GROUPMODES: LEARNING TO COUPLE GOALS AND INTERACTIONS FOR HUMAN TRAJECTORY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate human trajectory prediction is essential for autonomous driving, mobile robotics, and crowd monitoring. Yet, existing models can achieve low individual trajectory errors without necessarily generating coherent group-level futures that account for interactions among agents. To address this, we propose GroupModes, a trajectory prediction architecture that conditions agent-specific goals and evolving social interactions on a shared group-level mode. Position-aware encoding captures observed motion and spatial context. A structured mode bank generates observation-conditioned modes, augmented with bounded stochastic residuals to represent within-mode variation. Each mode specifies individual goals and modulates time-varying interaction logits through a low-rank parameterization. A goal-conditioned synchronous decoder jointly generates future trajectories for the target and its neighboring agents. Experimental results demonstrate that GroupModes achieves competitive prediction accuracy and that aligning goal and interaction realization under a shared mode improves target and joint forecasting. Further analyses indicate that the learned group modes organize individual trajectory candidates into coordinated joint hypotheses, providing benefits beyond individual candidate coverage. Our code is available at `https://anonymous.4open.science/r/groupmodes-review-0F6B`.

1 INTRODUCTION

Accurate human trajectory prediction is fundamental to autonomous driving, mobile robotics, and crowd monitoring (Rudenko et al., 2020). In interactive environments, however, people continually adapt their motion to others (Helbing & Molnar, 1995). The same observed history may lead to different future trajectories, which vary in both individual goals and the social interactions that unfold as agents move toward these goals. For example, agents approaching the same area may pass in different orders, with one proceeding while another slows down or yields. Predicting coherent group-level futures therefore requires considering individual goals together with the interactions among agents.

Modeling social interactions has long been central to human trajectory prediction. Early methods aggregate neighboring motion through social pooling (Alahi et al., 2016), while subsequent methods capture richer interaction patterns using graph-based architectures, attention mechanisms, or group-level representations (Salzmann et al., 2020; Yuan et al., 2021; Xu et al., 2023; Bae et al., 2022). Related advances in multimodal and joint prediction incorporate interactions into the generation of future trajectories, accounting for dependencies among agents beyond their observed motion (Varadarajan et al., 2022; Ngiam et al., 2021; Xu et al., 2023). These studies provide a foundation for modeling how agents influence one another during prediction.

Future uncertainty, in parallel, has been modeled through various stochastic generative strategies. Conditional latent-variable and adversarial models sample plausible futures from learned distributions (Gupta et al., 2018; Salzmann et al., 2020). Goal-conditioned methods generate trajectories toward sampled or predicted destinations (Zhao et al., 2021; Mangalam et al., 2020), and diffusion models produce multimodal futures through iterative denoising (Mao et al., 2023). Group-aware generative models further incorporate group-level dependencies into multimodal predic-

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.

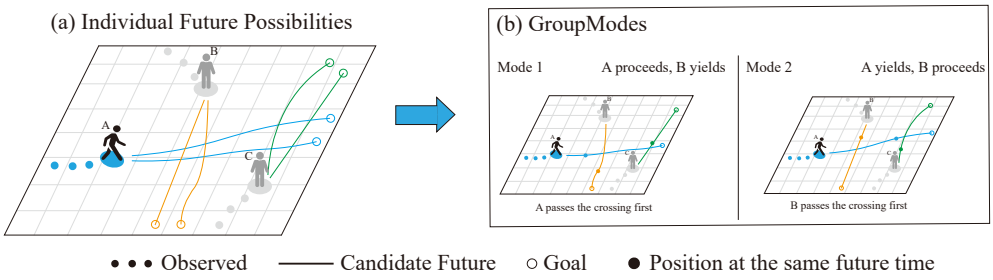

Figure 1: Motivation illustration. (a) The same observed history admits multiple individual futures, whose plausibility alone does not ensure joint consistency. (b) GroupModes conditions agent-specific goals and evolving interactions on a shared group-level mode to generate joint hypotheses. Two illustrative modes yield different crossing orders, with either A or B proceeding first. Highlighted future positions indicate the same prediction time within each hypothesis.

tion (Xu et al., 2022a). These advances motivate investigating how a shared future mode can jointly shape individual goals and social interactions during prediction. Goals specify where agents are predicted to arrive. Evolving interactions shape how they move in response to one another.

To this end, we introduce GroupModes, a trajectory prediction architecture that jointly conditions agent-specific goals and time-varying social interactions on a shared group-level mode. Position-aware encoding captures observed motion and target-relative spatial context. A learnable mode bank generates a finite set of observation-conditioned modes, while bounded stochastic residuals model variation within each mode. Each resulting mode specifies individual goals and modulates time-varying interaction logits through a low-rank parameterization. A goal-conditioned synchronous decoder uses these goals and interactions to jointly generate trajectories for the target and its neighboring agents. Each candidate index thus identifies one joint future with distinct agent-specific goals under shared mode conditioning.

Our experimental results demonstrate competitive prediction accuracy and support the benefits of structured modes over a Conditional Gaussian baseline. Controlled ablations show that aligning goal and interaction realization improves target and joint forecasting. Independently shuffling cross-agent candidate indices preserves each agent's candidate set but increases joint prediction errors and the mean collision rate. These findings support the value of learned candidate correspondence beyond individual candidate coverage. Additional analyses show that the accuracy advantage over Conditional Gaussian persists under shortened observation histories and across neighborhood densities. Details are provided in Section 4.

In summary, our contributions are twofold. First, we propose GroupModes, which uses structured group modes to jointly condition agent-specific goals and evolving social interactions for synchronous joint prediction. Second, we provide controlled analyses of mode organization, goal–interaction coupling, and cross-agent candidate pairing. These analyses establish predictive benefits beyond individual candidate coverage and clarify the role of shared modes in organizing joint futures.

2 Related Work

2.1 Human Trajectory Prediction

Human trajectory prediction has seen considerable progress in recent years. Early recurrent models encode individual motion and aggregate neighboring motion through social pooling (Alahi et al., 2016). Subsequent methods model social interactions using spatio-temporal graphs, attention mechanisms, Transformers, equivariant architectures, and group-level representations (Huang et al., 2019; Mohamed et al., 2020; Yuan et al., 2021; Xu et al., 2022a; Bae et al., 2022; Xu et al., 2023; Wong et al., 2024a). Goal-conditioned approaches further guide trajectory generation through predicted destinations or waypoints (Mangalam et al., 2020; 2021; Xu et al., 2022b). These approaches pro-

vide complementary mechanisms for modeling observed motion, social interactions, and future goals.

Beyond motion and interaction representations, multimodal predictors differ in how they generate future hypotheses. GAN-based and CVAE-based models sample continuous latent variables to produce multiple plausible trajectories (Gupta et al., 2018; Sadeghian et al., 2019; Kosaraju et al., 2019; Salzmann et al., 2020; Xu et al., 2022c). More structured approaches organize candidate futures through specialized generators (Dendorfer et al., 2021) or retrieved intentions (Xu et al., 2022b). EigenTrajectory further combines low-rank trajectory descriptors with anchor-based refinement (Bae et al., 2023). Diffusion-based predictors generate trajectories through iterative denoising (Gu et al., 2022; Mao et al., 2023; Bae et al., 2024; Hetzel et al., 2026), while flow-based methods use normalizing flows or flow matching to model the predictive distribution (Maeda & Ukita, 2023; Fu et al., 2025). Complementing these generative strategies, joint predictors model dependencies among agents when constructing future hypotheses (Luo et al., 2023; Rowe et al., 2023).

2.2 Group Interaction and Joint Prediction

Interaction modeling has become a central component of human trajectory prediction. Graph-based methods represent agents as nodes and exchange social information through graph convolution, message passing, or attention (Huang et al., 2019; Mohamed et al., 2020; Shi et al., 2021; Kim et al., 2024). Hypergraph and hierarchical models further capture higher-order relations through hyperedges or explicit pedestrian–group hierarchies (Xu et al., 2022a; Bae et al., 2022). MART combines pairwise and hypergraph relational Transformers with adaptive group estimation to model individual and group relations (Lee et al., 2024). Other approaches learn relational structures from motion, including group memberships, interaction strengths and types, and evolving interaction graphs (Li et al., 2020; Xu et al., 2022a; Bae et al., 2022; Xu et al., 2024). These methods establish the value of modeling group relations alongside pairwise interactions and their evolution over time.

Joint prediction methods incorporate cross-agent dependencies into future generation. Representative approaches use latent variables, scene-centric attention, or marginal and conditional prediction to model multi-agent futures (Yuan et al., 2021; Ngiam et al., 2021; Sun et al., 2022). AutoBots combines learnable decoder seeds with temporal and social attention to generate joint future trajectories (Girgis et al., 2021). MotionLM formulates joint forecasting as autoregressive generation of discrete motion tokens, while JFP learns pairwise interactions within a structured graphical model (Seff et al., 2023; Luo et al., 2023). These formulations encode cross-agent dependence through different representations and factorizations of the joint predictive distribution.

Recent methods further incorporate predicted futures into interaction reasoning through learned interaction graphs, hierarchical intention–behavior fusion, and level-wise decoding (Rowe et al., 2023; Zhu et al., 2023; Huang et al., 2023). GroupModes uses shared group modes to jointly condition agent-specific goals and low-rank modulation of time-varying interaction logits for synchronous joint decoding.

3 Method

GroupModes treats interacting agents as the unit of generation and predicts multiple complete joint futures. As illustrated in Figure 2, the observed trajectories are first encoded in a target-centric frame to obtain agent representations H and a group context c. The group context conditions a structured mode bank that produces a finite set of observation-conditioned group modes. Each selected mode conditions both agent-specific goals and time-varying interactions through low-rank logit modulation. These goals and interactions guide a goal-conditioned synchronous decoder to jointly generate the agents' future trajectories.

3.1 Problem Formulation

Given a target agent and its selected neighboring agents, let N denote the number of valid agents jointly predicted. Here, a group refers to this set of agents rather than an explicitly identified social group. The observed and future trajectories of agent i are denoted by $X_i = [x_i^1, \ldots, x_i^{T_{\text{obs}}}] \in$

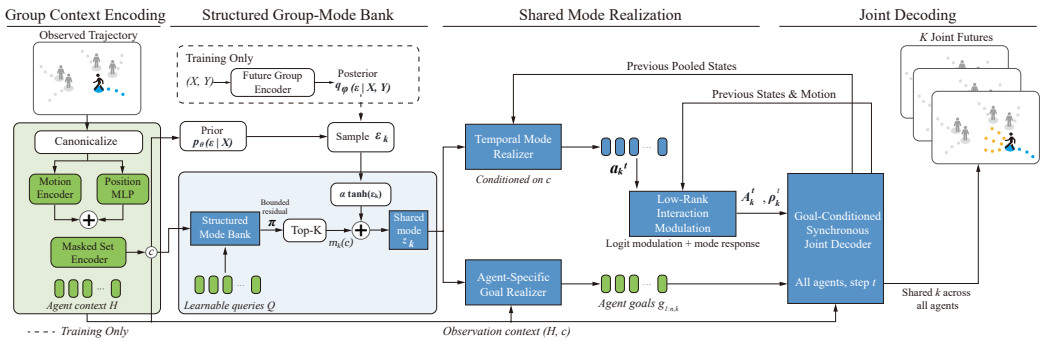

Figure 2: Overview of GroupModes. Structured group modes jointly condition agent-specific goals and time-varying interactions for synchronous joint decoding. The future encoder is used only during training.

$\mathbb{R}^{T_{\text{obs}}\times 2}$ and $Y_i = [y_i^1, \ldots, y_i^{T_f}] \in \mathbb{R}^{T_f\times 2}$, respectively, where $x_i^t, y_i^t \in \mathbb{R}^2$. The trajectories of these agents are collected as $X = [X_1, \ldots, X_N] \in \mathbb{R}^{N\times T_{\text{obs}}\times 2}$ and $Y = [Y_1, \ldots, Y_N] \in \mathbb{R}^{N\times T_f\times 2}$. We model the joint conditional distribution $p_\theta(Y \mid X)$ and generate K joint hypotheses $\widehat{Y} = [\widehat{Y}^{(1)}, \ldots, \widehat{Y}^{(K)}]$, where each $\widehat{Y}^{(k)} \in \mathbb{R}^{N\times T_f\times 2}$ contains the predicted trajectories of all N agents. The same hypothesis index k is shared across these agents, identifying one joint future for the group.

3.2 Target-Centric Group Context Encoding

For each scene, let $i^\star$ denote the target agent, $o^\star = x_{i^\star}^{T_{\text{obs}}}$ its last observed position, and $\phi^\star$ its observed heading. We define the target-centric rigid transformation and its inverse as

$$\mathcal{T}_{i^\star}(p) = R(-\phi^\star)(p - o^\star), \qquad \mathcal{T}_{i^\star}^{-1}(\bar{p}) = R(\phi^\star)\bar{p} + o^\star. \tag{1}$$

We apply $\mathcal{T}_{i^\star}$ to the observed trajectories of all agents and, during training, to the ground-truth future trajectories Y, using transformation parameters derived exclusively from X. Coordinates retain their original scale, and predicted trajectories are restored to the world frame through $\mathcal{T}_{i^\star}^{-1}$.

For the canonicalized trajectory $\bar{X}_i = [\bar{x}_i^1, \ldots, \bar{x}_i^{T_{\text{obs}}}]$, we define $\Delta\bar{x}_i^t = \bar{x}_i^t - \bar{x}_i^{t-1}$ and $\Delta\bar{X}_i = [\Delta\bar{x}_i^2, \ldots, \Delta\bar{x}_i^{T_{\text{obs}}}]$. A shared motion embedding and temporal positional encoding produce $S_i = \text{MLP}^m(\Delta\bar{X}_i) + P$, which is summarized as $r_i = E_{\text{obs}}(S_i; M_i^t)$. To retain target-relative spatial context, we use $\widetilde{r}_i = r_i + \text{MLP}^p(\bar{x}_i^{\text{last}})$, where $\bar{x}_i^{\text{last}}$ is the last valid canonicalized position.

A masked set encoder then aggregates the position-aware representations:

$$(H, c) = E_{\text{set}}\big([\widetilde{r}_1, \ldots, \widetilde{r}_N]; M^a\big), \qquad H = [h_1, \ldots, h_N]. \tag{2}$$

Here, $H \in \mathbb{R}^{N\times D}$ contains context-aware agent representations, $c \in \mathbb{R}^D$ denotes the group context, and M_i^t and M^a mask invalid observation steps and padded agents, respectively. Further implementation details are provided in Appendix A.1.

3.3 Structured Group-Mode Bank

We maintain $M = 20$ orthogonally initialized learnable mode queries $Q = [q_1, \ldots, q_M]$. Conditioned on the group context c, the mode bank maps these queries to observation-conditioned group modes $m_k(c) \in \mathbb{R}^{d_z}$ and corresponding logits ℓ_k:

$$\big\{m_k(c), \ell_k\big\}_{k=1}^{M} = B_{\text{mode}}(Q, c), \qquad \pi = \text{softmax}([\ell_1, \ldots, \ell_M]). \tag{3}$$

Each $m_k(c)$ defines a candidate group mode shared across all agents, while π_k denotes the corresponding probability conditioned on X.

To capture fine-grained within-mode variation, we introduce a Gaussian residual $\epsilon \in \mathbb{R}^{d_z}$. Its observation-conditioned prior is parameterized by c. During training, a Future Group Encoder addi-

tionally encodes $(\bar{X}, \bar{Y})$ as $g_Y = E_{\text{fut}}(\bar{X}, \bar{Y}; M^a, M^f)$, yielding the prior and posterior

$$\begin{aligned} p_\theta(\epsilon \mid X) &= \mathcal{N}\big(\mu_p(c), \mathrm{Diag}\big(\sigma_p^2(c)\big)\big)\,, \\ q_\phi(\epsilon \mid X, Y) &= \mathcal{N}\big(\mu_q(c, g_Y), \mathrm{Diag}\big(\sigma_q^2(c, g_Y)\big)\big)\,. \end{aligned} \tag{4}$$

Both use diagonal covariance and the reparameterization $\epsilon = \mu + \sigma \odot \xi$, where $\xi \sim \mathcal{N}(0, I)$. The posterior and E_{fut} are used only during training.

For mode k, the shared group-mode code is

$$z_k = m_k(c) + \alpha \tanh(\epsilon_k), \qquad \alpha = 0.15. \tag{5}$$

The bounded stochastic residual models within-mode variation by limiting the perturbation around each mode code $m_k(c)$. At inference, ϵ_k is sampled from $p_\theta(\epsilon \mid X)$. When $K = M$, every mode is realized once; when $K < M$, the K modes with the largest π_k are selected. During training, a posterior residual is combined with the best-matching mode $m_{k^\star}(c)$, where $k^\star$ is selected using the trajectory and endpoint errors defined in Section 3.6.

3.4 Temporal Low-Rank Interaction Realization

For each selected group mode k, the shared group-mode code z_k is realized as a sequence of time-varying interaction matrices over the prediction horizon. All states below are expressed in the canonical frame, with bars omitted for clarity. At step t, let $s_{i,k}^{t-1}$, $\hat{y}_{i,k}^{t-1}$, and $\Delta\hat{y}_{i,k}^{t-1}$ denote the decoder state, position, and displacement of agent i, initialized with $\hat{y}_{i,k}^0 = \bar{x}_i^{T_{\text{obs}}}$ and $\Delta\hat{y}_{i,k}^0 = \Delta\bar{x}_i^{T_{\text{obs}}}$.

A Temporal Mode Realizer pools the valid decoder states as $\bar{s}_k^{t-1} = \mathrm{Pool}_i(s_{i,k}^{t-1}; M^a)$ and updates

$$h_k^t = \mathrm{GRU}_{\text{mode}}\left([c, z_k, \bar{s}_k^{t-1}, p_t], h_k^{t-1}\right), \qquad a_k^t = f_a(h_k^t), \tag{6}$$

where $h_k^0 = f_{\text{init}}([c, z_k])$, p_t is a prediction-step embedding, and $a_k^t \in \mathbb{R}^r$ contains mode-specific, time-varying amplitudes shared across all valid edges.

For each directed edge $j \to i$, we form the relative-motion descriptor

$$e_{ij,k}^t = [\hat{y}_{j,k}^{t-1} - \hat{y}_{i,k}^{t-1},\ \Delta\hat{y}_{j,k}^{t-1} - \Delta\hat{y}_{i,k}^{t-1}]$$

and its relative-motion base logit $b_{ij,k}^t = f_{\text{base}}(e_{ij,k}^t)$. The receiver and sender states are projected into $u_{i,k}^t = f_U(s_{i,k}^{t-1})$ and $v_{j,k}^t = f_V(s_{j,k}^{t-1})$, both in $\mathbb{R}^r$. The resulting mode-conditioned interaction weight is

$$A_{ij,k}^t = \sigma\Big(b_{ij,k}^t + r^{-1/2}\left\langle u_{i,k}^t \odot \tanh(a_k^t), v_{j,k}^t \right\rangle\Big)\, M_{ij}^e. \tag{7}$$

Here, M^e masks self-connections, padded agents, and invalid edges. The asymmetric receiver–sender projections encode interaction directionality. Before sigmoid activation and edge masking, the additive mode-dependent logit modulation has rank at most r. The resulting interaction matrix $A_k^t = [A_{ij,k}^t]$ is reconstructed at every prediction step and passed to the synchronous joint decoder.

3.5 Goal-Conditioned Synchronous Joint Decoding

The same group-mode code z_k used for interaction modulation also conditions agent-specific goal prediction. For each agent, we predict a goal relative to constant-velocity extrapolation:

$$g_{i,k} = \bar{x}_i^{T_{\text{obs}}} + T_f \Delta\bar{x}_i^{T_{\text{obs}}} + f_g(h_i, c) + L_g(h_i, c) z_k, \tag{8}$$

where f_g produces a deterministic goal offset and L_g maps the shared group-mode code z_k to an agent-specific goal offset.

At prediction step t, the realized interaction matrix A_k^t yields a normalized interaction message $\psi_{i,k}^t$, while the temporal mode amplitude a_k^t induces an agent-specific mode response $\rho_{i,k}^t$:

$$\begin{aligned} \psi_{i,k}^t &= \frac{\sum_j A_{ij,k}^t f_v\Big([s_{j,k}^{t-1}, \hat{y}_{j,k}^{t-1} - \hat{y}_{i,k}^{t-1}]\Big)}{\varepsilon_{\text{num}} + \sum_j A_{ij,k}^t}, \\ \rho_{i,k}^t &= f_r\big(u_{i,k}^t \odot \tanh(a_k^t)\big)\,, \end{aligned} \tag{9}$$

where $\varepsilon_{\text{num}} > 0$ ensures numerical stability.

Let $v_{i,k}^{g,t} = (g_{i,k} - \hat{y}_{i,k}^{t-1})/(T_f - t + 1)$ denote the goal-directed displacement per remaining prediction step. We collect the previous motion, agent context h_i, interaction message $\psi_{i,k}^t$, mode response $\rho_{i,k}^t$, remaining goal displacement, $v_{i,k}^{g,t}$, and normalized time $\tau_t = t/T_f$ into $\omega_{i,k}^t$. The goal-conditioned synchronous joint decoder updates

$$\begin{aligned} s_{i,k}^t &= \mathrm{GRU}_{\mathrm{dec}}(\omega_{i,k}^t, s_{i,k}^{t-1}), \\ \hat{y}_{i,k}^t &= \hat{y}_{i,k}^{t-1} + v_{i,k}^{g,t} + \gamma_t f_{\mathrm{out}}(s_{i,k}^t). \end{aligned} \tag{10}$$

The residual gate γ_t decays over the prediction horizon and satisfies $\gamma_{T_f} = 0$, ensuring that each predicted trajectory reaches $g_{i,k}$ at the final step. Since every update depends only on states from step $t-1$, all agents are decoded synchronously without imposing an agent-wise order.

3.6 Learning Objective

All trajectory losses use Euclidean distances in the canonical frame, with bars omitted for clarity. Let D_k and F_k denote the target-agent ADE and FDE of prior-sampled joint hypothesis k, respectively. We optimize a best-of-many objective and determine the mode assignment using the combined target-agent error:

$$\begin{aligned} \mathcal{L}_{\mathrm{BoM}} &= \min_k D_k + \lambda_{\mathrm{end}} \min_k F_k, \\ E_k &= D_k + \lambda_{\mathrm{end}} F_k, \qquad k^\star = \arg\min_k E_k. \end{aligned} \tag{11}$$

The trajectory and endpoint terms select their winners independently, whereas $k^\star$ determines the mode used for posterior reconstruction and classification.

To encourage mode specialization, we apply the classification loss $\mathcal{L}_{\mathrm{cls}} = -\log \pi_{k^\star}$ and an assignment curriculum:

$$\mathcal{L}_{\mathrm{assign}} = \frac{1}{J_n} \sum_{k \in \mathcal{W}_n} E_k - \min_k E_k, \tag{12}$$

where $\mathcal{W}_n$ contains the J_n hypotheses with the smallest E_k. We decrease J_n from 4 to 1 over the first 6,000 optimization steps, gradually concentrating supervision on the best-matching hypothesis.

The posterior reconstruction loss $\mathcal{L}_{\mathrm{post}}$ supervises the target trajectory and endpoint, while a dimension-wise free-bits loss $\mathcal{L}_{\mathrm{KL}}$ regularizes the posterior residual distribution toward the observation-conditioned prior. The auxiliary loss $\mathcal{L}_{\mathrm{joint}}$ supervises all valid agents, with prior winners selected over complete joint hypotheses independently of $k^\star$. The temporal regularizer $\mathcal{L}_{\mathrm{temp}}$ penalizes second-order changes in the mode amplitudes. The complete objective is

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\mathrm{BoM}} &+ \lambda_{\mathrm{post}} \mathcal{L}_{\mathrm{post}} + \beta(n) \mathcal{L}_{\mathrm{KL}} \\ &+ \lambda_{\mathrm{assign}} \mathcal{L}_{\mathrm{assign}} + \lambda_{\mathrm{cls}} \mathcal{L}_{\mathrm{cls}} \\ &+ \lambda_{\mathrm{temp}} \mathcal{L}_{\mathrm{temp}} + \lambda_{\mathrm{joint}} \mathcal{L}_{\mathrm{joint}}. \end{aligned} \tag{13}$$

Loss definitions, masking, normalization, and weights are specified in Appendix A.2; the KL annealing schedule is provided in Appendix A.1.

4 Experiments

Datasets. (a) The **ETH–UCY** benchmark (Pellegrini et al., 2009; Lerner et al., 2007) contains five pedestrian subsets: ETH, HOTEL, UNIV, ZARA1, and ZARA2, with trajectories represented in meters. Following the standard leave-one-out protocol, we train on four subsets and evaluate on the remaining one using $(T_{\mathrm{obs}}, T_f, \Delta t) = (8, 12, 0.4\,\mathrm{s})$. (b) The **Stanford Drone Dataset** (SDD) (Robicquet et al., 2016) comprises 60 aerial videos of heterogeneous agents annotated in pixels. We split the videos into 60%/20%/20% for training, validation, and testing, respectively, and use $(T_{\mathrm{obs}}, T_f, \Delta t) = (8, 12, 0.4\,\mathrm{s})$. (c) **NBA SportVU** (Linou et al., 2016) contains trajectories of ten players and the ball captured during professional basketball games. We construct 50K samples using a 65%/10%/25% train/validation/test split and set $(T_{\mathrm{obs}}, T_f, \Delta t) = (5, 10, 0.4\,\mathrm{s})$, corresponding to 2.0-s observation and 4.0-s prediction horizons. Coordinates are converted to meters, and evaluation is performed on the ten players.

Metrics. We evaluate stochastic trajectory prediction under the best-of-20 protocol. The reported ADE/FDE correspond to minADE_{20}/minFDE_{20}, with the minimizing hypothesis selected independently for each metric. In the joint analyses, Target ADE/FDE applies the same protocol to the target agent. Following Weng et al. (2023), Joint ADE and Joint FDE first aggregate errors over all valid agents within each hypothesis and then independently select the minimizing joint hypothesis. We additionally report the mean collision rate CR_{mean}, defined as the percentage of predicted agents whose path comes within 0.2 m of at least one other agent in the same joint hypothesis. It is averaged over agents and all 20 candidates without best-of selection. Lower values are better. Formal definitions are provided in Appendix A.3.

4.1 Comparisons with State-of-the-Art Methods

ETH–UCY. Table 1 shows that GroupModes achieves competitive accuracy against closely related methods. Compared with EigenTrajectory-LB-EBM, EqMotion, MART, and SingularTrajectory, it reduces average ADE by 19.05% and FDE by 9.38%–17.14%. It also improves upon DD-MDN by 10.53% ADE and 6.45% FDE. GroupModes matches E-V^2-Net's lowest average ADE of 0.17, while its FDE is 3.57% higher, indicating strong trajectory accuracy with a small remaining endpoint gap.

SDD. In Table 1, GroupModes achieves the lowest ADE and second-lowest FDE among the listed methods, at 6.36/10.28. It reduces ADE/FDE by 14.40%/13.03% relative to MART and by 11.54%/13.03% relative to DD-MDN. Compared with W2W, its ADE is 14.29% lower, while its FDE is 1.48% higher. These results indicate a stronger advantage in average displacement accuracy, with endpoint performance remaining close to the best reported result.

Table 1: Comparison on ETH–UCY and SDD under the best-of-20 protocol. Models observe 8 time steps (3.2 s) and predict the next 12 time steps (4.8 s). Results are reported as ADE/FDE in meters for ETH–UCY and pixels for SDD.

Models (ETH-UCY)	eth	hotel	univ	zara1	zara2	ETH-UCY	Models (SDD)	SDD
SingularTrajectory (Bae et al., 2024)	0.35 / 0.42	0.13 / 0.19	0.25 / 0.44	0.19 / 0.32	0.15 / 0.25	0.21 / 0.32	FlowChain (Maeda & Ukita, 2023)	9.93 / 17.17
EigenTrajectory-LB-EBM (Bae et al., 2023)	0.36 / 0.53	0.12 / 0.19	0.24 / 0.43	0.19 / 0.33	0.14 / 0.24	0.21 / 0.34	SHENet (Meng et al., 2022)	9.01 / 13.24
EqMotion (Xu et al., 2023)	0.40 / 0.61	0.12 / 0.18	0.23 / 0.43	0.18 / 0.32	0.13 / 0.23	0.21 / 0.35	LED (Mao et al., 2023)	8.48 / 11.36
Trajectron++ (Salzmann et al., 2020)	0.43 / 0.86	0.12 / 0.19	0.22 / 0.43	0.17 / 0.32	**0.12** / 0.25	0.21 / 0.41	MID (Gu et al., 2022)	7.91 / 14.50
MoFlow (Fu et al., 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	MoFlow (Fu et al., 2025)	7.50 / 11.96
MART (Lee et al., 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	MART (Lee et al., 2024)	7.43 / 11.82
DD-MDN (Hetzel et al., 2026)	0.30 / 0.46	0.13 / 0.17	0.23 / 0.41	0.16 / 0.28	**0.12 / 0.21**	0.19 / 0.31	W2W (Xu et al., 2026)	7.42 / **10.13**
AgentFormer (Yuan et al., 2021)	0.26 / 0.39	0.11 / **0.14**	0.26 / 0.46	**0.15 / 0.23**	0.14 / 0.24	0.18 / 0.29	DD-MDN (Hetzel et al., 2026)	7.19 / 11.82
ULPTP-ID (Wang et al., 2026)	0.26 / 0.40	0.11 / 0.16	0.20 / 0.35	0.18 / 0.31	0.14 / 0.24	0.18 / 0.29	V^2-Net (Wong et al., 2022)	7.12 / 11.39
V^2-Net-SC+ (Wong et al., 2024b)	**0.25** / 0.40	**0.10** / 0.15	0.24 / 0.43	0.18 / 0.28	0.13 / 0.22	0.18 / 0.29	E-V^2-Net (Xia et al., 2025)	6.57 / 10.49
E-V^2-Net (Xia et al., 2025)	**0.25 / 0.38**	0.11 / 0.16	0.20 / 0.34	0.19 / 0.30	0.13 / 0.24	**0.17 / 0.28**	ULPTP-ID (Wang et al., 2026)	6.47 / 10.30
GroupModes	0.29 / 0.44	0.12 / 0.18	**0.17 / 0.32**	0.16 / 0.28	**0.12** / 0.22	**0.17** / 0.29	GroupModes	**6.36** / 10.28

Table 2: Comparison on NBA under the best-of-20 protocol ($T_{\text{obs}} = 5$, $T_f = 10$). ADE/FDE are reported in meters at prediction horizons of 2.0 s (5 frames) and 4.0 s (10 frames).

Models (NBA)	Metrics@2.0 s	Metrics@4.0 s
PECNet (Mangalam et al., 2020)	0.96 / 1.69	1.83 / 3.41
MemoNet (Xu et al., 2022b)	0.71 / 1.14	1.25 / 1.47
V^2-Net (Wong et al., 2022)	0.69 / 0.96	1.28 / 1.68
E-V^2-Net (Xia et al., 2025)	0.68 / 0.93	1.26 / 1.64
E-V^2-Net-SC+ (Wong et al., 2024b)	0.65 / 0.86	1.14 / **1.37**
GroupNet+CVAE (Xu et al., 2022a)	0.62 / 0.95	1.13 / 1.69
Re (Wong et al., 2025)	0.60 / **0.78**	**1.12** / 1.38
GroupModes	**0.58** / 0.79	1.18 / 1.75

NBA. Table 2 shows that GroupModes achieves the best ADE and second-best FDE among the listed methods at 2.0 s. It improves ADE/FDE over E-V^2-Net-SC+ by 10.77%/8.14%, while achieving 3.33% lower ADE and 1.28% higher FDE than Re. At 4.0 s, its ADE is 6.35% lower than E-V^2-Net's, but its FDE is 6.71% higher. Relative to Re, ADE/FDE increase by 5.36%/26.81%.

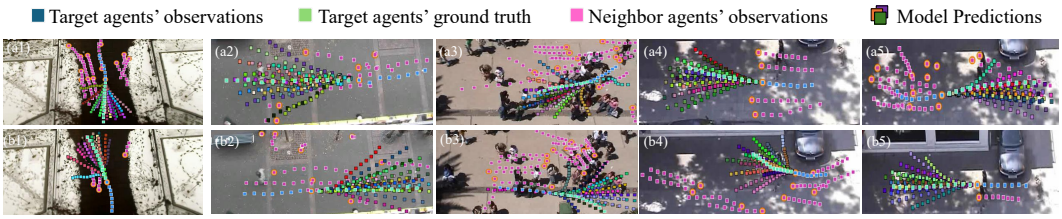

Figure 3: Qualitative trajectory predictions of GroupModes across different scenes in ETH–UCY.

Thus, GroupModes performs strongly at the shorter horizon, while long-horizon endpoint accuracy remains a limitation.

4.2 Qualitative Results

Figure 3 presents qualitative predictions on ETH–UCY, alongside the target agent's observed trajectory, ground-truth future, and neighboring agents' observations. The candidates exhibit different directions, endpoints, and degrees of spread across examples. Some predictions closely follow the ground-truth future, while others deviate in direction or travel distance, illustrating both the variation among candidate trajectories and the remaining prediction errors.

4.3 Ablation Studies and Quantitative Analyses

Analysis I: Validation of GroupModes. Table 3 shows that Conditional Gaussian incurs the largest ADE/FDE degradations: 47.06%/34.48% on ETH–UCY and 19.97%/30.93% on SDD. Fixed mode queries yield degradations of 41.18%/27.59% and 7.39%/10.02%, respectively, supporting the benefit of learning the queries within the bank. Removing mode residuals causes smaller degradations of 29.41%/10.34% and 3.30%/6.32%. The consistent ordering across datasets indicates that both learnable queries and stochastic residuals contribute to accuracy. Moreover, the residual-free bank still outperforms Conditional Gaussian, suggesting that the advantage of mode organization persists without within-mode stochastic variation. Appendix A.4.3 reports parameter counts and inference efficiency.

Table 3: Ablation of group-mode organization on ETH–UCY and SDD. No mode residual removes bounded mode residuals; Fixed mode queries freezes the initialized mode queries; Conditional Gaussian replaces the structured mode bank with a conditional Gaussian latent.

Variant	**ETH–UCY**	**SDD**
	ADE/FDE (Degradation ↓)	**ADE/FDE (Degradation ↓)**
GroupModes	0.17/0.29 (base)	6.36/10.28 (base)
No mode residual	0.22/0.32 (29.41%/10.34%)	6.57/10.93 (3.30%/6.32%)
Fixed mode queries	0.24/0.37 (41.18%/27.59%)	6.83/11.31 (7.39%/10.02%)
Conditional Gaussian	0.25/0.39 (47.06%/34.48%)	7.63/13.46 (19.97%/30.93%)

Analysis II: Goal–Interaction Mode Coupling. Table 4 examines whether agent-specific goals and temporal interactions should be conditioned on the same group mode. Goal-only conditioning increases target ADE/FDE by 23.53%/13.79% and joint ADE/FDE by 8.89%/8.33%, showing that mode-conditioned goals alone do not recover the full model's accuracy. Decoupling the goal and interaction modes leaves the reported target FDE unchanged but increases target ADE by 5.88% and joint ADE/FDE by 6.67%/3.57%. Similar target endpoint accuracy therefore does not imply equally accurate joint futures. Shared conditioning improves prediction accuracy across the jointly evaluated agents. Values of $\mathrm{CR}_{\mathrm{mean}}$ remain close across variants (8.81–8.93%), indicating that the

measured benefit of mode coupling primarily concerns target and joint displacement accuracy rather than collision reduction.

Table 4: Ablation of goal–interaction mode coupling on ETH–UCY. All variants use motion and position inputs. Goal-only conditioning removes direct mode input from the temporal branch, and Decoupled modes independently pair the goal and temporal modes.

Variant	**Target prediction**	**Joint prediction**	$\mathrm{CR}_{\mathrm{mean}}$ **(%)** $\downarrow$
	ADE/FDE (Degradation $\downarrow$**)**	**ADE/FDE (Degradation** $\downarrow$**)**	
Shared mode	0.17/0.29 (base)	0.45/0.84 (base)	8.81
Goal-only conditioning	0.21/0.33 ($23.53\%/13.79\%$)	0.49/0.91 ($8.89\%/8.33\%$)	8.93
Decoupled modes	0.18/0.29 ($5.88\%/0.00\%$)	0.48/0.87 ($6.67\%/3.57\%$)	8.90

Table 5: Effects of spatial input and cross-agent candidate pairing on ETH–UCY. Shuffling preserves each agent's candidate set but independently permutes the candidate indices of neighboring agents; results are averaged over 20 permutations.

Spatial input	**Pairing**	**Target prediction**	**Joint prediction**	$\mathrm{CR}_{\mathrm{mean}}$ **(%)** $\downarrow$
		ADE/FDE (Degradation $\downarrow$**)**	**ADE/FDE (Degradation** $\downarrow$**)**	
Motion + position	Original	0.17/0.29 (base)	0.45/0.84 (base)	8.81
Motion + position	Shuffled	0.17/0.29 ($0.00\%/0.00\%$)	0.47/0.90 ($4.44\%/7.14\%$)	12.41
Motion only	Original	0.20/0.32 ($17.65\%/10.34\%$)	0.47/0.87 ($4.44\%/3.57\%$)	8.91
Motion only	Shuffled	0.20/0.32 ($17.65\%/10.34\%$)	0.49/0.96 ($8.89\%/14.29\%$)	13.01

Analysis III: Spatial Context and Joint Candidate Pairing. Table 5 separates the effects of spatial context and cross-agent candidate correspondence. Under original pairing, adding position inputs reduces target ADE/FDE from 0.20/0.32 to 0.17/0.29 and joint ADE/FDE from 0.47/0.87 to 0.45/0.84, while $\mathrm{CR}_{\mathrm{mean}}$ changes only slightly from 8.91% to 8.81%. This result shows that spatial layout contributes directly to trajectory accuracy, addressing information that cannot be recovered from motion inputs alone. Independently shuffling neighbors' candidate indices preserves every agent's candidate set and therefore leaves target ADE/FDE unchanged. Nevertheless, joint ADE/FDE increases from 0.45/0.84 to 0.47/0.90 with position inputs and from 0.47/0.87 to 0.49/0.96 with motion-only inputs. The corresponding $\mathrm{CR}_{\mathrm{mean}}$ also rises by 3.60 and 4.10 percentage points, respectively. Because marginal candidates are unchanged in this control, the degradation provides direct evidence that the original mode indices encode meaningful cross-agent correspondence. The best overall results are obtained when spatial context is combined with the learned joint pairing, yielding more accurate joint hypotheses and a lower mean collision rate.

5 Conclusion

In this work, we introduced GroupModes, which uses a shared group-level mode to condition agent-specific goals and evolving social interactions. Position-aware encoding, a structured mode bank, low-rank interaction modulation, and synchronous decoding produce coordinated multimodal futures. Experiments on ETH–UCY, SDD, and NBA demonstrate competitive accuracy. Controlled ablations show that the shared mode also serves as a correspondence mechanism. Decoupling the goal and interaction modes preserves target endpoint accuracy but degrades joint prediction. Shuffling cross-agent candidates leaves target accuracy unchanged but worsens joint accuracy and increases the mean collision rate. These results indicate that the gains arise not only from marginal diversity or destination coverage, but also from organizing agents' futures into aligned scene-level hypotheses.

Limitations. GroupModes uses one shared mode per target–neighbor set. It neither models explicit subgroups nor enforces consistency between overlapping sets. Hierarchical modes could address these limitations.

AI USE STATEMENT

Generative AI tools assisted with literature discovery, manuscript polishing, feedback on experimental design and metric definitions, and interpretation of results. The authors reviewed the AI-assisted work and take full responsibility for the final manuscript, methods, and findings.

REPRODUCIBILITY STATEMENT

The model architecture and training objectives are detailed in Section 3. Section 4 and the Appendix describe the datasets, evaluation protocols, and implementation details. A link to the source code is provided in the paper.

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*, pp. 961–971, 2016.

Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Learning pedestrian group representations for multi-modal trajectory prediction. In *European Conference on Computer Vision*, pp. 270–289. Springer, 2022.

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, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17890–17901. IEEE, 2024.

Patrick Dendorfer, Sven Elflein, and Laura Leal-Taixé. Mg-gan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13138–13147. IEEE, 2021.

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.

Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. *arXiv preprint arXiv:2104.00563*, 2021.

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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2255–2264. IEEE, 2018.

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

Manuel Hetzel, Kerim Turacan, Hannes Reichert, Konrad Doll, and Bernhard Sick. DD-MDN: Human trajectory forecasting with diffusion-based dual mixture density networks and uncertainty self-calibration. *arXiv preprint arXiv:2602.11214*, 2026.

Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6271–6280. IEEE, 2019.

Zhiyu Huang, Haochen Liu, and Chen Lv. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3880–3890. IEEE, 2023.

Sungjune Kim, Hyung-gun Chi, Hyerin Lim, Karthik Ramani, Jinkyu Kim, and Sangpil Kim. Higher-order relational reasoning for pedestrian trajectory prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15251–15260. IEEE, 2024.

Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in neural information processing systems*, 32, 2019.

Seongju Lee, Junseok Lee, Yeonguk Yu, Taeri Kim, and Kyoobin Lee. Mart: Multiscale relational transformer networks for multi-agent trajectory prediction. In *European conference on computer vision*, pp. 89–107. Springer, 2024.

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

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.

Kostya Linou, Dzmitryi Linou, and Martijn de Boer. NBA Player Movements. GitHub repository, 2016. URL `https://github.com/linouk23/NBA-Player-Movements`.

Wenjie Luo, Cheol Park, Andre Cornman, Benjamin Sapp, and Dragomir Anguelov. Jfp: Joint future prediction with interactive multi-agent modeling for autonomous driving. In *Conference on Robot Learning*, pp. 1457–1467. PMLR, 2023.

Takahiro Maeda and Norimichi Ukita. Fast inference and update of probabilistic density estimation on trajectory prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9761–9771. IEEE, 2023.

Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European conference on computer vision*, pp. 759–776. Springer, 2020.

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.

Mancheng Meng, Ziyan Wu, Terrence Chen, Xiran Cai, Xiang Zhou, Fan Yang, and Dinggang Shen. Forecasting human trajectory from scene history. *Advances in Neural Information Processing Systems*, 35:24920–24933, 2022.

Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14412–14420. IEEE, 2020.

Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene transformer: A unified architecture for predicting multiple agent trajectories. *arXiv preprint arXiv:2106.08417*, 2021.

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*, pp. 261–268. IEEE, 2009.

Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pp. 549–565. Springer, 2016.

Luke Rowe, Martin Ethier, Eli-Henry Dykhne, and Krzysztof Czarnecki. Fjmp: Factorized joint multi-agent motion prediction over learned directed acyclic interaction graphs. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13745–13755. IEEE, 2023.

Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020.

Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 1349–1358. IEEE, 2019.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European conference on computer vision*, pp. 683–700. Springer, 2020.

Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8545–8556. IEEE, 2023.

Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8994–9003, 2021.

Qiao Sun, Xin Huang, Junru Gu, Brian C Williams, and Hang Zhao. M2i: From factored marginal trajectory prediction to interactive prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6533–6542. IEEE, 2022.

Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 international conference on robotics and automation (ICRA)*, pp. 7814–7821. IEEE, 2022.

Yunxi Wang, Xukai Zhai, Mingze Miao, and Zhishuai Yin. ULPTP-ID: A unified learning framework under imbalanced data distributions for pedestrian trajectory prediction. In *Intelligent Transportation Engineering*, volume 84 of *Advances in Transdisciplinary Engineering*, pp. 688–701. IOS Press, 2026. doi: 10.3233/ATDE251544.

Erica Weng, Hana Hoshino, Deva Ramanan, and Kris Kitani. Joint metrics matter: A better standard for trajectory forecasting. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20258–20269. IEEE, 2023.

Conghao Wong, Beihao Xia, Ziming Hong, Qinmu Peng, Wei Yuan, Qiong Cao, Yibo Yang, and Xinge You. View vertically: A hierarchical network for trajectory prediction via fourier spectrums. In *European Conference on Computer Vision*, pp. 682–700. Springer, 2022.

Conghao Wong, Beihao Xia, Ziqian Zou, Yulong Wang, and Xinge You. Socialcircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19005–19015. IEEE, 2024a.

Conghao Wong, Beihao Xia, Ziqian Zou, and Xinge You. Socialcircle+: Learning the angle-based conditioned interaction representation for pedestrian trajectory prediction. *arXiv preprint arXiv:2409.14984*, 2024b.

Conghao Wong, Ziqian Zou, and Beihao Xia. Resonance: Learning to predict social-aware pedestrian trajectories as co-vibrations. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25788–25799. IEEE, 2025.

Beihao Xia, Conghao Wong, Duanquan Xu, Qinmu Peng, and Xinge You. Another vertical view: A hierarchical network for heterogeneous trajectory prediction via spectrums. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):9749–9766, 2025.

Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6488–6497. IEEE, 2022a.

Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6478–6487. IEEE, 2022b.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1410–1420. IEEE, 2023.

Chenxin Xu, Yuxi Wei, Bohan Tang, Sheng Yin, Ya Zhang, Siheng Chen, and Yanfeng Wang. Dynamic-group-aware networks for multi-agent trajectory prediction with relational reasoning. *Neural Networks*, 170:564–577, 2024.

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

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.

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.

Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Conference on robot learning*, pp. 895–904. PMLR, 2021.

Yiyao Zhu, Di Luan, and Shaojie Shen. Biff: Bi-level future fusion with polyline-based coordinate for interactive trajectory prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8226–8237. IEEE, 2023.

A Appendix

A.1 Implementation Details

Training Details. GroupModes is implemented in PyTorch and trained on a single NVIDIA RTX 4090D GPU. We use hidden and latent dimensions of 128 and 16, an interaction rank of 16, and 20 modes with $K = 20$ hypotheses during training and validation. Training uses Adam with a batch size of 64, a learning rate of 5×10^{-5}, weight decay of 10^{-4}, gradient clipping at 1.0, and mixed precision. The maximum training budget is 50 epochs. ETH–UCY uses a plateau learning-rate scheduler and early stopping with a patience of 20 epochs. SDD and NBA use a 1,000-step warm-up followed by cosine decay, with NBA additionally capped at 15,000 optimization steps. The free-bits threshold is 0.025 per latent dimension. The KL weight remains zero for 4,000 steps and then increases linearly to 0.01 over 8,000 steps. Other loss weights are given in Appendix A.2.

All reported runs use training seed 42 and select the checkpoint with the lowest validation minADE at $K = 20$. Validation sampling seeds are fixed at 20260810 for ETH–UCY and 20260813 for SDD and NBA. Within each dataset, all ablation variants share the training budget, learning-rate schedule, and stopping criteria.

Position-aware observation encoding. A two-layer MLP with GELU maps each agent's last valid target-centric position to the motion-feature dimension. The resulting embedding is added to the motion representation before masked set encoding. Observation and agent masks identify the last valid position and exclude padded agents, respectively. The *Motion only* variant in Table 5 removes this embedding.

A.2 Loss Details

Trajectory errors and masking. For a fully observed target future, the trajectory and endpoint errors are

$$D_k = \frac{1}{T_f} \sum_{t=1}^{T_f} \big\| \hat{y}^t_{i^\star,k} - y^t_{i^\star} \big\|_2, \qquad F_k = \big\| \hat{y}^{T_f}_{i^\star,k} - y^{T_f}_{i^\star} \big\|_2. \tag{14}$$

For partially observed futures, D_k is averaged over valid future steps, and F_k uses the last valid future position. Unless otherwise specified, scene-wise losses are averaged over the minibatch. The discrete assignment $k^\star$ is treated as a fixed label during backpropagation. The assignment curriculum in Eq. 12 becomes zero when $J_n = 1$.

Posterior reconstruction and KL regularization. We sample $\epsilon^q \sim q_\phi(\epsilon \mid X, Y)$ and construct $z^q_{k^\star} = m_{k^\star}(c) + \alpha \tanh(\epsilon^q)$. The posterior prediction is supervised using $\mathcal{L}_{\text{post}} = \mathcal{L}^{\text{traj}}_{\text{post}} + \lambda_{\text{post,end}} \mathcal{L}^{\text{end}}_{\text{post}}$, with the same target-agent mask and Euclidean errors as the prior predictions. For a minibatch of size B, the free-bits KL loss is

$$\overline{\text{KL}}_l = \frac{1}{B} \sum_{b=1}^{B} D_{\text{KL}}\big(q_\phi(\epsilon_l \mid X_b, Y_b) \,\|\, p_\theta(\epsilon_l \mid X_b)\big), \qquad \mathcal{L}_{\text{KL}} = \sum_{l=1}^{d_z} \max\big\{\tau_{\text{fb}}, \overline{\text{KL}}_l\big\}. \tag{15}$$

Here, $\tau_{\text{fb}} = 0.025$. Free bits are applied to the batch-averaged KL divergence of each latent dimension before summation.

Joint supervision and temporal regularization. Let D^{all}_k denote the trajectory error averaged over all valid agent–time pairs, and let F^{all}_k denote the endpoint error averaged over agents with valid future observations. The joint loss is

$$\mathcal{L}_{\text{joint}} = \lambda_{\text{post}} \big(D^{\text{all}}_q + \lambda_{\text{post,end}} F^{\text{all}}_q\big) + \min_k D^{\text{all}}_k + \lambda_{\text{end}} \min_k F^{\text{all}}_k, \tag{16}$$

where q denotes the single posterior reconstruction. Each prior minimum selects one complete joint hypothesis for the valid agent set; the trajectory and endpoint winners are independent of each other and of $k^\star$.

For temporal regularization, we square the components of $a_k^{t+1} - 2a_k^t + a_k^{t-1}$ and average over the batch, hypothesis, temporal, and amplitude dimensions within each branch. The prior and posterior branch losses are then averaged with equal weight.

Loss weights. We use $\lambda_{\text{end}} = \lambda_{\text{post,end}} = 1$, $\lambda_{\text{post}} = 0.2$, $\lambda_{\text{assign}} = 0.25$, $\lambda_{\text{cls}} = 0.02$, $\lambda_{\text{temp}} = 0.005$, and $\lambda_{\text{joint}} = 0.1$.

A.3 Evaluation Metrics

Target-agent metrics. We evaluate trajectory prediction accuracy using average displacement error (ADE) and final displacement error (FDE). Let $i^\star$ denote the target agent, and let $\hat{y}^t_{i^\star,k}$ and $y^t_{i^\star}$ denote its predicted and ground-truth positions at future time step t, respectively. For hypothesis k, the two errors are defined as

$$\text{ADE}^{(k)} = \frac{1}{T_f}\sum_{t=1}^{T_f} \left\|\hat{y}^t_{i^\star,k} - y^t_{i^\star}\right\|_2, \qquad \text{FDE}^{(k)} = \left\|\hat{y}^{T_f}_{i^\star,k} - y^{T_f}_{i^\star}\right\|_2. \tag{17}$$

ADE measures prediction error over the full forecast horizon, whereas FDE measures endpoint error.

For multimodal prediction, we report the minimum error among $K = 20$ hypotheses:

$$\text{minADE}_K = \min_{1\le k\le K} \text{ADE}^{(k)}, \qquad \text{minFDE}_K = \min_{1\le k\le K} \text{FDE}^{(k)}. \tag{18}$$

The minima are computed independently for each evaluation instance and then averaged over the evaluation set. The hypotheses attaining minADE and minFDE may therefore differ. Unless otherwise specified, ADE/FDE in the main benchmark tables denote target-agent best-of-20 errors.

Joint prediction metrics. Joint metrics evaluate each complete hypothesis over all valid agents. Let $m_i^t \in \{0,1\}$ indicate a valid future position, $\mathcal{V} = \{i : \sum_t m_i^t > 0\}$, and $t_i^{\text{last}} = \max\{t : m_i^t = 1\}$. For hypothesis k, we define

$$\begin{aligned}\text{JADE}^{(k)} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_f} m_i^t \left\|\hat{y}^t_{i,k} - y^t_i\right\|_2}{\sum_{i=1}^{N}\sum_{t=1}^{T_f} m_i^t},\\ \text{JFDE}^{(k)} &= \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \left\|\hat{y}^{t_i^{\text{last}}}_{i,k} - y^{t_i^{\text{last}}}_i\right\|_2.\end{aligned} \tag{19}$$

We report

$$\text{JointADE} = \min_{1\le k\le K} \text{JADE}^{(k)}, \qquad \text{JointFDE} = \min_{1\le k\le K} \text{JFDE}^{(k)}, \tag{20}$$

with $K = 20$. Each metric selects one shared hypothesis for the entire valid agent set; the ADE and FDE winners may differ. Scene-level scores are aggregated using the same convention as the target-agent metrics.

Mean collision rate. Following Weng et al. (2023), we report the mean collision rate over all joint candidates. We evaluate collisions between predicted agents within the same joint candidate. Let $\hat{y}_{i,k}(\tau)$ denote the piecewise-linear interpolation of agent i's k-th predicted trajectory. For agents i and j, $\mathcal{U}_{ij}$ contains their common valid future timestamps and the intervals between consecutive timestamps at which both agents are valid. We define

$$C_{i,k} = \mathbf{1}\left[\exists j \in \mathcal{V}\setminus\{i\},\ \exists \tau \in \mathcal{U}_{ij} : \left\|\hat{y}_{i,k}(\tau) - \hat{y}_{j,k}(\tau)\right\|_2 < 2r_{\text{col}}\right], \tag{21}$$

where $\mathcal{V}$ is the set of valid evaluated agents. An agent is counted once per candidate if it collides with at least one other agent at any time. The mean collision rate, expressed as a percentage, is

$$\text{CR}_{\text{mean}} = \frac{100}{K|\mathcal{V}|}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}} C_{i,k}. \tag{22}$$

We use $K = 20$ and a pedestrian radius of $r_{\text{col}} = 0.1$ m on ETH–UCY. All candidates are evaluated without ground-truth-based selection; lower values are better.

A.4 Additional Analyses

A.4.1 Temporal Realization and Limited Observations

Table 6: Ablation of mode realization on ETH–UCY and SDD. The Static interactions variant reuses the first-step interaction weights throughout decoding.

Variant	ETH–UCY	SDD
	ADE/FDE (Degradation ↓)	ADE/FDE (Degradation ↓)
GroupModes	0.17/0.29 (base)	6.36/10.28 (base)
Static interactions	0.22/0.34 (29.41%/17.24%)	6.87/11.43 (8.02%/11.19%)

Table 6 complements the main mode-coupling analysis by examining whether social interactions should evolve throughout prediction. Reusing the first-step interaction weights increases ADE/FDE by 29.41%/17.24% on ETH–UCY and 8.02%/11.19% on SDD, supporting the benefit of temporally updating the interaction realization rather than treating it as static.

Table 7: Robustness under limited observation histories on ETH–UCY. Both models are trained with eight observed steps and evaluated without retraining using the most recent 8, 4, or 2 steps.

Observation history	GroupModes	Conditional Gaussian
	ADE/FDE (Degradation ↓)	ADE/FDE (Degradation ↓)
Full ($T_{\text{obs}} = 8$)	0.17/0.29 (base)	0.25/0.39 (base)
Half ($T_{\text{obs}} = 4$)	0.19/0.32 (11.76%/10.34%)	0.27/0.41 (8.00%/5.13%)
Quarter ($T_{\text{obs}} = 2$)	0.19/0.33 (11.76%/13.79%)	0.29/0.43 (16.00%/10.26%)

Table 7 further evaluates sensitivity to limited motion evidence. GroupModes maintains lower absolute ADE/FDE than Conditional Gaussian with 8, 4, and 2 observed steps. Under the most severe truncation, its relative ADE degradation is also lower (11.76% versus 16.00%), although its FDE degradation remains higher. These results indicate that the advantage of structured shared modes persists under shortened histories, while endpoint prediction remains sensitive to severe observation truncation.

A.4.2 Density Analysis

Table 8 examines whether the benefit of structured modes persists across different neighborhood densities, measured by the target's mean number of observed neighbors within 2 m. GroupModes achieves lower ADE/FDE than Conditional Gaussian in all three density groups. Replacing the structured mode bank with a conditional Gaussian latent increases ADE/FDE by 50.00%/50.00% in sparse neighborhoods ($\bar{N} < 2$), by 75.00%/55.56% in moderately populated neighborhoods ($2 \leq \bar{N} \leq 5$), and by 72.22%/34.38% in crowded neighborhoods ($\bar{N} > 5$). The magnitude of the difference varies across metrics and density levels, while the consistent accuracy advantage shows that the benefit of structured modes is not confined to crowded scenes.

Table 8: Density-conditioned analysis on ETH–UCY. $\bar{N}$ denotes the target's mean number of observed neighbors within 2 m. For each density stratum and the overall evaluation, ADE/FDE are first computed within each subset and then averaged equally across the five subsets.

Neighbors	GroupModes	Conditional Gaussian
	ADE/FDE	ADE/FDE (Degradation) ↓
Overall (100%)	0.17 / 0.29	0.25 / 0.39 (47.06%/34.48%)
$\bar{N} < 2$ (51.7%)	0.18 / 0.30	0.27 / 0.45 (50.00%/50.00%)
$2 \leq \bar{N} \leq 5$ (37.6%)	0.16 / 0.27	0.28 / 0.42 (75.00%/55.56%)
$\bar{N} > 5$ (10.7%)	0.18 / 0.32	0.31 / 0.43 (72.22%/34.38%)

A.4.3 PARAMETERS AND EFFICIENCY

We compare the inference speed and number of parameters of different models in Table 9. All models are evaluated on a single NVIDIA GeForce RTX 4090D GPU with a batch size of 1. For methods whose original implementations are not directly compatible with our evaluation environment, we reproduce their inference pipelines in PyTorch and mark them with "*" in the table. The inference time of GroupModes includes the generation of $K = 20$ joint trajectory hypotheses.

Table 9: Comparison of prediction accuracy, inference time, and model parameters on ETH–UCY.

Models	**ADE/FDE**	**Time**	**Params.**
EigenTrajectory-LB-EBM (Bae et al., 2023)	0.21 / 0.34	5.6 ms	2.34M
MART (Lee et al., 2024)	0.21 / 0.33	10.2 ms	1.54M
SingularTrajectory (Bae et al., 2024)	0.21 / 0.32	12.6 ms	1.93M
V^2-Net-SC+ (Wong et al., 2024b)	0.18 / 0.29	29 ms	1.92M
E-V^2-Net* (Xia et al., 2025)	0.17 / 0.28	21 ms	1.92M
GroupModes	0.17 / 0.29	26.5 ms	1.47M

GroupModes achieves an ADE/FDE of 0.17/0.29 with 1.47M parameters. It has the smallest parameter count among the compared methods, matches the best ADE, and remains within 0.01 FDE of the best result. Although its inference latency is higher, the absolute cost remains moderate while producing complete joint trajectory hypotheses. These results demonstrate a favorable balance between prediction accuracy and model compactness.