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

A. Training Details

Here, we outline the training process for GAP3DS, including GazeNet, Affordance-aware Pose Generator, and Dual-Prompted Motion Decoder. All components are trained on 8 NVIDIA RTX 4090 GPUs, achieving convergence in under 30 minutes.

A.1. Training GazeNet

Since the original datasets [1, 8] lack annotations for the interaction map $\boldsymbol{M}$, we enrich them by annotating each motion sequence with interaction information. GazeNet serves as a temporal aggregator, transforming the distance map $\boldsymbol{D}$ into an interaction map $\boldsymbol{M}$ to predict future interactive objects effectively. To ensure accurate predictions, the interaction map $\boldsymbol{M}$ is annotated based on two critical factors: the contact area between humans and objects $\boldsymbol{M}_{\text{cont}}$ and the proximity of humans to scene elements $\boldsymbol{M}_{\text{prox}}$:

$$\boldsymbol{M}_{\text{cont}} = \mathcal{G}_{\mathcal{N}(0,\sigma_1)}\big(\text{Dist3D}(\boldsymbol{S}, \boldsymbol{H}_{\text{contact}})\big), \tag{1}$$

$$\boldsymbol{M}_{\text{prox}} = \frac{1}{\Delta L}\sum_{t=1}^{\Delta L}\frac{t}{(\Delta L)^2}\mathcal{G}_{\mathcal{N}(0,\sigma_2)}\big(\text{Dist2D}(\boldsymbol{S}, \boldsymbol{Y}_t)\big), \tag{2}$$

$$\boldsymbol{M} = \lambda_{\text{cont}}\boldsymbol{M}_{\text{cont}} + \lambda_{\text{prox}}\boldsymbol{M}_{\text{prox}}, \tag{3}$$

where $\mathcal{G}_{\mathcal{N}(\mu,\sigma)}$ represent Gaussian functions with mean μ and standard deviations σ, modeling the spatial distributions of contact and proximity. We set $\sigma_1 = 0.5$ and $\sigma_2 = 1.0$ for balance. The functions Dist3D and Dist2D calculate the 3D distances and X-Z 2D distance functions (Y-axis denotes height) distances between scene points $\boldsymbol{S}$ and human interaction data $\boldsymbol{H}_{\text{contact}}$. The weights $\lambda_{\text{cont}} = 1$ and $\lambda_{\text{prox}} = 2$ balance the contributions of contact and proximity factors in the final interaction map $\boldsymbol{M}$.

After annotating the interaction map $\boldsymbol{M}$, GazeNet is trained to minimize the discrepancy between the predicted interaction map and the annotated map using a Kullback-Leibler (KL) divergence loss:

$$\mathcal{L}_{\text{gaze}} = \text{KL}\big(\boldsymbol{M} \parallel \text{GazeNet}(\boldsymbol{S}, \boldsymbol{G}_{1:L})\big), \tag{4}$$

where $\boldsymbol{G}_{1:L}$ represents the sequence of gaze points over L time steps. The KL divergence ensures that the predicted interaction map aligns closely with the ground truth, enabling GazeNet to robustly capture spatial-temporal relationships between human gaze and scene elements.

A.2. Training Affordance-aware Pose Generator

The Affordance-aware Pose Generator predicts the interactive poses $\bar{\boldsymbol{P}}^0_{\Delta L-w:\Delta L}$ directly instead of predicting noise (detailed in Section 3.3 of the main paper), following [12, 13, 20]. The training process includes a basic diffusion loss to ensure accurate reconstruction of interactive poses over the prediction window $[\Delta L - w, \Delta L]$:

$$\mathcal{L}_{\text{base}} = \|\bar{\boldsymbol{P}}^0_{\Delta L-w:\Delta L} - \boldsymbol{P}_{\Delta L-w:\Delta L}\|_2^2, \tag{5}$$

where $\bar{\boldsymbol{P}}^0_{\Delta L-w:\Delta L}$ denotes the predicted interactive poses, and $\boldsymbol{P}_{\Delta L-w:\Delta L}$ represents the corresponding ground truth.

To enhance the precision of destination prediction at the final timestep ΔL, we incorporate a destination-specific loss inspired by [22]:

$$\mathcal{L}_{\text{dest}} = \|\hat{\boldsymbol{P}}^0_{\Delta L} - \boldsymbol{P}^0_{\Delta L}\|_2^2, \tag{6}$$

where $\hat{\boldsymbol{P}}^0_{\Delta L}$ and $\boldsymbol{P}^0_{\Delta L}$ denote the predicted and ground truth poses at the final timestep, respectively.

The overall training loss is defined as a combination of these two components:

$$\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{dest}}. \tag{7}$$

This comprehensive loss function not only ensures the pose's accuracy but also enhances the precision of the destination prediction.

A.3. Training Dual-Prompted Motion Decoder

The Dual-Prompted Motion Decoder is trained to reconstruct both trajectories and poses while ensuring physically realistic motion. The trajectory reconstruction loss, $\mathcal{L}_{\text{traj\_p}}$, and the pose reconstruction loss, $\mathcal{L}_{\text{pose\_p}}$, are defined using L1 loss functions, while the joint reconstruction loss,

$\mathcal{L}_{\text{joint\_p}}$, is measured as the mean per joint position error [8]:

$$\mathcal{L}_{\text{traj\_p}} = \frac{1}{\Delta L}\sum_{t=1}^{\Delta L}\left(\|\hat{\boldsymbol{U}}_t - \boldsymbol{U}_t\|_1 + \|\hat{\boldsymbol{V}}_t - \boldsymbol{V}_t\|_1\right), \quad (8)$$

$$\mathcal{L}_{\text{pose\_p}} = \frac{1}{\Delta L}\sum_{t=1}^{\Delta L}\|\hat{\boldsymbol{Q}}_t - \boldsymbol{P}_t\|_1, \quad (9)$$

$$\mathcal{L}_{\text{joint\_p}} = \frac{1}{\Delta L}\cdot\frac{1}{N_j}\sum_{t=1}^{\Delta L}\sum_{j=1}^{N_j}\|\hat{\boldsymbol{Y}}_{t,j} - \boldsymbol{Y}_{t,j}\|_2^2, \quad (10)$$

where $\hat{\boldsymbol{U}}_t$ and $\hat{\boldsymbol{V}}_t$ are the predicted global translations and orientations at timestep t, and $N_j = 23$ represents the number of skeleton joints based on the SMPL-X model [10].

To enhance physical plausibility, we incorporate geometric losses following [12]. The foot contact loss, $\mathcal{L}_{\text{foot}}$, minimizes motion inconsistency for joints in contact with the ground, mitigating the foot-sliding effect [11]:

$$\mathcal{L}_{\text{foot}} = \frac{1}{\Delta L - 1}\cdot\frac{1}{N_j}\sum_{t=1}^{\Delta L-1}\sum_{j=1}^{N_j}\left\|\left(\hat{\boldsymbol{Y}}_{t+1,j} - \hat{\boldsymbol{Y}}_{t,j}\right)\cdot f_{t,j}\right\|_2^2, \quad (11)$$

where $f_{t,j} \in \{0, 1\}$ is a binary contact mask indicating whether joint j is in contact with the ground at timestep t. The velocity loss, $\mathcal{L}_{\text{vel}}$, enforces smooth motion dynamics by aligning predicted velocities with ground truth:

$$\mathcal{L}_{\text{vel}} = \frac{1}{\Delta L - 1}\cdot\frac{1}{N_j}\sum_{t=1}^{\Delta L-1}\sum_{j=1}^{N_j}\Big\|\left(\boldsymbol{Y}_{t+1,j} - \boldsymbol{Y}_{t,j}\right) - \left(\hat{\boldsymbol{Y}}_{t+1,j} - \hat{\boldsymbol{Y}}_{t,j}\right)\Big\|_2^2. \quad (12)$$

To prevent unrealistic human-scene intersections, we incorporate a penetration loss $\mathcal{L}_{\text{pen}}$, penalizing negative distances between human and the nearest scene points [5]:

$$\mathcal{L}_{\text{pen}} = \sum_{f^p<0}(f^p)^2, \quad (13)$$

where f^p denotes the signed distance between a human vertex and the nearest scene point.

The final training objective for the motion decoder is a weighted combination of all loss terms:

$$\begin{aligned}\mathcal{L}_{\text{decoder}} = {} & \lambda_{\text{traj\_p}}\cdot\mathcal{L}_{\text{traj\_p}} + \lambda_{\text{pose\_p}}\cdot\mathcal{L}_{\text{pose\_p}} + \lambda_{\text{joint\_p}}\cdot\mathcal{L}_{\text{joint\_p}} \\ & + \beta\left(\lambda_{\text{foot}}\cdot\mathcal{L}_{\text{foot}} + \lambda_{\text{vel}}\cdot\mathcal{L}_{\text{vel}}\right) \\ & + \lambda_{\text{pen}}\cdot\mathcal{L}_{\text{pen}},\end{aligned} \quad (14)$$

where the geometric loss weight β is set to 1.0, and the remaining weights are set as follows: $\lambda_{\text{traj\_p}} = \lambda_{\text{pose\_p}} = \lambda_{\text{joint\_p}} = 1$, $\lambda_{\text{foot}} = \lambda_{\text{vel}} = 0.5$, and $\lambda_{\text{pen}} = 0.1$. This training strategy ensures that the Dual-Prompted Motion Decoder generates physically consistent and semantically accurate human motions in 3D environments.

B. Comparison with Motion Synthesis

We recognize recent advances in human motion synthesis, particularly scene-aware methods [3, 4, 6, 7, 16, 21] that integrate environmental context to generate human-object interactions or text-guided motions. While effective for motion synthesis, these methods are not directly applicable to human motion prediction, which emphasizes predicting future motions based on past observations, requiring explicit modeling of temporal dependencies and fine-grained human-object interactions.

B.1. Adapting AffordMotion for HMP

AffordMotion [16] achieves state-of-the-art performance in scene-aware motion synthesis by leveraging scene affordances as intermediate representations to bridge environmental context and motion generation. Scene affordances focus on identifying potential interaction regions, enabling contextually relevant motion synthesis.

To adapt AffordMotion for HMP tasks, we encode the entire observed motion sequence $\boldsymbol{X}_L$ using a 2-layer transformer encoder [14], replacing its original language-guidance mechanism. This modification ensures temporal information from historical frames is preserved and directly contributes to motion predictions. Additionally, instead of generating motions from scratch, the synthesis process initializes from the last observed frame, providing a realistic starting point. These adaptations enable AffordMotion to utilize its affordance representations while maintaining compatibility with the predictive requirements of HMP.

B.2. Experimental Results

Table 1 presents the comparative performance of GAP3DS and AffordMotion on the GIMO dataset. AffordMotion excels in average trajectory prediction (Traj-P) by effectively capturing broad environmental structures through scene affordances. However, it struggles with precise endpoint trajectory alignment (Traj-I), reflecting its limitations in predicting motions that require fine-grained spatial precision and object-specific interactions.

In contrast, GAP3DS leverages gaze-informed affordances to infer human intent at a more granular level, enabling superior performance in both trajectory and pose prediction. GAP3DS achieves lower errors in detailed human-object interactions (MPJPE-I) and overall pose refinement (MPJPE-P), demonstrating its ability to produce physically consistent and semantically meaningful motions. This improvement is attributed to GAP3DS's interactive pose generation and dual-prompted mechanism, which seamlessly integrate trajectory guidance and pose refinement.

B.3. Discussion

While AffordMotion demonstrates strong capabilities in modeling average trajectory trends, its reliance on scene-

Method	Traj-P ↓	Traj-I ↓	MPJPE-P ↓	MPJPE-I ↓
AffordMotion [16]	**571**	627	151.7	180.3
GAP3DS	575	**623**	**141.2**	**171.4**

Table 1. Comparison between GAP3DS and AffordMotion [16] on GIMO [22]. The best results are highlighted in bold.

level affordances limits its ability to model object-specific interactions and refine poses. GAP3DS addresses these gaps by incorporating gaze-informed affordances and adaptive pose generation, resulting in accurate, contextually relevant, and physically plausible motion predictions. These results underscore the importance of integrating object-level affordances and dual-prompted mechanisms for advancing HMP tasks.

C. Experiments on GTA-IM

We evaluate GAP3DS against four state-of-the-art baselines: MDP [17], AuxFormer [18], GIMO [22], and SIF3D [8] on the synthetic dataset GTA-IM [1], focusing on both trajectory and pose prediction. Since GTA-IM does not include explicit gaze annotations, we approximate gaze points by calculating the intersection of rays originating from the human face with the 3D scene, following the methodology outlined in [8].

The experimental results in Table 2 underscore GAP3DS's superior trajectory accuracy, achieving the best performance in both Traj-path and Traj-interact. By leveraging gaze-informed affordances, GAP3DS predicts trajectories that closely align with ground truth, particularly in scenarios requiring precise spatial alignment for effective human-object interactions. In pose prediction, GAP3DS outperforms the baselines, achieving the lowest errors in both average pose error (MPJPE-path) and interaction pose error (MPJPE-interact). While SIF3D captures general trajectory patterns effectively, it struggles to refine poses and model object-specific interactions due to its reliance on gaze coordinates without affordance reasoning. GAP3DS addresses these limitations by integrating its dual-prompted motion decoder and affordance-aware pose generator, enabling semantically consistent and physically plausible motion predictions.

Overall, these experiments demonstrate GAP3DS's adaptability and effectiveness in synthetic 3D environments. By incorporating object-level affordance reasoning and adaptive pose generation, GAP3DS consistently outperforms existing SoTAs, delivering accurate, contextually relevant, and physically coherent trajectory and pose predictions.

Method	Trajectory Deviation		Pose MPJPE (in mm)	
	Traj-P↓	Traj-I↓	MPJPE-P↓	MPJPE-I↓
AuxFormer [18]	772	1072	182.0	250.2
GIMO [22]	683	903	164.6	234.2
MDP [17]	650	851	163.5	220.5
SIF3D [8]	626	836	164.9	227.7
GAP3DS	**614**	**802**	**150.1**	**210.5**

Table 2. Comparasion of trajectory deviation and pose MPJPE on GTA-IM [1]. The best results are highlighted in bold.

D. Ablation Study on GazeNet

GazeNet serves as a temporal aggregator, converting distance maps into interaction maps to predict future interactive objects with high accuracy. To further explore the optimal method for predicting interaction scores, we compare the default convolution approach with alternatives, including Max pooling, Average pooling, and recurrent operations.

The results in Table 3 highlight the superiority of the convolution approach, which effectively captures both spatial and temporal patterns through its convolution operations. In contrast, Average pooling fails to filter noise caused by gaze shifts, leading to diluted interaction scores. Max pooling highlights high interaction regions but is overly sensitive to individual points, ignoring surrounding contextual features and producing noisy predictions. Recurrent operations enhance temporal pattern recognition but struggle with spatial features and impose high computational costs, limiting scalability.

These findings emphasize the robustness and efficiency of convolution operations in GazeNet. By smoothing noise and accurately capturing interaction patterns, GazeNet consistently generates reliable interaction maps, making it the most suitable method for predicting future interactive objects in GAP3DS.

E. Experiments on Long-term Prediction

Long-term motion prediction plays a critical role in applications requiring extended temporal understanding [2, 9, 15, 19]. However, extending prediction horizons often results in challenges such as accumulated errors and diminished contextual coherence.

To evaluate the robustness and accuracy of GAP3DS in long-term motion prediction, we test its performance across 5 intervals, from 4200 ms to 5000 ms, with 200 ms increments. Table 4 reports MPJPE results, providing a fine-grained assessment of each model's capability to maintain prediction quality over time. GAP3DS demonstrates consistent superiority, achieving the lowest MPJPE at all time intervals. While SIF3D and MDP maintain reasonable per-

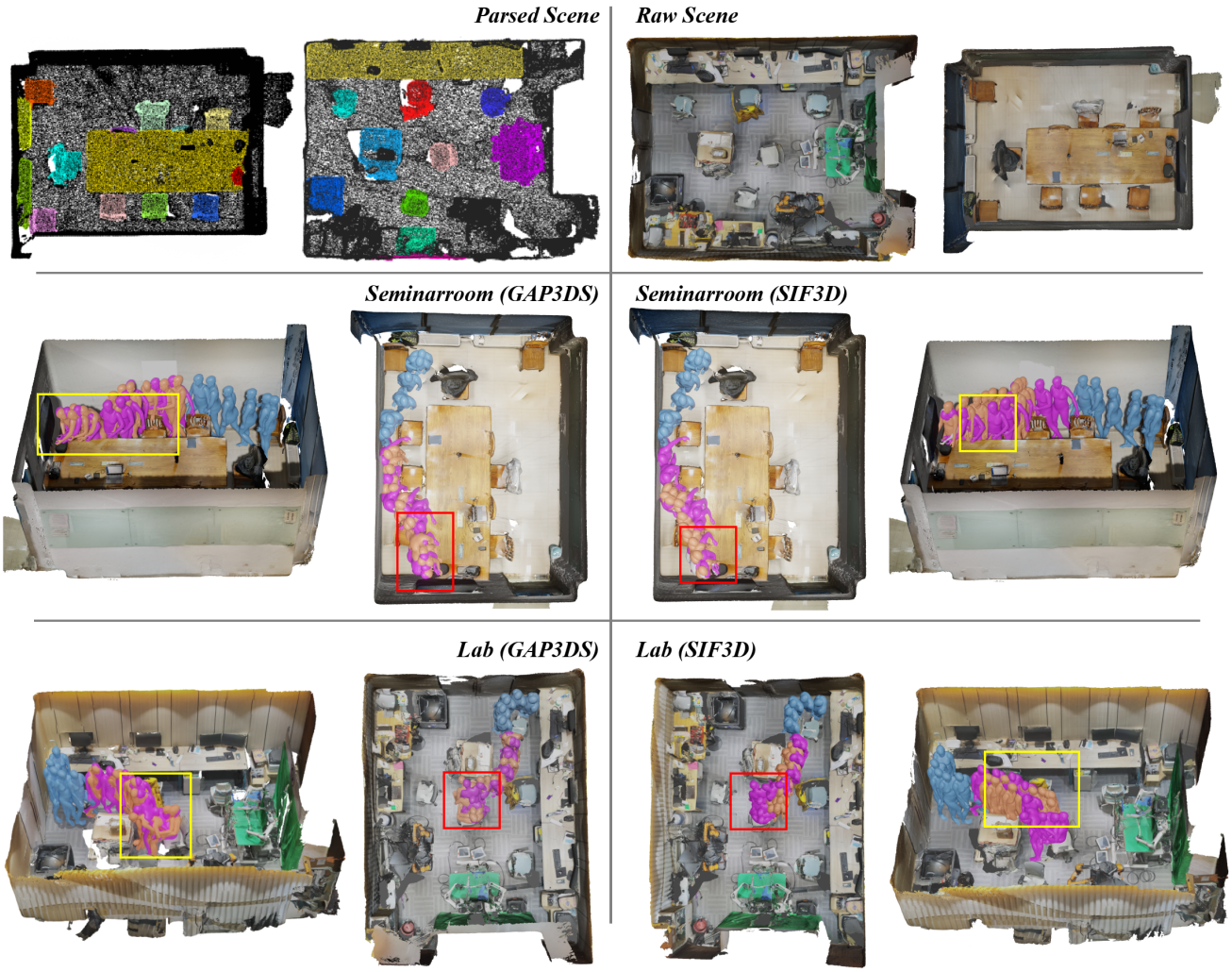

Figure 1. Additional visual comparison of GAP3DS with SoTA SIF3D [8] across two challenging indoor scenarios: Seminar Room and Lab. The **purple** human meshes represent the ground truth, the **brown** meshes indicate the model predictions, and the **blue** meshes represent historical motion sequences. GAP3DS outperforms SIF3D by achieving accurate trajectories, smooth and continuous poses, and physically consistent predictions. Its ability to model natural human-object interactions demonstrates the effectiveness of gaze-guided affordances and dual-prompted mechanisms, making it a robust solution for human motion prediction in real-world 3D environments.

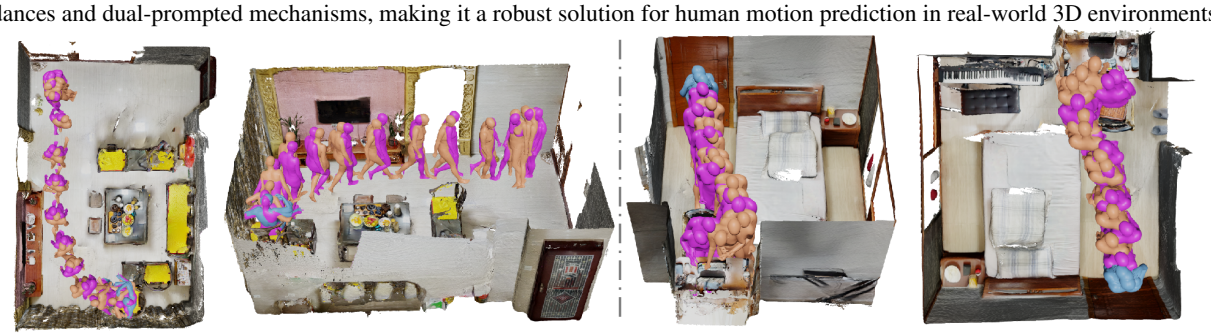

Figure 2. Visualization on non-interaction and multi-interaction cases. In rare non-interaction scenarios (left), the predicted motion aligns well with the expected trajectory, demonstrating temporal coherence when no clear affordance target is present. The failure case (right), where the model predicts direct interaction with the computer **(orange)** instead of first sitting down **(purple)**, reveals limitations in handling sequential interactions.

Aggregator	Traj-P ↓	Traj-I ↓	MPJPE-P ↓	MPJPE-I ↓
Ave. Pool.	585	637	150.9	182.5
Max Pool.	577	629	146.3	176.6
Rec. Op.	577	**621**	**140.5**	173.0
Conv. Op.	**576**	623	141.2	**171.4**

Table 3. Ablation study of GazeNet on GIMO [22]. Comparison of the default convolution operation with max pooling, average pooling, and recurrent approach for predicting interaction scores. The best results are highlighted in bold.

MPJPE	4.2s	4.4s	4.6s	4.8s	5.0s
AuxFormer [18]	174.1	185.5	186.4	199.0	217.1
BiFU [22]	173.2	183.0	187.7	198.2	214.8
MDP [17]	155.1	170.9	178.4	184.8	200.5
SIF3D [8]	150.6	168.1	174.9	189.2	206.0
GAP3DS	**143.4**	**153.9**	**158.3**	**174.3**	**185.2**

Table 4. Experiments on long-term motion prediction on GIMO [22]. The best results are highlighted in bold.

formance at shorter intervals, their errors escalate significantly at longer horizons, reflecting limitations in temporal consistency. These findings underscore GAP3DS's capacity to deliver robust and reliable long-term motion forecasts.

F. More Visualizations

Figure 1 showcases the visualization comparison between GAP3DS and SIF3D across two challenging indoor scenarios: Seminar Room and Lab. Purple represents ground-truth motions, brown indicates model predictions, and blue denotes observed trajectories. GAP3DS demonstrates robust scene parsing, accurately identifying and segmenting interactive objects such as desks, chairs, and boxes in the parsed scenes (top-left panels). This capability enables precise spatial reasoning and facilitates contextually aligned motion predictions.

In the seminar scenario, SIF3D accurately predicts the trajectory but encounters significant issues with pose continuity. The final pose appears disconnected from the preceding motions, resulting in an unnatural and disjointed sequence, especially when attempting to interact with the box on the table. In contrast, GAP3DS maintains not only trajectory accuracy but also seamless pose transitions throughout the sequence. The final pose aligns naturally with the earlier frames, effectively capturing the intended motion of picking up the box, demonstrating both contextual relevance and physical coherence.

In the lab scenario, SIF3D demonstrates limitations in both trajectory and pose prediction. The predicted trajectory deviates noticeably, failing to navigate effectively around scene obstacles. This limitation results in the motion prediction missing essential interactions with the target object. Additionally, SIF3D produces severe physical inconsistencies, such as the subject penetrating the table, highlighting a lack of respect for spatial boundaries and object geometries. In contrast, GAP3DS accurately navigates around the table to reach the target chair, delivering a consistent and contextually accurate motion prediction. Its trajectory remains well-aligned with the ground truth, and the generated poses accurately represent the "sitting" action, producing a seamless and physically coherent motion sequence that respects spatial constraints. These results underscore GAP3DS's capability to generate environment-aware and interaction-consistent human motion sequences in real-world 3D scenes.

In summary, GAP3DS outperforms SIF3D by achieving accurate trajectories, smooth and continuous poses, and physically consistent predictions. Its ability to model natural human-object interactions demonstrates the effectiveness of gaze-guided affordances and dual-prompted mechanisms, making it a robust solution for human motion prediction in real-world 3D environments.

G. Non-Interaction and Sequential Cases

Figure 2 illustrates non-interaction and sequential interaction failure cases in human motion prediction. For rare non-interaction scenarios, GAP3DS maintains temporal coherence, even when gaze does not clearly indicate a target object or affordance scores remain low. As shown in the first case, the predicted trajectory remains stable without noticeable anomalies. However, failure cases reveal limitations in multi-stage sequential interactions. As shown in the second case, the predicted motion (orange) incorrectly anticipates direct interaction with the computer instead of first sitting down before engagement, as observed in the ground truth (purple). This highlights the need for improved modeling of sequential human-object interactions.

H. Limitations

While GAP3DS excels in predicting human motion within 3D environments, it has limitations. The model relies heavily on accurate gaze data, which can be noisy in complex scenarios, and its affordance predictions are restricted to single objects, overlooking multi-object or sequential interactions. Furthermore, its generalization to unseen domains, such as outdoor environments, remains untested. Future work will focus on improving robustness to noisy inputs, expanding affordance modeling for multi-object and sequential interactions, and exploring cross-domain adaptability for diverse environments.

References

[1] 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. 1, 3

[2] 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. 3

[3] Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J Black. Wandr: Intention-guided human motion generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 927–936, 2024. 2

[4] 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. 2

[5] 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)*, 2021. b. 2

[6] 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

[7] 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)*, 2023. 2

[8] Zhenyu Lou, Qiongjie Cui, Haofan Wang, Xu Tang, and Hong Zhou. Multimodal sense-informed prediction of 3d human motions. In *IEEE IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 4, 5

[9] Thien-Minh Nguyen, Shenghai Yuan, Thien Hoang Nguyen, Pengyu Yin, Haozhi Cao, Lihua Xie, Maciej Wozniak, Patric Jensfelt, Marko Thiel, Justin Ziegenbein, and Noel Blunder. Mcd: Diverse large-scale multi-campus dataset for robot perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22304–22313, 2024. 3

[10] 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. 2

[11] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *Acm transactions on graphics (tog)*, 40(1):1–15, 2020. 2

[12] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human Motion Diffusion Model. *ArXiv*, abs/2209.14916, 2022. 1, 2

[13] 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. 1

[14] 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. 2

[15] 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. 3

[16] 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

[17] 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. 3, 5

[18] 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. 3, 5

[19] 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. 3

[20] 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. 1

[21] 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

[22] 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, 3, 5