistance was used for text translation and polishing, data aggregation, figure formatting, and coding support for research utilities and manuscript preparation.

ETHICS STATEMENT

This study uses existing research datasets and does not collect new motion or interaction recordings. The user study asks participants to rank videos of method outputs on test cases from these datasets. Use and redistribution of third-party datasets, assets, and body models remain subject to their original licenses and access conditions; any release of processed data will respect these restrictions. Dataset coverage may limit generalization across body types, objects, and interaction settings. Generated motions should not be presented as recordings of real individuals or used for deceptive impersonation. Moreover, geometric contact plausibility does not establish safe physical execution: deployment on robots requires separate dynamics and safety validation. Our intended use is research on motion synthesis and virtual human–object interaction.

REPRODUCIBILITY STATEMENT

Section 3 describes the motion representation, preprocessing, network, and refinement pipeline. Appendix A.1 details auxiliary pair generation, foot-refiner training, and geometric refinement, while Appendix A.2 specifies the architecture, training, and inference settings. Section 4 and Appendix A.3 document the datasets, evaluation protocol, and ablation configurations. Section 4.4 describes the preference-ranking protocol and score aggregation. The two main benchmark comparisons and the start-and-goal experiment use arithmetic means over sampling seeds 1–20. Other ablation protocols are described with their experiments. Appendix A.5 lists every FullBodyManipulation and 3D-FUTURE inference run, together with sample standard deviations and means. The full code, including preprocessing, inference, and evaluation scripts, will be made publicly available upon acceptance. Original datasets and licensed assets must be obtained under their respective access terms.

REFERENCES

German Barquero, Sergio Escalera, and Cristina Palmero. Seamless Human Motion Composition with Blended Positional Encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 457–469, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Barquero_Seamless_Human_Motion_Composition_with_Blended_Positional_Encodings_CVPR_2024_paper.html.

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*, pp. 30686–30695, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Cai_ViHOI_Human-Object_Interaction_Synthesis_with_Visual_Priors_CVPR_2026_paper.html.

Bin Cao, Sipeng Zheng, Hao Luo, Boyuan Li, Jing Liu, and Zongqing Lu. OpenT2M: No-frill Motion Generation with Open-source, Large-scale, High-quality Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30640–30649, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Cao_OpenT2M_No-frill_Motion_Generation_with_Open-source_Large-scale_High-quality_Data_CVPR_2026_paper.html.

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*, pp. 1577–1585, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Cha_Text2HOI_Text-guided_3D_Motion_Generation_for_Hand-Object_Interaction_CVPR_2024_paper.html.

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

Computer Vision and Pattern Recognition, pp. 18000–18010, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Chen_Executing_Your_Commands_via_Motion_Diffusion_in_Latent_Space_CVPR_2023_paper.html.

Peishan Cong, Ziyi Wang, Yuexin Ma, and Xiangyu Yue. SemGeoMo: Dynamic contextual human motion generation with semantic and geometric guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17561–17570, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cong_SemGeoMo_Dynamic_Contextual_Human_Motion_Generation_with_Semantic_and_Geometric_CVPR_2025_paper.html.

Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. MotionLCM: Real-time Controllable Motion Generation via Latent Consistency Model. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/2504_ECCV_2024_paper.php.

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*, 2024. URL https://www.christian-diller.de/projects/cg-hoi/.

Siyuan Fan, Bo Du, Xiantao Cai, Bo Peng, and Longling Sun. TextIM: Part-aware Interactive Motion Synthesis from Text. *arXiv preprint arXiv:2408.03302*, 2024. URL https://arxiv.org/abs/2408.03302.

Pengcheng Fang, Tengjiao Sun, Xiaoyu Zhan, Xiaohao Cai, and Dongjie Fu. MoGeFlow: Flowing through motion codebook geometry for text-to-motion generation. *arXiv preprint arXiv:2606.11656*, 2026. URL https://arxiv.org/abs/2606.11656.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3D-FUTURE: 3D furniture shape with TextURE. *International Journal of Computer Vision*, 129:3313–3337, 2021. doi: 10.1007/s11263-021-01534-z. URL https://link.springer.com/article/10.1007/s11263-021-01534-z.

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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Guo_Generating_Diverse_and_Natural_3D_Human_Motions_From_Text_CVPR_2022_paper.html.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024. URL https://ericguo5513.github.io/momask/.

Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. SALAD: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://seokhyeonhong.github.io/projects/salad/.

Yiheng Huang, Hui Yang, Chuanchen Luo, Yuxi Wang, Shibiao Xu, Zhaoxiang Zhang, Man Zhang, and Junran Peng. StableMoFusion: Towards Robust and Efficient Diffusion-based Motion Generation Framework. *arXiv preprint arXiv:2405.05691*, 2024. URL https://arxiv.org/abs/2405.05691.

Yihong Ji, Jinsong Zhang, He Hu, and Hongbo Xu. HOIMask: Towards Generative Masked Modeling for Human Object Interaction Generation. *arXiv preprint arXiv:2608.15141*, 2026. URL https://arxiv.org/abs/2608.15141.

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*, pp. 2253–2263, 2026. URL https://github.com/kuai-lab/cvpr26_Decoupled_HOI.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2151–2162, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Karunratanakul_Guided_Motion_Diffusion_for_Controllable_Human_Motion_Synthesis_ICCV_2023_paper.html.

Lucas Kovar, John Schreiner, and Michael Gleicher. Footskate cleanup for motion capture editing. In *Proceedings of the ACM SIGGRAPH Symposium on Computer Animation*, pp. 97–104, 2002. doi: 10.1145/545261.545277. URL https://graphics.cs.wisc.edu/Papers/2002/KSG02/.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics*, 42(6), 2023. URL https://lijiaman.github.io/projects/omomo/.

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*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/5720_ECCV_2024_paper.php.

Lei Li and Angela Dai. HOI-PAGE: Zero-Shot Human-Object Interaction Generation with Part Affordance Guidance. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62895.

Chenguo Lin and Yadong Mu. InstructScene: Instruction-Driven 3D Indoor Scene Synthesis with Semantic Graph Prior. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6d96950e6ce0b92e3d0fb2afa634e0f2-Abstract-Conference.html.

Xueyi Liu and Li Yi. GeneOH Diffusion: Towards Generalizable Hand-Object Interaction Denoising via Denoising Diffusion. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0ff30c4bf31db0119a6219e0d250e037-Abstract-Conference.html.

Lucas Mourot, Ludovic Hoyet, François Le Clerc, and Pierre Hellier. UnderPressure: Deep learning for foot contact detection, ground reaction force estimation and footskate cleanup. *Computer Graphics Forum*, 41(8):195–206, 2022. doi: 10.1111/cgf.14635. URL https://diglib.eg.org/items/def192e5-ad91-4409-b078-7d564fbaefb5.

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 Workshops*, 2025. URL https://neu-vi.github.io/HOI-Diff/.

Xiaogang Peng, Zeyu Han, Zichong Meng, Yiming Xie, Jihua Zhu, Gang Hua, and Huaizu Jiang. Surface Keypoint Representation for Multi-Object and Articulated Human-Object Interaction Generation. *arXiv preprint arXiv:2608.03158*, 2026. URL https://arxiv.org/abs/2608.03158.

Ilya A. Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. TriDi: Trilateral Diffusion of 3D Humans, Objects, and Interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5523–5535, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Petrov_TriDi_Trilateral_Diffusion_of_3D_Humans_Objects_and_Interactions_ICCV_2025_paper.html.

Roey Ron, Guy Tevet, Haim Sawdayee, and Amit H. Bermano. HOIDiNi: Human-Object Interaction through Diffusion Noise Optimization. *arXiv preprint arXiv:2506.15625*, 2025. URL https://arxiv.org/abs/2506.15625.

Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pp. 13263–13273, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Taheri_GOAL_Generating_4D_Whole-Body_Motion_for_Hand-Object_Grasping_CVPR_2022_paper.html.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Hao Wang, Shiqi Wang, and Qi Liu. MaMi-HOI: Harmonizing global kinematics and local geometry for human-object interaction generation. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.05756.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. PhysHOI: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023. URL https://arxiv.org/abs/2312.04393.

Lin Wu, Zhixiang Chen, and Jianglin Lan. HOI-Dyn: Learning interaction dynamics for human-object motion diffusion. In *Advances in Neural Information Processing Systems*, 2025a. URL https://arxiv.org/abs/2507.01737.

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*, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wu_Human-Object_Interaction_from_Human-Level_Instructions_ICCV_2025_paper.html.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D Latents for Scalable and Versatile 3D Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21469–21480, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xiang_Structured_3D_Latents_for_Scalable_and_Versatile_3D_Generation_CVPR_2025_paper.html.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10086–10096, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xiao_MotionStreamer_Streaming_Motion_Generation_via_Diffusion-based_Autoregressive_Model_in_Causal_ICCV_2025_paper.html.

Johnathan Xie, Stefan Stojanov, Cristobal Eyzaguirre, Daniel L. K. Yamins, and Jiajun Wu. Autoregressive Flow Matching for Motion Prediction. *arXiv preprint arXiv:2512.22688*, 2025. URL https://arxiv.org/abs/2512.22688.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control any joint at any time for human motion generation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7827290f07f63485b849b66cfa3e5dd0-Abstract-Conference.html.

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*, pp. 14928–14940, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Xu_InterDiff_Generating_3D_Human-Object_Interactions_with_Physics-Informed_Diffusion_ICCV_2023_paper.html.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. InterDreamer: Zero-Shot Text to 3D Dynamic Human-Object Interaction. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5eca2e4fe7858cbbfef4e08573cfcb25-Abstract-Conference.html.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12266–12277, 2025. URL https://github.com/Sirui-Xu/InterMimic.

Sirui Xu, Samuel Schulter, Morteza Ziyadi, Xialin He, Xiaohan Fei, Yu-Xiong Wang, and Liang-Yan Gui. InterPrior: Scaling Generative Control for Physics-Based Human-Object Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23516–23527, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Xu_InterPrior_Scaling_Generative_Control_for_Physics-Based_Human-Object_Interactions_CVPR_2026_paper.html.

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. PhyScene: Physically Interactable 3D Scene Synthesis for Embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16262–16272, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_PhyScene_Physically_Interactable_3D_Scene_Synthesis_for_Embodied_AI_CVPR_2024_paper.html.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language Guided Generation of 3D Embodied AI Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16227–16237, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Holodeck_Language_Guided_Generation_of_3D_Embodied_AI_Environments_CVPR_2024_paper.html.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://nvlabs.github.io/PhysDiff/.

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. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zeng_ChainHOI_Joint-based_Kinematic_Chain_Modeling_for_Human-Object_Interaction_Generation_CVPR_2025_paper.html.

Jiajun Zhang, Yuxiang Zhang, Liang An, Mengcheng Li, Hongwen Zhang, Zonghai Hu, and Yebin Liu. ManiDext: Hand-Object Manipulation Synthesis via Continuous Correspondence Embeddings and Residual-Guided Diffusion. *arXiv preprint arXiv:2409.09300*, 2024a. URL https://arxiv.org/abs/2409.09300.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14730–14740, 2023a. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Generating_Human_Motion_From_Textual_Descriptions_With_Discrete_Representations_CVPR_2023_paper.html.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. URL https://arxiv.org/abs/2208.15001.

Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. FineMoGen: Fine-grained spatio-temporal motion generation and editing. In *Advances in Neural Information Processing Systems*, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2d52879ef2ba487445ca2e143b104c3b-Abstract-Conference.html.

Wanyue Zhang, Rishabh Dabral, Vladislav Golyanik, Vasileios Choutas, Eduardo Alvarado, Thabo Beeler, Marc Habermann, and Christian Theobalt. BimArt: A Unified Approach for the Synthesis

of 3D Bimanual Interaction with Articulated Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27694–27705, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_BimArt_A_Unified_Approach_for_the_Synthesis_of_3D_Bimanual_CVPR_2025_paper.html.

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez-Pellitero, and Gerard Pons-Moll. FORCE: Physics-aware Human-object Interaction. *arXiv preprint arXiv:2403.11237*, 2024b. URL https://arxiv.org/abs/2403.11237.

Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion Mamba: Efficient and Long Sequence Motion Generation. In *European Conference on Computer Vision*, 2024c. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/100_ECCV_2024_paper.php.

Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A Diffusion-Based Autoregressive Motion Model for Real-Time Text-Driven Motion Control. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3b057de5a2e38bd8fa10201866c20dbf-Abstract-Conference.html.

Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. GEARS: Local Geometry-aware Hand-object Interaction Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20634–20643, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhou_GEARS_Local_Geometry-aware_Hand-object_Interaction_Synthesis_CVPR_2024_paper.html.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. MotionGPT3: Human Motion as a Second Modality. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1551c01d7a3d0bf21e2518331e9f7074-Abstract-Conference.html.

A Appendix

A.1 Foot–Ground Contact Refinement Details

This appendix supplements Section 3.5.2 with the complete foot-refinement pipeline: constructing aligned training pairs, learning support states, and converting those predictions into geometric corrections. The objective is to repair local contact artifacts while retaining the generated action and stepping rhythm. We train the auxiliary denoiser and FR on HumanML3D (Guo et al., 2022), a human-motion dataset without object interactions, using 69,616 training and 8,314 internal validation clips of 120 frames each.

A.1.1 Constructing Aligned Training Pairs

An auxiliary human-motion denoiser. We first train a separate, human-only diffusion network to reconstruct clean motion from low-noise inputs. It follows the part-based spatiotemporal architecture, without text, object, or waypoint conditioning. Reconstruction losses supervise joint positions and rotations, while velocity and acceleration losses preserve temporal evolution. We then fine-tune it with additional ankle/toe position, contact-frame velocity, and toe-height losses. This contact-aware training keeps its predictions close enough to the source motion for frame-aligned supervision. The trained network is frozen and used only to prepare FR data offline; it is not part of HOI inference.

Clean/low-noise pairs. For each clean clip, we add diffusion noise at a low noise level and use the frozen network to predict the clean-motion state in one pass. We restore the original root translation and rotation after prediction, retaining global transport while allowing local reconstruction errors. The resulting joint sequence and its unperturbed counterpart form a pair $(\mathbf{J}^-, \mathbf{J}^+)$. Low-noise reconstruction helps preserve step timing and motion semantics, unlike independently sampled motions whose footfalls may not correspond to the reference. FR therefore learns to recognize support despite local artifacts, rather than learning to replace or replan the action.

A.1.2 Learning Full-Sequence Support

Network inputs and outputs. FR takes the entire perturbed clip and extracts root displacement, root-relative lower-body positions, ankle/toe frame differences, foot heights, and normalized time. These features jointly describe leg configuration, global movement, and local foot behavior. A bidirectional temporal Transformer processes all 120 frames and predicts, for each foot, probabilities of swing, toe-support, and flat-support, together with a correction confidence. Full-sequence context helps resolve ambiguous transitions without introducing independent short-window boundaries. FR predicts neither final joint coordinates nor anchor positions.

Supervision and objective. Support labels are derived from the clean counterpart using toe height and horizontal displacement. Non-contact frames are labeled swing; during contact, a relatively stationary ankle indicates flat-support, while ankle movement indicates toe-support. Thus, toe-support permits ankle motion around the forefoot, whereas flat-support constrains both ankle and toe. The confidence target is the clipped horizontal discrepancy between the perturbed and clean ankle/toe positions during support. It represents the need for correction rather than contact probability alone. We also include clean-input examples with zero correction-confidence targets, replacing the perturbed input with probability 0.1, to discourage unnecessary changes. Flat-support requires the clean ankle's horizontal displacement to remain below 0.005 m per frame. Mode classification uses weights 2, 4, and 1 for swing, toe-support, and flat-support. Touchdown and liftoff positive labels are weighted by 20.

The training objective combines these complementary signals:

$$\begin{aligned}\mathcal{L}_{\mathrm{FR}} = \mathcal{L}_{\mathrm{mode}} + 0.5\mathcal{L}_{\mathrm{support}} + 0.25\mathcal{L}_{\mathrm{transition}} \\ + 0.5\mathcal{L}_{\mathrm{confidence}} + 0.05\mathcal{L}_{\mathrm{temporal}}.\end{aligned} \tag{6}$$

Weighted cross-entropy supervises the three support modes; binary cross-entropy supervises support, touchdown/liftoff transitions, and confidence. Temporal regularization penalizes changes between adjacent mode-probability vectors to reduce state flicker. The transition head provides auxil-

iary supervision; inference uses support-probability hysteresis instead of directly triggering anchors from transition logits.

The reported model is warm-started from a HumanML3D foot-refinement encoder and a windowed support predictor, then fine-tuned on complete sequences. The auxiliary pair generator and HOI denoiser remain fixed during FR training. Network fitting uses only HumanML3D-derived pairs, and the parameters remain frozen across all evaluations. FR consumes human root and lower-body motion features plus normalized time, but no object-specific input; the same frozen network is applied to generated HOI sequences without fine-tuning.

A.1.3 From Support Predictions to Refined Motion

World anchors and targets. After HOI sampling, we reconstruct the kinematic skeleton and run FR once over the generated sequence. Support probabilities are converted into continuous intervals with activation and release thresholds 0.60 and 0.35, avoiding rapid switching near uncertain contacts. Constraint weights ramp in over three frames, and the target displacement is limited to 0.04 m before weighting. When support starts, we initialize a world-XY toe anchor from the previous frame's constructed target, or its generated position if inactive, and retain it throughout the interval. Flat-support additionally initializes and retains an ankle anchor; swing activates neither constraint. Anchors come from the generated motion, not a clean reference or a learned coordinate-regression head.

We blend generated foot positions toward their anchors to obtain IK targets. The blend is controlled by support probability and correction confidence, with a gradual onset and bounded displacement to prevent abrupt snapping. The same weights determine constraint strength in IK. This separates learning *when* and *how strongly* contact should be enforced from geometrically determining *where* the foot should remain.

Temporal IK. Temporal DLS-IK updates hip, knee, and ankle rotations to approach these targets. In addition to per-frame target errors, it uses a pairwise horizontal-displacement residual across adjacent support frames to suppress sliding. Swing frames are excluded from this term. The default temporal weight is 0.5. Neighboring positions are refreshed across iterations, coupling refinement over time without constraining legitimate swing motion. Rotation-based updates preserve kinematic bone lengths; root translation, upper-body motion, object motion, and semantic hand-contact channels remain unchanged at this stage. Architecture and inference settings are summarized in Appendix A.2.

Foot-refinement ablation controls. Table 11 changes only the post-sampling foot-refinement stage; the generator and diffusion-time interaction guidance are fixed. The *w/o FR/AnchorIK* variant bypasses learned support prediction and the anchor-based IK correction. The *threshold-based support* variant replaces FR's learned support probabilities with rule-derived one-hot swing, toe-support, and flat-support states. It retains FR's learned correction confidence and the same subsequent anchor construction, target blending, and temporal IK. The *rule support, fixed strength* variant uses those same rule-derived states, anchors, and temporal IK, but replaces the learned correction confidence with a fixed correction strength, so no learned FR prediction remains. The *w/o temporal IK residual* variant retains the learned support and confidence predictions, anchors, targets, and per-frame IK constraints, but sets the adjacent-support-frame horizontal-displacement residual weight from 0.5 to zero. The other IK settings remain fixed.

Threshold-based support rule. This ablation applies a deterministic rule to forward-kinematics (FK) joints of the generated motion, without ground-truth joints; its thresholds reuse the defaults used to construct FR training labels. For each foot, toe coordinates are smoothed over three frames with weights $[0.25, 0.5, 0.25]$; ankle coordinates are not smoothed. We measure the smoothed toe's horizontal forward displacement, copying the penultimate value to the final frame, and the raw ankle's horizontal backward displacement, copying the second value to the first frame. Both are measured in m/frame rather than m/s.

The ground height is estimated from the generated toe trajectories. We sort smoothed-toe height samples whose 3D displacement is below 0.005 m/frame and split them whenever adjacent heights differ by more than 0.005 m. The median height of the lowest group containing at least five samples defines the ground; if no such group exists, we use the fifth percentile of all smoothed-toe heights. The signed toe height h_t is measured relative to this estimate.

Each foot starts in non-support. It enters support only when $h_t \leq 0.030$ m *and* toe horizontal displacement is at most 0.005 m/frame. Once supported, it remains so while $h_t \leq 0.045$ m *and* toe displacement is at most 0.010 m/frame; otherwise it becomes non-support. We then fill non-support gaps of at most two frames between support intervals and remove support runs shorter than three frames, in that order. A cleaned non-support frame is *swing*. A supported frame is *toe-support* when ankle horizontal displacement is at least 0.005 m/frame, including equality, and *flat-support* otherwise. The two feet are classified independently, so simultaneous support is allowed. Flat-support here is a toe-contact and low-ankle-motion heuristic, without a separate heel, sole-plane, or foot-tilt test. The three states replace FR's support output as one-hot probabilities. In the threshold-based row, the learned correction confidence and all downstream refinement stages remain unchanged. The rule-support, fixed-strength row reuses these states with a fixed correction strength.

A.2 Additional Implementation Details

Hardware and software. Our local workstation has an AMD Ryzen 7 7800X3D eight-core CPU and a single NVIDIA RTX PRO 6000 Blackwell Workstation Edition GPU, running Ubuntu 24.04.4 LTS. The implementation uses Python 3.10.20 and PyTorch 2.11.0 with CUDA 12.8.

Dataset protocol. Following CHOIS (Li et al., 2024), FullBodyManipulation uses 15 training subjects and two test subjects. Each test run has 482 cases for direct metrics and 480 for R_{prec}/FID. The 3D-FUTURE evaluation uses 17 unseen meshes paired with FullBodyManipulation test motions from the same object category. Planar object waypoints are extracted every 30 frames.

Training and inference. We train the main generator from scratch for 100,000 optimizer iterations using Adam, a constant learning rate 10^{-4}, batch size 128, and gradient clipping at 1.0. The denoiser predicts the clean motion under a cosine schedule of 1,000 steps. Reconstruction averages seven part-wise L_1 terms with equal weights. Semantic-contact, forward-kinematics, and object-keypoint losses use weights 1.0, 0.5, and 1.0. An exponential moving average of the weights, with decay 0.995 and an update every 10 steps, supplies the inference weights. Inference uses batch size 32. The recorded run takes approximately 5.6 hours with 19.6 GiB peak allocated GPU memory. The main FullBodyManipulation and 3D-FUTURE comparisons and the start-and-goal experiment report arithmetic means over sampling seeds 1–20. Sampling protocols for other ablations are given with their respective tables. Each seed is evaluated separately, and the trained networks remain fixed across runs. Neither network is retrained for 3D-FUTURE.

Main-network configuration. The generator produces 120-frame HOI sequences with 220 features per frame, organized into six human parts and one object part. Non-overlapping four-frame patches downsample each part stream to 30 temporal tokens. The denoiser stacks nine spatiotemporal blocks with hidden dimension 512, eight attention heads, feed-forward dimension 2,048, and dropout 0.05. Each object is represented by 512 surface points, pooled into 32 geometry tokens. Contextual text tokens come from a frozen CLIP ViT-B/32 encoder, and frame-aligned path tokens provide the route condition. The normalized waypoint-relative planar residual is clipped to $[-8, 8]$. Standard inference performs 1,000 reverse steps.

Reference path and state layout. The reference path and the 220-dimensional part layout follow CHOIS (Li et al., 2024). Horizontal waypoints are joined by a piecewise cubic Hermite interpolant that passes through the supplied knots; vertical translation is a smoothstep between the start and end heights. The path is built in meters and then normalized. Relative to CHOIS, only the horizontal object coordinates are stored as residuals to this path. Each component of $\mathbf{s}_{xy}$ is the 99.9th percentile of the corresponding absolute training-set residual and remains fixed during training and inference. Object height and the remaining channels stay absolute. The state remains object translation and rotation, joint positions, global 6D rotations, and four object-contact labels, grouped into the root, torso, arms, legs, and object.

FR configuration. The full-sequence FR takes 47-dimensional features for all 120 frames and uses a six-layer temporal Transformer with hidden dimension 256, eight attention heads, feed-forward dimension 1,024, and dropout 0.05. Following the initialization described in Appendix A.1, full-sequence fine-tuning uses AdamW, batch size 256, learning rate 5×10^{-5}, weight decay 10^{-4}, gradient clipping at 1.0. and 500-update fine-tuning run. At inference, AnchorIK uses refinement strength 0.25, temporal DLS weight 0.5, vertical weight scale 0.5, DLS damping 0.05, and eight iterations, with each joint rotation limited to 2° per iteration.

Interaction guidance. Guidance is applied on the final 25 reverse steps with gradient scale 1,000. Distance, palm-normal alignment, and penetration are first combined with coefficients 1.0, 0.04, and 0.05; the resulting hand–object energy is then multiplied by 10,000. Foot–ground guidance uses weight 275,000 and a target foot height of 2 cm. A predicted semantic contact is active above 0.95. The hand–object distance tolerance is 3 cm, and palm-orientation guidance is gated to hands within 8 cm of the surface. The local surface fit uses 32 neighbors and distance temperature 0.02. Contact distances are normalized by the number of active contacts, and penetration uses the same legacy surface definition as the reported P_{hand} metric. Object motion stays fixed at each guided update. Pre-contact palm-reach orientation guidance is disabled.

A.3 Additional Ablations

Sampling protocols for the additional ablations are stated with each experiment. The start-and-goal comparison reports the arithmetic mean over inference seeds 1–20.

A.3.1 Part-Structured Motion Modeling

Table 4: Part-token ablation on FullBodyManipulation (30/210/210 motion tokens per 120-frame sequence).

Variant	Params. (M)	FS ↓	H_{feet} ↓	T_{xy} ↓	R_{prec} ↑	FID ↓	C_{F1} ↑
Global 1 token	118.70	0.40	4.28	0.36	**0.74**	0.57	**0.80**
Free 7 tokens	119.55	<u>0.37</u>	<u>3.97</u>	<u>0.21</u>	<u>0.72</u>	<u>0.45</u>	**0.80**
TriFactorHOI	119.15	**0.23**	**2.99**	**0.13**	**0.74**	**0.34**	**0.80**

To test semantic assignment at fixed token count, we compare three variants. The global-token model encodes all 220 human–object motion channels into one token per four-frame patch and has no cross-part attention. The free seven-token model uses seven independent input projections, each reading the complete human–object patch with unchanged conditions and masks. Its token states pass through the same type of spatiotemporal blocks, then are concatenated and jointly decoded into all motion channels. Unlike TriFactorHOI, no input or output token is assigned to a body part or the object. All variants retain the seven-group loss, so the free model still receives part-wise training supervision.

The semantic and free variants each use 210 motion tokens over 120 frames. Their denoisers have 119.15M and 119.55M parameters, respectively, with hidden widths 512 and 488; both use nine blocks, eight attention heads, and 2,048-wide feed-forward layers. The global-token model uses 30 tokens, width 560, feed-forward width 2,240, and 118.70M parameters. Parameter counts exclude frozen CLIP and the separate foot-refinement network. The free and semantic models therefore match token count and approximately match parameter budget. All variants retain the same waypoint-relative representation, conditions, reconstruction and auxiliary losses, and full interaction-guidance and foot-refinement pipeline. Each is trained from scratch for 100K iterations with batch size 128 and learning rate 10^{-4}. Evaluation uses the fixed 100K-iteration checkpoints, sampling seeds 1–20, 1,000 denoising steps, batch size 32, and 482 direct-metric and 480 T2M-evaluator cases per run.

Table 4 shows that semantic assignment lowers foot sliding from 0.37 to 0.23 (37.2%) relative to free seven-token modeling. The foot-height metric falls from 3.97 to 2.99, waypoint error from 0.21 to 0.13, and text R-precision rises from 0.72 to 0.74. FID falls from 0.45 to 0.34. Contact F1 is 0.80 for both at the displayed precision. The matched-token comparison supports semantic input/output assignment for foot stability in the full pipeline. The free model already approaches TriFactorHOI

on FID, so the larger global-to-semantic FID gap cannot be ascribed solely to grouping: token count and other dimensions also change.

A.3.2 WAYPOINT-RELATIVE PATH DESIGN

Table 5: Path-design ablation on FullBodyManipulation. Absolute predicts absolute planar translation from directly masked waypoints; Absolute + path cond. additionally receives the interpolated reference path. TriFactorHOI uses the reference condition and residual prediction. Zero waypoint residual retrains that model with denoiser-output residuals at the supplied waypoints set to zero. Absolute, Absolute + path cond., and TriFactorHOI average 20 sampling runs with the same sparse waypoints, interaction guidance, and foot refinement.

Variant	$T_s \downarrow$	$T_e \downarrow$	$T_{xy} \downarrow$	FID $\downarrow$	$C_{\mathrm{F1}} \uparrow$	$P_{\mathrm{hand}} \downarrow$	FS $\downarrow$	$H_{\mathrm{feet}} \downarrow$	MPJPE $\downarrow$	$T_{\mathrm{root}} \downarrow$	$T_{\mathrm{obj}} \downarrow$	$R_{\mathrm{obj}} \downarrow$
Absolute	<u>0.57</u>	1.98	0.98	0.37	**0.80**	<u>0.55</u>	**0.22**	3.08	13.59	19.16	8.54	**0.73**
Absolute + path cond.	0.59	2.03	1.01	0.37	**0.80**	0.56	<u>0.23</u>	3.04	13.63	<u>19.04</u>	8.59	<u>0.74</u>
Zero waypoint residual	**0.35**	**1.39**	**0.00**	<u>0.36</u>	**0.80**	<u>0.55</u>	0.24	**2.93**	<u>13.42</u>	19.16	**8.34**	**0.73**
TriFactorHOI	**0.35**	<u>1.45</u>	<u>0.13</u>	**0.34**	**0.80**	**0.54**	<u>0.23</u>	<u>2.99</u>	**13.28**	**18.93**	<u>8.45</u>	**0.73**

Table 5 separates reference-path conditioning from residual prediction. Absolute predicts absolute planar object translation from directly masked sparse waypoints, without an interpolated reference condition. Absolute + path cond. adds that condition while retaining absolute prediction. TriFactorHOI retains the reference condition and instead predicts waypoint-relative residuals. Zero waypoint residual follows TriFactorHOI, but sets the denoiser-output residuals at the supplied waypoints to zero and retrains the network. Absolute, Absolute + path cond., and TriFactorHOI use the same sparse waypoints, interaction guidance, and foot refinement, and report means over sampling seeds 1–20.

Adding the interpolated path only as a condition does not reduce the waypoint errors: T_{xy} changes from 0.98 to 1.01 cm, while FID remains 0.37 at the reported precision. Holding this reference condition fixed, residual prediction reduces T_s, T_e, and T_{xy} from 0.59, 2.03, and 1.01 to 0.35, 1.45, and 0.13 cm, respectively, and FID from 0.37 to 0.34.

T_{xy} is evaluated at the supplied waypoints, and the reference path passes through them. The representation therefore aligns directly with these evaluation locations. CHOIS's Lin-OMOMO illustrates why waypoint error cannot be interpreted in isolation: it obtains $T_{xy} = 0$ through interpolation, but reports FID of 15.33 and contact F1 of 0.57, compared with 0.69 and 0.67 for CHOIS (Li et al., 2024). This comparison illustrates the limitation of the metric; it is not a controlled ablation of our path representation. In our matched comparison (Table 5), residual prediction also lowers FID while leaving contact F1 and foot sliding unchanged at the reported precision.

The retrained zero-residual row is the corresponding controlled comparison inside our model. It lowers T_{xy} from 0.13 to 0.00 cm, end error from 1.45 to 1.39 cm, object translation from 8.45 to 8.34 cm, and foot height from 2.99 to 2.93. Start error stays at 0.35 cm. Contact F1 remains 0.80, with contact precision, recall, and ratio unchanged at 0.81, 0.83, and 0.72, and object rotation unchanged at 0.73. The other motion metrics worsen: FID rises from 0.34 to 0.36, top-3 R-precision falls from 0.74 to 0.72, foot sliding from 0.23 to 0.24, hand penetration from 0.54 to 0.55, MPJPE from 13.28 to 13.42, and root translation from 18.93 to 19.16. Hard-setting the residual to zero can make learning harder, because the network must meet the waypoints exactly while still fitting the interaction and the body motion, and the retrained model does not recover the lost quality. The learned residual in TriFactorHOI leaves a waypoint error of only 0.13 cm, which is negligible for this generation task, while keeping the better text alignment, distribution match, pose, and penetration. We therefore retain that residual instead of clamping it to zero.

A.3.3 START-AND-GOAL CONDITIONING

DecHOI conditions generation on text, object geometry, the initial human–object state, and the final object position, without prescribed intermediate waypoints (Jung et al., 2026). Its FullBodyManipulation comparison evaluates CHOIS (Li et al., 2024), HOIFHLI (Wu et al., 2025b), and DecHOI under that shared input. We retrain TriFactorHOI for 100,000 iterations with the same restriction and

Table 6: Start-and-goal evaluation on FullBodyManipulation. No method receives intermediate waypoints. CHOIS, HOIFHLI, and DecHOI entries are the reproductions reported by DecHOI (Jung et al., 2026); TriFactorHOI reports the mean over 20 inference seeds. R_{prec} and FID are listed without rank marks.

	Condition Matching		Human Motion				Interaction				GT Difference			
Method	$T_s \downarrow$	$T_e \downarrow$	$H_{\text{feet}} \downarrow$	FS $\downarrow$	$R_{\text{prec}} \uparrow$	FID $\downarrow$	$C_{\text{prec}} \uparrow$	$C_{\text{rec}} \uparrow$	$C_{\text{F1}} \uparrow$	$P_{\text{hand}} \downarrow$	MPJPE $\downarrow$	$T_{\text{root}} \downarrow$	$T_{\text{obj}} \downarrow$	$R_{\text{obj}} \downarrow$
CHOIS (ECCV 2024)	1.92	8.01	6.15	0.41	0.68	1.58	0.74	0.53	0.58	0.66	18.86	44.04	51.86	1.23
HOIFHLI (ICCV 2025)	1.73	7.65	4.77	0.38	0.62	2.06	0.77	0.61	0.64	0.58	19.31	40.87	50.96	1.18
DecHOI (CVPR 2026)	1.59	6.91	4.42	0.38	0.72	0.33	**0.80**	0.64	0.67	**0.53**	15.27	**25.47**	22.96	0.86
TriFactorHOI	**0.40**	**1.45**	**2.98**	**0.22**	0.73	0.38	**0.80**	**0.82**	**0.79**	**0.53**	**14.85**	25.69	**22.44**	**0.81**

use it at inference. The reference path is interpolated from the object start position and the final object position. No intermediate waypoint observations enter the condition masks or the construction of reference-path tokens and residual targets. The goal does not include the final object orientation or the final human pose. Guidance and foot refinement remain enabled. Condition matching uses only the start and end object positions, since intermediate waypoints are not supplied.

Table 6 shows that this variant reaches start and end errors of 0.40 and 1.45 cm. Foot height and foot sliding are 2.98 and 0.22, and contact F1 is 0.79. Hand penetration is 0.53, matching DecHOI at the reported precision, and root translation is 25.69 cm, next to DecHOI's 25.47 cm. Object translation and rotation are 22.44 cm and 0.81. R_{prec} and FID are 0.73 and 0.38. We do not rank these two metrics against the tabulated baselines, because those entries use DecHOI's reported external evaluator and the evaluator weights have not been aligned.

A.3.4 Number of Optimization Steps

Table 7: Guided-step ablation with 1,000 denoising updates. AnchorIK remains enabled in all rows. Runtime is averaged per sequence with batch size 32; overhead is relative to zero guided steps.

Guided steps	$H_{\text{feet}} \downarrow$	FS $\downarrow$	$C_{\text{prec}} \uparrow$	$C_{\text{rec}} \uparrow$	$C_{\text{F1}} \uparrow$	$C_{\%} \uparrow$	$P_{\text{hand}} \downarrow$	FID $\downarrow$	Time/Seq. (s)$\downarrow$	Overhead$\downarrow$
0	4.85	0.24	0.80	0.58	0.63	0.50	0.65	0.65	**1.012**	**+0.0%**
10	2.99	**0.23**	**0.81**	0.81	0.79	0.71	0.58	0.36	1.338	+32.2%
25 (default)	2.99	**0.23**	**0.81**	**0.83**	**0.80**	**0.72**	0.54	**0.34**	1.811	+79.0%
50	**2.94**	0.24	**0.81**	0.82	0.79	0.71	**0.50**	0.36	2.616	+158.5%

We vary the number of guided reverse updates over 0, 10, 25, and 50 while fixing the 1,000-step diffusion sampler, generator, and full-sequence AnchorIK. All variants use sampling seeds 1–20.

Table 7 shows diminishing returns from additional guidance. Ten updates already improve contact quality and reduce foot sliding and hand penetration. The main contact gains are in recall and contact ratio, while precision remains comparable. Increasing the budget to 25 further improves recall, contact F1, and contact ratio, and lowers hand penetration and FID to the main-table values 0.54 and 0.34. Fifty updates leave contact F1 nearly unchanged and reduce penetration only modestly further, at approximately 44% higher runtime than 25 updates. Foot sliding and FID do not improve further. These results support our 25-step default as a quality–efficiency compromise, rather than a setting that is optimal on every metric.

Runtime in Tables 7 and 8 is the mean per-sequence computation time with batch size 32 on a single NVIDIA RTX PRO 6000 Blackwell Workstation Edition GPU, not single-request latency. It includes text encoding, noise construction, diffusion/guidance, AnchorIK, and predicted mesh reconstruction, but excludes initialization, data loading, ground-truth reconstruction, evaluation, and file I/O. Overhead is measured relative to zero guided updates. Quality metrics use two decimal places and times use three; bold and underlined entries denote the best and second-best distinct displayed values, with ties sharing a rank.

That guidance yields a clear improvement is a shared observation of prior methods, including CHOIS (Li et al., 2024) and HOI-Dyn (Wu et al., 2025a). A smaller gain from changing the network should not be read as evidence that the network contributes little. Network design remains a separate

contribution in this work and in other methods, even when guidance accounts for a larger share of the measured improvement.

A.3.5 Number of Denoising Steps

Table 8: Denoising-step ablation without retraining (seeds 1–20). Guidance is disabled in all four settings; AnchorIK remains fixed. Runtime is averaged per sequence with batch size 32; speedup is relative to the unguided 1,000-step setting.

Denoising steps	$T_{xy} \downarrow$	FID $\downarrow$	$C_{\mathrm{F1}} \uparrow$	FS $\downarrow$	Time/Seq. (s)$\downarrow$	Speedup$\uparrow$
1,000	0.13	0.65	0.63	**0.24**	1.013	1.00×
500	0.13	0.65	**0.65**	0.25	0.564	1.80×
100	0.13	0.64	0.64	0.25	0.199	5.09×
50	**0.12**	**0.63**	**0.65**	0.25	**0.154**	**6.57**×

Table 8 varies the number of denoising updates over 1,000, 500, 100, and 50 without retraining. Interaction guidance is disabled throughout to isolate the sampling budget from guidance allocation. The generator, sampling seeds 1–20, AnchorIK, hardware, batch size, and timing scope remain fixed. All configurations generate motions for the same 482 test sequences; direct metrics use 482 cases, while FID uses 480 evaluator cases.

The 1,000-step setting uses native DDPM; reduced budgets use the ancestral DDPM posterior spanning each skipped interval. Original training time indices are retained: for example, the 50-step schedule is $1000, 980, \ldots, 20$ in one-based notation, followed by the clean endpoint. Speedup is measured relative to the unguided 1,000-step configuration.

Reducing the budget substantially lowers runtime while preserving comparable motion and contact quality in this unguided setting. At 100 steps, runtime falls from 1.013 to 0.199 s per sequence, a 5.09× speedup. At 50 steps, it further decreases to 0.154 s, yielding a 6.57× speedup. Relative to 1,000 steps, the 50-step result has slightly lower FID (0.65 to 0.63) and higher contact F1 (0.63 to 0.65), while waypoint errors remain close to 0.13 cm across budgets. Foot sliding increases modestly from 0.24 to 0.25, indicating a small stability trade-off. These results suggest that fewer denoising updates can accelerate unguided generation without a substantial loss in the reported quality metrics; they do not establish the same trade-off for the full guidance-enabled pipeline. Unlike learned few-step models such as MotionLCM (Dai et al., 2024), this study changes only inference-time sampling and does not retrain or distill the generator.

A.3.6 Applying Our Guidance and Refinement to MaMi-HOI

Table 9: Released MaMi-HOI under TriFactorHOI's interaction guidance and foot refinement. The local row keeps MaMi-HOI's original pipeline. TriFactorHOI repeats Table 1.

	Condition Matching			Human Motion				Interaction					GT Difference			
Method	$T_s \downarrow$	$T_e \downarrow$	$T_{xy} \downarrow$	$H_{\text{feet}} \downarrow$	FS $\downarrow$	$R_{\text{prec}} \uparrow$	FID $\downarrow$	$C_{\text{prec}} \uparrow$	$C_{\text{rec}} \uparrow$	$C_{\text{F1}} \uparrow$	$C_{\%} \uparrow$	$P_{\text{hand}} \downarrow$	MPJPE $\downarrow$	$T_{\text{root}} \downarrow$	$T_{\text{obj}} \downarrow$	$R_{\text{obj}} \downarrow$
TriFactorHOI	**0.35**	**1.45**	**0.13**	**2.99**	**0.23**	**0.74**	**0.34**	**0.81**	**0.83**	**0.80**	**0.72**	**0.54**	**13.28**	**18.93**	**8.45**	**0.73**
MaMi-HOI + guidance/FR	1.30	3.77	2.25	4.20	0.34	0.73	0.36	**0.81**	0.81	0.77	0.70	0.57	13.85	21.04	9.31	0.81
MaMi-HOI (local)	1.40	3.98	2.30	4.16	0.36	0.73	0.38	**0.81**	0.74	0.75	0.63	0.60	14.12	21.17	10.53	0.82

MaMi-HOI is the clearly strongest baseline. Table 9 tests whether TriFactorHOI's advantage remains when the released MaMi-HOI (Wang et al., 2026) generator receives our inference processing. The processed row keeps that generator and applies the same interaction guidance and post-sampling foot refinement as TriFactorHOI: hand–object and foot–ground guidance during denoising, followed by full-sequence support prediction and AnchorIK. The local row reruns the released generator with MaMi-HOI's original pipeline. Table 1 cites the originally reported MaMi-HOI numbers; these two rows are separate local reruns. The TriFactorHOI row repeats the FullBodyManipulation mean from that table.

The advantage remains on trajectory, foot stability, and pose error. Against the processed row, TriFactorHOI reduces start, end, and planar waypoint errors from 1.30, 3.77, and 2.25 to 0.35, 1.45, and 0.13. Foot sliding falls from 0.34 to 0.23, and foot height from 4.20 to 2.99. Contact F1 rises

from 0.77 to 0.80, with contact precision tied at 0.81. FID falls from 0.36 to 0.34. Hand penetration and the joint, root, object-translation, and object-rotation errors are lower as well.

Relative to the local rerun, the shared guidance and refinement raise contact F1 from 0.75 to 0.79 and contact recall from 0.74 to 0.81, and lower hand penetration from 0.60 to 0.57 and foot sliding from 0.36 to 0.34. Waypoint error changes from 2.30 to 2.25. Applying the same inference processing therefore improves MaMi-HOI's contact and penetration, but does not close the gap between TriFactorHOI and MaMi-HOI.

A.3.7 PALM-ORIENTATION GUIDANCE

Table 10: Ablation of the palm-orientation term. Both rows use the same generator, contact and penetration guidance, and full-sequence AnchorIK.

Variant	$C_{\text{prec}} \uparrow$	$C_{\text{rec}} \uparrow$	$C_{\text{F1}} \uparrow$	$C_{\%} \uparrow$	$P_{\text{hand}} \downarrow$	$R_{\text{prec}} \uparrow$	FID $\downarrow$
TriFactorHOI	**0.81**	**0.83**	**0.80**	**0.72**	**0.54**	**0.74**	**0.34**
TriFactorHOI w/o palm	**0.81**	**0.83**	**0.80**	<u>0.71</u>	<u>0.56</u>	**0.74**	<u>0.35</u>

Table 10 isolates the palm-orientation term by removing it from the full TriFactorHOI pipeline. The generator, contact attraction, penetration penalty, and full-sequence AnchorIK stay fixed. Without the palm term, hand penetration rises from 0.54 to 0.56 and FID from 0.34 to 0.35, while the contact ratio falls from 0.72 to 0.71. Text alignment and contact precision, recall, and F1 are unchanged at the displayed precision.

A.3.8 FOOT–GROUND CONTACT REFINEMENT

Table 11: Ablation of foot refinement. All variants use the same generator and interaction guidance.

Variant	FS $\downarrow$	$H_{\text{feet}} \downarrow$	MPJPE $\downarrow$	$R_{\text{prec}} \uparrow$	FID $\downarrow$
TriFactorHOI	**0.23**	2.99	**13.28**	**0.74**	**0.34**
TriFactorHOI w/o FR/AnchorIK	0.30	**2.98**	<u>13.29</u>	**0.74**	**0.34**
Threshold-based support	<u>0.26</u>	2.99	13.32	<u>0.73</u>	**0.34**
Rule support, fixed strength	0.28	2.99	13.33	<u>0.73</u>	**0.34**
w/o temporal IK residual	0.28	**2.98**	13.32	<u>0.73</u>	**0.34**

Table 11 fixes the generator and interaction guidance while varying post-sampling foot refinement. The full pipeline lowers foot sliding (FS) from 0.30 without FR/AnchorIK to 0.23, a 21.3% reduction. Threshold-based support and removal of the temporal IK residual yield $\text{FS} = 0.26$ and 0.28, respectively, indicating that both learned support and temporal coupling contribute. Removing the learned predictor entirely, with rule-based support states, a fixed correction strength, and the same anchors and temporal IK, also yields $\text{FS} = 0.28$. That exceeds the threshold-based row, which still uses the learned correction confidence. The other reported metrics remain close across the five rows at the displayed precision. Appendix A.1 specifies the threshold rule and controls.

A.4 Additional Qualitative Results

Figure 5 expands the main-text comparison with seven interaction cases involving tables, boxes, and a clothes stand, including lifting, rotating, dragging, pushing, kicking, and pulling actions.

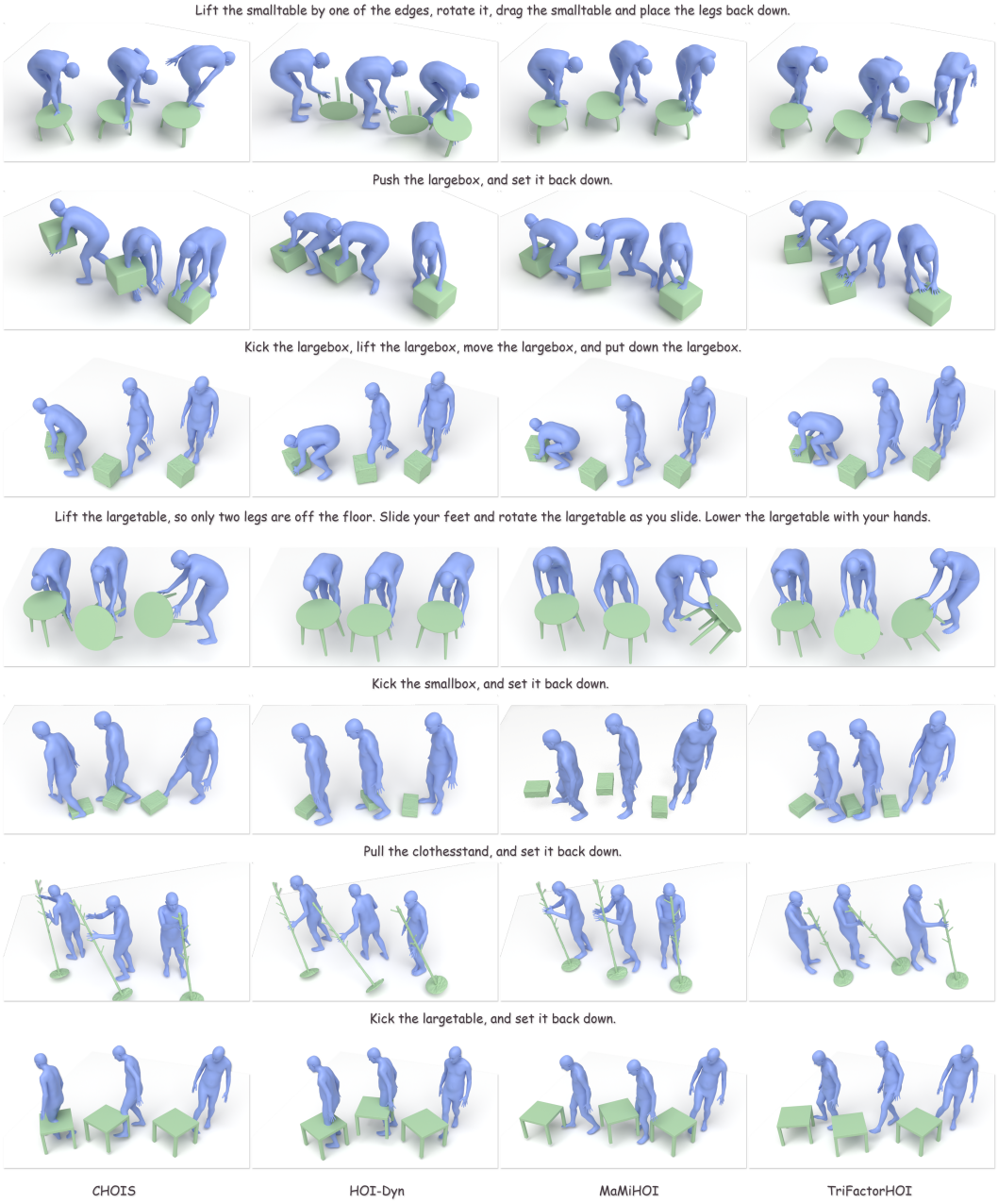

Figure 5: Extended qualitative comparison. Columns show CHOIS, HOI-Dyn, MaMi-HOI, and TriFactorHOI from left to right. Each row corresponds to a text instruction, and each panel shows representative motion frames. Humans and manipulated objects are shown in blue and green, respectively.

A.5 Additional Quantitative Results

Table 12 lists the 20 independent inference runs underlying Table 1, followed by their sample standard deviation (Std, $n-1$ denominator) and arithmetic mean (Mean). The trained networks are fixed across repeats. Each FullBodyManipulation run uses 482 cases for direct metrics and 480 for R_{prec}/FID. Table 13 lists the 20 independent 3D-FUTURE inference runs underlying Table 2, followed by their sample standard deviation (Std, $n-1$ denominator) and arithmetic mean (Mean). Values are shown to four decimal places without best/second-best highlighting.

Table 12: FullBodyManipulation: individual inference seeds 1–20, followed by sample standard deviation (Std) and arithmetic mean (Mean).

	Condition Matching			Human Motion				Interaction					GT Difference			
Seed	$T_s \downarrow$	$T_e \downarrow$	$T_{xy} \downarrow$	$H_{\text{feet}} \downarrow$	FS $\downarrow$	$R_{\text{prec}} \uparrow$	FID $\downarrow$	$C_{\text{prec}} \uparrow$	$C_{\text{rec}} \uparrow$	$C_{\text{F1}} \uparrow$	$C_{\%} \uparrow$	$P_{\text{hand}} \downarrow$	MPJPE $\downarrow$	$T_{\text{root}} \downarrow$	$T_{\text{obj}} \downarrow$	$R_{\text{obj}} \downarrow$
1	0.3513	1.5170	0.1276	2.9745	0.2331	0.7417	0.3566	0.8074	0.8237	0.7946	0.7147	0.5292	13.3876	19.0875	8.4684	0.7411
2	0.3342	1.4872	0.1300	2.9996	0.2333	0.7312	0.3272	0.8153	0.8238	0.7986	0.7138	0.5521	13.2570	18.9011	8.5373	0.7288
3	0.3546	1.4373	0.1275	2.9875	0.2346	0.7333	0.3387	0.8101	0.8267	0.7974	0.7147	0.5343	13.3668	19.0850	8.3649	0.7268
4	0.3411	1.3848	0.1326	2.9935	0.2313	0.7312	0.3495	0.8101	0.8239	0.7978	0.7162	0.5449	13.3901	18.9600	8.6052	0.7254
5	0.3647	1.5122	0.1297	2.9852	0.2336	0.7354	0.3553	0.8121	0.8244	0.7973	0.7132	0.5484	13.2303	18.7092	8.3993	0.7202
6	0.3394	1.4781	0.1291	2.9840	0.2335	0.7271	0.3213	0.8130	0.8265	0.7982	0.7168	0.5341	13.2960	19.0700	8.4947	0.7238
7	0.3390	1.4128	0.1325	2.9954	0.2361	0.7417	0.3183	0.8154	0.8296	0.8029	0.7159	0.5374	13.1659	19.1550	8.2358	0.7358
8	0.3589	1.4046	0.1262	2.9887	0.2318	0.7354	0.3785	0.8150	0.8258	0.7993	0.7157	0.5641	13.1436	18.6294	8.3429	0.7366
9	0.3348	1.3966	0.1304	2.9923	0.2343	0.7500	0.3315	0.8204	0.8276	0.8026	0.7143	0.5259	13.2444	18.8945	8.4197	0.7519
10	0.3410	1.4240	0.1302	2.9998	0.2331	0.7375	0.3457	0.8161	0.8260	0.7993	0.7176	0.5499	13.3344	19.2789	8.4239	0.7251
11	0.3602	1.3875	0.1286	2.9946	0.2324	0.7333	0.3383	0.8177	0.8231	0.7971	0.7142	0.5313	13.2024	19.0554	8.5270	0.7367
12	0.3466	1.5281	0.1314	2.9724	0.2312	0.7542	0.3531	0.8211	0.8262	0.8035	0.7122	0.5458	13.3710	19.1374	8.4726	0.7308
13	0.3562	1.4892	0.1297	2.9834	0.2307	0.7500	0.3158	0.8173	0.8262	0.8001	0.7187	0.5281	13.2711	18.8554	8.5562	0.7398
14	0.3502	1.4874	0.1290	2.9915	0.2316	0.7354	0.3612	0.8155	0.8218	0.7961	0.7094	0.5409	13.3631	18.9892	8.5636	0.7445
15	0.3435	1.4082	0.1304	2.9908	0.2341	0.7458	0.3469	0.8143	0.8223	0.7959	0.7127	0.5523	13.2356	18.6360	8.3468	0.7194
16	0.3296	1.5629	0.1275	2.9853	0.2336	0.7312	0.3502	0.8111	0.8309	0.8029	0.7195	0.5507	13.3014	18.9613	8.5252	0.7512
17	0.3470	1.4283	0.1301	2.9851	0.2365	0.7458	0.3683	0.8148	0.8289	0.8025	0.7186	0.5450	13.1012	18.6221	8.4473	0.7302
18	0.3548	1.4516	0.1305	2.9904	0.2356	0.7354	0.3353	0.8130	0.8267	0.8007	0.7159	0.5415	13.3828	19.0171	8.5010	0.7047
19	0.3336	1.4776	0.1295	2.9736	0.2331	0.7438	0.3368	0.8198	0.8231	0.7979	0.7145	0.5433	13.1739	18.4363	8.3998	0.7302
20	0.3264	1.4234	0.1291	2.9914	0.2322	0.7438	0.3419	0.8142	0.8250	0.7984	0.7166	0.5222	13.2859	19.2027	8.4441	0.7174
Std	0.0108	0.0517	0.0016	0.0078	0.0016	0.0075	0.0164	0.0036	0.0024	0.0026	0.0024	0.0107	0.0881	0.2246	0.0902	0.0116
Mean	0.3454	1.4549	0.1296	2.9879	0.2333	0.7392	0.3435	0.8147	0.8256	0.7992	0.7153	0.5411	13.2752	18.9342	8.4538	0.7310

Table 13: 3D-FUTURE: individual inference seeds 1–20, followed by sample standard deviation (Std) and arithmetic mean (Mean).

	Condition Matching			Human Motion				Interaction	
Seed	$T_s \downarrow$	$T_e \downarrow$	$T_{xy} \downarrow$	$H_{\text{feet}} \downarrow$	FS $\downarrow$	$R_{\text{prec}} \uparrow$	FID $\downarrow$	$C_{\%} \uparrow$	$P_{\text{hand}} \downarrow$
1	3.1778	5.5261	0.9852	2.4977	0.2493	0.7024	1.4576	0.7312	0.1317
2	3.1752	5.5988	0.9905	2.4916	0.2510	0.6830	1.3938	0.7323	0.1365
3	3.1842	5.5311	0.9914	2.5072	0.2521	0.6935	1.4364	0.7316	0.1391
4	3.1886	5.4882	0.9907	2.5019	0.2522	0.6935	1.4833	0.7303	0.1291
5	3.1830	5.5402	0.9875	2.4933	0.2515	0.6815	1.4302	0.7320	0.1365
6	3.1761	5.4968	0.9856	2.4887	0.2533	0.6964	1.4769	0.7327	0.1369
7	3.1766	5.5464	0.9998	2.4920	0.2537	0.6949	1.4657	0.7298	0.1371
8	3.1720	5.5149	0.9866	2.4996	0.2500	0.7024	1.4890	0.7304	0.1328
9	3.1752	5.5231	0.9979	2.5123	0.2526	0.7113	1.4422	0.7349	0.1392
10	3.1792	5.5366	0.9824	2.4738	0.2501	0.6935	1.3979	0.7322	0.1366
11	3.1847	5.5437	0.9817	2.4948	0.2516	0.6949	1.4804	0.7341	0.1393
12	3.1756	5.5094	0.9892	2.4862	0.2505	0.6830	1.4885	0.7322	0.1404
13	3.1703	5.5920	0.9988	2.5179	0.2510	0.6964	1.4401	0.7328	0.1352
14	3.1768	5.5101	0.9849	2.5212	0.2530	0.6875	1.4788	0.7328	0.1351
15	3.1746	5.5577	0.9853	2.5294	0.2539	0.6964	1.4298	0.7326	0.1355
16	3.1822	5.5129	0.9691	2.4687	0.2471	0.6920	1.4319	0.7344	0.1375
17	3.1762	5.5326	0.9882	2.5247	0.2538	0.6949	1.5226	0.7328	0.1358
18	3.1720	5.5192	0.9749	2.5047	0.2527	0.6979	1.5063	0.7295	0.1343
19	3.1881	5.5290	0.9957	2.5146	0.2530	0.6920	1.4688	0.7325	0.1344
20	3.1797	5.5259	0.9794	2.4915	0.2519	0.6964	1.4609	0.7315	0.1333
Std	0.0052	0.0274	0.0077	0.0161	0.0017	0.0070	0.0336	0.0014	0.0028
Mean	3.1784	5.5317	0.9872	2.5006	0.2517	0.6942	1.4591	0.7321	0.1358