A FOUNDATION MODEL-POWERED DUAL TRANSFORMER FRAMEWORK FOR CENTRALIZED MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent systems require explicit agent-wise feature representations and structured interaction modeling, while pretrained vision-language models primarily produce global group-wise multimodal feature representations whose token structure and feature dimensions do not naturally align with individual agents. To bridge this gap, we propose a unified foundation model-based dual-transformer framework **FMDT** for centralized multi-agent system. The framework combines a agent-wise state transformer encoder with a multimodal VLM encoder through cross-attention. The agent-centric features produced by the state encoder transformer serve as queries over global multimodal VLM features, enabling agent-wise reasoning while retaining rich scene-level context. We instantiate this framework as **TrajVLM** for centralized multi-agent long-horizon trajectory forecasting and **CMAVLA** for centralized multi-agent single-step decision making. TrajVLM is optimized through supervised fine-tuning followed by reinforcement learning-based refinement, whereas CMAVLA is trained through centralized multi-agent reinforcement learning for cooperative control. Beyond the algorithmic framework, we provide theoretical guarantees by deriving covering number and Rademacher complexity bounds for the proposed dual-transformer framework hypothesis class, which further yield uniform convergence and sample complexity. Experiments on ETH-UCY for trajectory forecasting and MaMuJoCo for cooperative multi-agent control demonstrate that the proposed framework achieves competitive or state-of-the-art performance against current representative baselines. These results suggest that FMDT provides a general mechanism for adapting foundation models to tackle centralized multi-agent learning problems. More details can be found in the project website: `https://sites.google.com/view/cmas-fmdt`.

ἀλλ᾽ ἔστι τι τὸ ὅλον παρὰ τὰ μόρια.
The whole is something beside the parts.

Aristotle, *Metaphysics*, VIII.6, 1045a8–10

1 INTRODUCTION

Recent vision-language foundation models have demonstrated strong capabilities in embodied intelligence, enabling agents to learn generalizable behaviors from large-scale multimodal data Kim et al. (2024); Li et al. (2025a). Benefiting from large-scale dataset and pretraining, these vision-language action models (VLAs) have demonstrated increasingly strong generalization across tasks, environments, and embodiments, establishing a general mechanism for connecting multimodal perception and semantic understanding with embodied behaviors Black et al. (2025b;a); Chen et al. (2026).

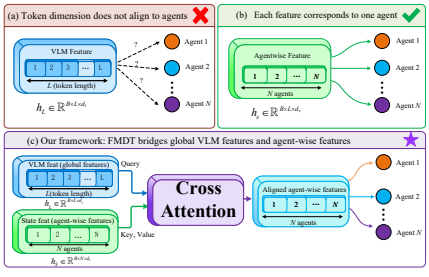


Figure 1: An illustration of FMDT.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Large language models (LLMs) have also been adapted into multi-agent systems Kannan et al. (2024); Zhao et al. (2024); Zhang et al. (2024); Liu et al. (2025a); Chen et al. (2025a). However, these approaches typically rely on textual language generation, resulting in planning process sensitive to generation errors, while such pipeline is generally not optimized end-to-end through rewards. Moreover, the substantial inference latency from the inference procedure of foudation model, limiting their applicability to real-time embodied control Li et al. (2026b). In contrast, recent reinforcement learning (RL)-based VLA methods demonstrate that VLA policies can be directly improved through reward feedback Lu et al. (2026); Li et al. (2026a); Zang et al. (2026a); Amin et al. (2026). Nevertheless, extending this paradigm to centralized multi-agent systems remains largely underexplored. Emerging systems have tired to incorporate VLAs into multi-robot learning Zang et al. (2026b); Doshi et al. (2026), but a inherent representation gap remains as shown in Fig. 1, where VLMs produce global multimodal token features $h_L \in \mathbb{R}^{B\times L\times d_l}$, whereas multi-agent decision making requires explicit agent-wise representations $h_S \in \mathbb{R}^{B\times N\times d_s}$. Since the VLM token dimension L has no explicit correspondence to the agent dimension N, directly adapting global VLM representations to centralized multi-agent policies does not naturally preserve agent identity.

To address aforementioned core challenge, we introduce a dual-transformer Vaswani et al. (2017) architecture in which a agent-wise expert transformer network extracts agent-centric features as queries, while the VLM provides global multimodal features as keys and values. Cross attention Lu et al. (2019) then aligns and fuses above two representations, allowing each agent to selectively retrieve relevant global multimodal context while preserving explicit agent-wise structure.

The main contributions of this paper are summarized as follows:

- We introduce a unified foundation model-based dual-transformer framework (FMDT) for centralized multi-agent systems, instantiated as **TrajVLM** for multi-step trajectory forecasting and **CMAVLA** for centralized single-step decision making. To bridge global VLM feature and multi-agent representations, agent-wise features serve as queries over VLM features through cross-attention.
- We develop task-specific adaptation fine-tuning strategies for the proposed framework. TrajVLM is optimized through supervised fine-tuning (SFT) followed by RL-based refinement, while CMAVLA is trained through a centralized MARL procedure for cooperative decision making.
- We establish theoretical guarantees for the proposed dual-transformer framework by deriving its covering number and Rademacher complexity, and further obtain uniform convergence and generalization bounds for centralized multi-agent learning.
- Extensive experiments demonstrate that the proposed framework achieves state-of-the-art or competitive performance against strong baselines across ETH-UCY Lerner et al. (2007) multi-agent trajectory prediction datasets, and MAMuJoCo Peng et al. (2021) MARL cooperative control benchmarks.

2 Literature Review

Early trajectory forecasting evolved from classical and graph-based methods, involving Social Force Helbing &Molnar (1995), Social-LSTM Alahi et al. (2016), Social-STGCN Mohamed et al. (2020), and Social-Implicit Mohamed et al. (2022), to attention and transformer-based models such as Social-Attention Vemula et al. (2018), EvolveGraph Li et al. (2020), Trajectron++ Salzmann et al. (2020), STAR Yu et al. (2020), and AgentFormer Yuan et al. (2021). To better represent and capture higher-order crowd interactions, current works such as GroupNet Xu et al. (2022), MART Lee et al. (2024), and Hyper-STTN Wang et al. (2026) construct crowd dynamics with hypergraph structures. Generative methods, typically including Social-GAN Gupta et al. (2018), MID Gu et al. (2022), LED Mao et al. (2023), and MoFlow Fu et al. (2025), have also shown promising potential for trajectory prediction. Apart from that, foundation model-based approaches Bae et al. (2024; 2025); Xu et al. (2026) have been introduced into multi-agent trajectory prediction tasks as well, which are fine-tuned by supervised fine-tuning (SFT) or RL fine-tuning paradigm. However, these methods still forecast future trajectories by parsing generated language-token sequences into numerical coordinates. This paradigm differs from recent RL-VLA approaches Amin et al. (2026); Zang et al.

(2026a), which leverage action expert networks to directly decode VLM features into output actions, and may introduce additional inference latency due to sequential language-token generation.

Multi-agent reinforcement learning (MARL) has been widely studied for multi-agent system decision making. Representative approaches include MADDPG Lowe et al. (2017), QMIX Rashid et al. (2020), MAPPO Yu et al. (2022), HASAC Liu et al. (2024), Sable Mahjoub et al. (2025), as well as diffusion-based methods such as MADiff Zhu et al. (2024) and DIMA Zhang et al. (2026). More recently, VLA-based approaches have begun to be explored for multi-agent system, such as RLinf-userZang et al. (2026b) and CHORUS Doshi et al. (2026). Compared with existing VLA-based approaches, CMAVLA not only integrates multimodal reasoning into cooperative learning, but also reduces computational and device overhead by sharing a single centralized VLM across all agents.

Statistical learning theory studies model generalization from finite samples, with uniform convergence Hanneke et al. (2023; 2024); Devale et al. (2026) characterizing the gap between empirical and population risks over a hypothesis class. Existing studies have analyzed transformer architectures from different theoretical perspectives, including the Lipschitz continuity of attention Kim et al. (2021), generalization benefits of multimodal fusion Huang et al. (2021), norm-based and sequence length independent generalization bounds Edelman et al. (2022); Trauger &Tewari (2024), and Rademacher complexity of next token prediction transformers Li et al. (2025b). More recent work further characterizes the VC dimension and sample complexity of transformers Yang et al. (2026). Herein, we leverage existing theoretical techniques to characterize the covering number and sample complexity of the general dual-transformer framework FMDT for multi-agent system.

3 METHODOLOGY

3.1 VLMS-BASED DUAL-TRANSFORMER FRAMEWORK

As shown in Fig. 2, we design a unified VLMs foundation model powered centralized multi-agent learning framework. The objective is to bridge a learnable mapping from multimodal system observations input to multi-agent system short-horizon or long-horizon decision-making. Recent advances in VLMs or VLAs foundation models have demonstrated strong potential for robotics and embodied intelligence. Representative approaches such as $\pi_{0.5}$ Black et al. (2025a) and RDT-1B Liu et al. (2025b) show that large-scale multimodal pretraining can improve generalization across tasks and embodiments, while recent VLA-RL studies further demonstrate that RL can refine pretrained multimodal policies through interaction Liu et al. (2026); Amin et al. (2026); Zang et al. (2026a). These advances motivate the use of foundation models in robotics to directly learn and reason about robotic behaviors from vision-language multimodal observations.

As discussed above, directly adapting VLM token representations to centralized multi-agent learning requires an explicit mechanism to align multimodal context with agent-wise representations. To address this, we propose the FMDT framework, which couples a state transformer encoder to capture agent-wise features with a foundation model for multimodal context. Specifically, the state encoder produces agent-indexed embeddings h_S, which are used as queries to retrieve relevant information from the VLM token features h_L through cross-attention. In this way, FMDT preserves the explicit agent dimension while incorporating global information, providing a general interface for foundation model-based centralized multi-agent learning. We instantiate FMDT into two methods who **TrajVLM** for centralized multi-agent long-horizon trajectory forecasting and **CMAVLA** for centralized multi-agent single-step decision making.

The FMDT framework of centralized multi-agent learning can be formulated as the mapping $\{F_\theta : (X_S, X_L) \mapsto \hat{Y}\}$, where X_L denotes the multimodal input to the foundation model, involving visual observations and language contexts, X_S denotes the structured multi-agent state input, and Y represents the task-dependent agent-indexed output. The mapping is realized by $F_\theta = D \circ C \circ (T_S, T_L)$, where the agent-wise features $h_S = T_S(X_S) \in \mathbb{R}^{B\times N\times d_s}$ and the global VLM features $h_L = T_L(X_L) \in \mathbb{R}^{B\times L\times d_l}$ are captured by is state transformer encoder T_S and large VLM encoder T_L, respectively. D corresponds to the decoder network, while C performs cross-attention fusion by using the agent-wise features as queries and the VLM features as keys and values as follows:

$$Q = h_S W_Q, \qquad K = h_L W_K, \qquad V = h_L W_V, \qquad h_C = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_c}}\right) V. \tag{1}$$

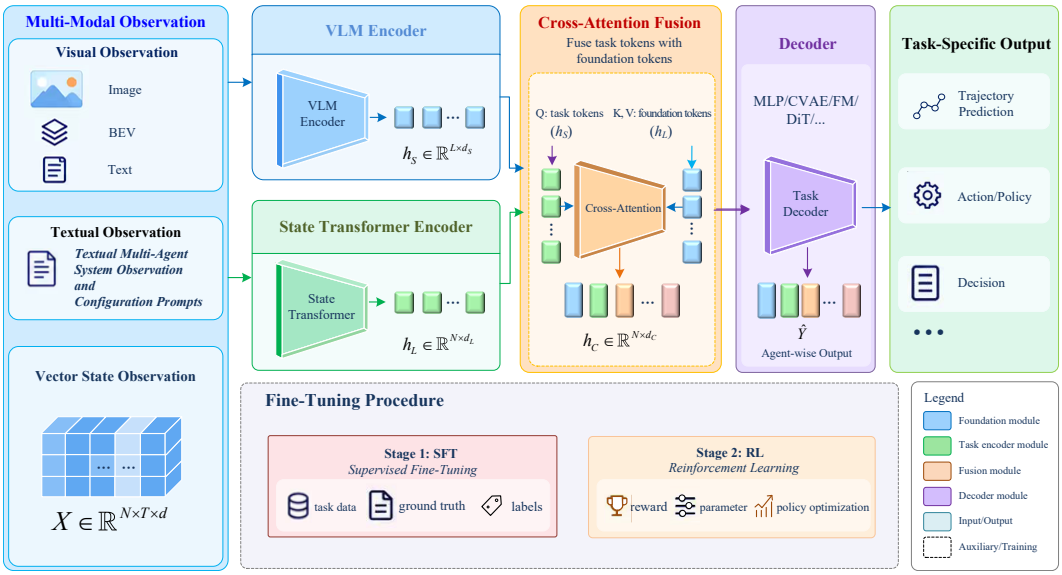


Figure 2: The architecture of FMDT.

3.2 FMDT for Centralized Multi-Agent System Long-Horizon Forecasting

TrajVLM instantiates the proposed FMDT framework for centralized multi-agent long-horizon forecasting $\hat{Y} = TrajVLM(X_S, X_L)$, where $\hat{Y} \in \mathbb{R}^{N\times T_{out}\times 2}$ and $X_S \in \mathbb{R}^{N\times T_{in}\times 2}$. Given the historical multimodal observations of all agents $\{X_S, X_L\}$, a state transformer encoder is adapted from existing trajectory forecaster Wang et al. (2026), while a foundation model Gemma4-12B Team et al. (2026) extracts global semantic representations from the image and text observations of scene $X_L = \{X_L^{img}, X_L^{txt}\}$. We design a conversion from state observation X_S to the multimodal image and text observation X_L, as shown in Appendix. A. The agent-wise features h_S then query the global VLM features h_L through cross-attention in Equation. 1, producing a fused representation h_C that preserves explicit agent identities while incorporating global multimodal context. Conditioned on h_C, a trajectory decoder models the joint future evolution of all agents over a long prediction horizon, where flow matching decoder Lipman et al. (2023) is employed to capture the conditional multimodal distribution of future trajectories. In this way, TrajVLM provides a centralized multi-agent forecasting framework that jointly leverages structured multi-agent interactions and foundation model knowledge for long-horizon trajectory prediction.

We remove the hypergraph encoder from Hyper-STTN Wang et al. (2026) and adopt its spatial-temporal transformer as the state encoder of TrajVLM. Spatial attention models inter-agent interactions at each step, while temporal attention captures each agent's motion dynamics across history. The attention mask excludes invalid or padded entries, preserving valid structures as follows:

$$\mathrm{Attn}(Q,K,V;M) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}} + M\right)V;\ i,j \in dim_2(QK^\top),\ M_{ij} = \begin{cases} 0, & \text{if } ij \text{ valid,} \\ -\infty, & \text{otherwise.} \end{cases} \tag{2}$$

After cross-attention, the fused multi-agent representation is denoted by h_C. We employ flow matching Lipman et al. (2023); Fu et al. (2025) as the TrajVLM decoder. Given a ground-truth future trajectory $Y_1 \sim q(Y_1 \mid h_C)$ from the distribution q, we define a conditional probability path $p_t(Y \mid Y_1, h_C)$, with the corresponding marginal path $p_t(Y \mid h_C)$ and vector field $u_t(Y \mid h_C)$:

$$\begin{aligned} p_t(Y \mid h_C) &= \int p_t(Y \mid Y_1, h_C)\, q(Y_1 \mid h_C)\, dY_1, \\ u_t(Y \mid h_C) &= \int u_t(Y \mid Y_1, h_C) \frac{p_t(Y \mid Y_1, h_C)\, q(Y_1 \mid h_C)}{p_t(Y \mid h_C)}\, dY_1. \end{aligned} \tag{3}$$

The conditional vector field is learned via the flow matching objective:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,Y_1,Y_t}\left[\left\| v_{\theta_{fm}}(Y_t, t, h_C) - u_t(Y_t \mid Y_1, h_C)\right\|_2^2\right], \tag{4}$$

where $v_{\theta_{fm}}$ is the learned vector field conditioned on h_C.

After SFT stage, we further optimize TrajVLM via a RL fine-tuning procedure using group relative policy optimization (GRPO) Shao et al. (2024). GRPO avoids an additional critic network by sampling total K future trajectories and using their relative prediction quality within each group to estimate the advantage, with the ADE and FDE-based Alahi et al. (2016) i-th agent's k-th sample's reward function $R_{i,k} = -\left(w_{\text{ADE}}\text{ADE}_{i,k} + w_{\text{FDE}}\text{FDE}_{i,k}\right)$. Where w_{ADE} and w_{FDE} control the relative importance of the two prediction objectives. The reward mean, standard deviation, and advantage of GRPO are defined as follows:

$$\mu_i = \frac{1}{K}\sum_{k=1}^{K} R_{i,k}; \quad \sigma_i = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(R_{i,k}-\mu_i\right)^2}; \quad A_{i,k} = \frac{R_{i,k}-\mu_i}{\sigma_i+\epsilon_A}. \tag{5}$$

where ϵ_A is a numerical constant. Moreover, for each sampled future agent position, the importance ratio between two updated policies is defined as $r_{i,k,t}(\theta) = \frac{\pi_\theta\left(\hat{Y}_{i,k,t}|X\right)}{\pi_{\theta_{\text{old}}}\left(\hat{Y}_{i,k,t}|X\right)}$ for the RL loss:

$$\mathcal{L}_{\text{RL}} = -\mathbb{E}_{i,k,t}\left[\min\left(r_{i,k,t}(\theta)A_{i,k}, \text{clip}\left(r_{i,k,t}(\theta), 1-\epsilon, 1+\epsilon\right)A_{i,k}\right)\right] + \beta D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right). \tag{6}$$

where ϵ controls the maximum policy update magnitude, β controls the strength of the KL regularization., and π_{ref} denotes the SFT fine-tuned policy.

3.3 FMDT for Centralized Multi-Agent System Single Step Decision-Making

We propose CMAVLA, a centralized multi-agent VLA framework for cooperative multi-agent control Peng et al. (2021). CMAVLA adopts a centralized-training and centralized-execution (CTCE) paradigm and formulates the multi-agent control problem as a MDP tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, N\rangle$, where $\mathcal{S}$ denotes the multi-agent system state space, $\mathcal{A} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$ is the joint action space of N agents, $\mathcal{P}$ denotes the transition dynamics, $\mathcal{R}$ is the shared team reward, and γ is the discount factor. The global multi-agent observations are jointly encoded by a centralized foundation model, while each agent is controlled by individual action expert conditioned on the centralized representation $\mathcal{C}(\mathcal{T}_S(\hat{s}_t), \mathcal{T}_L(\hat{s}_t^L)) = [h_t^1, ..., h_t^N]$. The joint policy is optimized using centralized multi-agent group relative policy optimization (CMAGRPO) Shao et al. (2024). For each agent i, the individual action expert is defined as $\pi_\theta^i(a_t^i \mid h_t^i) = \mathcal{N}\left(\mu_\theta^i(h_t^i), (\sigma_\theta^i)^2\right)$. The action is sampled from the corresponding policy as $a_t^i \sim \pi_\theta^i(\cdot \mid h_t^i)$. The joint policy of CMAVLA is factorized over all agent-wise action experts as $\hat{\pi}_\theta(\hat{a}_t \mid \hat{s}_t) = \prod_{i=1}^N \pi_\theta^i(a_t^i \mid h_t^i)$. The g-th group k-th sample t-th timestep return is $G_{g,k,t} = \sum_{\tau=t}^{T-1}\gamma^{\tau-t}\mathcal{R}_{g,k,\tau}$, and the reward mean, deviation, and advantage adhere to the equation. 5. Additionally, we employ an action entropy term as $\mathcal{H}\left(\pi_\theta^i\right) = \frac{1}{2}\sum_{d=1}^{D_i^{act}}\log\left(2\pi e\left(\sigma_{i,d}\right)^2\right)$ with the action dimension D^{act}. Eventually, the loss function is defined as follows:

$$\mathcal{L}_{MARL} = -\mathbb{E}_{g,k,t}\left[\min\left(r_{g,k,t}(\theta)A_{g,k,t}, \text{clip}\left(r_{g,k,t}(\theta), 1\pm\epsilon\right)A_{g,k,t}\right)\right] - \alpha\mathcal{H}\left(\hat{\pi}_\theta\right) + \beta D_{\text{KL}}\left(\hat{\pi}_\theta \| \hat{\pi}_{\text{ref}}\right). \tag{7}$$

4 Theoretical Results

Theorem 1 (Covering Number): For each trainable parameter block W_r, let W_r^0 be a fixed reference parameter block and define the nonempty admissible set $\Theta_r = \left\{W_r : \|W_r - W_r^0\|_{1,\text{entry}} \le B_r,\ \|W_r\|_{2\to 2} \le S_r\right\}$, where $0 \le B_r < \infty$ bounds the parameter perturbation around W_r^0, and $0 \le S_r < \infty$ bounds the spectral norm. Here, $\|\cdot\|_{1,\text{entry}}$ denotes the entrywise ℓ_1 norm, $\|\cdot\|_{2\to 2}$ denotes the spectral norm, and $\|\cdot\|_F$ denotes the Frobenius norm. Specifically, for a matrix W, $\|W\|_{1,\text{entry}} = \sum_{a,b}|W_{ab}|, \|W\|_{2\to 2} = \sup_{\|v\|_2=1}\|Wv\|_2, \|W\|_F = \left(\sum_{a,b}|W_{ab}|^2\right)^{1/2}$. Let $\mathcal{I}$ index all trainable parameter blocks, and let $\Theta = \prod_{r\in\mathcal{I}}\Theta_r$ denote the parameter space of the FMDT framework. For every covering radius $\tau_r > 0$, there exists a finite τ_r-net $\widehat{\Theta}_r(\tau_r) \subseteq \Theta_r$ under the Frobenius norm: for every $W_r \in \Theta_r$, there is a representative $\widehat{W}_r \in \widehat{\Theta}_r(\tau_r)$ such that $\|W_r - \widehat{W}_r\|_F \le \tau_r$. Applying Maurey's sparsification lemma Zhang (2002), following the matrix-vectorization argument of Trauger &Tewari (2024), and using nonexpansive Euclidean projection onto Θ_r to preserve admissibility, the net can be chosen to satisfy:

$$\log\mathcal{N}\left(\tau_r, \Theta_r, \|\cdot\|_F\right) \le \log\left|\widehat{\Theta}_r(\tau_r)\right| \le \frac{q_r}{\tau_r^2}, \qquad q_r = CB_r^2\log(2k_r+1), \tag{8}$$

where $\mathcal{N}(\tau_r, \Theta_r, \|\cdot\|_F)$ denotes the minimum cardinality of a τ_r-net contained in Θ_r, k_r is the number of scalar entries in W_r, $C > 0$ is a universal constant, and $\log$ denotes the natural logarithm.

For each component $j \in \{S, L, C, D\}$, define $\Theta_j = \prod_{r\in\mathcal{I}_j} \Theta_r$. Let M_{j,θ_j} denote the component mapping with parameters $\theta_j \in \Theta_j$, and let $\mathcal{F}_j = \{M_{j,\theta_j} : \theta_j \in \Theta_j\}$ be the corresponding function class. Define the uniform distance $d_j(M, M') = \sup_{u\in\mathcal{U}_j} \|M(u) - M'(u)\|_F$, where $\mathcal{U}_j$ is a fixed admissible input domain containing all inputs generated by admissible upstream components. For each $r \in \mathcal{I}_j$, assume that there exists a finite uniform parameter sensitivity bound $\Gamma_{j,r} \geq 0$ such that $d_j(M_{j,\theta_j}, M_{j,\theta'_j}) \leq \Gamma_{j,r}\|W_r - W'_r\|_F$, for all admissible parameter configurations θ_j and θ'_j that differ only in block r. By replacing the parameter blocks one at a time and applying the triangle inequality, block-wise approximations satisfying $\|W_r - \widehat{W}_r\|_F \leq \tau_r$ yield $d_j(M_{j,\theta_j}, M_{j,\widehat{\theta}_j}) \leq \sum_{r\in\mathcal{I}_j} \Gamma_{j,r}\tau_r$. For any component-level covering radius $\eta_j > 0$, choose the block-wise covering radii to satisfy $\sum_{r\in\mathcal{I}_j} \Gamma_{j,r}\tau_r \leq \eta_j$. The product of the parameter block covers then induces an η_j-cover of $\mathcal{F}_j$, giving $\log\mathcal{N}(\eta_j, \mathcal{F}_j, d_j) \leq \sum_{r\in\mathcal{I}_j} \log\mathcal{N}(\tau_r, \Theta_r, \|\cdot\|_F) \leq \sum_{r\in\mathcal{I}_j} \frac{q_r}{\tau_r^2}$. Minimizing this upper bound over the covering radii subject to the above error constraint yields:

$$\log\mathcal{N}(\eta_j, \mathcal{F}_j, d_j) \leq \frac{A_j}{\eta_j^2}, \qquad A_j = \left[\sum_{r\in\mathcal{I}_j} \left(\Gamma_{j,r}^2 q_r\right)^{1/3}\right]^3. \tag{9}$$

We define the hypothesis class of the proposed FMDT framework as $\mathcal{F}_{\mathrm{FMDT}} = \{F_\theta = D \circ C \circ (T_S, T_L) : T_S \in \mathcal{T}_S, T_L \in \mathcal{T}_L, C \in \mathcal{C}, D \in \mathcal{D}, \theta \in \Theta\}$, where $\mathcal{T}_S$ and $\mathcal{T}_L$ denote the function classes of the expert transformer encoder and the foundation model transformer, respectively, $\mathcal{C}$ denotes the cross-attention fusion function class, and $\mathcal{D}$ denotes the decoder function class.

Assume that the cross-attention fusion module $C \in \mathcal{C}$ is (β_S, β_L)-Lipschitz with respect to its two input representations, such that $\|C(h_S, h_L) - C(h'_S, h'_L)\|_F \leq \beta_S\|h_S - h'_S\|_F + \beta_L\|h_L - h'_L\|_F$. Given η_S- and η_L-cover representatives $\widehat{h}_S$ and $\widehat{h}_L$ satisfying $\|h_S - \widehat{h}_S\|_F \leq \eta_S$ and $\|h_L - \widehat{h}_L\|_F \leq \eta_L$, together with an η_C-cover representative $\widehat{C} \in \mathcal{C}$ of the fusion module, the resulting fusion perturbation satisfies $\|C(h_S, h_L) - \widehat{C}(\widehat{h}_S, \widehat{h}_L)\|_F \leq \beta_S\eta_S + \beta_L\eta_L + \eta_C$. Further assume that each decoder $D \in \mathcal{D}$ is L_D-Lipschitz with respect to its input representation, such that $\|D(h) - D(h')\|_F \leq L_D\|h - h'\|_F$. Combining this property with an η_D-cover representative $\widehat{D} \in \mathcal{D}$ of the decoder class, and defining $\widehat{F} = \widehat{D} \circ \widehat{C} \circ (\widehat{T}_S, \widehat{T}_L)$, the approximation error of the entire FMDT framework satisfies $\|F_\theta - \widehat{F}\|_F \leq L_D\beta_S\eta_S + L_D\beta_L\eta_L + L_D\eta_C + \eta_D \leq \epsilon$. Therefore, combining the component-wise covers yields $\log\mathcal{N}(\epsilon, \mathcal{F}_{\mathrm{FMDT}}, d_F) \leq A_S/\eta_S^2 + A_L/\eta_L^2 + A_C/\eta_C^2 + A_D/\eta_D^2$, subject to $L_D\beta_S\eta_S + L_D\beta_L\eta_L + L_D\eta_C + \eta_D \leq \epsilon$. Optimizing the allocation of $\{\eta_S, \eta_L, \eta_C, \eta_D\}$ under this constraint gives the final covering-number bound

$$\log\mathcal{N}\left(\epsilon, \mathcal{F}_{\mathrm{FMDT}}, d_F\right) \leq \frac{A_F}{\epsilon^2};\ A_F = \left[\left(L_D^2\beta_S^2 A_S\right)^{\frac{1}{3}} + \left(L_D^2\beta_L^2 A_L\right)^{\frac{1}{3}} + \left(L_D^2 A_C\right)^{\frac{1}{3}} + A_D^{\frac{1}{3}}\right]^3. \tag{10}$$

Theorem 2 (Rademacher Complexity): Let ℓ be an L_ℓ-Lipschitz loss function with $0 \leq \ell(\hat{Y}, Y) \leq b_\ell$, and let $\ell \circ \mathcal{F}_{\mathrm{FMDT}}$ denote the corresponding loss-composed hypothesis class. By above Theorem, $\log\mathcal{N}(\epsilon, \mathcal{F}_{\mathrm{FMDT}}, d_F) \leq A_F/\epsilon^2$. The Lipschitz property of ℓ implies $\log\mathcal{N}(\epsilon, \ell \circ \mathcal{F}_{\mathrm{FMDT}}, d_\ell) \leq L_\ell^2 A_F/\epsilon^2$. Applying the truncated Dudley entropy integral Dudley (1967) and optimizing the truncation scale yields with a universal constant $C_R > 0$:

$$\widehat{\mathrm{Rad}}_{S_n}\left(\ell \circ \mathcal{F}_{\mathrm{FMDT}}\right) \leq C_R L_\ell \sqrt{\frac{A_F}{n}} \left[1 + \log_+\left(\frac{b_\ell\sqrt{n}}{L_\ell\sqrt{A_F}}\right)\right]. \tag{11}$$

Theorem 3 (Uniform Convergence): Let ℓ be an L_ℓ-Lipschitz loss satisfying $0 \leq \ell \leq b_\ell$. For any $F \in \mathcal{F}_{\mathrm{FMDT}}$, define the population error as $\mathrm{er}_{\mathcal{P}}(F) = \mathbb{E}_{z\sim\mathcal{P}}[\ell(F; z)]$ and the empirical error as $\widehat{\mathrm{er}}_{S_n}(F) = \frac{1}{n}\sum_{i=1}^n \ell(F; z_i)$. Then, with probability at least $1 - \delta$ over an i.i.d. sample S_n, the following uniform convergence bound holds:

$$\sup_{F\in\mathcal{F}_{\mathrm{FMDT}}} |\mathrm{er}_{\mathcal{P}}(F) - \widehat{\mathrm{er}}_{S_n}(F)| \leq 2C_R L_\ell \sqrt{\frac{A_F}{n}} \left[1 + \log_+\left(\frac{b_\ell\sqrt{n}}{L_\ell\sqrt{A_F}}\right)\right] + b_\ell\sqrt{\frac{\log(\frac{4}{\delta})}{2n}}. \tag{12}$$

Theorem 4 (Sample Complexity): Under the conditions of Theorem 3, for any target generalization error $\varepsilon > 0$ and confidence level $\delta \in (0, 1)$, there exists a universal constant $C_{\mathrm{sc}} > 0$ such that the sample complexity for $\sup_{F \in \mathcal{F}_{\mathrm{FMDT}}} |\mathrm{er}_{\mathcal{P}}(F) - \widehat{\mathrm{er}}_{S_n}(F)| \leq \varepsilon$ with probability at least $1 - \delta$ is

$$\mathcal{M}_{UC}^{\mathcal{F}_{FMDT}}(\varepsilon, \delta) \leq \frac{C_{\mathrm{sc}}}{\varepsilon^2} \left[L_\ell^2 A_F \left(1 + \log_+ \frac{b_\ell}{\varepsilon} \right)^2 + b_\ell^2 \log \left(\frac{1}{\delta} \right) \right]. \tag{13}$$

5 Experiments

5.1 Experimental Setup

Tasks and Benchmarks: The experiments of FMDT are conducted on two representative centralized multi-agent learning tasks, as multi-agent trajectory forecasting and cooperative decision making. For trajectory prediction, we evaluate TrajVLM via the five sub-datasets ETH, Hotel, UNIV, Zara1, and Zara2 from ETH-UCY dataset Lerner et al. (2007), where each sample contains 8 observed time steps positions $X_S \in \mathbb{R}^{N \times 8 \times 2}$ and 12 future time steps trajectories $Y \in \mathbb{R}^{N \times 12 \times 2}$. The dataset pre-processing protocol of TrajVLM adheres to Fu et al. (2025); Bae et al. (2023). Additionally, the pedestrian mask matrix is constructed following Wang et al. (2026) to exclude invalid or padded agents for state transformer encoder component. For multi-agent decision making, we directly implement the original MaMuJoCo formulation Peng et al. (2021); Liu et al. (2024), including state, action, and reward definitions, without additional pre-processing. The main tested simulators involve Ant-2×4, Walker2d-2×3, HalfCheetah-6×1, Hopper-3×1, and Swimmer-2×1. These environments emphasize cooperative locomotion learning among multiple agent nodes, requiring coordinated control across different body structures and motion dynamics.

Implementation Details: We instantiate the VLM encoder $T_L \in \mathcal{T}_L$ as Gemma-4-12B Team et al. (2026) for both TrajVLM and CMAVLA frameworks. For the state encoder $T_S \in \mathcal{T}_S$, TrajVLM adopts the spatial-temporal transformer from Wang et al. (2026) with the hypergraph module removed, whereas CMAVLA employs a lightweight transformer encoder Vaswani et al. (2017). Both frameworks share the same cross-attention layer $C \in \mathcal{C}$ to fuse the state features and VLM representations. For the decoder $D \in \mathcal{D}$, TrajVLM employs a flow matching network Lipman et al. (2023); Fu et al. (2025), while CMAVLA leverages an MLP expert head.

Training Protocol: TrajVLM is trained in two stages, consisting of SFT procedure from equation. 4 followed by GRPO refinement from equation. 6. CMAVLA is trained directly through centralized MAGRPO from equation. 7 to optimize multi-agent cooperative decision making. In the training procedure, we first freeze the pretrained VLM and optimize the remaining components at beginning stage until the training performance becomes stable. We then unfreeze and fine-tune the VLM jointly with the entire framework. For CMAVLA, the state transformer encoder is initialized from the pretrained Trans-CMAGRPO ablation model checkpoints, while its decoder parameters are not transferred. After this initialization, CMAVLA follows the same staged training strategy, first optimizing with the VLM frozen and subsequently fine-tuning the VLM together with the full policy.

Evaluation Metrics: For TrajVLM, we evaluate trajectory prediction performance using average displacement error ADE_{20} and final displacement error FDE_{20} Alahi et al. (2016), which measures the average Euclidean distance between the predicted and ground-truth trajectories over all future time steps or at the final prediction step, among 20 stochastic trajectory samples. For CMAVLA, we evaluate cooperative control performance using the sum of average episodic reward from all agents.

Baselines and Ablation Study: For trajectory prediction, all baseline results are directly taken from the corresponding original papers under their reported evaluation protocols. For multi-agent control, only DIMA results are taken from its original paper Zhang et al. (2026), while all other MARL baselines are reproduced and evaluated by us from [1] under a total 1M-environmental step protocol. The ablation models Trans-CMAGRPO and MLP-CMAGRPO remove the VLM and cross-attention modules from CMAVLA, retaining only the transformer-based or MLP-based state encoder and decoder. Lastly, all experiments are conducted via eight NVIDIA H100 GPUs.

Notably, CMAGRPO samples multiple parallel trajectories from the same policy at each update to estimate its advantages. Since these trajectories are collected before a single policy update, they are

[1] https://github.com/PKU-MARL/HARL

Table 1: Quantitative comparison of trajectory prediction performance on the ETH-UCY benchmark. ADE_{20} / FDE_{20} are reported, with lower values indicating better performance. Bold values indicate the best results within the two categories.

Model	**Venue**	**Dataset** ($\mathrm{ADE}_{20}\downarrow$ / $\mathrm{FDE}_{20}\downarrow$)					
		ETH	**HOTEL**	**UNIV**	**ZARA1**	**ZARA2**	**AVG**
Non-Foundation Model-based Frameworks							
Social-LSTM Alahi et al. (2016)	CVPR 2016	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
Social-GAN Gupta et al. (2018)	CVPR 2018	0.87/1.62	0.67/1.37	0.76/1.52	0.35/0.68	0.42/0.84	0.61/1.21
Social-Attention Vemula et al. (2018)	ICRA 2018	1.39/2.39	2.51/2.91	1.25/2.54	1.01/2.17	0.88/1.75	1.41/2.35
Social-STGCNN Mohamed et al. (2020)	CVPR 2020	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75
Trajectron++ Salzmann et al. (2020)	ECCV 2020	0.61/1.03	0.20/0.28	0.30/0.55	0.24/0.41	0.18/0.32	0.31/0.52
STAR Yu et al. (2020)	ECCV 2020	0.36/0.65	0.17/0.36	0.31/0.62	0.26/0.55	0.22/0.46	0.26/0.53
AgentFormer Yuan et al. (2021)	ICCV 2021	0.46/0.80	0.14/0.22	0.25/0.45	0.18/0.30	0.14/0.24	0.23/0.40
Y-Net Mangalam et al. (2021)	ICCV 2021	0.28/0.33	**0.10**/0.14	0.24/0.41	0.17/0.27	0.13/0.22	0.18/0.27
GroupNet Xu et al. (2022)	CVPR 2022	0.46/0.73	0.15/0.25	0.26/0.49	0.21/0.39	0.17/0.33	0.25/0.44
LED Mao et al. (2023)	CVPR 2023	0.39/0.58	0.11/0.17	0.26/0.43	0.18/0.26	0.13/0.22	0.21/0.33
MART Lee et al. (2024)	ECCV 2024	0.35/0.47	0.14/0.22	0.25/0.45	0.17/0.29	0.13/0.22	0.21/0.33
MoFlow Fu et al. (2025)	CVPR 2025	0.40/0.57	0.11/0.17	0.23/0.39	0.15/0.26	0.12/0.22	0.20/0.32
SocialMOIF Chen et al. (2025b)	CVPR 2025	**0.26/0.31**	**0.10/0.12**	**0.11/0.17**	**0.10/0.17**	**0.09/0.14**	**0.13/0.18**
Hyper-STTN Wang et al. (2026)	ICRA 2026	0.35/0.59	0.14/0.19	0.23/0.42	0.18/0.29	0.15/0.23	0.21/0.34
Foundation Model-based Frameworks							
LMTraj-ZERO Bae et al. (2024)	CVPR 2024	0.80/1.64	0.20/0.37	0.37/0.77	0.33/0.66	0.24/0.50	0.39/0.79
LMTraj-SUP Bae et al. (2024)	CVPR 2024	0.41/0.51	0.12/0.16	0.22/0.34	0.20/0.32	0.17/0.27	0.22/0.32
VLMTraj-ZERO Bae et al. (2025)	TPAMI 2025	0.74/1.57	0.17/0.30	0.39/0.84	0.35/0.75	0.25/0.56	0.38/0.80
VLMTraj-SUP Bae et al. (2025)	TPAMI 2025	**0.24/0.37**	0.12/**0.15**	0.22/0.34	0.18/0.30	**0.13/0.22**	**0.18/0.27**
W2W Xu et al. (2026)	CVPR 2026	0.35/0.41	0.12/**0.15**	**0.20/0.32**	0.19/0.29	0.17/0.26	0.21/0.29
TrajVLM (Ours)	Under Review	0.41/0.62	**0.11**/0.17	0.22/0.38	**0.15/0.28**	**0.13/0.22**	0.20/0.33

Table 2: **Comparison of evaluation returns across MaMuJoCo benchmarks.** We report the episode return at 1M timesteps. For CTDE MARL methods, results are reported as the mean $\pm$ standard deviation over four distinct random seeds, whereas for centralized MARL methods, we report the maximum return obtained from a single seed.

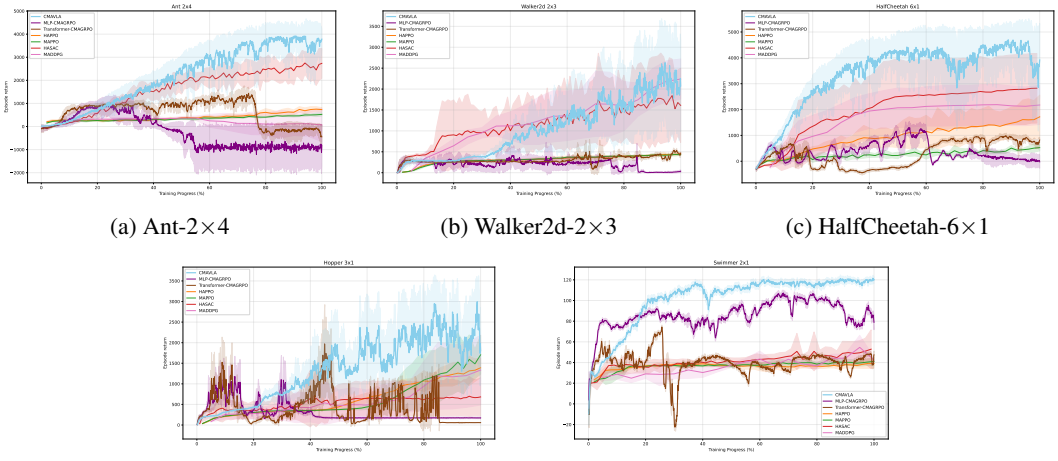
Methods	Venue	Steps	Updates	MaMuJoCo Tasks				
				Ant-2 × 4	Walker2d-2 × 3	HalfCheetah-6 × 1	Hopper-3 × 1	Swimmer-2 × 1
CTDE MARL								
MADDPG Lowe et al. (2017)	NeurIPS 2017	1M	2000	$-183.5_{\pm 673.9}$	$2314.3_{\pm 594.8}$	$2284.2_{\pm 714.5}$	$1784.0_{\pm 1236.1}$	$38.2_{\pm 2.6}$
MAPPO Yu et al. (2022)	NeurIPS 2022	1M	250	$831.2_{\pm 70.4}$	$914.9_{\pm 538.8}$	$851.9_{\pm 307.2}$	$1651.2_{\pm 522.7}$	$42.2_{\pm 2.3}$
HAPPO Kuba et al. (2022)	ICLR 2022	1M	250	$2211.7_{\pm 376.0}$	$824.1_{\pm 405.1}$	$2575.4_{\pm 1288.2}$	$1956.5_{\pm 538.4}$	$38.6_{\pm 3.6}$
HASAC Liu et al. (2024)	ICLR 2024	1M	1000	$3012.2_{\pm 493.0}$	$1830.6_{\pm 812.3}$	$3149.7_{\pm 1277.3}$	$706.5_{\pm 426.7}$	$53.1_{\pm 11.8}$
DIMA Zhang et al. (2026)	NeurIPS 2026	1M	–	**4881**$_{\pm 756}$	**3329**$_{\pm 1056}$	**5643**$_{\pm 163}$	–	–
Centralized MARL								
Trans-CMAGRPO	Under Review	1M Nominal	1000	1679.1	582.1	2302.9	1089.4	77.9
MLP-CMAGRPO	Under Review	1M Nominal	1000	1310.6	804.1	2012.4	2397.1	112.1
CMAVLA (Ours)	Under Review	1M Nominal	1000	**4515.1**	**4089.4**	**5242.8**	**3435.3**	**130.8**

treated as grouped samples rather than additional policy update steps. We therefore count only the primary trajectories toward the reported 1M timestep budget. CMAGRPO is trained for 1000 policy updates, matching the HASAC update schedule under 1M environment steps.

6 RESULTS

Table 1 presents the quantitative comparison on the ETH-UCY benchmark using ADE_{20}/FDE_{20}. Overall, TrajVLM achieves competitive trajectory forecasting performance across the five subsets, with an average ADE_{20}/FDE_{20} of $0.20/0.33$. Compared with conventional predictors, TrajVLM remains competitive with strong recent methods, although SOTA model SocialMOIF achieve lower overall error. This suggests that foundation model does not automatically improve every trajectory forecasting scenario, especially when compact task-specific models are highly optimized. Moreover, TrajVLM performs competitively among foundation model-based methods, achieving the best ADE_{20} on HOTEL (0.11) and ZARA1 (0.15) and ZARA2 (0.13). The gains vary across scenes, suggesting that multimodal semantic fusion is more beneficial in some interaction contexts than others. These results support the effectiveness of FMDT for multi-agent long-horizon trajectory prediction, showing that dual-transformer fusion preserves fine-grained motion dynamics while incorporating complementary high-level contextual information.

Table 2 and Fig. 3 summarize the quantitative results and training dynamics on MaMuJoCo. CMAVLA achieves maximum returns of 4515.1, 4089.4, 5242.8, 3435.3, and 130.8 on Ant,



(a) Ant-2×4 (b) Walker2d-2×3 (c) HalfCheetah-6×1

(d) Hopper-3×1 (e) Swimmer-2×1

Figure 3: Learning curves on five MaMuJoCo benchmarks. Solid lines denote the mean episode return, while shaded regions indicate the corresponding standard deviation.

Walker2d, HalfCheetah, Hopper, and Swimmer, respectively. CMAVLA consistently exceeds the HASAC mean plus one standard deviation reference across all five tasks, corresponding to improvements of approximately 28.8%, 54.7%, 18.4%, 203.2%, and 101.5%, respectively, while remaining competitive with the recently reported DIMA results on the available benchmarks. For the ablation studies, compared with the Trans-CMAGRPO and MLP-CMAGRPO ablations, CMAVLA also exhibits stronger final performance and generally more stable learning curves, indicating that incorporating VLM representations preserves stable training behavior in centralized multi-agent decision making. For both TrajVLM and CMAVLA, we adopt a staged frozen-to-LoRA training strategy, progressively adapting the foundation model through SFT-RL or MARL fine-tuning. For trajectory prediction, all modules are initialized without pretrained checkpoints. For MARL, CMAVLA initializes only the state transformer encoder from the pretrained model, while the action decoder remains randomly initialized. Consequently, its initial policy performance remains close to zero before progressively improving during training. These results demonstrate the feasibility of FMDT across both prediction and control tasks. Nevertheless, introducing a VLM incurs substantially higher GPU memory consumption and longer training time than conventional lightweight trajecotry networks and MARL models. Overall, the experiments show that the proposed dual-transformer coupling enables VLM-based frameworks to effectively learn multi-agent feature representations and achieve competitive performance across both long-horizon trajectory prediction and short-horizon multi-agent decision making.

7 Conclusion

In this work, we proposed FMDT, a unified foundation model-based dual-transformer framework for centralized multi-agent learning. By using agent-wise state features to query global VLM representations through cross-attention, FMDT bridges the representation gap between foundation models and structured multi-agent systems. We instantiated FMDT as TrajVLM for long-horizon trajectory forecasting and CMAVLA for centralized multi-agent decision making, with task-specific SFT-RL and MARL fine-tuning strategies. We further established theoretical guarantees for the dual-transformer hypothesis class through covering number, Rademacher complexity, uniform convergence, and sample complexity analyses. Experiments on ETH-UCY and MaMuJoCo demonstrate that FMDT achieves competitive performance across both prediction and control tasks, while CMAVLA shows substantial improvements over its state-only ablations. These results demonstrate the feasibility of dual-transformer coupling for adapting foundation models to centralized multi-agent systems. A remaining limitation is the additional GPU memory and training cost introduced by large VLMs, improving the computational efficiency of FMDT therefore represents an important direction for future work.

REFERENCES

Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–971, 2016.

Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, et al. $\pi^*_{0.6}$: A vla that learns from experience. In *Robotics: Science and Systems*, 2026.

Inhwan Bae, Jean Oh, and Hae-Gon Jeon. Eigentrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9983–9995. IEEE, 2023.

Inhwan Bae, Junoh Lee, and Hae-Gon Jeon. Can language beat numerical regression? language-based multimodal trajectory prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 753–766. IEEE, 2024.

Inhwan Bae, Junoh Lee, and Hae-Gon Jeon. Social reasoning-aware trajectory prediction via multimodal language model. *IEEE transactions on pattern analysis and machine intelligence*, 48(6): 6035–6052, 2025.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *9th Annual Conference on Robot Learning*, 2025a. URL https://openreview.net/forum?id=vlhoswksBO.

Kevin Black, Noah Brown, Danny Driess, et al. π_0: A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems*, 2025b.

Junting Chen, Checheng Yu, Xunzhe Zhou, Tianqi Xu, Yao Mu, Mengkang Hu, Wenqi Shao, Yikai Wang, Guohao Li, and Lin Shao. Emos: Embodiment-aware heterogeneous multi-robot operating system with llm agents. In *International Conference on Learning Representations*, volume 2025, pp. 56670–56690, 2025a.

Kai Chen, Xiaodong Zhao, Yujie Huang, Guoyu Fang, Xiao Song, Ruiping Wang, and Ziyuan Wang. Socialmoif: Multi-order intention fusion for pedestrian trajectory prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22465–22475. IEEE, 2025b.

Shizhe Chen, Paul Pacaud, and Cordelia Schmid. PointACT: Vision-Language-Action Models with Multi-Scale Point-Action Interaction. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.073.

Tanmay Devale, Pramith Devulapalli, and Steve Hanneke. Uniform convergence beyond glivenko-cantelli. In *37th International Conference on Algorithmic Learning Theory*, 2026. URL https://openreview.net/forum?id=EWtWBzDoe7.

Ria Doshi, Tian Gao, Annie Chen, Chelsea Finn, and Jeannette Bohg. Chorus: Decentralized multi-embodiment collaboration with one vla policy. *CoRL*, 2026.

Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pp. 5793–5831. PMLR, 2022.

Yuxiang Fu, Qi Yan, Lele Wang, Ke Li, and Renjie Liao. Moflow: One-step flow matching for human trajectory forecasting via implicit maximum likelihood estimation based distillation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17282–17293. IEEE, 2025.

Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17092–17101. IEEE, 2022.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2255–2264, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Steve Hanneke, Shay Moran, Vinod Raman, Unique Subedi, and Ambuj Tewari. Multiclass online learning and uniform convergence. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 5682–5696. PMLR, 2023.

Steve Hanneke, Shay Moran, and Waknine Tom. List sample compression and uniform convergence. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 2360–2388. PMLR, 2024.

Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51 (5):4282, 1995.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

Shyam Sundar Kannan, Vishnunandan L. N. Venkatesh, and Byung-Cheol Min. Smart-llm: Smart multi-agent robot task planning using large language models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 12140–12147, 2024.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pp. 5562–5571. PMLR, 2021.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=ZMnD6QZAE6.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EcGGFkNTxdJ.

Seongju Lee, Junseok Lee, Yeonguk Yu, Taeri Kim, and Kyoobin Lee. Mart: Multiscale relational transformer networks for multi-agent trajectory prediction. In *Computer Vision – ECCV 2024*, volume 15124 of *Lecture Notes in Computer Science*, pp. 89–107. Springer, 2024. doi: 10.1007/978-3-031-72848-8_6.

Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pp. 655–664. Wiley Online Library, 2007.

Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Yang Zhaohui, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, Dehui Wang, Dingxiang Luo, Yuchen Fan, Youbang Sun, Jia Zeng, Jiangmiao Pang, Shanghang Zhang, Yu Wang, Yao Mu, Bowen Zhou, and Ning Ding. SimpleVLA-RL: Scaling VLA training via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=TQhSodCM4r.

Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *Advances in neural information processing systems*, 33:19783–19794, 2020.

Peihan Li, Zijian An, Shams Abrar, and Lifeng Zhou. Large language models for multi-robot systems: A survey. *Autonomous Robots*, 50(3):30, 2026b.

Xiang Li, Cristina Mata, Jongwoo Park, Kumara Kahatapitiya, Yoo Jang, Jinghuan Shang, Kanchana Ranasinghe, Ryan Burgert, Mu Cai, Yong Jae Lee, et al. Llara: Supercharging robot learning data for vision-language policy. In *International Conference on Learning Representations*, volume 2025, pp. 57463–57509, 2025a.

Zhihao Li, Xue Jiang, Liyuan Liu, Xuelin Zhang, Hong Chen, and Feng Zheng. On the generalization ability of next-token-prediction pretraining. In *ICML*, 2025b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jiarong Liu, Yifan Zhong, Siyi Hu, Haobo Fu, Qiang Fu, Xiaojun Chang, and Yaodong Yang. Maximum entropy heterogeneous-agent reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 57333–57374, 2024.

Jijia Liu, Feng Gao, Bingwen Wei, Xinlei Chen, Qingmin Liao, Yi Wu, Chao Yu, and Yu Wang. What can rl bring to vla generalization? an empirical study. *Advances in Neural Information Processing Systems*, 38:97121–97151, 2026.

Kehui Liu, Zixin Tang, Dong Wang, Zhigang Wang, Xuelong Li, and Bin Zhao. Coherent: Collaboration of heterogeneous multi-robot system with large language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10208–10214. IEEE, 2025a.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 29982–30009, 2025b.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. VLA-RL: Towards masterful and general robotic manipulation with scalable reinforcement learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2026.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Omayma Mahjoub, Sasha Abramowitz, Ruan John de Kock, Wiem Khlifi, Simon Verster Du Toit, Jemma Daniel, Louay Ben Nessir, Louise Beyers, Juan Claude Formanek, Liam Clark, and Arnu Pretorius. Sable: a performant, efficient and scalable sequence model for MARL. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=CAbuWU44ky.

Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 15213–15222. IEEE, 2021.

Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5517–5526. IEEE, 2023.

Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14424–14432, 2020.

Abduallah Mohamed, Deyao Zhu, Warren Vu, Mohamed Elhoseiny, and Christian Claudel. Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In *European Conference on Computer Vision*, pp. 463–479. Springer, 2022.

Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. Facmac: Factored multi-agent centralised policy gradients. *Advances in neural information processing systems*, 34:12208–12221, 2021.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. *arXiv preprint arXiv:2001.03093*, 2020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *International Conference on Artificial Intelligence and Statistics*, pp. 1405–1413. PMLR, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE international Conference on Robotics and Automation (ICRA)*, pp. 4601–4607. IEEE, 2018.

Weizheng Wang, Baijian Yang, Sungeun Hong, Wenhai Sun, and Byung-Cheol Min. Hyper-sttn: Hypergraph augmented spatial-temporal transformer network for trajectory prediction. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6498–6507, 2022.

Zirui Xu, Biao Yang, Rongrong Ni, Zhongkai Zhou, and Shaobo Shen. W2w: Language-model-based trajectory prediction with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23538–23548, 2026.

Chenxiao Yang, Nathan Srebro, and Zhiyuan Li. Tight sample complexity of transformers. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 6887–6923. PMLR, 29 Jun–03 Jul 2026. URL https://proceedings.mlr.press/v336/yang26a.html.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *European conference on computer vision*, pp. 507–523. Springer, 2020.

Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9813–9823, 2021.

Hongzhi Zang, Mingjie Wei, Si Xu, Yongji Wu, Zhen Guo, Yuanqing Wang, Hao Lin, Liangzhi Shi, Yuqing Xie, Zhexuan Xu, Zhihao Liu, Kang Chen, Wenhao Tang, Quanlu Zhang, Weinan Zhang, Chao Yu, and Yu Wang. Rlinf-vla: A unified and efficient framework for reinforcement learning of vision-language-action models. In *Robotics: Science and Systems*, 2026a.

Hongzhi Zang, Shu'ang Yu, Hao Lin, Tianxing Zhou, Zefang Huang, Zhen Guo, Xin Xu, Jiakai Zhou, Yuze Sheng, Shizhe Zhang, Feng Gao, Wenhao Tang, Yufeng Yue, Quanlu Zhang, Xinlei Chen, Chao Yu, and Yu Wang. Rlinf-user: A unified and extensible system for real-world online policy learning in embodied ai. *Proceedings of Robotics: Science and Systems*, 2026b.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. In *International Conference on Learning Representations*, volume 2024, pp. 19373–19401, 2024.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Yang Zhang, Xinran Li, Jianing Ye, Shuang Qiu, Delin Qu, Xiu Li, Chongjie Zhang, and Chenjia Bai. Revisiting multi-agent world modeling from a diffusion-inspired perspective. *Advances in Neural Information Processing Systems*, 38:65439–65470, 2026.

Mandi Zhao, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. In *IEEE International Conference on Robotics and Automation*, 2024.

Zhengbang Zhu, Minghuan Liu, Liyuan Mao, Bingyi Kang, Minkai Xu, Yong Yu, Stefano Ermon, and Weinan Zhang. Madiff: Offline multi-agent learning with diffusion models. *Advances in Neural Information Processing Systems*, 37:4177–4206, 2024.

SUPPLEMENTARY MATERIALS OF FMDT

A Implementation Details **16**

A.1 ETH-UCY Dataset 16

A.2 TrajVLM Multi-Modal Input 16

A.3 MAMuJoCo Benchmark 19

A.4 CMAVLA Multi-Modal Input 20

B Theoretical Analysis and More Implementation Details **22**

A Implementation Details

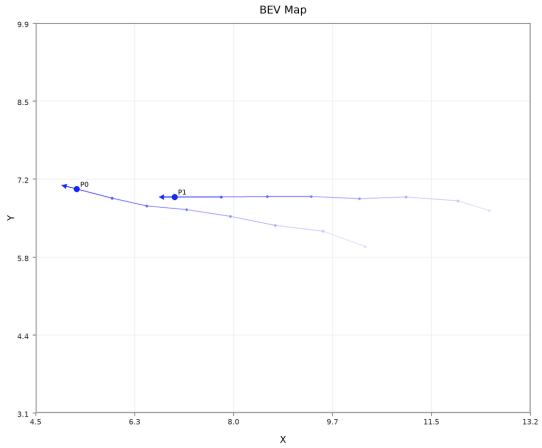
A.1 ETH-UCY Dataset

ETH-UCY Lerner et al. (2007) is a real-world pedestrian trajectory benchmark composed of five heterogeneous scenes that include ETH, HOTEL, UNIV, ZARA1, and ZARA2. These scenes cover diverse environments, including university campuses, hotel surroundings, and public shopping areas, with substantially different pedestrian densities, walking patterns, and interaction structures. The trajectories are captured from fixed cameras and transformed from image coordinates into real-world ground-plane coordinates through scene-specific homography matrices. Consequently, the benchmark contains diverse behaviors such as individual walking, group motion, crossing, following, and collision avoidance. Such variations in scene geometry and crowd interactions make ETH-UCY a representative benchmark for evaluating the generalization capability of multi-agent trajectory forecasting models across different environments.

A.2 TrajVLM Multi-Modal Input

TrajVLM constructs its multimodal input directly from the historical multi-agent trajectories. For each scene, the original numerical trajectory tensor is retained as the state-transformer input, while the same observations are converted into a bird's-eye-view (BEV) image and a textual motion description for the VLM. The BEV representation visualizes the global spatial configuration and historical motion of all pedestrians, where earlier trajectory points are rendered lighter and recent observations are emphasized. The current position is highlighted together with a heading arrow, enabling the VLM to capture crowd geometry, relative positions, and motion directions.

The textual input provides an explicit description of each pedestrian P_i, including its current position, velocity, speed, heading direction, and historical motion states. Agent identities are strictly aligned across the three representations, i.e., $\mathrm{BEV}(P_i) \leftrightarrow \mathrm{Text}(P_i) \leftrightarrow X_S[i]$, such that the VLM and state transformer observe consistent agent-wise information. The complete scene, rather than only the target pedestrian, is provided to the VLM to preserve social-context information and inter-agent interactions.

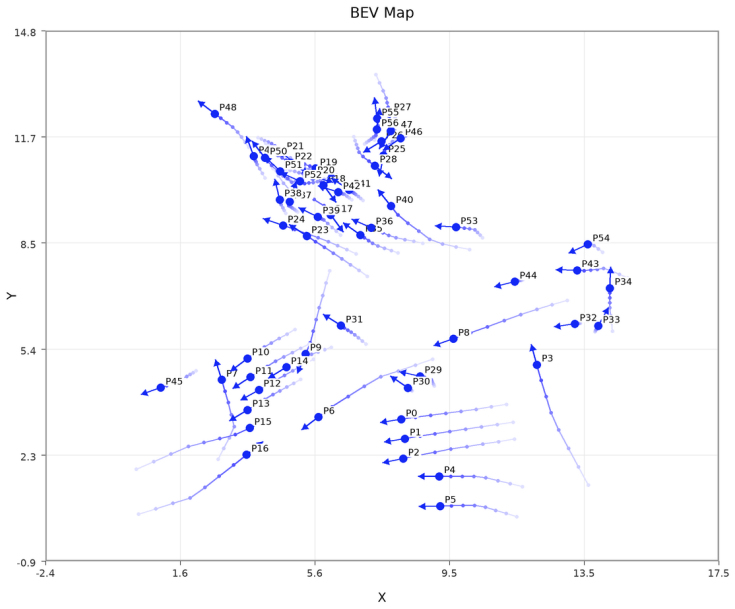


A sparse-scene BEV input to TrajVLM, containing 2 observed pedestrians.

```
<|im_start|>user
<|vision_start|><|image_pad|><|vision_end|>

Task Prompt:

You are performing multi-agent trajectory prediction.

Use the historical observations from T=1 to T=8 to understand the
motion of all pedestrians in the scene. T=8 is the current timestep.
```

```
The scene uses a 2D X-Y Cartesian coordinate system. Each pedestrian
is identified as P_i. Velocity is defined from consecutive positions
as vx_t=(x_t-x_(t-1))/delta_t and vy_t=(y_t-y_(t-1))/delta_t, with
delta_t=1 for this input.

BEV Observation Prompt:

The BEV shows the global spatial configuration and historical
trajectories of all pedestrians. Lighter trajectory marks indicate
earlier observations and darker marks indicate more recent
observations. The emphasized point with a heading arrow is the
current position and heading at T=8.

Textual Observation Prompt:

The textual observation describes each pedestrian's current state
at T=8 followed by historical motion from T=7 to T=1, including
position, velocity, speed, and heading.

Maintain strict correspondence across modalities:
BEV P_i = textual observation P_i = scene_past_abs[i].

The scene contains 2 observed pedestrians.
The downstream sample predicts pedestrian P1; the complete scene is
provided for social-context reasoning.

Pedestrian P0:
T=8: location=(5.240, 6.980), velocity=(-0.620, 0.160),
speed=0.640, heading theta=165.530 degrees.
T=7: location=(5.860, 6.820), velocity=(-0.610, 0.140),
speed=0.626, heading theta=167.074 degrees.
T=6: location=(6.470, 6.680), velocity=(-0.700, 0.060),
speed=0.703, heading theta=175.101 degrees.
T=5: location=(7.170, 6.620), velocity=(-0.770, 0.120),
speed=0.779, heading theta=171.142 degrees.
T=4: location=(7.940, 6.500), velocity=(-0.790, 0.160),
speed=0.806, heading theta=168.551 degrees.
T=3: location=(8.730, 6.340), velocity=(-0.840, 0.100),
speed=0.846, heading theta=173.211 degrees.
T=2: location=(9.570, 6.240), velocity=(-0.740, 0.270),
speed=0.788, heading theta=159.955 degrees.
T=1: location=(10.310, 5.970), velocity=(0.000, 0.000),
speed=0.000, heading theta=undefined.

Pedestrian P1:
T=8: location=(6.960, 6.840), velocity=(-0.820, 0.000),
speed=0.820, heading theta=180.000 degrees.
T=7: location=(7.780, 6.840), velocity=(-0.810, -0.010),
speed=0.810, heading theta=180.707 degrees.
T=6: location=(8.590, 6.850), velocity=(-0.770, 0.000),
speed=0.770, heading theta=180.000 degrees.
T=5: location=(9.360, 6.850), velocity=(-0.850, 0.040),
speed=0.851, heading theta=177.306 degrees.
T=4: location=(10.210, 6.810), velocity=(-0.820, -0.030),
speed=0.821, heading theta=182.095 degrees.
T=3: location=(11.030, 6.840), velocity=(-0.910, 0.070),
speed=0.913, heading theta=175.601 degrees.
T=2: location=(11.940, 6.770), velocity=(-0.550, 0.170),
speed=0.576, heading theta=162.824 degrees.
T=1: location=(12.490, 6.600), velocity=(0.000, 0.000),
speed=0.000, heading theta=undefined.

Use the complete scene history to understand pedestrian motion,
spatial relationships, and social interactions.
```

```
<|im_end|>
```



A crowded-scene BEV input to TrajVLM, containing 57 observed pedestrians.

```
<|im_start|>user
<|vision_start|><|image_pad|><|vision_end|>

Task Prompt:

You are performing multi-agent trajectory prediction.

Use the historical observations from T=1 to T=8 to understand the
motion of all pedestrians in the scene. T=8 is the current timestep.

The scene uses a 2D X-Y Cartesian coordinate system. Each pedestrian
is identified as P_i. Velocity is defined from consecutive positions
as vx_t=(x_t-x_(t-1))/delta_t and vy_t=(y_t-y_(t-1))/delta_t,
with delta_t=1 for this input.

BEV Observation Prompt:

The BEV shows the global spatial configuration and historical
trajectories of all pedestrians. Lighter trajectory marks indicate
earlier observations and darker marks indicate more recent
observations. The emphasized point with a heading arrow is the
current position and heading at T=8.

Textual Observation Prompt:

The textual observation describes each pedestrian's current state
at T=8 followed by historical motion from T=7 to T=1, including
position, velocity, speed, and heading.

Maintain strict correspondence across modalities:
BEV P_i = textual observation P_i = scene_past_abs[i].

The scene contains 57 observed pedestrians.
```

```
The downstream sample predicts pedestrian P4; the complete scene
is provided for social-context reasoning.

...

Representative pedestrian P1:

Current state:
T=8: location=(8.227, 2.737), velocity=(-0.442, -0.072),
speed=0.448, heading theta=189.201 degrees.

Previous historical motion:
T=7: location=(8.669, 2.809), velocity=(-0.442, -0.072),
speed=0.448, heading theta=189.204 degrees.
T=6: location=(9.111, 2.881), velocity=(-0.442, -0.072),
speed=0.448, heading theta=189.201 degrees.
T=5: location=(9.553, 2.952), velocity=(-0.467, -0.068),
speed=0.472, heading theta=188.264 degrees.
T=4: location=(10.020, 3.020), velocity=(-0.469, -0.067),
speed=0.474, heading theta=188.159 degrees.
T=3: location=(10.489, 3.087), velocity=(-0.470, -0.067),
speed=0.474, heading theta=188.157 degrees.
T=2: location=(10.959, 3.155), velocity=(-0.470, -0.067),
speed=0.474, heading theta=188.156 degrees.
T=1: location=(11.428, 3.222), velocity=(0.000, 0.000),
speed=0.000, heading theta=undefined.

...

The remaining pedestrians P0, P2, ..., P56 follow the same
agent-wise textual representation.

Use the complete scene history to understand pedestrian motion,
spatial relationships, and social interactions.

<|im_end|>
```

A.3 MAMuJoCo Benchmark

MaMuJoCo Peng et al. (2021) extends the standard MuJoCo continuous-control environments to cooperative multi-agent settings by decomposing a single articulated robot into multiple agents. Each agent controls a subset of the robot joints and receives its corresponding observations, while all agents must coordinate their actions to accomplish a shared locomotion objective. Since the agents jointly determine the dynamics of the same physical body, independently reasonable actions may still lead to globally suboptimal or unstable behaviors, making coordinated representation learning and policy optimization essential.

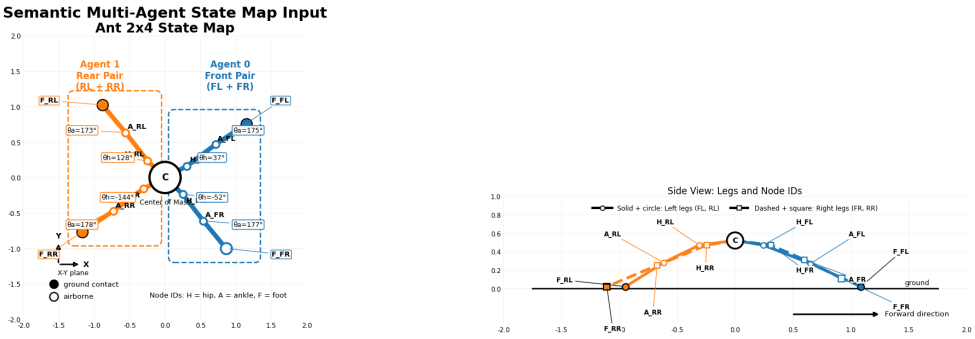
The benchmark provides multiple agent-partition configurations across different robot morphologies. For example, Ant-2 $\times$ 4 divides the eight actuated joints between two agents, whereas HalfCheetah-6 $\times$ 1 assigns individual joints to six agents. Walker2d-2 $\times$ 3, Hopper-3 $\times$ 1, and Swimmer-2 $\times$ 1 introduce additional coordination patterns with different action dimensions, body structures, and locomotion dynamics. These heterogeneous configurations make MaMuJoCo suitable for evaluating whether a multi-agent policy can generalize its coordination mechanism across substantially different embodiments.

Compared with trajectory forecasting, MaMuJoCo additionally requires closed-loop sequential decision making, where each joint action changes the subsequent system state and therefore affects all agents' future observations and rewards. This property makes it a complementary benchmark for evaluating FMDT in short-horizon multi-agent control, in addition to its long-horizon forecasting capability on ETH-UCY.

A.4 CMAVLA MULTI-MODAL INPUT

For CMAVLA, we convert the numerical multi-agent state into two complementary semantic images and jointly pass them to the VLM together with the textual state description. The first image is a body-centered X-Y state map that represents the global robot configuration, agent partition, joint geometry, and foot-contact states. The second image provides a side-view representation of the robot posture relative to the ground, emphasizing leg extension, body height, contact relationships, and forward direction.

These two visual representations provide complementary geometric perspectives of the same multi-agent state: the top-view map captures spatial and inter-agent structure, while the side-view map emphasizes posture and ground-contact dynamics. Together with the textual semantic state, the VLM is therefore conditioned on global posture, front-rear and left-right coordination, leg geometry, and contact patterns for centralized multi-agent decision making.



Two semantic image inputs to the CMAVLA VLM: the body-centered multi-agent state map and the side-view posture map.

```
<|im_start|>user

<|vision_start|><|image_pad|><|vision_end|>
<|vision_start|><|image_pad|><|vision_end|>

[TASK]

Control the multi-agent Ant to move forward while maintaining
stable posture, coordinated leg motion, and effective ground contact.


[MULTI-AGENT CONFIGURATION]

The Ant is decomposed into 2 agents. Agent 0 controls the front
pair of legs: FL and FR, with four controlled joints in total.
Agent 1 controls the rear pair of legs: RL and RR, with four
controlled joints in total.


[VISUAL INPUT INTERPRETATION]

Two complementary semantic state maps are provided.

Image 1 presents a body-centered X-Y top view of the Ant.
It describes the global robot configuration, leg geometry,
agent partition, joint orientations, and foot-contact states.

Image 2 presents a side view of the same robot state relative
to the ground. It describes body posture, leg extension,
ground-contact relationships, and forward direction.

C denotes the robot center/root reference.
```

```
For each leg, H denotes the hip joint, A denotes the ankle
joint, and F denotes the foot/end node.

theta_h denotes the semantic hip-link orientation relative
to the robot forward axis, while theta_a denotes the ankle
joint geometric angle.

A filled foot marker indicates ground contact, while a hollow
foot marker indicates that the foot is airborne.

In the Ant 2x4 configuration, legs controlled by the same
agent share the same color.


[CURRENT SEMANTIC STATE]

Center/root height above ground: 0.520 m.
Robot heading in world coordinates: 8.0 degrees.

FL (Agent 0):
theta_h=37.0 deg,
theta_a=175.0 deg,
foot=ground contact.

FR (Agent 0):
theta_h=-52.1 deg,
theta_a=176.6 deg,
foot=airborne.

RL (Agent 1):
theta_h=127.8 deg,
theta_a=173.3 deg,
foot=ground contact.

RR (Agent 1):
theta_h=-143.8 deg,
theta_a=177.6 deg,
foot=ground contact.


[CONTROL INTERPRETATION]

Interpret the two semantic images and textual state jointly.
Focus on whole-body stability, front-rear coordination,
left-right coordination, leg extension/flexion, and contact
patterns.

Use the multimodal observations to construct a centralized
multi-agent representation for cooperative action generation.

<|im_end|>
```

Table 3: Structural Comparison among Current MARL Algorithm Paradigms.

Component	**Existing Approaches Single-Agent RL**	**Existing Approaches CTDE MARL**	**Existing Approaches CTCE MARL**	**FMDT CTCE MARL**
System Formulation	single agent	N agents	N agents	N agents
Input	single robot state	local observation	global state	multimodal context + global state
Encoder	single agent encoder	heterogeneous or shared encoders	centralized encoder	centralized VLM encoder + state encoder
Encoder Number	1	1 shared or N agent-specific	1	1
VLM number	None	None	None	1 centralized VLM
Action Decoder	single decoder	decentralized actor decoders	shared sequential actor decoder	independent agent-specific actor decoders
Action Decoder Number	1	N	1	N
Representative Instances	PPO Schulman et al. (2017) SAC Haarnoja et al. (2018)	MADDPG Lowe et al. (2017) MAPPO Yu et al. (2022)	Sable Mahjoub et al. (2025)	**CMAVLA**

Table 3 summarizes the structural differences among representative single-agent RL, CTDE MARL, CTCE MARL, and the proposed FMDT-based CTCE paradigm. Unlike conventional approaches that rely solely on numerical states or local observations, FMDT jointly processes the global multi-agent state and multimodal context through a centralized state encoder and a shared VLM. The fused representation is then mapped to individual actions through N agent-specific action decoders. This design preserves centralized multimodal reasoning while maintaining explicit agent-wise action generation, distinguishing CMAVLA from both decentralized actor architectures and conventional centralized sequential decoding frameworks.

B Theoretical Analysis and More Implementation Details

The assumptions and theoretical proofs with more implementation details can be found via the project website: `https://sites.google.com/view/cmas-fmdt`.