Multi-modal Knowledge Distillation-based Human Trajectory Forecasting

Jaewoo Jeong[1], Seohee Lee[1], Daehee Park[2†], Giwon Lee[1], and Kuk-Jin Yoon[1]
[1]Visual Intelligence Lab., KAIST, Korea
[2]Intelligent Systems and Learning Lab., DGIST, Korea

Abstract

Pedestrian trajectory forecasting is crucial in various applications such as autonomous driving and mobile robot navigation. In such applications, camera-based perception enables the extraction of additional modalities (human pose, text) to enhance prediction accuracy. Indeed, we find that textual descriptions play a crucial role in integrating additional modalities into a unified understanding. However, online extraction of text requires the use of VLM, which may not be feasible for resource-constrained systems. To address this challenge, we propose a multi-modal knowledge distillation framework: a student model with limited modality is distilled from a teacher model trained with full range of modalities. The comprehensive knowledge of a teacher model trained with trajectory, human pose, and text is distilled into a student model using only trajectory or human pose as a sole supplement. In doing so, we separately distill the core locomotion insights from intra-agent multi-modality and inter-agent interaction. Our generalizable framework is validated with two state-of-the-art models across three datasets on both ego-view (JRDB, SIT) and BEV-view (ETH/UCY) setups, utilizing both annotated and VLM-generated text captions. Distilled student models show consistent improvement in all prediction metrics for both full and instantaneous observations, improving up to ∼13%. The code is available at `github.com/Jaewoo97/KDTF`.

1. Introduction

Trajectory forecasting aims to predict the future 2D trajectory of an agent based on its historical trajectory [10, 14, 15, 33, 44, 51, 59, 60, 63, 64]. This predictive capability is widely applied in fields such as autonomous driving [9, 16, 17, 24, 35, 36, 43–45, 66, 70], mobile robot navigation [4, 50, 52], and surveillance systems [1, 32, 46]. These applications depend on accurately forecasting the trajectories of human or vehicles to prevent collisions, enable smooth navigation, and improve situational awareness.

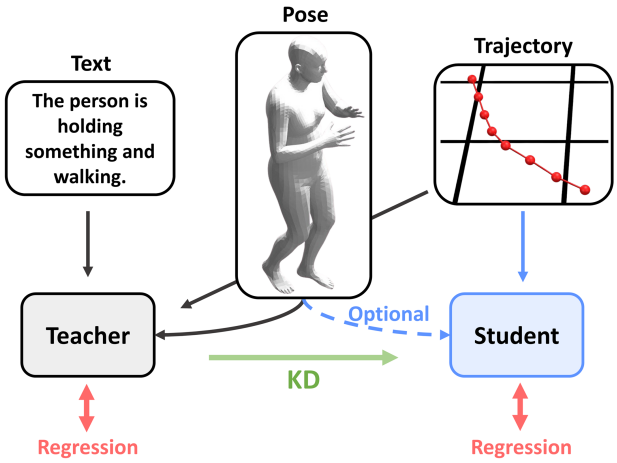

Figure 1. Multi-modal data such as human pose and text greatly improve trajectory forecasting performance. However, expensive modalities such as text are not readily available during application. Thus, we transfer the extensive knowledge from full modalities to a student model operating on a limited set of modalities.

The main challenge in trajectory forecasting lies in modeling the agent's future intent in locomotion. A clear objective on the purpose of locomotion such as a short-term (1∼2 seconds) destination significantly improves the prediction accuracy [20, 33]. In that sense, human agents naturally interact via visual signals to convey such intention to nearby agents during their maneuvers. Vehicles blink signal lights, bikers raise their arms, pedestrians rotate their torsos, all of which indicate their intent in locomotion to avoid collision with others. Diverse applications of trajectory forecasting allow for an effortless acquisition of these visual cues. In autonomous driving, vehicles employ various complementary sensors alongside RGB cameras as the primary tool for perception [5, 53]. Mobile robots and surveillance systems also rely on RGB cameras for perception and interaction with its surrounding environment and agents [4, 39].

Recent works attempt to extract and utilize additional modalities from these visual inputs. Extrinsics such as 2D, 3D human pose and bounding boxes have been examined to complement understanding human motion dynamics [50]. For vehicle trajectory forecasting, vision language models

[†]This work was completed as a Ph.D candidate at KAIST.

(VLM) are utilized to leverage textual descriptions of the perceived environment and surrounding agents for generalization [2, 41, 43, 57]. Compared to uni-modal trajectory forecasting with a sequence of mere 2D numerical coordinates, reasoning with multi-modal features allows for a more comprehensive understanding of agent's motion intent and a correspondingly accurate prediction.

Additionally, we find that textual descriptions on locomotion are essential for integrating additional modalities into a cohesive insight. For pedestrian trajectory forecasting, text effectively bridges the domain gap between trajectory and human pose, yielding the most competent model. This advantage manifests most distinctively in instantaneous prediction with only few input frames, as visual cues supplement the semantic context lost due to incomplete observation.

However, acquiring and processing multi-modal data substantially increases the demand in computational resources, which may exceed the capabilities of mobile systems. In particular, acquisition of textual information requires the use of a VLM which requires significant computational resources for a reliable performance [7, 25].

Then, how might we leverage textual insights while mitigating the computational demands associated with their acquisition? In this regard, we propose a knowledge distillation (KD) [18] framework where a student model operating with limited modalities is trained by a teacher model that utilizes the full range of modalities as shown in Fig. 1. KD resolves the computational burden, as only the affordable student model modalities are used during inference, while costly modalities are required solely during training. Specifically, the multi-modal insight from the teacher is distilled into the modality-limited student model, enabling it to acquire language-driven comprehensive knowledge. To fully harness the benefits of KD, we focus on pedestrian trajectory forecasting, leveraging three key aspects: i) Dynamic human pose exhibits a diverse yet coherent correlation with locomotion, ii) Pedestrian scenes span diverse indoor and outdoor environments, driving various agent-agent and agent-scene interactions, iii) Ego-view trajectory forecasting faces frequent occlusion, making it ideal for exploiting supplementary modalities.

Our framework is validated on three pedestrian trajectory forecasting datasets (JRDB [39], SIT [4], and ETH[47]/UCY [27]) and with two state-of-the-art trajectory forecasting models (HiVT [69] and MART [26]). To thoroughly assess the model's ability to integrate multi-modal data, we conduct holistic training and evaluation on both full and instantaneous observations. The teacher model leverages trajectory, 3D human pose, and text, while the student model inputs are restricted to trajectory or 3D human pose as the sole supplement. While text annotations on agent's behavior and inter-agent interaction are used for JRDB, we use VLM [62] to extract agent descriptions on SIT. For both datasets, a SOTA 3D pose extractor [54] is used to extract the 3D pose. We first train the teacher model with regression loss on both full observation and instantaneous input setups. The student model is then distilled upon the frozen teacher model, incorporating regression loss and additional distillation loss aligning the core semantic embedding spaces. We treat **intra**-agent and **inter**-agent embeddings as the core embeddings. Intra-agent embeddings capture coarse locomotive implications from agent-specific modalities. These are further encoded to account for inter-agent relationships, forming complete representations on future motion intent. Aligning the core latent spaces across both scales enables the student model to generalize its predictive capabilities on the teacher's extensive multi-modality. As a result, our generalized KD framework improves the distilled student model performance up to 13% on an averaged metric. In summary, our contributions are three-fold:

- We show that textual descriptions play a crucial role in blending the modalities into a comprehensive knowledge on human motion.

- For the first time, we propose a knowledge distillation framework for human trajectory forecasting where a modality-limited student model acquires the teacher's full-modal knowledge, including VLM-generated text.

- To validate our framework, we parse two datasets that include 3D human pose and text leveraging VLM.

2. Related works

2.1. Multi-modal trajectory forecasting

Leveraging multi-modal data such as text or human pose has proven essential in recent research for both pedestrian and vehicle trajectory forecasting, particularly in environments where occlusions and complex interactions are prevalent [41, 43, 50, 57]. For vehicle trajectory forecasting, recent works discovered that a language-based reasoning of motion is robust and generalizable [8, 11, 37]. In doing so, leveraging VLM for perception [41, 43, 57] or by directly interacting with a VLM/LLM to predict and plan future behavior has been studied [56, 65]. However, exploiting the robust knowledge of text on human motion has been less visited. Instead, recent works focuses on leveraging visual cues such as bounding boxes and human poses [50, 52]. To the best of our knowledge, we are first to leverage the inter-modality relationship between human pose, text, and trajectory for modeling motion intent.

2.2. Multi-modal knowledge distillation

While use of additional modalities greatly enhances the performance of models on a plethora of tasks [34, 50, 61, 68], these are often limited due to the computational limitations

or requirements of bulky, expensive sensors. Knowledge distillation serves as its solution, namely by conveying the broad knowledge of a model trained on full modalities to a student model with limited input modalities A common protocol is to construct a competent teacher model that utilizes the full range of input modalities, upon which a student model is distilled [12, 22, 23, 31, 48, 58]. Two most crucial component are defining the target knowledge to transfer and specifically tailoring a distillation method for the target modality [6, 28, 55, 67]. Building on these principles, we are first to propose a KD framework crafted to distill the comprehensive multi-modal knowledge on agent's motion intent for trajectory forecasting.

2.3. Instantaneous trajectory forecasting

trajectory forecasting in mobile systems, such as autonomous vehicles and robots, must handle unexpected agents and occlusions, relying on motion prediction from limited frames. This poses a considerably disparate, ill-posed nature compared to conventional protocol [3, 10, 42]. A key approach to addressing this challenge is incorporating past trajectories into the prediction process. The missing past trajectory is predicted in addition to its future counterpart, thereby teaching the model to implicitly leverage the past trajectories towards predicting the future [29, 30]. Another line of approach related to our work is utilizing knowledge distillation, where a teacher model trained on full past trajectory distills its full knowledge on its instantaneous counterpart [40, 64]. We take a step forward by leveraging multi-modality for knowledge distillation. Visual and textual cues greatly supplement the model's instantaneous prediction performance, and we show that such knowledge could also be effectively conveyed to the student model.

3. Method

3.1. Problem definition

trajectory forecasting is a task of learning a mapping function between the observed 2D trajectory sequence of N agents: $\mathcal{X} : \{\mathbf{x}_n^t\}_N^{-T_p:0}$, and future trajectory $\mathcal{Y} : F \times \{\mathbf{y}_n^t\}_N^{0:T_f}$, where F is the number of future trajectory proposals. We also exploit 3D human pose and text descriptions which are given in each frame, respectively notated as: $\mathcal{P} : \{\mathbf{p}_n^t\}_N^{-T_p:0}$ and $\mathcal{S} : \{\mathbf{s}_n^t\}_N^{-T_p:0}$. These modalities are collectively notated as $(\mathcal{X}, \mathcal{P}, \mathcal{S}) \in \mathcal{M}$. For JRDB dataset with text annotation on agent interaction, $\mathcal{S}$ includes s_{A} and also s_{R}, respectively for agent behavior and interaction relationship. For the remaining method section, notations ϕ, ψ denote transformer and graph networks, respectively.

3.2. Overall framework

The crux of our KD framework shown in Fig. 2 lies in two aspects: i) Constructing a competent teacher that musters a comprehensive knowledge on motion intent from three modalities: $\mathcal{X}$, $\mathcal{P}$, and $\mathcal{S}$. ii) Conveying such insight towards student models treating $\mathcal{X}$ or $\mathcal{X}+\mathcal{P}$. Additionally, we generalize the framework to be applicable to any regression-based model. In this light, we re-implement two baseline models focusing on two components: a local encoder for encoding the intra-agent modalities and a global encoder for modeling the inter-agent interaction. We distill each encoder's representation to its student counterpart, conveying insights on intra-agent intention and inter-agent interaction.

Specifically, both teacher and student models share the same design with a sole difference on the number of input modalities. Each have their respective end-to-end models. We first train the teacher model, followed by student model distilled on the frozen pre-trained teacher model. While we detail each component for KD baselines HiVT [69] and MART [26], other baselines are implemented in a similar fashion, detailed in supplementary materials.

3.2.1 Modality embedder

We use separate encoders $E_{\text{M}}^{\mathcal{X}}, E_{\text{M}}^{\mathcal{P}}, E_{\text{M}}^{\mathcal{S}}$ to encode each modality. Following conventional approaches, trajectory embedding is obtained using a MLP. Depending on the model, normalized global position and temporal difference of positions are both encoded or selectively encoded as trajectory embedding: $z_{x_n}^t = E_M^{\mathcal{X}}(\mathcal{X})$. In addition, $E_M^{\mathcal{X}}$ also encodes heading angle for trajectory-only student models to acquire context on instantaneous forecasting given 1 frame. For text, we use a pre-trained TinyBERT encoder [21] to obtain the sentences' class tokens as text embeddings, thereby introducing semantic context: $z_{s_n}^t = E_M^{\mathcal{S}}(\mathcal{S})$. For human pose, we use an MLP to encode the SMPL theta parameters into pose embedding: $z_{p_n}^t = E_M^{\mathcal{P}}(\mathcal{P})$. We use simple MLP networks for embedders to focus on the effects of cross-modality fusion via these embeddings, not the processing of obtaining the embeddings itself. As a result, we obtain modality-specific embeddings for $\mathcal{X}$, $\mathcal{P}$, and $\mathcal{S}$ as:

$$Z = \{z_{x_n}^t, z_{p_n}^t, z_{s_{A}n}^t, z_{s_{R}n}^t\} \in \mathbb{R}^{N_A \times T_p \times D} \tag{1}$$

3.2.2 Local encoder

The local encoder E_{L} fuses modalities in both the modality ($\mathcal{M}$) and temporal (T_p) dimensions, capturing a detailed representation of each agent's behavior. Since all baseline models are transformer-based, we use cross-attention to fuse modalities. MART originally processed $\mathcal{X}$ with an MLP, which we replace with a transformer module E_{L} to encode both $\mathcal{M}$ and T_p axes via holistic attention. A class token $\overline{q_n}$ is added with the modalities to acquire a representative embedding for each agent, similar to BERT [13].

$$q_n = E_{\text{L}}^{\text{MART}}(Z) = \phi_{\mathcal{M},T_p}(\overline{q_n}, z_x, z_p, z_s) \tag{2}$$

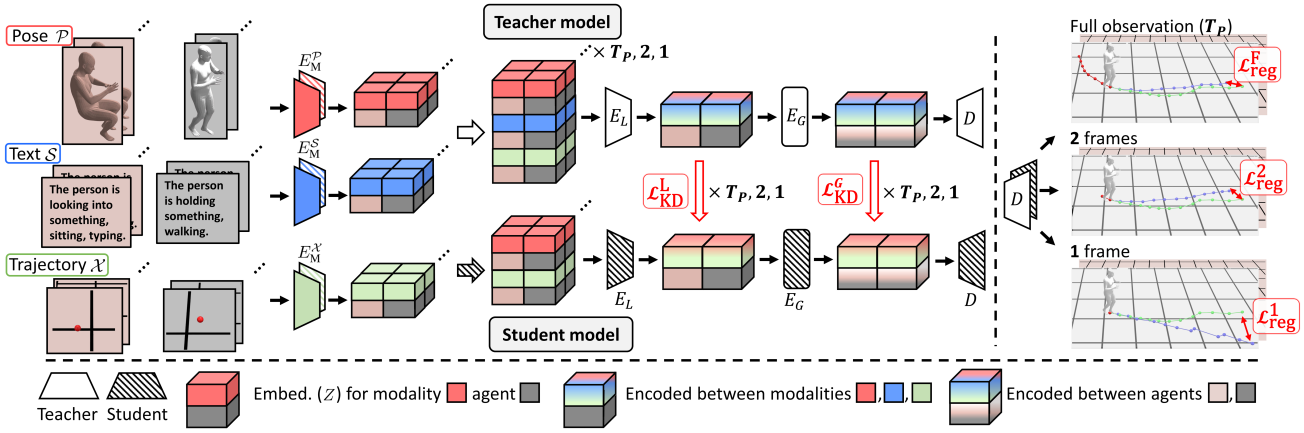

Figure 2. We first pre-train a teacher model that leverages the full range of modalities, upon which a student model with limited modalities ($\mathcal{X}+\mathcal{P}$ or $\mathcal{X}$) is distilled from scratch. Regression losses for three observation settings ($T_p, 2, 1$) are applied to both teacher and student, while additional KD losses guide the student to robustly encode intra-agent modalities (Q) and inter-agent interactions (H).

For HiVT, modality and temporal dimensions are processed sequentially to support graph-based interaction modeling per frame. Unlike standard transformer encoders that apply attention across embeddings simultaneously, HiVT's graph-based approach models relationships between each agent and its neighbors individually. Prior to encoding, HiVT rotates each neighbor's trajectory to the ego vehicle's heading vector coordinates, achieving rotation-invariant embeddings. Such design aims to encode a more thorough agent representation by adaptively handling each inter-agent relationship. We extend this design to incorporate human pose for neighbors within a threshold τ, addressing pose correlations in a rotation-invariant manner to more effectively capture motion intents from subtle interactions and nuances reflected in pose.

For each frame, agent modalities $(\mathcal{X}^t, \mathcal{P}^t, \mathcal{S}^t)_i$ and the neighbor modalities $(\mathcal{X}^t, \mathcal{P}^t)_j$ are encoded as follows:

$$\begin{aligned}\{z_x, z_p, z_s\}_i &= \mathcal{E}(\{x^t, p^t, s^t\}_i) \\ \{z_x, z_p\}_i &= \mathcal{E}(\{R_{ji}x^t, R_{ji}p^t\}_j)\end{aligned} \tag{3}$$

$$\begin{aligned}q_n^t &= E_{\mathrm{L}}^{\mathrm{HiVT}}(\mathcal{M}) \\ &= \psi_{\mathcal{M}}\left([(z_x, z_p, z_s)_i, (z_x, z_p, z_s)_j, (v_{ji})_e]\right)\end{aligned} \tag{4}$$

$\psi_{\mathcal{M}}[()_i, ()_j, ()_e]$ denotes a graph operation encoding the ego (i) feature via cross attention with neighbor (j) and edge (e) features. v_{ji} denotes a vector from j to i position at time $t=0$. R_{ji} represents a rotation matrix that maps the neighbor trajectory from global coordinates to agent's heading vector coordinates. Then, the temporal domain is encoded by a transformer as in $E_{\mathrm{L}}^{\mathrm{MART}}$. The resulting embedding $q_n \in \mathbb{R}^N \times D$ comprehensively represents the coarse motion intents inferred from multi-modal input.

3.2.3 Global encoder

The global encoder E_{G} subsequently attends to inter-agent interaction among q_n, forming complete implications on socially compliant locomotion. MART employs a group prediction-based transformer model to incorporate inter-agent interactions, a design that aligns well with the integration of additional modalities. Leveraging these modalities enhances the robustness of inter-agent relationship inferences, thereby supporting more reliable group assignments.

$$H = E_{\mathrm{G}}^{\mathrm{MART}}(Q) = \phi_N(Q) \tag{5}$$

For HiVT, we again exploit the advantages of a graph-based model by respectively leveraging text for each interaction. JRDB explicitly includes text annotations on the relationship between agents, such as "They are standing together and having conversation." These text embeddings are used to encode edge attributes along with a spatial cue of 2D vector at t=0 across agent-neighbor pairs as follows:

$$H = E_{\mathrm{G}}^{\mathrm{HiVT}}(Q) \quad = \psi_N\left([(q_n)_i, (q_n)_j, (v_{ji}, s_{R,ji})]\right) \tag{6}$$

3.2.4 Decoder

We use a standard MLP decoder D that maps the global encoder output $H \in \mathbb{R}^{N_A \times D}$ to future trajectory forecasts with multiple proposals, $\mathcal{Y} \in \mathbb{R}^{F \times T_f \times 2}$.

3.3. Training objectives

3.3.1 Regression loss

We regress each model with their corresponding default losses for both teacher and student models: L2 for MART and Negative Log Likelihood for HiVT. Compared to a conventional protocol of only training with full past observation, we also train the models with instantaneous observations (1 or 2 frames) to thoroughly assess the model's ability

in integrating multi-modality for challenging situations. All observations other than the last 2 or 1 frames are padded for the instantaneous prediction setting. These losses are notated as $L^{\text{F}}_{\text{reg}}$, L^{2}_{reg}, L^{1}_{reg}, each corresponding to loss with full, 2 frames, and 1 frame of observation. The teacher model is first trained via these regression losses.

3.3.2 Knowledge distillation loss

We distill the two core knowledges in trajectory forecasting: 1. Q, the fusion of intra-agent multi-modalities into a coarse motion intent, 2. H, the refined implications on future locomotion incorporating agent interactions. In doing so, we apply KL divergence between the teacher ($Q_{\mathcal{T}}$, $H_{\mathcal{T}}$) and student latents ($Q_{\mathcal{S}}$, $H_{\mathcal{S}}$) to align their distributions.

$$\begin{aligned}\mathcal{L}_{\text{KD}} &= \mathcal{L}^{\text{L}}_{\text{KD}} + \mathcal{L}^{\text{G}}_{\text{KD}} \\ &= \mathcal{L}_{\text{KL}}(Q_{\mathcal{T}} \| Q_{\mathcal{S}}) + \mathcal{L}_{\text{KL}}(H_{\mathcal{T}} \| H_{\mathcal{S}})\end{aligned} \tag{7}$$

For HiVT, we add an additional regularization term for training stability and use cosine similarity for KD:

$$\begin{aligned}\mathcal{L}^{\text{L}}_{\text{KD}} &= \lambda_{\cos}\mathcal{L}_{\cos}(Q_{\mathcal{T}}, Q_{\mathcal{S}}) + \mathcal{L}_{\text{KL}}(\mathcal{N} \| Q_{\mathcal{S}}) \\ \mathcal{L}^{\text{G}}_{\text{KD}} &= \lambda_{\cos}\mathcal{L}_{\cos}(H_{\mathcal{T}}, H_{\mathcal{S}}) + \mathcal{L}_{\text{KL}}(\mathcal{N} \| H_{\mathcal{S}})\end{aligned} \tag{8}$$

Again, $\mathcal{L}_{\text{KL}}$ is computed for all three types of observations: $\mathcal{L}^{\text{F}}_{\text{KD}}$, $\mathcal{L}^{2}_{\text{KD}}$, $\mathcal{L}^{1}_{\text{KD}}$, each denoting full, 2, and 1 frame of observation. The overall loss for a student model is:

$$\mathcal{L} = \lambda_{\text{reg}} L^{\text{F}}_{\text{reg}} + L^{2}_{\text{reg}} + L^{1}_{\text{reg}} + \mathcal{L}^{\text{F}}_{\text{KD}} + \mathcal{L}^{2}_{\text{KD}} + \mathcal{L}^{1}_{\text{KD}} \tag{9}$$

Where λ denotes scaling factor: 0.5 for $\lambda_{\cos}$ and 3 for λ_{reg}. These losses guide the student to predict future trajectories while aligning its core latents with the teacher. By mimicking the teacher's behavior while learning to forecast the future, the student is encouraged to operate as robustly or even surpass the teacher via improved generalization.

4. Experiment

4.1. Dataset

We test our model on three datasets: JRDB [39], SIT [4], and ETH [47]/UCY [27]. JRDB and SIT are the largest datasets for human tracking and prediction from a robot perspecitve, suitable for developing our framework with their affluent visual features. In addition, traditional ETH/UCY benchmark is also tested upon to show the versatility of our approach even on limited visual features.

For JRDB and SIT, we utilize 3D human pose and text captions as an additional modality of visual feature. We extract the SMPL theta parameters of 3D human pose using [54] for each visible agent. Unlike previous works [20, 50, 52], we use SMPL representation for human pose for improved generalizability as discussed in supplementary materials. For text captions, annotated text descriptions [19] are parsed to describe intra-agent behavior and inter-agent interaction for JRDB dataset. For SIT dataset, we use PLLaVa [62] to caption agent descriptions from consecutive frames of cropped images of agents, using prompt "What is the person doing?". As JRDB dataset includes more diverse scenes and agent behaviors, with up to 38 agents and a 2× longer duration, we treat JRDB as our primary dataset for KD and the ablation studies. Both JRDB and SIT are parsed with 2.5FPS to predict 12 frames given 8 frames, matching the ETH/UCY dataset configuration.

For ETH/UCY dataset, its BEV-view and low resolution present disparate limitations compared to JRDB and SIT datasets. While human pose has been used as visual feature for JRDB and SIT, neither 3D or 2D pose are able to be accurately extracted due to low visibility of human pose. Therefore, we resort to 2D image features of cropped images from pretrained CLIP image encoder [49] as visual features. For text captions, VLM models were incompetent in accurately captioning agent behavior for BEV-views. Instead, we implement a rule-based approach to generate text captions that describe the map context, such as "There is an obstacle in the right." Further details regarding datasets are scrutinized in the supplementary materials.

4.2. Metrics

We use the widely used metrics for trajectory forecasting, Average Displacement Error (ADE) and Final Displacement Error (FDE). For all experimentation, we predict 6 modes (F) of future trajectories to assess the model's competence in predicting a few accurate trajectories rather than over-populating similar trajectories. We also evaluate performance on instantaneous forecasting, where only the most recent one or two frames of the past trajectory are given, denoted as ADE_1 and ADE_2.

4.3. Baselines

We choose the following works as baselines and additionally perform KD on underlined models. We precisely modify each model to have similar number of parameters. Further details are found in the supplementary materials.

HiVT [69] is a graph-based network designed with attention-based computation. We choose HiVT for KD due to its SOTA performance and flexibility in incorporating additional modalities across both ego and interaction dimensions, enabled by its graph-based architecture.

MART [26] is a transformer-based network and is one of most recent SOTA model for pedestrian trajectory forecasting. We choose MART for KD to demonstrate generalizing our KD framework on any transformer-based models, even ones not specifically designed to handle multi-modality.

ST [50] (socialTransmotion) is a transformer-based model for human trajectory forecasting designed to handle addi-

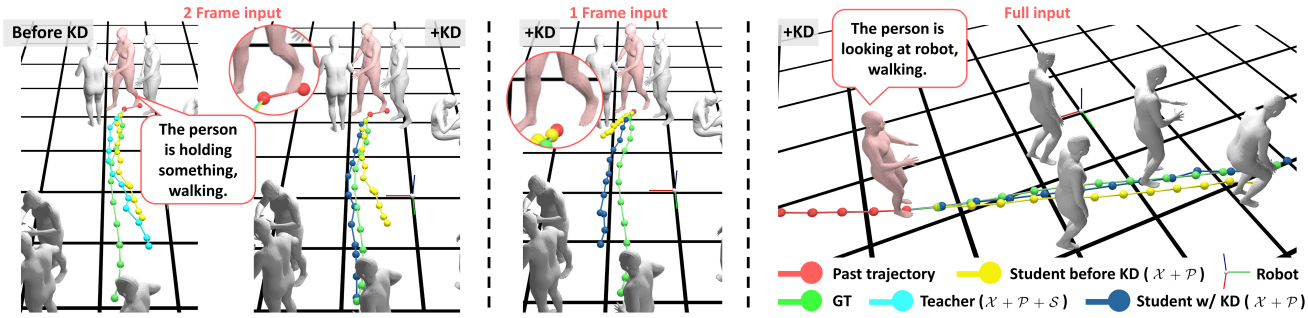

Figure 3. Qualitative results on JRDB with $\mathcal{X}+\mathcal{P}$ HiVT model. The model outperforms its baseline counterpart with KD. Improved accuracy is demonstrated on all instantaneous and full observations. Origin denotes robot position, and bubbles represent text annotations.

Table 1. Prediction results on JRDB with multimodal inputs: human pose $\mathcal{P}$ and text $\mathcal{S}$. **Bold** denotes best. Lower is better.

$\mathcal{M}$	HiVT							MART						
	ADE	ADE_2	ADE_1	FDE	FDE_2	FDE_1	Ave. +%[1]	ADE	ADE_2	ADE_1	FDE	FDE_2	FDE_1	Ave. +%
$\mathcal{X}$	0.221	0.240	0.342	0.432	0.467	0.632	-	0.286	0.275	0.395	0.545	0.526	0.753	-
$\mathcal{X}+\mathcal{P}$	0.229	0.242	0.364	0.441	0.465	0.659	-2.84	0.287	0.282	0.366	0.543	0.538	0.682	+2.02
$\mathcal{X}+\mathcal{S}$	0.224	0.230	0.285	0.432	0.444	0.540	+6.53	0.261	0.265	0.301	0.496	0.506	0.564	+12.41
$\mathcal{X}+\mathcal{P}+\mathcal{S}$	**0.220**	**0.227**	**0.280**	**0.423**	**0.435**	**0.521**	**+8.38**	**0.258**	**0.256**	**0.289**	**0.491**	**0.486**	**0.538**	**+14.98**

Table 2. KD on ETH/UCY (average) with HiVT student model using $\mathcal{X}$. % measures improvement with KD. **Bold** denotes best. Grey background highlights results with our KD framework.

Model	w/ KD	ADE	ADE_2	ADE_1	FDE	FDE_2	FDE_1	Ave. + %
LED	-	0.312	0.334	0.600	1.008	1.017	1.427	-
ST	-	0.319	0.319	0.605	0.853	0.855	1.224	-
MART		0.364	0.369	0.571	0.852	0.862	1.166	**+3.80**
	✓	0.347	0.353	0.553	0.807	0.817	1.141	
HiVT		0.314	0.320	0.555	0.718	**0.726**	1.132	+1.55
	✓	**0.304**	**0.317**	**0.546**	**0.698**	0.727	**1.122**	

tional human modalities such as 2D, 3D pose and bounding boxes. This aligns with the motivations of our framework and makes it a suitable candidate for comparison.

LED [38] is a transformer-based diffusion model achieving SOTA performance on pedestrian trajectory forecasting, chosen to compare our work on various types of learning-based models including diffusion models.

5. Results

5.1. Quantitative results

5.1.1 Multi-modal teacher model

As with any KD framework, we start off by constructing a competent teacher model via leveraging diverse modalities. Table 1 demonstrates the performance improvement with the use of additional modalities. Both HiVT and MART gradually improves over all metrics with more modality supplements, achieving up to 14.98% improvement in performance. Such significant improvement demonstrates that $\mathcal{P}$ and $\mathcal{S}$ greatly aids understanding motion intent and interpreting the corresponding future motion.

[1]Average of % improvement for each metric. Same for all Ave. +% notations in the remainder of the paper.

Specifically, $\mathcal{S}$ plays a key role in building robust knowledge. For both models, the addition of $\mathcal{P}$ results in limited improvement, primarily due to the noise introduced during pose extraction. In contrast, all metrics considerably improve with $\mathcal{S}$, reflecting the strong correlation between human motion and language. Leveraging this correlation fully exploits the potential of human pose as shown by further improvement from $\mathcal{X}+\mathcal{P}$ to $\mathcal{X}+\mathcal{P}+\mathcal{S}$, even when $\mathcal{X}$ to $\mathcal{X}+\mathcal{P}$ has shown degradation for HiVT. This shows that language bridges the domain gap between trajectory and noisy human pose, forming a more comprehensive understanding of motion intent. Use of modalities is particularly beneficial on instantaneous settings, as additional modalities supplement the missing context of instantaneous trajectory, improving up to 27% for ADE_1 on MART from $\mathcal{X}$ to $\mathcal{X}+\mathcal{P}+\mathcal{S}$.

5.1.2 Multi-modal knowledge distillation

Full modality $\mathcal{X}+\mathcal{P}+\mathcal{S}$ model is used as a teacher for all datasets. Table 2 and 3 show the improvements made with KD. For ego-view datasets JRDB and SIT, we choose Student model using $\mathcal{X}$ and $\mathcal{X}+\mathcal{P}$. A consistent improvement over most metrics on all settings is observed, improving up to 13% on SIT dataset. Notably, the performance of both modality type students improves with KD, even when only $\mathcal{X}$ is used as input modality. This improvement shows the unlocking of the potential of numerical trajectories to grasp the semantic context-based motion intent from the knowledge of other modalities.

For JRDB dataset which contains considerable amount of training data, HiVT's vanilla $\mathcal{X}$ model performs considerably well with full T_F input. In such case, the advantages of KD is spotlighted on instantaneous prediction where the

Table 3. KD results on JRDB and SIT dataset. Full modality $\mathcal{X}+\mathcal{P}+\mathcal{S}$ model is chosen as teacher model. KD is performed on student models using modality $\mathcal{X}$ and $\mathcal{X}+\mathcal{P}$. % measures improvement with KD. **Bold** represents best performance.

Modality	Model	w/ KD	JRDB dataset							SIT dataset						
			ADE	ADE_2	ADE_1	FDE	FDE_2	FDE_1	Avg. +%	ADE	ADE_2	ADE_1	FDE	FDE_2	FDE_1	Avg. +%
$\mathcal{X}$	LED		0.358	0.426	0.447	0.760	0.897	0.934	-	0.479	0.507	0.546	1.045	1.057	1.093	-
	ST		0.324	0.337	0.427	0.614	0.636	0.780	-	0.531	0.544	0.699	1.037	1.065	1.310	-
	MART		0.286	0.275	0.395	0.545	0.526	0.753	**+7.61**	0.468	0.483	0.556	0.909	0.937	1.073	+3.25
		✓	0.259	0.262	0.366	0.495	0.496	0.684		0.441	0.465	0.570	0.832	0.885	1.090	
	HiVT		0.221	0.240	0.342	**0.432**	0.467	0.632	+2.38	0.455	0.457	0.526	0.869	0.884	0.988	**+6.37**
		✓	**0.220**	**0.232**	**0.326**	0.438	**0.453**	**0.604**		**0.404**	**0.427**	**0.503**	**0.799**	**0.833**	**0.965**	
$\mathcal{X}+\mathcal{P}$	LED		0.340	0.358	0.446	0.715	0.822	0.949	-	0.469	0.491	0.548	1.051	1.045	1.085	-
	ST		0.345	0.362	0.437	0.637	0.663	0.782	-	0.541	0.576	0.751	1.046	1.119	1.410	-
	MART		0.287	0.282	0.366	0.543	0.538	0.682	**+6.63**	0.473	0.484	0.574	0.909	0.937	1.119	+5.19
		✓	0.266	0.261	0.337	0.519	0.507	0.633		0.448	0.465	0.550	0.849	0.883	1.059	
	HIVT		**0.229**	0.242	0.364	**0.441**	0.465	0.659	+4.98	0.518	0.519	0.531	0.979	0.979	1.006	**+13.03**
		✓	0.232	**0.239**	**0.308**	0.445	**0.464**	**0.560**		**0.414**	**0.444**	**0.500**	**0.789**	**0.853**	**0.951**	

improvement averages around 5∼10%. $\mathcal{X}+\mathcal{P}$ for HiVT also shows a similar tendency, showing significant improvement on instantaneous prediction. Repeated improvement on $\mathcal{X}+\mathcal{P}$ again showcases the maximal exploitation of human pose, guided by multi-modal knowledge.

Compared to the JRDB dataset, SIT dataset is a smaller dataset with less number of agents and shorter total duration. For such scarce setting, the base model suffered to establish a strong understanding between $\mathcal{X}$ and $\mathcal{P}$ and resulted in inferior performance even with $\mathcal{P}$ as additional modality. The strength of KD is particularly evident in this setting, showing the greatest improvement of 13%. consistent improvement with KD on $\mathcal{X}+\mathcal{P}$ confirms the crucial role of text in establishing a firm understanding on the correlation between human pose and locomotion intent via KD.

Our KD framework's competence is also manifested on ETH/UCY dataset as shown in Tab. 2. As mentioned in experiment section, cropped 2D image feature is used in place of human pose and text describes the nearby obstacles. Since validation is performed on a different scene, use of image features does not inherently guarantee any generalizability. On the other hand, text contains generalizable map context information in unified format such as "There is an obstacle on the right of the person." This provides hints on reducing the agent motion intent's degree of freedom. The teacher model learns to leverage these hints for forecasting, especially improving their instantaneous prediction performance. Upon KD, $\mathcal{X}$ student models were able to learn such knowledge and improved between 1∼ 4%. Consistent improvement across all ego, BEV-view datasets demonstrate the generalizability of our multi-modal KD framework.

5.2. Qualitative results

Figure 3 shows predictions made with both instantaneous and full observations on JRDB dataset. The leftmost two figures visualize predictions before and after KD. Surprisingly, the student model outperforms the teacher model after KD on prediction with 2 frame input. This shows that the student not only mimics the teacher model, but generalizes itself to the task based on the comprehensive knowledge of the teacher. In the extreme of a single frame input, the student before KD collides into a nearby agent. After KD, however, the model accurately predicts a collision-free trajectory. Notably, the model successfully infers the direction of avoidance, demonstrating a solid understanding of pose orientation after KD. Overall, student model shows distinct improvements across all input settings with KD. More qualitative results are found in the supplementary materials.

Table 4. Ablation on KD components, namely E_L and E_G outputs. Models use $\mathcal{X}+\mathcal{P}$ modalities. Experimented on JRDB.

$\mathcal{L}_{KD}^{L}$	$\mathcal{L}_{KD}^{G}$	ADE	ADE_2	ADE_1	FDE	FDE_2	FDE_1	Avg. +%
HiVT								
		0.229	0.242	0.364	0.441	0.465	0.659	-
	✓	0.228	0.244	0.352	0.447	0.472	0.647	+0.26
✓		0.228	0.241	0.327	0.445	0.468	0.601	+3.09
✓	✓	**0.232**	**0.239**	**0.308**	**0.445**	**0.464**	**0.560**	**+4.98**
MART								
		0.287	0.282	0.366	0.543	0.538	0.682	-
	✓	0.274	0.272	0.380	0.520	0.514	0.711	+1.49
✓		0.275	0.271	0.366	0.529	0.520	0.683	+2.33
✓	✓	**0.266**	**0.261**	**0.337**	**0.519**	**0.507**	**0.633**	**+6.63**

5.3. Ablation studies

Herein, we delve into deeper questions regarding the dual foundation of our KD framework: 5.1.1, 5.1.2. Further ablations are found in the supplementary materials.

How do we ensure an effective knowledge distillation? The common core of various KD frameworks lies in aligning the latent space across different modalities. Latent spaces sharing a comparable semantic meaning is precisely aligned with latent-specific losses [6, 31, 48]. For example, L2 loss is used to align BEV feature space where teacher and student features share a geometric coordinate, while KL divergence loss is used for distributional latents [6]. In addition to adhering to such philosophy, we attempt to keep our framework generalizable. In that sense, we have chosen the core distillation features by their general context: intra-agent modality fusion Q and inter-agent interaction H.

Table 4 studies the effects of each $\mathcal{L}_{KD}^{L}$ and $\mathcal{L}_{KD}^{G}$ on overall performance gain. First, $\mathcal{L}_{KD}^{L}$ guides the student model

to learn and encode the comprehensive knowledge of $Q_{\mathcal{T}}$ on agent's motion intent with its limited modality. $\mathcal{L}_{\text{KD}}^{\text{G}}$ then further reinforces the student model by inducing the model to effectively encode inter-agent interactions on the heterogenous latents acquired from $Q_{\mathcal{T}}$. For both HiVT and MART, $\mathcal{L}_{\text{KD}}^{\text{L}}$ is shown to be most important as it aligns the most fundamental latents that represent agents' motion intent. Additional use of $\mathcal{L}_{\text{KD}}^{\text{G}}$ corroborates the distillation by guiding the model to learn incorporating interaction among these intents. As a result, incorporating both $\mathcal{L}_{\text{KD}}^{\text{L}}$ and $\mathcal{L}_{\text{KD}}^{\text{G}}$ results in greatest improvement.

Does language assist in modeling human interactions? Human locomotion depends on two major factors: intra-agent motion intent and inter-agent interaction. While recent works attempt to leverage agent-descriptive multi-modality in modeling motion intent [50, 52], no works have explored utilizing explicit interaction cues other than global positional vectors.

To the best of our knowledge, we are the first to investigate the impact of incorporating explicit text as supplementary guidance in modeling agent interaction. Table 5 demonstrates the improvements with the use of interaction text caption $\mathcal{S}_{\text{R}}$ on various models. The most significant improvement is observed with the $\mathcal{X}$ model, which operates with least modalities. In this setting where data is scarce for inferring agent motion intent, the interaction text compensates for missing information by conveying implicit action intent cues embedded within the behavioral descriptions. Other models also consistently improve with $\mathcal{S}_{\text{R}}$, with the performance gain decreasing as the number of modalities increases. For KD, building a $\mathcal{X}+\mathcal{P}+\mathcal{S}$ model with the text-based guidance improved the teacher model by 0.9%, followed by 0.51% gain in its corresponding student. Overall, we observe the broad impact of interaction text from building a competent teacher to distilling its knowledge.

What kind of language description is most helpful? From previous sections, we discovered that language plays the most important role in constructing a comprehensive teacher model. Then, what kind of text caption helps the most? To answer this question, we use VLM to generate captions based on diverse prompts shown in Tab. 6. We generate text captions for each visible agent in the text annotation-free SIT dataset. Each prompt is selected to leverage different abilities of VLM. Prompt 1 is the same format as JRDB annotation, describing the past behavior. Prompt 2 incorporates the scene context by specifying any surrounding obstacle as text captions for ETH/UCY. Prompt 3 takes a predictive stance on human motion by captioning the VLM's prediction on future human motion.

Evaluation on $\mathcal{X}+\mathcal{S}$ teacher model shows that incorporating the scene context yields the best performance. Indeed, hints about the presence of an obstacle offer insight into the agent's future trajectory direction, whereas descriptive captions primarily focus on the agent's dynamic or static nature. Comparing between prompt 1 and prompt 3, the $\mathcal{X}+\mathcal{S}$ teacher model shows comparable performance, showing that VLM's prediction of agent behavior accurately reflects the corresponding past motion. With KD, however, knowledge from prompt 1 resulted in most superior performance due to the straightforward correlation between past pose and text describing the past motion. For prompt 2, map context was unable to be inferred upon KD since there is no distinctive correlation between past pose and map context. While KD with prompt 3 shows nearly comparable result compared to prompt 1, a direct correlation between pose and text as in prompt 1 yields greatest improvement upon KD between these latents.

Table 5. Ablation study on interaction text $\mathcal{S}_{\text{R}}$ using the HiVT model under various modality protocols. Experimented on JRDB. $\mathcal{X}$ column indicates the additional modalities used alongside $\mathcal{X}$.

$+\mathcal{S}_{\text{R}}$	w/ KD	$\mathcal{X}$	ADE	ADE_2	ADE_1	FDE	FDE_2	FDE_1	Avg. +%
		-	0.221	0.240	0.342	**0.432**	0.467	0.632	**+4.57**
✓		-	**0.218**	**0.228**	**0.316**	**0.432**	**0.445**	**0.576**	
		$+\mathcal{P}$	0.229	0.242	0.364	0.441	0.465	0.659	+3.59
✓		$+\mathcal{P}$	**0.226**	**0.237**	**0.334**	**0.436**	**0.458**	**0.614**	
		$+\mathcal{S}_{\text{A}}$	0.233	0.238	0.275	0.450	0.458	0.516	+0.98
✓		$+\mathcal{S}_{\text{A}}$	**0.224**	**0.230**	0.285	**0.432**	**0.444**	0.540	
		$+\mathcal{P}$ $+\mathcal{S}_{A}$	0.225	0.231	**0.278**	0.431	0.442	**0.515**	+0.90
✓		$+\mathcal{P}$ $+\mathcal{S}_{A}$	**0.220**	**0.227**	0.280	**0.423**	**0.435**	0.521	
	✓	$+\mathcal{P}$	**0.227**	**0.238**	0.317	**0.444**	**0.461**	0.580	+0.51
✓	✓	$+\mathcal{P}$	0.232	0.239	**0.308**	0.445	0.464	**0.560**	

Table 6. Ablation on VLM prompts for acquiring text captions on SIT dataset. HiVT $\mathcal{X}+\mathcal{P}$ is distilled from $\mathcal{X}+\mathcal{S}$ for each.

	ADE	ADE_2	ADE_1	FDE	FDE_2	FDE_1
Prompt 1	What is the person doing?					
$\mathcal{X}+\mathcal{S}$	0.521	0.525	0.533	0.978	0.984	1.005
$\mathcal{X}+\mathcal{P}+$ KD	**0.414**	**0.444**	**0.500**	**0.789**	**0.853**	**0.951**
Prompt 2	Is there any obstacle in front of the person?					
$\mathcal{X}+\mathcal{S}$	**0.506**	**0.508**	0.531	**0.950**	**0.959**	**0.991**
$\mathcal{X}+\mathcal{P}+$ KD	0.418	0.444	**0.500**	0.795	0.854	0.960
Prompt 3	What will the person do in the future?					
$\mathcal{X}+\mathcal{S}$	0.520	0.522	**0.530**	0.989	0.990	1.003
$\mathcal{X}+\mathcal{P}+$ KD	**0.414**	0.448	0.535	0.796	0.865	1.032

6. Conclusion

In this work, we present a multi-modal knowledge distillation framework for human trajectory forecasting. First, we demonstrate that incorporating additional modalities such as 3D pose and text significantly enhances forecasting performance. However, as expensive modality such as text is not readily available in most applications, we additional perform KD on models with limited modalities. Student models with only trajectory input or pose as sole additional modality is distilled upon a teacher trained with full range of modalities. Student models consistently improve in both instantaneous and full observation metrics on all ego and BEV-view datasets, confirming the generalized competence of our KD framework.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF2022R1A2B5B03002636).

References

[1] Sk Arif Ahmed, Debi Prosad Dogra, Samarjit Kar, and Partha Pratim Roy. Trajectory-based surveillance analysis: A survey. *IEEE transactions on circuits and systems for video technology*, 29(7):1985–1997, 2018. 1

[2] Inhwan Bae, Junoh Lee, and Hae-Gon Jeon. Can language beat numerical regression? language-based multimodal trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 753–766, 2024. 2

[3] Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17890–17901, 2024. 3

[4] Jong Wook Bae, Jungho Kim, Junyong Yun, Changwon Kang, Jeongseon Choi, Chanhyeok Kim, Junho Lee, Jungwook Choi, and Jun Won Choi. Sit dataset: socially interactive pedestrian trajectory dataset for social navigation robots. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 5

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[6] Yujeong Chae, Hyeonseong Kim, Changgyoon Oh, Minseok Kim, and Kuk-Jin Yoon. Lidar-based all-weather 3d object detection via prompting and distilling 4d radar. In *Computer Vision – ECCV 2024*, pages 368–385, Cham, 2025. Springer Nature Switzerland. 3, 7

[7] Kaibing Chen, Dong Shen, Hanwen Zhong, Huasong Zhong, Kui Xia, Di Xu, Wei Yuan, Yifei Hu, Bin Wen, Tianke Zhang, et al. Evlm: An efficient vision-language model for visual understanding. *arXiv preprint arXiv:2407.14177*, 2024. 2

[8] Yuan Chen, Zi-han Ding, Ziqin Wang, Yan Wang, Lijun Zhang, and Si Liu. Asynchronous large language model enhanced planner for autonomous driving. In *European Conference on Computer Vision*, pages 22–38. Springer, 2025. 2

[9] Jie Cheng, Xiaodong Mei, and Ming Liu. Forecast-mae: Self-supervised pre-training for motion forecasting with masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8679–8689, 2023. 1

[10] Pranav Singh Chib and Pravendra Singh. Pedestrian trajectory prediction with missing data: Datasets, imputation, and benchmarking. *arXiv preprint arXiv:2411.00174*, 2024. 1, 3

[11] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. *IEEE Intelligent Transportation Systems Magazine*, 2024. 2

[12] Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*, 2022. 3

[13] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[14] Yang Gao, Po-Chien Luan, and Alexandre Alahi. Multi-transmotion: Pre-trained model for human motion prediction. *arXiv preprint arXiv:2411.02673*, 2024. 1

[15] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17113–17122, 2022. 1

[16] Xunjiang Gu, Guanyu Song, Igor Gilitschenski, Marco Pavone, and Boris Ivanovic. Producing and leveraging online map uncertainty in trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14521–14530, 2024. 1

[17] Je-Seok Ham, Dae Hoe Kim, NamKyo Jung, and Jinyoung Moon. Cipf: Crossing intention prediction network based on feature fusion modules for improving pedestrian safety. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3665–3674, 2023. 1

[18] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[19] Simindokht Jahangard, Zhixi Cai, Shiki Wen, and Hamid Rezatofighi. Jrdb-social: A multifaceted robotic dataset for understanding of context and dynamics of human interactions within social groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22087–22097, 2024. 5

[20] Jaewoo Jeong, Daehee Park, and Kuk-Jin Yoon. Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1617–1628, 2024. 1, 5

[21] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019. 3

[22] Sanmin Kim, Youngseok Kim, Sihwan Hwang, Hyeonjun Jeong, and Dongsuk Kum. Labeldistill: Label-guided cross-modal knowledge distillation for camera-based 3d object detection. In *European Conference on Computer Vision*, pages 19–37. Springer, 2025. 3

[23] Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman Narayanan, Senthil Yogamani, and Fatih Porikli. X3kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13343–13353, 2023. 3

[24] Zhiqian Lan, Yuxuan Jiang, Yao Mu, Chen Chen, and Shengbo Eben Li. Sept: Towards efficient scene representation learning for motion prediction. In *The Twelfth International Conference on Learning Representations*, 2023. 1

[25] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. 2

[26] Seongju Lee, Junseok Lee, Yeonguk Yu, Taeri Kim, and Kyoobin Lee. Mart: Multiscale relational transformer networks for multi-agent trajectory prediction. *arXiv preprint arXiv:2407.21635*, 2024. 2, 3, 5

[27] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, pages 655–664. Wiley Online Library, 2007. 2, 5

[28] Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12458–12468, 2024. 3

[29] Rongqing Li, Changsheng Li, Dongchun Ren, Guangyi Chen, Ye Yuan, and Guoren Wang. Bcdiff: Bidirectional consistent diffusion for instantaneous trajectory prediction. *Advances in Neural Information Processing Systems*, 36: 14400–14413, 2023. 3

[30] Rongqing Li, Changsheng Li, Yuhang Li, Hanjie Li, Yi Chen, Ye Yuan, and Guoren Wang. Itpnet: Towards instantaneous trajectory prediction for autonomous driving. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1643–1654, 2024. 3

[31] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640, 2023. 3, 7

[32] Haozhe Lin, Chunyu Wei, Li He, Yuchen Guo, Yunqi Zhao, Shanglong Li, and Lu Fang. Gigatraj: Predicting long-term trajectories of hundreds of pedestrians in gigapixel complex scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19331–19340, 2024. 1

[33] Xiaotong Lin, Tianming Liang, Jianhuang Lai, and Jian-Fang Hu. Progressive pretext task learning for human trajectory prediction. *arXiv preprint arXiv:2407.11588*, 2024. 1

[34] Xun Lin, Shuai Wang, Rizhao Cai, Yizhong Liu, Ying Fu, Wenzhong Tang, Zitong Yu, and Alex Kot. Suppress and rebalance: Towards generalized multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 211–221, 2024. 2

[35] Haochen Liu, Li Chen, Yu Qiao, Chen Lv, and Hongyang Li. Reasoning multi-agent behavioral topology for interactive autonomous driving. *arXiv preprint arXiv:2409.18031*, 2024. 1

[36] Mengmeng Liu, Hao Cheng, Lin Chen, Hellward Broszio, Jiangtao Li, Runjiang Zhao, Monika Sester, and Michael Ying Yang. Laformer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2039–2049, 2024. 1

[37] Yunsheng Ma, Can Cui, Xu Cao, Wenqian Ye, Peiran Liu, Juanwu Lu, Amr Abdelraouf, Rohit Gupta, Kyungtae Han, Aniket Bera, et al. Lampilot: An open benchmark dataset for autonomous driving with language model programs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15141–15151, 2024. 2

[38] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5517–5526, 2023. 6

[39] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6748–6765, 2021. 1, 2, 5

[40] Alessio Monti, Angelo Porrello, Simone Calderara, Pasquale Coscia, Lamberto Ballan, and Rita Cucchiara. How many observations are enough? knowledge distillation for trajectory forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6553–6562, 2022. 3

[41] Seokha Moon, Hyun Woo, Hongbeen Park, Haeji Jung, Reza Mahjourian, Hyung-gun Chi, Hyerin Lim, Sangpil Kim, and Jinkyu Kim. Visiontrap: Vision-augmented trajectory prediction guided by textual descriptions. *arXiv preprint arXiv:2407.12345*, 2024. 2

[42] Achintya Nath, Paritosh Kabra, Ishu Gupta, Pravendra Singh, et al. Ms-tip: Imputation aware pedestrian trajectory prediction. In *Forty-first International Conference on Machine Learning*, 2024. 3

[43] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024. 1, 2

[44] Daehee Park, Hobin Ryu, Yunseo Yang, Jegyeong Cho, Jiwon Kim, and Kuk-Jin Yoon. Leveraging future relationship reasoning for vehicle trajectory prediction. *arXiv preprint arXiv:2305.14715*, 2023. 1

[45] Daehee Park, Jaeseok Jeong, Sung-Hoon Yoon, Jaewoo Jeong, and Kuk-Jin Yoon. T4p: Test-time training of trajectory prediction via masked autoencoder and actor-specific token memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15065–15076, 2024. 1

[46] Armin Danesh Pazho, Ghazal Alinezhad Noghre, Vinit Katariya, and Hamed Tabkhi. Vt-former: An exploratory study on vehicle trajectory prediction for highway surveil-

lance through graph isomorphism and transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5651–5662, 2024. 1

[47] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*, pages 261–268. IEEE, 2009. 2, 5

[48] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5213–5224, 2023. 3, 7

[49] 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. 5

[50] Saeed Saadatnejad, Yang Gao, Kaouther Messaoud, and Alexandre Alahi. Social-transmotion: Promptable human trajectory prediction. *arXiv preprint arXiv:2312.16168*, 2023. 1, 2, 5, 8

[51] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer, 2020. 1

[52] Tim Salzmann, Lewis Chiang, Markus Ryll, Dorsa Sadigh, Carolina Parada, and Alex Bewley. Robots that can see: Leveraging human pose for trajectory prediction. *IEEE Robotics and Automation Letters*, 8(11):7090–7097, 2023. 1, 2, 5, 8

[53] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1

[54] 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. 2, 5

[55] Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. Egodistill: Egocentric head motion distillation for efficient video understanding. *Advances in Neural Information Processing Systems*, 36:33485–33498, 2023. 3

[56] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 2

[57] Jhih-Ciang Wu Wang, Hong-Han Shuai, and Wen-Huang Cheng. Trajprompt: Aligning color trajectory with vision-language representations. 2

[58] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1828–1838, 2020. 3

[59] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1410–1420, 2023. 1

[60] Guowei Xu, Jiale Tao, Wen Li, and Lixin Duan. Learning semantic latent directions for accurate and controllable human motion prediction. In *European Conference on Computer Vision*, pages 56–73. Springer, 2025. 1

[61] Haoran Xu, Peixi Peng, Guang Tan, Yuan Li, Xinhai Xu, and Yonghong Tian. Dmr: Decomposed multi-modality representations for frames and events fusion in visual reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26508–26518, 2024. 2

[62] 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. 2, 5

[63] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *European Conference on Computer Vision*, pages 511–528. Springer, 2022. 1

[64] Yi Xu and Yun Fu. Adapting to length shift: Flexilength network for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15226–15237, 2024. 1, 3

[65] Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024. 2

[66] Haichao Zhang, Yi Xu, Hongsheng Lu, Takayuki Shimizu, and Yun Fu. Oostraj: Out-of-sight trajectory prediction with vision-positioning denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14802–14811, 2024. 1

[67] Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, and Jungong Han. Efficient rgb-t tracking via cross-modality distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5404–5413, 2023. 3

[68] Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5116–5125, 2023. 2

[69] 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 Con-*

ference on Computer Vision and Pattern Recognition, pages 8823–8833, 2022. 2, 3, 5

[70] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17863–17873, 2023. 1