Multimodal Sense-Informed Forecasting of 3D Human Motions

Zhenyu Lou[1]   Qiongjie Cui[2*]

Haofan Wang[3]   Xu Tang[3]   Hong Zhou[1]

[1]Zhejiang University, [2]Nanjing University of Science and Technology, [3]Xiaohongshu Inc

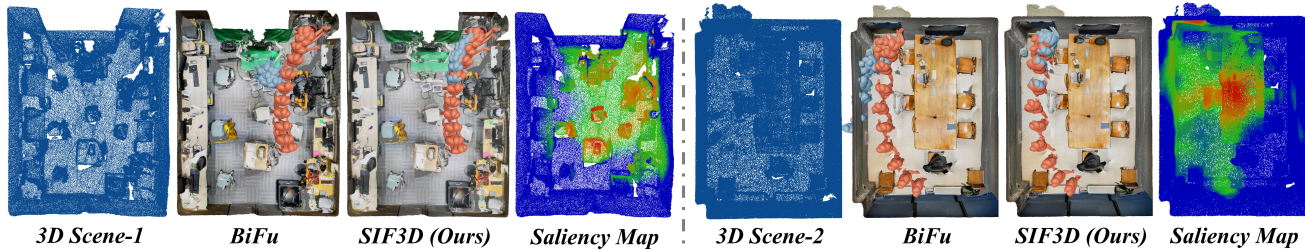11915044@zju.edu.cn   cuiqiongjie@126.com

Figure 1. ***The proposed SIF3D:** multimodal Sense-**I**nformed **F**orecasting of **3D** human motions. Our SIF3D takes the observed motion sequence, as well as the 3D scene point cloud as input modalities, and is able to identify salient points (redder) and underlying ones (bluer), to generate the accurate trajectory and high-fidelity future poses within given 3D scenarios. In contrast, the state-of-the-art baseline of BiFu [73] equally considers the global scene embedding, and thus cannot distinguish the saliency of the 3D scene, leading to the physically implausible motions, e.g., human mesh intersecting or distorting with the 3D environment, violating any physical constraints.*

Abstract

Predicting future human pose is a fundamental application for machine intelligence, which drives robots to plan their behavior and paths ahead of time to seamlessly accomplish human-robot collaboration in real-world 3D scenarios. Despite encouraging results, existing approaches rarely consider the effects of the external scene on the motion sequence, leading to pronounced artifacts and physical implausibilities in the predictions. To address this limitation, this work introduces a novel multi-modal sense-informed motion prediction approach, which conditions high-fidelity generation on two modal information: external 3D scene, and internal human gaze, and is able to recognize their salience for future human activity. Furthermore, the gaze information is regarded as the human intention, and combined with both motion and scene features, we construct a ternary intention-aware attention to supervise the generation to match where the human wants to reach. Meanwhile, we introduce semantic coherence-aware attention to explicitly distinguish the salient point clouds and the underlying ones, to ensure a reasonable interaction of the generated sequence with the 3D scene. On two real-world benchmarks, the proposed method achieves state-of-the-art performance both in 3D human pose and trajectory prediction. More detailed results are available on the page: https://sites.google.com/view/cvpr2024sif3d.

*Corresponding author

1. Introduction

Forecasting future human poses, from the observed ones, stands as a fundamental application in the domain of machine intelligence, autonomous vehicles, and human-robot collaboration [11, 12, 18, 33, 39, 40, 42, 48, 65, 70, 73].

Human motion is fundamentally intertwined with and constrained by its environment [36, 55, 59], a reality often overlooked in previous motion prediction methods [2, 33, 41, 42]. This oversight could lead to noticeable anomalies and unrealistic outcomes in the predictions. We notice this issue and aim to solve it.

In robotics, scene information is commonly represented as a point cloud [6, 35, 57, 71], where a large number of points are constructed as a collection to describe a 3D environment. Existing methods perform motion prediction/generation by encoding the entire scene information into a unified global embedding [5, 22, 73]. Despite their good performance, we note that not all information within the point cloud is equally relevant to the motion prediction task; instead, only a small subset is salient. Moreover, the global embedding fails to capture the intricate local details of the scene; therefore, the generation may deviate from the intended scene semantics. In addition, the human gaze is a valuable manifestation, providing an intention that the human is likely to reach or a direction that the human is likely to move toward [16, 63, 72]. By jointly considering the 3D scene and human gaze, it is possible to capture the human's behavior of wanting to move toward a specific location, and

thus generate more reasonable motion sequences.

Towards this end, we propose a novel multimodal sense-informed predictor for real-world 3D scenarios, named SIF3D. It integrates past motion sequences with critical data from external 3D scenes and the observer's internal gaze to enhance scene-aware human activity representations [5,8,22,73]. Following the recent progress, both motion encoder and scene encoder are introduced to extract the motion embedding and scene embedding [53], respectively. To further refine point cloud data utilization and focus on key points affecting actions, we propose two novel scene-aware cross-modal attention mechanisms: ternary intention-aware attention (TIA) integrates human gaze, motion, and scene data through a comprehensive analysis to concentrate on globally salient points; semantic coherence-aware attention (SCA) focuses on identifying the local salient scene points by analyzing their semantic connections with human poses. SIF3D adeptly identifies salient points and the underlying ones in the 3D scene (as in Figure 1), where the salient points are more likely to interact with the human, and thus the generated motion sequence is more realistic. Additionally, inspired by [67], we apply a geometry discriminator to boost the overall authenticity of the predicted motions.

Our contributions are as follows: (1) We introduce SIF3D which considers both the external scene and the internal human gaze, and is capable of generating the accurate future motion, as well as the trajectory, within given 3D scenarios. (2) Both semantic coherence-aware attention and ternary intention-aware attention are proposed, to explicitly distinguish salient point clouds to the local human pose representation, and the one to the global trajectory planning. (3) We show that, on two benchmarks GIMO and GTA-1M, the proposed SIF3D achieves state-of-the-art performance both in human pose and trajectory prediction.

2. Related Work

3D human motion prediction has witnessed substantial advancements, particularly with the rise of deep end-to-end approaches [4,20,27,30,56]. Prior works typically leverage RNNs [15,18,19,43,44,61] to capture the motion temporal dependence, facing the limitations of discontinuity and error accumulation. To address it, recent alternatives, CNN-based [32] and transformer-based [1] methods have been proposed. To effectively exploit both spatial and temporal dependencies, [1] designs a spatial-temporal attention schema, while [70] proposes a dependencies modeling Aux-Former. Notably, GCNs-based predictors are the dominant techniques and still evolve nowadays [9–11,31,33,34,42].

We also notice that current methods often fail to explicitly consider real 3D scenes, leading to physically implausible and artifactual predicted trajectories and poses. To solve these limitations, our SIF3D is proposed.

Scene-aware motion generation. Considering the connection between human activities and scene context, scene-aware motion generation has become a focal point of research [5, 8, 22, 73]. Early attempts [5, 8, 67] incorporate object bounding boxes, 2D scene images, or depth maps to contextualize scenes. [22] further introduces the object's geometry to estimate the interaction and generate interactive motions, while adaptations of the A∗ algorithm have been proposed for collision-averse trajectory planning [22,66].

The adoption of directly accessible 3D point clouds for scene representation marks a new trend in scene-aware motion generation [25, 68, 73, 74]. [68] proposes to condition the motion synthesis through text descriptions, while [25] contributes a conditional generative diffusion model for scene understanding, and [74] advances a structure-aware motion framework for active object tracking. Notably, BiFu [73] introduces a bidirectional fusion strategy for motion prediction. However, its neglect of local scene intricacies limits salient point detection, leading to ambiguity in cross-modal motion encoding across the scene. In contrast, our SIF3D adeptly isolates critical scene points, enhancing the integration of pivotal scene data into motion forecasting.

3D scene understanding. 3D Point cloud, a vital geometric data structure, has been traditionally transformed into regular 3D voxel grids or collections of images, because of its irregular representation [45,52,52,58,69]. However, these methods are either computationally expensive or task-specific, making them hard to transfer into downstream tasks. To address it, PointNet and PointNet++ [51, 53] introduces a unified architecture directly consuming point clouds, proving highly efficient and effective. Since then, various methods have been proposed for different tasks, *e.g*., 3D object detection [21, 49, 50], semantic segmentation [7, 17, 62], and instance segmentation [14, 23, 24, 28].

Although 3D scenes are crucial for information, effectively utilizing them in human motion prediction presents a challenge. Previous works [51, 53] have shown that many point cloud points are often superfluous or irrelevant. SIF3D advances this understanding by selectively focusing on salient points within the 3D scene, leading to enhanced prediction quality for human poses and global trajectories.

3. Proposed Method

Figure 2 illustrates the overall architecture of our SIF3D. It begins with the proposed MotionEncoder and the vanilla PointNet++ [53], to encode the observed sequence, and 3D point clouds. Then, ternary intention-aware attention (TIA) and semantic coherence-aware attention (SCA) are used to extract crossmodal features from salient scene points, followed by a TrajectoryPlanner and a PosePredictor to plan future trajectories and predict poses, respectively. Ultimately, the predicted motion sequence is generated by a MotionDecoder, supervised by the geometric discriminator.
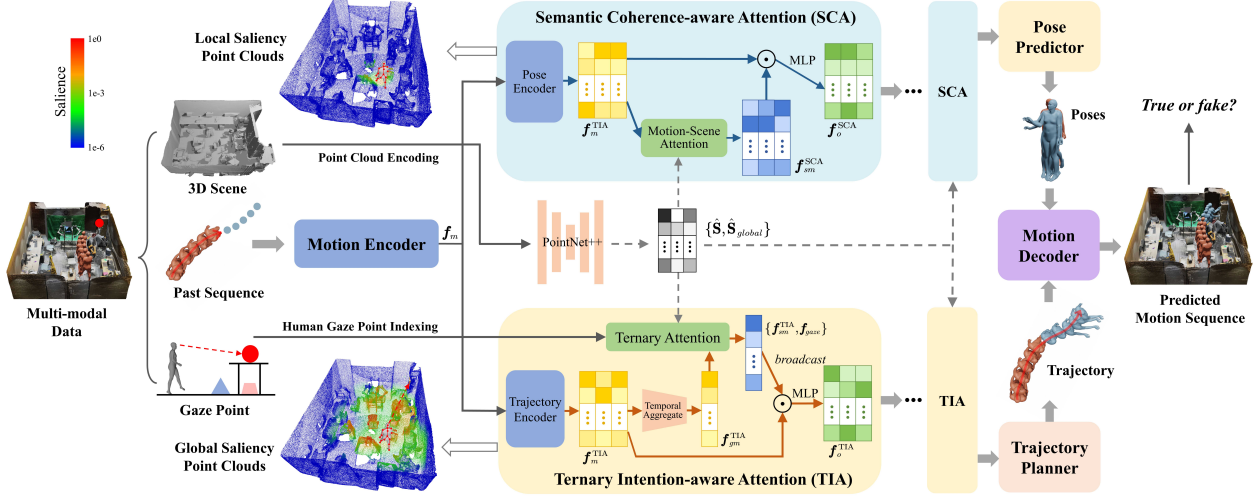
Figure 2. **The architecture of SIF3D.** SIF3D incorporates three modalities of input, the past motion sequence, the 3D scene point cloud, and the human gaze. First, MotionEncoder encodes past motion sequence into a motion embedding $\boldsymbol{f}_m$, and the 3D scene $\boldsymbol{S}$ is encoded into $\{\hat{\boldsymbol{S}}, \hat{\boldsymbol{S}}_{global}\}$ through PointNet++ [53]. Then, our TIA mechanism compresses motion embedding in the temporal dimension and searches for global salient points in the scene for trajectory planning. In addition, human gaze point $\boldsymbol{G}$ is introduced to index the scene point cloud for gaze point scene feature extraction. The SCA mechanism, on the other hand, is designed to capture local salient points in the scene for each independent pose. A TrajectoryPlanner and a PosePredictor are applied to predict trajectory and poses, respectively. And finally, the predicted motion sequence is generated through a MotionDecoder, which is supervised by the geometric discriminator.

3.1. Problem Setup

Given T historical motion $\boldsymbol{X}_{1:T} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T\}$ in the scene cloud $\boldsymbol{S} \in \mathbb{R}^{n \times 3}$, corresponding with the human gaze point sequence $\boldsymbol{G}_{1:T} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_T\}$, we aim to predict the future motion $\boldsymbol{Y}_{1:\Delta T}$. Here, n denotes the size of the scene point clouds, $\boldsymbol{g}_k \in \mathbb{R}^3$ is the position of human gaze (denoted as the intersection of the human eye ray and 3D scene). Each pose is described as $\boldsymbol{x}_k = (\boldsymbol{t}_k, \boldsymbol{o}_k, \boldsymbol{p}_k)$ [73], where $\boldsymbol{t}_k \in \mathbb{R}^3$ denotes the global translation, $\boldsymbol{o}_k \in SO(3)$ denotes the orientation, and $\boldsymbol{p}_k \in \mathbb{R}^{32}$ refers to the body pose embedding. The SMPL-X body mesh and the joints position $\boldsymbol{j}_k \in \mathbb{R}^{23 \times 3}$ can be then obtained via VPoser [47]. Following [73], we set $T = 3\text{-}sec$ and $\Delta T = 5\text{-}sec$ to achieve a more challenging long-term prediction compared with the previous tasks ($\Delta T = 1\text{-}sec$) [33, 42].

In summary, with the model parameter θ, our goal is:

$$\arg\max P(\boldsymbol{Y}_{1:\Delta T} | \boldsymbol{X}_{1:T}, \boldsymbol{S}, \boldsymbol{G}_{1:T}; \theta). \qquad (1)$$

Note that the motion trajectory involves both global translation $\boldsymbol{T}_{1:T+\Delta T}$ and orientation $\boldsymbol{O}_{1:T+\Delta T}$. Our primary focus in trajectory evaluation centers on $\boldsymbol{T}_{1:T+\Delta T}$ due to its significant impact on the overall motion. The orientation is implicitly reflected in the joints positions $\boldsymbol{J}_{1:T+\Delta T}$.

In contrast to numerous prior literature [26, 60], our task incorporates 3D scene and human gaze information, with the objective of producing semantically reasonable, and physically stable motion predictions within a given 3D scene. Our effort aims to bridge the divide between motion prediction and real-world scenarios, fostering diverse applications grounded in realistic 3D scenes.

3.2. Multimodal Encoding

In addition to the observed motion sequence $\boldsymbol{X}_{1:T}$, SIF3D integrates two additional input modalities: the external 3D scene $\boldsymbol{S}$, and the internal human's gaze $\boldsymbol{G}_{1:T}$. The motion and the scene are first encoded into a unified space, while the gaze point is used for scene point indexing to obtain the gaze point feature (see Sec-3.3).

A transformer-based MotionEncoder [64] is proposed to transfer the passing motion sequence $\boldsymbol{X}_{1:T}$ into a motion embedding $\boldsymbol{f}_m$. Note that, instead of using the original observation $\boldsymbol{X}_{1:T}$ with T frames, we pad it with the last observed pose $\boldsymbol{x}_T$ repeated ΔT times to construct the virtual sequence $\{\boldsymbol{X}, \boldsymbol{x}_T, ..., \boldsymbol{x}_T\} \in \mathbb{R}^{(T+\Delta T) \times c_m}$, which is taken as the input of the MotionEncoder:

$$\boldsymbol{f}_m = \text{MotionEncoder}(\{\boldsymbol{X}, \boldsymbol{x}_T, ..., \boldsymbol{x}_T\}), \qquad (2)$$

where the padded frames need to be predicted, and will be replaced after the inference to achieve the final prediction. Here, $\boldsymbol{f}_m \in \mathbb{R}^{(T+\Delta T) \times c_m}$, and c_m is the dimension.

Meanwhile, 3D point cloud $\boldsymbol{S}$ is encoded using Point-Net++ [53] to obtain per-point scene features $\hat{\boldsymbol{S}}$ and the global scene embedding $\hat{\boldsymbol{S}}_{global}$:

$$\hat{\boldsymbol{S}}, \hat{\boldsymbol{S}}_{global} = \text{PointNet++}(\boldsymbol{S}), \qquad (3)$$

where $\hat{\boldsymbol{S}} \in \mathbb{R}^{n \times c_s}$, $\hat{\boldsymbol{S}}_{global} \in \mathbb{R}^{c_s}$, and c_s represents the dimension of the scene embedding.

3.3. Ternary Intention-Aware Attention

Scene information serves as a crucial constrain in predicting the long-term trajectory [5, 8, 67]. Moreover, as

people often direct their gaze toward the objectives of their movements, human gaze becomes an essential cue, reflecting the potential destination of the motion trajectory [73].

To enhance trajectory planning by effectively utilizing scene and gaze information, we propose ternary intention-aware attention (TIA). TIA is designed to capture global salient scene features and analyze human intention through 3 distinct modalities: motion embedding $\boldsymbol{f}_m$, 3D scene $\{\hat{\boldsymbol{S}}, \hat{\boldsymbol{S}}_{global}\}$, and human gaze $\boldsymbol{G}_{1:T}$.

Since long-term trajectory planning hinges on the semantics of the entire motion sequence, TIA first extracts a comprehensive global motion representation. It encodes the input motion embedding to $\boldsymbol{f}_m^{\text{TIA}}$, and then condenses it into a global motion representation $\boldsymbol{f}_{gm}^{\text{TIA}}$ through a temporal aggregator. Various methods can be employed to design this aggregator, while we simply adopt the last motion embedding as the global representation in our case, aligning with a widely-used technique in NLP [54]:

$$\boldsymbol{f}_m^{\text{TIA}} = \text{TrajectoryEncoder}(\boldsymbol{f}_m), \qquad (4)$$

$$\boldsymbol{f}_{gm}^{\text{TIA}} = \text{Aggregate}(\boldsymbol{f}_m^{\text{TIA}}). \qquad (5)$$

The global salient scene features are grounded in the global scene embedding $\hat{\boldsymbol{S}}_{global}$ as a foundation. To achieve the extraction of effective global salient scene features, TIA assesses the global salience $s_g \in \mathbb{R}^n$ for each scene point to $\boldsymbol{f}_{gm}^{\text{TIA}}$, focusing only on the salient points. To exclude underlying points, a simple method arises by multiplying the salience with the corresponding scene point features. This process yields the global salience $s_g \in \mathbb{R}^n$ and the cross-modal global scene-motion features $\boldsymbol{f}_{sm}^{\text{TIA}}$ through a cross-modal attention mechanism as follows:

$$\boldsymbol{Q}^{\text{TIA}}; (\boldsymbol{K}^{\text{TIA}}, \boldsymbol{V}^{\text{TIA}}) = \text{Linear}(\boldsymbol{f}_{gm}^{\text{TIA}}); \text{Linear}(\hat{\boldsymbol{S}}), \quad (6)$$

$$s_g = \text{softmax}\left(\frac{\boldsymbol{Q}^{\text{TIA}} * (\boldsymbol{K}^{\text{TIA}})^T}{\sqrt{c}}\right), \quad (7)$$

$$\boldsymbol{f}_{sm}^{\text{TIA}} = \text{broadcast}(\hat{\boldsymbol{S}}_{global} + s_g \cdot \boldsymbol{V}^{\text{TIA}}), \quad (8)$$

where c is the attention dimension. We broadcast the global salient scene features to the same shape as $\boldsymbol{f}_m$ to disperse it across the entire motion sequence.

Moreover, recognizing the informative nature of the gaze regarding the subject's intent, we also introduce the human gaze into cross-modal attention. The gaze point $\boldsymbol{G}_{1:T}$, defined as the intersection between the human eye's line of sight and the 3D scene, correlates with a specific scene point. By indexing the scene point cloud features $\hat{\boldsymbol{S}}$ with the gaze point, we derive the gaze-related scene feature, representing the potential human intention. This T-frame gaze feature (padded with last frame for ΔT times to fit the length) is then encoded to align with the length of motion sequence using a GazeEncoder:

$$\boldsymbol{f}_{gaze} = \text{GazeEncoder}(\{\hat{\boldsymbol{S}}[\boldsymbol{G}], \hat{\boldsymbol{S}}[\boldsymbol{g}_T], ..., \hat{\boldsymbol{S}}[\boldsymbol{g}_T]\}), \quad (9)$$

where $[\cdot]$ represents the indexing operation, $\boldsymbol{f}_{gaze}$ represents the encoded gaze features. Finally, TIA integrates the features from the three modalities: $\boldsymbol{f}_m$ for motion, $\boldsymbol{f}_{sm}^{\text{TIA}}$ for scene, and $\boldsymbol{f}_{gaze}$ for gaze. The output of TIA is then computed through a standard two-layer MLP [64]:

$$\boldsymbol{f}_o^{\text{TIA}} = \text{MLP}(\text{concat}(\boldsymbol{f}_m, \boldsymbol{f}_{sm}^{\text{TIA}}, \boldsymbol{f}_{gaze})). \qquad (10)$$

3.4. Semantic Coherence-Aware Attention

We introduce semantic coherence-aware attention (SCA) to integrate local salient scene details into per-frame pose prediction. In contrast to TIA, SCA computes scene point salience for each frame independently. It initiates with a PoseEncoder to encode the input motion embedding $\boldsymbol{f}_m$ into pose embedding $\boldsymbol{f}_m^{\text{SCA}}$. Then, the local scene point salience $s_l \in \mathbb{R}^{(T+\Delta T) \times n}$ is evaluated between the scene features $\hat{\boldsymbol{S}}$ and each of the $T + \Delta T$ pose embedding:

$$\boldsymbol{f}_m^{\text{SCA}} = \text{PoseEncoder}(\boldsymbol{f}_m), \qquad (11)$$

$$\boldsymbol{Q}^{\text{SCA}}; \boldsymbol{K}^{\text{SCA}}, \boldsymbol{V}^{\text{SCA}} = \text{Linear}(\boldsymbol{f}_m^{\text{SCA}}), \text{Linear}(\hat{\boldsymbol{S}}), \quad (12)$$

$$s_l = \text{softmax}\left(\frac{\boldsymbol{Q}^{\text{SCA}} * (\boldsymbol{K}^{\text{SCA}})^T}{\sqrt{c}}\right), \qquad (13)$$

where c represents the attention dimension.

SCA places a greater emphasis on fine-grained local scene information, exhibiting heightened sensitivity to spatial details. In practice, points situated closer to the subject and aligned with the human body are empirically considered more salient. To adopt this spatial attention mechanism, we incorporate a spatial salience bias to s_l, denoted as $s_{spatial} \in \mathbb{R}^{(T+\Delta T) \times n}$, inspired by [13, 37, 38]. The absolute scene points position $\boldsymbol{S}$ is normalized to the relative positions $\boldsymbol{S}_{rel} \in \mathbb{R}^{(T+\Delta T) \times n \times 3}$ according to the predicted human translation $\hat{\boldsymbol{T}}$ and orientation $\hat{\boldsymbol{O}}$ from TrajectoryPlanner. Then, the spatial salience bias $s_{spatial}$ is determined based on $\boldsymbol{S}_{rel}$ through a standard two-layer MLP:

$$\boldsymbol{S}_{rel} = \text{Normalize}(\boldsymbol{S}; \hat{\boldsymbol{T}}, \hat{\boldsymbol{O}}), \qquad (14)$$

$$s_{spatial} = \text{MLP}(\boldsymbol{S}_{rel}). \qquad (15)$$

The local salient scene features are then extracted based on locally salient scene points, while the low-salience underlying points are ignored through salience multiplication:

$$\boldsymbol{f}_{sm}^{\text{SCA}} = (s_l + s_{spatial}) \cdot \boldsymbol{V}^{\text{SCA}}. \qquad (16)$$

Finally, SCA produces output by combining the motion features $\boldsymbol{f}_m$ with the local salient scene features $\boldsymbol{f}_{sm}^{\text{SCA}}$:

$$\boldsymbol{f}_o^{\text{SCA}} = \text{MLP}(\text{concat}(\boldsymbol{f}_m^{\text{SCA}}, \boldsymbol{f}_{sm}^{\text{SCA}})). \qquad (17)$$

3.5. Motion Sequence Generation

Predicting future motion involves considering both the trajectory and poses. In our way of multimodal feature extraction, TIA adds global scene features and human intention, while SCA brings in local scene details. Since trajectory planning looks further ahead, and human poses are

more related to the local scene semantics, it makes sense in practice that TIA could enhance trajectory prediction, while SCA helps predict poses. Thus we use the features from TIA to predict the global translation $\hat{T}$ and orientation $\hat{O}$, while relying on SCA for prediction poses $\hat{P}$:

$$\hat{T}, \hat{O} = \text{TrajectoryPlanner}(\boldsymbol{f}_o^{\text{TIA}}), \quad (18)$$

$$\hat{P} = \text{PosePredictor}(\boldsymbol{f}_o^{\text{SCA}}), \quad (19)$$

where TrajectoryPlanner/PosePredictor are linear layers.

The predicted trajectory and pose are then combined to generate the final motion sequence $\hat{Y}$ through a motion decoder. To achieve precise adjustment of joint positions, the motion decoder applies a GCN architecture [29], and we calculate the SMPL-X human skeleton based on the predicted trajectory and pose embedding through the frozen pre-trained SMPL-X model and VPoser:

$$\hat{J} = \text{SMPL-X}(\hat{T}, \hat{O}, \hat{P}, \text{VPoser}(\hat{P})), \quad (20)$$

$$\hat{Y} = \text{MotionDecoder}(\hat{J}), \quad (21)$$

where $\hat{J}$ represents the reconstructed joints positions, and $\hat{Y}$ denotes the final predicted motion sequence.

To further improve the physical stability of the predicted motion sequence, a geometric discriminator is introduced following [67] to classify the fake and true motion sequences to the ground truth within the 3D scenarios.

3.6. Implement Details

Architecture Details: In our experiments, we employ a 6-layer transformer encoder for the MotionEncoder, a single-layer self-attention mechanism for both the TrajectoryEncoder and the PoseEncoder, and a 6-layer graph convolutional network for the MotionDecoder. We configured the stack size of both TIA and SCA to be 2, indicating that SIF3D incorporates 2 TIA-blocks and 2 SCA-blocks. Furthermore, we set all embedding dimensions, including attention dimension c, motion embedding dimension c_m, and scene embedding dimension c_s, to 256. The MLPs in TIA and SCA are both 2-layer feed-forward with a hidden dimension of 1024. Finally, the geometry discriminator is designed with a 3-layer transformer decoder.

Training Setup: We train our model using an AdamW optimizer with an initial learning rate of 0.0004. An exponential learning rate scheduler is applied with a decay rate of 0.98. The models are all trained for 100 epochs with a batch size of 8 on a single NVIDIA RTX3090 GPU.

4. Experiments

4.1. Experimental Setup

Dataset-1: GIMO [73] records three modalities, including (1) full-body SMPL-X poses with $\approx 129K$ frames. (2) 3D geometry scene scanned from LiDAR sensors. (3) human gaze is recorded as a 3D coordinate. We down-sample

the motion sequences to 2fps for long-term prediction and maintain the original train-test split [73].

Dataset-2: GTA-1M [5] comprises over $1000K$ frames, which collects the human pose from real game engine rendering. The dataset has clean 3D human pose and camera annotations, and large diversity in human appearances, camera views: 10 large houses, 13 weathers, 50 human models, 22 walking styles, and various actions.

Baselines: Our SIF3D is compared with 4 recent methods, *i.e.*, LTD [42], SPGSN [33], BiFu [73] and AuxFormer [70]. BiFu [73] and AuxFormer [70] are transformer-based, while LTD [42] and SPGSN [33] are variants of GCNs. For a fair comparison, for the methods that do not inherently consider 3D scenes or human gaze, we concatenate the global scene embedding $\hat{S}_{global}$ to the motion sequence and include the gaze embedding $\boldsymbol{f}_{gaze}$. Other aspects align with their open-source codes and original settings.

Metrics: We evaluate SIF3D from 2 main aspects: trajectory and human pose. The trajectory evaluation [3,46,73] quantifies the difference between the predicted trajectory and the ground truth one. The MPJPE [2,33,42,73] evaluates the mean error in the position across all human joints. Specifically, the result is analyzed from the intermediate poses and the final destination: **(1) Traj-path**: the average trajectory deviation over all predicted poses; **(2) Traj-dest**: the trajectory deviation of the end pose; **(3) MPJPE-path**: the average MPJPE over the predicted sequence; **(4) MPJPE-dest**: the MPJPE of the destination.

4.2. Ablation of Multiple Modalities

The results in Table 1 highlight the vital role of gaze and 3D scene information in scene-conditioned motion prediction. Note that, as GTA-1M lacks human gaze information, we approximate it using the intersection point of the ray facing the human face and the 3D scene. We observe that the introduction of gaze and scene information improves the performance of all methods, including the estimated gaze information in GTA-1M, with SIF3D benefiting the most.

SIF3D excels by adeptly utilizing 3D scene information, surpassing baseline methods that show only marginal gains. This success is attributed to SIF3D's proficient integration of scene features from both global and local perspectives through TIA and SCA. On the other hand, the human gaze significantly enhances trajectory planning by providing key insights into the subject's intentions and closely aligning with the motion sequence's destination, resulting in a notable decrease in trajectory deviation.

While the introduction of both scene and gaze enhances the performance individually, their simultaneous incorporation yields optimal results across the board. It evidences the effectiveness of collaborating external scene and internal gaze features, establishing a robust foundation for improved motion prediction.

Dataset	GIMO [73]				GTA-1M [5]			
	Traj-path	Traj-dest	MPJPE-path	MPJPE-dest	Traj-path	Traj-dest	MPJPE-path	MPJPE-dest
LTD [42]	681	890	158.7	204.6	691	1051	175.0	251.8
SPGSN [33]	739	910	159.5	203.1	805	1115	167.3	241.4
AuxFormer [70]	688	893	173.5	205.4	688	1078	187.8	262.0
SIF3D	737	888	169.6	205.6	772	1072	182.0	250.2
LTD [42]+Scene	673	890	156.4	203.0	690	1029	175.8	245.6
SPGSN [33]+Scene	700	893	157.0	202.2	767	1054	167.8	236.9
AuxFormer [70]+Scene	732	902	165.1	202.4	745	1028	167.0	246.0
SIF3D w/ Scene	700	823	164.2	199.6	746	967	175.2	231.4
LTD [42]+Gaze	655	807	157.6	203.5	687	996	169.1	246.0
SPGSN [33]+Gaze	681	857	158.2	201.7	748	1031	164.4	234.8
AuxFormer [70]+Gaze	627	755	165.8	202.8	672	932	176.0	256.6
SIF3D w/ Gaze	603	714	163.1	199.1	680	917	171.5	233.0
LTD [42]+Scene+Gaze	655	801	**155.6**	202.0	702	982	167.8	241.9
SPGSN [33]+Scene+Gaze	702	982	156.9	202.0	737	1018	**162.7**	238.1
AuxFormer [70]+Scene+Gaze	622	746	161.5	201.1	672	916	165.3	236.8
BiFu [73] w/ Scene+Gaze	661	727	167.8	205.0	683	903	164.6	234.2
SIF3D w/ Scene+Gaze	**590**	**666**	156.6	**195.7**	**626**	**836**	164.9	**227.7**

Table 1. **Comparison of trajectory deviation and MPJPE** (in mm) over the sequences of the GIMO [73] and GTA-1M [5] datasets. The best result is highlighted in bold. From the results, we observe that the 3D scene and gaze information can boost all the methods, and the proposed SIF3D obtains both lower trajectory deviation and smaller MPJPE in almost all scenarios compared to the previous methods.

	Method	Gaze&Scene	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
Traj	LTD [42]	×	109	459	1147	195	415	613	149	578	1076	100	457	481	184	644	1165	127	624	1212	190	529	851
		✓	**108**	**411**	966	192	422	524	147	496	560	**88**	528	589	149	516	878	133	**551**	1297	**170**	527	750
	SPGSN [33]	×	144	492	922	189	**396**	564	**129**	556	813	117	639	660	181	640	1014	153	703	1454	182	589	866
		✓	112	475	965	180	407	497	172	540	661	156	640	677	**120**	515	919	135	649	1258	174	514	753
	AuxFormer [70]	×	185	541	1162	319	442	529	232	630	972	172	509	625	248	745	1113	151	621	1030	273	606	881
		✓	212	472	813	228	427	532	231	536	577	225	544	555	246	**511**	**621**	273	647	1082	222	546	735
	BiFu [73]	✓	232	540	838	332	442	689	218	532	659	275	467	582	273	551	681	340	607	936	417	659	629
	SIF3D	✓	183	441	554	155	424	431	200	**472**	533	176	471	444	245	540	730	257	602	**794**	219	**440**	**611**
MPJPE	LTD [42]	×	92.9	142.8	263.5	98.0	160.4	169.7	73.6	231.1	209.1	86.6	102.3	141.4	70.3	99.1	222.6	97.6	128.6	252.5	109.5	150.8	204.2
		✓	85.5	134.9	236.6	110.1	192.8	159.8	67.8	227.0	204.6	91.0	97.7	128.3	74.3	104.1	237.1	111.3	140.7	271.7	117.2	156.0	199.9
	SPGSN [33]	×	89.0	131.4	232.8	102.9	170.5	160.5	**62.3**	218.8	205.5	87.5	97.1	131.6	66.8	98.5	240.1	103.6	124.1	285.2	110.5	149.2	191.5
		✓	89.4	137.6	249.5	101.7	176.1	147.8	67.1	233.2	202.6	**85.5**	**92.1**	120.9	71.8	98.7	238.2	109.1	124.9	281.3	112.8	145.4	193.1
	AuxFormer [70]	×	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	259.5	130.9	160.7	204.9
		✓	90.5	142.3	240.7	110.2	183.8	170.5	91.5	232.7	203.9	103.8	106.4	115.5	76.3	112.9	218.2	113.5	127.1	271.5	105.7	146.0	191.6
	BiFu [73]	✓	**67.0**	115.0	258.0	**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	**128.7**	209.9
	SIF3D	✓	82.0	**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**	218.1	108.9	**121.1**	**188.4**	**101.8**	137.3	**179.4**

Table 2. **Performance details on the GIMO test set** [73] using trajectory deviation and MPJPE (in mm). Our SIF3D demonstrates strong performance across all scenarios, particularly excelling in long-term predictions, and achieves the best result on average.

4.3. Detailed Results

Next, we analyze the performance of each method under different timestamps, as illustrated in Table 2 and Table 3.

From the result, we observe that AuxFormer [70] demonstrates substantial improvement, especially in long-term trajectory predictions. BiFu [73] outperforms baseline methods lacking scene and gaze information and remains competitive with methods that include such information. We attribute this success to the better ability of cross-modal feature extraction of transformers. Similarly, our SIF3D also applies an attention mechanism (TIA and SCA), ensuring the effective extraction of multi-modal features.

GCN-based methods (LTD [42], SPGSN [33]) show proficiency in short-term predictions. However, challenges arise in longer-term scenarios, particularly in 5-second predictions, where Traj-dest averages over 400mm for both LTD and SPGSN. This difficulty arises due to the challenges GCN-based methods encounter in obtaining comprehensive whole-person pose features, making it challenging

	Method	0.5s	1.0s	2.0s	3.0s	5.0s
Traj	AuxFormer [70]	212	322	500	771	1078
	AuxFormer [70]+Scene+Gaze	198	274	411	747	916
	BiFu [73]	192	**251**	417	647	903
	SIF3D	**179**	265	**392**	**579**	**836**
MPJPE	AuxFormer [70]	105.3	135.7	163.7	211.5	262.0
	AuxFormer [70]+Scene+Gaze	**95.5**	127.3	160.7	203.4	236.8
	BiFu [73]	100.3	131.0	158.8	196.6	234.2
	SIF3D	97.9	**125.7**	**155.9**	**192.4**	**227.7**

Table 3. **Performance details on the GTA-1M test set** [5] using trajectory deviation and MPJPE (in mm). We observe that our SIF3D consistently outperforms the baseline methods in motion prediction within game engine-rendered 3D scenarios.

to estimate global human intention and establish effective cross-modal interaction with scene features.

In contrast, our SIF3D achieves superior performance across all scenarios, particularly achieving the best metrics in almost all long-term (5-sec) evaluations. This success is evident in both global trajectory and local pose, highlighting the effectiveness of TIA's intention understanding
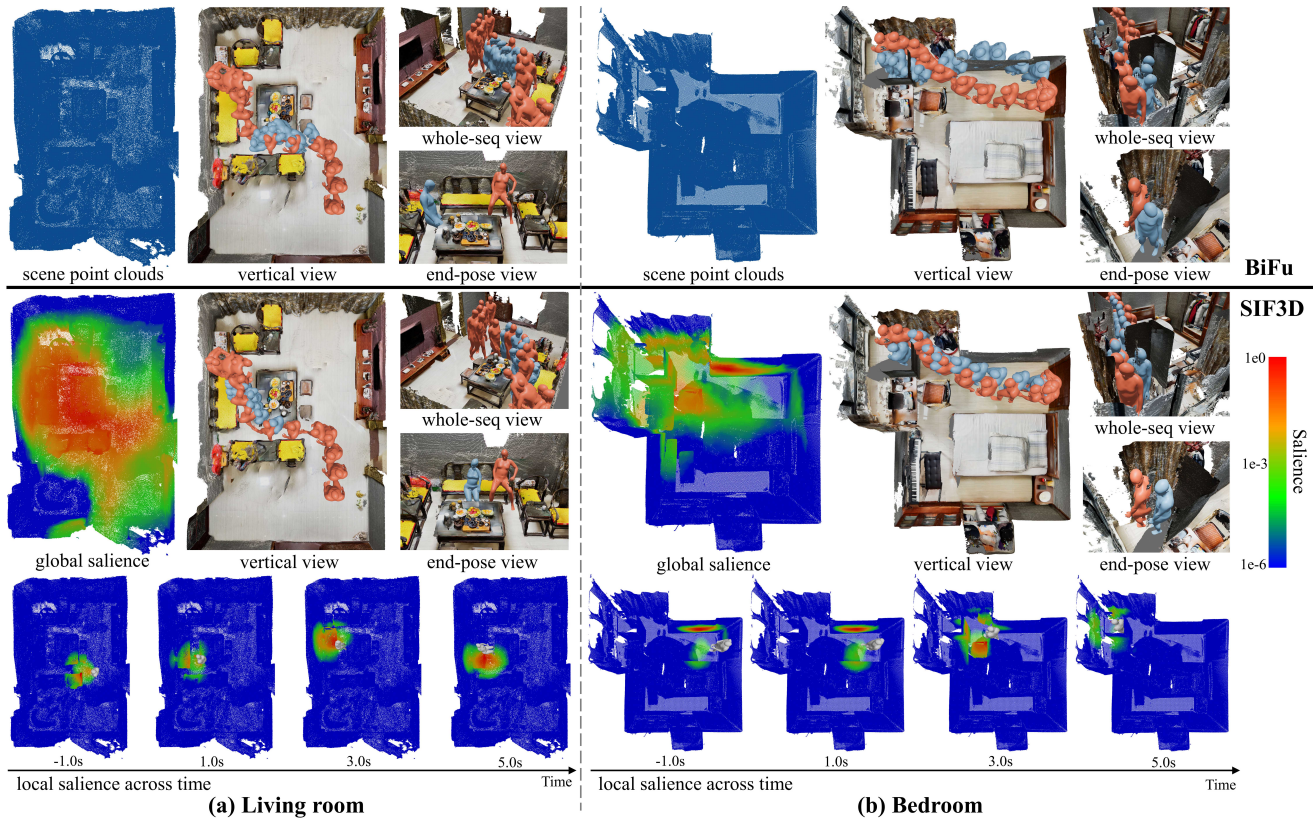
Figure 3. **Visualizations of our SIF3D compared with the SoTA BiFu,** under the scenarios of (a) living room and (b) bedroom. The top is the results of BiFu [73], which equally treats all scene points; in contrast, the middle row is our SIF3D, where the salient points are highlighted in red, and the underlying points are in blue. For the sake of clarity, the predicted sequence is presented from the vertical view, whole-seq view, and end-pose view. We note that the red human meshes are the ground truth, while the blue ones indicate the predictions. At the bottom, we present the local scene salience heatmap across time for SIF3D, with a time interval of 2 seconds.

and SCA's locally scene-aware pose prediction. It also evidences the importance of scene-awareness and line of gaze-guidance for motion prediction in 3D scenes.

4.4. Visualizations

To delve deeper into SIF3D, we provide the prediction visualizations in two different scenarios, where BiFu [73] is used as a baseline, due to its built-in integration of scene and gaze, resulting in superior numerical performance.

In the first scene (Figure 3(**a**)), the person navigates through the aisle between the chair and table, proceeding toward the sofa. While BiFu estimates the overall movement trend, the predicted motion has the person passing through the coffee table. Moreover, in the final pose, the person appears suspended before the chair, deviating from physical constraints. In contrast, SIF3D demonstrates exhibits precise predictions of the individual's motion. The global scene point salience heatmap derived highlights SIF3D's ability to comprehend the 3D scene, concentrating on key interactive objects (chairs, sofa, and coffee table) aligned with the observed motion sequence and corresponding human gaze. This focus enables SIF3D to predict a clear hu-

man intention and devise a plausible trajectory plan.

In the second scene (Figure 3(**b**)), the individual is poised to traverse the hallway leading to the main entrance, reaching the upper left corner intending to open the window. BiFu [73] predicts an accurate destination, but the motion sequence presents a significant physical inconsistency within the scene, penetrated by the wall (frames 2-5) and the screen (frames 7-8). The global salience heatmap of SIF3D reveals the understanding of the semantics of the entire 3D scene, even if the upper left corner of the room is not visible at the end of the observation, providing a better action prediction. In addition, the local salience heatmap also shows SIF3D's awareness of obstacles like the walls.

The visualizations demonstrate that SIF3D effectively predicts motions that align closely with the ground truth, accurately capturing both global trajectories and local poses.

4.5. Ablation Studies

Ablation of each component. We first investigate the impact of key components within our SIF3D, including (**1**) **TIA**, (**2**) **SCA**, (**3**) **MotionDecoder**, (**4**) **Discriminator**, and (**5**) **PointNet++**, as shown in Table 4.

	Traj-path	Traj-dest	MPJPE-path	MPJPE-dest
w/o TIA	701	850	167.9	203.8
w/o SCA	640	703	164.8	200.4
w/o MotionDecoder	669	699	159.2	198.9
w/o Discriminator	604	677	158.2	196.8
w/o PointNet++	766	918	170.4	218.5
SIF3D	**590**	**666**	**156.6**	**195.7**

Table 4. **SIF3D performance with ablations of components**, evaluated using trajectory deviation and MPJPE (in mm).

Aggregate Method	**Last**	Mean	Max	Conv	Transformer
Traj-path (mm)	**594**	595	597	602	605
MPJPE-path (mm)	**157.1**	157.7	159.2	161.3	160.8

Table 6. Performance of SIF3D with various temporal aggregators.

The removal of TIA and SCA results in a significant performance decline. Specifically, *w/o* TIA, SIF3D experiences a significant increase of 111mm in Traj-dest, underscoring a reduced capability to fully understand the 3D scene and human intentions. Note that SCA's spatial salience bias $s_{spatial}$ depends on the predicted trajectory, removing TIA also negatively impacts SCA due to the inaccurate predicted trajectory, which is evident in increases of 11.3mm and 8.1mm in MPJPE-path and MPJPE-dest. In addition, *w/o* SCA, SIF3D cannot capture local scene details, resulting in a 37mm increase in Traj-dest and a 4.7mm rise in MPJPE-dest.

TIA and SCA make independent predictions for global trajectory and local pose. Eliminating the motion decoder makes the predicted trajectory and pose identical, which introduces a 20.0mm increase in Traj-dest and a 3.2mm rise in MPJPE-dest. The discriminator also contributes to a 7.8mm decrease in Traj-dest and a 1.1mm decrease in MPJPE-dest.

Moreover, the scene encoder PointNet++, plays a crucial role as a vital link connecting our motion predictor. Notably, when PointNet++ is removed (replaced with an MLP), there is a remarkable increase in Traj-dest and MPJPE-dest. In summary, these results highlight the effectiveness of each component in our SIF3D.

Point cloud size	512	1024	2048	**4096**	8192	16384	32768
Traj-path (mm)	626	605	594	590	588	587	587
MPJPE-path (mm)	165.4	160.2	157.7	156.6	155.6	155.8	155.4
VRAM (MB)	3560	3700	4358	5246	7286	11916	20876
Speed (sample/s)	99.8	96.1	88.5	82.4	64.2	42.2	24.1

Table 5. SIF3D performance with various scene point cloud sizes.

(6) Scene point cloud size. The original size of scene point clouds from LiDAR sensors typically surpasses 500K [73]. To optimize the use of scene point clouds, we initially down-sample them. However, this approach risks omitting critical scene details. To strike a balance between computational overhead and performance, we conduct experiments by varying the scene point cloud sizes to assess SIF3D's performance. Table 5 reveals that the performance improves with the increase in scene point cloud size until the size exceeds 4096. While the performance continues to advance with the larger point cloud size, a critical issue emerges in terms of computational overhead and memory consumption (The VRAM and speed are measured on a NVIDIA RTX3090 GPU during training). Therefore, in SIF3D, we opt for a trade-off point cloud size of 4096.

(7) Motion aggregator in TIA. In TIA, we consolidate the motion sequence into a unified motion embedding using a temporal motion aggregator. To investigate the influence of various action sequence aggregators on performance, we conduct experiments involving 5 aggregation strategies: (1) Last, capturing the last motion embedding in the sequence. (2) Mean, averaging across the entire motion sequence. (3) Max, employing max-pooling throughout the entire motion sequence. (4) Conv, applying a three-layer convolutional network to downsample the sequence. (5) Transformer, introducing a single-layer transformer decoder into the aggregation process.

The results in Table 6 reveal that the performance of SIF3D is not significantly affected by the choice of motion aggregator, with the first aggregator (Last) demonstrating the best performance. However, the two weighted methods (Conv and Transformer) surprisingly yield inferior results. We attribute this counterintuitive finding to the limited size of the dataset and the insufficient supervision for the aggregated global motion embedding, making it challenging for the weighted methods to learn effectively.

5. Conclusion

This work presents a pioneering multi-modal sense-informed framework, SIF3D, designed for human motion prediction within real-world 3D scenes. By incorporating external 3D scene points and internal human gaze, SIF3D acquires the ability to perceive the scene and comprehend human intention using SCA and TIA. In evaluations conducted on two datasets of GIMO and GTA-1M, SIF3D achieves an average destination trajectory deviation of 666mm and 836mm on each dataset, and a 195.7mm and 227.7mm average destination MPJPE, establishing a new state-of-the-art performance both in global trajectory and local pose prediction. Consequently, our findings underscore the significance of scene and gaze information in scene-based motion prediction. Furthermore, we posit that delving into high-fidelity diverse human motion generation in real-world 3D scenarios holds promise as a compelling avenue for future exploration.

References

[1] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A Spatio-temporal Transformer for 3D Human Motion Prediction. In *3DV*, pages 565–574. IEEE, 2021. 2

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *3DV*, pages 565–574. IEEE, 2021. 1, 5

[3] Florent Altché and Arnaud de La Fortelle. An lstm network for highway trajectory prediction. In *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)*, pages 353–359. IEEE, 2017. 5

[4] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: Probabilistic 3D Human Motion Prediction via GAN. In *CVPR*, pages 1418–1427, 2018. 2

[5] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 387–404. Springer, 2020. 1, 2, 3, 5, 6

[6] Jingdao Chen, Zsolt Kira, and Yong K Cho. Deep learning approach to point cloud scene understanding for automated scan to 3d reconstruction. *Journal of Computing in Civil Engineering*, 33(4):04019027, 2019. 1

[7] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 2

[8] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware Human Motion Prediction. In *CVPR*, pages 6992–7001, 2020. 2, 3

[9] Qiongjie Cui and Huaijiang Sun. Towards Accurate 3D Human Motion Prediction From Incomplete Observations. In *CVPR*, pages 4801–4810, June 2021. 2

[10] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning Dynamic Relationships for 3D Human Motion Prediction. In *CVPR*, pages 6519–6527, 2020. 2

[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 *ICCV*, pages 11467–11476, 2021. 1, 2

[12] Christian Diller, Thomas Funkhouser, and Angela Dai. Forecasting characteristic 3d poses of human actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15914–15923, 2022. 1

[13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. 4

[14] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9031–9040, 2020. 2

[15] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent Network Models for Human Dynamics. In *ICCV*, pages 4346–4354, 2015. 2

[16] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, volume 42, pages 1–12. Wiley Online Library, 2023. 1

[17] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 2

[18] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José M. F. Moura. Adversarial Geometry-Aware Human Motion Prediction. In *ECCV*, pages 786–803, 2018. 1, 2

[19] Liang-Yan Gui, Kevin Zhang, Yu-Xiong Wang, Xiaodan Liang, José M. F. Moura, and Manuela Veloso. Teaching Robots to Predict Human Motion. *IROS*, pages 562–567, 2018. 2

[20] Xiao Guo and Jongmoo Choi. Human Motion Prediction via Learning Local Structure Representations and Temporal Dependencies. In *AAAI*, pages 2580–2587, 2019. 2

[21] JunYoung Gwak, Christopher Choy, and Silvio Savarese. Generative sparse detection networks for 3d single-shot object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 297–313. Springer, 2020. 2

[22] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021. 1, 2

[23] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4421–4430, 2019. 2

[24] Ji Hou, Angela Dai, and Matthias Nießner. Revealnet: Seeing behind objects in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2098–2107, 2020. 2

[25] 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*, pages 16750–16761, 2023. 2

[26] 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*, 36(7):1325–1339, 2013. 3

[27] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. In *CVPR*, pages 5308–5317, 2016. 2

[28] Haiyong Jiang, Feilong Yan, Jianfei Cai, Jianmin Zheng, and Jun Xiao. End-to-end 3d point cloud instance segmentation

without detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12796–12805, 2020. 2

[29] Thomas N Kipf and Max Welling. Semi-supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*, 2016. 5

[30] Jogendra Nath Kundu, Maharshi Gor, and R. Venkatesh Babu. BiHMP-GAN: Bidirectional 3D Human Motion Prediction GAN. In *AAAI*, volume 33, pages 8553–8560, 2019. 2

[31] Bin Li, Jian Tian, Zhongfei Zhang, Hailin Feng, and Xi Li. Multitask Non-Autoregressive Model for Human Motion Prediction. *IEEE Transactions on Image Processing*, 2020. 2

[32] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional Sequence to Sequence Model for Human Dynamics. In *CVPR*, pages 5226–5234, 2018. 2

[33] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-Parted Graph Scattering Networks for 3D Human Motion Prediction. In *ECCV*, pages 18–36. Springer, 2022. 1, 2, 3, 5, 6

[34] 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 *CVPR*, pages 214–223, 2020. 2

[35] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 529–537, 2019. 1

[36] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7577–7586, 2021. 1

[37] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 4

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[39] Zhenguang Liu, Kedi Lyu, Shuang Wu, Haipeng Chen, Yanbin Hao, and Shouling Ji. Aggregated Multi-GANs for Controlled 3D Human Motion Prediction. In *AAAI*, volume 35, pages 2225–2232, 2021. 1

[40] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6437–6446, 2022. 1

[41] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively Generating Better Initial Guesses Towards Next Stages for High-Quality Human Motion Prediction. In *CVPR*, pages 6437–6446, 2022. 1

[42] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning Trajectory Dependencies for Human Motion Prediction. In *ICCV*, pages 9489–9497, 2019. 1, 2, 3, 5, 6

[43] Julieta Martinez, Michael J Black, and Javier Romero. On Human Motion Prediction using Recurrent Neural Networks. In *CVPR*, pages 2891–2900, 2017. 2

[44] Angel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2276–2284, 2021. 2

[45] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 922–928. IEEE, 2015. 2

[46] Nishant Nikhil and Brendan Tran Morris. Convolutional neural network for trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 5

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

[48] AJ Piergiovanni, Anelia Angelova, Alexander Toshev, and Michael S Ryoo. Adversarial Generative Grammars for Human Activity Prediction. In *ECCV*, pages 507–523. Springer, 2020. 1

[49] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4404–4413, 2020. 2

[50] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2

[51] 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*, pages 652–660, 2017. 2

[52] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016. 2

[53] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2, 3

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[55] Noha Radwan, Wolfram Burgard, and Abhinav Valada. Multimodal interaction-aware motion prediction for autonomous street crossing. *The International Journal of Robotics Research*, 39(13):1567–1598, 2020. 1

[56] Alejandro Hernandez Ruiz, Juergen Gall, and Francesc Moreno-Noguer. Human Motion Prediction via Spatio-Temporal Inpainting. In *CVPR*, pages 7134–7143, 2018. 2

[57] Radu Bogdan Rusu, Nico Blodow, Zoltan Csaba Marton, and Michael Beetz. Close-range scene segmentation and reconstruction of 3d point cloud maps for mobile manipulation in domestic environments. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1–6. IEEE, 2009. 1

[58] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 2

[59] Zhaoxin Su, Gang Huang, Sanyuan Zhang, and Wei Hua. Crossmodal transformer based generative framework for pedestrian trajectory prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2337–2343. IEEE, 2022. 1

[60] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 3

[61] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-Term Human Motion Prediction by Modeling Motion Context and Enhancing Motion Dynamics. In *IJCAI*, 2018. 2

[62] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 2

[63] Susanne Trick, Dorothea Koert, Jan Peters, and Constantin A Rothkopf. Multimodal uncertainty reduction for intention recognition in human-robot interaction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7009–7016. IEEE, 2019. 1

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *NeurIPS*, pages 5998–6008, 2017. 3, 4

[65] Hongsong Wang, Jian Dong, Bin Cheng, and Jiashi Feng. PVRED: A Position-Velocity Recurrent Encoder-Decoder for Human Motion Prediction. *IEEE Transactions on Image Processing*, 30:6096–6106, 2021. 1

[66] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20460–20469, 2022. 2

[67] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12206–12215, 2021. 2, 3, 5

[68] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022. 2

[69] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 2

[70] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Xinchao Wang, and Yanfeng Wang. Auxiliary tasks benefit 3d skeleton-based human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9509–9520, 2023. 1, 2, 5, 6

[71] Bisheng Yang, Zhen Dong, Fuxun Liang, and Yuan Liu. Automatic registration of large-scale urban scene point clouds based on semantic feature points. *ISPRS Journal of Photogrammetry and Remote Sensing*, 113:43–58, 2016. 1

[72] Zhibin Yu, Sangwook Kim, Rammohan Mallipeddi, and Minho Lee. Human intention understanding based on object affordance and action classification. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2015. 1

[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*, pages 676–694. Springer, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[74] Fangwei Zhong, Xiao Bi, Yudi Zhang, Wei Zhang, and Yizhou Wang. Rspt: reconstruct surroundings and predict trajectory for generalizable active object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3705–3714, 2023. 2