EgoHMP: Achieving Precise Human Motion Prediction in 3D Scenes via Egocentric Cues

Xin Zhao[1], **Pengzhan Zhou**[1]*, **Ziyi Li**[1], **Riheng Jia**[2], **Zhida Qin**[3], **Yu Liu**[4]
[1]College Of Computer Science, Chongqing University
[2]School of Computer Science and Technology, Zhejiang Normal University
[3]School of Computer Science and Technology, Beijing Institute of Technology
[4]Department of Computing, Hong Kong Polytechnic University
xzhao@stu.cqu.edu.cn, pzzhou@cqu.edu.cn, shanksrlrobot@stu.cqu.edu.cn
rihengjia@zjnu.edu.cn, zanderqin@bit.edu.cn, yu-y.liu@polyu.edu.hk

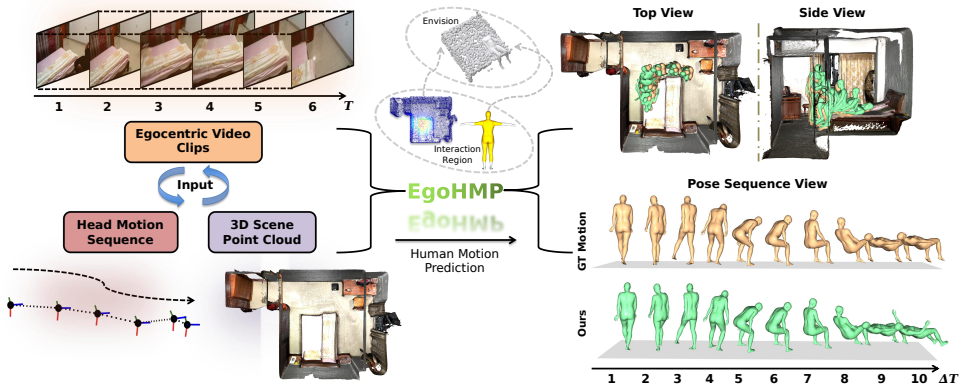


Figure 1: EgoHMP captures human intent through synergistic egocentric observations and head motions, and reveals interaction concepts by combining the scene and human interaction regions, thereby envisioning the interaction, which enables precise human motion prediction in 3D scenes.

Abstract

The complex interactive dependencies between human behavior and 3D scenes make accurately predicting human intent a major challenge for human motion prediction (HMP). While recent studies use human gaze to infer intent, gaze signals suffer from ambiguity, drift, and a reliance on eye-tracking hardware. Research in cognitive science indicates that egocentric views inherently encode potential interactions, while head motion serves as a crucial precursor in motion planning. Motivated by this, we propose EgoHMP, a novel framework that uses egocentric views and head motions as robust carriers of interaction intent. By translating this intent into an envisioning of future interactions, EgoHMP achieves precise HMP in 3D scenes. EgoHMP first employs a vision-motion alignment encoder to align the visual and motion feature spaces. Based on these aligned features, an egocentric intent predictor with a modality-aware modulation mechanism adaptively suppresses invalid interaction cues to predict an accurate interaction probability map. Furthermore, to enhance the stability and realism of the generated motions, we propose a contact-aware motion decoder. Specifically, inspired by the human cognitive process of planning trajectories prior to action execution, we first predict a global trajectory to serve as a self-prompt, effectively mitigating the optimization conflict

*The corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

in diffusion models. Subsequently, we introduce a multi-stage denoiser, which relieves motion artifacts by incorporating human contact modeling through a multi-stage optimization strategy. Finally, a GCN-based motion decoder is employed to synthesize physically plausible and semantically consistent motions. Extensive experiments demonstrate that EgoHMP achieves state-of-the-art performance. Our code is available at `https://github.com/zhaoxinv587/EgoHMP`.

1 Introduction

Human motion prediction (HMP) aims to infer future motion sequences based on historical observations [1–4]. It holds significant value in applications such as augmented and virtual reality (AR/VR) [5–7], human-computer interaction[8, 9], and embodied AI [10–13]. Unlike motion prediction in unconstrained spaces[14–16], realistic human behaviors in 3D scenes are not isolated motion sequences; rather, they are highly intent-driven and tightly coupled with the environment. Consequently, to accurately predict motions, a model must reliably infer the underlying interaction intent. However, precisely capturing this interactive intent remains a formidable challenge.

To tackle this challenge, recent studies [5, 7, 17] have attempted to use human gaze as a proxy for intent. While intuitively appealing, this approach relies on the strict premise that individuals look exclusively at their intended targets prior to interaction. In reality, the inherent drift and ambiguity of gaze signals frequently mislead models. Because distinct gaze points can correspond to entirely different motion directions and interaction targets, this ambiguity easily results in unreliable motion predictions[7]. Furthermore, acquiring high-quality 3D gaze data necessitates specialized eye-tracking devices. Consequently, these critical limitations severely restrict the predictive performance of such methods and impede their deployment in real-world scenarios.

Research in cognitive science shows that humans synthesize egocentric behaviors by coordinating the visual cortex, cerebellum, and cerebrum. This coordination intrinsically bridges visual observations, ego-motion, and conceptual understanding, thereby connecting the physical body to its surrounding environment and unveiling latent interaction intents [18–20]. Concurrently, previous studies demonstrate that head motion serves as a crucial kinematic precursor in motor planning, consistently anticipating and guiding subsequent full-body execution [14, 21, 22]. Motivated by these insights, we argue that the egocentric view, coupled with its corresponding head motion sequence, acts as a more robust and accessible medium for human interaction intent than gaze signals. Specifically, the egocentric view provides rich semantic context regarding the user's focus, while head motion supplies the essential 3D spatial cues to anchor these 2D semantics into the physical world [14, 22, 23].

To instantiate this insight, we propose EgoHMP, a novel framework that achieves precise human motion prediction in 3D scenes using egocentric videos, head motion sequences, and 3D scene point clouds (Figure 1). Specifically, EgoHMP first employs a vision-motion alignment encoder to bridge the semantic gap between visual observations and head motions. To enable the model to extract effective interaction contexts from diverse interaction cues, we propose a modality-aware modulation mechanism. This mechanism manipulates the backpropagated gradient flow, enabling the network to adaptively suppress invalid interaction cues and distill a reliable global intent token. Subsequently, an egocentric intent predictor uses scene point features to query this intent representation, predicting an accurate interaction probability map. To enhance the stability and realism of the generated motions, we propose a contact-aware motion decoder. Specifically, guided by the interaction probability map and inspired by the human cognitive process of planning paths prior to action execution, we first infer a global trajectory to serve as a self-prompt. By anchoring the diffusion process with this trajectory, the model is freed to focus on fine-grained pose synthesis. Subsequently, conditioned on multi-modal representations and local scene contexts, we introduce a multi-stage denoiser. Inspired by the insight that human contact learning is the key to bridging perception and action for envisioning interactive poses [22], this denoiser progressively optimizes the generated poses by explicitly learning human contact regions, thereby generating semantically consistent and physically plausible motions.

In summary, our main contributions can be summarized as follows:

- We propose a new perspective for human motion prediction in 3D scenes, demonstrating that integrating egocentric views with head motion sequences serves as a robust proxy for inferring human interaction intent, without relying on gaze information.

- We introduce EgoHMP, a novel framework comprising three core components: a vision-motion alignment encoder to align 2D vision and 3D head motion features, an egocentric intent predictor to precisely ground interaction intents, and a contact-aware motion decoder to synthesize context-consistent and physically plausible motions.
- Extensive experiments on the GIMO [5], GTA-IM [24], and Ego-Exo4D [25] benchmarks demonstrate that EgoHMP achieves state-of-the-art performance in both pose and trajectory prediction accuracy, effectively alleviating geometric artifacts such as floating and penetration.

2 Related Work

Human Motion Prediction in 3D Scenes. Recent methods incorporate elements such as 3D scene geometry features, human gaze information, and object affordance into the prediction framework, enhancing prediction accuracy by integrating external context [5, 7, 8, 17, 26, 27]. SIF3D [17] uses human gaze as a motion intention and models the motion-constrained space using salient point clouds derived from 3D scenes. However, it overlooks humans' subjective intention to interact with the environment, leading to predictions that lack targeted interaction behaviors and fail to accurately infer human-object or human-computer interaction actions. Subsequent works [7, 11] further incorporate object affordance, inferring interaction gestures by leveraging objects' functional attributes in conjunction with human motion history. GAP3DS [7] constructs an object-action mapping using predefined affordance labels to enhance the semantic rationality of predicted interactions. However, the above methods neglect the inherent drift and ambiguity of gaze information, rendering them prone to trajectory-prediction biases and misidentification of the intended interactive object. In contrast, EgoHMP circumvents the reliance on gaze signals. By coupling egocentric views with head motion sequences, we provide a more robust proxy for human interaction intent in 3D scenes.

Egocentric Intent Understanding. The egocentric perspective inherently captures the user's visual attention and functional focus, making it a powerful tool for intent understanding. Recently, egocentric intent understanding has achieved remarkable progress [14, 22, 28–32]. However, a critical limitation of these approaches is their lack of explicit 3D scene perception. They primarily perceive intents within 2D visual spaces, forcing downstream tasks—such as motion prediction—to operate exclusively within the local camera coordinate system, leaving them completely oblivious to the global physical world. Consequently, predicting physically plausible and collision-free interactions in real-world 3D environments remains highly challenging for these methods. EgoHMP elegantly bridges this gap. By seamlessly integrating egocentric signals with 3D scene point clouds, EgoHMP successfully grounds 2D egocentric intents into 3D scenes.

Diffusion Models for Motion Generation. Recently, denoising diffusion probabilistic models [7, 11, 33–35] have emerged as the dominant paradigm for human motion generation due to their exceptional generative diversity and fidelity. However, we observe that forcing a single diffusion model to simultaneously regress global locomotion and local joint articulation introduces an optimization conflict. Furthermore, while most existing methods incorporate object affordances to improve physical plausibility, their success depends on the accuracy of scene segmentation. More importantly, they largely overlook the human-centric contact. To address these limitations, we first propose a self-prompting strategy. By predicting an intent-aligned global trajectory as prompt to pose synthesis, we relieve the optimization conflict. Second, EgoHMP employs a denoiser driven by a multi-stage optimization strategy. It leverages egocentric vision-motion cues to model human-centric contact, utilizing these contact priors to refine motions. This effectively resolves physical artifacts, empowering EgoHMP to generate context-consistent and physically plausible motions.

3 Method

Figure 2 shows EgoHMP's overall architecture. EgoHMP accurately infers human intent and generates physically plausible, semantically consistent human motion predictions.

3.1 Problem Setup

Given an egocentric video clip $V_{1:T} \in \mathbb{R}^{T\times H\times W\times 3}$, its corresponding head motion sequence $H_{1:T} = \{(t_1^h, o_1^h), \ldots, (t_T^h, o_T^h)\}$ (where $t^h \in \mathbb{R}^3$ and $o^h \in \mathrm{SO}(3)$ denote the head translation and

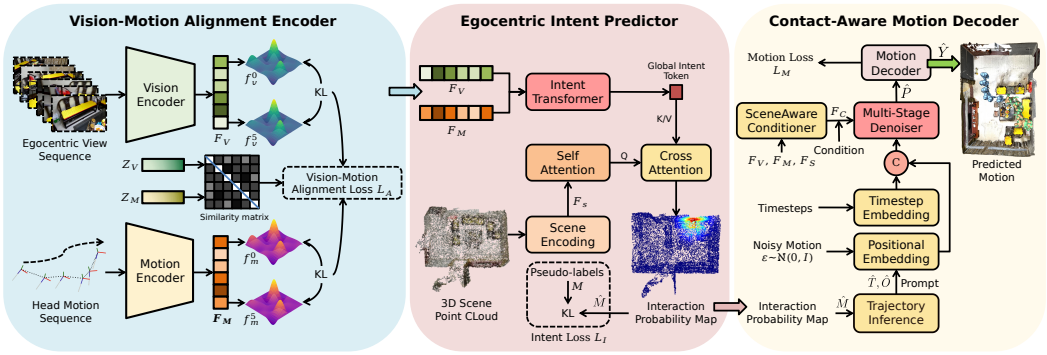


Figure 2: **Pipeline of EgoHMP.** EgoHMP first employs a pre-trained vision-motion alignment encoder to extract aligned features by minimizing the discrepancy between visual and motion disparities (Sec. 3.2). Subsequently, the egocentric intent predictor uses these aligned features to query the 3D scene point clouds, inferring an interaction probability map (Sec. 3.3). Finally, guided by this probability map, a contact-aware motion decoder synthesizes the final human motion sequence (Sec. 3.4).

orientation, respectively), and a 3D scene point cloud $S = \{p_1, \ldots, p_N\} \in \mathbb{R}^{N\times 6}$ with RGB, our objective is to forecast the future motion sequence $\hat{Y}_{T+1:T+\Delta T} = \{x_{T+1}, \ldots, x_{T+\Delta T}\}$. Following [36], each pose is parameterized as $x_k = (t_k, o_k, p_k)$, comprising the translation $t_k \in \mathbb{R}^3$, the orientation $o_k \in \mathrm{SO}(3)$, and a body pose embedding $p_k \in \mathbb{R}^{32}$. Following [7], we define the historical observation window as $T = 6$ frames and the future prediction horizon as $\Delta T = 10$ frames to achieve long-term prediction. Formally, we cast inference as a conditional generative process. Parameterized by θ, our model f_θ synthesizes future motion by iteratively denoising a pure Gaussian noise sequence $\mathcal{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ conditioned on the multi-modal inputs:

$$\hat{Y}_{T+1:T+\Delta T} = \arg\max f_\theta\big(\mathcal{Z} \mid V_{1:T}, H_{1:T}, S\big). \tag{1}$$

3.2 Vision-Motion Alignment Encoder

Vision Encoder. The vision encoder aims to extract context-rich visual representations from egocentric video clips. Given an input sequence $V_{1:T} \in \mathbb{R}^{T\times H\times W\times 3}$, each frame is encoded using a pre-trained DINOv2 [37], followed by a linear projection to obtain frame-level embeddings $F_V \in \mathbb{R}^{T\times D}$, where D denotes the feature dimension. To capture temporal dependencies in human motion, we introduce a learnable temporal positional encoding into F_V and feed the resulting sequence into a multi-layer Transformer encoder to model correlations among frame-level features. Meanwhile, we apply average pooling over F_V to obtain a sequence-level global representation, which is then projected into a lower-dimensional space $Z_V \in \mathbb{R}^{D_{proj}}$ via MLP layers, where D_{proj} denotes the projection dimension. Finally, the encoder outputs the feature representations $\{F_V, Z_V\}$.

Motion Encoder. The motion encoder extracts high-dimensional motion representations from the head motion sequence $H_{1:T} \in \mathbb{R}^{T\times 12}$. We first project $H_{1:T}$ via MLP layers into high-dimensional embeddings $F_M \in \mathbb{R}^{T\times D}$. To maintain consistency with the vision encoder, we then inject a learnable temporal positional encoding and feed the sequence into a lightweight Transformer encoder. Since head motion signals exhibit lower complexity than vision signals, a lightweight Transformer is sufficient to capture correlations among frame-level features. Following the same design paradigm as the vision encoder, we further aggregate F_M via average pooling and project it into a lower-dimensional latent space Z_M. The motion encoder thus outputs the representations $\{F_M, Z_M\}$.

Vision-Motion Alignment Loss. To align visual and motion modalities, we propose a vision-motion alignment loss $\mathcal{L}_A$ that correlates cross-modal features by restricting the local distances and global semantic alignment between their respective disparities in the feature space, allowing for mutual sensitivity to modal variations. Given F_V, F_M and Z_V, Z_M, the objective is defined as:

$$\mathcal{L}_A = \left\| \mathrm{KL}\big(\mathbf{d}_M^{(1)} \,\|\, \mathbf{d}_M^{(T)}\big) - \mathrm{KL}\big(\mathbf{d}_V^{(1)} \,\|\, \mathbf{d}_V^{(T)}\big) \right\|^2 + \lambda \cdot \frac{1}{2}\left[\mathrm{CE}(\mathbf{S}, \mathbf{y}) + \mathrm{CE}(\mathbf{S}^\top, \mathbf{y})\right], \tag{2}$$

where $\mathbf{d}^{(t)} = \mathrm{softmax}(F^{(t)}/\tau_{kl})$ is the temperature-scaled feature distribution. $\mathbf{S} = (Z_M Z_V^\top)/\tau_{ce}$ denotes the temperature-scaled pairwise cosine similarity matrix within a batch, and CE denotes

cross-entropy guided by the matching labels $\mathbf{y}$ to attract paired vision-motion samples while repelling unmatched ones within a batch [38]. τ_{kl} and τ_{ce} are temperature scalars, and λ balances the objectives.

3.3 Egocentric Intent Predictor

Scene Encoding. We employ PointNet++ [39] as our backbone to encode the 3D scene point cloud $S \in \mathbb{R}^{N\times 6}$ into point-wise features $F_S \in \mathbb{R}^{N\times D_e}$ (D_e denotes the feature dimension).

Intent Transformer. To synergize cross-modal features, we first project F_V and F_M into a shared latent space $\mathbb{R}^{T\times D_e}$. In egocentric scenarios, visual observations and head motions convey distinct intent cues with varying degrees of reliability. Indiscriminate fusion often causes reliable intent cues to be overshadowed by irrelevant noise. To address this issue and enable the network to adaptively adopt reliable cues for intent modeling, we propose a modality-aware modulation mechanism.

Specifically, we introduce learnable modulation tokens $\tau_V, \tau_M \in \mathbb{R}^{D_e}$ to scale the respective features via the Hadamard product ($\odot$), which function as a gradient interception mechanism during backpropagation. Formally, let $Z'_V = \tau_V \odot F_V$ denote the modulated visual features. The gradient of the total loss $\mathcal{L}$ with respect to any upstream mapping parameter θ can be derived via the chain rule, and the parameters are updated accordingly via gradient descent, as:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \left(\frac{\partial \mathcal{L}}{\partial Z'_V} \odot \tau_V\right) \frac{\partial F_V}{\partial \theta}, \theta_{t+1} \leftarrow \theta_t - \eta \left(\frac{\partial \mathcal{L}}{\partial Z'_V} \odot \tau_V\right) \frac{\partial F_V}{\partial \theta_t} \tag{3}$$

where η is the learning rate. Through this modulation, the network is able to adaptively suppress gradients from unreliable interaction cues during optimization. To finally distill a global intent, we introduce a learnable intent token $X_{intent} \in \mathbb{R}^{1\times D_e}$ and construct a feature sequence $Z_{in} \in \mathbb{R}^{(2T+1)\times D_e}$ by incorporating temporal positional embeddings E_{pos} as:

$$Z_{in} = \left[X_{intent} \,\middle\|\, (\tau_V \odot F_V + E_{pos}) \,\middle\|\, (\tau_M \odot F_M + E_{pos})\right], \tag{4}$$

where $[\cdot\|\cdot]$ denotes concatenation. The sequence is processed by a multi-layer Transformer, and the updated intent token X_{intent} is used as the global interaction representation $Z_{global} \in \mathbb{R}^{D_e}$.

Interaction Probability Map. We first use scene point features F_S to query the processed intent sequence via cross-attention, producing fused features $F_{fused} \in \mathbb{R}^{N\times D_e}$. Then, we compute a semantic matching score $S_{sim} = (F_{fused} \cdot Z_{global})/\sqrt{D_e}$ to measure the alignment between each point and the global intent. In parallel, a projection head predicts a point-wise interaction score S_{mlp} from F_{fused}. Finally, the interaction probability map is defined as: $\hat{M} = S_{mlp} + S_{sim} \in \mathbb{R}^N$.

Intent Loss. In existing benchmarks, human interaction intent lacks explicit annotations [40]. To provide supervision for the intent predictor, we formulate a pseudo-label $M \in \mathbb{R}^N$ as follows:

$$M_i = \lambda_c \exp\left(-\frac{\|p_i - p_{target}\|_2^2}{2\sigma_1^2}\right) + \lambda_m \frac{1}{\Delta T} \sum_{l=T+1}^{T+\Delta T} \left(\frac{l-T}{\Delta T}\right)^2 \exp\left(-\frac{\|p_i - t_l^h\|_2^2}{2\sigma_2^2}\right), \tag{5}$$

where p_{target} denotes the gt human-object contact position. The first term represents the interaction contact prior. The second term is the head motion prior, modulated by a temporal weight $(t/\Delta T)^2$ to capture the dynamic convergence of human intent. This design ensures that as time progresses toward the interaction moment, the user's head position serves as a stronger indicator of the intent. The deviations σ_1 and σ_2 are set to 0.3 and 1.0, respectively. Finally, to ensure valid probability distributions, the intent loss is defined as: $\mathcal{L}_I = \mathrm{KL}(\mathrm{softmax}(M) \,\|\, \mathrm{softmax}(\hat{M}))$.

3.4 Contact-Aware Motion Decoder

Trajectory Inference. To relieve optimization conflicts between local poses and global trajectories in diffusion models [8, 41], we decouple them by first inferring a global trajectory (translations $\hat{T}$, rotations $\hat{O}$) as a self-prompt:

$$(\hat{T}, \hat{O}) = \Psi_{\text{traj}}\big([\text{Linear}(H_{1:T}) \parallel \text{Linear}(H_{mid}) \parallel \text{Linear}(\hat{p}_{target})], F_{fused} \odot W_{dist}\big), \tag{6}$$

where Ψ_{traj} denotes the Transformer-based decoder (detailed in the Appendix A.3). The first term represents the coarse trajectory embeddings, constructed by concatenating the projected historical

head motion Linear($H_{1:T}$), the replicated intermediate states Linear(H_{mid}), and the predicted human-object contact position Linear($\hat{p}_{target}$) ($\hat{p}_{target} \in \mathbb{R}^3$ is extracted as the point with the highest probability from $\hat{M}$). To emphasize task-relevant regions, the F_{fused} are weighted by $W_{dist} \in \mathbb{R}^N$, where $W_{dist,i} = \mathcal{R}^*\big(\|p_i - \hat{p}_{target}\|_2\big)$ using the reverse min-max normalization operator $\mathcal{R}^*$.

Scene-Aware Conditioner. Following [33], we formulate the pose generation as a conditional denoising diffusion process. Guided by the trajectory prompt, the diffusion model focuses primarily on the local context. Therefore, we define the condition as: $F_C = [F_V \parallel F_M \parallel F_{\text{local}}]$, where F_{local} represents the local scene context. For each frame, we apply a cylindrical mask to isolate scene points along the trajectory. A small set of learnable latent queries then cross-attends to these points, efficiently distilling the point cloud into the compact local context F_{local}.

Multi-Stage Denoiser. During training, we use the ground-truth body poses as a clean sequence $P_0 \in \mathbb{R}^{\Delta T \times 32}$. The forward diffusion process progressively injects Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ into P_0 to yield a noisy sequence P_t at a random timestep t. To synthesize physically plausible motions, we propose a multi-stage denoiser f_ϕ. The denoiser f_ϕ is built upon stacked Transformer blocks that extract hidden states $\mathbf{H} \in \mathbb{R}^{\Delta T \times D}$ from P_t, conditioned on F_C and prompt $(\hat{T}, \hat{O})$. The generation process is designed into global coarse prediction stage and local contact-aware refinement stage.

In the first stage, a linear head maps $\mathbf{H}$ to a coarse pose sequence $\hat{P}_c \in \mathbb{R}^{\Delta T \times 32}$ to capture the fundamental motion manifold. In the second stage, recognizing that inferring specific human contact regions is crucial for enhancing egocentric motion understanding [22], we introduce local contact-aware refinement. Specifically, a parallel linear head maps the embedding $\mathbf{H}$ to a human contact probability distribution $\hat{\mathcal{A}} \in [0,1]^{\Delta T \times 10475}$, where 10475 denotes the number of SMPL-X vertices [36]. To supervise $\hat{\mathcal{A}}$, we apply a contact loss $\mathcal{L}_{con} = \text{KL}(\mathcal{A} \parallel \hat{\mathcal{A}})$ (detailed in Appendix). Subsequently, we project $\hat{\mathcal{A}}$ into a 22-dimensional kinematic space using SMPL-X skinning weights $\mathbf{W}$, calculating the contact importance for different joints $\hat{J} \in \mathbb{R}^{\Delta T \times 22 \times 3}$. A linear encoder Enc then maps this representation into contact features $\mathbf{C} = \text{Enc}(\hat{\mathcal{A}}\mathbf{W})$. To inject $\mathbf{C}$ into the generative process without distorting the latent space, we apply Adaptive Layer Normalization (AdaLN) to obtain the modulated feature $\mathbf{H}_{mod} = \text{LayerNorm}(\mathbf{H}) \odot (1 + \gamma(\mathbf{C})) + \beta(\mathbf{C})$. Here, $\gamma(\cdot)$ and $\beta(\cdot)$ denote the learnable scale and shift parameters derived from $\mathbf{C}$. Finally, the refined pose is defined as $\hat{P}_0 = \hat{P}_c + \Delta P$, where the residual displacement ΔP is computed by an MLP-based refiner f_l:

$$\Delta P = f_l\big([\mathbf{H}_{mod} \parallel \text{stop-gradient}(\hat{P}_c) \parallel \hat{T} \parallel \hat{O}]\big). \tag{7}$$

Motion Decoder. The final motion sequence $\hat{Y}_{T+1:T+\Delta T}$ is reconstructed via a GCN-based decoder [17]. Specifically, we use the pre-trained VPoser to decode the poses $\hat{P}_0$ into joint articulations, and use the SMPL-X [36] to integrate these articulations with the trajectory $(\hat{T}, \hat{O})$. The decoder then fuses these representations to a GCN model to refine the final 3D body mesh geometry.

Motion Loss. The decoder is optimized using a joint motion loss $\mathcal{L}_M$: $\mathcal{L}_M = \lambda_t \mathcal{L}_{traj} + \lambda_p \mathcal{L}_{pose} + \lambda_c \mathcal{L}_{con} + \lambda_j \mathcal{L}_{joint} + \lambda_s \mathcal{L}_{slide}$. Specifically, $\mathcal{L}_{traj}$ computes the L2 distance for root translations to ensure trajectory alignment. The pose loss $\mathcal{L}_{pose}$ is defined as: $\mathcal{L}_{pose} = \mathcal{L}_{coarse} + \mathcal{L}_{refine}$. $\mathcal{L}_{coarse}$ and $\mathcal{L}_{refine}$ compute the mean squared error between ground truth and coarse pose $\hat{P}_{coarse}$ as well as refined pose $\hat{P}_0$, respectively. Notably, we propose a joint adaptive reconstruction loss $\mathcal{L}_{joint}$:

$$\mathcal{L}_{joint} = \frac{1}{\Delta T \times 22} \sum_{t=T+1}^{T+\Delta T} \sum_{j=1}^{22} \big(1 + (\mathcal{A}\mathbf{W})_{t,j}\big) \big\| \hat{J}_{t,j} - J_{t,j} \big\|_2^2, \tag{8}$$

where $\mathcal{L}_{joint}$ uses the contact importance $(\mathcal{A}\mathbf{W})_{t,j}$ to adaptively penalize joints exhibiting higher contact probabilities. Furthermore, to relieve sliding artifacts, we introduce a slide loss $\mathcal{L}_{slide}$. It penalizes the inter-frame velocity of joints conditioned on their contact importance:

$$\mathcal{L}_{slide} = \frac{1}{(\Delta T - 1) \times 22} \sum_{t=T+2}^{T+\Delta T} \sum_{j=1}^{22} (\mathcal{A}\mathbf{W})_{t,j} \big\| \hat{J}_{t,j} - \hat{J}_{t-1,j} \big\|_2^2. \tag{9}$$

Table 1: **Quantitative Results. T-P** and **T-E** denote Traj-Path and Traj-End. **M-P** and **M-E** denote MPJPE-Path and MPJPE-End. Intent localization is evaluated via **TDE@*K*** for $K \in \{1, 50, 100\}$. **Pen.** denotes penetration error. All metrics are reported in millimeters (mm) except for **Pen.** (in %). Lower values indicate better performance (↓). The best results are highlighted in **bold**.

Method	**GIMO** [5]								**GTA-IM** [24]							
	Motion Metrics ↓				Intent Metrics ↓			Phys. ↓	Motion Metrics ↓				Intent Metrics ↓			Phys. ↓
	T-P	T-E	M-P	M-E	@1	@50	@100	Pen.	T-P	T-E	M-P	M-E	@1	@50	@100	Pen.
Gaze-based																
BiFu [5]	661	727	167.8	205.0	415.8	405.3	398.7	3.73	683	1066	186.4	252.6	455.2	445.6	438.1	3.95
SIF3D [17]	590	666	156.6	195.7	345.2	336.8	330.4	4.59	626	836	164.9	227.7	385.6	378.2	371.9	4.82
GAP3DS [7]	575	623	141.2	171.4	210.5	205.1	201.3	0.95	610	799	155.0	210.1	235.4	228.6	224.5	1.12
Non-Gaze-based																
Auxformer [42]	688	893	173.5	205.4	620.5	612.3	605.4	6.65	688	1078	187.8	262.0	685.2	678.5	670.3	7.10
MDP [27]	612	679	157.1	190.4	512.4	502.8	495.6	4.77	640	860	162.5	222.5	560.3	551.2	545.8	5.05
EgoHMP (Full)	**481**	**545**	**114.9**	**138.0**	**161.3**	**153.2**	**148.6**	**0.81**	**522**	**623**	**129.6**	**177.2**	**182.5**	**176.4**	**174.0**	**0.91**

4 Experiment

4.1 Experimental Setup

Datasets: (1) GIMO [5] has ∼129K frames of human pose data, including three core types: whole-body SMPL-X poses, LiDAR-scanned 3D geometric scenes (14 scenes, 10–20 objects/scene, ∼500K vertices/scene), and 3D gaze coordinates. Following [7, 17], we downsample sequences to 2fps and adopt the original train-test split. **(2) GTA-IM** [24] has ∼1M frames of human pose data, with segmentation and camera pose annotations, covering 10 indoor scenes, 50 unique characters, 22 walking styles, and diverse actions. **(3) Ego-Exo4D** [25] is a large-scale dataset containing over 1,400 hours of synchronized egocentric and exocentric videos. It captures skilled human activities across diverse real-world environments, providing rich annotations including high-fidelity 3D whole-body and hand poses, eye gaze, camera trajectories, and 3D scene point clouds.

Metrics. We use six metrics to evaluate EgoHMP: (1) **Traj-Path** and (2) **Traj-End** [5] measure the average and final endpoint deviations of the trajectory. (3) **MPJPE-Path** and (4) **MPJPE-End** [43] measure the mean joint positional errors of the whole pose sequence and final pose. (5) **TDE@*K*** evaluates intent localization accuracy via the minimum L2 distance between the ground-truth interaction point and the top-K predicted interaction points. (6) **Penetration Error** calculates the percentage of body intersecting the scene via SDF [44], quantifying physical plausibility.

Baselines. We compare EgoHMP against a comprehensive set of SoTA methods, which can be categorized into two main groups. For scene-aware methods, we compare against both non-gaze-based methods (Auxformer [42], MDP [27]) and gaze-based methods (BiFu [5], SIF3D [17], GAP3DS [7]). To further validate EgoHMP, we benchmark against the latest egocentric-driven approaches (EgoEgo [15], UniEgoMotion [14], EgoCast [31], EgoAllo [16]). Notably, these egocentric-driven approaches do not explicitly condition on 3D scene. Therefore, to ensure fairness comparison, we first evaluate the pose error within the local coordinate space (w/o scene-aware). Subsequently, we re-evaluate them after aligning their predicted local poses with the ground-truth global trajectories in the 3D scene (w/ scene-aware). EgoHMP is evaluated using its own predicted trajectories.

4.2 Main Experimental Results and Analysis

Table 1 presents a quantitative comparison between the proposed EgoHMP and state-of-the-art baselines on both datasets. The results demonstrate that EgoHMP outperforms all competing methods across all evaluated metrics, validating the effectiveness of our approach. Notably, compared to the best-performing gaze-based baseline, GAP3DS [7], EgoHMP achieves consistent improvements across all metrics and secures the lowest intent localization er-

Table 2: **Quantitative Results on Ego-Exo4D** [25].

Method	Scene-Aware	M-P (mm) ↓	M-E (mm) ↓
EgoEgo [15]	✗	315.0	401.3
EgoEgo [15]	✓	263.8	335.4
EgoCast [31]	✗	225.8	281.6
EgoCast [31]	✓	153.6	193.7
EgoAllo [16]	✗	**207.4**	**268.8**
EgoAllo [16]	✓	144.7	178.4
UniEgoMotion [14]	✗	216.4	285.5
UniEgoMotion [14]	✓	158.5	192.5
EgoHMP (Base)	✗	213.5	276.2
EgoHMP (Full)	✓	**128.6**	**157.0**

Table 3: **Quantitative Ablations.** F_V: egocentric view signal; **G**: using the gaze signal instead of the egocentric view signal; τ_V, τ_M: modality-aware modulation; S_{pm}: self-prompted mechanism; F_{local}: scene-aware conditioner; $\mathcal{D}_{ms}$: multi-stage denoiser; D_m: GCN-based motion decoder; $\mathcal{L}_A/\mathcal{L}_{joint}/\mathcal{L}_{slide}$: loss functions. ✗ means without.

Metric ↓	✗ F_V	G	✗ τ_V, τ_M	✗ S_{pm}	✗ F_{local}	✗ $\mathcal{D}_{ms}$	✗ D_m	✗ $\mathcal{L}_A$	✗ $\mathcal{L}_{joint}$	✗ $\mathcal{L}_{slide}$	**Full**
T-P	676	589	501	612	482	483	483	558	484	483	**481**
T-E	784	643	576	691	547	548	546	639	546	547	**545**
M-P	168.9	145.9	130.4	150.4	135.3	134.7	118.6	132.6	122.8	115.4	**114.9**
M-E	216.8	171.7	150.6	182.2	167.2	165.8	142.0	161.4	153.5	140.9	**138.0**

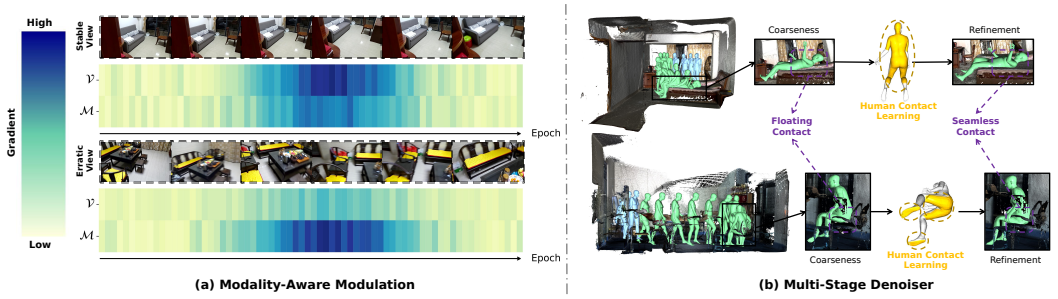


Figure 3: **Visualization Ablations.** (a) The gradient flow of the visual ($\mathcal{V}$) and motion ($\mathcal{M}$) modalities across 70 sampled training epochs. (b) The refinement process of the multi-stage denoiser.

ror (TDE@K) on both benchmarks. While gaze-based methods generally outperform non-gaze approaches, they remain constrained by the inherent drift and ambiguity of human gaze signals, limiting the potential for further accuracy gains. Conversely, non-gaze methods, such as MDP [27], yield significantly higher trajectory and intent errors due to a complete lack of explicit intent localization. By substituting gaze signals with egocentric vision-motion cues, EgoHMP successfully constructs accurate interaction probability maps. This inferred intent serves as an anchor for trajectory prediction, thereby generating precise and collision-free trajectories. Subsequently, acting as a self-prompt, this trajectory alleviates the global generation burden on the diffusion model, enabling it to focus on synthesizing context-consistent and physically plausible local poses.

As shown in Table 2, we further evaluate EgoHMP on Ego-Exo4D [25]. Under the scene-agnostic setting (✗), our model does not exhibit an absolute advantage. In this context, kinematics-focused baselines like EgoAllo [16] achieve robust results, as our framework relies on 3D scene geometry to generate context-consistent motions. However, to demonstrate the importance of 3D scene understanding, we establish the scene-aware setting (✓). Since most baselines cannot ingest 3D scene geometry, we simulate an ideal scene constraint for them by aligning their predictions with the ground-truth trajectories. In contrast, EgoHMP uses its predicted trajectories. Although ground-truth trajectories significantly improve baseline accuracy, their pose errors remain higher than ours. This demonstrates EgoHMP effectively integrates scene context to constrain and rectify predicted poses, thereby achieving better performance.

4.3 Ablation Study

We conduct extensive ablation studies on GIMO [5] and the results are summarized in Table 3.

Input Signals. Removing the egocentric view signal (✗ F_V) blinds the model to the scene semantics, causing a deterioration in prediction accuracy. Furthermore, replacing F_V with gaze signals (**G**) results in sub-optimal performance. This demonstrates that gaze signals are easily distracted by task-irrelevant objects, whereas the egocentric view provides a more robust context for intent modeling.

Intent Modeling. We first ablate the vision-motion alignment loss (✗ $\mathcal{L}_A$). Its removal causes a significant drop in accuracy, indicating that bridging the semantic gap between visual and motion disparities is crucial for reliable intent prediction. Additionally, without the modality-aware modulation (✗ τ_V, τ_M), the model fails to intercept gradients from unreliable cues, leading to degradation in prediction accuracy. As illustrated in Figure 3 (a), under a stable egocentric view, the visual

Table 4: **Robustness Evaluation.** (1) Foot Sliding (FS) penalizes abnormal horizontal shifts of ankle and toe vertices when they fall below contact height thresholds (80mm and 40mm). (2) Ground Contact Deviation (GCD) captures floating/penetration artifacts via the absolute minimum vertical distance between supporting vertices and the local scene surface. (3) Human Contact Distance (HCD) evaluates interaction quality by measuring the average Euclidean distance from the intended contact vertices to the intended scene surface.

Setting	**Perturbation**	T-P ↓	T-E ↓	M-P ↓	M-E ↓	Pen. ↓	FS ↓	GCD ↓	HCD ↓
Clean Input	Original Data	**481**	**545**	**114.9**	**138.0**	**0.81**	**3.70**	**36.4**	**43.5**
Head-Pose Noise	Gaussian, $\sigma = 0.05$	487	553	120.2	142.1	**0.81**	3.71	36.6	44.9
	Gaussian, $\sigma = 0.10$	505	572	128.6	155.7	0.83	3.74	37.0	47.3
Calibration Error	Offset (5 cm, 2°)	511	579	131.8	161.5	0.83	3.72	36.9	51.5
	Offset (10 cm, 5°)	524	594	139.3	170.4	0.88	3.76	37.5	54.0
Incomplete Geo.	30% Points Dropped	490	565	119.3	144.2	0.86	3.71	36.5	44.7
	50% Points Dropped	533	621	145.4	174.0	0.91	3.74	36.8	56.2

observations are clean and goal-directed, generating a high gradient flow in the visual branch. This enables both modalities to update normally and maintain a healthy balance. However, under an erratic view characterized by drastic viewpoint shifts, the visual observations fail to maintain a stable focus on the target object and become cluttered with multiple irrelevant objects, rendering the interaction intent ambiguous. Regulated by the modality-aware modulation mechanism, the network dampens magnitude of the visual gradients. Consequently, the optimization focus shifts toward head motions to extract reliable interaction contexts, thereby achieving robust intent modeling.

Motion Generation and Refinement. When the self-prompted mechanism is removed (✗ S_{pm}), the diffusion model suffers from optimization conflicts, resulting in a significant drop in motion metrics. Removing the scene-aware conditioner (✗ F_{local}) causes the pose errors to surge sharply. This validates that the fine-grained local scene geometry is indispensable for generating physically plausible local poses. Removing the multi-stage denoiser (✗ $\mathcal{D}_{ms}$) deprives the model of human contact modeling, causing noticeable pose error. As shown in Figure 3 (b), in the initial stage, the predicted poses suffer from artifacts such as floating. By incorporating human contact learning, the model leverages perceived human contact regions to correct local joint articulations, thereby generating physically plausible poses. Moreover, ablating the joint adaptive reconstruction loss (✗ $\mathcal{L}_{joint}$) allows the network to bypass penalties on highly-probable contact regions. Similarly, without the slide loss (✗ $\mathcal{L}_{slide}$) or the GCN-based decoder (✗ D_m), the model loses micro-level geometric refinements, which are essential for relieving slide and penetration.

4.4 Robustness Evaluation

To further evaluate the robustness of EgoHMP, we simulate three types of perturbations: (1) Head-Pose Noise, where Gaussian noise is injected into head motions; (2) Calibration Error, where spatial offsets are applied to simulate sensor misalignment; and (3) Incomplete Geometry, where the scene points are randomly removed to simulate incomplete scene scans.

As shown in Table 4, EgoHMP maintains relatively stable performance across different perturbation levels, demonstrating its robustness to noisy head poses, calibration errors, and incomplete scene geometry. This robustness benefits from the modality-aware modulation, which dynamically suppresses unreliable modalities by adjusting gradient flows, allowing complementary cues to maintain stable intent inference. Furthermore, the intent predictor and contact decoder are supervised via Gaussian-smoothed continuous probability maps rather than hard labels, preventing the network from overfitting to granular 3D details and enabling it to generalize well to sparse or lower-quality point clouds. The scene-aware conditioner further improves robustness by extracting local scene features within a bounded spatial region around the predicted trajectory, making the model less sensitive to occlusions and missing regions outside the interaction area.

4.5 Visualization Analysis

As shown in Figure 4, we compare our EgoHMP with the state-of-the-art gaze-based method, GAP3DS [7], across diverse scenarios. The rightmost columns for each scene visualize the predicted

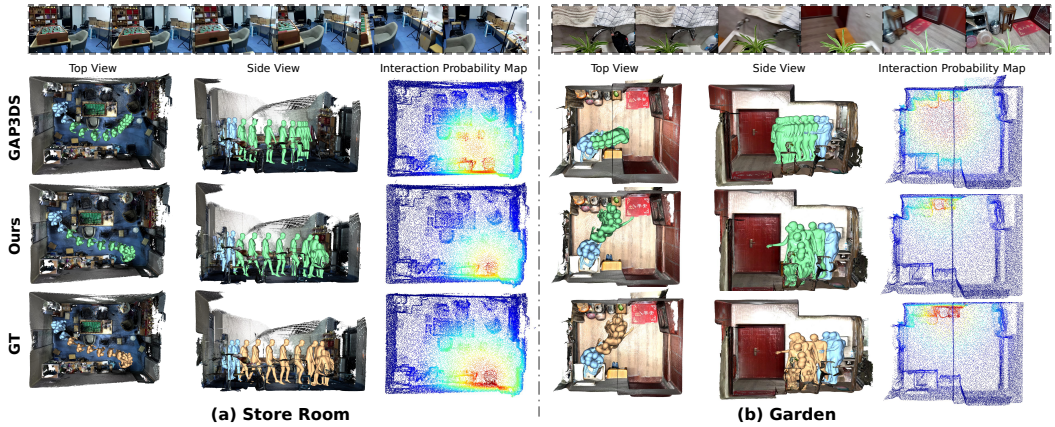


Figure 4: **Visualization Comparison.** The blue, green, and tan human meshes represent the historical, predicted, and ground-truth motion sequences, respectively. Zoom in for better visualization.

interaction probability maps. Due to the ambiguity and drift of human gaze signals, GAP3DS is easily misled by task-irrelevant visual regions, resulting in scattered and mislocalized interaction probability maps. Consequently, as observed in the top views and side views, this erroneous intent estimation causes GAP3DS to generate incorrect motion sequences. In contrast, EgoHMP accurately infers the user's interaction intent by exploring the egocentric visual context coupled with head motion cues. Guided by the precise interaction probability map, EgoHMP successfully generates context-consistent and physically plausible motion sequences.

5 Conclusion and Discussion

In this paper, we propose EgoHMP, a novel framework for precise human motion prediction (HMP) in 3D scenes. Recognizing the ambiguity, drift, and hardware dependency of gaze signals, we introduce a paradigm shift that uses egocentric observations and head motions as a more robust proxy for human interaction intent. To achieve this, EgoHMP bridges the cross-modal semantic gap and employs a modality-aware modulation mechanism to adaptively distill reliable interaction contexts. Crucially, to bridge scene perception and action execution, our framework simulates human "cognitive envisioning" by simultaneously predicting scene interaction and human contact regions. Guided by this dual-intent representation, our contact-aware motion decoder decouples global trajectory formulation from local pose synthesis. By utilizing a global trajectory as a self-prompt and progressively refining poses via a multi-stage denoiser, EgoHMP effectively mitigates geometric artifacts. Extensive experiments demonstrate that EgoHMP significantly outperforms existing methods, synthesizing 3D motions that are both semantically coherent and physically plausible.

Limitations and Future Work. Currently, while egocentric views and head motions offer more robust intent modeling compared to gaze signals, extreme real-world cases (where both visual observations and head motions exhibit highly deceptive cues) will mislead the model's intent inference. Furthermore, EgoHMP predicts motion of a fixed sequence length. When the actual sequence length varies, it tends to truncate ongoing actions or generate stationary poses. Future work will explore long-term temporal causal reasoning to disambiguate short-term deceptive cues, as well as variable-length motion prediction paradigms to dynamically adapt to varying sequence length.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62672066, 62272417, and 62672037. The authors declare no competing interests.

References

[1] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2891–2900, 2017.

[2] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6519–6527, 2020.

[3] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 214–223, 2020.

[4] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. Quaternet: A quaternion-based recurrent model for human motion. *International Journal of Computer Vision*, 128:855–872, 2020.

[5] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision (ECCV)*, pages 676–694. Springer, 2022.

[6] Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J Black. Wandr: Intention-guided human motion generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 927–936, 2024.

[7] Ting Yu, Yi Lin, Jun Yu, Zhenyu Lou, and Qiongjie Cui. Vision-guided action: Enhancing 3d human motion prediction with gaze-informed affordance in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12335–12346, 2025.

[8] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021.

[9] Wei Mao, Richard I Hartley, Mathieu Salzmann, et al. Contact-aware human motion forecasting. In *Advances in Neural Information Processing Systems*, volume 35, pages 7356–7367, 2022.

[10] Yuanzhi Liang, Xiaohan Wang, Linchao Zhu, and Yi Yang. Maal: Multimodality-aware autoencoder-based affordance learning for 3d articulated objects. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 217–227, 2023.

[11] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 433–444, 2024.

[12] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14531–14542, 2024.

[13] Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3086–3096, 2024.

[14] Chaitanya Patel, Hiroki Nakamura, Yuta Kyuragi, Kazuki Kozuka, Juan Carlos Niebles, and Ehsan Adeli. Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10318–10329, 2025.

[15] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023.

[16] Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7072–7084, 2025.

[17] Zhenyu Lou, Qiongjie Cui, Haofan Wang, Xu Tang, and Hong Zhou. Multimodal sense-informed prediction of 3d human motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[18] Randy L Buckner. The cerebellum and cognitive function: 25 years of insight from anatomy and neuroimaging. *Neuron*, 80(3):807–815, 2013.

[19] Jeff Pelz, Mary Hayhoe, and Russ Loeber. The coordination of eye, head, and hand movements in a natural task. *Experimental brain research*, 139(3):266–277, 2001.

[20] Arash Fassihi, Athena Akrami, Francesca Pulecchi, Vinzenz Schönfelder, and Mathew E Diamond. Transformation of perception from sensory to motor cortex. *Current Biology*, 27(11): 1585–1596, 2017.

[21] Ralph Adolphs. Cognitive neuroscience of human social behaviour. *Nature reviews neuroscience*, 4(3):165–178, 2003.

[22] Yuhang Yang, Wei Zhai, Chengfeng Wang, Chengjun Yu, Yang Cao, and Zheng-Jun Zha. Egochoir: Capturing 3d human-object interaction regions from egocentric views. *Advances in Neural Information Processing Systems*, 37:54529–54557, 2024.

[23] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16284–16295, 2024.

[24] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020.

[25] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[26] Xiao Han, Yiming Ren, Yichen Yao, Yujing Sun, and Yuexin Ma. Towards practical human motion prediction with lidar point clouds. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7629–7638, 2024.

[27] Chaoyue Xing, Wei Mao, and Miaomiao Liu. Scene-aware human motion forecasting via mutual distance prediction. In *European Conference on Computer Vision*, pages 128–144. Springer, 2024.

[28] Esteve Valls Mascaró, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6048–6057, 2023.

[29] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer for egocentric action anticipation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6740–6750, 2024.

[30] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing*, 30:1143–1152, 2020.

[31] Maria Escobar, Juanita Puentes, Cristhian Forigua, Jordi Pont-Tuset, Kevis-Kokitsi Maninis, and Pablo Arbelaez. Egocast: Forecasting egocentric human pose in the wild. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5831–5841. IEEE, 2025.

[32] Jian Wang, Rishabh Dabral, Diogo Luvizon, Zhe Cao, Lingjie Liu, Thabo Beeler, and Christian Theobalt. Ego4o: Egocentric human motion capture and understanding from multi-modal input. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22668–22679, 2025.

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[34] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022.

[35] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16010–16021, 2023.

[36] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[39] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[40] Zhenyu Lou, Qiongjie Cui, Tuo Wang, Zhenbo Song, Luoming Zhang, Cheng Cheng, Haofan Wang, Xu Tang, Huaxia Li, and Hong Zhou. Harmonizing stochasticity and determinism: Scene-responsive diverse human motion prediction. *Advances in Neural Information Processing Systems*, 37:39784–39811, 2024.

[41] German Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2317–2327, 2023.

[42] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Xinchao Wang, and Yanfeng Wang. Auxiliary tasks benefit 3d skeleton-based human motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9509–9520, 2023.

[43] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *International Conference on 3D Vision (3DV)*, pages 565–574, 2021.

[44] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[45] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020.

[46] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13309–13318, 2021.

[47] Jianwei Tang, Hong Yang, Tengyue Chen, and Jian-Fang Hu. Stochastic human motion prediction with memory of action transition and action characteristic. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1883–1893, 2025.

[48] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023.

[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[50] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Appendix

A Architecture and Training Details

All training detailed in this section were conducted on a single NVIDIA RTX 3090 GPU. Notably, within the EgoHMP framework, only the Vision-Motion Alignment Encoder requires pre-training, whereas the Egocentric Intent Predictor and Contact-Aware Motion Decoder are optimized end-to-end. The necessity for this pre-training phase stems from the profound semantic and dimensional gap between 2D egocentric visual observations and 3D head motion kinematics. Jointly training such heterogeneous modalities from scratch alongside a complex downstream generative task can easily lead to representation collapse or suboptimal convergence. We conduct an experiment to validate this point. Please refer to Sec. B.1

A.1 Pre-trained Vision-Motion Alignment Encoder

Vision Encoder. The vision encoder takes egocentric video clips uniformly sampled at 2 fps. Each frame is resized to 224×224 and processed by a frozen pre-trained DINOv2 (ViT-L/14) [37] model to extract frame-level features with a dimension of $D = 1024$. To model temporal correlations, we inject a learnable temporal positional embedding (initialized from a normal distribution $\mathcal{N}(0, 0.02)$) and process the sequence through a 4-layer Transformer encoder with 8 attention heads and a feed-forward network dimension of 2048. The global representation is obtained via average pooling along the temporal axis and then mapped to a latent space of $D_{proj} = 256$ using a projection head consisting of an MLP (Linear(1024, 512) $\rightarrow$ BatchNorm1d $\rightarrow$ GELU $\rightarrow$ Linear(512, 256)).

Motion Encoder. For the motion encoder, the raw input is a 12-dimensional relative motion vector for each frame (comprising a 3D translation vector and a 3×3 flattened rotation matrix relative to the first frame). This input is first lifted to a 1024-dimensional space via a spatial MLP (Linear(12, 512) $\rightarrow$ LayerNorm $\rightarrow$ GELU $\rightarrow$ Linear(512, 1024)). Similar to the vision branch, temporal positional embeddings are added before feeding the sequence into a 1-layer lightweight Transformer encoder (4 attention heads, 2048 FFN dimension). The global motion representation is similarly average-pooled and mapped to $D_{proj} = 256$ via an identical projection head.

The vision-motion alignment encoder requires robust pre-training prior to the joint optimization of EgoHMP. To maximize alignment generalization and avoid data leakage, we propose a massive self-supervised data mining strategy. Specifically, we apply a dense sliding window with a stride of 10 frames across all training sequences to construct a large-scale paired vision-motion pre-training dataset. During this process, any sliding window that intersects with the test set sequences (or a 15-frame safety margin surrounding them) is strictly excluded.

Training Details. During pre-training, the model is optimized using the vision-motion alignment loss $\mathcal{L}_A$ defined in Eq. (2) of the main manuscript. The temperature hyper-parameters for the KL divergence and the InfoNCE loss (Cross-Entropy) are empirically set to $\tau_{kl} = 1.0$ and $\tau_{ce} = 0.7$, respectively. The balancing weight for the contrastive objective is set to $\lambda = 5.0$. The alignment encoder is pre-trained for 80 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 64. We implement an offline caching mechanism for data pre-processing and feature extraction, which requires a one-time overhead of approximately 20 minutes. Once cached, the in-memory data loading reduces the time per epoch to merely 3 seconds, allowing the core pre-training phase to be completed in roughly 4 minutes.

A.2 Egocentric Intent Predictor

Intent Transformer and Modality-Aware Modulation. To adaptively extract reliable interaction contexts and suppress invalid cues, we apply the modality-aware modulation mechanism as formulated in the main manuscript. Specifically, we define learnable modality-specific modulation tokens, $\tau_V, \tau_M \in \mathbb{R}^{256}$, which scale the respective projected features via the Hadamard product ($\odot$). This multiplicative gating effectively functions as a gradient interception mechanism during backpropagation. Subsequently, learnable temporal positional embeddings ($E_{pos} \in \mathbb{R}^{1\times6\times256}$) are added to the modulated features to preserve temporal dynamics. The resulting visual and motion sequences are then concatenated along the temporal dimension. Finally, a learnable global intent token ($X_{intent} \in \mathbb{R}^{1\times1\times256}$) is prepended to the concatenated features, yielding a unified 13-token

sequence. This sequence is processed by a 2-layer Transformer encoder (8 attention heads, 1024 FFN dimension) to establish dense cross-modal interactions, and the updated X_{intent} token is extracted as the global intent representation.

Scene Encoding and Interaction Probability Map Generation. For the 3D scenes, we employ PointNet++ [39] as our backbone to downsample and encode the raw 3D scene point clouds into point-wise geometric features, which are then mapped to the shared latent dimension ($D_e = 256$). To capture complex intra-scene topological relations, these point features are first refined by a 1-layer self-attention module (8 attention heads). Subsequently, we introduce a 3-layer cross-attention mechanism where the refined scene point features act as queries to attend to the 13-token intent representation (acting as keys and values). Ultimately, a prediction head consisting of a 1D convolution block (Conv1d(256, 64) $\rightarrow$ ReLU $\rightarrow$ Conv1d(64, 1)) projects the fused features into point-wise probability, yielding the final interaction probability map.

Intent Pseudo-Label and Intent Loss. Since human cognitive intent is inherently unobservable and lacks explicit annotations in existing datasets, we formulate a pseudo-label generation strategy (Eq. (5) of the main manuscript.) to provide probabilistic supervision. To address potential concerns regarding the empirical fidelity and robustness of these hand-crafted heuristics, we ground our design in fundamental cognitive science and biomechanical principles. First, the pseudo-label leverages the future physical contact point as an absolute anchor, representing the ultimate physical manifestation of intent. Second, it incorporates head motion proximity. Extensive studies in motor control demonstrate that head orientation acts as a reliable kinematic precursor to hand and body interactions. Crucially, we introduce a temporal decay weight $(t/\Delta T)^2$ to modulate the head motion prior. This reflects the cognitive reality that as a human approaches an interaction moment, their head trajectory becomes increasingly deterministic and strongly correlated with the target, thereby minimizing ambiguity. Furthermore, our method is inherently robust to potential errors or ambiguities in these heuristics. Rather than employing rigid, one-hot hard labels, Eq. (5) yields a continuous, Gaussian-smoothed probability distribution map. Supervising the network via Kullback-Leibler (KL) divergence with these soft distributions acts as a form of label smoothing. It prevents the model from overfitting to isolated heuristic inaccuracies and encourages the network to learn the underlying multi-modal distribution of plausible intents.

A.3 Contact-Aware Motion Decoder

Trajectory Inference. To decouple global locomotion from local pose articulation, we first predict a coarse future trajectory. The predicted interaction target (extracted via the argmax of the intent probability map) serves as the spatial goal. The historical 6-frame trajectory and the goal are individually embedded into a high-dimensional latent space ($D_{latent} = 768$) using linear projections. To maintain temporal continuity, we interpolate the intermediate frames, constructing a complete 16-frame sequence. This sequence undergoes successive cross-modal fusions: an 8-layer cross-attention module queries the global scene features, followed by another 8-layer cross-attention module that selectively queries non-obstacle geometric features. Finally, a 4-layer Transformer encoder (8 attention heads) refines these temporal features to output the predicted 6D global trajectories (3D translation and 3D axis-angle orientation). This predicted trajectory explicitly serves as a self-prompt for the subsequent diffusion process.

Scene-Aware Conditioner. To effectively ground the local pose generation within the 3D environment, we introduce a local scene conditioner. For each predicted frame, we transform the global point cloud into the local coordinate system of the current predicted root trajectory. We apply a cylindrical spatial mask—defined by a radius of 1.0 meter and a vertical range from -1.2 to 1.0 meters relative to the root—to filter out task-irrelevant geometries, retaining only the potential contact surfaces. To compress this local geometry into a compact representation, we initialize 4 learnable latent queries per frame. A Multi-Head Attention pooling layer (8 heads) aggregates the masked scene features into these queries, resulting in a highly efficient, contact-aware local context embedding sequence.

Multi-Stage Denoiser. The local pose generation is formulated as a conditional denoising diffusion probabilistic model. The denoising backbone is a 12-layer Transformer Decoder (12 attention heads, $D_{latent} = 768$, dropout $= 0.1$) that processes noisy latent poses conditioned on temporal embeddings, cross-modal features, and local scene contexts. To relieve physical artifacts, we design a multi-stage optimization strategy instead of a direct mapping. In the first stage, the network maps the latent features to predict a coarse 32D VPoser latent pose alongside a dense, vertex-level human

contact probability map (10,475 vertices). In the second stage, we introduce an Explicit Kinematic Pooling mechanism. Utilizing official SMPL-X Linear Blend Skinning weights, the 10,475 vertex probabilities are rigorously pooled into 22 joint-level contact priors. These kinematic priors are then encoded to generate scale (γ) and shift (β) parameters for an Adaptive Layer Normalization (AdaLN), which modulates the latent representations to inject explicit contact awareness. Crucially, the AdaLN generator is zero-initialized to prevent noisy intent signals from disrupting the generative manifold during early training. In the final stage, the modulated latent features, concatenated with the global trajectory and the stop-gradient coarse pose, are processed by a residual refiner to predict a fine-grained latent displacement (ΔP). The final refined pose is obtained by adding this zero-initialized residual to the coarse pose, successfully alleviating geometric artifacts such as floating and penetration while preserving the fundamental motion manifold.

Human Contact Pseudo-Label and Contact Loss. To provide supervision for human contact learning, we derive pseudo-labels for human-scene contact, as explicit contact annotations are absent in datasets. Specifically, we decode the ground-truth motion sequence into explicit 3D human meshes using the SMPL-X model. For each frame t, we compute the shortest Euclidean distance between each of the $V = 10,475$ human mesh vertices, denoted as $v_{t,i}$, and the 3D scene point cloud S. We then formulate the contact pseudo-label $\mathcal{A}_{t,i}$ by mapping this distance into a soft probability distribution via a Gaussian kernel:

$$\mathcal{A}_{t,i} = \exp\left(-\frac{\min_{p\in S} \|v_{t,i} - p\|_2^2}{2\sigma^2}\right). \tag{10}$$

Crucially, employing a Gaussian kernel rather than binary threshold simulates the physical nature of real-world human-scene interactions. For instance, when a human grasps a cup, the entire hand does not conform perfectly and uniformly to the surface; rather, the interaction manifests as focal contact points surrounded by a continuous spatial decay. By setting the standard deviation σ to 0.02 meters, our formulation accurately captures this proximity-based interaction pattern. Finally, after applying softmax normalization to ensure valid probability distributions, the contact loss $\mathcal{L}_{con}$ is defined as the KL divergence between the predicted contact map $\hat{\mathcal{A}}$ and the pseudo-label $\mathcal{A}$:

$$\mathcal{L}_{con} = \frac{1}{\Delta T} \sum_{t=T+1}^{T+\Delta T} \mathrm{KL}(\mathrm{softmax}(\mathcal{A}_t) \,||\, \mathrm{softmax}(\hat{\mathcal{A}}_t)). \tag{11}$$

Through the contact region learning, EgoHMP effectively captures the continuous spatial relationship between the human body and the 3D environment, which is paramount for alleviating geometric artifacts such as floating and penetration.

Motion Decoder. To ensure physical plausibility and alleviate foot-sliding artifacts, the predicted 32D VPoser latent codes and the predicted 6D trajectories are decoded into 23 3D joint coordinates (69D) via the pre-trained VPoser and SMPL-X [36] models. The 3D joints and the latent codes are concatenated to form a 101D mixed node representation. This sequence is then processed by a 9-stage Graph Convolutional Network (GCN). To ensure temporal smoothness, the GCN uses Discrete Cosine Transform (DCT) and Inverse DCT (IDCT). The refined joints are ultimately aligned with the predicted trajectory to prevent root drift.

Training Details. The Egocentric Intent Predictor and Contact-Aware Motion Decoder are optimized in an end-to-end joint training paradigm. We employ the AdamW optimizer with an initial learning rate of 1×10^{-4}, governed by a cosine annealing scheduler across 120 epochs. Each epoch takes approximately 45 seconds, with the total training process completing in roughly 1.5 hours.

B Additional Experiments

B.1 Further Analysis of Vision-Motion Alignment Encoder

Pre-training. As shown in Table 5 (a), when directly training the entire framework end-to-end (E2E) without pre-training phase, the model's performance drops drastically, highlighted by a near-zero cross-modal retrieval recall. This degradation fundamentally stems from the severe risk of dimensional collapse during joint optimization. Specifically, the continuous fine-tuning of the visual backbone causes the discrepancies among visual features to fluctuate violently. Consequently, to minimize the penalty, the alignment loss forces the visual disparities to zero, which in turn drives

Table 5: **Quantitative Results.** Evaluated on the GIMO [5]. We analyze the necessity of the pre-training for the vision-motion alignment encoder and the impact of different objectives within the alignment loss. E2E denotes training the entire framework end-to-end. $\mathcal{L}_A$ denotes the vision-motion alignment loss. kl denotes the KL divergence loss in $\mathcal{L}_A$. nce denotes the InfoNCE contrastive loss in $\mathcal{L}_A$. R@1, R@5, and R@10 denote the cross-modal retrieval recall metrics (%). $\mathcal{M} \rightarrow \mathcal{V}$ denotes motion-to-vision retrieval, and $\mathcal{V} \rightarrow \mathcal{M}$ denotes vision-to-motion retrieval.

Strategy	$\mathcal{M} \rightarrow \mathcal{V} \uparrow$			$\mathcal{V} \rightarrow \mathcal{M} \uparrow$			Motion Metrics ↓				Intent Metrics ↓			Phys. ↓
	R@1	R@5	R@10	R@1	R@5	R@10	T-P	T-E	M-P	M-E	@1	@50	@100	Pen.
(a) E2E	0.00	0.06	0.12	0.03	0.09	0.11	588	672	163.2	200.4	312.5	305.7	298.4	3.15
(b) Pre-train w/o kl	57.53	90.90	95.53	55.95	92.23	96.77	580	647	151.6	189.1	295.4	288.6	282.5	2.42
(c) Pre-train w/o nce	0.00	0.04	0.11	0.01	0.10	0.13	490	558	119.4	144.2	188.5	180.2	174.0	0.89
(d) Full Pre-training	**66.37**	**94.75**	**97.84**	**63.25**	**95.44**	**98.15**	**481**	**545**	**114.9**	**138.0**	**161.3**	**153.2**	**148.6**	**0.81**

Table 6: **Quantitative Results.** Trajectory deviation (Traj) and pose error (MPJPE) across seven 3D scenes on the GIMO [5]. Time horizons (0.5, 2.0, 5.0) are reported in (s). Best results are in **bold**.

Method	**Metric**	**Bedroom**			**Classroom**			**Garden**			**Kitchen**			**Lab**			**Livingroom**			**Seminarroom**		
		0.5	2.0	5.0	0.5	2.0	5.0	0.5	2.0	5.0	0.5	2.0	5.0	0.5	2.0	5.0	0.5	2.0	5.0	0.5	2.0	5.0
BiFu [5]	Traj	215	495	735	312	425	605	202	500	578	257	440	515	256	515	595	317	580	702	394	622	558
	MPJPE	**62.0**	106.5	162.6	84.1	136.0	151.3	83.2	193.0	223.5	91.5	100.7	118.1	**49.5**	106.2	203.2	**89.4**	119.8	210.2	100.7	131.0	161.5
SIF3D [17]	Traj	171	415	485	**145**	405	482	186	425	565	165	445	395	231	505	645	241	565	698	205	415	542
	MPJPE	77.5	98.4	162.6	93.3	138.0	141.3	**68.5**	176.2	175.1	93.1	97.4	101.5	61.4	88.1	191.9	102.2	**115.3**	166.2	94.7	131.1	158.6
GAP3DS [7]	Traj	168.5	**350**	415	179	391	413	183	**422**	472	163	**377**	368	235	465	560	173	**515**	576	174	375	531
	MPJPE	88.4	**93.8**	152.4	93.2	117.3	125.2	84.2	**164.3**	169.0	84.4	**85.2**	88.6	53.6	87.9	140.0	96.2	119.0	155.8	96.8	119.4	153.1
Auxformer [42]	Traj	171	519	1022	295	424	465	211	592	851	162	510	552	232	675	977	140	588	905	252	569	785
	MPJPE	106.1	168.3	228.5	128.0	169.9	143.5	97.5	217.6	203.1	119.1	108.5	112.6	105.3	132.1	205.2	146.3	127.5	230.1	122.9	151.8	181.5
MDP [27]	Traj	**154**	515	592	165	415	506	181	455	565	155	425	390	244	575	602	230	575	738	**158**	481	561
	MPJPE	107.0	169.0	247.6	89.4	155.8	144.5	94.3	177.2	185.3	85.9	100.0	110.5	74.5	**87.4**	203.9	95.7	124.1	185.9	96.7	133.0	166.1
EgoHMP (Full)	Traj	163	351	**403**	157	**278**	**352**	**170**	435	**461**	**155**	395	**359**	**221**	**430**	**540**	**161**	528	**554**	166	**352**	**492**
	MPJPE	86.0	94.0	**146.1**	**82.5**	**95.2**	**119.9**	80.0	166.2	**168.4**	**81.4**	85.9	**88.1**	53.5	**87.4**	**132.7**	89.6	126.9	**144.7**	**91.1**	**108.0**	**136.5**

the motion disparities to zero as well. Furthermore, 2D visual observations and 3D head kinematics inherently reside in highly heterogeneous semantic spaces. Forcing joint optimization without prior feature alignment leads to severe multi-modal feature collision, thereby destroying the discriminability of the feature space and providing ambiguous, noisy cues to the downstream generation modules.

Vision-Motion Alignment Loss. Table 5 (b) and (c) demonstrates the indispensable and complementary roles of our composite objectives. Discarding the InfoNCE contrastive loss (w/o nce) leads to a failure in cross-modal retrieval. This confirms that InfoNCE provides the primary driving force for establishing instance-level alignment between visual and kinematic signals. Conversely, relying solely on InfoNCE without the KL divergence loss (w/o kl) yields high retrieval metrics but severely degrades downstream motion generation and physical plausibility. Without the distribution regularization provided by KL divergence, the latent feature manifold becomes sharply overfitted to the retrieval objective, failing to provide the smooth and continuous prior essential for the downstream motion generator. Ultimately, the full composite loss achieves a profound synergy: KL divergence smooths the latent space to support robust generation, which in turn alleviates feature overfitting.

B.2 Analysis of Prediction Horizons

We evaluate the performance of EgoHMP against several baselines across various prediction horizons. As illustrated in Table 6, while most methods maintain reasonable accuracy in short-term forecasting due to physical inertia, baseline performance deteriorates significantly as the temporal horizon expands, leading to substantial error accumulation. In contrast, EgoHMP achieves superior performance across all metrics in both mid- and long-term horizons. This improvement is attributed to our construction of precise interaction probability maps, which effectively mitigates the error accumulation that plagues other methods due to inaccurate initial intent localization.

B.3 Discussion on Stochastic and Deterministic Paradigms

In this section, we provide a comprehensive discussion of stochastic and deterministic human motion prediction (HMP) methods to contextualize EgoHMP within existing paradigms.

Stochastic Scene-Agnostic Paradigms. Methods such as DLow [45], SmoothDMP [46], BeLFusion [41], and STAB-ACB [47] prioritize modeling the multimodal distribution of future poses. By

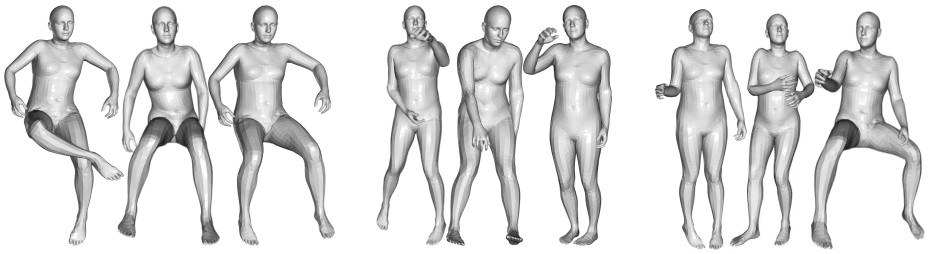


Figure 5: **Qualitative Visualization.** Given specific interaction intents, EgoHMP synthesizes diverse poses, confirming effective exploration of the valid interaction subspace without mean pose collapse.

leveraging generative backbones (e.g., VAEs, Diffusion, or Flow-based models), they synthesize diverse motion hypotheses based primarily on historical sequences. While achieving high diversity scores, these methods often function as "blind" generators. Due to the absence of explicit grounding in 3D scene geometry, the generated diversity frequently manifests as physically implausible random drifts or collisions. In essence, they prioritize kinematic variety at the expense of interaction fidelity.

Deterministic and Regression-based Scene-Aware Paradigms. Conversely, methods like Auxformer [42], MDP [27], BiFu [5], and SIF3D [17] integrate scene geometry to guide prediction through regression-based decoders. While these methods achieve better stability, they suffer from mean pose problem. By mapping complex poses to a single deterministic sequence, they tend to converge to the most generic interaction, failing to capture the rich variations of human behavior.

Reconciling Stochastic and Deterministic Paradigms. To effectively reconcile the trade-off between stochastic exploration and deterministic guidance, recent advanced methods like GAP3DS [7] and our EgoHMP adopt a two-stage, intent-guided generation paradigm. Unlike stochastic methods, this paradigm first establishes a deterministic global trajectory derived from predicted intent, effectively pruning task-irrelevant hypotheses. Building upon this, the diffusion-based decoder functions as a stochastic engine to synthesize kinematically diverse poses compliant with the intent. This paradigm allows model to explore effective interaction subspaces without collapsing into a singular mean pose.

Dataset: CIRCLE [48] comprises 10 hours of high-fidelity full-body motion sequences captured from 5 subjects across nine apartment scenes. Leveraging a 12-camera Vicon system at 120 FPS integrated with the AI Habitat VR environment, it ensures precise synchronization between human motion and scene geometry. The dataset utilizes integrated apartment meshes, treating each room as a distinct scene, and contains over 7,000 sequences covering 128 diverse tasks.

Baselines. We benchmark EgoHMP against five methods categorized into two paradigms. To represent purely stochastic scene-agnostic methods, we select DLow [45] (Flow-based), SmoothDMP [46] (VAE-based), BeLFusion [41] (Diffusion-based), and STAB-ACB [47] (Memory-based), which encompass the state-of-the-art in diverse motion generation. To represent the paradigm that reconciles stochastic exploration with deterministic guidance, we select the GAP3DS [7].

Metrics. Following DLow [45], we quantify the prediction diversity through the average pairwise distance (APD), computed as the mean L2 distance between the generated samples. To evaluate reconstruction accuracy, we employ average displacement error (ADE) and final displacement error (FDE) to measure whole-sequence and endpoint quality, respectively. Furthermore, we report MMADE and MMFDE, to assess the fidelity of the best-matched hypothesis in diverse scenarios.

Results and Analysis. As shown in Table 7, purely stochastic and scene-agnostic methods (e.g., SmoothDMP, STAB-ACB) achieve high diversity scores by unconstrainedly exploring the latent motion space. However, they suffer from severe interaction infidelity. Despite advancements in diffusion (BeLFusion) and memory-based priors that

Table 7: **Quantitative Results.** ↑: higher is better, ↓: lower is better. The best results are highlighted in **bold**.

Method	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓
DLow [45]	49.37	13.49	14.95	11.70	14.46
SmoothDMP [46]	**57.38**	11.31	13.57	9.75	12.04
BeLFusion [41]	39.46	9.56	10.04	7.91	9.30
STAB-ACB [47]	52.11	11.49	13.95	11.06	12.14
GAP3DS [7]	42.47	5.37	6.51	4.85	5.62
EgoHMP (Full)	44.10	**4.98**	**6.17**	**4.12**	**5.08**

Table 8: **Quantitative Results.** VRAM denotes the GPU memory consumption (in MB), and Speed represents the inference speed (in samples/s). The best results are highlighted in **bold**.

Metric	Vision Backbone (with 16384 points)					Scene Points (with DINOv2-L)			
	ResNet-50 [49]	CLIP [38]	DINOv2-B [37]	DINOv2-L [37]	DINOv3-L [50]	4096	8192	16384	32768
T-P (↓)	572	534	512	481	**479**	538	511	481	**480**
M-P (↓)	138.6	133.2	124.5	114.9	**113.2**	129.4	121.5	114.9	**114.3**
VRAM (↓)	**12100**	12650	12680	14610	14850	**7840**	11120	14610	18067
Speed (↑)	**17.5**	16.8	16.7	15.1	14.8	**27.4**	20.8	15.1	8.9

enhance kinematic consistency, their elevated deterministic (ADE/FDE) and multimodal errors (MMADE/MMFDE) confirm that internal motion priors are insufficient without explicit environmental grounding. In essence, they function as "blind" generators, producing physically implausible random drifts to inflate diversity. In contrast, the intent-guided paradigms (GAP3DS and EgoHMP) drastically reduce displacement errors. Notably, EgoHMP achieves state-of-the-art performance across all fidelity metrics. EgoHMP uses egocentric visual and motion cues to construct a precise intent-driven trajectory. This robust trajectory prevents the model from diverging into invalid physical spaces. Consequently, EgoHMP maintains competitive local pose diversity while strictly confining the generative manifold to the true intent distribution. As demonstrated in Figure 5, given specific interaction intents (e.g., sitting, reaching, or grasping), EgoHMP synthesizes diverse poses, confirming effective exploration of the valid interaction subspace without mean pose collapse.

B.4 Analysis of Vision Backbone and Computational Efficiency

Vision Backbone. To investigate the impact of visual feature extractors on EgoHMP, we evaluate various backbones at a fixed scale of 16,384 scene points. As shown in the left section of Table 8, standard supervised architectures like ResNet-50 [49] struggle to capture the fine-grained semantic required for human-scene interactions, yielding the highest trajectory and pose errors. While vision-language models like CLIP [38] provide stronger high-level semantics, they lack dense geometric understanding. Notably, employing self-supervised foundation models (DINOv2 [37]) significantly enhances predictive performance. Scaling from Base (DINOv2-B) to Large (DINOv2-L) further reduces the prediction error, albeit with a moderate increase in VRAM. Although the latest DINOv3-L [50] achieves the best prediction accuracy, its performance gain over DINOv2-L is marginal compared to the additional computational overhead. Therefore, to achieve the optimal trade-off between accuracy and efficiency, we adopt DINOv2-L as our vision backbone.

Scene Points. As shown in the right section of Table 8, we further explore the impact of the sampled scene point number on model accuracy and inference performance. As the number of points increases from 4,096 to 16,384, the model's perception of the 3D environment improves significantly, leading to a sharp decrease in prediction errors. However, this performance gain eventually saturates. Forcing the point count to 32,768 yields only a marginal improvement in fidelity, but introduces a severe computational penalty and inference speed drops drastically. Consequently, we employ 16,384 points to strike the optimal balance between generative accuracy and inference efficiency.

C More Visualizations

In this section, we provide more visualization results to demonstrate the capability of EgoHMP across various complex scenarios. Figure 6 illustrates the generated motion sequences in five scene types: lab, bedroom, kitchen, livingroom, and seminarroom.

As shown in the top view and side view, our model effectively perceives the 3D scene geometry and generates motions that strictly adhere to environmental constraints. In the kitchen scenario, the model successfully predicts a path through a narrow corridor between counters, avoiding collisions with the surrounding cabinets. Similarly, in the seminarroom, the generated character navigates alongside a long conference table, maintaining a natural distance from the furniture. The pose sequence strips at the bottom of each sub-figure demonstrate the kinematic quality of the predictions. The transition from the historical observation (blue meshes) to the predicted future (green meshes) is seamless, exhibiting no jitter or discontinuity. Whether the action involves straight walking or complex turning maneuvers, the generated poses maintain contextual consistency and physical plausibility.

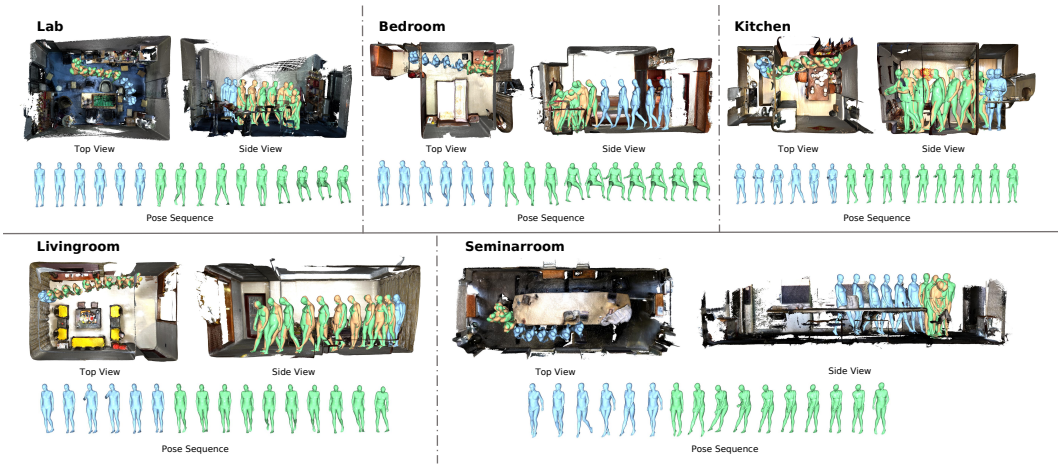


Figure 6: **Qualitative Visualization.** The blue, green, and tan human meshes represent the historical, predicted, and ground-truth motion sequences, respectively. Zoom in for better visualization.

D Potential Broader Impacts

Enhanced Human-Robot Collaboration. EgoHMP can improve human-robot collaboration. By accurately predicting human motions, collaborative robots can proactively avoid collisions and mitigate unsafe interactions, thereby fostering safer environments for both humans and machines.

Improved AR/VR and Immersive Experiences. In augmented and virtual reality, this framework can drive more realistic and responsive interactions with virtual elements. By deeply understanding and predicting how a user might move and interact within a scene, AR/VR systems can offer highly immersive, natural experiences that seamlessly align with user expectations, ultimately enhancing user engagement.

Advancements in Wearable Assistive Technologies. For elderly individuals or people with physical disabilities, wearable assistive devices equipped with EgoHMP can anticipate daily needs or critical actions (e.g., reaching for an object or potential loss of balance). This capability enables timely physical or cognitive assistance, enhancing users' independence and overall quality of life.

Applications in Autonomous Systems. Integrating robust human motion prediction frameworks into broader autonomous systems (such as embodied AI and intelligent vehicles) can improve safety and traffic management. By proactively anticipating pedestrian or user movements, these autonomous agents can navigate complex, dynamic urban and indoor environments more safely and efficiently.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims and contributions outlined in the Abstract and Introduction (Section 1) are explicitly supported by the methodological formulation in Section 3 and the comprehensive experimental evaluations presented in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We analyze the limitation of the proposed EgoHMP in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper focuses on network design and evaluation for egocentric human motion prediction, and does not claim any formal theoretical results or mathematical proofs. The main results of this paper are experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All necessary information for reproduction, including network architecture, datasets, evaluation metrics, and training hyperparameters, is provided in Section 4 and Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplemental material provides the code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All necessary information is provided in Section 4 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The performance margins between our method and the baselines are enough to demonstrate the effectiveness of our proposed approach.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Appendix A and Table 8.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have ensured that our research conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to Appendix D for a detailed discussion on the potential broader impacts of EgoHMP.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not present any of the aforementioned risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please refer to the Reference.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: There is no crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: There is no research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this paper does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.