

Vision-Guided Action: Enhancing 3D Human Motion Prediction with Gaze-informed Affordance in 3D Scenes

Ting Yu¹ Yi Lin¹ Jun Yu² Zhenyu Lou³ Qiongjie Cui^{4†}

¹Hangzhou Normal University, ²Harbin Institute of Technology (Shenzhen)

³Zhejiang University, ⁴Singapore University of Technology and Design

yut@hznu.edu.cn

qiongjie_cui@sutd.edu.sg

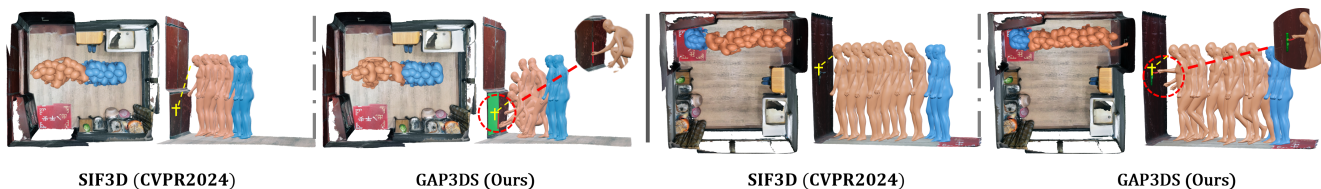


Figure 1. **The proposed GAP3DS: Gaze-informed Affordance for human pose Prediction in 3D Scenes.** Compared to the state-of-the-art SIF3D [41] utilizing the position of human gaze (yellow cross), GAP3DS is able to identify the intended object (green) and infer its affordance, where both position and affordance information are used to guide the motion prediction. It therefore achieves more accurate predicted poses, and performs significant gains for those poses where human-object interaction is expected (as highlighted in circled areas).

Abstract

Recent advances in human motion prediction (HMP) have shifted focus from isolated motion data to integrating human-scene correlations. In particular, the latest methods leverage human gaze points, using their spatial coordinates to indicate intent—where a person might move within a 3D environment. Despite promising trajectory results, these methods often produce inaccurate poses by overlooking the semantic implications of gaze, specifically the affordances of observed objects, which indicate the possible interactions. To address this, we propose GAP3DS, an affordance-aware HMP model that utilizes gaze-informed object affordances to improve HMP in complex 3D environments. GAP3DS incorporates a gaze-guided affordance learner to identify relevant objects in the scene and infer their affordances based on human gaze, thus contextualizing future human-object interactions. This affordance information, enriched with visual features and gaze data, conditions the generation of multiple human-object interaction poses, which are subsequently decoded into final motion predictions. Extensive experiments on two datasets demonstrate that GAP3DS outperforms state-of-the-art methods in both trajectory and pose accuracy, producing more physically consistent and contextually grounded predictions. For more details and code, please refer to [the project page](#).

1. Introduction

Human motion prediction (HMP) aims to forecast future human movements based on observed sequences of a person’s action. It is a crucial task in fields such as autonomous driving, and human-robot interaction, which facilitates the seamless and immediate responses of intelligent systems to human behaviors [8, 36, 41, 42, 45, 71, 73].

Standard HMP approaches have primarily relied on single-modality data, *i.e.*, predicting future sequences based solely on human motion, without considering additional contextual cues or external information [9, 32, 41, 42, 44, 45, 66]. Within this framework, researchers develop various architectures, including RNNs [6, 14, 44], CNNs [35, 52], and MLPs [18], to capture temporal dependencies and spatial dynamics in motion data. Notably, Graph Convolutional Networks (GCNs) [9, 36, 42, 67] have demonstrated particular effectiveness in HMP due to their ability to model the human skeleton’s topological structure. Although these single-modality methods have become standard, their isolated approach limits their effectiveness in complex real-world scenarios, where human actions are frequently influenced by surrounding objects and environmental context.

Recent studies have increasingly focused on incorporating contextual cues—such as environmental structures [20, 22, 65] and human gaze [41, 78]—to enhance the accuracy and realism of motion prediction. Notably, [41] utilizes real-captured point clouds to represent physical scenes

[†]The corresponding author.

and employs human gaze direction as an intent indicator to guide predictions within 3D environments. By interpreting gaze as a directional cue that suggests where a person is likely to move, [41] achieves improved trajectory predictions in 3D space. However, this approach lacks a nuanced understanding of the affordances associated with the object of focus [10, 40, 63], particularly regarding the specific actions a person intends to perform with that object. This limitation often leads to physically implausible predictions, where human-object interactions fail to align with semantic consistency, as shown in Figure 1.

To address these limitations, we propose GAP3DS, an affordance-aware human motion prediction (HMP) model that enhances realism and accuracy in real-world 3D environments. Unlike traditional methods that treat gaze as a mere directional signal [41, 78], GAP3DS interprets gaze to extract semantic insights, specifically by identifying the affordances of objects within the user’s line of sight. Understanding both where a person is likely to move and what actions they intend to perform with specific objects [24, 63] provides a richer context for motion prediction. GAP3DS begins with a gaze-guided affordance learner that identifies relevant objects and infers their affordances to predict potential human-object interactions. This step ensures that GAP3DS links gaze with actionable intent, moving beyond simple trajectory predictions. Next, the interactive pose generator produces realistic human-object interaction poses based on identified affordances and spatial orientation, effectively simulating the transitional phases of movement to ensure physical coherence and alignment with real-world constraints. This design helps to prevent issues like scene intersections or unnatural poses often seen in simpler models. Ultimately, GAP3DS employs a dual-prompted motion decoder comprising a trajectory and pose decoder, which refines these interaction poses into a cohesive, continuous motion sequence [41, 58]. This module ensures consistency between the predicted motion, observed motion, and environmental context, addressing the common issues of fragmented or unrealistic predictions found in previous methods. As a result, GAP3DS generates the physically plausible sequences that are contextually grounded and semantically aligned with the external scene.

Our main contributions are as follows: (1) We propose GAP3DS, an affordance-aware HMP model that leverages gaze-informed object affordances, producing accurate and contextually consistent predictions within 3D environments. (2) GAP3DS introduces a novel gaze-guided affordance learner that interprets gaze as an indicator of human intent, capturing both movement direction and intended interaction, thus enhancing the predictive context for human motion. (3) Extensive experiments demonstrate that GAP3DS surpasses state-of-the-art methods in trajectory and pose accuracy, delivering physically plausible and se-

mantically consistent predictions.

2. Related Work

Human Motion Prediction (HMP) has been extensively explored, with early approaches such as RNN-based models [14, 17, 44, 45] effectively capturing temporal dependencies in motion sequences but often struggling with discontinuity and error accumulation over longer horizons. Recent advancements with CNN [33], transformer [1, 2], and GCN-based models [9, 36, 42] have improved prediction accuracy and sequence coherence.

Despite these advancements, most models overlook the crucial role of 3D environmental context in motion prediction. Scene-aware models [3, 7, 43, 65, 78] attempt to address this by emphasizing human-scene interactions. For example, MDP [65] enhances contextual grounding by modeling spatial relationships between humans and their surroundings, resulting in more realistic trajectory and pose predictions. Incorporating human gaze as an indicator of intent further enriches prediction accuracy. Models such as BiFU [78] refine predictions using gaze to navigate 3D scenes, while SIF3D [41] introduces a multimodal framework that leverages gaze information to improve prediction outcomes. However, these methods generally treat gaze as a simple spatial cue, missing the deeper semantic insights that gaze can provide about human-object interactions. Our approach, GAP3DS, addresses this limitation by using gaze-guided affordance learning to inform interaction-aware predictions, advancing HMP in real-world 3D settings.

Scene-aware Motion Synthesis focuses on generating realistic human motions within 3D environments based on specific instructions [7, 22, 78]. Recent advancements include whole-body grasping with small objects [54] and dynamic object interactions [37]. Techniques such as diffusion models [23] and reinforcement learning [77] have further refined adaptive interactions and navigation. Additionally, multimodal approaches that integrate scene context with text have provided greater control over motion generation [27, 62, 77], with models like WANDR [11] utilizing intention functions for guided synthesis. Affordance-informed methods, such as AffordMotion [63], bridge 3D scene grounding and conditional motion synthesis by using affordances to guide interactions. Our approach builds on these advances by learning the affordances of objects targeted by human gaze, enabling motion predictions that align more closely with human intent within 3D scenes.

Scene and Affordance Understanding. Scene understanding [10, 21, 72, 76] is critical for applications in robotics [4] and embodied AI [60, 69, 70]. Affordances [15], typically defined as the potential actions an environment or object enables, play a key role in scene understanding [29, 49, 68]. Early studies on affordances focused on object manipulation from image-level observa-

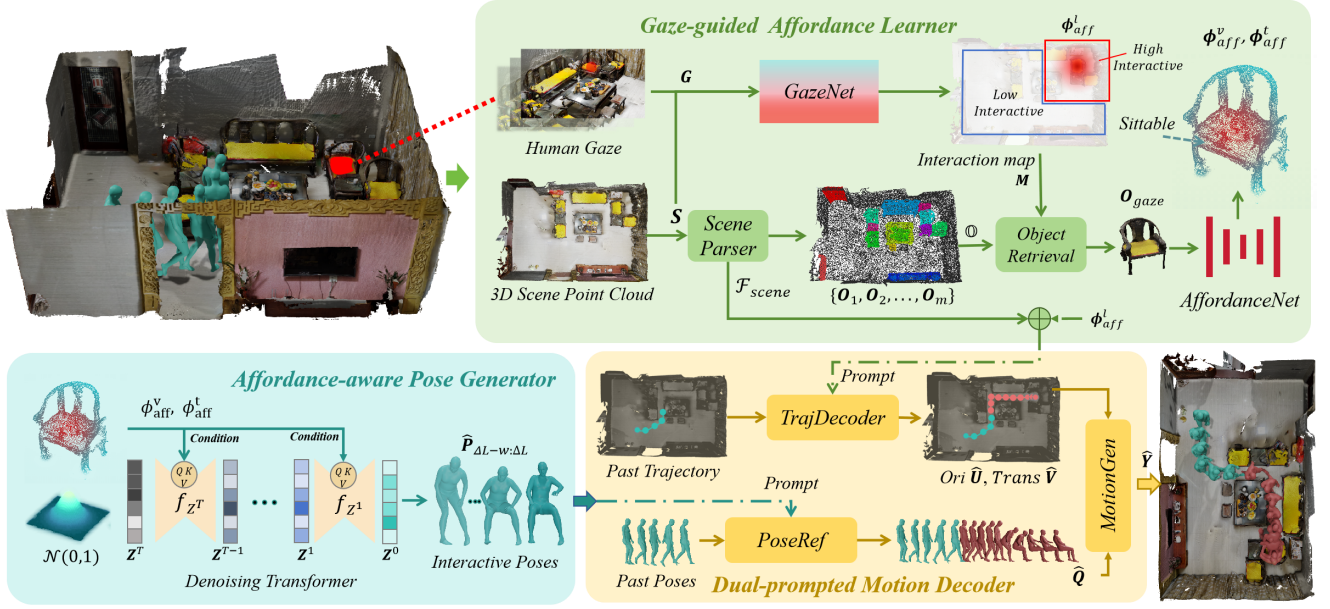


Figure 2. Illustration of the proposed GAP3DS. GAP3DS integrates three main components to generate affordance-aware human motion predictions in 3D scenes. (1) Gaze-guided Affordance Learner leverages human gaze $\mathbf{G}$ to identify the intended object $\mathbf{O}_{\text{gaze}}$ and infer its affordance $\Phi_{\text{aff}} = (\phi_{\text{aff}}^l, \phi_{\text{aff}}^v, \phi_{\text{aff}}^t)$, with respect to the interaction location $\Phi_{\text{aff}} = (\phi_{\text{aff}}^l, \text{visual semantics } \phi_{\text{aff}}^v, \text{and textual context } \phi_{\text{aff}}^t)$; (2) Affordance-aware Pose Generator uses the predicted affordance ϕ_{aff}^v and ϕ_{aff}^l to produce interaction poses $\hat{\mathbf{P}}_{\Delta L-w:\Delta L}$; (3) Dual-Prompted Motion Decoder, taking the interactive location ϕ_{aff}^l and the generated interactive poses $\hat{\mathbf{P}}_{\Delta L-w:\Delta L}$ as the input, can achieve the translation $\hat{\mathbf{U}}$ and orientation $\hat{\mathbf{V}}$, as well as predicted poses $\hat{\mathbf{Q}}$, which are then used to predict the final future motions $\hat{\mathbf{Y}}_{1:\Delta L}$.

tions [13, 16, 19, 30, 31, 38, 47, 48, 61, 79, 80], while later research began to incorporate human-object interactions [24, 34, 55, 64] to enhance 3D scene understanding [5, 10, 21, 76].

Recent work has shifted toward explicit 3D affordance representations, such as contact maps for grasping [25, 28, 39, 40] and distance maps for scene-aware motion synthesis [63, 75]. Unlike traditional affordance learning methods that apply affordances across an entire scene, our approach uniquely leverages gaze-informed affordances to highlight objects most relevant to human intention.

3. Proposed Method

Figure 2 illustrates the overall procedure of GAP3DS, a framework that enhances human motion prediction by using gaze to infer object affordances in 3D scenes. Unlike previous methods that treat gaze as a simple directional cue, GAP3DS interprets gaze as an indicator of interaction intent, enabling predictions that are both physically plausible and semantically aligned with real-world interactions.

3.1. Problem Setup

Given an observed motion sequence $\mathbf{X}_{1:L} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_L\}$, each pose frame $\mathbf{x}_k = (\mathbf{t}_k, \mathbf{r}_k, \mathbf{p}_k)$ at time step k consists of global translation $\mathbf{t}_k \in \mathbb{R}^3$, global orientation $\mathbf{r}_k \in SO(3)$, and body pose embedding $\mathbf{p}_k \in \mathbb{R}^{32}$, allowing full body mesh reconstruction via the

SMPL-X model [51]. The surrounding 3D environment is represented as a point cloud $\mathbf{S} \in \mathbb{R}^{N \times 3}$, where N denotes the number of sampled points, and we incorporate human gaze information, represented as a sequence $\mathbf{G} = \{\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_L\} \in \mathbb{R}^{L \times 3}$, where each $\mathbf{g}_k$ is the 3D intersection coordinate indicating the gaze target. With the network parameter θ , our task is to predict the future motion $\hat{\mathbf{Y}}_{1:\Delta L} = \{\hat{\mathbf{x}}_{L+1}, \hat{\mathbf{x}}_{L+2}, \dots, \hat{\mathbf{x}}_{L+\Delta L}\}$, $\mathbf{Y}_{1:\Delta L}$ being the ground truth, for a horizon ΔL , formulated as:

$$\arg \max_{\theta} P(\mathbf{Y}_{1:\Delta L} | \mathbf{X}_{1:L}, \mathbf{S}, \mathbf{G}; \theta). \quad (1)$$

3.2. Gaze-guided Affordance Learner (GAL)

GAL uses human gaze as a reliable indicator of interaction intent, identifying the primary object of focus and extracting its affordance attributes to enhance contextual understanding in motion prediction. While gaze may sometimes drift [78], sustained focus on a specific object—such as a chair—often signals an intent to engage with it.

To quantify gaze interaction with scenes, we compute the distance matrix $\mathbf{D} \in \mathbb{R}^{N \times L}$ between each gaze point $\mathbf{g}_j \in \mathbf{G}$ and each scene point $\mathbf{s}_i \in \mathbf{S}$:

$$\mathbf{D}_{ij} = \|\mathbf{s}_i - \mathbf{g}_j\|_2, \quad (2)$$

where N and L denote the number of scene points and gaze frames. We then apply a 3-layer 1D CNN, GazeNet (details in supp. mat.), to aggregate gaze data and generate an

interaction map $M \in \mathbb{R}^{N \times 1}$:

$$M = \text{GazeNet}(\mathcal{D}), \quad (3)$$

where higher values in M indicate the regions of likely interaction.

To determine the intended target, we encode the scene into a compact feature representation $\mathcal{F}_{\text{scene}} = \text{SceneEncoder}(\mathcal{S})$ and parse it into a set of objects $\mathbb{O} = \{\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_m\}$, where m is the total number of objects in the point clouds of scene. We employ the off-the-shelf SoftGroup [59] to encode and parse the scene. The object with the highest cumulative interaction score in M is selected as the intended target $\mathcal{O}_{\text{gaze}} \in \mathbb{R}^{n_b \times 3}$, with n_b being the number of point cloud of the object:

$$\mathcal{O}_{\text{gaze}} = \arg \max_{\mathcal{O}_i \in \mathbb{O}} \sum_{p \in \mathcal{O}_i} M[p], \quad (4)$$

where $M[p]$ is the interaction score for each point p within a object $\mathcal{O}_i$. For the target $\mathcal{O}_{\text{gaze}}$, AffordanceNet [12] is used to derive visual features $\phi_{\text{aff}}^v \in \mathbb{R}^{n_b \times c_e}$ and textual affordance attributes $\phi_{\text{aff}}^t \in \mathbb{R}^{1 \times c_h}$:

$$\phi_{\text{aff}}^v, \phi_{\text{aff}}^t = \text{AffordanceNet}(\mathcal{O}_{\text{gaze}}). \quad (5)$$

Here, c_e and c_h are the dimensions of visual and textual affordance features.

The resulting affordance triplet $\Phi_{\text{aff}} = (\phi_{\text{aff}}^l, \phi_{\text{aff}}^v, \phi_{\text{aff}}^t)$ includes: ϕ_{aff}^l : The interaction map M serves as a spatial anchor to guide trajectory prediction, and for the sake of uniformity, it can also be expressed as ϕ_{aff}^l . ϕ_{aff}^v : Visual characteristics (e.g., shape) of the object. ϕ_{aff}^t : High-level semantics that describe the object's affordances. Each component provides complementary information to guide accurate trajectory and pose predictions in the subsequent stage.

3.3. Affordance-aware Pose Generator (APG)

Human gaze often signals intent, guiding actions beyond simple directional cues [78]. To generate interaction poses that align with inferred affordance attributes, APG employs a motion diffusion model, the affordance-aware pose diffusion model. This model uses forward and reverse processes to create human-object interaction poses $\hat{P}_{\Delta L-w:\Delta L}$, where w is the number of generated frames. By conditioning pose generation on affordance cues, APG ensures the resulting motion is both contextually and physically plausible, enhancing GAP3DS's capacity for realistic, intention-driven motion prediction in 3D environments.

For simplicity, we denote the sequence $P_{\Delta L-w:\Delta L}$ at noising step t as Z^t . As $t \rightarrow \infty$, Z^t converges to an isotropic Gaussian distribution. The forward process is:

$$q(Z^t | Z^0) = \mathcal{N}(Z^t; \sqrt{\bar{\alpha}_t} Z^0, (1 - \bar{\alpha}_t) \mathbf{I}), \quad (6)$$

where $\bar{\alpha}_t = \cos^2 \left(\frac{t+0.008}{1.008} \cdot \frac{\pi}{2} \right)$ represents the cumulative α at time step t .

In the affordance-guided reverse denoising process, APG progressively denoises Z^t . Following [56, 57, 74], we directly predict the denoised sample iteratively across reverse steps rather than predicting noise. In each reverse step t , the model predicts the denoised sample $\bar{Z}^0$ from the current state Z^t , then adds Gaussian noise for the $(t-1)$ -th step, simulating reverse diffusion. This process iteratively refines the state from $t = T$ down to $t = 0$, producing the final clean sample $\hat{Z}^0$. The reverse diffusion process is conditioned on the visual feature ϕ_{aff}^v and textual affordance ϕ_{aff}^t :

$$\bar{Z}^0 = \text{PoseGenerator}(Z^t, \phi_{\text{aff}}^v, \phi_{\text{aff}}^t), \quad (7)$$

where the PoseGenerator is a dual-layer transformer decoder. After T denoising steps, APG produces a cohesive pose sequence $\hat{Z}^0$, namely $\hat{P}_{\Delta L-w:\Delta L}$.

3.4. Dual-Prompted Motion Decoder (DPM)

To ensure that the predicted motion is consistent with the observed motion and the environmental context, and to address issues of fragmentation and unrealistic motion, we utilize interactive location ϕ_{aff}^l and interactive poses $\hat{P}_{\Delta L-w:\Delta L}$ to refine the poses into a cohesive, continuous motion sequence.

For trajectory consistency, the interactive location ϕ_{aff}^l serves as a prompt for global translation and orientation, anchoring the motion trajectory within the scene layout. Specifically, we adopt a standard transformer-based decoder, TrajDecoder [41], to combine the observed motion sequence $X_{1:L}$ with the interactive location and scene context, producing the global translation $\hat{U}$ and orientation $\hat{V}$:

$$\hat{U}, \hat{V} = \text{TrajDecoder}(X_{1:L}, [\mathcal{F}_{\text{scene}}, \phi_{\text{aff}}^l]). \quad (8)$$

Here, $[\cdot]$ denotes a concatenation operation. For the final pose result, interactive poses $\hat{P}_{\Delta L-w:\Delta L}$ are used to prompt refinement of the final pose sequence, ensuring that predicted poses align with intended actions. The final pose sequence $\hat{Q}$ is decoded by PoseRef [41]:

$$\hat{Q} = \text{PoseRef}(X_{1:L}, \hat{P}_{\Delta L-w:\Delta L}). \quad (9)$$

The combined predictions of $\hat{U}$, $\hat{V}$, and $\hat{Q}$ are processed through the SMPL-X [51] to reconstruct the human body, with VPoser [51] refining the pose. The final continuous motion sequence $\hat{Y}$ is assembled by MotionGen [41]:

$$\hat{Y} = \text{MotionGen}(\text{SMPL-X}(\hat{U}, \hat{V}, \hat{Q}), \text{VPoser}(\hat{Q})). \quad (10)$$

By using the interactive location ϕ_{aff}^l and interactive poses $\hat{P}_{\Delta L-w:\Delta L}$ to guide both trajectory and pose prediction, GAP3DS achieves contextually consistent future motion.

Method	GAL	APG	Trajectory Deviation (in <i>mm</i>)		Pose MPJPE (in <i>mm</i>)		Penetration Error [%] ↓
			Traj-path↓	Traj-interact ↓	MPJPE-path ↓	MPJPE-interact ↓	
AuxFormer [66]	×	×	688	893	173.5	205.4	6.65
BiFU [78]	×	×	661	727	167.8	205.0	3.73
MDP [65]	×	×	612	679	157.1	190.4	4.77
SIF3D [41]	×	×	590	666	156.6	195.7	4.59
AuxFormer [66]	✓	×	637	688	159.8	196.4	4.84
BiFU [78]	✓	×	650	696	165.1	198.9	3.79
MDP [65]	✓	×	601	658	157.0	190.8	4.56
SIF3D [41]	✓	×	583	656	155.1	194.6	3.61
AuxFormer [66]	✓	✓	621	672	156.1	193.1	3.64
BiFU [78]	✓	✓	630	677	160.9	194.1	3.60
MDP [65]	✓	✓	597	649	152.0	184.1	4.51
SIF3D [41]	✓	✓	580	644	150.2	186.1	3.24
GAP3DS	✓	✓	575	623	141.2	171.4	0.95

Table 1. Performance comparison on trajectory deviation, MPJPE, and penetration error in the GIMO dataset [78]. This table evaluates the impact of the Gaze-guided Affordance Learner (GAL) and Affordance-aware Pose Generator (APG) on various baselines. By incorporating both GAL and APG, GAP3DS achieves the best performance across all metrics, demonstrating its effectiveness in producing accurate, physically consistent human motion predictions within 3D environments. Bold values indicate the best results.

4. Experiments

4.1. Experimental Setup

Datasets: (1) **GIMO** [78] consists of $\approx 129K$ frames of high-quality human motion data with gaze information in real-world 3D environments. It includes SMPL-X pose representations across 14 scenes, each captured with a 3D LiDAR sensor, producing detailed 3D point clouds with 10 20 objects and about 500K vertices per scene. Following SIF3D [41], we down-sample sequences to 2fps and adopt the official scene-based splits for training and testing. (2) **GTA-IM** [3] provides 1M frames of human poses in indoor settings, along with segmentation and camera pose data. It features 50 unique characters in 10 indoor scenes, supporting diverse interaction activities valuable for human motion analysis. We note that due to the space limitation, the result of GTA-IM is reported in Supp.mat.

Baselines: We compare GAP3DS with 4 methods: Auxformer [66], BiFu [78], MDP [65], and SIF3D [41]. Auxformer [66] integrates spatial-temporal dependencies in an auxiliary-adapted transformer framework to improve motion prediction. MDP [65] enhances human-scene interactions by leveraging the mutual distance between the human body and the surrounding environment. BiFU [78] incorporates gaze coordinates to predict future human activities, offering insights into ego-centric interactions. SIF3D [41] achieves state-of-the-art performance by implementing a multimodal sense-informed predictor tailored for real-world 3D scenarios. Moreover, we discuss scene-aware motion synthesis methods in Supp.mat, even if they are not directly applicable to the HMP task.

Metrics: Following SIF3D [41], we evaluate GAP3DS with three primary metrics: (1) Trajectory Metrics, where

Traj-path measures average trajectory deviation to assess path alignment, and Traj-interaction evaluates final pose accuracy for target-reaching precision; (2) Pose Metrics, using Mean Per Joint Position Error (MPJPE)[36, 42], with MPJPE-path measuring overall pose consistency, and MPJPE-interaction assessing joint accuracy at the endpoint to emphasize interaction precision; and (3) Penetration Error, calculated as the percentage of intersecting points between the human body and scene point cloud using the Signed Distance Function (SDF)[50], to assess physical realism by penalizing intersections with scene objects.

4.2. Main Results and Analysis

Table 1 showcases the effectiveness of GAP3DS compared to baselines across multiple evaluation metrics, including trajectory deviation, pose accuracy, and physical consistency. The scene-independent model AuxFormer [66] demonstrates relatively low prediction accuracy and higher penetration error, suggesting that its lack of scene-awareness limits its capacity to generate spatially aligned and contextually consistent predictions. Scene-aware models, such as BiFU [78], MDP [65], and SIF3D [41], achieve notable improvements in trajectory and pose accuracy, with SIF3D outperforming others by leveraging gaze location as an indicator of human intent. However, even with gaze cues, SIF3D struggles to capture detailed interaction information, as indicated by its MPJPE-interact score. GAP3DS addresses this limitation through its gaze-guided affordance learning, which provides a richer semantic context for motion prediction. By integrating affordances with gaze, GAP3DS achieves the lowest scores in trajectory deviation, MPJPE, and penetration error, emphasizing its advantage in predicting accurate and physically

Method	Scene Time	Bedroom			Classroom			Garden			Kitchen			Lab			Livingroom			Seminarroom		
		0.5s	2.0s	5.0s	0.5s	2.0s	5.0s	0.5s	2.0s	5.0s	0.5s	2.0s	5.0s	0.5s	2.0s	5.0s	0.5s	2.0s	5.0s	0.5s	2.0s	5.0s
AuxFormer [66]	Traj	185	541	1162	319	442	529	232	630	972	172	509	625	248	745	1113	151	621	1030	273	606	881
	MPJPE	114.9	173.2	259.7	137.1	178.7	162.4	105.3	230.5	229.9	128.5	108.0	125.1	113.1	144.0	231.9	156.3	132.7	259.8	130.9	160.7	204.9
AuxFormer [66] w/ GAL & APG	Traj	179	471	766	282	392	472	183	547	666	165	468	537	236	672	863	146	587	864	248	579	623
	MPJPE	105.7	152.6	237.6	125.6	134.6	154.5	97.6	210.3	223.9	108.2	95.6	114.5	98.5	124.0	215.5	122.1	113.1	226.6	112.4	144.5	195.6
MDP [65]	Traj	165	543	673	176	433	576	194	481	642	165	451	439	259	611	685	246	629	832	343	504	620
	MPJPE	114.9	174.8	281.4	95.6	165.4	157.4	101.0	188.3	210.5	91.0	104.5	124.2	79.5	91.0	230.6	102.3	130.5	210.0	103.5	140.5	186.7
MDP [65] w/ GAL & APG	Traj	159	530	660	165	419	556	190	470	599	158	432	428	251	594	669	230	599	786	329	496	605
	MPJPE	104.2	159.9	253.6	93.1	157.2	145.7	93.3	183.6	199.9	90.8	99.5	106.7	82.4	92.4	216.7	99.4	124.5	204.2	102.1	135.6	181.0
BiFU [78]	Traj	232	540	838	332	442	689	218	532	659	275	467	582	273	551	681	340	607	794	417	659	629
	MPJPE	67.0	115.0	185.9	86.6	144.2	171.9	89.4	205.0	251.7	98.0	104.0	133.0	53.9	110.8	229.8	95.7	126.1	236.1	107.2	137.3	179.4
BiFU [78] w/ GAL & APG	Traj	212	529	644	298	413	663	222	527	644	266	489	591	270	534	664	310	563	767	402	631	622
	MPJPE	66.0	109.1	176.0	82.5	135.5	160.6	89.2	203.1	223.7	95.6	99.4	120.5	53.7	105.4	218.4	93.2	124.4	228.7	106.0	134.7	175.6
SIF3D [41]	Traj	183	441	554	155	424	554	200	472	533	176	471	444	245	540	730	257	602	794	219	440	611
	MPJPE	82.3	105.7	185.9	100.3	147.1	160.1	74.0	190.4	199.0	100.1	101.5	114.3	65.4	94.3	218.3	108.9	121.1	188.4	101.8	137.3	179.4
SIF3D [41] w/ GAL & APG	Traj	165	413	496	158	401	532	196	466	521	166	464	431	238	531	713	250	588	706	210	420	598
	MPJPE	81.1	99.5	179.5	93.2	135.3	146.0	73.6	185.5	189.1	98.6	96.4	103.0	64.9	93.4	214.5	102.6	115.2	185.9	100.9	134.5	173.5
Ours	Traj	180.1	369	478	192	425	356	196	444	489	174	393	415	250	496	622	185	542	649	186	407	589
	MPJPE	95.0	101.4	175.5	100.2	125.9	139.1	90.4	180.5	185.6	90.4	94.2	99.5	56.7	94.2	150.5	103.4	125.6	176.2	103.0	126.7	172.0

Table 2. Experimental results of trajectory deviation and pose MPJPE (in *mm*) across diverse scenarios on GIMO [78] dataset. Our GAP3DS demonstrates strong performance across most scenarios, particularly excelling in longer-term forecasting.

consistent human motion within 3D environments.

Table 2 further evaluates each model’s performance across different timestamps and diverse sub-scenarios, assessing the robustness of each model in both short-term and long-term prediction contexts. While baseline models generally maintain reasonable accuracy in short-term predictions, their performance degrades over long-term (5-second) forecasting, with declines in trajectory coherence and pose precision. GAP3DS consistently outperforms baselines across both pose and trajectory metrics in all scenarios, particularly in long-term predictions where its accuracy remains robust. This long-term consistency is attributed to GAP3DS’s dual-prompted mechanism, which combines interactive location as a trajectory anchor with affordance-informed poses for enhanced motion refinement. Additionally, GAP3DS demonstrates notable strengths in complex environments, such as kitchens and bedrooms, where interactions are frequent and require fine-grained predictive detail. Its ability to anticipate these complex interactions effectively underscores the impact of integrating affordance cues, which provide spatial and contextual depth essential for navigating intricate settings. Overall, these results validate GAP3DS’s capability to generate motion sequences that are not only accurate and realistic but also contextually and physically aligned within challenging 3D environments.

Overall, GAP3DS demonstrates that affordance-aware human motion prediction not only enhances trajectory and pose accuracy but also ensures contextual relevance and physical plausibility in challenging 3D environments.

4.3. Visualization Comparison

Figure 3 illustrates the advantages of GAP3DS over SIF3D [41] in terms of interaction realism, trajectory fi-

delity, and physical consistency across various indoor scenarios. **(1)** In the **kitchen scenario** (subfigure a), GAP3DS achieves precise alignment with target objects, generating realistic interaction poses, such as reaching towards specific items or avoiding nearby obstacles. These poses demonstrate GAP3DS’s capacity to adaptively shape motions based on the affordances of objects in the scene. In contrast, SIF3D only approximates target positions without creating meaningful interactions, often resulting in generalized or non-interactive motions, as highlighted by the red boxes. **(2)** GAP3DS effectively maintains **trajectory alignment** over extended sequences, showing robustness in longer-term predictions. For example, in the **living room scenario** (subfigure b), GAP3DS consistently follows the intended trajectory, reflecting the model’s ability to integrate contextual cues and maintain spatial coherence throughout the sequence. By comparison, SIF3D frequently deviates from the intended path as the sequence progresses, leading to drift and noticeable misalignment with the target trajectory, as indicated in the yellow boxes. **(3)** In terms of **physical consistency and collision avoidance**, GAP3DS generates motion sequences that respect spatial boundaries, avoiding collisions with scene objects. For instance, in the **bedroom scenario** (subfigure c), GAP3DS adjusts poses to prevent intersection with the bed, generating physically plausible interactions that respect environmental constraints. Conversely, SIF3D often produces poses that intersect with objects such as the bed frame, resulting in physically implausible configurations, as illustrated in the green boxes.

We conclude that, GAP3DS brings significant improvements over the state-of-the-art SIF3D [41], with producing accurate, contextually consistent trajectories, physically plausible interactions, and more natural sequence.

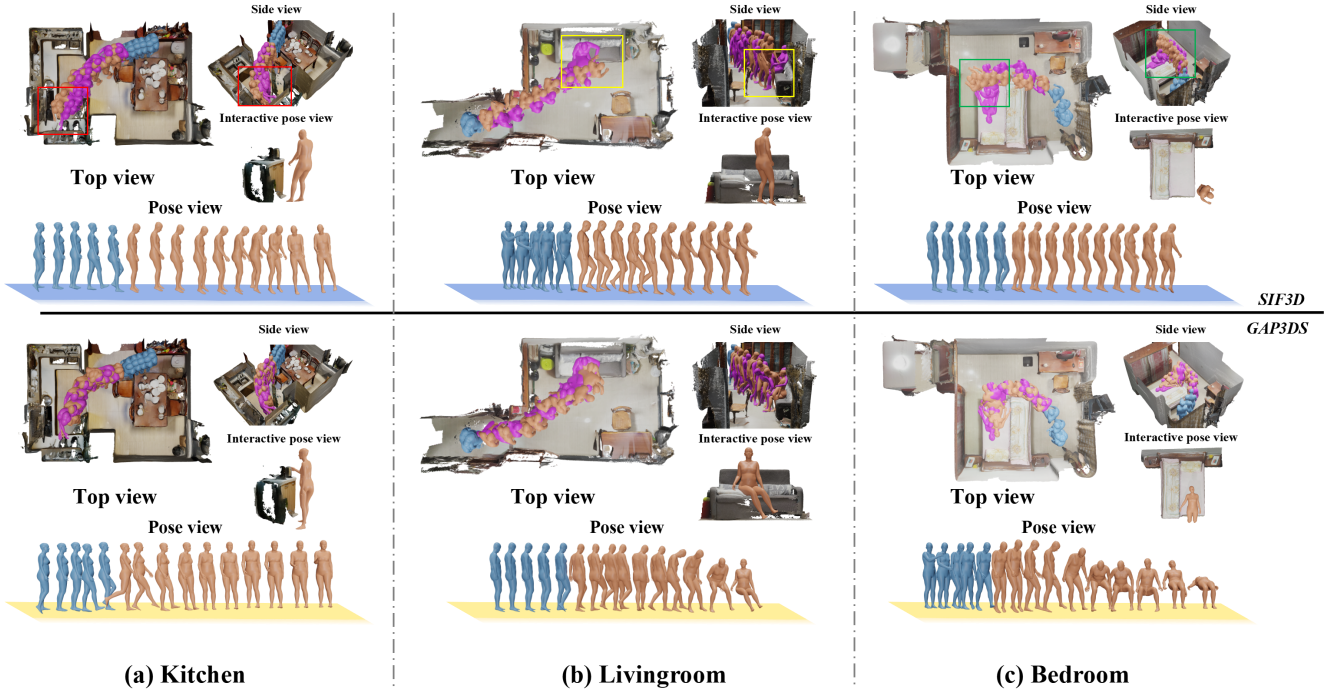


Figure 3. Visual comparison of GAP3DS with the state-of-the-art SIF3D [41] across various indoor scenarios: (a) kitchen, (b) living room, and (c) bedroom. Each scene is shown from *top view*, *side view*, *interactive pose view*, and *pose view*. The **purple** human meshes denote the ground truth, the **brown** meshes show model predictions, and the **blue** meshes represent historical motion sequences. GAP3DS exhibits improved accuracy and natural motion prediction over SIF3D, particularly in trajectory alignment, collision avoidance, and generating realistic human-object interactions within the scene.

4.4. Ablation Studies

We conduct extensive ablation studies to assess the contributions of each core component in GAP3DS.

Gaze-guided Affordance Learner (GAL). Table 3a highlights the role of GAL in enhancing GAP3DS’s performance. Removing GAL results in a marked increase in trajectory deviation and pose error, which underlines its essential function in interpreting affordance cues from gaze direction, thus refining motion accuracy.

Further analysis in Table 3b investigates the influence of each attribute in the affordance triplet. The interactive location is shown to have a stronger effect on trajectory alignment, providing crucial spatial context, while the visual and textual attributes contribute to pose accuracy by capturing interaction details. Together, these attributes complement one another in accurate motion prediction.

Moreover, adapting GAL to baseline models (*e.g.*, MDP [65], BiFU [78], SIF3D [41]), as presented in Table 1 and 2, consistently enhances their trajectory and pose accuracy. This demonstrates that gaze-informed affordances, as introduced in GAL, not only strengthen GAP3DS but also provide a flexible and valuable context for existing baseline models, facilitating a more detailed understanding of the scene and human intention.

Affordance-aware Pose Generator (APG). We evaluate the influence of APG in generating interactive poses,

particularly in relation to pose accuracy. Removing APG, as shown in Table 3a, significantly increases pose prediction error, emphasizing its essential role in capturing interaction details necessary for accurate final pose generation.

Further analysis of the number of interactive poses, presented in Table 3d, indicates that 3 interactive poses achieve an optimal balance by enhancing pose detail without adding redundancy. This setup not only improves pose accuracy but also indirectly supports trajectory coherence by enabling smoother translations.

Lastly, we assess APG’s adaptability when integrated into baseline models, as detailed in Table 1 and 2. The addition of APG enhances pose accuracy, particularly in frequent interaction scenarios. This improvement highlights APG’s effectiveness in establishing a solid foundation for coherent human-object interactions that contribute to smooth motion sequences.

Dual-Prompted Mechanism. We first study the impact of the DPM decoder in improving GAP3DS’s performance. Table 3a shows an immediate performance drop in both Traj and MPJPE when we replace DPM with one basic decoder, demonstrating DPM’s effectiveness for trajectory alignment and pose accuracy. Table 3c further assesses the impact of the dual-prompted mechanism performed on the DPM decoder. Four configurations are analyzed: (i) single pose-prompt, where interactive poses alone guide the pose de-

Components	Traj-P	Traj-I	MPJPE-P	MPJPE-I	Afford. Type	Traj-P	Traj-I	MPJPE-P	MPJPE-I	Prom. Mech.	Traj	MPJPE
w/o GAL	642	701	170.1	206.4	w/o ail	587	668	146.7	174.2	single pose-prom.	585	148.3
w/o APG	586	632	164.6	193.2	w/o avs	578	627	152.1	186.4	single afford-prom.	586	164.1
w/o DPM	582	632	144.2	175.7	w/o ats	577	627	154.2	189.0	reverse-prom.	637	171.2
full model	575	623	141.2	171.4	full triplet	575	623	141.2	171.4	standard dual-prom.	575	141.2

(a) Model components analysis. Traj-P is for Traj-pose, Traj-I for Traj-interaction, MPJPE-P for MPJPE-pose, and MPJPE-I for MPJPE-interaction metric. We observe that GAL, APG, and DPM contributes to performance, and the full model achieves the best result.

(b) Affordance triplets study. Combining all affordance triplet components: location (ail), visual (avs), and textual (ats) attributes, yields the best results, indicating their complementary roles in enhancing both trajectory alignment and interaction accuracy.

(c) Dual-prompted mechanism. Using interactive location for trajectory and interactive poses for pose prediction, achieves superior performance.

Inter. Pose	Traj-P	Traj-I	MPJPE-P	MPJPE-I	Scene Points	512	2048	8192	32768	SceneParser	Traj	MPJPE
1	580	626	146.2	173.8	Traj-P (mm)	631	592	575	575	SPFormer [46]	575	150.2
3	575	623	141.2	171.4	MPJPE-P (mm)	166.4	156.2	141.2	141.0	SphericalMask [53]	580	146.8
5	575	624	142.9	172.6	VRAM (MB)	5420	6124	11249	21085	PointGroup [26]	576	143.3
7	577	625	143.0	174.5	Speed (sample/s)	29.6	24.5	10.3	3.1	SoftGroup [59]	575	141.2

(d) Interactive poses number analysis. Generating 3 interactive poses provides the optimal balance between spatial coherence and interaction precision, improving both trajectory and pose accuracy without redundancy.

(e) Scene density and computation efficiency. Increasing scene point density improves prediction accuracy but comes at the cost of higher VRAM and slower processing. A moderate density of 8192 points strikes a balance between efficiency and accuracy.

(f) SceneParser analysis. High-quality SceneParsers improve accuracy by capturing finer object boundaries, though GAP3DS maintains robustness even with simpler SceneParser models.

Table 3. **Ablation Studies for GAP3DS.** All results are evaluated on the GIMO [78] testing set. Each subtable investigates the contributions of individual model components, the affordance attributes, the effectiveness of dual-prompt mechanisms, the optimal number of interactive poses, the trade-offs between scene density and computation efficiency, and the impact of segmentation quality on model performance.

coder; (ii) single affordance-prompt, where only interactive location informs the trajectory decoder; (iii) reverse-prompt, with interactive location guiding the pose decoder and interactive poses prompting the trajectory; and (iv) the standard dual-prompt setup, combining interactive location for trajectory prediction with interactive poses for final pose generation.

The findings reveal that: (1) Using only interactive location improves trajectory alignment yet lacks pose details, while relying solely on interactive poses enhances pose accuracy at the expense of trajectory coherence. This indicates that interactive location is more effective for global motion guidance, while interactive poses provide critical details for accurate human-object interactions; (2) Applying prompts in reverse (affordance for pose and interactive poses for trajectory) results in lower accuracy across both trajectory and pose, highlighting the specialized role of each prompt; (3) The optimal configuration affirms the dual-prompted mechanism’s effectiveness in harmonizing trajectory and pose predictions, resulting in accurate and physically consistent motion sequences.

Scene Density and Computation Efficiency. Table 3e evaluates the impact of scene point density on GAP3DS’s performance and computational efficiency. As scene point density increases, GAP3DS demonstrates improved trajectory and pose predictions, highlighting the value of denser environmental representations for spatial awareness. However, this accuracy gain comes at the cost of increased VRAM usage and slower processing speeds. Moderate densities (8192) achieve an effective balance, enabling GAP3DS to retain high performance without excessive computational overhead, making it suitable for real-time ap-

plications while maintaining precision.

SceneParser Analysis. Table 3f study the impact of SceneParser quality on GAP3DS’s performance. Given the importance of clear object boundaries in informing motion and interaction, SceneParser quality directly affects the model’s environmental understanding. Results indicate that using high-resolution SceneParsers, such as SoftGroup [59], enhances GAP3DS’s ability to accurately detect interaction regions, leading to better trajectory alignment and pose accuracy. Notably, GAP3DS demonstrates robustness across varying segmentation qualities, maintaining certain performance even with simpler SceneParsers.

5. Conclusion

In this paper, we present GAP3DS, an affordance-aware model for human motion prediction in complex 3D environments. By leveraging gaze-guided affordance cues, GAP3DS advances beyond traditional approaches that treat gaze merely as a directional signal, enabling a deeper understanding of interaction intent by identifying target objects and interpreting their affordances. This capability allows GAP3DS to generate human-object interaction poses that are contextually accurate and physically coherent, providing a strong foundation for final motion decoding. Through extensive evaluations, GAP3DS consistently outperforms state-of-the-art methods across key metrics, achieving superior trajectory alignment, pose accuracy, and collision avoidance. Our approach sets a new benchmark for affordance-informed motion prediction, demonstrating the importance of integrating semantic cues for generating realistic, context-aware human motions in 3D scenes.

Acknowledgements

This work was supported by the National Natural Science Foundation of China No. 62125201, U24B20174, 62002314, and the Zhejiang Provincial Natural Science Foundation of China under Grant No. LY23F020005.

References

- [1] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A Spatio-temporal Transformer for 3D Human Motion Prediction. In *International Conference on 3D Vision (3DV)*, pages 565–574, 2021. 2
- [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)*, pages 565–574, 2021. 2
- [3] 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)*, pages 387–404. Springer, 2020. 2, 5
- [4] Zhihao Cao, Zidong Wang, Siwen Xie, Anji Liu, and Lifeng Fan. Smart help: Strategic opponent modeling for proactive and adaptive robot assistance in households. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18091–18101, 2024. 2
- [5] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ Scene Understanding: Single-view 3D Holistic Scene Parsing and Human Pose Estimation with Human-object Interaction and Physical Commonsense. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8648–8657, 2019. 3
- [6] Hsu-Kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-Agnostic Human Pose Forecasting. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432, 2019. 1
- [7] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware Human Motion Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6992–7001, 2020. 2
- [8] Qiongjie Cui and Huaijiang Sun. Towards Accurate 3D Human Motion Prediction From Incomplete Observations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4801–4810, 2021. 1
- [9] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning Dynamic Relationships for 3D Human Motion Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6519–6527, 2020. 1, 2
- [10] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14531–14542, 2024. 2, 3
- [11] Markos Diomatari, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J Black. Wandr: Intention-guided human motion generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 927–936, 2024. 2
- [12] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018. 4
- [13] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2139–2147, 2018. 3
- [14] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent Network Models for Human Dynamics. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4346–4354, 2015. 1, 2
- [15] JJ Gibson. The theory of affordances. *Perceiving, acting and knowing: Towards an ecological psychology*/Erlbaum, 1977. 2
- [16] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1529–1536, 2011. 3
- [17] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José M. F. Moura. Adversarial Geometry-Aware Human Motion Prediction. In *European Conference on Computer Vision (ECCV)*, pages 786–803, 2018. 2
- [18] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4809–4819, 2023. 1
- [19] Abhinav Gupta, Scott Satkin, Alexei A. Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1961–1968, 2011. 3
- [20] Xiao Han, Yiming Ren, Yichen Yao, Yujing Sun, and Yuexin Ma. Towards practical human motion prediction with lidar point clouds. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7629–7638, 2024. 1
- [21] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3d scenes by learning human-scene interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14703–14713. 2, 3
- [22] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11374–11384, 2021. 1, 2
- [23] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [24] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *IEEE/CVF*

- International Conference on Computer Vision (ICCV)*, pages 14713–14724, 2023. 2, 3
- [25] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *International Conference on Computer Vision (ICCV)*, pages 11087–11096, 2021. 3
- [26] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4866–4875, 2020. 8
- [27] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2151–2162. 2
- [28] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning task-oriented grasping from human activity datasets. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):3352–3359, 2020. 3
- [29] Lingdong Kong, Youquan Liu, Lai Xing Ng, Benoit R. Coteau, and Wei Tsang Ooi. Openess: Event-based semantic scene understanding with open vocabularies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15686–15698, 2024. 2
- [30] Hema S. Koppula and Ashutosh Saxena. Physically grounded spatio-temporal object affordances. In *European Conference on Computer Vision (ECCV)*, pages 831–847. 3
- [31] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A. Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3
- [32] Jogendra Nath Kundu, Maharshi Gor, and R. Venkatesh Babu. BiHMP-GAN: Bidirectional 3D Human Motion Prediction GAN. In *AAAI*, pages 8553–8560, 2019. 1
- [33] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional Sequence to Sequence Model for Human Dynamics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5226–5234, 2018. 2
- [34] Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3086–3096, 2024. 3
- [35] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 214–223, 2020. 1
- [36] Maosen Li, Siheng Chen, Zijong 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)*, pages 18–36. Springer, 2022. 1, 2, 5
- [37] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. *arXiv preprint*, 2023. 2
- [38] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3
- [39] Yuyang Li, Bo Liu, Puhao Li, Yaodong Yang, Yixin Zhu, Tengyu Liu, and Siyuan Huang. Grasp multiple objects with one hand. *IEEE Robotics and Automation Letters (RA-L)*, 9(5):4027–4034, 2024. 3
- [40] Yuanzhi Liang, Xiaohan Wang, Linchao Zhu, and Yi Yang. Maal: Multimodality-aware autoencoder-based affordance learning for 3d articulated objects. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 217–227, 2023. 2, 3
- [41] Zhenyu Lou, Qiongjie Cui, Haofan Wang, Xu Tang, and Hong Zhou. Multimodal sense-informed prediction of 3d human motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 4, 5, 6, 7
- [42] 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)*, pages 9489–9497, 2019. 1, 2, 5
- [43] Wei Mao, Richard I Hartley, Mathieu Salzmann, et al. Contact-aware human motion forecasting. *Advances in Neural Information Processing Systems*, 35:7356–7367, 2022. 2
- [44] Julieta Martinez, Michael J Black, and Javier Romero. On Human Motion Prediction using Recurrent Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2891–2900, 2017. 1, 2
- [45] Angel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2276–2284, 2021. 1, 2
- [46] Jieru Mei, Liang Chieh Chen, Alan Yuille, and Cihang Xie. Spformer: Enhancing vision transformer with superpixel representation. *arXiv preprint arXiv:2401.02931*, 2024. 8
- [47] Lorenzo Mur-Labadia, Jose J. Guerrero, and Ruben Martinez-Cantin. Multi-label affordance mapping from ego-centric vision. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5238–5249, 2023. 3
- [48] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8687–8696. 3
- [49] Trong-Thuan Nguyen, Pha Nguyen, and Khoa Luu. Hig: Hierarchical interlacement graph approach to scene graph generation in video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18384–18394, 2024. 2
- [50] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. 5
- [51] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and

- Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 3, 4
- [52] Dario Pavullo, Christoph Feichtenhofer, Michael Auli, and David Grangier. Quaternet: A Quaternion-based Recurrent Model for Human Motion. *International Journal of Computer Vision*, (128):855–872, 2020. 1
- [53] Sangyun Shin, Kaichen Zhou, Madhu Vankadari, Andrew Markham, and Niki Trigoni. Spherical mask: Coarse-to-fine 3d point cloud instance segmentation with spherical representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4060–4069, 2024. 8
- [54] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13253–13263, 2022. 2
- [55] Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibe Yang. Cotdet: Affordance knowledge prompting for task driven object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3068–3078, 2023. 3
- [56] G Tevet, S Raab, B Gordon, Y Shafir, D Cohen-Or, and AH Bermano. Human motion diffusion model. arxiv 2022. *arXiv preprint arXiv:2209.14916*. 4
- [57] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 448–458, 2023. 4
- [58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 2
- [59] Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2698–2707, 2022. 4, 8
- [60] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19757–19767, 2024. 2
- [61] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2596–2605, 2017. 3
- [62] 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)*, pages 14959–14971, 2022. 2
- [63] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 433–444, 2024. 2, 3
- [64] Ruihai Wu, Chuanruo Ning, and Hao Dong. Learning foresightful dense visual affordance for deformable object manipulation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10947–10956, 2023. 3
- [65] Chaoyue Xing, Wei Mao, and Miaomiao Liu. Scene-aware human motion forecasting via mutual distance prediction. In *European Conference on Computer Vision (ECCV)*, pages 128–144. Springer, 2025. 1, 2, 5, 6, 7
- [66] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Xinchao Wang, and Yanfeng Wang. Auxiliary tasks benefit 3d skeleton-based human motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9509–9520, 2023. 1, 5, 6
- [67] Haitao Yan, Qiongjie Cui, Jiexin Xie, and Shijie Guo. Forecasting of 3d whole-body human poses with grasping objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1726–1736, 2024. 1
- [68] Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19823–19832, 2024. 2
- [69] Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. Physcene: Physically interactable 3d scene synthesis for embodied ai. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16262–16272, 2024. 2
- [70] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language guided generation of 3d embodied ai environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16227–16237, 2024. 2
- [71] Ting Yu, Kunhao Fu, Jian Zhang, Qingming Huang, and Jun Yu. Multi-granularity contrastive cross-modal collaborative generation for end-to-end long-term video question answering. *IEEE Transactions on Image Processing*, 33:3115–3129, 2024. 1
- [72] Ting Yu, Xiaojun Lin, Shuhui Wang, Weiguo Sheng, Qingming Huang, and Jun Yu. A comprehensive survey of 3d dense captioning: Localizing and describing objects in 3d scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3):1322–1338, 2024. 2
- [73] Ting Yu, Kunhao Fu, Shuhui Wang, Qingming Huang, and Jun Yu. Prompting video-language foundation models with domain-specific fine-grained heuristics for video question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1615–1630, 2025. 1
- [74] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16010–16021, 2023. 4

- [75] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. In *International Conference on 3D Vision (3DV)*, pages 642–651, 2020. [3](#)
- [76] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3d people in scenes without people. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6194–6204, 2020. [2](#), [3](#)
- [77] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14738–14749, 2023. [2](#)
- [78] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision (ECCV)*, pages 676–694, 2022. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [79] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European Conference on Computer Vision (ECCV)*, pages 408–424, 2014. [3](#)
- [80] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. Inferring forces and learning human utilities from videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3823–3833, 2016. [3](#)