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†]

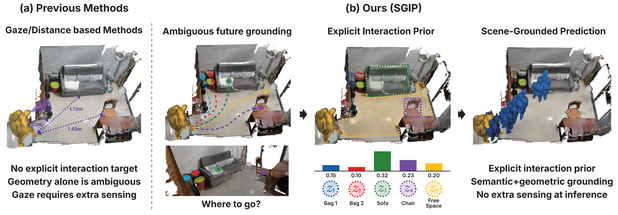
[1] DGIST [2] UNIVA * Equal contribution † Corresponding author

NEURAL INFORMATION PROCESSING SYSTEMS

ISL Lab

Project ISL Lab

Motivation



The same observed motion can lead to different targets or free-space navigation. (a) Prior methods rely on distance, which is ambiguous among nearby objects, or on gaze, which requires extra sensors. (b) SGIP predicts an explicit prior over scene tokens without extra sensing.

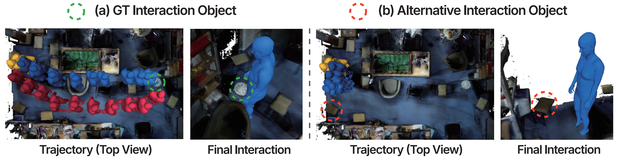
Contributions

Scene-grounded interaction prior. An explicit distribution over object-instance and free-space tokens makes the grounding of future motion a supervisable target.

Privileged intent distillation. During training only, VLM-extracted intent and an LLM rubric guide a teacher. The student learns the same prior from observed motion and the 3D scene alone.

Gaze-free, LLM-free inference. Without intent, gaze, a VLM, or an LLM at test time, the student outperforms observation-only baselines on GIMO and PROX in both trajectory and pose accuracy.

Multimodal futures



Decoding the top-K tokens separately gives distinct, ranked futures (GIMO, root trajectory, m)

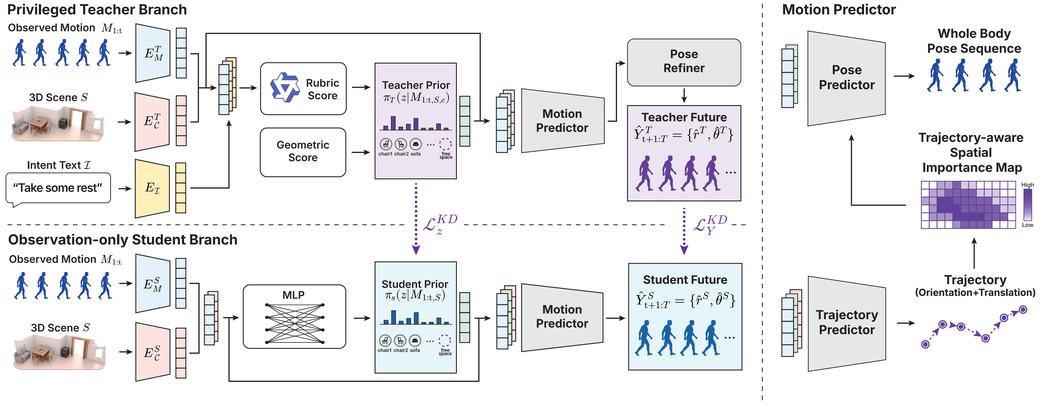
K	APD ↑	minADE@K ↓	minFDE@K ↓	FDE@minADE ↓
1	--	0.549	0.681	0.681
2	0.884	0.517	0.601	0.605
3	**0.927**	**0.467**	**0.523**	**0.571**

Method

1 **Privileged teacher (training only).** Given intent extracted by a VLM from training videos, an LLM scores how well each token affords that intent. A geometric score from observed motion identifies which same-class instance is being approached. Their fusion forms the teacher prior; an intent-aware refiner distinguishes sitting from standing.

2 **Observation-only student.** Through privileged intent distillation, a permutation-equivariant MLP over motion and token geometry learns the teacher's prior via forward KL (where motion is grounded), and its trajectory and pose via motion KD (how it moves there). No intent, gaze, VLM, or LLM is needed at inference.

3 **Motion predictor.** The top token's embedding steers the trajectory branch. Its 3D mask and the predicted path bias scene attention in the pose branch, keeping the full-body pose consistent with the chosen target.



Ablation studies

(a) Prior KD is a prerequisite for motion KD

KD z	KD Y	Traj ↓ path	Traj ↓ dest	MPJPE ↓ path	MPJPE ↓ 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) Committing to one token beats blending top-k tokens

Aggregation	Traj ↓ path	Traj ↓ dest	MPJPE ↓ path	MPJPE ↓ 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

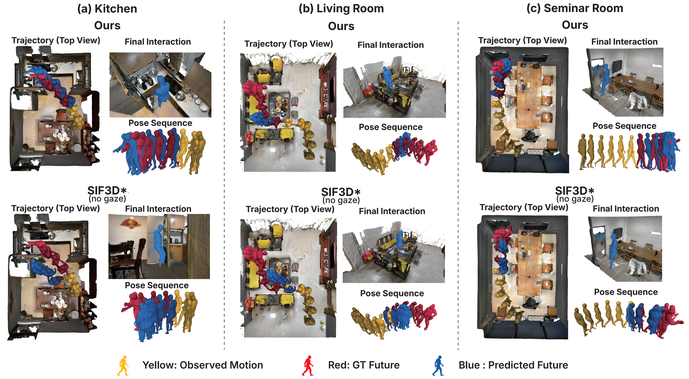
(c) Fusing intent (R) and geometry (G) gives the most accurate grounding

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

GIMO, 5 seeds; errors in mm. z/Y: prior/motion KD; R/G: rubric/geometry. Motion KD alone yields only 43% teacher-student token agreement (70% with prior KD).

Results

Best on all four metrics among observation-only models. Traj-path: 549 vs. 642 mm on GIMO (-14.5%); Traj-dest: 576 vs. 659 mm on PROX (-12.6%).



Main comparison (mm); GIMO SGIP: 5-seed mean ± std. Privileged intent beats gaze on Traj.

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

g: gaze; I: intent. † Diagnostic privileged branch; * gaze-free re-evaluation;
‡ original gaze-free variant. Bold / underline: best / second-best within block.