Multi-modal Knowledge Distillation-based Human Trajectory Forecasting

Supplementary Material

1. Overview

The supplementary material provides further insight to our paper. The contents are as follows:
-Section 2: Details on model implementation
-Section 3: Details on dataset
-Section 4: Further results and ablations
-Section 5: More qualitative results
-Section 6: Limitations

2. Details on model implementation

2.1. LED

The denoising process of original LED model is analogous to a latent diffusion model [3], where the latent encoded from historical information is used for each denoising step. For both $\mathcal{X}$ and $\mathcal{X}+\mathcal{P}$ models, we add a heading vector (at t=0) embedding encoded from MLP for instantaneous predictions when encoding the latent for denoising. For $\mathcal{X}+\mathcal{P}$ model, we additionally acquire pose embeddings with MLP followed by cross attention across trajectory and pose embeddings to acquire the final historical latent used for denoising. Attention across trajectory and pose embeddings for all timesteps and heading vector embedding are holistically considered. Total number of parameters is 3.6M.

2.2. socialTransmotion

Similarly, we add heading vector embedding for both $\mathcal{X}$ and $\mathcal{X}+\mathcal{P}$ models as a additional token to be considered for cross attention. We preserve all other network designs for both models. Total number of parameters is 3.9M.

2.3. MART

Details are found in the main paper. Total number of parameters is 3.0M.

2.4. HiVT

Details are found in the main paper. Total number of parameters is 3.3M.

3. Details on dataset

3.1. JRDB/SIT dataset

For both JRDB and SIT datasets, we first extract the global location from 3D bounding box annotations and use them as trajectories for each agent in each frame. We adjust the FPS to 2.5 frames per second and predict 12 frames given 8 frames to match the setting for ETH/UCY dataset. Human pose is extracted for agents within a distance ζ and with no occlusion via SOTA 3D pose estimation algorithm [4]. The body orientation around the height axis is rotated based on the z-axis rotation angle provided in 3D bounding box annotation. ζ is set as 5m for JRDB and 10m for SIT to minimize noise incorporated with human pose. Single image resolution for JRDB is inferior with 752 × 480 while SIT image resolution acquired as 1920 × 1200, allowing for accurate pose acquisition on agents further away for SIT dataset. For JRDB dataset, text annotation $\mathcal{S}$ includes both agent caption $\mathcal{S}_{\text{A}}$ describing what the agent is doing and interaction relationship caption $\mathcal{S}_{\text{R}}$ between certain agents involved in a same activity. For SIT dataset, no text annotations are provided, so we use vision language model (VLM) [5] to acquire agent caption $\mathcal{S}_{\text{A}}$ for each visible agent. Specifically, we use cropped images around the bounding box for all historical timesteps, therefore a total of 8 frames are used for captioning an agent of a given timestep. Algorithm 1 describes the process of constructing the JRDB and SIT dataset.

Algorithm 1 JRDB/SIT dataset construction algorithm

```
for $t = 1, t{+}{+}$, while $t < T$ do
    $N \leftarrow$ number of persons with 3D bounding box
            annotation
    for $n = 1, n{+}{+}$, while $n <= N$ do
        $\mathcal{M}_n^t \leftarrow X_n^t$
        if $L2(X_{robot}^t, X_n^t) < \zeta$ and No occlusion then
            $\mathcal{M}_n^t \leftarrow \phi_{\mathcal{P}}(I_n^t)$
            if SIT then
                $\mathcal{M}_n^t \leftarrow \phi_{\mathcal{S}}(\phi_{\text{VLM}}(I_n^{t-T_F:t}))$
            end if
        end if
        if JRDB then
            $\mathcal{M}_n^t \leftarrow \phi_{\mathcal{S}}(\mathcal{S}_{\text{A}})$
            $\mathcal{M}_n^t \leftarrow \phi_{\mathcal{S}}(\mathcal{S}_{\text{R}})$
        end if
    end for
end for
```

In algorithm 1, I_n^t denotes a cropped image for agent n on timestep t, and $\mathcal{M}_n^t$ denotes modality information available for agent n on timestep t. Also, while only agent description caption $\mathcal{S}_{\text{A}}$ is acquired for SIT dataset, we additionally utilize agent relationship caption $\mathcal{S}_{\text{R}}$ from annotations for JRDB dataset. $\phi_{\mathcal{P}}$ denotes pose extractor [4], $\phi_{\mathcal{S}}$ denotes pre-trained BERT encoder [1], and ϕ_{VLM} denotes vision-language model (VLM) [5] used to acquire caption for each agent. While the default prompt to acquire caption is set as ”What is the person doing?”, we also compare using different prompts as shown in table 6 of main paper.

The exact prompts used are as follows:

Prompt 1
Shortened: ”what is the person doing?”
Full: ”The input consists of camera view from a mobile robot. In the images, people are standing still or walking, alone or in a group. Your job is to describe the behavior of the person marked with red bounding box. What is the person doing? In one full sentence. Concisely, less than 20 words. Do not talk about clothing of person.”

Prompt 2
Shortened: ”Is there any obstacle in front of the person?”
Full: ”The input consists of camera view from a mobile robot. In the images, people are standing still or walking, alone or in a group. Is there any obstacle in the way of person with red bounding box? In one full sentence. Concisely, less than 20 words.”

Prompt 3
Shortened: ”What will the person do in the future?”
Full: ”The input consists of camera view from a mobile robot. In the images, people are standing still or walking, alone or in a group. What will the person in red bounding box do in the future? In one full sentence. Concisely, less than 20 words.”

Figure 1 compares the acquired agent captions $\mathcal{S}_\mathrm{A}$ with different prompts.

3.2. ETH/UCY dataset

While overall dataset construction process for ETH/UCY is analogous to that of JRDB/SIT, the main difference is the form of representation for pose and how text captions are obtained. For human pose, we use 2D CLIP vision encoder [2] features instead of 3D pose, as an accurate 3D pose was unable to be obtained due to the low resolution and limited visibility of BEV (Bird eye view) perspective. For text, neither agent-specific captions are available or VLM able to obtain text captions. Specifically, as VLM failed to generate accurate captions with a low resolution BEV input, we resort to a rule-based method to generate text captions that captures the surrounding obstacle information for each agent. In doing so, we use 2D walkable area segmentation map in image coordinates to generate text captions that contain map information for each agent. Text is generated based on agent’s speed and surrounding obstacles. The following algorithm is used to construct the dataset.

In algorithm 2, heading vector is the vector from X_n^{t-1} to X_n^t, which is same as the heading vector used for rotation-invariant encoding of HiVT [6]. Obstacle vector is vector from X_n^t to the point of closest obstacle.

Algorithm 2 ETH/UCY dataset construction algorithm

```
for $t = 1, t{+}{+}$, while $t < T$ do
    $N \leftarrow$ number of persons with location annotation
    for $n = 1, n{+}{+}$, while $n <= N$ do
        $\mathcal{M}_n^t \leftarrow X_n^t$
        $\mathcal{M}_n^t \leftarrow \phi_\mathcal{P}(I_n^t)$
        $\mathcal{S}_n^t =$ ‘ ’
        if $L2(X_n^t, X_n^{t-1}) < 1.5$pixels then
            $\mathcal{S}_n^t \leftarrow$ ‘The person is standing still.’
        else if $L2(X_n^t, X_n^{t-1}) < 20$pixels then
            $\mathcal{S}_n^t \leftarrow$ ‘The person is walking slowly.’
        else
            $\mathcal{S}_n^t \leftarrow$ ‘The person is walking.’
        end if
        if $L2(\text{Nearest obstacle}, X_n^t) < \zeta$ then
            $\sigma = \text{Heading vector} - \text{Obstacle vector}$
            if $\sigma > 30°$ and $\sigma < 100°$ then
                $\mathcal{S}_n^t \leftarrow$ ‘There is an obstacle on the right.’
            else if $\sigma < 30°$ and $\sigma > -30°$ then
                $\mathcal{S}_n^t \leftarrow$ ‘There is an obstacle in front.’
            else if $\sigma < -30°$ and $\sigma > -100°$ then
                $\mathcal{S}_n^t \leftarrow$ ‘There is an obstacle on the left.’
            else
                $\mathcal{S}_n^t \leftarrow$ ‘There is no obstacle in the heading direction of the person.’
            end if
        else
            $\mathcal{S}_n^t \leftarrow$ ‘There is no obstacle around.’
        end if
        $\mathcal{M}_n^t \leftarrow \phi_\mathcal{S}(\mathcal{S}_n^t)$
    end for
end for
```

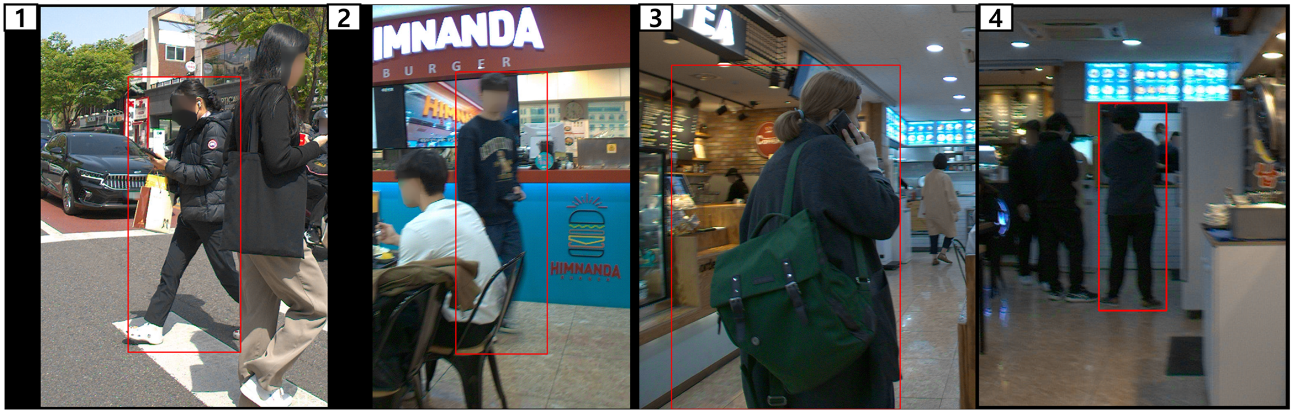
3.3. Dataset visualization

Figures 3, 4 visualize two scenes from JRDB dataset, one from indoor and one from outdoor. For JRDB scenes, we show agent and interaction text captions for only a few selected agents for compact visualization. For SIT dataset visualized in Fig. 2, the captions are acquired with the default prompt “What is the person doing?”. Origin denotes position of robot which captured the scene with omnidirectional cameras.

3.4. Scene parsing for Train/Validation sets

As only training sets of JRDB/SIT are provided with full annotations (3D bounding box, 2D bounding box, text annotation), we parse the training set of the initially provided dataset into train (85%) and validation sets (15%). The parsed validation set contains a similar ratio of indoor and outdoor scenes.
Scenes for JRDB train set parsing:

Prompt 1: What is the person doing?
Prompt 2: Is there any obstacle in front of the person?
Prompt 3: What will the person do in the future?



1. Person is walking across the street.
No obstacle in the path of the person with the red bounding box.
The person in red bounding box is about to cross the street.

2. Man with red box standing and walking with phone.
No obstacle visible.
Person in red bounding box will continue to walk in same direction.

3. Talking on cell phone.
No obstacle in way of person with red bounding box.
Keep walking while talking on phone.

4. The person marked with a red bounding box appears to be standing still, perhaps waiting in line.
No obstacle.
The person in red bounding box is likely to walk up to counter and place an order, based on the context of the scene which suggests he is at a restaurant or food service place.

Figure 1. Examples of images used for captioning via VLM and the acquired agent captions $\mathcal{S}_{\mathrm{A}}$ for SIT dataset. Colors denote the type of prompts used on the VLM. Each image denotes the last frame of 8 frames used for captioning. Person of interest is marked by the red bounding box.

- clark-center-2019-02-28_0
- clark-center-2019-02-28_1
- clark-center-intersection-2019-02-28_0
- cubberly-auditorium-2019-04-22_0
- forbes-cafe-2019-01-22_0
- gates-159-group-meeting-2019-04-03_0
- gates-ai-lab-2019-02-08_0
- gates-basement-elevators-2019-01-17_1
- hewlett-packard-intersection-2019-01-24_0
- huang-2-2019-01-25_0
- huang-basement-2019-01-25_0
- gates-to-clark-2019-02-28_1
- memorial-court-2019-03-16_0
- meyer-green-2019-03-16_0
- nvidia-aud-2019-04-18_0
- packard-poster-session-2019-03-20_1
- packard-poster-session-2019-03-20_2
- stlc-111-2019-04-19_0
- svl-meeting-gates-2-2019-04-08_0
- svl-meeting-gates-2-2019-04-08_1
- tressider-2019-03-16_0
- tressider-2019-03-16_1
- tressider-2019-03-16_2

Scenes for JRDB validation set parsing:

- bytes-cafe-2019-02-07_0
- huang-lane-2019-02-12_0
- jordan-hall-2019-04-22_0
- packard-poster-session-2019-03-20_0

Scenes for SIT train set parsing:

- Cafe_street_2-001
- Cafeteria_1-006
- Cafeteria_2-005
- Cafeteria_5-003
- Corridor_2-007
- Corridor_3-004
- Corridor_5-003
- Corridor_7-001
- Corridor_8-006
- Corridor_9-011
- Corridor_10-009
- Corridor_11-002
- Courtyard_1-009
- Courtyard_2-002
- Courtyard_4-008
- Courtyard_5-005
- Courtyard_6-007

Table 1. Statistical comparison between JRDB and SIT datasets.

	Duration (s)	Indoor #	Outdoor #	Sample #	Max N	Avg. N
JRDB	1860	17	10	24788	38	16
SIT	940	4	6	5141	19	6

- Courtyard_8-001
- Courtyard_9-003
- Crossroad_1-001
- Hallway_1-011
- Hallway_2-003
- Hallway_3-001
- Hallway_4-012
- Hallway_6-009
- Hallway_7-010
- Hallway_8-005
- Hallway_9-008
- Hallway_10-006
- Hallway_11-002
- Lobby_2-009
- Lobby_3-007
- Lobby_4-002
- Lobby_5-008
- Lobby_7-004
- Lobby_8-003
- Outdoor_Alley_2-003
- Subway_Entrance_2-004
- Subway_Entrance_4-003
- Three_way_Intersection_3-006
- Three_way_Intersection_4-001
- Three_way_Intersection_5-005
- Three_way_Intersection_8-002

Scenes for SIT validation set parsing:

- Cafe_street_1-002
- Cafeteria_3-004
- Corridor_1-010
- Lobby_6-001
- Outdoor_Alley_3-002
- Three_way_Intersection_4-001
- Subway_Entrance_2-004

Furthermore, Tab. 1 compares the statistics between JRDB and SIT dataset. JRDB dataset consists of more complex scenes with larger average and maximum number of agents in a given sequence. In addition, total duration of JRDB is 2× than SIT dataset, making JRDB a more challenging and thorough dataset.

4. Further results and ablations

4.1. Multi-modal teacher model on ETH/UCY

Table 2 highlights the performance enhancements achieved by incorporating additional modalities on the ETH/UCY dataset. Since pose is extracted from cropped images, its generalizability across various scenes is limited. Similar to the JRDB/SIT datasets, text has proven to be the most impactful modality, even when generated using a rule-based approach. These additional modalities are particularly beneficial for instantaneous predictions. Although the absolute percentage improvement is smaller than that observed in the JRDB/SIT datasets due to the limited richness of available modalities, the teacher model leveraging all modalities still outperforms the base model $\mathcal{X}$, indicating potential for further gains through KD.

4.2. Can we raise a competent teacher while training only on full observation regression?

In addition to training on all observation settings with $\mathcal{L}^{\text{F}}_{\text{reg}}, \mathcal{L}^{2}_{\text{reg}}, \mathcal{L}^{1}_{\text{reg}}$ as shown in table 1 of main paper, we experiment only training with full observation $\mathcal{L}^{\text{F}}_{\text{reg}}$. As shown in Tab. 3, using additional modalities improved the prediction performance even when only trained with $\mathcal{L}^{\text{F}}_{\text{reg}}$. Again, text has been the most crucial modality. When only trained on $\mathcal{L}^{\text{F}}_{\text{reg}}$ with sufficient past trajectory information, the generalizable nature of text is most valued. Interestingly, compared to $(\mathcal{X}+\mathcal{P}+\mathcal{S})$ improving over $(\mathcal{X}+\mathcal{S})$ when trained also on $\mathcal{L}^{2,1}_{\text{reg}}$ as shown in main paper table 1, Tab. 3 does not show such improvement. This shows that eliciting a deeper understanding on human intent with pose is most effectively done by training the model on both full ($\mathcal{L}^{\text{F}}_{\text{reg}}$) and instantaneous predictions ($\mathcal{L}^{2,1}_{\text{reg}}$), therefore building a more competent teacher model.

4.3. Which representation of human pose is most effective?

Table 4 compares the prediction performance of HiVT $\mathcal{X}+\mathcal{P}$ model with SMPL theta or 3D joint representations. For both MART and HiVT, using SMPL as the pose representation leads to improvements in performance, whether or not KD is applied. The angle-based representation of SMPL clearly captures the inter-joint relationships, making it more resilient to noise and more generalizable compared to the 3D joint representation, where these relationships are not explicitly defined. In particular, a more significant improvement is seen in models with KD, highlighting that a general angle-based representation can effectively integrate the diverse knowledge from textual descriptions.

4.4. Can we teach instantaneous prediction without explicit loss during KD?

Table 5 presents a comparison of the regression loss combinations used for training the teacher model and the student model during KD. The upper two rows compare using different teacher models: teacher additionally trained on instantaneous prediction or only trained on full observation. Then, the student model is distilled with regression loss based only on full observation. The performance of

Table 2. Prediction results on ETH/UCY (Average) with different modality ($\mathcal{M}$) combinations, with human pose ($\mathcal{M}$) and text ($\mathcal{S}$) as additional input. Trained with all regression losses ($\mathcal{L}^{\text{F}}_{\text{reg}}$,$\mathcal{L}^{2}_{\text{reg}}$,$\mathcal{L}^{1}_{\text{reg}}$). Bold denotes best. Lower is better.

$\mathcal{M}$	HiVT							MART						
	ADE	ADE$_2$	ADE$_1$	FDE	FDE$_2$	FDE$_1$	Avg. +%	ADE	ADE$_2$	ADE$_1$	FDE	FDE$_2$	FDE$_1$	Avg. +%
$\mathcal{X}$	**0.314**	**0.320**	0.555	**0.718**	**0.726**	1.132	-	0.364	0.369	0.571	0.852	0.862	1.166	-
$\mathcal{X}+\mathcal{S}$	0.331	0.340	0.490	0.748	0.760	1.005	+0.45	0.355	**0.360**	0.521	**0.823**	**0.828**	1.083	+4.69
$\mathcal{X}+\mathcal{P}+\mathcal{S}$	0.339	0.350	**0.444**	0.761	0.772	**0.920**	**+1.57**	**0.352**	0.367	**0.472**	0.852	0.871	**1.064**	**+4.81**

Table 3. Prediction results on JRDB with different modality ($\mathcal{M}$) combinations, with human pose ($\mathcal{M}$) and text ($\mathcal{S}$) as additional input. Only trained on regression loss with full past observation, $\mathcal{L}^{\text{F}}_{\text{reg}}$. Bold denotes best. Lower is better.

$\mathcal{M}$	HiVT							MART						
	ADE	ADE$_2$	ADE$_1$	FDE	FDE$_2$	FDE$_1$	Avg. +%	ADE	ADE$_2$	ADE$_1$	FDE	FDE$_2$	FDE$_1$	Avg. +%
$\mathcal{X}$	0.223	0.246	0.718	0.434	0.476	1.231	-	0.274	0.288	0.629	0.526	0.547	1.127	-
$\mathcal{X}+\mathcal{P}$	0.229	0.251	0.573	0.448	0.486	1.015	+4.65	0.291	0.294	0.596	0.549	0.551	1.097	-0.90
$\mathcal{X}+\mathcal{S}$	**0.220**	**0.240**	0.449	**0.430**	**0.462**	0.848	+12.72	**0.267**	**0.265**	**0.347**	**0.517**	0.513	**0.643**	**+17.75**
$\mathcal{X}+\mathcal{P}+\mathcal{S}$	0.223	0.242	**0.437**	0.438	0.469	**0.782**	**+12.98**	0.278	0.270	0.403	0.528	**0.512**	0.721	+13.79

Table 4. Ablation on representation format for pose $\mathcal{P}$. Exprimented on JRDB dataset with HiVT $\mathcal{X}+\mathcal{P}$ model.

Model	$\mathcal{P}$ type	w/ KD	ADE	ADE$_2$	ADE$_1$	FDE	FDE$_2$	FDE$_1$	Avg. +%
MART	3D joint		0.288	0.286	0.394	0.553	0.547	0.727	
		✓	0.271	0.268	0.381	0.520	0.517	0.721	-
	SMPL		**0.287**	**0.282**	**0.366**	**0.543**	**0.538**	**0.682**	+3.08
		✓	**0.266**	**0.261**	**0.337**	**0.519**	**0.507**	**0.633**	**+5.07**
HiVT	3D joint		0.232	0.248	**0.354**	0.446	0.480	**0.646**	-
		✓	**0.229**	0.240	0.325	0.446	0.464	0.598	-
	SMPL		**0.229**	**0.242**	0.364	**0.441**	**0.465**	0.659	+0.56
		✓	0.232	**0.239**	**0.308**	**0.445**	**0.464**	**0.560**	**+1.87**

Table 5. Ablation on using different combinations for regression loss ($\mathcal{L}^{\text{F}}_{\text{reg}}$, $\mathcal{L}^{1}_{\text{reg}}$, $\mathcal{L}^{2}_{\text{reg}}$) on student, teacher models. Experimented on JRDB.

Teacher		Student		Metrics					
$\mathcal{L}^{\text{F}}_{\text{reg}}$	$\mathcal{L}^{2,1}_{\text{reg}}$	$\mathcal{L}^{\text{F}}_{\text{reg}}$	$\mathcal{L}^{2,1}_{\text{reg}}$	ADE	ADE$_2$	ADE$_1$	FDE	FDE$_2$	FDE$_1$
✓		✓		0.227	0.282	0.725	0.442	0.546	1.330
✓	✓	✓		**0.222**	**0.277**	**0.498**	**0.431**	**0.534**	**0.907**
✓		✓	✓	**0.222**	**0.233**	0.319	**0.437**	**0.452**	0.578
✓	✓	✓	✓	0.232	0.239	**0.308**	0.445	0.464	**0.560**

Table 6. Ablation on number of modes F for KD. Experimented on JRDB with $\mathcal{X}+\mathcal{P}$ HiVT model.

F	w/ KD	ADE	ADE$_2$	ADE$_1$	FDE	FDE$_2$	FDE$_1$	Avg. +%
1		**0.381**	0.403	0.764	**0.729**	0.772	1.341	+1.70
	✓	**0.381**	**0.389**	**0.737**	**0.729**	**0.757**	**1.324**	
6		**0.229**	0.242	0.364	**0.441**	0.465	0.659	+4.98
	✓	0.232	**0.239**	**0.308**	0.445	**0.464**	**0.560**	
10		**0.196**	0.203	0.280	0.389	0.401	0.525	+1.56
	✓	**0.196**	**0.201**	**0.273**	**0.386**	**0.397**	**0.505**	
20		0.160	0.165	0.235	0.314	0.322	0.428	+2.42
	✓	**0.157**	**0.163**	**0.223**	**0.307**	**0.321**	**0.415**	

instantaneous prediction improves significantly when distilled using a teacher model trained on instantaneous predictions. This indicates that the teacher's knowledge of human intent through instantaneous observations is transferable, even without explicit instantaneous loss for the student model. Furthermore, when the student model is also trained with $\mathcal{L}^{2,1}_{\text{reg}}$, the student's performance in instantaneous prediction again improves when distilled by a teacher model also trained on instantaneous predictions. This once more supports the idea that training on instantaneous predictions fosters a deeper understanding of human motion intent, and that this knowledge is transferable through distillation.

4.5. Does number of modes have an influence on KD performance?

Table 6 compares the KD performance on different number of modes of future trajectory forecasting. KD consistently enhances the performance of the student model, demonstrating that the knowledge of human motion intent is transferable, regardless of the F configuration used to decode the locomotion intent into multiple trajectories. This also shows winner-takes-all regression training approach on multiple predictions is effective in transferring knowledge on human motion intent.

5. More qualitative results

Figure 5 visualizes the effect of KD (left two figures) and the effect of using interaction relationship text $\mathcal{S}_{\text{R}}$ on JRDB dataset. For KD comparison, the model originally predicts a collision with nearby agents. After KD from $\mathcal{X}+\mathcal{P}+\mathcal{S}$ model, the student model predicts a trajectory that avoids collision with other agents, even with only one or two frames of historical observation. This capability is developed by training the student model leveraging the teacher's extensive knowledge of human motion intent across multiple modalities.

Analysis of interaction relation caption is performed on the rightmost scene of Fig. 5 where two agents are waiting in line, having a conversation, and then walking together. Without utilizing relationship context, the group dynamic between these two agents is not explicitly accounted for, leading to diverging trajectories. In contrast, when the relationship is considered, the two agents initially walk together, and then the front agent moves ahead as the agent is in front in line. Interestingly, prediction on the agent behind these two agents explicitly described an interaction also improves when using interaction text. As the model predicts these two agents to walk together, the agent behind them follows the general locomotion flow made by the two agents, resulting in an improved prediction.

Figure 6 visualizes the improvement with KD on SIT dataset. The figure shows a crowded indoor scene with multiple agents in motion. Predictions made with KD improves upon student without KD for all agents. Again, using a $\mathcal{X} + \mathcal{P} + \mathcal{S}$ model as a teacher for a $\mathcal{X} + \mathcal{P}$ student enhanced the understanding of locomotion intentions in complex scenes from the comprehensive knowledge gained by the complementary modalities.

6. Limitations

Our study highlights the potential of incorporating textual information—an effective modality for representing human motion, accessible through Vision-Language Models (VLM)—into trajectory forecasting. This approach allows for a more comprehensive utilization of human pose and trajectory data to understand an agent's locomotion intent. As demonstrated, language can capture not only the intent inferred from past pose sequences but also the agent's interactions with the surrounding environment.

While our work focuses on leveraging text via Knowledge Distillation (KD), we did not explicitly integrate the 3D scene information provided by LiDAR. Including 3D scene data as an additional modality along with human pose, text, and trajectory could further enhance understanding of motion intent and improve predictions of human behavior.

References

[1] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

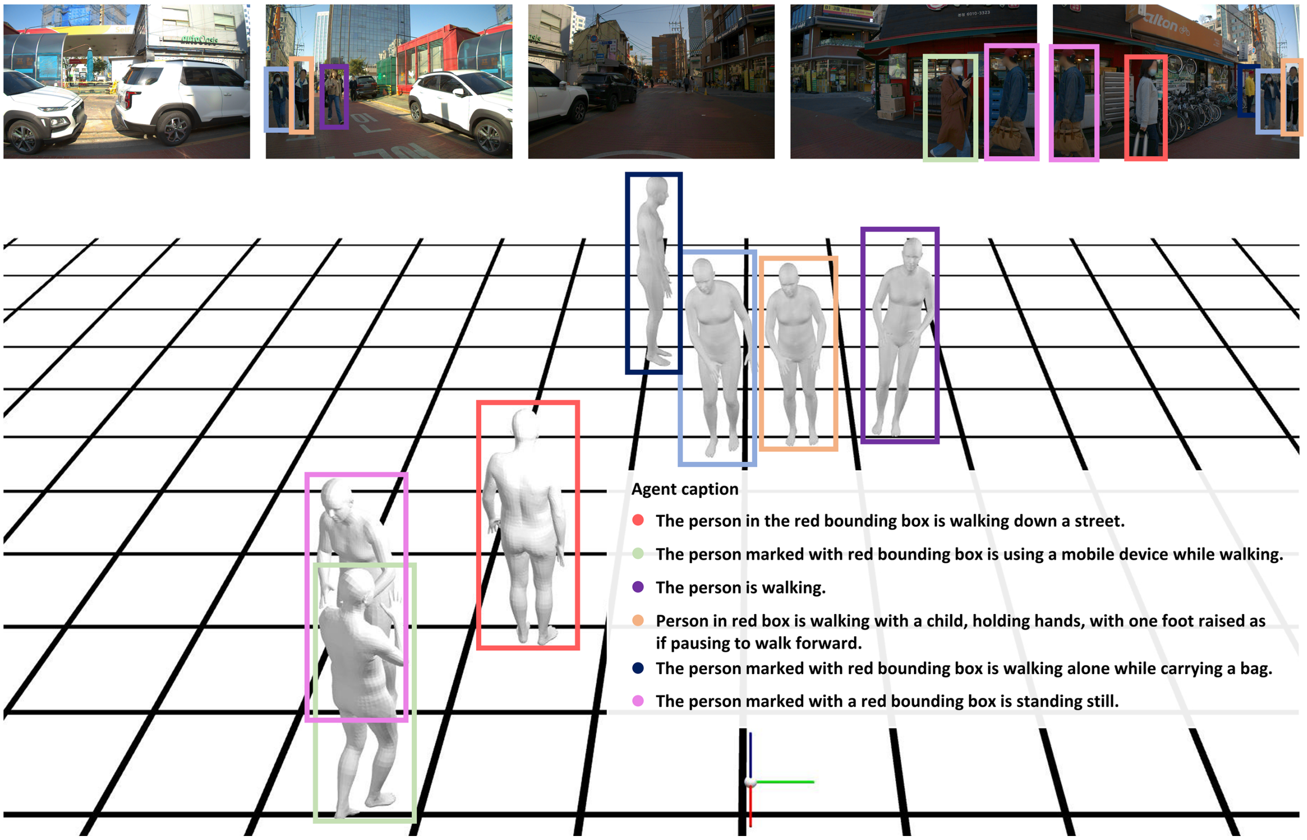
[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[4] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 1

[5] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 1

[6] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022. 2



Agent caption

- ● The person in the red bounding box is walking down a street.
- ● The person marked with red bounding box is using a mobile device while walking.
- ● The person is walking.
- ● Person in red box is walking with a child, holding hands, with one foot raised as if pausing to walk forward.
- ● The person marked with red bounding box is walking alone while carrying a bag.
- ● The person marked with a red bounding box is standing still.

Figure 2. An example scene for SIT dataset. The agent caption $\mathcal{S}_\text{A}$ are acquired from VLM with prompt ”What is the person doing?”.

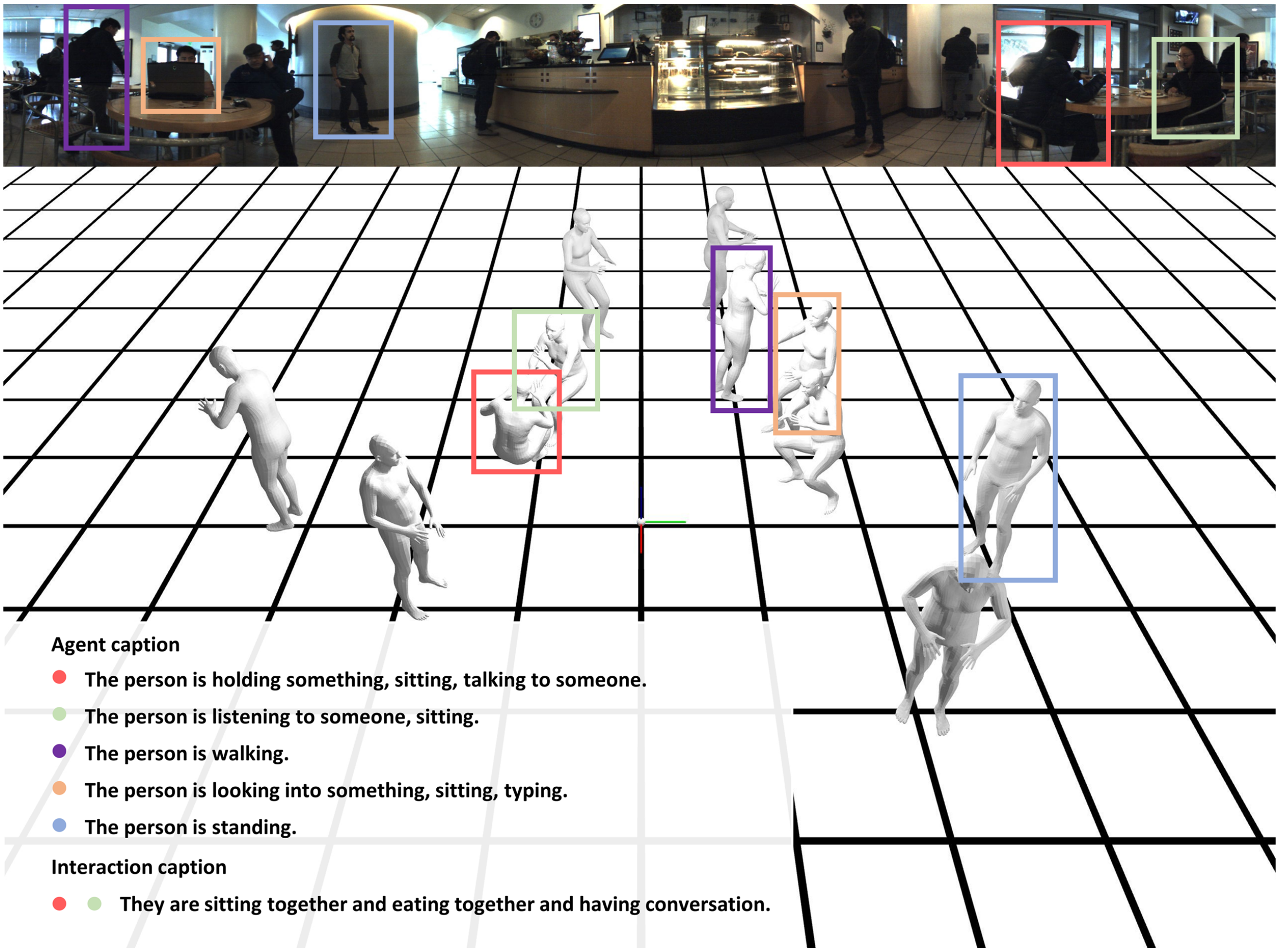

Agent caption
● The person is holding something, sitting, talking to someone.
● The person is listening to someone, sitting.
● The person is walking.
● The person is looking into something, sitting, typing.
● The person is standing.
Interaction caption
● ● They are sitting together and eating together and having conversation.

Figure 3. Example indoor scene from JRDB dataset. Agent caption $\mathcal{S}_{\mathrm{A}}$ and interaction caption $\mathcal{S}_{\mathrm{R}}$ are parsed from annotation. Captions of only a few agents are described.

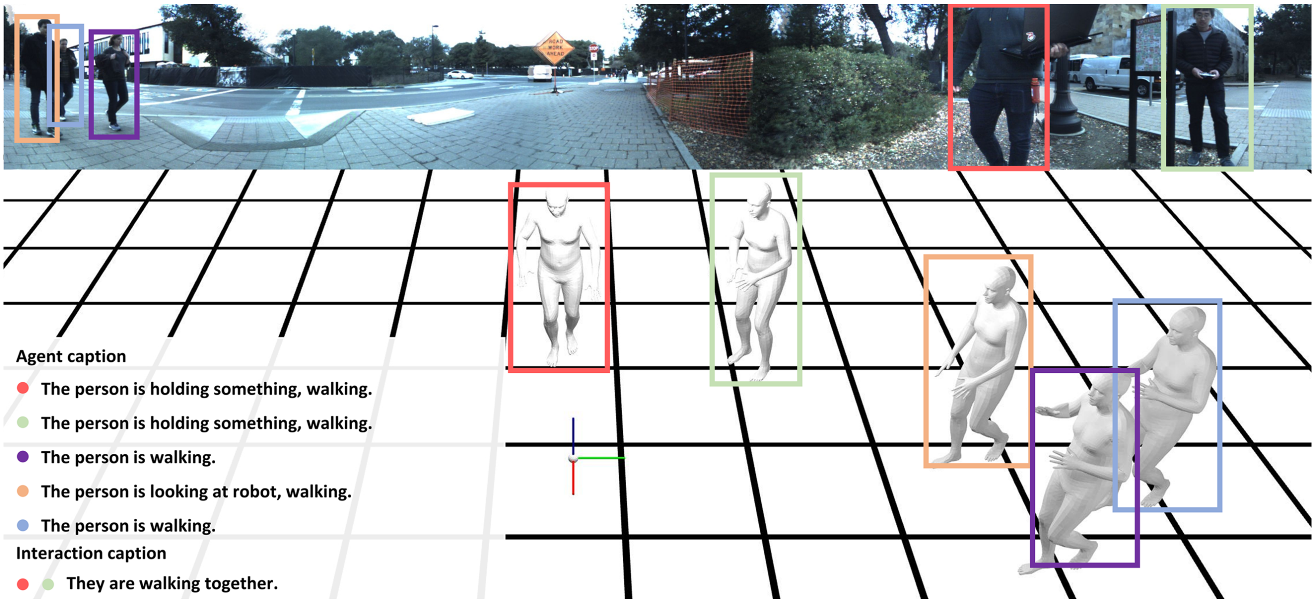

Agent caption
● The person is holding something, walking.
● The person is holding something, walking.
● The person is walking.
● The person is looking at robot, walking.
● The person is walking.
Interaction caption
● ● They are walking together.

Figure 4. Example outdoor scene from JRDB dataset. Agent caption $\mathcal{S}_{\mathrm{A}}$ and interaction caption $\mathcal{S}_{\mathrm{R}}$ are parsed from annotation.

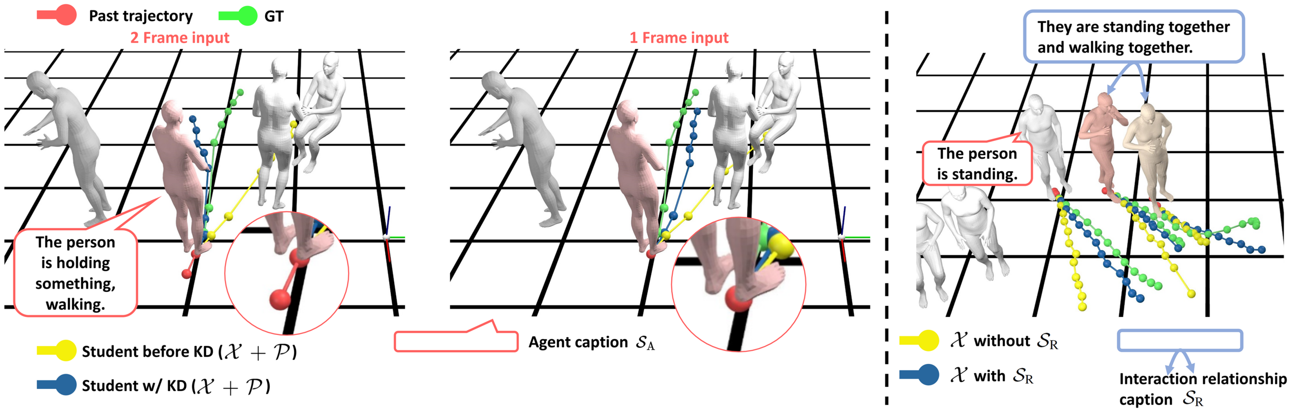

Past trajectory
GT
2 Frame input
1 Frame input
The person is holding something, walking.
Student before KD ($\mathcal{X}+\mathcal{P}$)
Student w/ KD ($\mathcal{X}+\mathcal{P}$)
Agent caption $\mathcal{S}_{\mathrm{A}}$
They are standing together and walking together.
The person is standing.
$\mathcal{X}$ without $\mathcal{S}_{\mathrm{R}}$
$\mathcal{X}$ with $\mathcal{S}_{\mathrm{R}}$
Interaction relationship caption $\mathcal{S}_{\mathrm{R}}$

Figure 5. Additional qualitative result on JRDB dataset. The left two figures compare prediction result with and without KD on student HiVT model using $\mathcal{X}+\mathcal{P}$ modalities. The right figure compares the effect of using interaction text $\mathcal{S}_{\mathrm{R}}$ on HiVT model only using $\mathcal{X}$ modality.

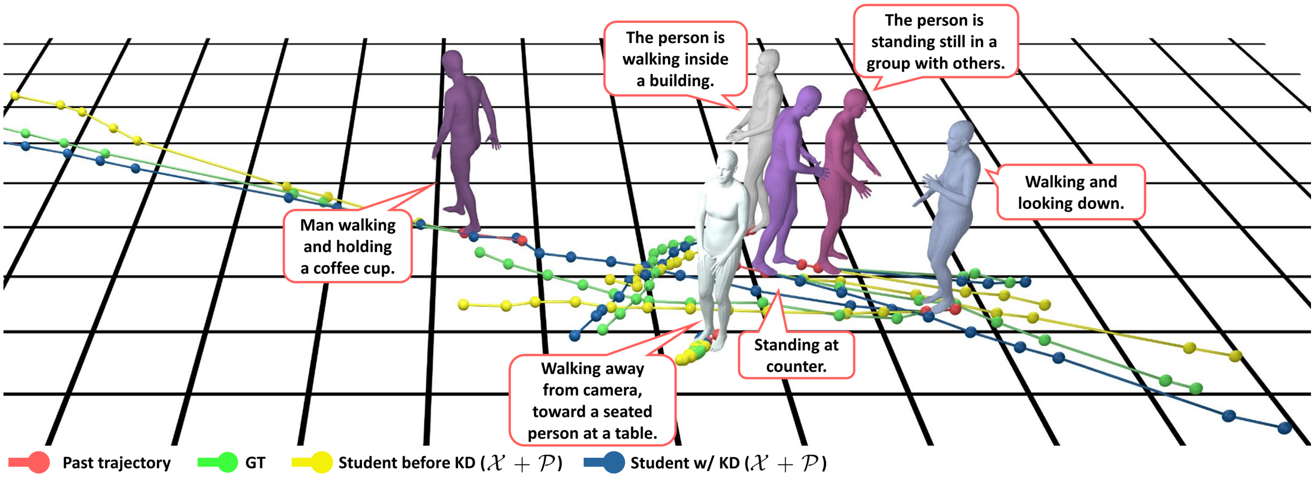

The person is walking inside a building.
The person is standing still in a group with others.
Walking and looking down.
Man walking and holding a coffee cup.
Standing at counter.
Walking away from camera, toward a seated person at a table.
Past trajectory
GT
Student before KD ($\mathcal{X}+\mathcal{P}$)
Student w/ KD ($\mathcal{X}+\mathcal{P}$)

Figure 6. Qualitative result on SIT dataset. Agent text captions are acquired from VLM with prompt ”What is the person doing?”. Prediction results between HiVT models using $\mathcal{X}+\mathcal{P}$ modalities with and without KD are compared.