PAMI: PART ANCHORED MOTION FOR TEXT TO HUMAN-OBJECT INTERACTION GENERATION

Anonymous authors
Paper under double-blind review

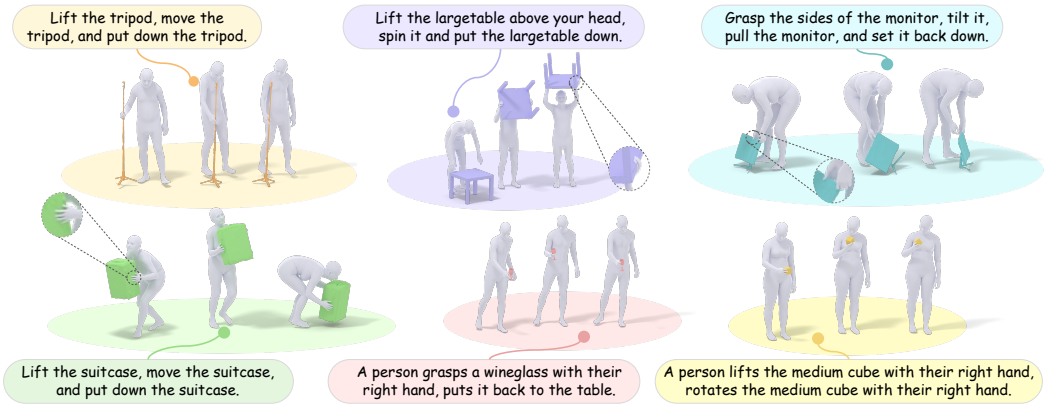


Figure 1: **Text-conditioned full-body HOI generation.** Given a text prompt and a canonical object mesh, PAMI generates coordinated human and object motion across diverse one-handed, two-handed, and full-body interactions. Insets highlight fine-grained contact.

ABSTRACT

Text-conditioned full-body human–object interaction (HOI) generation requires synthesizing human motion and object trajectories that match the input text while remaining precisely coordinated over time. Most methods represent the human and object as separate trajectories and predict the global human-object couplings. Learning this complex, dynamically changing relationship implicitly, however, often yields object drift, missed contact, and penetration. We introduce PAMI, a Part-Anchored Motion framework for Interaction generation. Inspired by the classic Hough Transform, our key idea is to localize object motion by letting body-part anchors vote for it: we express object motion relative to multiple body-part anchors and use PamiVAE to learn an interaction latent space, decoding frame-wise weights that aggregate these part-specific votes. Building on this representation, PAMI generates interactions in a coarse-to-fine hierarchy. PamiGen first generates a coarse human–object interaction from text in this structured latent space, and PamiRefiner then recursively resolves fine-grained contact geometry using a hybrid surface-sensing representation, combining long-range probes that capture overall body-part influence with short-range sensors that resolve detailed contacts near the object surface. Experiments on InterAct show that PAMI generates more faithful interactions and more accurate human-relative object motion than previous methods, achieving 14.5% higher contact recall than the previous state of the art. Extensive ablations validate the contributions of both the part-anchored voting representation and hybrid surface-sensing refinement.

1 INTRODUCTION

Humans constantly interact with objects, making the ability to generate human–object interaction (HOI) a cornerstone of physical intelligence and embodied AI, with direct implications for robotics,

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.

animation, and AR/VR. Driving this generation from natural language is especially valuable, as text offers an intuitive and scalable interface for specifying the rich space of interactions without motion capture or manual authoring. Text-conditioned full-body HOI generation is however non-trivial: it must produce a human motion and an object trajectory that match the text and stay coordinated over time. This makes HOI more than the independent generation of two plausible motions, as small relative errors can produce a floating grasp, a drifting object, missed contact, or penetration.

Most full-body text-to-HOI methods diffuse explicit per-frame human and object states Xu et al. (2023; 2025a); Wang et al. (2026b). In text-to-motion, learned temporal representations have shown that motion can instead be generated in a compact latent space (Chen et al., 2023; Zhang et al., 2023; Jiang et al., 2023; Li et al., 2024). Recent work HOIMask Ji et al. (2026) is the first to use vector quantized latents for full body interaction. However, it represents interaction as two separate trajectories, which cannot capture the complex human-object relations in latent space. Similarly, most other methods Wang et al. (2026b) also use the separate human-object trajectories as representation and rely on feature exchange Wang et al. (2026b), contact prediction Diller & Dai (2024), or guidance based correction Peng et al. (2023) to reason about the interactions. Some methods explore predicting contact distances Xu et al. (2025a) or weighted combination of multiple predictions based on distances Cha et al. (2024); Xu et al. (2025a). But such feature space based correlation reasoning or hard coded weights is suboptimal as the network needs to reason about the complete object geometry under all dynamic pose transformations and even from body regions that are not affecting the object, making it hard to learn the essential correlations efficiently. In this paper, we ask *what representation is suitable for generating interaction motion in latent space while respecting the intricate human-object interaction and contacts?*

To this end, we introduce **PAMI**: **P**art **A**nchored **M**otion voting for **I**nteraction motion generation in latent space. We take inspiration from the classic Hough Transform (Ballard, 1981) in computer vision, where a global structure that is hard to detect directly is instead localized by letting many local elements each vote for it and taking their consensus, an aggregation that is robust to noise and to irrelevant votes. We cast object motion in the same way: the object pose is the global structure, and the body-part anchors are the local voters. We express object translation relative to each body-part anchor in its local frame, so every anchor votes for where the object should be, and let the decoder predict how strongly each anchor votes per frame. Combining these votes with learned weights forms a soft, differentiable Hough aggregation that naturally down-weights parts not involved in the interaction. We embed this representation with PamiVAE, a variational autoencoder (VAE), into a compact interaction latent that preserves subtle geometric relations yet is easier to generate.

Human–object interaction motion is highly dynamic and complex especially from only text as input. We therefore adopt a hierarchical coarse-to-fine design, first recovering an approximate interaction and then resolving its fine-grained contact geometry. In the coarse stage, we train PamiGen, a text-conditioned generator that produces interaction latents decoded into a human–object motion that captures the text, the overall movement, and the anchor-voted human–object relation. In the fine stage, PamiRefiner recursively refines this motion through hybrid local geometry sensing: long-range probes capture overall body-part influence, while short-range sensors capture detailed contact in the immediate vicinity of the surface. This decomposition lets latent voting establish the coarse interaction while surface sensing recovers the fine, contact-rich geometry.

Experiments on InterAct Xu et al. (2025a) dataset show that our method generates more faithful interaction motions with more accurate object motions relative to the human than prior arts. Notably, PAMI achieves 14.5% higher contact recall than previous state-of-the-art method. Extensive ablations also validate that our part-anchored representation is important to learn a meaningful latent space for HOI generation and that our hybrid surface sensing and refinement improves geometric correctness of the generated motions. In summary, our main contributions are:

- We present PAMI, to our knowledge, the **first part-based latent generative framework** for text-conditioned full-body HOI motion generation. Our generation is hierarchical from coarse to fine and achieves the best performance on the benchmark and generates high-quality interaction motion from text. Our code and model will be publicly released.
- We introduce a **body-part based object voting representation** that anchors object motion to body parts and lets anchors vote for object movement via learned weights. This allows efficient generation in latent space while preserving subtle human-object relations, yielding semantically well aligned coarse interaction motion.

- We design a **hybrid local surface sensing module** to train a RefineNet that reasons fine human-object geometry and contacts to improve the coarse interaction motion from latent generation. Trained with mixed data from generation and ground truth, our RefineNet improves not only our latent generation results but also baseline generations zero-shot.

2 Related Work

Text-to-HOI and latent motion generation. Text-conditioned interaction generation has been studied for hand–object motion (Cha et al., 2024; Christen et al., 2024; Song et al., 2026) and for full-body human–object motion. Early full-body methods condition generation on intent or object trajectories (Ghosh et al., 2022; Li et al., 2023b;a); more recent diffusion and related generative models generate interactions from text and incorporate contact, object geometry, or human–object relation features (Xu et al., 2023; Peng et al., 2023; Diller & Dai, 2024; Wu et al., 2024; Song et al., 2024; Xu et al., 2024; Zeng et al., 2025; Xue et al., 2025; Geng et al., 2025; Li et al., 2025b; Jung et al., 2026; Cai et al., 2026; Wang et al., 2026a; Wu et al., 2025; Di et al., 2026; Zou et al., 2026). HIMO (Lv et al., 2024) handles multiple objects, InterAct (Xu et al., 2025a) introduces a marker-based representation and a large-scale benchmark, and LIGHT (Wang et al., 2026b) denoises body, hand, and object streams at different paces. Most full-body methods operate directly on per-frame motion states, coupling the human and object through feature exchange, relation or contact prediction, or sampling guidance. LatentHOI (Li et al., 2025a) uses latent diffusion for hand–object motion, but does not model full-body interaction. In text-to-motion, learned tokenizers and temporal autoencoders have made generation in compact latent spaces common (Chen et al., 2023; Zhang et al., 2023; Jiang et al., 2023; Li et al., 2024; Nazarenus et al., 2026). Whole-body and part-based models further separate the dynamics of different body regions (Lu et al., 2023; Li et al., 2026a). We design a novel part-anchored voting representation and embed it into a compact latent space for faithful interaction generation.

Local and relative object representations. Relative object representations encode human–object coordination directly (Li et al., 2023b; Zeng et al., 2025). Local modeling also appears in STAR (Osman et al., 2020), which restricts pose-dependent body deformation to sparse joint neighborhoods rather than coupling every joint to the full surface. In HOI generation, InterAct (Xu et al., 2025a) uses vectors from body markers to the nearest object surface and weights relative object-motion guidance by marker–object distance. CG-HOI (Diller & Dai, 2024) predicts object transformations associated with body-surface markers and combines them using contact-distance weights. In both cases, distance determines which human locations influence the object. Dex2HOI (Pratikaki et al., 2026) instead predicts mixture weights for left-wrist-relative, right-wrist-relative, and free object trajectories. Our representation extends this learned composition from wrists to full-body and embed them in latent space.

Guidance and refinement for HOI. Physical errors in generated interactions can be addressed during training, sampling, or after generation. Contact-aware losses and auxiliary contact or affordance prediction supervise the generator (Li et al., 2023a; Ma et al., 2024; Peng et al., 2023; Diller & Dai, 2024; Cha et al., 2024; Song et al., 2026; Di et al., 2026), while guidance and relation intervention modify diffusion sampling (Karunratanakul et al., 2023; Wu et al., 2024; Xue et al., 2025; Wang et al., 2026b; Li & Dai, 2025; Li et al., 2026b; Ron et al., 2026). InterDiff (Xu et al., 2023) performs kinematics-informed iterative correction, and OOD-HOI (Zhang et al., 2024b) uses predicted contact regions during diffusion inference. Other work applies optimization Xie et al. (2022; 2023); Lee et al. (2026), learned Zhou et al. (2022); Taheri et al. (2024); Xie et al. (2024); Wang et al. (2026a) or physics-based correction (Zhang et al., 2024a; Xu et al., 2024; 2025b; 2026; Lin et al., 2025). GEARS (Zhou et al., 2024) and GRAFT (YM et al., 2026) use local geometry for hand–object synthesis and human–scene fitting, respectively. We follow the widely used hierachical design and use latent generation for coarse and local geometry sensing for fine interaction generation.

3 Method

Given a text prompt c, a canonical object mesh $\mathcal{M}$, our goal is to generate a human motion $\mathbf{x}^h_{1:T}$ and a rigid-object trajectory $\mathbf{x}^o_{1:T}$ of T frames that form one coordinated interaction. We use SMPL-H Loper et al. (2015); Romero et al. (2017) to represent human motion. Since the generation must

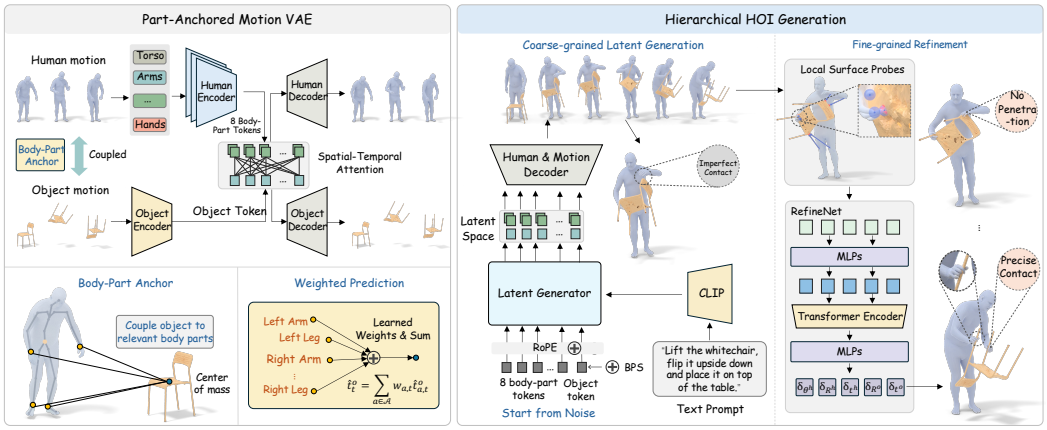


Figure 2: **Overview of PAMI.** Body-part anchors vote for the object motion, and PamiVAE embeds this part-anchored interaction into a latent space (Sec. 3.1). Generation then proceeds coarse-to-fine: PamiGen samples a coarse interaction from text in the latent space (Sec. 3.2), which PamiRefiner refines in explicit space using hybrid long- and short-range surface-sensing probes to sharpen contacts (Sec. 3.3).

jointly satisfy the text, the human motion, and their fine-grained coupling, we adopt a coarse-to-fine design, illustrated in Fig. 2. PamiVAE first learns a compact interaction latent space built on our part-anchored voting representation (Sec. 3.1). A text-conditioned generator PamiGen then produces a coarse interaction in this space (Sec. 3.2). Finally, PamiRefiner operates in explicit space and recursively refines this interaction to recover fine-grained contact geometry (Sec. 3.3).

3.1 Part-Anchored Interaction VAE

We aim at compressing the Euclidean-space interaction into a compact latent capturing the human-object motions and their coupling. We use a variational autoencoder to encode the two with separate body-part and object encoders then couple them via spatiotemporal attention, and decodes the human motion together with a part-anchored voting representation of object motion, detailed next.

Part-anchored object motion representation. Because object motion is most strongly coupled to contacting or supporting body parts, we predict it in the spirit of the Hough Transform: each body part votes for object motion, and per-frame weights aggregate the votes. For detached objects, no body part should drive motion, so we add a free anchor that votes independently of the human.

For each sequence, we define the world coordinate frame as the human root coordinate in the first frame and express all subsequent global human and object motion relative to this frame. We write the rigid-object trajectory to be generated as $\mathbf{x}_t^o = (\mathbf{R}_t^o, \mathbf{t}_t^o)$, where $\mathbf{R}_t^o \in \mathrm{SO}(3)$ and $\mathbf{t}_t^o \in \mathbb{R}^3$ denote the object orientation and translation at frame t, respectively. Let $\mathcal{A}$ denote a set of $K = 6$ body-part anchors, which includes the selected body joints Li et al. (2026a) of SMPL model, and one additional free anchor that is the same as the world space. Let $\mathbf{p}_{a,t}^h, \mathbf{R}_{a,t}^h$ be the local coordinate frame of an anchor $a \in \mathcal{A}$, we represent the object translation relative to each anchor as: $\mathbf{\Delta}_{a,t}^o = \left(\mathbf{R}_{a,t}^h\right)^\top \left(\mathbf{t}_t^o - \mathbf{p}_{a,t}^h\right)$. For anchors belonging to the body joints, we perform forward kinematic to obtain $\mathbf{p}_{a,t}^h, \mathbf{R}_{a,t}^h$ and for the free anchor we have $\mathbf{p}_{\text{free},t}^h = 0, \mathbf{R}_{\text{free},t}^h = \mathbf{I}$. Hence $\mathbf{t}_{\text{free},t}^o = \mathbf{t}_t^o$. The object motion representation at each frame can hence be written as: $\{\mathbf{R}_t^o, \mathbf{\Delta}_{a,t}^o, a \in \mathcal{A}\}$.

Our object VAE decoder reconstructs the object rotation $\mathbf{R}_t^o$, the relative translations $\widehat{\mathbf{\Delta}}_{a,t}^o$, and routing logits $\boldsymbol{\alpha}_t \in \mathbb{R}^K$ for the K anchors. Each body-part-anchored trajectory is then $\widehat{\mathbf{t}}_{a,t}^o = \widehat{\mathbf{p}}_{a,t}^h + \widehat{\mathbf{R}}_{a,t}^h \widehat{\mathbf{\Delta}}_{a,t}^o$. The final object translation is obtained through learned soft voting:

$$\widehat{\mathbf{t}}_t^o = \sum_{a \in \mathcal{A}} w_{a,t} \widehat{\mathbf{t}}_{a,t}^o, \qquad \mathbf{w}_t = \mathrm{softmax}(\boldsymbol{\alpha}_t). \tag{1}$$

At decoding time, the anchor coordinate frames are computed from the decoded human motion, coupling each trajectory dynamically to the anchors via learned voting. We show in Tab. 2 that this representation produces better coordinated object motion relative to human.

Part-factored human–object interaction encoding. We encode human and object motion separately. For human, we represent the motion following Xiao et al. (2025) except we use global rotation $\mathbf{R}^h$ and translation $\mathbf{t}^h$ instead of relative ones, and adopt the semantic body-part partition used in FrankenMotion (Li et al., 2026a). Let $\mathcal{B}$ denote the set of human body parts and $\mathbf{x}^{h,b}_{1:T}$ the motion features of part $b \in \mathcal{B}$. We introduce a temporal variational encoder E^h_b for each part, producing $\mathbf{z}^{h,b} = E^h_b\left(\mathbf{x}^{h,b}_{1:T}\right) \in \mathbb{R}^{T' \times d_z}, b \in \mathcal{B}$. For object motion, a temporal variational encoder E^o takes the rigid-object rotation $\mathbf{R}^o_{1:T}$ and anchor relative translations $\mathbf{\Delta}^o_{a,1:T}$ defined above as input: $\mathbf{z}^o = E^o\left(\mathbf{R}^o_{1:T}, \{\mathbf{\Delta}^o_{a,1:T}\}_{a\in\mathcal{A}}\right)$.

All encoders downsample the input by a factor of four, such that each latent token represents four motion frames and $T' = T/4$. These encoders extract human and object motion features separately, we then apply several layers of spatiotemporal attention to allow them to exchange information and learn the coupling of interactions. Each layer applies self attention with all human-part and object tokens stacked together, followed by causal temporal attention over the sequence. Denoting F_{ST} as the spatiotemporal layers, the resulting VAE latent is:

$$\mathbf{Z} = F_{\mathrm{ST}}\Big([\mathbf{z}^{h,1}, \ldots, \mathbf{z}^{h,|\mathcal{B}|}, \mathbf{z}^o]\Big) \in \mathbb{R}^{T' \times (|\mathcal{B}|+1) \times d_z}, \tag{2}$$

Since the attention layers keep the original token positions, we simply split $\mathbf{Z}$ along the token dimension into $\mathbf{Z}^h \in \mathbb{R}^{T' \times |\mathcal{B}| \times d_z}$, which contains the latent sequence of every human body part, and $\mathbf{Z}^o \in \mathbb{R}^{T' \times d_z}$, which contains the object latent sequence.

Interaction decoding. To predict interaction motions, we upsample $\mathbf{Z}^h, \mathbf{Z}^o$ and use separate modules D_h, D_o to decode human and object motions respectively: $\widehat{\mathbf{x}}^h_{1:T} = D_h(\mathbf{Z}^h), \widehat{\mathbf{x}}^o_{1:T} = D_o(\mathbf{Z}^o; \widehat{\mathbf{x}}^h_{1:T})$. Here, D_o denotes the object decoder prediction followed by the part-anchored composition in Eq. 1. Architectural details are provided in Appendix B.

Training objective. We train the VAE with standard L2 distance between predicted and ground truth items, KL divergence loss $\mathcal{L}_{\mathrm{KL}}$ and auxiliary losses $\mathcal{L}_{\mathrm{comp}}, \mathcal{L}_w$ to supervise anchor weights prediction:

$$\mathcal{L}_{\mathrm{VAE}} = \lambda_h \mathcal{L}^h_{\mathrm{rec}} + \lambda_o \mathcal{L}^o_{\mathrm{rec}} + \lambda_{\mathrm{KL}} \mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{comp}} \mathcal{L}_{\mathrm{comp}} + \lambda_{\mathrm{w}} \mathcal{L}_{\mathrm{w}}. \tag{3}$$

where $\mathcal{L}^h_{\mathrm{rec}}, \mathcal{L}^o_{\mathrm{rec}}$ are the L2 loss on the reconstructed human and object motions respectively. Because these terms do not directly supervise the result after anchor composition, we additionally compare the composed translation from Eq. 1 with the observed object translation: $\mathcal{L}_{\mathrm{comp}} = \frac{1}{3T}\sum_{t=1}^{T}\left\|\widehat{\mathbf{t}}^o_t - \mathbf{t}^o_t\right\|_2^2$. To warm start and regularize the weight prediction, we also introduce L2 loss between the predicted weights $\mathbf{w}^a_t$ and pseudo ground truth weights $\tilde{\mathbf{w}}^a_t$. We assign $\tilde{\mathbf{w}}^a_t$ to one if any vertex of this body part has contact with the object and zero otherwise. For frames with no contacts to any parts, we assign the free anchor weight $\tilde{\mathbf{w}}^{\text{free}}_t$ as one and keep others as zero. The weights are then normalized via softmax and contact is defined as distance smaller than 1cm. Hence $\mathcal{L}_w = \frac{1}{T|\mathcal{A}|}\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\|\mathbf{w}^a_t - \widetilde{\mathbf{w}}^a_t\|_2^2$. We provide details of the loss weights in Appendix.

3.2 Text-conditioned latent interaction generation

PamiVAE provides a compact latent space for coupled human-object motion. We cast interaction generation as a coarse-to-fine process, first generating coarse interactions via a flow matching model PamiGen to generate latents $\mathbf{Z}$ conditioned on text prompt c and canonical object mesh $\mathcal{M}$.

Rectified-flow training. For a clean interaction latent $\mathbf{Z}_0$, we draw $\tau \sim \mathcal{U}[0,1]$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and construct noisy latent via: $\mathbf{Z}_\tau = (1-\tau)\mathbf{Z}_0 + \tau\boldsymbol{\epsilon}, \mathbf{v}^\star = \boldsymbol{\epsilon} - \mathbf{Z}_0$. The transformer predicts $\mathbf{v}_\theta(\mathbf{Z}_\tau, \tau, c, \mathcal{M})$ and we train it with standard rectified flow loss $\mathcal{L}_{\mathrm{RF}}$ Liu et al. (2023).

Explicit Euclidean space motion supervision. The rectified-flow objective supervises latent velocity but does not directly constrain the object trajectories recovered in physical space. We therefore estimate the clean latent and decode it to compute losses defined in explicit Euclidean space.

The clean latent is computed as: $\widehat{\mathbf{Z}}_0 = \mathbf{Z}_\tau - \tau \mathbf{v}_\theta(\mathbf{Z}_\tau, \tau, c, \mathcal{M})$. Decoding this latent yields predictions of human and object motions which we denote as $\hat{\cdot}$. We introduce self-consistency loss $\mathcal{L}_{\text{align}}$ between object motion $\hat{\mathbf{t}}^{\text{o}}_{\text{free},t}$ anchored to free trajectory and motions anchored to body parts and Huber distance loss $\mathcal{L}_{\text{traj}}$ on the predicted object trajectories. Hence the total training loss is: $\mathcal{L} = \lambda_{\text{RF}}\mathcal{L}_{\text{RF}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{traj}}\mathcal{L}_{\text{traj}}$, where $\mathcal{L}_{\text{align}}$ forces the predicted free trajectory $\hat{\mathbf{t}}^{\text{o}}_{\text{free},t}$ to be close to the body part anchored trajectories $\hat{\mathbf{t}}^{o}_{a,t}$: $\mathcal{L}_{\text{align}} = \sum_{t\in\mathcal{C}} \left[\sum_{a\in\mathcal{A}} \overline{w}_{a,t} H_\delta(\widehat{\mathbf{t}}^{o}_{\text{free},t}, \widehat{\mathbf{t}}^{o}_{a,t})\right]$, where $t \in \mathcal{C}$ denotes the loss is applied only to frames where object is in contact with human and H_δ denotes Huber distance. We adopt Huber distance to allow stochastic generation of motions that satisfy interaction constrain but deviate from ground truth trajectory. Exact masks, weights, and the noise interval are provided in Appendix B.

Sampling. At inference, we start from Gaussian noise at $\tau = 1$, integrate backward to $\tau = 0$ with Classifier-free guidance, and decode the resulting interaction latent.

3.3 Recursive geometric refinement

The second stage of our hierarchical generation is fine-grained generation by training PamiRefiner using a hybrid local surface sensing representation to iteratively update the fine interaction details.

Hybrid surface sensors. Our sensors includes surface geometric probes to reason both long range and short range influence of body to the object for more geometrically correct interaction. We denote $F_{\text{long},t}, F_{\text{short},t}$ as the features extracted from long and short probes at frame t respectively and $p^h_{j,t}$ as the SMPL-H skeleton joint at index j. Similar to GRAFT (YM et al., 2026), the long range probe features F_{long} include the direction and distance from $J_b = 22$ body joints to its closest object surface point $p^o_{j,t}$, together corresponding surface normal at $p^o_{j,t}$: $F_{\text{long},t} = \{(\mathbf{v}_{j,t}, ||\mathbf{v}_{j,t}||_2, \mathbf{n}(p^o_{j,t})), \mathbf{v}_{j,t} = p^o_{j,t} - p^h_{j,t}, \forall j \in \{1, ..., J_b\}\}$. These features provide information of how overall body pose affects the object movement but it cannot reason the fine-grained grasp pose which requires more detailed information about the object surface geometry near the fingers. To this end, we introduce short range probes on the hand joints and fingertips that includes not only the direction, distance and object surface normal, but also the object surface points $\mathcal{P}^o_{j,t}$ that are inside a given radius r around these hand points, see visualization in Fig. 2 stage 3 topleft. Denoting J_h as the total number of hand joints and fingertips, we have $F_{\text{short},t} = \{(\mathbf{v}_{j,t}, ||\mathbf{v}_{j,t}||_2, \mathbf{n}(p^o_{j,t}), \mathcal{P}^o_{j,t}), \forall j \in \{1, ..., J_h\}\}$ where $\mathcal{P}^o_{j,t} = \{p^o_t | \quad \forall p^o_t \in \mathcal{M}_t, ||p^o_t - p^h_{j,t}||_2 < r\}$. Following GEARS Zhou et al. (2024), we transform object points $\mathcal{P}^o_{j,t}$ to the local coordinate of its query point and compute geometric features like distance and surface normals as input to our PamiRefiner. We uniformly sample points on the object surface to obtain $\mathcal{M}_t$ and use PointNet Qi et al. (2017) like encoder to map them into a fixed dimensional patch feature. Empty spheres are zero-padded. More details can be found in appendix.

Refinement network PamiRefiner. The input to our network includes the hybrid probe features $F_{\text{long},t}, F_{\text{short},t}$, human shape parameter, global 6D human pose and object 6D pose. We use an MLP to extract features and then use transformer to compute attention across all tokens and through time. The prediction heads then output delta updates to the coarse human-object interaction motion decoded from generated latents: $\delta_{\theta^h}, \delta_{\mathbf{R}^h}, \delta_{\mathbf{t}^h}, \delta_{\mathbf{R}^o}, \delta_{\mathbf{t}^o}$.

Training with mixed data. We train our network to refine both corrupted GT and generated motions, denoted as *recovery stream* and *generation stream* respectively. This ensures it learns the error pattern of our latent generator while also maintaining generalization. To obtain recovery stream data, we use our trained VAE to encode ground truth interaction motions, and add temporally smooth noise to the decoded motions. For these kinds of input, the GT motions are available hence we train directly using L2 distance between refined output and GT motions. More details in Sec. B.5.

The generation stream refines samples from the latent generator. Since a text prompt can admit multiple valid motions, direct supervision against a single ground-truth motion is inappropriate. Instead, we optimize geometric objectives targeting motion and geometric quality:

$$\mathcal{L}_{\text{gen}} = \lambda_c \mathcal{L}_{\text{contact}} + \lambda_p \mathcal{L}_{\text{penetration}} + \lambda_s \mathcal{L}_{\text{support}} + \lambda_m \mathcal{L}_{\text{smooth}} + \lambda_n \mathcal{L}_{\text{preserve}}. \tag{4}$$

Here, $\mathcal{L}_{\text{contact}}$ pulls relevant body parts toward the object and reduces slip during contact, while $\mathcal{L}_{\text{penetration}}$ penalizes joints inside the object. $\mathcal{L}_{\text{support}}$ prevents planted feet from sliding, drifting,

Table 1: **Quantitative evaluation on InterAct** (Xu et al., 2025a). R-precision is evaluated with a batch size of 64. $\rightarrow$ indicates that closer to the ground truth is better. $\pm$ indicates a 95% confidence interval. Our method achieves best performance on almost all metrics.

Method	R-Precision (Top 1)$^{\uparrow}$	FID$^{\downarrow}$	Multimodal Dist.$^{\downarrow}$	FSR$^{\downarrow}$	Pene$^{\downarrow}$	Contact$^{\rightarrow}$	Interaction (hands)$^{\uparrow}$		
							C_{prec}	C_{rec}	C_{F1}
Ground Truth	$0.859^{\pm 0.000}$	$0.000^{\pm 0.000}$	$1.475^{\pm 0.003}$	$0.051^{\pm 0.000}$	$0.060^{\pm 0.000}$	$0.219^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$
HOI-Diff	$0.703^{\pm 0.015}$	$0.698^{\pm 0.017}$	$2.906^{\pm 0.002}$	$0.042^{\pm 0.000}$	$0.092^{\pm 0.004}$	$0.090^{\pm 0.001}$	$0.754^{\pm 0.009}$	$0.531^{\pm 0.007}$	$0.574^{\pm 0.009}$
CHOIS	$0.717^{\pm 0.005}$	$0.544^{\pm 0.003}$	$2.659^{\pm 0.006}$	$0.072^{\pm 0.001}$	$0.122^{\pm 0.005}$	$0.141^{\pm 0.002}$	$0.823^{\pm 0.003}$	$0.701^{\pm 0.007}$	$0.723^{\pm 0.004}$
InterDiff	$0.779^{\pm 0.000}$	$0.302^{\pm 0.018}$	$2.364^{\pm 0.004}$	$0.069^{\pm 0.001}$	$0.110^{\pm 0.006}$	$0.150^{\pm 0.001}$	$0.834^{\pm 0.008}$	$0.718^{\pm 0.003}$	$0.741^{\pm 0.004}$
Text2HOI	$0.710^{\pm 0.007}$	$0.284^{\pm 0.012}$	$2.551^{\pm 0.011}$	$\mathbf{0.035}^{\pm 0.001}$	$0.104^{\pm 0.002}$	$0.108^{\pm 0.000}$	$0.840^{\pm 0.010}$	$0.617^{\pm 0.007}$	$0.670^{\pm 0.006}$
InterAct	$0.805^{\pm 0.008}$	$0.280^{\pm 0.016}$	$2.652^{\pm 0.029}$	$0.065^{\pm 0.009}$	$0.119^{\pm 0.028}$	$0.121^{\pm 0.005}$	$0.668^{\pm 0.004}$	$0.552^{\pm 0.011}$	$0.555^{\pm 0.010}$
LIGHT	$0.719^{\pm 0.003}$	$0.153^{\pm 0.019}$	$2.627^{\pm 0.023}$	$0.050^{\pm 0.001}$	$0.121^{\pm 0.001}$	$0.152^{\pm 0.002}$	$0.832^{\pm 0.003}$	$0.730^{\pm 0.003}$	$0.751^{\pm 0.003}$
PAMI	$\mathbf{0.840}^{\pm 0.003}$	$\mathbf{0.092}^{\pm 0.003}$	$\mathbf{2.195}^{\pm 0.016}$	$0.078^{\pm 0.001}$	$\mathbf{0.062}^{\pm 0.001}$	$\mathbf{0.193}^{\pm 0.001}$	$\mathbf{0.843}^{\pm 0.002}$	$\mathbf{0.836}^{\pm 0.006}$	$\mathbf{0.821}^{\pm 0.004}$

or penetrating the ground; $\mathcal{L}_{\text{smooth}}$ regularizes abrupt human and object motion; and $\mathcal{L}_{\text{preserve}}$ discourages unnecessary changes to non-interacting body regions, the root, and object trajectory. All losses are evaluated at each refinement step and averaged over three iterations.

At inference, the same feed-forward PamiRefiner is applied recursively for a small fixed number of steps. We show in Tab. 3 that our mixed data training is important for more faithful generation. More implementation details are provided in Appendix B.

4 Experiments

We compare against existing baselines and ablate key components: our method outperforms prior approaches on most metrics, and the ablations confirm the value of each design. PamiRefiner also generalizes zero-shot to interactions produced by other methods, improving their geometry.

Experimental and evaluation setup. We conduct experiments on the standard InterAct (Xu et al., 2025a) benchmark and follow LIGHT (Wang et al., 2026b) to report metrics along three dimensions: **semantic alignment** using R-precision and multimodal distance (MM Dist); **motion quality** using FID, diversity, and foot-skating ratio (FSR); and **interaction geometry** using penetration depth (Pene), contact ratio (Contact), and interaction precision, recall, and F1. While LIGHT evaluates the latter three metrics only on hand joints, we report them for both hand and all body joints. Additional metrics are reported in the appendix.

4.1 Baseline Comparison

We compare our method against existing baselines quantitatively in Table 1 and qualitatively in Figure 4. Table 1 shows that our method achieves large improvements in both semantic alignment (R-1) and interaction geometry, particularly in penetration and contact. As shown in Figure 4, LIGHT (Wang et al., 2026b) models human and object motion as separate trajectories, so the object often fails to follow the human (chair, tripod). InterAct (Xu et al., 2025a) relies on body-surface markers and their object distances, spreading attention across markers that may be far from the object hence it often misses contact, leaving gaps or unrealistic relative motion. Our anchor representation directly captures human–object coupling in a compact latent space for coherent generation, while local surface sensing refines interaction geometry. Even on the slender tripod (third row), our model keeps a secure, natural grasp without penetration.

4.2 Ablation Studies

We ablate the main components of our method along two dimensions: the part-anchored representation and latent structure in Table 2, and the design of PamiRefiner in Table 3.

4.2.1 Part anchored voting representation

Table 2 evaluates the representation designs after refinement. For a fair comparison, each generator variant is paired with a PamiRefiner trained on its own outputs. The corresponding coarse gener-

Table 2: **Ablation for part anchored voting representation**. Our part factorization, anchor design with learned routing and absolute root representation all contribute to better generation.

	R-Precision (Top 1)↑	FID↓	Pene↓	Contact→	Interaction (full body)↑			Interaction (hands)↑		
					C_{prec}	C_{rec}	C_{F1}	C_{prec}	C_{rec}	C_{F1}
Ground Truth	0.859	0.000	0.060	0.219	1.000	1.000	1.000	1.000	1.000	1.000
Our full PAMI model	**0.840**$^{\pm 0.003}$	**0.092**$^{\pm 0.003}$	**0.062**$^{\pm 0.001}$	**0.193**$^{\pm 0.001}$	**0.502**$^{\pm 0.003}$	**0.477**$^{\pm 0.003}$	**0.457**$^{\pm 0.004}$	0.843$^{\pm 0.002}$	**0.836**$^{\pm 0.006}$	**0.821**$^{\pm 0.004}$
(a) w/o part factorization	0.822$^{\pm 0.002}$	0.171$^{\pm 0.006}$	0.086$^{\pm 0.001}$	0.129$^{\pm 0.002}$	0.466$^{\pm 0.002}$	0.294$^{\pm 0.005}$	0.327$^{\pm 0.004}$	0.823$^{\pm 0.007}$	0.657$^{\pm 0.006}$	0.697$^{\pm 0.006}$
(b) w/o leg and root anchors	**0.840**$^{\pm 0.002}$	0.097$^{\pm 0.006}$	0.087$^{\pm 0.002}$	0.174$^{\pm 0.001}$	0.499$^{\pm 0.005}$	0.425$^{\pm 0.004}$	0.425$^{\pm 0.004}$	**0.849**$^{\pm 0.002}$	0.806$^{\pm 0.004}$	0.804$^{\pm 0.003}$
(c) w/o anchors	0.834$^{\pm 0.003}$	0.102$^{\pm 0.004}$	0.076$^{\pm 0.003}$	0.116$^{\pm 0.001}$	0.478$^{\pm 0.015}$	0.277$^{\pm 0.007}$	0.313$^{\pm 0.009}$	0.830$^{\pm 0.006}$	0.647$^{\pm 0.003}$	0.686$^{\pm 0.001}$
(d) w/o learned voting weights	0.831$^{\pm 0.001}$	0.099$^{\pm 0.005}$	0.089$^{\pm 0.003}$	0.175$^{\pm 0.002}$	0.495$^{\pm 0.005}$	0.423$^{\pm 0.002}$	0.423$^{\pm 0.002}$	0.844$^{\pm 0.003}$	0.807$^{\pm 0.009}$	0.802$^{\pm 0.007}$
(e) w/o absolute root	0.780$^{\pm 0.005}$	0.401$^{\pm 0.011}$	0.063$^{\pm 0.002}$	0.126$^{\pm 0.002}$	0.478$^{\pm 0.006}$	0.312$^{\pm 0.005}$	0.342$^{\pm 0.003}$	0.829$^{\pm 0.006}$	0.698$^{\pm 0.002}$	0.720$^{\pm 0.002}$

Table 3: **PamiRefiner ablations**. All variants refine the same coarse generations. Mixed data training and hybrid sensing are important to obtain high-quality fine interaction details.

	R-Precision (Top 1)↑	FID↓	Pene↓	Contact→	Interaction (full body)↑			Interaction (hands)↑		
					C_{prec}	C_{rec}	C_{F1}	C_{prec}	C_{rec}	C_{F1}
Ground Truth	0.859	0.000	0.060	0.219	1.000	1.000	1.000	1.000	1.000	1.000
(a) no refinement	0.839$^{\pm 0.004}$	0.103$^{\pm 0.004}$	0.115$^{\pm 0.003}$	0.142$^{\pm 0.001}$	0.500$^{\pm 0.006}$	0.364$^{\pm 0.005}$	0.389$^{\pm 0.004}$	**0.849**$^{\pm 0.001}$	0.758$^{\pm 0.003}$	0.775$^{\pm 0.002}$
(b) w/o generation stream	**0.841**$^{\pm 0.003}$	0.101$^{\pm 0.005}$	0.091$^{\pm 0.001}$	0.155$^{\pm 0.000}$	0.496$^{\pm 0.003}$	0.388$^{\pm 0.002}$	0.404$^{\pm 0.001}$	0.847$^{\pm 0.002}$	0.788$^{\pm 0.003}$	0.794$^{\pm 0.002}$
(c) w/o corruption stream	**0.841**$^{\pm 0.003}$	0.104$^{\pm 0.003}$	0.085$^{\pm 0.003}$	0.166$^{\pm 0.001}$	0.496$^{\pm 0.005}$	0.414$^{\pm 0.004}$	0.419$^{\pm 0.002}$	0.847$^{\pm 0.005}$	0.786$^{\pm 0.004}$	0.793$^{\pm 0.004}$
(d) w/o short-range probes	0.840$^{\pm 0.003}$	0.100$^{\pm 0.004}$	0.083$^{\pm 0.002}$	0.186$^{\pm 0.000}$	0.496$^{\pm 0.002}$	0.462$^{\pm 0.001}$	0.446$^{\pm 0.000}$	0.845$^{\pm 0.002}$	0.828$^{\pm 0.003}$	0.816$^{\pm 0.002}$
(e) w/o long-range probes	0.840$^{\pm 0.003}$	0.101$^{\pm 0.004}$	0.111$^{\pm 0.003}$	0.137$^{\pm 0.000}$	0.500$^{\pm 0.007}$	0.351$^{\pm 0.003}$	0.379$^{\pm 0.004}$	**0.849**$^{\pm 0.007}$	0.750$^{\pm 0.003}$	0.770$^{\pm 0.003}$
(f) PamiRefiner	0.840$^{\pm 0.003}$	**0.092**$^{\pm 0.003}$	**0.062**$^{\pm 0.001}$	**0.193**$^{\pm 0.001}$	**0.502**$^{\pm 0.003}$	**0.477**$^{\pm 0.003}$	**0.457**$^{\pm 0.004}$	0.843$^{\pm 0.002}$	**0.836**$^{\pm 0.006}$	**0.821**$^{\pm 0.004}$

ations before refinement are reported in Appendix 7. Together, these evaluations examine whether the advantages of the proposed representation persist after geometric refinement.

Part-factored human encoder. We use a continuous part-based VAE that encodes each body part into a separate latent stream and decodes all jointly. Replacing the part encoders with a single full-body encoder, while keeping all else fixed, degrades all metrics (Tab. 2a). The benefit appears already in autoencoder where part factorization lowers reconstruction error by 60.8% (Tab. 6).

Part-anchored object motion representation. Our object representation encodes part-anchored trajectories in the object latent and combines them through learned voting. We ablate the design choices to understand the importance of anchors, learned weights, and the coordinate representation of global human and object root trajectories. The performance degrades gradually after removing non-wrist anchors (Tab. 2b) or all anchors (Tab. 2c), indicating the importance of our anchor representation and reasoning anchors beyond hands. Our model predicts dynamic per-frame voting weights to combine anchored predictions, rather than averaging them uniformly (Tab. 2d). This better models the local influence of each anchor on the object motion, leading to better hand and body contacts. We also include the absolute poses of the human and object roots in the representation, placing them in the same coordinate system instead of using the commonly adopted frame-relative root representation (Xiao et al., 2025). As shown in Tab. 2e, all metrics substantially degrade without absolute pose, as errors accumulate independently for the human and object, leading to divergent trajectories. See the supplementary website for qualitative examples.

4.2.2 Fine-grained refinement

Refinement quality. As shown in Rows (a) and (f) of Table 3, applying four PamiRefiner steps substantially improves the interaction geometry of the same coarse generations. It nearly halves penetration, brings Contact closer to the ground-truth distribution, and considerably improves both full-body and hand interaction F1. Meanwhile, R-1 remains stable and FID improves, supporting the role of PamiRefiner as a fine geometric generator that preserves the semantics of the generated interaction. See qualitative examples on the supplementary website.

Training with mixed data. We train PamiRefiner using controlled corruptions to learn ground-truth contact geometry and generation stream to improve geometry errors from the latent generator. Without either one, the performance degrades (Tab. 3b, c) hence the two streams are complementary.

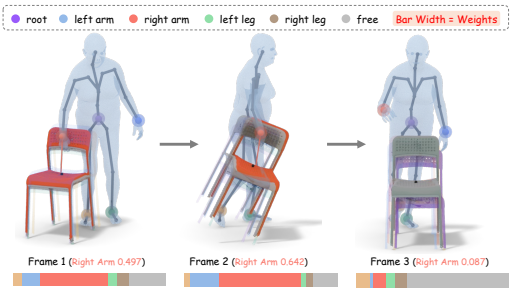


Figure 3: **Body-part anchor weights** strongly correlate with contact body parts.

Table 4: **Generalization of PamiRefiner** to LIGHT and InterAct. Full metrics are in Table 9.

	Pene$^{\downarrow}$	Contact$^{\rightarrow}$	F1$^{\uparrow}_{\text{body}}$	F1$^{\uparrow}_{\text{hand}}$
Ground Truth	0.060	0.219	1.000	1.000
LIGHT	0.134	0.128	0.328	0.704
+ w/o corr.	0.094	0.151	0.361	0.712
+ full	**0.069**	**0.181**	**0.408**	**0.755**
InterAct	0.126	0.122	0.278	0.658
+ w/o corr.	0.098	0.150	0.315	0.687
+ full	**0.077**	**0.188**	**0.374**	**0.738**

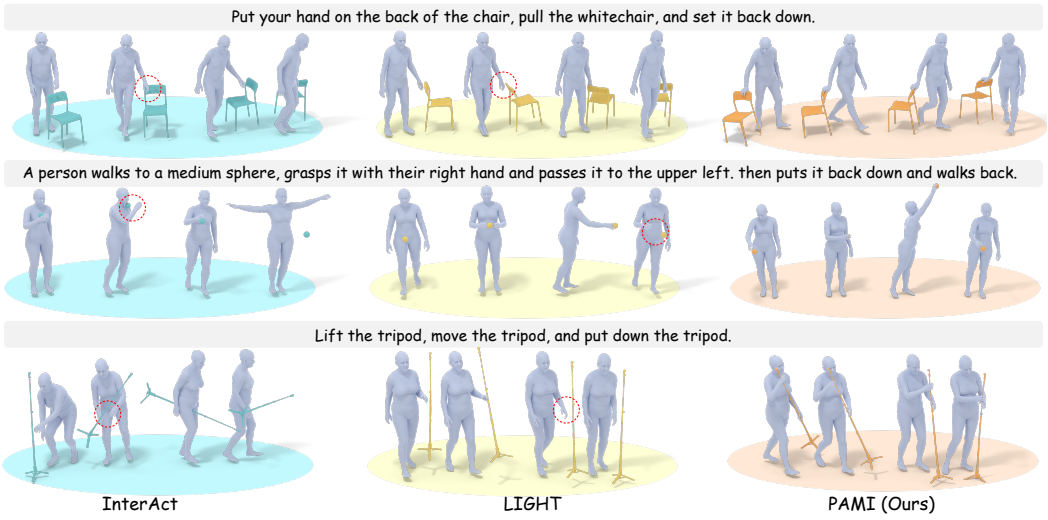


Figure 4: **Qualitative comparison.** Each row shows four frames from InterAct, LIGHT, and PAMI. InterAct and LIGHT often produce incorrect object orientation or interaction with missing contacts. PAMI better follows the described motion while maintaining coherent contact.

Hybrid surface sensors. We combine long-range probes for overall body–object relations with short-range probes for precise hand grasping. Tab. 3d, e shows that removing long-range probes hurts overall interaction quality, while removing short-range probes mainly degrades hand contacts.

Generalization across HOI generators. Our PamiRefiner can be used as a general fine interaction reasoner to improve interaction motions from other generators. We convert the outputs of InterAct and LIGHT into compatible input to our PamiRefiner via inverse kinematics and apply the same frozen PamiRefiner. As can be seen in Tab. 4, our model improves all baseline generations, showing its transferable capability to reason fine-grained interaction geometries.

5 Conclusion and Limitations

We presented PAMI for text-conditioned full-body HOI generation. Following the Hough-Transform idea, we represent object motion as votes from body-part anchors and learn per-frame weights that aggregate the relevant votes, and we generate the interaction in a compact part-based latent space. Generation is coarse-to-fine: PamiGen produces a coarse interaction in this latent space, and PamiRefiner then refines local interaction poses from hybrid surface geometry sensing. On InterAct, PAMI achives best performance in almost all metrics, achieving 14.5% higher contact recall than the previous art, and our ablations confirm that the part-based latent, anchor voting, learned weights, and PamiRefiner each contribute. PamiRefiner further improves LIGHT and InterAct outputs without retraining. Our code and models will be publicly released.

AI USE STATEMENT

We used large language models (LLMs) to assist with code implementation and to polish the writing of this paper. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

We will publicly release our full code, including training, inference, and evaluation scripts, to ensure that all results in this paper are fully reproducible.

REFERENCES

Dana H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

Songjin Cai, Linjie Zhong, Ling Guo, and Changxing Ding. Vihoi: Human-object interaction synthesis with visual priors. *arXiv preprint arXiv:2603.24383*, 2026.

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *CVPR*, 2024.

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 *CVPR*, 2023.

Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. *arXiv preprint arXiv:2403.17827*, 2024.

Yan Di, Yuheng Li, Yaoxing Wang, Mengge Liu, Shan Gao, and Xiangyang Ji. Pamotion: Physics-aware motion generation for full-body interaction with multiple objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30661–30674, 2026.

Christian Diller and Angela Dai. CG-HOI: Contact-guided 3d human-object interaction generation. In *CVPR*, 2024.

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

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. IMoS: Intent-driven full-body motion synthesis for human-object interactions. *arXiv preprint arXiv:2212.07555*, 2022.

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`.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *NeurIPS*, 2023.

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.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. GMD: Controllable human motion synthesis via guided diffusion models. In *ICCV*, 2023.

Jiye Lee, Yonghun Choi, and Jungdam Won. Mochi: Motion enhancement of collaborative human-object interactions. *ACM Transactions on Graphics (TOG)*, 45(4):1–18, 2026.

Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. Unimotion: Unifying 3d human motion synthesis and understanding. *arXiv preprint arXiv:2409.15904*, 2024.

Chuqiao Li, Xianghui Xie, Yong Cao, Andreas Geiger, and Gerard Pons-Moll. FrankenMotion: Part-level human motion generation and composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. *arXiv preprint arXiv:2312.03913*, 2023a.

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

Lei Li and Angela Dai. Hoi-page: Zero-shot human-object interaction generation with part affordance guidance. *arXiv preprint arXiv:2506.07209*, 2025.

Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 17416–17425, June 2025a.

Shujia Li, Haiyu Zhang, Xinyuan Chen, Yaohui Wang, and Yutong Ban. Genhoi: Generalizing text-driven 4d human-object interaction synthesis for unseen objects. *arXiv preprint arXiv:2506.15483*, 2025b.

Zhe Li, Jiakun Li, Mingqi Gao, Jinyu Yang, Wei Wang, and Feng Zheng. Conta-hoi: Towards physically plausible human-object interaction generation via contact-aware modeling. In *2026 International Conference on 3D Vision (3DV)*, pp. 946–955. IEEE, 2026b.

Yuhang Lin, Yijia Xie, Jiahong Xie, Yuehao Huang, Ruoyu Wang, Jiajun Lv, Yukai Ma, and Xingxing Zuo. Simgenhoi: Physically realistic whole-body humanoid-object interaction via generative modeling and reinforcement learning. *arXiv preprint arXiv:2508.14120*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics*, 2015.

Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arxiv:2310.12978*, 2023.

Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, et al. Himo: A new benchmark for full-body human interacting with multiple objects. In *European Conference on Computer Vision*, pp. 300–318. Springer, 2024.

Sihan Ma, Qiong Cao, Jing Zhang, and Dacheng Tao. Contact-aware human motion generation from textual descriptions. *arXiv preprint arXiv:2403.15709*, 2024.

Eric Nazarenus, Chuqiao Li, Yannan He, Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. ActionPlan: Future-aware streaming motion synthesis via frame-level action planning. *arXiv preprint*, 2026.

Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, pp. 598–613, 2020. URL `https://star.is.tue.mpg.de`.

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. *arXiv preprint arXiv:2312.06553*, 2023.

Chrysa Pratikaki, Pablo Ruiz-Ponce, Jiankang Deng, Stefanos Zafeiriou, and Rolandos Alexandros Potamias. Dex2hoi: Dexterous bimanual two-object interaction generation. *arXiv preprint arXiv:2605.30444*, 2026.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *ICCV*, 2019.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017.

Roey Ron, Guy Tevet, Haim Sawdayee, and Amit Bermano. Hoidini: Human-object interaction through diffusion noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5739–5749, 2026.

Mingyeong Song, Jungbin Cho, Jisoo Kim, Ananya Bal, Kartik Sharma, Youngjae Yu, Laszlo A Jeni, and Junhyug Noh. Jointhoi: Jointly generating contact maps enhances hand–object interaction generation. In *European Conference on Computer Vision*, pp. 97–114. Springer, 2026.

Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao Chen, Ning Li, and Hong Qin. Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *CVPR*, 2024.

Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J Black. Grip: Generating interaction poses using spatial cues and latent consistency. In *3DV*, 2024.

Hao Wang, Shiqi Wang, and Qi Liu. Mami-hoi: Harmonizing global kinematics and local geometry for human-object interaction generation. *arXiv preprint arXiv:2605.05756*, 2026a.

Ziyin Wang, Sirui Xu, Chuan Guo, Bing Zhou, Jiangshan Gong, Jian Wang, Yu-Xiong Wang, and Liang-Yan Gui. Unleashing guidance without classifiers for human-object interaction animation. In *ICLR*, 2026b.

Lin Wu, Zhixiang Chen, and Jianglin Lan. Hoi-dyn: Learning interaction dynamics for human-object motion diffusion. *NeurIPS*, 2025.

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

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 (ICCV)*, pp. 10086–10096, October 2025.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *ECCV*, 2022.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *CVPR*, 2023.

Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10003–10015, 2024.

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

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *arXiv preprint arXiv:2403.19652*, 2024.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. InterAct: Advancing large-scale versatile 3d human-object interaction generation. In *CVPR*, 2025a.

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 *CVPR*, 2025b.

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 *CVPR*, 2026.

Mengqing Xue, Yifei Liu, Ling Guo, Shaoli Huang, and Changxing Ding. Guiding human-object interactions with rich geometry and relations. *arXiv preprint arXiv:2503.20172*, 2025.

Pradyumna YM, Yuxuan Xue, Yue Chen, Nikita Kister, István Sárándi, and Gerard Pons-Moll. GRAFT: Geometric refinement and fitting transformer for human scene reconstruction. Project paper, 2026. URL `https://pradyumnaym.github.io/graft/`.

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. *arXiv preprint arXiv:2503.13130*, 2025.

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.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023.

Yixuan Zhang, Hui Yang, Chuanchen Luo, Junran Peng, Yuxi Wang, and Zhaoxiang Zhang. Oodhoi: Text-driven 3d whole-body human-object interactions generation beyond training domains. *arXiv preprint arXiv:2411.18660*, 2024b.

Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *ECCV*. Springer, 2022.

Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. GEARS: Local geometry-aware hand-object interaction synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

Yude Zou, Junji Gong, Xing Gao, Zixuan Li, Tianxing Chen, and Guanjie Zheng. Infbagel: Human-object-scene interaction generation with dynamic perception and iterative refinement. *arXiv preprint arXiv:2604.04843*, 2026.

Appendix

Table of Contents

A Additional experiments **14**
A.1 Number of refinement steps 14
A.2 Runtime and memory 14

B Implementation details **15**
B.1 Motion representation and coordinates 15
B.2 Anchor voting implementation 15
B.3 PamiGen: Text-Conditioned Latent Interaction Generation 16
B.4 PamiRefiner architecture and rollout 16
B.5 PamiRefiner training losses 17

C Limitations and future work **17**

A Additional experiments

Table 5 reports the retrieval and diversity metrics omitted from the main table.

Table 5: **Additional metrics on InterAct.** Top-2/3 R-precision and Diversity complement Table 1.

Method	R-Prec. Top 2$^{\uparrow}$	R-Prec. Top 3$^{\uparrow}$	Diversity$^{\rightarrow}$
Ground Truth	$0.967^{\pm 0.001}$	$0.992^{\pm 0.002}$	$7.764^{\pm 0.020}$
HOI-Diff	$0.881^{\pm 0.012}$	$0.929^{\pm 0.003}$	$7.577^{\pm 0.054}$
CHOIS	$0.880^{\pm 0.004}$	$0.934^{\pm 0.002}$	$7.777^{\pm 0.046}$
InterDiff	$0.930^{\pm 0.002}$	$0.968^{\pm 0.005}$	$7.738^{\pm 0.058}$
Text2HOI	$0.869^{\pm 0.005}$	$0.926^{\pm 0.000}$	$7.730^{\pm 0.009}$
LIGHT	$0.877^{\pm 0.007}$	$0.930^{\pm 0.001}$	$7.737^{\pm 0.062}$
InterAct	$0.935^{\pm 0.007}$	$0.967^{\pm 0.006}$	$\mathbf{7.775}^{\pm 0.026}$
PAMI	$\mathbf{0.958}^{\pm 0.002}$	$\mathbf{0.980}^{\pm 0.001}$	$7.693^{\pm 0.007}$

A.1 Number of refinement steps

We vary the number of recursive updates at inference to select the refinement depth.

Most contact improvement occurs within four updates. Further refinement gives little contact gain and eventually worsens FID. We therefore use $K{=}4$ throughout.

A.2 Runtime and memory

PamiVAE, PamiGen, and PamiRefiner have 755M, 55M, and 4.4M parameters, respectively. On one H100, PamiGen trains for 240k iterations in about 16 hours and PamiRefiner for 44k iterations in about 8 hours. For a 231-frame sequence at batch size 1, generation takes 0.13 s and 5.7 GB; four refinement steps add 0.66 s and use 2.8 GB.

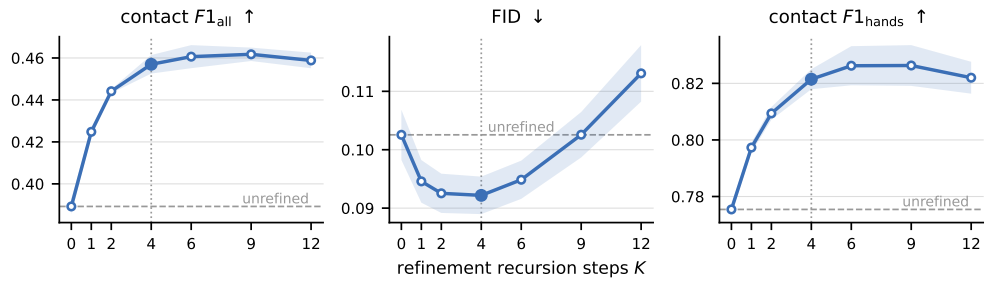


Figure 5: **Effect of recursion depth K.** From left to right: full-body contact F1, FID, and hand-contact F1. Curves show the mean over three generation draws; bands denote 95% confidence intervals.

Table 6: **PamiVAE reconstruction.** Validation errors from the final checkpoint of each variant. Body and hand errors are in mm; object translation and rotation errors are in cm and degrees.

Variant	MPJPE (mm)↓	hand (mm)↓	object comp. (cm)↓	object pos. (cm)↓	object rot. (°)↓
part-factored (ours)	15.2	1.3	2.54	1.66	2.2
+ contact-tuned decoder (final)	16.7	1.3	2.49	1.26	1.6
whole-body, param.-matched	38.8	2.3	2.70	1.38	2.0
wrist-only anchors	16.5	1.3	2.59	1.02	2.2
no anchors	17.7	1.2	–	1.41	1.2
w/o spatiotemporal attention	15.4	1.2	2.94	2.10	2.5

B Implementation Details

B.1 Motion Representation and Coordinates

Motion is processed at 30 fps in windows of at most 300 frames. For each sequence, the human root in the first frame defines the world coordinate frame, and all subsequent global human and object motion is expressed relative to this frame. We follow MotionStreamer (Xiao et al., 2025) for the 272-dimensional body representation, but replace its frame-relative root variables with global root orientation and translation in this first-frame coordinate system. The remaining features contain root-relative joint positions and velocities, and parent-relative 6D rotations. The two hands are represented separately by wrist-local positions and parent-relative 6D rotations for 15 finger joints per side.

PamiVAE uses eight causal body-part encoders for the root, torso, left/right arms, legs, and hands. The object encoder receives its global 6D rotation, free translation, and anchor-relative offsets. All encoders downsample time by four and produce 16-dimensional tokens. Six spatiotemporal-attention layers exchange information across tokens and time. One joint human decoder fuses all body-part tokens, and two upsampling stages restore the original frame rate. Table 6 compares VAE reconstruction quality, while Table 7 reports the corresponding generators before refinement.

B.2 Anchor Voting Implementation

The representation uses five body-part anchors—the root, left/right wrists, and left/right ankles—and one free anchor. Each body-part vote transforms a decoded relative offset through its decoded human anchor; the free vote predicts object translation directly.

The voting target $\widetilde{\mathbf{w}}_t$ assigns weight to body parts within 1 cm of the object. Multiple active votes are normalized; when no body-part contact is detected, the free branch receives the full vote. We use $\mathcal{L}_{\mathrm{w}} = \mathrm{MSE}(\mathbf{w}_t, \widetilde{\mathbf{w}}_t)$ with weight 0.01. At inference, the decoder predicts the votes and softmax weights without the target. Object orientation is decoded separately in 6D. Table 8 reports the additional metrics for the refined representation ablations.

Table 7: **Generator ablations before refinement** ($K=0$). These coarse outputs isolate the representation choices from geometric refinement. We report the mean and 95% confidence interval over three generation draws.

	Top-1$^{\uparrow}$	Top-2$^{\uparrow}$	Top-3$^{\uparrow}$	FID$^{\downarrow}$	MM Dist$^{\downarrow}$	Diversity$^{\rightarrow}$	FSR$^{\downarrow}$
Ground Truth	0.859	0.967	0.992	0.000	1.475	7.764	0.051
Our full PAMI model	$0.839^{\pm 0.004}$	$0.957^{\pm 0.002}$	$0.980^{\pm 0.002}$	$\mathbf{0.103}^{\pm 0.004}$	$\mathbf{2.198}^{\pm 0.015}$	$7.683^{\pm 0.008}$	$0.137^{\pm 0.002}$
(a) w/o part factorization	$0.816^{\pm 0.002}$	$0.946^{\pm 0.001}$	$0.972^{\pm 0.002}$	$0.330^{\pm 0.009}$	$2.375^{\pm 0.004}$	$7.535^{\pm 0.016}$	$0.262^{\pm 0.001}$
(b) w/o leg and root anchors	$\mathbf{0.840}^{\pm 0.001}$	$\mathbf{0.960}^{\pm 0.003}$	$\mathbf{0.984}^{\pm 0.001}$	$0.109^{\pm 0.006}$	$2.215^{\pm 0.014}$	$7.706^{\pm 0.006}$	$0.137^{\pm 0.002}$
(c) w/o anchors	$0.832^{\pm 0.004}$	$0.953^{\pm 0.003}$	$0.977^{\pm 0.001}$	$0.119^{\pm 0.007}$	$2.230^{\pm 0.004}$	$\mathbf{7.723}^{\pm 0.013}$	$\mathbf{0.128}^{\pm 0.002}$
(d) w/o learned voting weights	$0.830^{\pm 0.003}$	$0.950^{\pm 0.001}$	$0.974^{\pm 0.001}$	$0.116^{\pm 0.006}$	$2.237^{\pm 0.007}$	$7.689^{\pm 0.017}$	$0.135^{\pm 0.001}$
(e) w/o absolute root	$0.775^{\pm 0.002}$	$0.909^{\pm 0.006}$	$0.948^{\pm 0.005}$	$0.368^{\pm 0.026}$	$2.684^{\pm 0.042}$	$7.418^{\pm 0.015}$	$0.147^{\pm 0.001}$

	Pene$^{\downarrow}$	Contact$^{\rightarrow}$	Interaction (full body)$^{\uparrow}$			Interaction (hands)$^{\uparrow}$		
			C_{prec}	C_{rec}	C_{F1}	C_{prec}	C_{rec}	C_{F1}
Ground Truth	0.060	0.219	1.000	1.000	1.000	1.000	1.000	1.000
Our full PAMI model	$0.115^{\pm 0.003}$	$\mathbf{0.142}^{\pm 0.001}$	$\mathbf{0.500}^{\pm 0.006}$	$\mathbf{0.364}^{\pm 0.005}$	$\mathbf{0.389}^{\pm 0.004}$	$\mathbf{0.849}^{\pm 0.001}$	$\mathbf{0.758}^{\pm 0.003}$	$\mathbf{0.775}^{\pm 0.002}$
(a) w/o part factorization	$\mathbf{0.083}^{\pm 0.002}$	$0.069^{\pm 0.001}$	$0.434^{\pm 0.005}$	$0.151^{\pm 0.004}$	$0.199^{\pm 0.003}$	$0.813^{\pm 0.002}$	$0.447^{\pm 0.001}$	$0.532^{\pm 0.002}$
(b) w/o leg and root anchors	$0.115^{\pm 0.003}$	$0.140^{\pm 0.002}$	$0.489^{\pm 0.003}$	$0.353^{\pm 0.005}$	$0.375^{\pm 0.004}$	$0.848^{\pm 0.003}$	$0.754^{\pm 0.004}$	$0.773^{\pm 0.003}$
(c) w/o anchors	$0.100^{\pm 0.006}$	$0.091^{\pm 0.001}$	$0.451^{\pm 0.012}$	$0.210^{\pm 0.006}$	$0.254^{\pm 0.009}$	$0.828^{\pm 0.005}$	$0.575^{\pm 0.005}$	$0.633^{\pm 0.003}$
(d) w/o learned voting weights	$0.115^{\pm 0.004}$	$0.132^{\pm 0.001}$	$0.485^{\pm 0.003}$	$0.316^{\pm 0.003}$	$0.350^{\pm 0.001}$	$0.848^{\pm 0.000}$	$0.725^{\pm 0.008}$	$0.754^{\pm 0.006}$
(e) w/o absolute root	$0.112^{\pm 0.002}$	$0.115^{\pm 0.002}$	$0.469^{\pm 0.003}$	$0.290^{\pm 0.002}$	$0.323^{\pm 0.004}$	$0.826^{\pm 0.007}$	$0.660^{\pm 0.003}$	$0.694^{\pm 0.007}$

Table 8: **Additional metrics for the refined representation ablations in Table 2** ($K=4$). Mean and 95% confidence interval over three runs.

	Top-2$^{\uparrow}$	Top-3$^{\uparrow}$	MM Dist$^{\downarrow}$	Diversity$^{\rightarrow}$	FSR$^{\downarrow}$
Ground Truth	0.967	0.992	1.475	7.764	0.051
Our full PAMI model	$0.958^{\pm 0.002}$	$0.980^{\pm 0.001}$	$\mathbf{2.195}^{\pm 0.016}$	$7.693^{\pm 0.007}$	$\mathbf{0.078}^{\pm 0.001}$
(a) w/o part factorization	$0.948^{\pm 0.000}$	$0.973^{\pm 0.001}$	$2.308^{\pm 0.003}$	$7.651^{\pm 0.019}$	$0.115^{\pm 0.001}$
(b) w/o leg and root anchors	$\mathbf{0.960}^{\pm 0.002}$	$\mathbf{0.985}^{\pm 0.001}$	$2.207^{\pm 0.013}$	$7.718^{\pm 0.004}$	$0.112^{\pm 0.001}$
(c) w/o anchors	$0.953^{\pm 0.002}$	$0.976^{\pm 0.001}$	$2.218^{\pm 0.003}$	$\mathbf{7.727}^{\pm 0.010}$	$0.092^{\pm 0.001}$
(d) w/o learned voting weights	$0.950^{\pm 0.002}$	$0.974^{\pm 0.001}$	$2.228^{\pm 0.006}$	$7.695^{\pm 0.014}$	$0.103^{\pm 0.001}$
(e) w/o absolute root	$0.916^{\pm 0.008}$	$0.951^{\pm 0.009}$	$2.705^{\pm 0.038}$	$7.384^{\pm 0.026}$	$0.096^{\pm 0.001}$

B.3 PamiGen: Text-Conditioned Latent Interaction Generation

The latent grid contains eight human-part tokens and one object token per latent step. The transformer has 10 layers, width 512, eight attention heads, and feed-forward width 2048. It applies self-attention to the full token grid and cross-attention to frozen caption features. Rotary embeddings encode time, while token embeddings distinguish body parts and the object. The object token is additionally conditioned on basis-point-set features of the canonical mesh (Prokudin et al., 2019). All normalization statistics are computed on the training split; human statistics are computed separately for each part token. To prevent the object loss from being diluted by the larger number of human-part tokens, we compute the human and object errors separately:

$$\mathcal{L}_{\mathrm{RF}} = \mathbb{E}_{\tau,\boldsymbol{\epsilon}}\left[\left\|\mathbf{v}_\theta^h - \mathbf{v}^{\star,h}\right\|_2^2 + \lambda_o \left\|\mathbf{v}_\theta^o - \mathbf{v}^{\star,o}\right\|_2^2\right]. \tag{5}$$

We use classifier-free guidance with scale 2.5 and 50 integration steps. Decoded losses use Huber distance and are applied only for $0.1 < \tau < 0.5$. The free-position, free-velocity, and composed translation terms each have weight 5.0; alignment has weight 2.5. Alignment is evaluated only on frames with an active body-part vote, after removing the predicted free weight and renormalizing the remaining weights. Targets come from decoding the clean latent with the same VAE.

B.4 PamiRefiner architecture and rollout

For one object, PamiRefiner uses 54 tokens per frame: 22 body joints, 30 finger joints, one object, and one global token. Joint tokens contain 6D rotation and long-range surface probes. Body probes

Table 9: **Full metrics for Table 4.** Zero-shot refinement of LIGHT and InterAct outputs with $K=4$; mean and 95% confidence interval over two sampling campaigns. Bold marks the best result in each generator block.

	Top-1↑	Top-2↑	Top-3↑	FID↓	MM Dist↓	Diversity→	FSR↓	Pene↓	Contact→	Interaction (full body)↑			Interaction (hands)↑		
										C_{proc}	C_{rec}	C_{F1}	C_{proc}	C_{rec}	C_{F1}
Ground Truth	0.859	0.967	0.992	0.000	1.475	7.764	0.051	0.060	0.219	1.000	1.000	1.000	1.000	1.000	1.000
LIGHT	$0.763^{\pm0.013}$	$0.911^{\pm0.009}$	$\mathbf{0.948}^{\pm0.004}$	$\mathbf{0.225}^{\pm0.073}$	$2.643^{\pm0.013}$	$7.593^{\pm0.012}$	$0.105^{\pm0.003}$	$0.134^{\pm0.005}$	$0.128^{\pm0.000}$	$0.468^{\pm0.001}$	$0.308^{\pm0.004}$	$0.328^{\pm0.000}$	$\mathbf{0.808}^{\pm0.003}$	$0.686^{\pm0.013}$	$0.704^{\pm0.010}$
LIGHT + PamiRefiner w/o corruption stream	$0.762^{\pm0.009}$	$0.910^{\pm0.009}$	$\mathbf{0.948}^{\pm0.002}$	$0.241^{\pm0.068}$	$2.644^{\pm0.011}$	$\mathbf{7.606}^{\pm0.022}$	$0.061^{\pm0.001}$	$0.094^{\pm0.003}$	$0.151^{\pm0.000}$	$0.459^{\pm0.001}$	$0.360^{\pm0.006}$	$0.361^{\pm0.001}$	$0.795^{\pm0.002}$	$0.704^{\pm0.011}$	$0.712^{\pm0.008}$
LIGHT + PamiRefiner	$\mathbf{0.765}^{\pm0.010}$	$\mathbf{0.912}^{\pm0.010}$	$\mathbf{0.948}^{\pm0.006}$	$0.235^{\pm0.076}$	$\mathbf{2.642}^{\pm0.009}$	$7.603^{\pm0.016}$	$\mathbf{0.059}^{\pm0.002}$	$\mathbf{0.069}^{\pm0.003}$	$\mathbf{0.181}^{\pm0.000}$	$\mathbf{0.476}^{\pm0.000}$	$\mathbf{0.429}^{\pm0.001}$	$\mathbf{0.408}^{\pm0.002}$	$\mathbf{0.808}^{\pm0.003}$	$\mathbf{0.765}^{\pm0.009}$	$\mathbf{0.755}^{\pm0.006}$
InterAct	$0.803^{\pm0.010}$	$0.928^{\pm0.004}$	$0.962^{\pm0.005}$	$0.328^{\pm0.013}$	$2.524^{\pm0.037}$	$\mathbf{7.827}^{\pm0.047}$	$0.101^{\pm0.002}$	$0.126^{\pm0.005}$	$0.122^{\pm0.001}$	$0.418^{\pm0.007}$	$0.249^{\pm0.005}$	$0.278^{\pm0.001}$	$\mathbf{0.777}^{\pm0.003}$	$0.631^{\pm0.000}$	$0.658^{\pm0.002}$
InterAct + PamiRefiner w/o corruption stream	$0.806^{\pm0.005}$	$0.929^{\pm0.004}$	$\mathbf{0.964}^{\pm0.004}$	$0.320^{\pm0.019}$	$2.507^{\pm0.039}$	$7.836^{\pm0.050}$	$0.076^{\pm0.004}$	$0.098^{\pm0.001}$	$0.150^{\pm0.001}$	$0.421^{\pm0.003}$	$0.303^{\pm0.002}$	$0.315^{\pm0.001}$	$\mathbf{0.777}^{\pm0.001}$	$0.670^{\pm0.000}$	$0.687^{\pm0.001}$
InterAct + PamiRefiner	$\mathbf{0.808}^{\pm0.006}$	$\mathbf{0.930}^{\pm0.005}$	$0.963^{\pm0.006}$	$\mathbf{0.300}^{\pm0.023}$	$\mathbf{2.504}^{\pm0.040}$	$7.842^{\pm0.053}$	$\mathbf{0.058}^{\pm0.001}$	$\mathbf{0.077}^{\pm0.001}$	$\mathbf{0.188}^{\pm0.000}$	$\mathbf{0.430}^{\pm0.001}$	$\mathbf{0.389}^{\pm0.004}$	$\mathbf{0.374}^{\pm0.003}$	$\mathbf{0.777}^{\pm0.004}$	$\mathbf{0.749}^{\pm0.001}$	$\mathbf{0.738}^{\pm0.001}$

use the root frame, while finger probes use the wrist frame. The object token contains object pose; the global token contains root pose and a 16D body-shape feature. Long-range probes use 512 surface points.

Short-range probes use a separate 16,384-point area-weighted surface sample. For each finger joint and fingertip, we encode at most 64 points within 4 cm using a shared pointwise network and masked-mean pooling. Empty neighborhoods produce zero features. Both probe types are recomputed after every refinement step.

Features are projected to width 256 and processed by four spatial–temporal transformer blocks. Zero-initialized heads predict bounded updates to joint rotations, root motion, and object $\mathrm{SE}(3)$. Training backpropagates through a three-step rollout. Object updates are capped at 5 cm and 10°; root updates at 10 cm and 5° in yaw. Table 9 reports the complete cross-generator transfer metrics.

B.5 PamiRefiner training losses

Recovery and generation batches alternate 1:1 while the VAE and generator remain frozen. A recovery input is a corrupted VAE round trip, supervised by the original interaction. We sample clean, root, torso, upper-body, lower-body, object, and mixed groups with probabilities $0.2/0.1/0.1/0.2/0.1/0.2/0.1$; mixed corruption combines two or three groups.

Temporal noise follows an Ornstein–Uhlenbeck process with $\alpha \approx 0.895$. Root noise is 4.7 cm in xz, 3° in yaw, and a 4 cm height offset. Rotation ranges are 2–5° for the torso, 3–7° for hips/knees, 2–5° for ankles, 3–12° for the upper body, and 4–15° for fingers. Object noise uses 1–4 cm translation and 3–12° rotation. Persistent biases are 1–6° for the upper body, 2–8° for fingers, and 1–5 cm/3–15° for the object. The corruption distribution is fixed across the three training steps; generated samples are precomputed.

The implementation terms follow the five groups in Eq. 4. $\mathcal{L}_{\mathrm{contact}}$ contains surface attraction and contact slip; $\mathcal{L}_{\mathrm{penetration}}$ penalizes human–object intersection; $\mathcal{L}_{\mathrm{support}}$ constrains planted feet, ground penetration, and lower-body drift; $\mathcal{L}_{\mathrm{smooth}}$ penalizes abrupt motion and updates; and $\mathcal{L}_{\mathrm{preserve}}$ protects non-interacting body regions, the root, and the object. The five group weights $\lambda_c, \lambda_p, \lambda_s, \lambda_m, \lambda_n$ are all 1; their internal coefficients and the separate recovery-stream weights are reported in Table 11. Contact masks and segment anchors are computed once from the coarse generation and kept fixed during recursion. Table 10 reports the additional metrics for the RefineNet ablations.

Parameter counts, optimizer schedules, and the scalar loss weights are listed in Table 11, read directly from the final checkpoints' recorded configurations.

C Limitations and future work

Our experiments focus on the single-human, single-rigid-object setting of InterAct, and the method assumes access to a known canonical object mesh for shape conditioning and surface sensing, which may limit its applicability when accurate object geometry is unavailable. Moreover, PamiRefiner is designed as a bounded local corrector: it can improve contact when the coarse interaction is already reasonable, but cannot recover from an incorrect action, severe temporal misalignment, or a large trajectory error. Future work will extend the framework to interactions involving multiple objects and improve its generalization to novel object shapes beyond those observed during training.

Table 10: **Additional metrics for the PamiRefiner ablations in Table 3** (K=4). Mean and 95% confidence interval over three runs.

	Top-2$^{\uparrow}$	Top-3$^{\uparrow}$	MM Dist$^{\downarrow}$	Diversity$^{\rightarrow}$	FSR$^{\downarrow}$
Ground Truth	0.967	0.992	1.475	7.764	0.051
(a) no refinement	$0.957^{\pm 0.002}$	$\mathbf{0.980}^{\pm 0.002}$	$2.198^{\pm 0.015}$	$7.683^{\pm 0.008}$	$0.137^{\pm 0.002}$
(b) w/o generation stream	$\mathbf{0.958}^{\pm 0.002}$	$\mathbf{0.980}^{\pm 0.002}$	$2.208^{\pm 0.015}$	$\mathbf{7.696}^{\pm 0.009}$	$0.134^{\pm 0.001}$
(c) w/o corruption stream	$\mathbf{0.958}^{\pm 0.002}$	$\mathbf{0.980}^{\pm 0.003}$	$\mathbf{2.190}^{\pm 0.014}$	$7.692^{\pm 0.008}$	$0.085^{\pm 0.000}$
(d) w/o short-range probes	$\mathbf{0.958}^{\pm 0.002}$	$\mathbf{0.980}^{\pm 0.002}$	$2.198^{\pm 0.014}$	$7.691^{\pm 0.006}$	$0.107^{\pm 0.001}$
(e) w/o long-range probes	$\mathbf{0.958}^{\pm 0.003}$	$\mathbf{0.980}^{\pm 0.003}$	$2.202^{\pm 0.016}$	$7.683^{\pm 0.007}$	$0.105^{\pm 0.002}$
(f) PamiRefiner	$\mathbf{0.958}^{\pm 0.002}$	$\mathbf{0.980}^{\pm 0.001}$	$2.195^{\pm 0.016}$	$7.693^{\pm 0.007}$	$\mathbf{0.078}^{\pm 0.001}$

Table 11: Hyperparameters used by the final checkpoints.

PamiVAE	
Human part tokens / object tokens	8 / 1
Latent dim per token	16
Part encoder / decoder width	1024 / 1024
Spatiotemporal attention	$d = 256$, 6 layers, 8 spatial / 14 temporal heads
Anchor set	root, left/right wrist, left/right ankle + free slot
Base loss weights (direct recon. / free vel. / composed trans. / voting / human KL / object KL)	1.0 / 1.0 / 1.0 / 0.01 / 10^{-4} / 10^{-4}
Auxiliary loss weights (hand / root vel. / foot slide / joint vel. / FK cons. / contact)	1.0 / 50.0 / 30.0 / 30.0 / 1.0 / 1.0
Training iterations	200k; contact-tuned decoder +25k (encoder frozen, $\lambda_{\mathrm{ctc}} = 1.0$)
Parameters	755M
PamiGen	
Generator backbone	$d = 512$, 10 layers, 8 heads, FF 2048, dropout 0.1
Conditioning	text (CLIP ViT-B/32 tokens), object BPS, body-shape joint rest pose
Classifier-free guidance (train drop / test scale) / sampling steps	0.1 / 2.5 / 50
Optimizer	AdamW, lr 0.0001, warmup 2500, cosine to 1e-06 over 350k
Decoded-space losses (free pos. / free vel. / composed / alignment)	5.0 / 5.0 / 5.0 / 2.5
Training iterations	240k
Parameters	55M
PamiRefiner	
Backbone	$d = 256$, 4 layers
Short-range surface sensor	up to 64 points per query, radius 4 cm, dense surface 16,384
Streams	corruption : generation = 1 : 1 (real_every=2)
Recursion depth K (train / test)	3 / 4
Recovery weights (absolute / centered vertices / rotations / object / root xz / root y)	7.0 / 5.0 / 1.0 / 1.0 / 1.0 / 1.0
Generation group weights ($\lambda_c/\lambda_p/\lambda_s/\lambda_m/\lambda_n$)	1.0 / 1.0 / 1.0 / 1.0 / 1.0
Contact and penetration terms (attraction / slip / penetration)	1.0 / 0.5 / 1.0
Support terms (foot / pin and drift / leg trust)	2.0 / 5.0 / 1.0
Smoothness terms (motion / update)	0.1 / 0.5
Preservation terms (body / object / root)	1.0 / 1.0 / 1.0
Correction authority (object)	5.0 cm, 10.0°
Optimizer	AdamW, lr 0.0003, batch 4, warmup 12k, 60k schedule (checkpoint at 44k)
Parameters	4.4M