BeTop-ens: Ensemble Behavioral Topology Reasoning for Interactive Prediction

Haochen Liu[†], Li Chen[◇], Hongyang Li[◇], Chen Lv[†]
Nanyang Technological University[†], The University of Hong Kong[◇]
haochen002@e.ntu.edu.sg, ilnehc@connect.hku.hk
hongyang@hku.hk, lyuchen@ntu.edu.sg

Abstract

Accurately forecasting the future behaviors of interactive traffic participants is a critical capability for autonomous driving systems. While interactive prediction offers an effective paradigm for modeling multi-modal futures of agent pairs, it remains challenging to ensure consistency between interacting agents and to combat with uncertainty across diverse behavioral modes. In this work, we propose ***BeTop-ens****, an ensemble-based behavioral topology reasoning framework that extends the BeTop [6] formulation for robust interactive motion prediction. At the core of BeTop-ens is BeTopNet, a synergistic learning model that represents interactive behaviors through a hierarchical topology graph. This graph captures both low-level motion features and high-level behavioral modes, enabling structured reasoning over future interactions. BeTopNet employs a dual-branch network architecture to separately model local spatio-temporal motion cues and global interaction semantics, which are fused via cross-attention for joint future prediction. To address epistemic uncertainty and improve robustness, BeTop-ens aggregates multiple independently trained BeTopNet models using a refined ensemble learning strategy. This not only enhances the calibration of predictive uncertainty, but also yields notable improvements in the accuracy and consistency of interactive forecasting.*

1. Introduction

Understanding and forecasting the interactive behaviors of multiple traffic agents is a fundamental challenge for autonomous driving [2]. Accurate multi-agent prediction requires not only modeling the individual dynamics of each actor but also reasoning over their complex interactions, potential future intentions, and diverse behavioral outcomes. Existing methods often struggle to maintain consistency across interacting agents and to capture the uncertainty inherent in socially plausible futures.

A wide range of prior arts have explored architectural designs for interactive motion prediction. A common approach is to use social pooling or graph-based interaction to aggregate context from neighboring agents. These works represent interactions Through goal anchors [4] or message passing [10], it enables mutual influence among agents. However, these methods often lack structured reasoning over joint behavior modes, leading to inconsistencies in predicted futures between interacting agents. Another line employs transformer-based models or scene-centric encoders [11] . These architectures fuse spatial and temporal information in a data-driven manner to capture interactions at scale. Despite their expressive capacity, they typically rely on dense attention mechanisms without explicitly modeling behavioral dependencies. Relation modeling, in contrast, explicitly captures future behavioral interactions using mechanisms such as attention [8], game theory [5], or braided structures [9]. However, these methods struggle to scale effectively in capturing high-level social semantics due to significant computational overhead. Another challenge lies in the epistemic uncertainty, which stems from model ambiguity. This can result in overconfident predic-

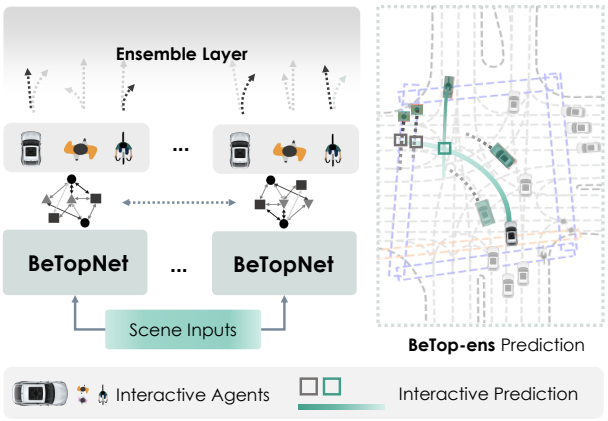

Figure 1. **BeTop-ens Overview:** We propose an ensemble framework that couples the interactive motion prediction task with multi-agent behavior topology (BeTop) reasoning. The method is established upon our preceding work [6]. Our framework reduces uncertainty, and delivers compliant interactive predictions.

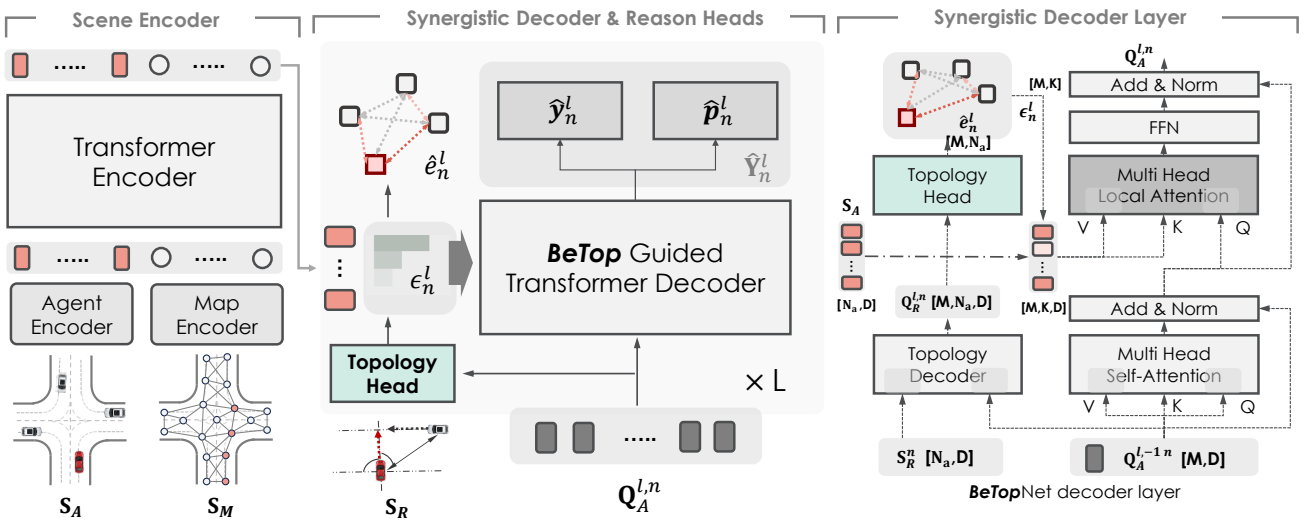

Figure 2. **The BeTopNet Architecture.** BeTop establishes an integrated network for topological behavior reasoning. Scene encoder generates scene-aware attributes for agent $\mathbf{S}_A$ and map $\mathbf{S}_M$. Initialized by $\mathbf{S}_R$ and $\mathbf{Q}_A$, synergistic decoder reasons edge topology $\hat{e}_n^l$ and trajectories $\hat{\mathbf{Y}}_n^l$ iteratively from topology-guided local attention.

tions in ambiguous or safety-critical scenarios.

In this study, we aimed to address these challenges through an ensemble learning framework upon our preceding interactive formulation, termed as **BeTop-ens**. BeTop introduce a behavioral topology reasoning framework that formulates multi-agent interactions as a structured graph of latent behavioral modes. While BeTop improves consistency and interpretability, it does not explicitly address epistemic uncertainty, which arises from rare interaction patterns. Hence, an ensemble extension of the BeTop framework are designed to enhance uncertainty modeling. At its core is BeTopNet, a dual-branch neural architecture that captures both local motion dynamics and global interaction semantics through a hierarchical topology graph. BeTop-ens constructs a diverse set of BeTopNet instances trained with different initialization and learning dynamics. This ensemble strategy enables the model to better represent the distribution over plausible futures. Evaluated on WOMD interactive prediction benchmarks, BeTop-ens achieves improved accuracy and consistency in interactive prediction.

2. Method

2.1. BeTop Formulation

We leverage the braid theory [1], which probes explicit formulations for compliant multi-agent interactions from future data $\mathbf{Y}_{1:N_a}$. Intuitively, it denotes a transform process for $\mathbf{Y}_{1:N_a}$ with respective agent coordinates, and then gathers each future forward intertwine (occupancy) as joint interactions. Formally, consider the braid group $\mathbf{B}_{N_a} = \{\sigma_n\}$ by N_a primitive braids σ_n, each of which $\sigma_n = (f_1^n, \cdots, f_{N_a}^n)$ denotes a tuple of monotonically increased functions $f : \mathbb{R}^3 \times \mathbf{Y} \to \mathbb{R}^2 \times I$ mapping from Cartesian $(\vec{x}, \vec{y}, \vec{t})$ to lateral coordinate $(\vec{y}, \vec{t})$ for agent future $\mathbf{Y}$. Specifically, the function f_i^n in σ_n is defined as $f_i^n \to (\mathbf{Y}_i - \mathbf{b}_n)\mathbf{R}_n; 1 \le i, n \le N_a$, where $\mathbf{b}_n$ and $\mathbf{R}_n$ denote the left-hand transform matrix to local coordinate of agent A_n. The joint interactive behaviors are identified as a set of braids having intertwines $\{\sigma_n^{\pm}\} \subset \mathbf{B}_{N_a}$ over others [9]. The goal of BeTop is to reason a topological graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for multi-agent future behaviors. Expressly, node topology $\mathcal{V} = \{\mathbf{Y}_n\}$ is denoted by multi-agent future trajectories. We can then reformulate the braid set $\{\sigma_n^{\pm}\}$ as an edge topology $e_{ij} \to \mathcal{E} \in \mathbb{R}^{N_a \times N_a}; 1 \le i, j \le N_a$ for future interactive behaviors. Each topology element e_{ij} can be defined by two braid functions $f_i^i, f_j^i \in \sigma_i$ assessing the future interwines along with $\mathbf{Y}_i, \mathbf{Y}_j$ as: $e_{ij} = \max_t \mathbf{I}\left(f_i^i(\mathbf{y}_i^t), f_j^i(\mathbf{y}_j^t)\right)$. Here $\mathbf{I}$ is an intertwine indicator by segment intersection under lateral coordinates. We can formulate the reasoning task as: $\mathcal{G}^* = (\max \hat{\mathcal{V}}, \max \hat{\mathcal{E}})$. Agent future $\hat{\mathbf{Y}}$ in node term $\hat{\mathcal{V}}$ is defined by Gaussian mixtures (GMM). The edge reasoning $\hat{\mathcal{E}}$ can be specified as minimizing the binary cross entropy (BCE):

$$\max \hat{\mathcal{E}} = \min \sum_i \sum_j \mathrm{BCE}(e_{ij}, \hat{e_{ij}}), \tag{1}$$

where $1 \le i, j \le N_a$. Synergistic reasoning structures are then established optimizing $\mathcal{G}^*$.

2.2. BeTopNet

As presented in Fig. 2, we introduce the synergistic learning framework reasoning BeTop. It encompasses a Transformer backended encoder-decoder network. With encoded scene semantics $\mathbf{X}; \mathbf{M}$, the proposed network fea-

tures a synergistic decoder which reasons and guides BeTop. Reason heads for topology $\hat{\mathcal{E}}$ and Prediction for $\hat{\mathcal{V}}$ comprise the behavioral graph $\mathcal{G}$.

2.3. Scene Encoder

We leverage a scene-centric coordinate system [7], it comprises historical agent states $\mathbf{X} \in \mathbb{R}^{N_a \times T_h \times D_a}$ and map polyline inputs $\mathbf{M} \in \mathbb{R}^{N_m \times L_m \times D_m}$, where we portion N_m map segments with length L_m from full scene map. Both attributes are encoded separately as $\mathbf{S}_A \in \mathbb{R}^{N_a \times D}$ and $\mathbf{S}_M \in \mathbb{R}^{N_m \times D}$ and concatenated as scene features $\mathbf{S} = [\mathbf{S}_A; \mathbf{S}_M] \in \mathbb{R}^{(N_a+N_m)\times D}$. A stack of Transformer encoders with local attention are directly employed from encoded scene semantics $\mathbf{S}_A, \mathbf{S}_M$.

2.4. Synergistic Decoder

Retaining encoded scene features $\mathbf{S}_A, \mathbf{S}_M$, we zoom in the decoding strategy. We introduce the iterative process of N Transformer decoder layers contributed to all agents, pursuing the basis from [11]. To iron out the scene uncertainties, a multi-modal set of M decoding queries $\mathbf{Q}_A^0 \in \mathbb{R}^{M \times D}$ are initialized for multi-agent future trajectories. Meanwhile, relative attributes $\mathbf{S}_R \in \mathbb{R}^{N_a \times N_a \times D_R}$ are deployed through MLPs as topology features $\mathbf{Q}_R^0 \in \mathbb{R}^{N_a \times N_a \times D}$ for edge topology reasoning.

Next, we devise dual streams to the iterative decoding process for $\hat{\mathcal{V}}$ of future trajectories and $\hat{\mathcal{E}}$ of future topology. Given agent A_n, the decoding process in layer l follows:

$$\begin{aligned} \mathbf{Q}_R^{l,n} &= \texttt{TopoDecoder}\left(\mathbf{Q}_A^{l-1,n}, \mathbf{Q}_R^{l-1,n}, \mathbf{S}_A\right), \\ \hat{e}_n^l &= \texttt{TopoHead}\left(\mathbf{Q}_R^{l,n}\right), \end{aligned} \tag{2}$$

where both future trajectories $\hat{\mathbf{Y}}_n \in \hat{\mathcal{V}}$ and interactive topology $\hat{e}_n \in \hat{\mathcal{E}}$ in BeTop are decoded in parallel:

$$\begin{aligned} \mathbf{Q}_A^{l,n} &= \texttt{Decoder}\left(\mathbf{Q}_A^{l-1,n}, \mathbf{S}_A, \mathbf{S}_M, \hat{\mathbf{Y}}_n^{l-1}, \hat{e}_n^l\right), \\ \hat{\mathbf{Y}}_n^l &= \texttt{Head}(\mathbf{Q}_A^{l,n}), \end{aligned} \tag{3}$$

Reasoned edge topology $\hat{e}_n^l \in \mathbb{R}^{M \times N_a}$ are garnered by topological decoder with query broadcasting $\mathbf{Q}_A^{l-1,n}$; Reasoning nodes for $\hat{\mathbf{Y}}_n$, a Transformer decoder with topology-guided local attention are drafted serving $\hat{e}_n^l$ as priors.

2.5. Topology-Guided Local Attention

Querying whole-scene agent semantics results in misaligned interactive agents and sparse attention. This motivates our design for local attention guided by the reasoned topology $\hat{e}_n^l \in \mathbb{R}^{M \times N_a}$ as priors. Specifically, we retrieve the top-K index $\epsilon_n^l \in \mathbb{R}^{M \times K}$ priored from $\hat{e}_n^l$ for eventual interactive agents behaviors with A_n. Interactive indices are directly leveraged in gathering $\mathbf{S}_A$ selectively for local cross-attention. This process is formed as:

$$\begin{aligned} \mathbf{C}_A^{l,n} =& \texttt{TopoAttn}\left(\mathbf{Q}_A^{l-1,n}, \mathbf{S}_A, \hat{e}_n^l\right) \rightarrow \\ & \texttt{MHA}\left(q = \mathbf{Q}_A^{l-1,n}; k, v = \mathbf{S}_A^{i \in \epsilon_n^l}\right), \end{aligned} \tag{4}$$

where $\epsilon_n^l = \arg\max_K(\hat{e}_n^l)$. Topology-guided agent features $\mathbf{C}_A^{l,n}$ are then aggregated in each layer.

Reason heads Given respective decoding features $\mathbf{Q}_R^{l,n}$ and $\mathbf{Q}_A^{l,n}$ for each layer, we affix reason heads accustomed to corresponding formulations for $\hat{e}_n$ and $\hat{\mathbf{Y}}_n$. Referred in Eq. (2), the topology head and prediction head are jointly devised by stacked MLPs in reasoning BeTop results. For agent A_n in each layer, reason heads decode GMM components of future states $\hat{\mathbf{y}}_n \in \mathbb{R}^{M \times T_f \times 5}$ (referring to $(\mu_x, \mu_y, \log\sigma_x, \log\sigma_y, \rho)$ per step) with mixture score $\hat{\mathbf{p}}_n \in \mathbb{R}^M$,$\{\hat{\mathbf{y}}_n, \hat{\mathbf{p}}_n\} \in \hat{\mathbf{Y}}_n$, as well as interactive edge topology $\hat{e}_n^l \in \mathbb{R}^{M \times N_a}$ for BeTop.

2.6. Learning and Ensemble Strategy

Learning objectives: We start by pursuing the target in Sec. 2.1. Objectives are firstly established in regulating multi-agent behavioral states $\{\hat{\mathbf{Y}}_n\} \subset \hat{\mathcal{V}}$ while maximizing their interactive distributions $\hat{\mathcal{E}}$. The imitative objective for $\hat{\mathbf{Y}}$ is defined by the negative log-likelihood (NLL) from best-reasoned components m^* closest to ground-truths, as denoted: $\mathcal{L}_{\mathcal{V}} = \sum_t^{T_f} \mathcal{L}_{\text{NLL}}(\hat{\mathbf{y}}_n^{m^*,t}, \hat{\mathbf{p}}_n^{m^*}, \mathbf{Y}_n)$. The topology are computed by BCE given gathered $\hat{e}_n^{m^*} \in \mathbb{R}^{N_a}$, formulated as $\mathcal{L}_{\mathcal{E}} = \sum_j^{N_a} \mathcal{H}(\hat{e}_{n,j}^{m^*}, e_{n,j})$ over $N_a = 2$ interactive agents jointly.

Ensemble Strategy: Given F BeTopNet ensemble with different configurations, we firstly derive interactive modalities by top-$M^{(i)}$ pair-wise joint scoring for each BeTopNet: $P_J^{(i)} = \max_{M^{(i)}} \prod_n^{N_I} \hat{\mathbf{p}}_n^{(i)}$, where $i \in F$. BeTop-ens then applies non-maximum suppression (NMS) strategy [11] to retrieve the top-M joint modalities from the concatenated ensemble joint scores $P_J^{(i)}$ produced by each model:

$$\hat{\mathbf{Y}}_J, P_J = \arg\max_M \texttt{NMS}([P_J^{(1)}; \cdots; P_J^{(F)}]) \tag{5}$$

where $[;]$ denotes concatenation, and $\hat{\mathbf{Y}}_J$ terms for the final interactive prediction outputs.

2.7. Implementation Details

BeTop-ens is deviced by $F = 5$ BeTopNet variants, where each model are configured following our preceding work [6], varying by $L \in [2, 4, 6]$, $D \in [256, 512]$ and $M \in [6, 64]$. Each BeTopNet in BeTop-ens is trained from scratch by the WOMD [3] training set without augmentations. We use a distributed training strategy on 8 A100 GPUs with batch size of 256. AdamW optimizer is used with learning rate as 1e-4. Training epochs are set to 30.

Table 1. Testing performance on the Waymo 2025 Interaction Prediction Leaderboard

Method Name	minADE ↓	minFDE ↓	Miss Rate ↓	Overlap Rate ↓	mAP v2 ↑	**Soft-mAP** ↑
Parallel ModeSeq	0.7707	1.6897	0.3782	0.1896	0.2949	0.2978
IMPACT [13]	0.9738	2.2734	0.4316	0.1684	0.2659	0.2718
RetroMotion (SMoE hybrid)	0.9256	2.0890	0.4347	0.1927	0.2519	0.2562
RMP YOLO [12]	0.9274	2.1131	0.4167	0.1695	0.2313	0.2486
AutoDiffuser-Draft	0.9422	2.1759	0.4885	0.1636	0.2211	0.2249
infgen-base-xl	1.2059	2.6427	0.5649	0.2402	0.0714	0.1236
Waymo LSTM baseline [3]	1.9056	5.0278	0.7750	0.3407	0.0648	0.0693
BeTop (2024) [6]	0.9744	2.2744	0.4355	0.1696	0.2412	0.2466
BeTop-ens	0.9779	2.2805	0.4376	0.1688	0.2511	0.2573

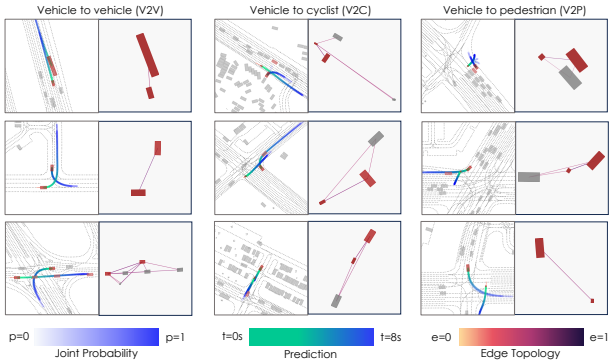

Figure 3. Qualitative results on WOMD Validation Interactive Sets. Our method could accurately reason the annotated interactive pairs, while delivering consistent joint motion predictions.

3. Result and Conclusion

Table 1 presents the quantitative results on the 2025 WOMD Interaction Prediction Leaderboard. BeTop-ens achieves a competitive balance across all metrics. Specifically, it obtains a minADE of 0.9779 and a minFDE of 2.2805, which are on par with other leading approaches such as IMPACT and RetroMotion. Notably, BeTop-ens achieves better mAP score (0.2511) than diffusion-based models like AutoDiffuser-Draft (0.2211) and large-scale transformers such as infgen-base-xl (0.0714). Qualitative results in Fig. 3 also highlight the strength of BeTop-ens in reasoning accurate and deiverse interaction patterns

We propose BeTop-ens, an ensemble reasoning model of multi-agent behavioral topology for interactive prediction. Through model ensembling, our method reports a solid improvement in consistent interactive prediction.

References

[1] Emil Artin. Theory of braids. *Annals of Mathematics*, 1947. 2

[2] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE TPAMI*, 2024. 1

[3] Scott Ettinger et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *ICCV*, pages 9710–9719, 2021. 3, 4

[4] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *ICCV*, pages 15303–15312, 2021. 1

[5] Zhiyu Huang, Haochen Liu, and Chen Lv. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *ICCV*, 2023. 1

[6] Haochen Liu, Li Chen, Yu Qiao, Chen Lv, and Hongyang Li. Reasoning multi-agent behavioral topology for interactive autonomous driving. In *NeurIPS*, 2024. 1, 3, 4

[7] Haochen Liu, Zhiyu Huang, Wenhui Huang, Haohan Yang, Xiaoyu Mo, and Chen Lv. Hybrid-prediction integrated planning for autonomous driving. *IEEE TPAMI*, 2025. 3

[8] Wenjie Luo, Cheol Park, Andre Cornman, Benjamin Sapp, and Dragomir Anguelov. Jfp: Joint future prediction with interactive multi-agent modeling for autonomous driving. In *CoRL*, pages 1457–1467. PMLR, 2023. 1

[9] Christoforos Mavrogiannis, Jonathan A DeCastro, and Siddhartha S Srinivasa. Abstracting road traffic via topological braids: Applications to traffic flow analysis and distributed control. *IJRR*, 43(9):1299–1321, 2024. 1, 2

[10] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *ECCV*, pages 683–700. Springer, 2020. 1

[11] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *NeurIPS*, 2022. 1, 3

[12] Jiawei Sun, Jiahui Li, Tingchen Liu, Chengran Yuan, Shuo Sun, Zefan Huang, Anthony Wong, Keng Peng Tee, and Marcelo H Ang Jr. Rmp-yolo: A robust motion predictor for partially observable scenarios even if you only look once. *ICRA*, 2025. 4

[13] Jiawei Sun, Xibin Yue, Jiahui Li, Tianle Shen, Chengran Yuan, Shuo Sun, Sheng Guo, Quanyun Zhou, and Marcelo H Ang Jr. Impact: Behavioral intention-aware multimodal trajectory prediction with adaptive context trimming. *arXiv preprint arXiv:2504.09103*, 2025. 4