SGIP: Learning Scene-Grounded Interaction Priors for Scene-Aware Human Motion Prediction

Donghyeon Soon[1]* **Jiil Yoon**[1]* **Hyeondong Kim**[1] **Daehee Park**[1,2]†
[1]DGIST [2]UNIVA
{dhsoon,wldlf2819,hdkim,dhpark}@dgist.ac.kr

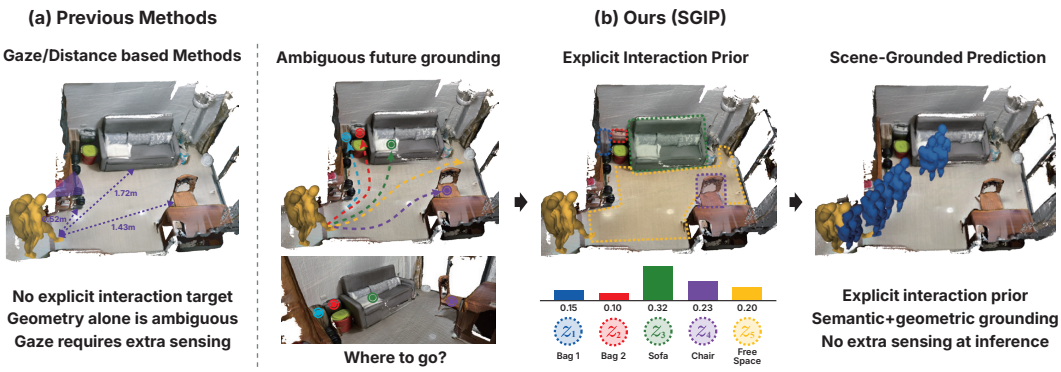


Figure 1: **Overview of the proposed method.** Scene-aware human motion prediction (HMP) requires resolving *future-motion ambiguity*. **(a)** Previous methods handle this ambiguity *implicitly*, relying on motion–scene distance or gaze without an explicit interaction target; geometry alone is ambiguous when several objects lie at similar distances, and gaze requires extra sensing. **(b)** Given the same observed motion, a person may approach a bag, sit on the sofa or chair, or freely navigate the room. *SGIP* resolves this by predicting an explicit *scene-grounded interaction prior*, a distribution over candidate scene tokens covering both object instances and free space, from the observed motion and 3D scene alone. Distilled at training time from an intent-aware teacher that combines semantic and geometric scores, the prior places the highest probability on the sofa token, which conditions both trajectory and full-body pose prediction to correctly anticipate walking toward and sitting on the sofa, with no extra sensing at inference.

Abstract

Human motion prediction (HMP) aims to forecast future movement from an observed motion sequence, and recent methods have incorporated 3D scene geometry and object semantics to improve long-term plausibility. However, the destination of future motion remains ambiguous in cluttered scenes: the same observed motion can lead to different interaction targets or navigation modes. Existing methods either handle this ambiguity implicitly or rely on auxiliary signals like gaze that are hard to obtain in practice. To address this, we propose *SGIP*, which explicitly predicts where in the scene future motion will be grounded across both objects and open navigation regions, and uses it to condition future trajectory and full-body pose prediction. This scene-grounded interaction prior is learned via intent distillation: *at training time only*, a teacher model scores candidate scene regions by combining an LLM prompted with a textual description of the person's intent with geometric motion-scene features. A student model learns to reproduce these

*Equal contribution.
†Corresponding author.

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

scores from only the observed motion and 3D scene, making *intent and LLM unnecessary at test time*. Across multiple benchmark datasets, SGIP outperforms prior scene-aware baselines on both trajectory and pose accuracy in a realistic deployment setting where neither gaze nor intent is available at inference, demonstrating the effectiveness of the proposed framework. The project page is available at `https://sgip2026.github.io`.

1 Introduction

Human motion prediction (HMP) aims to forecast future human poses and trajectories from an observed motion sequence, and is essential for autonomous systems, embodied agents, and human–robot collaboration [17, 40, 41]. While early approaches largely modeled motion dynamics in isolation [41, 45], recent HMP has shifted toward 3D scenes, where predictions must respect both the physical layout and the semantic affordances of the environment [20, 56]. These scene-aware methods improve long-term plausibility by incorporating human–object context, future contact, and gaze-based intention cues [10, 40, 44, 66].

However, future motion in such scenes is inherently ambiguous: the same observed motion can correspond to multiple valid ways of interacting with or navigating through the scene. For example, a person approaching a dining area may sit on one of several chairs or pass through the space toward another room; each possibility induces a different trajectory, body orientation, and contact pattern. Prior work has attempted to resolve this ambiguity by relying on auxiliary cues such as gaze [40, 66] or manually specified interaction targets [19, 59]. However, such cues require specialized sensors or per-frame supervision and are difficult to obtain in deployment. A practical HMP model should therefore make scene-aware predictions from the observed motion and 3D scene alone (Figure 1).

To address this, we propose SGIP, which makes scene-interaction reasoning an explicit, supervisable target by predicting a scene-grounded interaction prior, a distribution over candidate scene tokens covering both interactive objects and navigable regions. SGIP predicts this prior from the observed motion and 3D scene alone, and uses it to condition future trajectory and full-body pose prediction. However, learning the prior directly from observation is challenging since the observed motion alone often does not specify a clear interaction target. We address this through *privileged intent distillation*. At training time, we leverage a textual description of the person's intent (e.g., "take some rest"), which is practical to obtain through a vision-language model rather than requiring per-sample human annotation. A teacher prompts a large language model with this intent to score how well each candidate scene region aligns with the person's goal. Yet textual intent alone is insufficient when a scene contains multiple instances of the same object class, where the LLM cannot distinguish which specific instance the person is heading toward. We therefore combine the LLM's semantic score with a geometric compatibility score derived from motion–scene distance and orientation, which uses the observed motion to disambiguate between same-class instances. Together, these form the teacher's intent-conditioned distribution over scene tokens. At inference, where neither textual intent nor an LLM is available, a lightweight MLP-based student reproduces this distribution from motion and scene alone, inheriting the teacher's intent-aware scene-interaction reasoning through distillation.

Our contributions are summarized as follows: (1) We introduce a *scene-grounded interaction prior*, an explicit distribution over candidate scene tokens covering both object-centric interactions and free-space navigation, that turns scene-interaction reasoning into a supervisable target. (2) We propose a *privileged intent distillation* framework in which textual intent and an LLM supervise a teacher at training, while the student reproduces the resulting prior from motion and scene alone, requiring no language model or auxiliary cue at inference. (3) On GIMO and PROX, the observation-only student consistently improves trajectory and pose accuracy over scene-aware baselines without auxiliary cues, with the privileged teacher additionally serving as a performance ceiling.

2 Related Work

2.1 Human Motion Prediction

Human motion prediction (HMP) forecasts future human poses or trajectories from observed motion and underpins applications such as autonomous driving, human–robot collaboration, and animation [3,

13, 29, 41, 45, 50]. Architectures span recurrent [13, 45], graph-based [11, 37, 41, 52], and attention- or Transformer-based [2, 42, 62, 63] formulations, with probabilistic variants (VAEs, flows, GANs, diffusion) modeling multi-future ambiguity [4, 6, 9, 43, 51, 67]. Yet motion-only predictors overlook the surrounding 3D environment and frequently violate scene constraints in deployment. Scene-aware HMP addresses this by injecting scene context, evolving from 2D image- or bounding-box-based formulations [7, 10] to point-cloud fusion [26, 44, 73], typically built on PointNet-style backbones [47, 48] with diverse scene-grounded predictors [14, 57, 61]; gaze further serves as a behavioral proxy for intent in cluttered scenes [40, 66, 73]. Recent methods derive intention features from gaze to drive separate trajectory and pose branches [38], or model human–human and human–scene interactions in dynamic multi-person environments [53] captured in datasets such as HIK [54] and HOI-M^3 [69]. A complementary line synthesizes scene-aware motion conditioned on object affordances [20, 32, 33, 36, 56] or textual action descriptions and activity labels [8, 58, 59], rather than forecasting motion from observed histories.

2.2 Language Models for Scene and Motion Understanding

Large language models (LLMs) have recently been used to inject commonsense knowledge into 3D scene and motion tasks. Motion–language alignment models adapt LLMs to translate textual descriptions into human motion sequences [31], and scene-aware variants further ground the generated motion in 3D environments through textual scene or activity descriptions [8, 58, 59, 65]. Closely related, several methods route language conditions through explicit *affordance* or *intent* representations as intermediate structure between language and motion [12, 16, 60], reflecting a broader shift toward semantically structured intermediates rather than direct text-to-motion mapping. Beyond generation, LLMs serve as reasoners for 3D scene and human–scene understanding by aligning object semantics with language [24, 25, 70], and as grounding modules that connect language goals to spatial affordances for embodied agents [1]. A complementary line of work treats LLMs as judges or rubric-based scorers [72], where natural-language criteria elicit interpretable, score-level outputs over candidate options. Despite this flexibility, deploying LLMs at inference incurs substantial computational and latency cost, which is poorly suited to real-time motion forecasting in deployment.

2.3 Privileged Information and Knowledge Distillation

A natural alternative is to use such expensive or hard-to-obtain supervision only at training time. Learning with privileged information [39, 55] formalizes this setting by exploiting signals available during training but absent at inference, while knowledge distillation [22, 49] transfers such information from a teacher to a more deployable student through output- or intermediate-feature-level imitation. A concrete realization in vision is *modality hallucination* or *modality distillation*, where a privileged modality (e.g., depth or optical flow) supervises a student that operates from a single observable modality at inference [15, 23, 71]; in human trajectory forecasting, multi-modal teachers have similarly been distilled into students that observe fewer modalities [30]. In the context of human motion in 3D scenes, behavioral or semantic cues such as gaze [40, 66, 73], language goals [8, 59], or explicitly stated intent [16] can disambiguate human–scene behavior but are rarely accessible in practical deployment, making them natural candidates for privileged supervision. We adopt this paradigm with two privileged training-time signals—textual intent and an LLM-based rubric scorer—and distill the resulting *scene-grounded interaction prior* into an observation-only student, rather than transferring gaze- or language-conditioned motion outputs as in prior distillation.

3 Proposed Method

SGIP is a two-branch teacher–student framework (Figure 2). The teacher fuses an LLM-based rubric score with a geometric compatibility score into an intent-conditioned interaction prior, and a student approximates this prior from motion and scene alone, serving as the deployed model. Section 3.1 formalizes the problem; Section 3.2 defines the scene-grounded interaction prior and its teacher and student forms; Section 3.3 describes how the prior conditions trajectory and full-body pose prediction; and Section 3.4 describes the distillation procedure. Implementation details (VLM and LLM calls, hyperparameters, training schedule) are deferred to Appendices E–G.

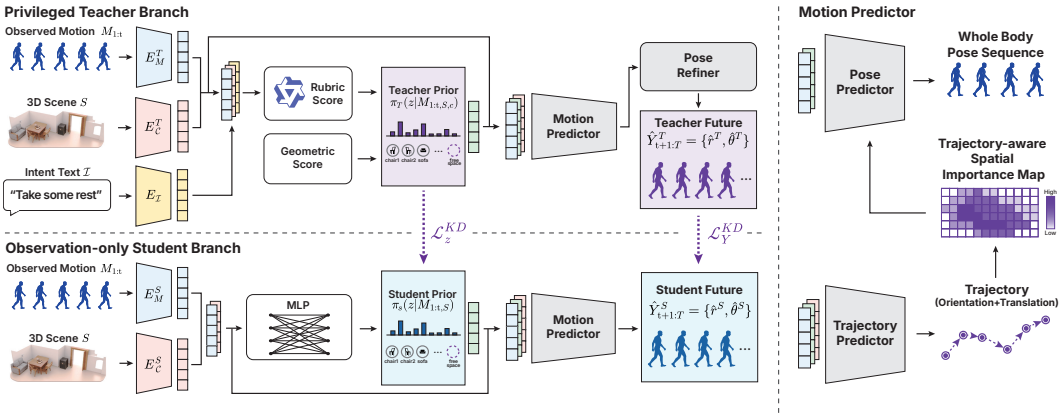


Figure 2: **Overall pipeline of *SGIP*.** A privileged teacher branch (top, training only) constructs an intent-conditioned prior $\pi^{\mathcal{T}}$ over candidate grounding tokens by fusing an LLM-based rubric score with a geometric compatibility score, with a Pose Refiner further intent-aligning the teacher's pose. An observation-only student branch (bottom) predicts $\pi^{\mathcal{S}}$ from motion $M_{1:t}$ and scene S alone via a scoring head $F_{\mathcal{S}}^{\text{grd}}$ (a permutation-equivariant MLP), trained via prior distillation $\mathcal{L}_z^{\text{KD}}$ and motion-level distillation $\mathcal{L}_Y^{\text{KD}}$. The selected token then conditions the motion predictor (right) that outputs future trajectory and full-body pose.

3.1 Problem Formulation

We are given an observed motion sequence $M_{1:t} = \{x_1, \ldots, x_t\}$, where each frame $x_\tau = (\boldsymbol{r}_\tau, \boldsymbol{\theta}_\tau)$ consists of root motion $\boldsymbol{r}_\tau = (\boldsymbol{t}_\tau, \boldsymbol{R}_\tau)$, with translation $\boldsymbol{t}_\tau \in \mathbb{R}^3$ and orientation $\boldsymbol{R}_\tau \in SO(3)$, together with SMPL-X body pose $\boldsymbol{\theta}_\tau$ [46]. The surrounding 3D environment is a point cloud $S \in \mathbb{R}^{N\times 3}$. Our goal is to predict the future motion $\hat{Y}_{t+1:T} = \{\hat{x}_{t+1}, \ldots, \hat{x}_T\}$ over a horizon $\Delta T = T - t$. To make scene-interaction reasoning explicit, we extract a set of candidate *grounding tokens* $\mathcal{C} = \mathcal{C}^{\text{obj}} \cup \mathcal{C}^{\text{nav}}$, where $\mathcal{C}^{\text{obj}}$ contains object instances and $\mathcal{C}^{\text{nav}}$ is a single scene-global token representing object-agnostic navigable regions. Since multiple tokens in $\mathcal{C}$ can afford plausible futures from the same observed motion, we introduce a *scene-grounded interaction prior* π over $\mathcal{C}$:

$$\pi_i = p(z = i \mid M_{1:t}, S), \quad i \in \mathcal{C}, \tag{1}$$

where the latent z indexes the token on which future motion is grounded.

At training time, we obtain a textual intent $\mathcal{I}$ (e.g., *"take some rest"*) by applying a frozen vision-language model (VLM) to training videos and corresponding 3D motion visualizations (Appendix E.1). The intent and an LLM-based rubric scorer construct a privileged teacher prior $\pi^{\mathcal{T}}$, which is distilled into the observation-only student prior $\pi^{\mathcal{S}}$. The most probable token under $\pi^{\mathcal{S}}$ conditions the motion predictor, augmented at training with prior- and motion-level distillation losses against the teacher. At inference, $\mathcal{I}$, the VLM, and the LLM are discarded.

3.2 Scene-Grounded Interaction Prior

We model future-motion grounding as a categorical distribution π over the candidate set $\mathcal{C}$, and construct two views of this distribution: a privileged *teacher* prior $\pi^{\mathcal{T}}$ that exploits intent and an LLM at training time, and an observation-only *student* prior $\pi^{\mathcal{S}}$ that approximates $\pi^{\mathcal{T}}$ from motion and scene alone at inference.

Token representations. Object instances $\mathcal{C}^{\text{obj}}$ are extracted via zero-shot 3D instance segmentation [27] on S. We retain object candidates that intersect a 120° forward-facing cone whose axis bisects the observed head- and torso-facing directions, aggregated over the observation window. This pruning uses only the observed motion (no gaze) and is shared by the teacher, the student, and all SGIP ablations (Appendix D.2).

Each token $c_i = (\boldsymbol{\phi}_i, \boldsymbol{e}_i, \boldsymbol{m}_i)$ consists of three representations that capture complementary aspects of the token:

$$\boldsymbol{\phi}_i = [\boldsymbol{\phi}_{\text{geo}}(c_i),\ \boldsymbol{\phi}_{\text{rel}}(c_i, M_{1:t})] \qquad \text{(geometric descriptor)}, \tag{2}$$

$$\boldsymbol{e}_i = E_{\mathcal{C}}(c_i) \qquad \text{(token embedding)}, \tag{3}$$

$$\boldsymbol{m}_i \subseteq S \qquad \text{(point-cloud mask)}. \tag{4}$$

$\boldsymbol{\phi}_i$ captures where the token is and how the observed motion approaches it, which identifies the specific instance among same-class candidates being headed toward; $\boldsymbol{e}_i$ is a learned semantic representation of the interaction the token affords; and $\boldsymbol{m}_i$ provides its 3D shape, grounding physically plausible body configurations at the target. The observed motion sequence is encoded into a single embedding $\boldsymbol{h}_M^b = E_M^b(M_{1:t}) \in \mathbb{R}^{d_M}$ for branch $b \in \{\mathcal{T}, \mathcal{S}\}$, where E_M^b denotes the branch's motion encoder.

Teacher prior: combining semantic and geometric evidence. At training, the teacher constructs $\pi^{\mathcal{T}}$ by combining two complementary scores. A *semantic relevance score* obtained from an LLM rates each candidate against the textual intent $\mathcal{I}$:

$$s_i^{\text{rubric}} = \text{LLM}(\mathcal{I}, c_i), \quad i \in \mathcal{C}, \tag{5}$$

where the LLM is queried with a fixed prompt scoring two criteria—intent affordance (how well a candidate semantically supports the intent) and intent usability (how easily the intent can be realized using that candidate, given the scene). The rubric, LLM call, and output validation are given in Appendix E.3, and the intent taxonomies in Appendix E.2.

A *geometric compatibility score* aggregates fixed directional alignments between the observed motion and each candidate (details in Appendix D.5):

$$s_i^{\text{geo}} = w_{\text{face}}\, \text{align}_{\text{face}}(\boldsymbol{\phi}_i, M_{1:t}) + w_{\text{body}}\, \text{align}_{\text{body}}(\boldsymbol{\phi}_i, M_{1:t}) + w_{\text{line}}\, \text{align}_{\text{line}}(\boldsymbol{\phi}_i, M_{1:t}). \tag{6}$$

The two scores are complementary. The LLM captures intent–class affordance (*a chair affords sitting*) but cannot disambiguate among multiple instances of the same class within a scene; the geometric score grounds the semantic prior in the observed motion, identifying *which* specific instance is being approached. We fuse them via temperature softmax with fusion weight α and temperature τ:

$$\pi_i^{\mathcal{T}} = \frac{\exp\big((s_i^{\text{rubric}} + \alpha\, s_i^{\text{geo}})/\tau\big)}{\sum_{j \in \mathcal{C}} \exp\big((s_j^{\text{rubric}} + \alpha\, s_j^{\text{geo}})/\tau\big)}. \tag{7}$$

This yields an intent-conditioned prior that the student is trained to match.

Student prior: observation-only approximation. At inference, neither $\mathcal{I}$ nor the LLM is available, so the student reproduces π from motion and the geometric token descriptors alone through a learnable scoring head $F_{\mathcal{S}}^{\text{grd}}$:

$$\pi_i^{\mathcal{S}} = \frac{\exp F_{\mathcal{S}}^{\text{grd}}(\boldsymbol{h}_M^{\mathcal{S}}, \boldsymbol{\phi}_i)}{\sum_{j \in \mathcal{C}} \exp F_{\mathcal{S}}^{\text{grd}}(\boldsymbol{h}_M^{\mathcal{S}}, \boldsymbol{\phi}_j)}. \tag{8}$$

Because $\mathcal{C}$ varies in size across scenes and tokens have no canonical order, $F_{\mathcal{S}}^{\text{grd}}$ is implemented as a permutation-equivariant MLP over candidates [68]. The student is trained to match $\pi^{\mathcal{T}}$ via knowledge distillation (Section 3.4), inheriting the teacher's intent-aware grounding without any auxiliary signal at inference.

3.3 Prior-Conditioned Motion Prediction

Given the prior π^b from branch $b \in \{\mathcal{T}, \mathcal{S}\}$, the motion predictor D^b produces future motion $\hat{Y}_{t+1:T}^b = \{\hat{\boldsymbol{r}}_{t+1:T}^b, \hat{\boldsymbol{\theta}}_{t+1:T}^b\}$ conditioned on the most probable grounding token,

$$i_b^* = \arg\max_{i \in \mathcal{C}} \pi_i^b, \tag{9}$$

whose embedding $\boldsymbol{e}_{i_b^*}$ and point-cloud mask $\boldsymbol{m}_{i_b^*}$ are read off from the selected token $c_{i_b^*}$. For multi-future prediction, we separately decode the top-K tokens ranked by $\pi^{\mathcal{S}}$ while keeping the observed motion, scene, and model parameters fixed. Only the selected token's embedding and mask change, yielding K token-conditioned future motions without stochastic sampling or privileged inputs. As shown on the right side of Figure 2, D^b follows a two-stage trajectory-then-pose paradigm [40].

Trajectory branch. The trajectory branch first encodes the 3D scene S into per-point features via a scene encoder E_S, then cross-attends the motion embedding $\boldsymbol{h}_M^b$ to these features while conditioning on the selected token's embedding $\boldsymbol{e}_{i_b^*}$, which steers the predicted path toward the chosen interaction target (Appendix F.2):

$$\hat{\boldsymbol{r}}_{t+1:T}^b = D_{\text{traj}}^b(\boldsymbol{h}_M^b, S, \boldsymbol{e}_{i_b^*}). \tag{10}$$

The predicted trajectory then induces a spatial importance map P^b aggregating path proximity, forward-facing alignment, and proximity to the selected token's region:

$$P^b = f_{\text{spatial}}(\hat{\boldsymbol{r}}_{t+1:T}^b, S, \boldsymbol{m}_{i_b^*}). \tag{11}$$

Pose branch. P^b is added as a bias to the pose branch's scene-attention logits, keeping body pose consistent with the planned path and interaction target:

$$\tilde{\boldsymbol{\theta}}_{t+1:T}^b = D_{\text{pose}}^b(\boldsymbol{h}_M^b, \hat{\boldsymbol{r}}_{t+1:T}^b, S, P^b). \tag{12}$$

Architectural details (f_{spatial}, encoders, attention blocks) are in Appendix F. For the teacher only, a pose refiner Refine that leverages $\mathcal{I}$ is applied to recover intent-conditioned pose semantics beyond motion reconstruction (e.g., distinguishing sitting from standing in front of the same chair); details are in Appendix F.3.

3.4 Privileged Intent Distillation

To deploy the teacher's intent-aware reasoning at inference, we distill it into the student via two complementary signals: *prior distillation* aligns *where* future motion is grounded ($\pi^{\mathcal{T}} \to \pi^{\mathcal{S}}$), and *motion-level distillation* aligns *how* the model moves toward that target ($\hat{Y}^{\mathcal{T}} \to \hat{Y}^{\mathcal{S}}$).

Prior distillation. We align $\pi^{\mathcal{S}}$ with $\pi^{\mathcal{T}}$ via forward KL divergence:

$$\mathcal{L}_z^{\text{KD}} = D_{\text{KL}}\big(\text{sg}[\pi^{\mathcal{T}}] \,\|\, \pi^{\mathcal{S}}\big) = \sum_{i \in \mathcal{C}} \pi_i^{\mathcal{T}} \log \frac{\pi_i^{\mathcal{T}}}{\pi_i^{\mathcal{S}}}, \tag{13}$$

where $\text{sg}[\cdot]$ stops gradients to $\pi^{\mathcal{T}}$. The mass-covering forward direction forces the student to cover every token the teacher considers, preserving the full intent-conditioned grounding structure rather than only its top-1 and preventing mode collapse.

Motion-level distillation. We distill the teacher's motion via L_1 loss on trajectory and pose:

$$\mathcal{L}_Y^{\text{KD}} = \lambda_{\text{traj}} \big\| \hat{\boldsymbol{r}}_{t+1:T}^{\mathcal{S}} - \text{sg}[\hat{\boldsymbol{r}}_{t+1:T}^{\mathcal{T}}] \big\|_1 + \lambda_{\text{pose}} \big\| \tilde{\boldsymbol{\theta}}_{t+1:T}^{\mathcal{S}} - \text{sg}[\hat{\boldsymbol{\theta}}_{t+1:T}^{\mathcal{T}}] \big\|_1, \tag{14}$$

where $\hat{\boldsymbol{\theta}}_{t+1:T}^{\mathcal{T}}$ is the refined teacher pose, and $\lambda_{\text{traj}}, \lambda_{\text{pose}}$ balance the two terms. Crucially, since the teacher pose is refined with intent, the intent-conditioned pose semantics (e.g., sitting versus standing in front of the same chair) can be distilled into the student.

Training schedule. The teacher is trained first with its motion loss $\mathcal{L}_{\text{mot}}^{\mathcal{T}}$ and additional Refine-specific terms (Appendix G). The student then freezes the entire teacher and optimizes

$$\mathcal{L}^{\mathcal{S}} = \mathcal{L}_{\text{mot}}^{\mathcal{S}} + \lambda_z \, \mathcal{L}_z^{\text{KD}} + \lambda_Y \, \mathcal{L}_Y^{\text{KD}}, \tag{15}$$

where $\mathcal{L}_{\text{mot}}^{\mathcal{S}}$ shares the form of $\mathcal{L}_{\text{mot}}^{\mathcal{T}}$.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate *SGIP* on GIMO [73], a real-world indoor dataset with full-body SMPL-X motion and reconstructed 3D scenes, following the standard scene-aware HMP protocol [40, 66, 73]: the model observes a motion prefix in a 3D scene and predicts the future global trajectory and full-body pose. We additionally evaluate on PROX [18], which has no gaze annotations, so gaze-using baselines are evaluated without gaze. For both datasets, training-time pseudo-intents are extracted

Table 1: Main comparison on GIMO and PROX. The "Auxiliary Cue" column denotes auxiliary signals beyond $M + S$: g=gaze, $\mathcal{I}$=textual intent. †: diagnostic privileged branch on GIMO (not deployed); $*$: our gaze-free re-evaluation of gaze-using baselines; ‡: gaze-free variant reported in the original paper. SGIP results on GIMO are mean±std over 5 seeds. **Bold**/underline: best/second-best within block.

	Method	Auxiliary Cue	Traj-path ↓	Traj-dest ↓	MPJPE-path ↓	MPJPE-dest ↓
GIMO	*Models with additional or privileged cues*					
	LTD [41] w/ gaze	g	655	801	155.6	202.0
	SPGSN [37] w/ gaze	g	702	982	156.0	202.0
	AuxFormer [62] w/ gaze	g	622	746	161.5	201.1
	BiFu [73]	g	661	727	167.8	205.0
	SIF3D [40]	g	590	666	156.6	195.7
	GAP3DS [66]	g	575	<u>623</u>	**141.2**	**171.4**
	HuMo3D [38]	g	543	651	<u>146.1</u>	<u>185.5</u>
	SGIP$^{\dagger}_{\text{priv}}$	g	<u>528</u>	635	149.9	190.6
	SGIP$^{\dagger}_{\text{priv}}$	$\mathcal{I}$	**485**	**556**	151.2	187.8
	Observation-only models					
	LTD [41]	–	673	890	<u>156.4</u>	203.0
	SPGSN [37]	–	700	893	157.0	202.2
	AuxFormer [62]	–	732	902	165.1	202.4
	SIF3D* [40]	–	700	823	164.2	<u>199.6</u>
	GAP3DS* [66]	–	<u>642</u>	<u>701</u>	170.1	206.4
	HuMo3D‡ [38]	–	<u>642</u>	731	178.3	210.1
	SGIP	–	**549** (±5.0)	**681** (±7.6)	**153.1** (±1.8)	**195.1** (±2.6)
PROX	LTD [41]	–	677	753	223.4	226.6
	SPGSN [37]	–	595	663	286.3	271.3
	AuxFormer [62]	–	553	727	<u>199.2</u>	240.8
	SIF3D* [40]	–	<u>488</u>	<u>659</u>	228.5	<u>217.4</u>
	SGIP	–	**470**	**576**	**194.6**	**197.7**

offline by a frozen VLM with dataset-specific taxonomies (Appendix E.1), and SGIP is trained and evaluated separately on each. We further compare SGIP in multi-human, dynamic scenes on HIK [54] and HOI-M^3 [69] following the HUMOF protocol [53]; these results are reported in Appendix B.5.

Baselines. *Pose-only* baselines (LTD [41], SPGSN [37], AuxFormer [62]) forecast from body dynamics alone. *Scene- and gaze-aware* baselines (BiFu [73], SIF3D [40], GAP3DS [66], HuMo3D [38]) condition on the scene with gaze; to match our deployment assumption, we also report their gaze-free variants. We additionally compare SGIP with a conditional diffusion-based motion generation model adapted to HMP in Appendix B.3. We report the deployed student as *SGIP* and the privileged teacher as *SGIP*$^{\dagger}_{\text{priv}}$ (diagnostic only); in the privileged variants, intent ($s_i^{\text{rubric}} + \alpha s_i^{\text{geo}}$) or gaze (proximity and alignment of each candidate to the final-frame 3D gaze point) enters only through the candidate scores, so all variants share the same predictor (Appendix B.4).

Metrics. *Traj-path*/*Traj-dest* measure global trajectory error averaged over the horizon and at the final frame, and *MPJPE-path*/*MPJPE-dest* the mean per-joint position error [28, 41] on the same horizons [40, 66, 73]. For multi-future evaluation, we report minADE@K/minFDE@K over K predictions, FDE@minADE (the final error of the prediction with minimum ADE), and the average pairwise distance (APD) for diversity.

Implementation. Object tokens are obtained via OpenIns3D [27]; intents are extracted with Qwen3-VL-8B [5] and rubric scores with Qwen3-14B [64], both run once offline and cached. After a one-time preprocessing of a static scene (61.7 s), the student runs at 132.8 ms per observation (7.5 FPS) on one RTX 3090, below the 500 ms observation interval of GIMO (Appendix B.8).

4.2 Main Results

Among observation-only models, *SGIP* achieves the best result on all four metrics on both GIMO and PROX (Table 1), indicating that the gains do not stem from GIMO-specific tuning. Against the strongest observation-only baseline in each metric, SGIP reduces Traj-path by 93 mm on GIMO

Table 2: Multi-future evaluation on GIMO. SGIP separately decodes its top-K grounding tokens; root-trajectory errors and APD are in meters (Appendix B.3).

K	APD ↑	minADE@K ↓	minFDE@K ↓	FDE@minADE ↓
1	–	0.549	0.681	0.681
2	0.884	0.517	0.601	0.605
3	**0.927**	**0.467**	**0.523**	**0.571**

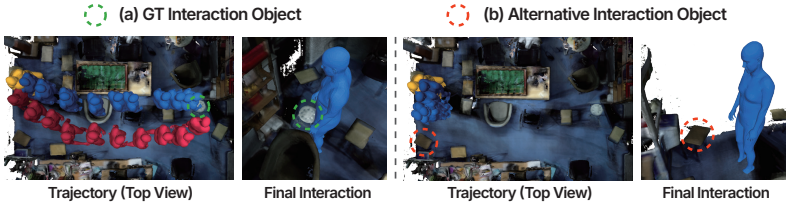


Figure 3: **Multimodal futures from $\pi^{\mathcal{S}}$.** We condition the same fixed student predictor $D^{\mathcal{S}}$ on two different grounding tokens drawn from $\pi^{\mathcal{S}}$: **(a)** the ground-truth interaction object (green dashed circle), and **(b)** an alternative high-probability token of $\pi^{\mathcal{S}}$ (red dashed circle). Each panel visualizes the resulting predicted trajectory (top-down view) and the final interaction pose; only the selected token $(\boldsymbol{m}_i, \boldsymbol{e}_i)$ fed into $D^{\mathcal{S}}$ varies.

(549 vs. 642 mm) and Traj-dest by 83 mm on PROX (576 vs. 659 mm). Pose-only methods exhibit large trajectory errors (673–732 mm Traj-path on GIMO), and scene-aware baselines still trail *SGIP*: scene-level conditioning alone does not resolve *which* object the future motion engages, and gaze encodes *where the subject looks* rather than *what the subject intends to do*. We attribute the gains to (1) the **explicit grounding prior**, whose selected token provides a spatial mask and semantic embedding that anchor the predictor to a specific target, yielding the largest gains on trajectory metrics where grounding ambiguity dominates; and (2) **privileged intent distillation**, which transfers intent-aware grounding and motion structure to the student, with $\textit{SGIP}^{\dagger}_{\text{priv}}$ serving as a supervision upper bound rather than a deployable model.

Privileged cues. With the same predictor, intent yields lower trajectory error than gaze (485/556 vs. 528/635 mm), indicating that knowing *what* the person intends is more informative for grounding than knowing where they look. GAP3DS remains stronger in privileged pose error (141.2/171.4 vs. 149.9/190.6 mm with gaze) because it feeds an explicitly generated interaction pose to its decoder, whereas SGIP's cue only determines *where* the interaction occurs (Appendix B.4).

4.3 Multi-Future Evaluation

We decode the top-K tokens of $\pi^{\mathcal{S}}$ separately, $K \in \{1, 2, 3\}$, with the observed motion, scene, and model parameters fixed. Increasing K from 1 to 3 reduces minADE/minFDE from 0.549/0.681 to 0.467/0.523 m and FDE@minADE from 0.681 to 0.571 m, while APD reaches 0.927 m (Table 2): lower-ranked tokens provide distinct alternatives, some closer to the ground truth than the top-1, and each alternative is associated with a ranked, interpretable scene token rather than obtained by stochastic sampling. Because FDE@minADE also decreases, the alternative that is closest on average also ends closer to the ground-truth destination, so the gain does not merely come from spreading predictions over more of the scene. As Figure 3 illustrates, conditioning $D^{\mathcal{S}}$ on another high-probability token yields a distinct, scene-consistent trajectory and arrival pose, i.e., $D^{\mathcal{S}}$ has learned a token-conditional motion distribution. A comparison with a conditional diffusion model under the same budget of three predictions is reported in Appendix B.3.

4.4 Ablation Studies

Distillation components. Without distillation, the student performs poorly on trajectory (Table 3a), confirming that scene conditioning alone does not resolve grounding ambiguity; prior distillation $\mathcal{L}_z^{\text{KD}}$ yields the largest single-component gain, reducing Traj-path/dest from 740/838 to 574/688 mm. Applying $\mathcal{L}_Y^{\text{KD}}$ alone *degrades* trajectory accuracy below the no-distillation baseline, which we

Table 3: Ablations on GIMO (5 seeds). **(a)** Distillation components $\mathcal{L}_z^{\mathrm{KD}}$ (prior) and $\mathcal{L}_Y^{\mathrm{KD}}$ (motion). **(b)** Hard vs. soft top-k token aggregation at inference with the same trained student. **(c)** Top-1 accuracy (%) on the GIMO-annotated target for teacher-side priors (R: rubric, G: geometry) and for the same student head distilled from different targets; bold marks the best within each group.

(a) Distillation components.

KD		Traj ↓		MPJPE ↓	
z	Y	path	dest	path	dest
		740	838	169.1	204.8
	✓	856	896	164.1	198.9
✓		574	688	**151.8**	198.6
✓	✓	**549**	**681**	153.1	**195.1**

(b) Hard vs. soft top-k.

Aggregation	Traj ↓		MPJPE ↓	
	path	dest	path	dest
Hard (top-1)	**549**	**681**	**153.1**	**195.1**
Soft (top-2)	584	736	158.3	210.3
Soft (top-3)	715	877	176.5	210.2

(c) Grounding accuracy (Acc@1, %).

	Teacher-side prior			Student (identical head), distillation target			
	R	G	$G{+}R$ ($\pi^{\mathcal{T}}$)	None	KL(R)	KL(G)	KL($G{+}R$)
Acc@1	54	51	**86**	41	48	51	**76**

attribute to *grounding misalignment*: the teacher's output is conditioned on its token $i^*_{\mathcal{T}}$, while a student without prior supervision selects an inconsistent $i^*_{\mathcal{S}}$, so imitating the teacher's motion pulls predictions toward incompatible targets. Consistently, the two branches agree on only 43% of Top-1 tokens under $\mathcal{L}_Y^{\mathrm{KD}}$ alone versus 70% once $\mathcal{L}_z^{\mathrm{KD}}$ is added (Appendix B.2). Prior distillation is thus a prerequisite for output-level distillation, which then further reduces trajectory error.

Hard vs. soft top-k aggregation. SGIP conditions the predictor on the single token selected by the argmax of $\pi^{\mathcal{S}}$ (Eq. 9). A natural alternative is to aggregate the top-k tokens softly: their masks and embeddings are averaged with the renormalized probabilities $\tilde{\pi}_i^{\mathcal{S}}$, i.e., $\bar{\boldsymbol{m}} = \sum_i \tilde{\pi}_i^{\mathcal{S}} \boldsymbol{m}_i$ and $\bar{\boldsymbol{e}} = \sum_i \tilde{\pi}_i^{\mathcal{S}} \boldsymbol{e}_i$ over the top-k set, and the same trained predictor consumes $(\bar{\boldsymbol{m}}, \bar{\boldsymbol{e}})$ in place of $(\boldsymbol{m}_{i^*_{\mathcal{S}}}, \boldsymbol{e}_{i^*_{\mathcal{S}}})$ (Appendix B.9). As shown in Table 3b, soft aggregation degrades all four metrics, and the degradation grows with k: Traj-path/dest increase from 549/681 to 584/736 mm with $k = 2$ and to 715/877 mm with $k = 3$, and MPJPE-dest rises from 195.1 to about 210 mm. When the top-ranked tokens correspond to spatially separated targets (e.g., two chairs on opposite sides of a table), their weighted mixture describes no physically existing target: the averaged embedding steers the trajectory branch toward a location between the candidates, and the mixed mask spreads the spatial importance map P^b (Eq. 11) over several disjoint regions, so the pose branch is no longer anchored to a single contact surface. Moreover, since the predictor is trained with hard tokens, the mixed conditioning lies outside the distribution it has seen during training. This mode-averaging behavior is precisely what explicit, discrete grounding is designed to avoid; the ambiguity among the top-ranked tokens is better exploited by decoding them separately as alternative futures (Section 4.3) than by blending them into a single prediction.

Attribution of grounding accuracy. On the GIMO-annotated target (used only for evaluation), fusing rubric and geometry (86%) clearly exceeds either alone (54%/51%; Table 3c). To separate intent-conditioned supervision from the learnable head and candidate preprocessing, we train the same student head with identical inputs, candidates, and architecture, varying only the distillation target: it reaches 41% without prior distillation and 51% when distilling the geometry-only prior, but 76% when distilling the fused, intent-conditioned prior. This 25-point gap under an identical architecture isolates the contribution of intent-conditioned supervision (Appendix B.1).

4.5 Qualitative Results

Figure 4 compares *SGIP* with the gaze-free SIF3D* on three cluttered GIMO scenes. In the kitchen, SGIP follows the ground-truth path toward the counter, whereas SIF3D* turns toward the refrigerator side and ends at a different location. In the living room, SIF3D* stalls near its starting position

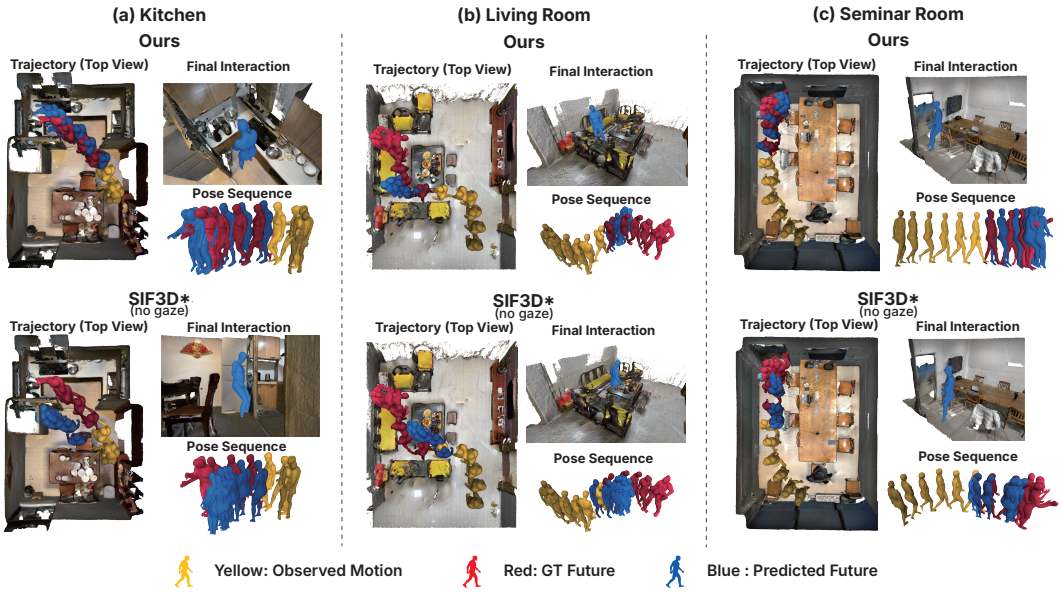


Figure 4: Qualitative comparison on three GIMO scenes (Kitchen, Living Room, Seminar Room). For each scene, the top row shows *SGIP* (ours) and the bottom row shows the gaze-free baseline SIF3D*. Each row visualizes the predicted trajectory (top-down view), the final interaction frame, and the full-body pose sequence. Yellow: observed motion; red: GT future; blue: predicted future.

between the table and chairs, while SGIP advances along the ground-truth direction toward the sofa. In the seminar room, both methods move along the wall, but SIF3D* falls short of the destination, whereas SGIP reaches the region where the ground-truth motion ends. The gap is largest when several regions could plausibly attract the future motion: by committing to an explicit grounding token, SGIP produces a trajectory and arrival pose consistent with a single target instead of drifting between candidates, mirroring the trajectory gains in Table 1. Moreover, even when its prediction departs from the recorded motion, SGIP commits to a different but reachable target and completes a coherent interaction with it, yielding an intent-consistent alternative rather than an averaged or stalled motion (Appendix C.2). Appendix C also compares the SGIP teacher and student with SIF3D (with and without gaze); videos are in the supplementary material.

5 Conclusion and Limitations

We presented *SGIP*, which makes human motion prediction explicitly grounded: it predicts where in the scene the future motion will arrive—at an object instance or in an open navigation region—and what it will interact with, and conditions trajectory and full-body pose prediction on this grounding. The grounding is distilled from an LLM-guided, intent-aware teacher only at training time, so the deployed model requires neither intent annotations nor an LLM at inference. In this sense, SGIP connects language-scale semantics to physical behavior: commonsense knowledge about what a space affords is transferred into a lightweight predictor that operates on the observed motion and 3D scene alone. On GIMO and PROX, SGIP outperforms scene-aware HMP baselines on both trajectory and pose accuracy without gaze or intent at inference, and provides accurate, interpretable alternative futures. We believe this matters beyond forecasting: robots that move in human living spaces must understand what can be done where, and must retain this understanding in deployment, where no language annotation is available; scene-grounded predictions of this kind could also turn recorded human motion into a useful teacher for learning humanoid full-body locomotion.

Limitations. SGIP depends on the quality of candidate scene tokens (Appendix B.6), requires a one-time instance segmentation for each new scene (Appendix B.8), and represents neighboring humans as static tokens, which limits pose accuracy in multi-human scenes (Appendix B.5). Modeling time-varying human and object states and sampling futures directly from the full predicted prior are natural next steps.

Acknowledgments and Disclosure of Funding

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2022.

[2] 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)*, 2021.

[3] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5223–5232, 2020.

[5] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[6] 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 (ICCV)*, pages 2317–2327, 2023.

[7] 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 (ECCV)*, 2020.

[8] Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3D scenes from text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1855–1866, 2024.

[9] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. HumanMAC: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9544–9555, 2023.

[10] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6992–7001, 2020.

[11] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. MSR-GCN: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11467–11476, 2021.

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

[13] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[14] Xuehao Gao, Yang Yang, Yang Wu, Shaoyi Du, and Guo-Jun Qi. Multi-condition latent diffusion network for scene-aware neural human motion prediction. *IEEE transactions on Image Processing (TIP)*, pages 3907–3920, 2024.

[15] Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *European Conference on Computer Vision (ECCV)*, 2018.

[16] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. IMoS: Intent-driven full-body motion synthesis for human-object interactions. *Computer Graphics Forum*, 42(2):1–12, 2023.

[17] Liang-Yan Gui, Kevin Zhang, Yu-Xiong Wang, Xiaodan Liang, José MF Moura, and Manuela Veloso. Teaching robots to predict human motion. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 562–567. IEEE, 2018.

[18] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019.

[19] 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 (ICCV)*, 2021.

[20] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14708–14718, 2021.

[21] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194, 2005.

[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[23] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–834, 2016.

[24] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3D world into large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[25] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.

[26] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3D scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16750–16761, 2023.

[27] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In *European Conference on Computer Vision (ECCV)*, 2024.

[28] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7), 2014.

[29] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-RNN: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5308–5317, 2016.

[30] Jaewoo Jeong, Seohee Lee, Daehee Park, Giwon Lee, and Kuk-Jin Yoon. Multi-modal knowledge distillation-based human trajectory forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24222–24233, 2025.

[31] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[32] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024.

[33] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. NIFTY: Neural object interaction fields for guided human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[34] Michael F Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25-26):3559–3565, 2001.

[35] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: Insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

[36] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision (ECCV)*, 2024.

[37] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks for 3D human motion prediction. In *European Conference on Computer Vision (ECCV)*, 2022.

[38] Yi Lin, Wangjie Xia, and Ting Yu. HuMo3D: Intention-driven dual-branch multimodal human motion prediction in 3D scenes. *IEEE Signal Processing Letters*, 33:614–618, 2026. doi: 10.1109/LSP.2026.3652960.

[39] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

[40] Zhenyu Lou, Qiongjie Cui, Haofan Wang, Xu Tang, and Hong Zhou. Multimodal sense-informed forecasting of 3D human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2144–2154, 2024.

[41] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[42] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[43] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[44] Wei Mao, Miaomiao Liu, Richard I. Hartley, and Mathieu Salzmann. Contact-aware human motion forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 7356–7367, 2022.

[45] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[46] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. 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 (CVPR)*, pages 10975–10985, 2019.

[47] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[48] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[49] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015.

[50] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision (ECCV)*, 2020.

[51] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6457–6466, 2022.

[52] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11209–11218, 2021.

[53] Caiyi Sun, Yujing Sun, Xiao Han, Zemin Yang, Jiawei Liu, Xinge Zhu, Siu Ming Yiu, and Yuexin Ma. Humof: Human motion forecasting in interactive social scenes. In *International Conference on Learning Representations*, volume 2026, pages 6511–6534, 2026.

[54] Julian Tanke, Oh-Hun Kwon, Felix B. Mueller, Andreas Doering, and Jürgen Gall. Humans in kitchens: A dataset for multi-person human motion forecasting with scene context. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, 2023.

[55] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009.

[56] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3D human motion and interaction in 3D scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9401–9411, 2021.

[57] Tuo Wang, Zhenbo Song, Zhenyu Lou, Qiongjie Cui, Luoming Zhang, Cheng Cheng, Haofan Wang, Xu Tang, Huaxia Li, and Hong Zhou. Harmonizing stochasticity and determinism: Scene-responsive diverse human motion prediction. volume 37, pages 39784–39811, 2024.

[58] Wenjia Wang, Liang Pan, Zhiyang Dou, Jidong Mei, Zhouyingcheng Liao, Yuke Lou, Yifan Wu, Lei Yang, Jingbo Wang, and Taku Komura. Sims: Simulating stylized human-scene interactions with retrieval-augmented script generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14117–14127, October 2025.

[59] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3D scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

[60] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 433–444, 2024.

[61] Chaoyue Xing, Wei Mao, and Miaomiao Liu. Scene-aware human motion forecasting via mutual distance prediction. In *European Conference on Computer Vision (ECCV)*, 2024.

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

[63] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. EqMotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1410–1420, 2023.

[64] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[65] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7694–7701. IEEE, 2024.

[66] 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 (CVPR)*, pages 12335–12346, 2025.

[67] Ye Yuan and Kris Kitani. DLow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision (ECCV)*, 2020.

[68] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[69] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. HOI-M^3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 516–526, 2024.

[70] Jiahe Zhao, Ruibing Hou, Zejie Tian, Hong Chang, and Shiguang Shan. His-gpt: Towards 3d human-in-scene multimodal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4317–4327, October 2025.

[71] Peisen Zhao, Lingxi Xie, Ya Zhang, and Qi Tian. Progressive privileged knowledge distillation for online action detection. *Pattern Recognition*, 129, 2022.

[72] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[73] 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)*, 2022.

A Overview and Information Availability

This appendix is organized as follows. Appendix B reports additional quantitative results: the attribution of grounding accuracy (B.1), teacher–student token agreement (B.2), the multi-future protocol and the comparison with diffusion-based generation (B.3), privileged-cue variants (B.4), comparisons in multi-human, dynamic scenes (B.5), robustness to segmentation noise (B.6), sensitivity to the intent-extraction VLM (B.7), runtime (B.8), and soft token aggregation (B.9). Appendix C provides additional qualitative results. Appendices D–G detail candidate-token construction, privileged supervision, the architecture, and training. Appendix H discusses broader impacts.

Table 4 summarizes which inputs each component of SGIP observes. The VLM, the LLM, and the privileged teacher are used only to construct training-time supervision; only the student is deployed.

Table 4: Information availability across SGIP components. $c_i = (\boldsymbol{\phi}_i, \boldsymbol{e}_i, \boldsymbol{m}_i)$ denotes a candidate token.

Component	**Used at**	**Inputs**
VLM intent extractor (offline)	training only	training video, 3D motion visualization, intent taxonomy
LLM rubric scorer (offline)	training only	intent $\mathcal{I}$, candidate semantic labels, scene context
Privileged teacher	training only	$M_{1:t}$, S, $\{c_i\}$, $\mathcal{I}$, cached $\{s_i^{\text{rubric}}\}$
Observation-only student	training + inference	$M_{1:t}$, S, $\{c_i\}$
GT interaction target	evaluation only	never input to any network

The teacher never observes the ground-truth grounding token; it observes only the candidate tokens, their semantic labels, the textual intent $\mathcal{I}$, and the geometric descriptors $\{\boldsymbol{\phi}_i\}$. Distillation therefore transfers *intent-aware reasoning* over candidate properties rather than ground-truth label leakage.

B Additional Quantitative Results

All SGIP results in this section are averaged over five random seeds unless stated otherwise. P/D denote path/destination.

B.1 Attribution of Grounding Accuracy

Evaluation target. Grounding accuracy is measured against the GIMO-annotated interaction target, which is used only for evaluation (Table 4). We match each annotated target to OpenIns3D proposals by IoU on point membership (threshold 0.5). After cone filtering, the matched target remains in the candidate set for every evaluation sequence (Appendix D.2).

Matched controls. To separate the effect of intent-conditioned supervision from that of the learnable head and candidate preprocessing (main Table 3c), all student variants use the same grounding MLP, input features, candidate sets, cone filtering, downstream architecture, and training protocol; only the prior-distillation target changes. *No prior KD* removes $\mathcal{L}_z^{\text{KD}}$ and trains the same head with the remaining objectives; $\text{KL}(R)$ and $\text{KL}(G)$ distill the rubric-only and geometry-only priors (Eq. 7 with only one score), and $\text{KL}(G{+}R)$ distills the fused teacher prior $\pi^{\mathcal{T}}$. With the identical head, distilling G yields 51% Acc@1, whereas distilling $G{+}R$ yields 76%; the No-prior-KD result (41%) further shows that the head alone does not account for the accuracy of the deployed student.

Rank distribution. Table 5 reports, for the fixed priors and the deployed student, the fraction of sequences in which the annotated target is ranked at position k. The unimodal priors reflect two complementary failure modes: the rubric score cannot disambiguate same-category instances (e.g., several chairs), while the geometric score is intent-blind and confuses objects that are equally reachable but not intent-compatible. Fusing both concentrates mass on the target (86% at rank 1), and the distilled student retains most of this concentration (76%).

Table 5: Rank distribution of the GIMO-annotated target (%). "Rank k" is the fraction of sequences in which the target is ranked exactly at position k; "≤ 3" is their sum.

Prior	Rank 1	Rank 2	Rank 3	≤ 3
Rubric only (R)	54	24	2	80
Geometry only (G)	51	16	11	78
Teacher $\pi^{\mathcal{T}}$ (G+R)	**86**	5	3	**94**
Student $\pi^{\mathcal{S}}$	76	5	3	84

B.2 Teacher–Student Token Agreement

To test the grounding-misalignment explanation of Section 4.4, we measure exact Top-1 token agreement $i^*_{\mathcal{T}} = i^*_{\mathcal{S}}$ between the frozen teacher and the student. Agreement counts identical selections regardless of whether the selected token matches the ground truth.

Table 6: Teacher–student Top-1 token agreement, student grounding accuracy, and trajectory error on GIMO (mm).

Student loss	T–S Agree@1	GT Acc@1	Traj. P/D ↓
$\mathcal{L}_Y^{\mathrm{KD}}$ only	43%	41%	856/896
$\mathcal{L}_z^{\mathrm{KD}}$ only	70%	76%	574/688
$\mathcal{L}_z^{\mathrm{KD}} + \mathcal{L}_Y^{\mathrm{KD}}$	70%	76%	**549/681**

Under $\mathcal{L}_Y^{\mathrm{KD}}$ alone, the teacher and student select different tokens in 57% of the sequences, so the student imitates trajectories conditioned on a grounding it does not share. Adding $\mathcal{L}_z^{\mathrm{KD}}$ raises agreement to 70% and accuracy to 76%. Once grounding is aligned, adding $\mathcal{L}_Y^{\mathrm{KD}}$ further reduces trajectory error from 574/688 to 549/681 mm without changing either grounding metric, i.e., motion-level distillation helps only after prior distillation has aligned the two branches.

B.3 Multi-Future Protocol and Comparison with Diffusion-Based Generation

Metrics. Let $\{\hat{\boldsymbol{p}}_\tau^{(k)}\}_{k=1}^K$ be the predicted global root positions of K futures and $\boldsymbol{p}_\tau$ the ground truth, $\tau = t{+}1, \ldots, T$. With $\mathrm{ADE}^{(k)} = \frac{1}{\Delta T}\sum_\tau \|\hat{\boldsymbol{p}}_\tau^{(k)} - \boldsymbol{p}_\tau\|_2$ and $\mathrm{FDE}^{(k)} = \|\hat{\boldsymbol{p}}_T^{(k)} - \boldsymbol{p}_T\|_2$,

$$\mathrm{minADE@}K = \min_k \mathrm{ADE}^{(k)}, \qquad \mathrm{minFDE@}K = \min_k \mathrm{FDE}^{(k)}, \tag{16}$$

$$\mathrm{FDE@minADE} = \mathrm{FDE}^{(k^\star)}, \quad k^\star = \arg\min_k \mathrm{ADE}^{(k)}, \tag{17}$$

$$\mathrm{APD} = \frac{2}{K(K-1)} \sum_{k<l} \frac{1}{\Delta T} \sum_\tau \big\|\hat{\boldsymbol{p}}_\tau^{(k)} - \hat{\boldsymbol{p}}_\tau^{(l)}\big\|_2. \tag{18}$$

FDE@minADE evaluates whether the trajectory that is best on average also arrives at the right place, which minFDE alone does not guarantee.

SGIP futures. SGIP decodes each of the top-K tokens of $\pi^{\mathcal{S}}$ separately; the observed motion, scene, and model parameters are fixed and only the selected token's mask and embedding change. SGIP therefore produces K futures without privileged cues or stochastic sampling, each associated with an interpretable scene token. Main Table 2 reports SGIP for $K \in \{1, 2, 3\}$. Diffusion-based generation models offer an alternative route to multiple futures, obtaining them by stochastic sampling rather than from an explicit distribution over scene tokens; below, we compare SGIP with such a model under the same inputs and the same prediction budget.

AffordMotion adaptation. AffordMotion [60] is a two-stage conditional diffusion model that first predicts a scene affordance map and then generates motion conditioned on it. Since it is designed for language-guided motion generation rather than forecasting, we adapt it to HMP following GAP3DS [66] by replacing the language condition with an encoding of the observed motion, so that it receives the same inputs ($M + S$) as SGIP; both methods use the same GIMO split, windows,

and evaluation code. For single-future evaluation, AffordMotion produces one diffusion sample per sequence and SGIP decodes its Top-1 token; for $K = 3$, AffordMotion draws three independent samples and SGIP decodes its Top-3 tokens.

Table 7: Comparison with the diffusion-based AffordMotion adapted to HMP. Both methods use $M + S$ at inference. Single-future errors are in mm; multi-future root-trajectory errors are in m.

	Single future		Multi-future ($K = 3$)		
Method	Traj. P/D ↓	MPJPE P/D ↓	minADE ↓	minFDE ↓	FDE@minADE ↓
AffordMotion	577/714	160.9/**194.1**	0.597	0.728	0.730
SGIP	**549/681**	**153.1**/195.1	**0.467**	**0.523**	**0.571**

Under the matched setting, SGIP achieves lower single-future trajectory errors and MPJPE-path, and with three predictions it attains lower minADE, minFDE, and FDE@minADE than three diffusion samples. Rather than relying on stochastic sampling, SGIP exposes an explicit distribution over interpretable scene tokens, provides a deterministic Top-1 prediction, and generates controllable alternatives by changing the selected token.

B.4 Privileged-Cue Variants

Cue interface. All privileged variants inject their cue only through scores over the same candidate tokens, followed by the same Top-1 selection and the same token-conditioned predictor; no cue is fed to the motion predictor directly. For intent, $s_i^{\mathcal{I}} = s_i^{\text{rubric}} + \alpha s_i^{\text{geo}}$ (Eq. 7). For gaze, we use the single 3D gaze point at the final observed frame and compute a proximity score $s_i^{\text{gaze-dist}}$ between candidate c_i and the gaze point and a directional-alignment score $s_i^{\text{gaze-align}}$ between the head-to-gaze and head-to-candidate directions:

$$s_i^{g} = \alpha s_i^{\text{geo}} + \lambda_d \, s_i^{\text{gaze-dist}} + \lambda_a \, s_i^{\text{gaze-align}}, \tag{19}$$

$$s_i^{g+\mathcal{I}} = s_i^{\mathcal{I}} + \lambda_d \, s_i^{\text{gaze-dist}} + \lambda_a \, s_i^{\text{gaze-align}}, \tag{20}$$

with $\lambda_d = \lambda_a = 3$ selected on the training split. The scores are normalized before selection.

Table 8: Privileged-cue variants on GIMO (mm). All variants share the same token-conditioned predictor.

Variant	Cue	Traj. P/D ↓	MPJPE P/D ↓
SGIP (deployed)	–	549/681	153.1/195.1
$SGIP^{\dagger}_{\text{priv}}$	gaze	528/635	149.9/190.6
$SGIP^{\dagger}_{\text{priv}}$	intent	485/556	151.2/187.8
$SGIP^{\dagger}_{\text{priv}}$	gaze+intent	**475/539**	**149.7/180.8**

Intent yields lower trajectory error than gaze, and combining both cues improves all four metrics over either cue alone, consistent with intent providing semantic evidence about *which* target will be engaged and gaze providing spatial evidence for localizing it. Because the cues act only through the candidate scores, heterogeneous privileged cues can be combined without modifying the downstream predictor.

B.5 Comparison in Multi-Human, Dynamic Scenes

GIMO and PROX contain a single person per scene. To examine SGIP in scenes where other people move around the target, we compare it with HUMOF [53], a forecasting model designed for interactive social scenes, on HIK [54] and HOI-M^3 [69].

Protocol. We follow the target-centric forecasting setting of HUMOF [53], in which each sample predicts the future motion of one target person in a multi-person scene. For HIK [54], recordings A–C are used for training and D for testing, with 25 observed and 50 future frames at 25 FPS. For HOI-M^3 [69], we use a subset of 9/3 train/test sequences (3,332/689 samples) with 30 observed and

60 future frames at 30 FPS. We train HUMOF with its official implementation on the same samples as SGIP, and evaluate both with the released Path/Pose metrics, using the last observed root of the target as the local origin.

Inputs. HUMOF observes the temporal histories of neighboring humans (H) in addition to the target history M and the scene S. SGIP uses only the target history and represents neighboring humans and objects as static tokens at the last observed frame. Object tokens on HIK and HOI-M^3 are constructed from native object annotations and geometry at the last observed frame. HIK uses native action annotations as training-time intents for a Qwen3-14B text teacher. For HOI-M^3, Qwen3-VL-8B infers intents from a video and a reference image identifying the target person; future frames provide privileged teacher supervision.

Table 9: Comparison in multi-human, dynamic scenes on HIK and HOI-M^3 (mm). Both methods are trained and evaluated on identical samples.

			Path error					Pose error				
Dataset	Method	H	0.5s	1.0s	1.5s	2.0s	mean	0.5s	1.0s	1.5s	2.0s	mean
HIK	HUMOF [53]	✓	96.9	231.3	400.3	546.6	251.9	**73.5**	**99.3**	**116.7**	**123.7**	**90.6**
	SGIP	✗	**89.2**	**130.9**	**167.7**	**205.7**	**127.1**	81.5	114.9	130.7	138.8	103.1
HOI-M^3	HUMOF [53]	✓	**104.7**	**238.0**	**385.8**	**538.4**	**253.9**	**117.2**	163.0	189.3	210.7	148.9
	SGIP	✗	134.9	280.7	437.4	583.0	291.1	121.1	**159.2**	**179.1**	**200.9**	**145.0**

Discussion. On HIK, SGIP achieves lower path error than HUMOF at every horizon, reducing the mean from 251.9 to 127.1 mm without observing neighboring-human histories, while HUMOF, which explicitly models these histories, attains lower pose error. On HOI-M^3, SGIP obtains lower pose error from 1.0 s onward and on average. These results indicate that the scene-grounded token representation transfers to multi-human, dynamic scenes even without neighbor histories; modeling time-varying human and object states is a natural extension.

B.6 Robustness to Segmentation Noise

Protocol. We corrupt the OpenIns3D output at test time on all GIMO evaluation sequences while keeping the trained model fixed. We perturb instance-mask logits with Gaussian noise $\sigma \in \{0.5, 1.0, 2.0\}$ and simulate under- and over-segmentation by merging the target with its nearest object and splitting it into two contiguous instances, respectively. Gaussian results are averaged over three corruption seeds; merge and split are deterministic. *Target recall* indicates whether a target-derived candidate survives corruption and cone filtering. *Target Acc@1* is computed over all evaluation sequences, counting missing targets as incorrect; for merge/split, the merged candidate or either fragment counts as target-derived.

Table 10: Segmentation-noise ablation on GIMO. Gaussian rows report mean±std over three corruption seeds; motion errors are in mm.

Condition	Target recall (%)	Target Acc@1 (%) ↑	Traj. P/D ↓	MPJPE P/D ↓
Clean	100.00	75.68	549.0/681.0	153.1/195.1
Gaussian, $\sigma = 0.5$	100.00 ± 0.00	69.37 ± 1.56	$559.0 \pm 14.7/708.0 \pm 17.4$	$158.1 \pm 1.1/211.1 \pm 1.2$
Gaussian, $\sigma = 1.0$	99.10 ± 1.56	65.77 ± 1.56	$545.0 \pm 15.0/690.0 \pm 35.5$	$158.5 \pm 0.9/212.1 \pm 1.5$
Gaussian, $\sigma = 2.0$	81.08 ± 4.68	53.15 ± 1.56	$581.0 \pm 20.3/749.0 \pm 34.1$	$158.1 \pm 1.6/210.3 \pm 0.4$
Target–neighbor merge	100.00	56.76	661.0/875.0	159.2/211.3
Target two-way split	100.00	54.05	644.0/908.0	157.1/208.8

Under mild noise ($\sigma = 0.5$), target recall remains 100% while Target Acc@1 decreases from 75.68% to 69.37% and trajectory error increases from 549/681 to 559/708 mm. Under severe noise ($\sigma = 2.0$), target recall and Target Acc@1 drop to 81.08% and 53.15%. Merge and split preserve a target-derived candidate but raise destination error to 875 and 908 mm, since the merged or fragmented instance no longer delineates the interaction surface. Severe and structural segmentation errors therefore cause substantially larger degradation than mild mask corruption, which motivates the limitation discussed in Section 5.

B.7 Sensitivity to the Intent-Extraction VLM

The main experiments use Qwen3-VL-8B [5] to extract sequence-level intents for both GIMO and PROX. We replace it with Qwen3-VL-4B and Idefics3-8B [35], regenerate the intent labels, and retrain the same teacher–student pipeline while keeping the dataset-specific taxonomy, candidate tokens, architecture, optimization, and evaluation protocol fixed.

Table 11: Sensitivity to the VLM used for training-time intent. Agreement is the exact intent-category agreement with Qwen3-VL-8B within each dataset (mm).

	GIMO			PROX		
VLM	Agreement	Traj. P/D	MPJPE P/D	Agreement	Traj. P/D	MPJPE P/D
Qwen3-VL-8B (default)	–	549/681	153.1/195.1	–	470/576	194.6/197.7
Qwen3-VL-4B	90.7%	527/753	152.8/196.8	96.6%	459/570	195.0/209.5
Idefics3-8B	92.1%	584/723	154.9/190.6	97.0%	493/564	181.1/202.9

The alternative extractors agree with Qwen3-VL-8B on 90.7–92.1% of GIMO and 96.6–97.0% of PROX sequences. No extractor consistently improves or degrades all metrics across both datasets, indicating reasonable robustness to VLM scale and family, with larger variation in destination errors.

B.8 Runtime

We measure one-time scene preprocessing and per-observation student inference separately on a single NVIDIA RTX 3090 with batch size 1 and FP32. Scene initialization, including point-cloud loading, OpenIns3D instance segmentation, and construction of the static instance inventory, takes 61.73 ± 7.18 s per scene across the 19 GIMO scenes; the resulting masks and inventory are cached.

Table 12: Runtime on one RTX 3090 (batch size 1, FP32).

Stage	Frequency	Runtime
One-time preprocessing		
OpenIns3D scene initialization	once per scene	61.73 ± 7.18 s
Online student inference		
Motion-dependent descriptors	per observation	59.72 ± 18.99 ms
Student prior + Top-1 selection	per observation	6.10 ± 0.40 ms
Host-to-device transfer	per observation	0.38 ± 0.07 ms
Student motion prediction	per observation	66.54 ± 1.04 ms
Total online inference	per observation	132.75 ms (7.53 FPS)

After scene initialization, SGIP requires 132.75 ms per observation, below the 500 ms observation interval of the 2-FPS GIMO protocol, so it supports online prediction in a static scene; processing a new scene incurs the initialization cost again. No VLM or LLM is invoked at inference.

B.9 Hard vs. Soft Grounding Token Aggregation

The motion predictor consumes a single grounding token i_b^* via its mask $\boldsymbol{m}_{i_b^*}$ and embedding $\boldsymbol{e}_{i_b^*}$ (Eqs. 10–12). A natural alternative replaces the hard argmax (Eq. 9) with a probability-weighted mixture of the top-k tokens:

$$\bar{\boldsymbol{m}} = \sum_{i \in \mathcal{K}_k} \tilde{\pi}_i^{\mathcal{S}} \, \boldsymbol{m}_i, \qquad \bar{\boldsymbol{e}} = \sum_{i \in \mathcal{K}_k} \tilde{\pi}_i^{\mathcal{S}} \, \boldsymbol{e}_i, \tag{21}$$

where $\mathcal{K}_k = \mathrm{TopK}_k(\pi^{\mathcal{S}})$ and $\tilde{\pi}_i^{\mathcal{S}} = \pi_i^{\mathcal{S}} / \sum_{j \in \mathcal{K}_k} \pi_j^{\mathcal{S}}$; $D^{\mathcal{S}}$ then consumes $(\bar{\boldsymbol{m}}, \bar{\boldsymbol{e}})$ in place of $(\boldsymbol{m}_{i_{\mathcal{S}}^*}, \boldsymbol{e}_{i_{\mathcal{S}}^*})$. The soft mask $\bar{\boldsymbol{m}}$ enters the spatial importance map in place of $\boldsymbol{m}_{i_{\mathcal{S}}^*}$ (Eq. 26). All variants share the same trained student; only the token selection differs at inference.

Soft aggregation smooths the conditioning signal under ambiguity but dilutes the geometry of the selected region by blending in less likely tokens. When the blended tokens lie in different parts of

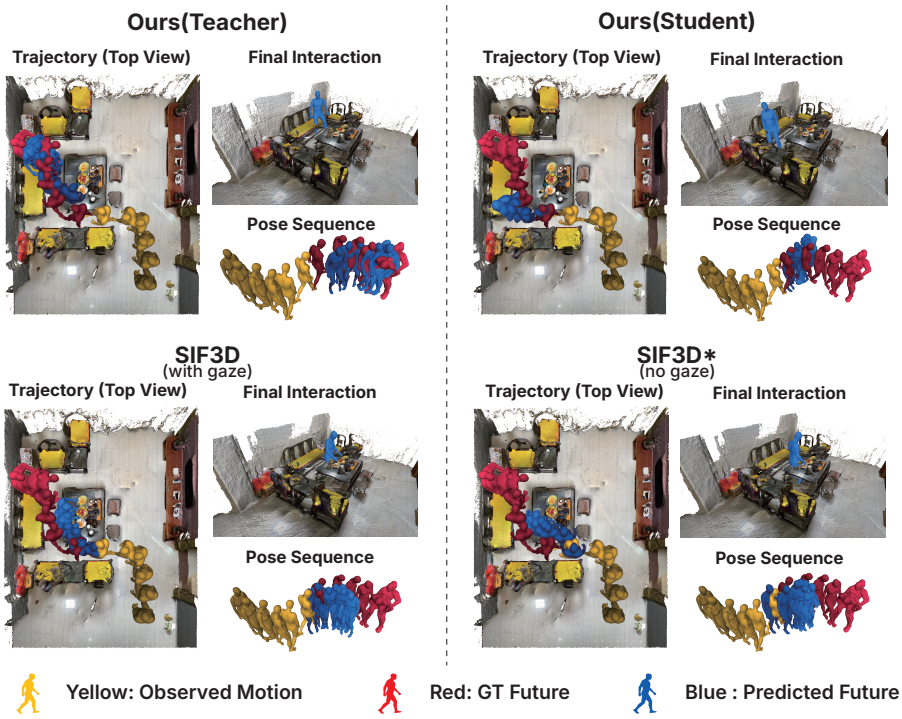


Figure 5: **Qualitative comparison across SGIP branches and SIF3D.** For one cluttered GIMO test scene, given the same observed motion: **top left**, the privileged SGIP teacher ($SGIP^{\dagger}_{\text{priv}}$, textual intent); **top right**, the deployed observation-only SGIP student; **bottom left**, SIF3D [40] with gaze; **bottom right**, SIF3D* without gaze. Each panel shows the top-down trajectory, the final interaction frame, and the pose sequence. Yellow: observed motion; red: GT future; blue: predicted future.

the scene, $\bar{e}$ no longer corresponds to any single target and $\bar{m}$ covers several disjoint regions, so the predicted trajectory tends toward an average of the candidate destinations and the arrival pose is not anchored to one contact surface. Accordingly, the trajectory degradation in main Table 3b grows as more tokens are blended ($k = 2 \rightarrow 3$). When multiple futures are desired, decoding the top-k tokens separately (Section 4.3) preserves the multimodal structure of $\pi^{\mathcal{S}}$ without blending.

C Additional Qualitative Results

C.1 Teacher, Student, and Gaze-Based Baselines

Figure 5 compares the two SGIP branches with SIF3D on the same scene. Without gaze, SIF3D drifts toward an incorrect region and produces an inconsistent arrival pose; with gaze, it recovers the rough direction but still over- or under-shoots the target, indicating that gaze alone does not fully disambiguate the interaction target in cluttered scenes. Both SGIP branches ground the trajectory on the correct target and produce a coherent arrival pose, and the observation-only student closely tracks the privileged teacher despite using neither gaze nor textual intent at inference.

C.2 Intent-Consistent Alternative Interactions

Single-future metrics count every deviation from the recorded motion as an error, yet in cluttered scenes the same observed motion often leaves several intentions open. Figure 6 shows cases in which the student's top-ranked token differs from the recorded target. Rather than drifting between candidates, stopping short, or averaging several destinations, SGIP commits to an alternative object that is reachable from the observed approach and that affords a common everyday activity, and it completes the interaction with that object: the trajectory ends at the selected object and the arrival pose is oriented toward it. Such predictions are consistent with an intent that the observation alone does



Figure 6: **Intent-consistent alternative interactions on GIMO.** For each scene, the recorded future (red) and the SGIP prediction (blue) are shown as top-down trajectories (top) and final-frame poses (bottom). **(a)** Seminar room: the recorded person goes to retrieve an eraser from the blackboard, while SGIP selects a desk as its grounding target. **(b)** Kitchen: the ground-truth target is the cabinet, while SGIP selects a pot on the dining table. These examples illustrate intent-aware grounding of plausible alternative futures beyond the recorded interaction. In both cases, SGIP commits to a different but reachable target and completes a coherent interaction with it.

not rule out, and they arise mainly when several targets are similarly compatible with the observed approach and the observation window ends before the person commits to one of them. This behavior follows from making grounding explicit: each prediction corresponds to one interpretable target together with a physically coherent interaction, which is also why SGIP exposes ranked alternatives (Section 4.3) instead of a single averaged future.

D Candidate Grounding Tokens

This appendix details the construction of the candidate grounding token set $\mathcal{C} = \mathcal{C}^{\text{obj}} \cup \mathcal{C}^{\text{nav}}$ used by the prior (Section 3.1).

D.1 Object-Instance Tokens

Object instances are obtained by zero-shot 3D instance segmentation on the scene point cloud S with OpenIns3D [27], producing per-instance point-cloud masks $\boldsymbol{m}_i \subseteq S$ with semantic labels (Figure 7). Same-category objects (e.g., two chairs) share a label but remain distinct tokens with distinct masks; the geometric compatibility score (Appendix D.5) disambiguates among them. The unfiltered inventory contains 14.68 ± 6.01 object tokens per GIMO evaluation scene.

D.2 Forward-Facing Cone Filtering

The cone is an SGIP-specific candidate-pruning step computed only from the observed motion; it uses neither gaze nor target annotations. At each observed frame τ, we define a cone $C(x_\tau)$ originating at the head whose axis bisects the head- and torso-facing directions, with a fixed aperture of $120°$, and take the union over the observation window:

$$C(M_{1:t}) = \bigcup_{\tau=1}^{t} C(x_\tau). \tag{22}$$

An object token is retained if its mask intersects the aggregated cone, i.e., $\boldsymbol{m}_i \cap C(M_{1:t}) \neq \emptyset$. Since upcoming interactions are typically preceded by head and body alignment toward the target [21, 34, 73], the $120°$ cone approximates a broad forward-facing interaction region and removes rearward objects that are unlikely to be immediate interaction targets. The teacher, the student, and all SGIP ablations use the same filtered candidate set.

Because the filtered set depends on the observed motion, the number of retained candidates varies across sequences. We first average sequence-level counts within each scene and then summarize them across the 19 GIMO evaluation scenes (Table 13). Together with the always-retained navigation

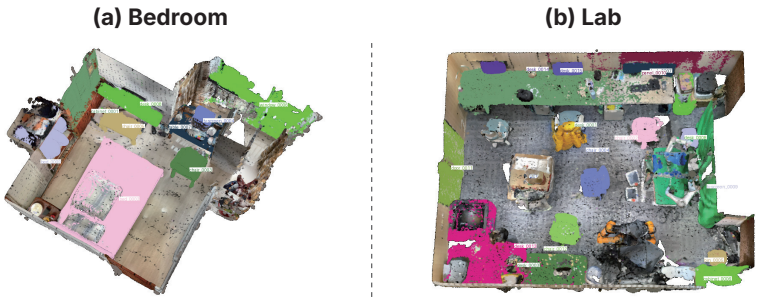


Figure 7: **Zero-shot 3D instance segmentation by OpenIns3D [27]** on two GIMO scenes. Each color denotes a distinct object instance $c_i \in \mathcal{C}^{\text{obj}}$; same-category objects (e.g., multiple chairs) are separated into distinct instances.

token (Appendix D.3), this yields 13.60 ± 4.62 grounding candidates on average, and the annotated target remains in the filtered set for every evaluation sequence (100% target recall).

Table 13: Cone-filtered object candidates on GIMO. Sequence-level counts are first averaged within each scene; values are mean±std across the 19 evaluation scenes.

Candidate set	Object tokens	Retention (%)
Before cone filtering	14.68 ± 6.01	100.00
After cone filtering (120°)	12.60 ± 4.62	88.83 ± 11.07

D.3 Navigation-Region Token

The navigation token $\mathcal{C}^{\text{nav}}$ represents object-agnostic free space (e.g., open floor, doorways, corridors). Unlike object tokens, it is scene-global and is not cone-filtered, so that the prior can represent free-space motion when the observed motion is not yet aligned with any specific object. Its mask is the complement of all object masks within an interaction-relevant region.

D.4 Geometric Descriptor

The geometric descriptor concatenates two groups of features, $\boldsymbol{\phi}_i = [\boldsymbol{\phi}_{\text{geo}}(c_i),\ \boldsymbol{\phi}_{\text{rel}}(c_i, M_{1:t})] \in \mathbb{R}^F$ with $F = 16$. $\boldsymbol{\phi}_{\text{geo}}(c_i)$ encodes token-intrinsic spatial properties (centroid, oriented bounding-box size, height above the floor, and vertical orientation of the contact normal). $\boldsymbol{\phi}_{\text{rel}}(c_i, M_{1:t})$ encodes motion-relative quantities at the last observed frame: the signed distance from the root to c_i, the root-to-token approach angle, root-horizontal alignment, face-normal alignment, and torso alignment. Features are z-score normalized with training-set statistics.

D.5 Geometric Compatibility Score

Each alignment term in Eq. (6) measures the cosine similarity between a motion-derived direction at the last observed frame and the direction from the root toward token c_i, normalized to $[0, 1]$: $\text{align}_{\text{face}}$ uses the head-facing direction, $\text{align}_{\text{body}}$ the torso-facing direction, and $\text{align}_{\text{line}}$ the recent trajectory direction. We fix $w_{\text{face}} = 0.5$, $w_{\text{body}} = 2$, and $w_{\text{line}} = 0.5$ across all experiments, selected on a held-out portion of the training split.

E Privileged Supervision

E.1 VLM-Based Intent Extraction

We label the intent of each training sequence with a frozen Qwen3-VL-8B [5]. Because a person's intent is difficult to infer from a single viewpoint, the VLM jointly receives two videos of the full training clip. On GIMO, these are the **egocentric video**, which conveys what the person looks at

and immediate interaction targets, and a **rendering of the person's motion in the reconstructed 3D scene**, which conveys where the action takes place. On PROX, which has no egocentric video, the first input is the RGB video recorded by the static PROX camera. The prompt includes (i) the definitions of the dataset's intent categories, (ii) the role of each video, and (iii) an instruction to answer *unknown* when no category applies. The predicted category is converted into a textual intent $\mathcal{I}$ (e.g., "take some rest.") that is used only by the privileged teacher. The VLM, the videos, and the pseudo-intents are used only to construct training-time supervision.

E.2 Intent Taxonomies

GIMO (16 categories). *sit*, *sleep*, *drink*, *eat*, *read*, *use*, *take*, *open*, *light*, *clean*, *dress*, *throw*, *cook*, *write*, *fix*, and *unknown*.

PROX (7 categories). *sit-up*, *sit-down*, *lie*, *walk*, *touch*, *jump*, and *unknown*.

Design principles. Categories are intended to be mutually exclusive while covering the main behavioral patterns of each dataset, and each corresponds to a unit action that is visually distinguishable from the available views, so that the VLM can classify it from visual cues alone. The explicit *unknown* class prevents ambiguous sequences from being forced into an incorrect category. Robustness to the choice of VLM is reported in Appendix B.7.

E.3 LLM-Based Rubric Scorer

Rubric. A reasoning LLM scores each candidate token c_i on two integer criteria:

$a_i \in \{1, \dots, 5\}$ (*intent affordance*: how well does c_i semantically support $\mathcal{I}$?)
$u_i \in \{1, \dots, 5\}$ (*intent usability*: given the scene, how easily can $\mathcal{I}$ be realized using c_i?)

The total $s_i^{\text{rubric}} = a_i + u_i \in [2, 10]$ enters the temperature softmax of Eq. (7). Geometric alignment is already available to the student through $\boldsymbol{\phi}_i$ and $\boldsymbol{h}_M$, so the rubric deliberately focuses on intent-aware semantic judgment, which is the teacher's distinctive contribution.

LLM call. We use Qwen3-14B [64] with NF4 4-bit quantization and greedy decoding (temperature 0, top-$p = 1$, at most 4096 new tokens, fixed seed). The prompt enforces a two-stage output: a natural-language rationale (used to encourage scoring consistency, not parsed as supervision) followed by a strict JSON block listing per-token integer scores and the recomputed total. Each (sequence, candidate set) pair is queried once and cached offline; no LLM call occurs during training updates or inference.

JSON validation and exclusion. Each JSON block is validated against a schema that checks per-criterion ranges and recomputes $s_i^{\text{rubric}} = a_i + u_i$ when the reported total disagrees. Out-of-range scores, missing entries, or duplicate token identifiers trigger a single retry; duplicated score values or repeated object names are legitimate (e.g., two chairs) and do not. The protocol was designed to exclude the entire sample in the event of a JSON parsing failure, since removing individual tokens could alter the candidate set and eliminate the target. In practice, no parsing failures occurred, and no training samples were excluded ($N = 0$).

F Architecture Details

F.1 Encoders and Student Scoring Head

Motion encoder E_M^b. A 4-layer Transformer with hidden dimension 256 and 4 heads ingests the per-frame SMPL-X state of $M_{1:t}$ and outputs the motion feature $\boldsymbol{h}_M^b$.

Token encoder $E_{\mathcal{C}}$. For each token c_i, a shared PointNet++ [48] backbone (2 set-abstraction layers) extracts a point-cloud feature from the points of $\boldsymbol{m}_i$, which is concatenated with a semantic-label embedding and a projection of $\boldsymbol{\phi}_{\text{geo}}(c_i)$ and passed through a two-layer MLP (GELU) to produce $\boldsymbol{e}_i \in \mathbb{R}^{d_e}$, $d_e = 16$.

Student scoring head $F_{\mathcal{S}}^{\mathrm{grd}}$. A permutation-equivariant set encoder over candidates. Given $\boldsymbol{h}_M^{\mathcal{S}}$ and $\{\boldsymbol{\phi}_i\}$:

1. Per-token MLP: $\boldsymbol{u}_i = \mathrm{MLP}_{\mathrm{cand}}([\boldsymbol{\phi}_i; \boldsymbol{h}_M^{\mathcal{S}}]) \in \mathbb{R}^{d_h}$, $d_h = 256$, GELU.
2. Masked mean and max pooling over candidates: $\bar{\boldsymbol{u}}^{\mathrm{mean}}, \bar{\boldsymbol{u}}^{\mathrm{max}} \in \mathbb{R}^{d_h}$.
3. Context projection: $\bar{\boldsymbol{c}} = \mathrm{MLP}_{\mathrm{ctx}}([\bar{\boldsymbol{u}}^{\mathrm{mean}}; \bar{\boldsymbol{u}}^{\mathrm{max}}])$.
4. Per-token logit: $F_{\mathcal{S}}^{\mathrm{grd}}(\boldsymbol{h}_M^{\mathcal{S}}, \boldsymbol{\phi}_i) = \mathrm{MLP}_{\mathrm{head}}([\boldsymbol{u}_i; \bar{\boldsymbol{c}}]) \in \mathbb{R}$.

The logits are normalized with a masked softmax over valid candidates (Eq. 8).

F.2 Motion Predictor

The motion predictor D^b follows the trajectory-then-pose design of Section 3.3; the student and the teacher use the same architecture and differ only in the source of the selected token.

Scene encoding. The scene point cloud S is encoded by PointNet++ into per-point features $F_S \in \mathbb{R}^{N \times d}$.

Trajectory predictor. Each block applies self-attention to the motion tokens H_M, cross-attention to the scene features, and an MLP fusion with the selected object's time-indexed latent sequence:

$$H' = \mathrm{SelfAttn}(H_M), \tag{23}$$

$$C^{\mathrm{TP}} = \mathrm{softmax}\big(Q' K_S^\top / \sqrt{d}\big) V_S, \tag{24}$$

$$H^{\mathrm{TP}} = \mathrm{MLP}\big([H', C^{\mathrm{TP}}, \boldsymbol{E}_{i_b^*}^{\mathrm{obj}}]\big), \tag{25}$$

where $Q' = W_Q H'$, $K_S = W_K F_S$, and $V_S = W_V F_S$. Here, $\boldsymbol{E}_{i_b^*}^{\mathrm{obj}}$ is produced by encoding the selected object's point cloud into a sequence of time-indexed latents. At each temporal position, the corresponding object latent is concatenated with the attention features before MLP fusion. The same object latent sequence is supplied to each trajectory block. A linear head maps the output of the last block to the future root trajectory $\hat{\boldsymbol{r}}_{t+1:T}^b$ (translation and orientation).

Spatial importance map. The predicted trajectory induces a spatial importance map over scene points (Eq. 11):

$$P^b = f_{\mathrm{dist}}(S, \hat{\boldsymbol{r}}^b) + f_{\mathrm{dir}}(S, \hat{\boldsymbol{r}}^b) + f_{\mathrm{mlp}}(S, \hat{\boldsymbol{r}}^b) + \delta_{\mathrm{obj}}\, f_{\mathrm{obj}}(S, \boldsymbol{c}_{i_b^*}, \hat{\boldsymbol{r}}^b), \tag{26}$$

where f_{dist} is a path-proximity prior, f_{dir} a forward-facing alignment prior, f_{mlp} a learned spatial bias, and f_{obj} an object-direction alignment prior scaled by a learned coefficient. It computes the horizontal cosine similarity between the root-to-scene-point and root-to-object-center directions, where $\boldsymbol{c}_{i_b^*}$ is the centroid of the selected object's point cloud. The gate $\delta_{\mathrm{obj}} \in \{0, 1\}$ equals 1 when a valid point cloud for the selected object is available and 0 otherwise. This gate affects only the object-direction term; the remaining spatial-prior terms are retained.

Pose predictor. The pose branch first injects the predicted trajectory into the motion tokens and then performs scene cross-attention with $P^b \in \mathbb{R}^{\Delta T \times N}$ as an additive bias:

$$\tilde{H} = H_M + W_{\mathrm{ori}} \hat{\boldsymbol{R}}^b + W_{\mathrm{trans}} \hat{\boldsymbol{t}}^b, \tag{27}$$

$$H'' = \mathrm{SelfAttn}(\tilde{H}), \tag{28}$$

$$C^{\mathrm{PP}} = \mathrm{softmax}\big(Q'' K_S^\top / \sqrt{d} + P^b\big) V_S, \tag{29}$$

$$H^{\mathrm{PP}} = \mathrm{MLP}\big([H'', C^{\mathrm{PP}}]\big). \tag{30}$$

A final decoder assembles the predicted poses and trajectory into the SMPL-X future motion $\hat{Y}_{t+1:T}^b$.

Configuration. We stack 3 trajectory blocks and 3 pose blocks with hidden dimension 256, 4 attention heads, and dropout 0.1.

Comparison to gaze-conditioned decoders. Gaze-conditioned scene-aware HMP models [40, 66] inject a per-frame 3D gaze point as the scene-conditioning signal, which most directly affects the trajectory branch. Our predictor instead consumes the selected token i_b^*, inferred from $(M_{1:t}, S)$ through the prior and requiring no auxiliary signal at inference; the token participates in both branches (its embedding in the trajectory branch and its mask through P^b in the pose branch), so the inferred grounding shapes trajectory and pose jointly.

F.3 Teacher Pose Refiner

Architecture. The teacher-only pose refiner (Section 3.3) adjusts the initial pose $\tilde{\boldsymbol{\theta}}^{\mathcal{T}}_{t+1:T}$ from $D^{\mathcal{T}}_{\text{pose}}$ using the textual intent $\mathcal{I}$ and the selected token's embedding:

$$\hat{\boldsymbol{\theta}}^{\mathcal{T}}_{t+1:T} = \mathrm{Refine}\big(\tilde{\boldsymbol{\theta}}^{\mathcal{T}}_{t+1:T},\, \mathcal{I},\, \boldsymbol{e}_{i^*_{\mathcal{T}}}\big). \tag{31}$$

To preserve the geometrically grounded trajectory, Refine keeps the predicted translation fixed and allows only a small, zero-initialized orientation residual $\Delta\boldsymbol{R}^{\mathcal{T}}_{t+1:T}$, composed as $\hat{\boldsymbol{R}}^{\mathcal{T}}_{\tau} \leftarrow \hat{\boldsymbol{R}}^{\mathcal{T}}_{\tau} \cdot \Delta\boldsymbol{R}^{\mathcal{T}}_{\tau}$ in $SO(3)$; this permits limited re-orientation (e.g., turning slightly toward an object before sitting) without disrupting the global path. Since the student has no access to $\mathcal{I}$, Refine is teacher-only; the student's pose comes directly from $D^{\mathcal{S}}_{\text{pose}}$.

Intent-alignment loss. Refine is supervised by an intent-alignment loss that pulls the refined pose toward $\mathcal{I}$ in a shared text–pose feature space:

$$\mathcal{L}_{\text{intent-pose}} = \lambda_{\text{end}}\, d\big(\psi_\theta(\hat{\boldsymbol{\theta}}^{\mathcal{T}}_{T}),\, \psi_c(\mathcal{I})\big) + \lambda_{\text{seq}}\, d\big(\psi_\theta(\hat{\boldsymbol{\theta}}^{\mathcal{T}}_{t+1:T}),\, \psi_c(\mathcal{I})\big), \tag{32}$$

where ψ_c is a text encoder, ψ_θ a pose-sequence encoder mapping body pose into the same latent space, and $d(\cdot,\cdot)$ a cosine distance. The first term enforces alignment at the destination frame, which corresponds to the interaction state described by $\mathcal{I}$, while the second propagates intent semantics across the sequence. The orientation residual is regularized by $\lambda_{\text{ori}}\, \|\Delta\boldsymbol{R}^{\mathcal{T}}_{t+1:T}\|$, keeping orientation close to the trajectory anchor. Values are listed in Table 14.

G Training and Implementation Details

G.1 Losses

Motion reconstruction loss. For branch $b \in \{\mathcal{T}, \mathcal{S}\}$,

$$\mathcal{L}^b_{\text{mot}} = \lambda_{\text{joint}}\, \mathcal{L}_{\text{joint}} + \lambda_{\text{vp}}\, \mathcal{L}_{\text{vposer}} + \lambda_{\text{mpjpe}}\, \mathcal{L}_{\text{mpjpe}} + \lambda_{\text{vel}}\, \mathcal{L}_{\text{vel}} + \lambda_{\text{foot}}\, \mathcal{L}_{\text{foot}} + \lambda_{\text{sem}}\, \mathcal{L}_{\text{sem}}, \tag{33}$$

comprising joint-level reconstruction on 3D joint positions, VPoser-latent regularization toward the pose prior [46], non-root MPJPE, inter-frame velocity consistency, a foot-contact penalty that suppresses sliding during ground contact, and semantic-prompt alignment; individual terms follow [40, 66].

Optimized parameters.

- Teacher trajectory pass: updates $\{E^{\mathcal{T}}_M, E^{\mathcal{T}}_{\mathcal{C}}, D^{\mathcal{T}}\}$ with $\mathcal{L}^{\mathcal{T}}_{\text{mot}}$.
- Teacher refinement pass: freezes $\{E^{\mathcal{T}}_M, E^{\mathcal{T}}_{\mathcal{C}}, D^{\mathcal{T}}_{\text{traj}}\}$ and updates Refine (with $\Delta\boldsymbol{R}^{\mathcal{T}}$) using $\mathcal{L}_{\text{intent-pose}}$, pose reconstruction, and the residual penalty.
- Student: freezes the teacher and updates $\{E^{\mathcal{S}}_M, E^{\mathcal{S}}_{\mathcal{C}}, F^{\text{grd}}_{\mathcal{S}}, D^{\mathcal{S}}\}$ using $\mathcal{L}^{\mathcal{S}}$ (Eq. 15), treating $\pi^{\mathcal{T}}$ and $\hat{Y}^{\mathcal{T}}_{t+1:T}$ as stop-gradient targets.

G.2 Optimization and Inference

The trajectory and pose branches are jointly optimized with AdamW (initial learning rate 3×10^{-4}, weight decay 10^{-4}) and an exponential learning-rate decay ($\gamma = 0.98$ per epoch), with batch size 8 and a training budget of up to 100 epochs.

At inference, the VLM, the LLM, textual intent, and the teacher are not used: given only $M_{1:t}$ and S, the student predicts $\pi^{\mathcal{S}}$, selects $i^*_{\mathcal{S}}$ (Eq. 9), and forecasts

$$\hat{Y}^{\mathcal{S}}_{t+1:T} = D^{\mathcal{S}}\big(\boldsymbol{h}^{\mathcal{S}}_M,\, S,\, \boldsymbol{m}_{i^*_{\mathcal{S}}},\, \boldsymbol{e}_{i^*_{\mathcal{S}}}\big). \tag{34}$$

G.3 Hyperparameters

Table 14: Hyperparameters used in all GIMO and PROX experiments.

Group	Hyperparameter	Value
Candidate tokens	cone aperture	120°
	GT–proposal matching IoU (evaluation)	0.5
Teacher prior	fusion weight α, temperature τ (Eq. 7)	0.79, 1.0
	$w_{\text{face}}, w_{\text{body}}, w_{\text{line}}$ (Eq. 6)	$0.5, 2, 0.5$
Gaze variant	λ_d, λ_a	$3, 3$
Encoders	E_M	4-layer Transformer, $d = 256$, 4 heads
	$E_{\mathcal{C}}$	PointNet++ (2 SA layers), $d_e = 16$
	descriptor dimension F	16
Scoring head	hidden size, MLPs	256; 2-layer MLPs, GELU
Motion predictor	trajectory / pose blocks	3 / 3
	hidden size, heads, dropout	256, 4, 0.1
Student loss	λ_z, λ_Y (Eq. 15)	1.0, 0.5
	λ_{traj}, λ_{pose} (Eq. 14)	1.0, 0.5
Motion loss	λ_{vel}	0.2
	$\lambda_{\text{vp}}, \lambda_{\text{mpjpe}}, \lambda_{\text{foot}}, \lambda_{\text{sem}}$	1.0, 1.0, 0.1, 0.1
Pose refiner	λ_{end}, ψ_c, ψ_θ	0.1, $(1536 \to 256 \to 256)$, $(66 \to 256 \to 256)$
Optimization	optimizer, lr, weight decay	AdamW, 3×10^{-4}, 10^{-4}
	schedule, batch size, epochs	exponential (γ=0.98), 8, 100

G.4 Compute, Seeds, Assets, and Code

All experiments run on a single NVIDIA RTX 3090 (24 GB). Each training stage takes approximately 2 hours. The VLM intent extraction and the LLM rubric scoring are run once per dataset and cached (the LLM pass over the GIMO training set takes approximately 4 hours); no VLM or LLM call occurs during training updates or inference. SGIP results are averaged over five random seeds (7, 42, 100, 984, 1004). We use GIMO [73], PROX [18], HIK [54], and HOI-M^3 [69], together with OpenIns3D [27], Qwen3 [64], Qwen3-VL [5], and Idefics3 [35], under their respective licenses and terms of use for non-commercial research.

H Broader Impacts

Anticipating where a person will move and how they will interact with the scene can make assistive robots and embodied agents safer and more responsive, and SGIP's observation-only deployment avoids gaze trackers or user-provided intent. The same capability could, however, be used to track or anticipate people's behavior without their consent. SGIP is trained on public research datasets of consenting participants and does not identify individuals, but deployments that observe people should obtain consent and comply with applicable privacy regulations. Predictions can also differ from a person's actual choice in ambiguous scenes (Appendix C.2), so safety-critical systems should not rely on a single predicted future; the ranked alternatives of Section 4.3 can support more conservative planning.

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 claims in the abstract and introduction are supported by the method (Section 3) and by the results in Section 4 (Tables 1–3); the scope (observation-only deployment, results on GIMO and PROX, comparisons in multi-human, dynamic scenes on HIK and HOI-M^3) is stated explicitly.

   Guidelines:

   - The answer NA 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 NA 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: Section 5 discusses limitations (dependence on candidate-token quality, per-scene instance segmentation, and static neighbor tokens in multi-human scenes), supported by Appendices B.6, B.8, and B.5; the privileged pose gap and sensitivity to the intent-extraction VLM are reported in Appendices B.4 and B.7.

   Guidelines:

   - The answer NA 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: The paper does not include theoretical results.

   Guidelines:

   - The answer NA 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: The paper specifies the architecture in Section 3 and Appendix F, candidate construction in Appendix D, VLM/LLM supervision (models, decoding, rubric, validation) in Appendix E, losses, optimization, hyperparameters, and compute in Appendix G, and the protocols of all additional experiments in Appendix B; the source code and reproduction scripts are also publicly released (see the next question).

   Guidelines:

   - The answer NA 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 source code, training and evaluation scripts, and instructions with the exact commands and environment needed to reproduce the main results are publicly available at `https://github.com/dhsoon02/SGIP`. All datasets (GIMO, PROX, HIK, HOI-M^3) are available from their original sources under their respective licenses.

   Guidelines:

   - The answer NA 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: Data splits, protocols, and training details are given in Section 4 and Appendices B and G, and the full configurations are included in the released code.

   Guidelines:

   - The answer NA 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:[Yes]

   Justification: SGIP results are averaged over five random seeds, and Table 1 reports the standard deviation on GIMO; corruption experiments report mean$\pm$std over three corruption seeds (Table 10).

Guidelines:

- The answer NA 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: All experiments use a single NVIDIA RTX 3090; training time, offline VLM/LLM cost, and inference runtime are reported in Appendices G.4 and B.8.

   Guidelines:

   - The answer Na 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 reviewed the NeurIPS Code of Ethics and ensured that our research conforms to it.

   Guidelines:

   - The answer NA 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: Positive and negative societal impacts, including privacy concerns of anticipating human behavior, are discussed in Appendix H.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA 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: There is no high risk for misuse in the data or models.

Guidelines:

- The answer NA 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: All datasets (GIMO, PROX, HIK, HOI-M^3) and models (OpenIns3D, Qwen3, Qwen3-VL) are cited, and their use under their respective licenses is stated in Appendix G.4.

Guidelines:

- The answer NA 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: There are no new assets introduced in the paper.

    Guidelines:

    - The answer NA 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: The paper does not involve any crowdsourcing or new research with human subjects; all human motion data comes from publicly released benchmarks (GIMO, PROX, HIK, HOI-M^3).

    Guidelines:

    - The answer NA 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: The paper does not involve any crowdsourcing or new research with human subjects; all human motion data comes from publicly released benchmarks (GIMO, PROX, HIK, HOI-M^3).

Guidelines:

- The answer NA 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: [Yes]

Justification: LLMs are core methodological components used only at training time: Qwen3-14B is the rubric-based scorer of the privileged teacher (Eq. 5, Appendix E.3), and Qwen3-VL-8B extracts sequence-level intent labels (Appendix E.1); Qwen3-VL-4B and Idefics3-8B are used in the sensitivity study (Appendix B.7).

Guidelines:

- The answer NA 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.