BEVFormer++: Temporal Amplified BEVformer with Explicit Parameter Prediction for Automatic Trajectory Prediction

Jiabin Fang[1] , **Xu Zhang**[1] , **Zhuoming Ding**[1] , **Xuan Liu**[1] , **Meifang Zhang**[3]
Jin Yuan[1] * and **Yuyi Wang**[2] *
[1]Hunan University
[2]CRRC(China) Tengen Intelligence Institute
[3]The Public Security Bureau of Changsha
{fangjiabin, xuzhang1211, dromin, xuan_liu, yuanjin}@hun.edu.cn, yuyiwang920@gmail.com,
372917175@qq.com

Abstract

Vision-based trajectory prediction with BEV representations has achieved promising results, yet existing methods often suffer from limited temporal modeling and insufficient characterization of motion dynamics, resulting in unsatisfactory trajectory prediction results. To address these issues, we propose a temporally enhanced framework with explicit motion parameter prediction. Specifically, we introduce BEVFormer++, which leverages multi-view images and BEV features from multiple preceding timesteps to generate more robust BEV representations, along with BEV differential features to capture temporal variations. Moreover, we propose a motion-parameter-decoupled tracking module that explicitly estimates velocity, acceleration, and heading angle, providing informative motion cues for trajectory prediction. Extensive experimental results demonstrate that our method outperforms state-of-the-art approaches and can be seamlessly integrated into existing vision-based frameworks, consistently yielding performance improvements.

1 Introduction

Accurate trajectory prediction of surrounding agents is essential for autonomous vehicles to achieve safe and efficient navigation in complex and highly dynamic urban environments [Madjid *et al.*, 2025]. It demands a deep understanding of multi-agent social interactions and the constraints imposed by traffic scenes, making it highly challenging and often leading to suboptimal performance.

Most existing trajectory prediction models are data-driven [Madjid *et al.*, 2025] and can be broadly categorized into trajectory-based and vision-based approaches. Trajectory-based methods [Ngiam *et al.*, 2022; Shi *et al.*, 2022] predict future motion from historical trajectories and are computationally efficient, but their limited scene semantic understanding restricts performance in complex traffic environments. In contrast, vision-based methods [Gu *et al.*, 2023; Zhang *et al.*, 2025a] perform end-to-end prediction from multi-camera inputs, enabling richer scene reasoning and better adaptability to complex scenarios. Generally, vision-based approaches adopt a two-stage pipeline of scene perception and motion prediction. Multi-view images are first transformed into a Bird's-Eye-View (BEV) representation for 3D object detection [Li *et al.*, 2022; Wang *et al.*, 2021] and map construction [Liu *et al.*, 2024b; Li *et al.*, 2024], after which the motion prediction module leverages BEV features together with detection results to forecast future trajectories of surrounding agents. This end-to-end paradigm allows joint optimization of perception and prediction, effectively reducing error propagation.

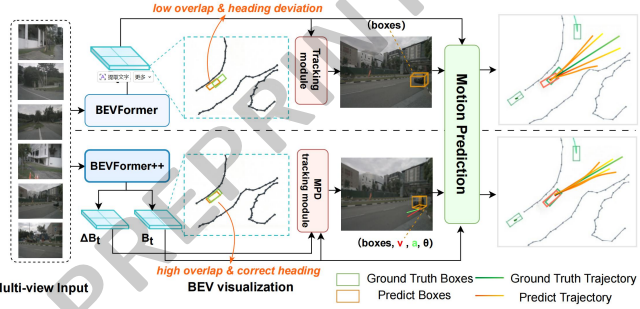

Figure 1: Illustration of the difference between existing vision-based methods (top) and our approach (bottom). Our method leverages BEVFormer++ to explicitly predict dynamic motion parameters—including velocity, acceleration, and heading angle—which are used to enhance BEV feature generation and trajectory prediction.

Nevertheless, the effectiveness of the end-to-end paradigm is largely constrained by the quality of the BEV features, as errors in the BEV representation can propagate through the pipeline and severely degrade both 3D detection and trajectory prediction performance. As illustrated in Fig. 1 (see the upper part), existing methods often suffer from low spatial overlap and heading deviations in BEV representations. These inaccuracies are subsequently propagated to the trajectory prediction module, resulting in significant deviations in both the direction and position of predicted object trajectories.

*Corresponding author.

Although existing BEV generation methods [Li *et al.*, 2022; Liu *et al.*, 2023b] aim to capture complex traffic scenes and incorporate temporal information, they remain suboptimal for trajectory prediction. First, most BEV representations primarily focus on modeling static scene elements and exhibit limited capability in encoding dynamic motion information, resulting in weak temporal expressiveness and inaccurate characterization of velocity and acceleration. Second, BEV features are typically learned in a weakly supervised manner using downstream tasks such as 3D detection and trajectory prediction, which favors accurate localization but fails to explicitly capture fine-grained motion dynamics and directional attributes, such as acceleration and yaw angle, that are crucial for reliable trajectory prediction.

Motivated by the above analysis, we propose a temporally enhanced BEVFormer (BEVFormer++) with explicit motion parameter prediction for automatic trajectory prediction. Following the vision-based pipeline, BEVFormer++ generates enhanced BEV representations by jointly leveraging multi-view images and BEV features from the previous two timesteps, rather than relying solely on the immediate past. By exploiting a broader temporal context, BEVFormer++ more effectively captures scene dynamics and occluded targets, thereby improving the robustness of BEV feature generation. Moreover, BEVFormer++ produces BEV differential features that explicitly model temporal variations between neighboring timesteps, providing informative cues for estimating dynamic motion attributes such as acceleration. Building upon these enhanced representations, the proposed motion-parameter-decoupled tracking module jointly performs object tracking and motion parameter prediction—including velocity, acceleration, and heading angle—by leveraging both BEV and BEV differential features. The predicted motion parameters are subsequently incorporated into the motion prediction module to facilitate trajectory forecasting. Unlike existing vision-based approaches, explicitly supervising motion parameter learning enforces more accurate scene- and motion-aware BEV representations, leading to improved robustness. Furthermore, the incorporation of explicit dynamic parameters enables the trajectory prediction model to better capture both the position and orientation of moving agents, thereby significantly enhancing trajectory prediction accuracy.

The main contributions are summarized as follows:

- To the best of our knowledge, we are the first to introduce explicit motion parameter estimation into vision-based trajectory prediction frameworks, effectively addressing their limitations in capturing temporal dynamics. The proposed module is plug-and-play and can be seamlessly integrated into existing vision-based trajectory prediction pipelines to substantially improve trajectory prediction accuracy.
- We introduce BEVFormer++, which enhances BEV representation learning by exploiting multi-timestep temporal context. The generated BEV and BEV differential features better model both the static and dynamic characteristics of traffic scenes, thereby providing more robust support for subsequent trajectory prediction.
- We propose a motion-parameter-decoupled tracking module that explicitly predicts fine-grained motion states alongside object tracking, serving as a plug-and-play component that enhances motion modeling and trajectory prediction in vision-based frameworks.

2 Related Work

.

2.1 Bird's-Eye-View (BEV) Representation

The shift from perspective-view to Bird's-Eye-View (BEV) representations has significantly advanced multi-sensor 3D perception by providing a unified top-down space for fusion and planning. Early methods such as LSS [Philion and Fidler, 2020] and BEVDet [Huang *et al.*, 2021] employ explicit depth distribution estimation to project 2D features into 3D space, establishing a geometric basis for view transformation. In contrast, Transformer-based approaches like PETR [Liu *et al.*, 2022] implicitly learn this transformation via spatial cross-attention or 3D coordinate embeddings, effectively reducing error propagation caused by inaccurate depth cues. To address the sparsity and occlusion inherent in single-frame observations, temporal modeling has become crucial for achieving temporally consistent perception. BEVFormer [Li *et al.*, 2022] introduces recurrent temporal self-attention to aggregate historical BEV features, while subsequent works such as VideoBEV [Han *et al.*, 2024] and StreamPETR [Wang *et al.*, 2023] explore longer temporal sequences or memory buffers to improve velocity estimation and object recall. Despite these advances, most existing BEV backbones are optimized for low-level detection or static occupancy tasks, where temporal fusion mainly smooths spatial features, and thus fail to capture higher-order temporal variations—such as sudden acceleration or subtle heading changes—required for complex motion forecasting. Our BEVFormer++ addresses this limitation by leveraging multi-timestep differential features.

2.2 Trajectory Prediction

Early methodologies primarily formulated trajectory prediction as a sequence extrapolation problem based on historical coordinates. However, due to the absence of scene constraints, relying solely on past motion often leads to socially or geometrically implausible trajectories. To address this limitation, subsequent studies incorporated scene constraints to restrict the prediction space [Liu *et al.*, 2024a], [Pei *et al.*, 2025a], [Zhang *et al.*, 2025b].

Despite leveraging scene context, these methods still struggle to capture the full semantic complexity of dynamic environments, motivating the development of vision-based end-to-end frameworks such as UniAD [Hu *et al.*, 2023] and VAD [Jiang *et al.*, 2023]. UniAD introduces a unified query-based architecture that jointly models detection, tracking, mapping, and prediction within a single transformer to reduce error propagation. In contrast, VAD adopts a vectorized representation to transform multi-view images into Bird's-Eye-View (BEV) features for joint depth reasoning and agent interaction modeling. Nevertheless, both frameworks exhibit limitations in long-term temporal modeling. Although recent

LLM-based approaches[Xu *et al.*, 2025] attempt to address these issues by exploiting large-scale reasoning capabilities for complex social understanding, they remain computationally expensive and are typically limited to isolated sub-tasks rather than full-stack integration.

In contrast, BEVFormer++ advances vision-based trajectory prediction by explicitly grounding high-level scene understanding in physical motion constraints. Unlike prior methods that rely on black-box latent regressions or purely reasoning-driven LLMs, our approach decouples motion modeling into supervised kinematic parameters, including velocity, acceleration, and heading angle. By anchoring trajectory prediction to these explicit motion cues, BEVFormer++ produces forecasts that are both semantically informed and physically consistent with vehicle dynamics.

3 Methodology

3.1 Overview

Fig. 2 demonstrates the overall framework of our approach. Given multi-view image inputs $\mathcal{F}_t = \{I_t^1, \ldots, I_t^6\}$, existing trajectory prediction methods typically first generate the Bird's-Eye-View (BEV) feature B_t at the current timestep t, and then predict future trajectories $\hat{Y}_t$ as follows:

$$\begin{aligned} B_t &= \textit{BEVFormer}(\mathcal{F}_t, \mathcal{P}, B_{t-1}), \\ \hat{Y}_t &= \Phi_{traj}(\Phi_{track}(B_t), B_t, \mathcal{M}), \end{aligned} \tag{1}$$

where $\mathcal{P}$ denotes the camera geometry parameters, and $\mathcal{M}$ represents the high-definition (HD) map [Liu *et al.*, 2023a; Liao *et al.*, 2023] features. The tracking network Φ_{track} performs 3D object tracking [Pang *et al.*, 2023; Ding *et al.*, 2024] and provides indexed object bounding boxes to the trajectory prediction network Φ_{traj}. In contrast, as illustrated in Fig. 2, our approach introduces a novel BEVFormer++ to extract more informative BEV representations:

$$[B_t, \Delta B_t] = \textit{BEVFormer++}(\mathcal{F}_t, \mathcal{P}, B_{t-2}, B_{t-1}), \tag{2}$$

where ΔB_t denotes the BEV differential feature, which explicitly captures the temporal variation between BEV features at neighboring timesteps. Building upon these enhanced BEV representations, we further redesign the tracking network by introducing a motion parameter decoupling mechanism:

$$[\hat{M}_t, \hat{O}_t] = \Phi_{MPDtrack}(B_t, \Delta B_t), \tag{3}$$

where $\hat{M}_t = \{a, v, \theta\}$ denotes the estimated motion states of agents, including acceleration a, velocity v, and heading angle θ, and $\hat{O}_t$ represents the indexed bounding boxes of predicted motion agents. Finally, the predicted motion states $\hat{M}_t$ are explicitly incorporated into the trajectory prediction module to enhance future trajectory estimation:

$$\hat{Y}_t = \Phi_{traj}(\hat{O}_t, \hat{M}_t, B_t). \tag{4}$$

in the following sections, we provide detailed descriptions of BEVFormer++, the motion-parameter-decoupled tracking network $\Phi_{MPD_{track}}$, and the trajectory prediction network.

3.2 BEVFormer++

The traditional BEVFormer constructs the BEV representation at the current timestep by fusing multi-camera visual features with the BEV feature propagated from the previous timestep. This design facilitates object velocity estimation and partially mitigates occlusion effects. However, relying only on visual information from two adjacent timesteps is often insufficient for robustly identifying heavily occluded objects, which may in turn propagate errors into downstream trajectory prediction. Furthermore, due to variations in vehicle speed during driving, feature differences between consecutive timesteps provide merely a coarse approximation of velocity and are incapable of capturing higher-order motion dynamics such as acceleration, thereby potentially degrading the accuracy of trajectory prediction.

Motivated by these observations, we propose a temporally enhanced BEVFormer, termed BEVFormer++. BEVFormer++ jointly exploits the temporal correlations between the current multi-view features and the BEV features from the two preceding timesteps, leading to more accurate and temporally consistent BEV representations. Furthermore, by explicitly comparing BEV features across different timesteps, BEVFormer++ derives BEV differential features that encode higher-order motion information, providing informative cues for object acceleration estimation.

Specifically, given the grid-shaped learnable BEV queries $Q \in \mathbb{R}^{H \times W \times C}$ at the current timestamp t, together with the historical BEV features B_{t-1} and $B_{t-2} \in \mathbb{R}^{H \times W \times C}$ preserved from the two preceding timesteps, where H and W denote the spatial resolution of the BEV plane and C is the channel dimension, we first align B_{t-1} and B_{t-2} to Q according to the ego-motion. This alignment ensures that features at the same BEV grid correspond to the same real-world locations, yielding two ego-motion–compensated historical BEV features B'_{t-1} and B'_{t-2}. On this basis, the traditional BEVFormer models the temporal association of the same objects between timesteps $t-1$ and t through a temporal self-attention (TSA [Vaswani *et al.*, 2017]) layer, which can be expressed as:

$$\begin{aligned} \mathbf{TSA}(Q_p, Q, B'_{t-1}) &= \sum_{V \in \{Q, B'_{t-1}\}} \textit{DeformAttn}(Q_p, p, V) \\ &= \sum_{i=1}^{N_{head}} \mathbf{W}_i [\sum_{j=1}^{N_{point}} A_{ij} \mathbf{W}'_i V(p + \Delta p_{ij})], \end{aligned} \tag{5}$$

where $Q_p \in \mathbb{R}^{1 \times C}$ denotes the BEV query located at $p = (x, y)$, and V is the input feature maps concatenated by Q and B'_{t-1}. A_{ij} and Δp_{ij} are the attention weights and predicted offsets both obtained through a linear transformation on Q_p. $\mathbf{W}'_i$ and $\mathbf{W}_i$ are learnable weights. N_{head} and N_{point} indicates the total number of attention heads and sampling points in each head, respectively. Differently, BEVformer++ further considers the temporal association of the BEV features between the timestep $t-2$ and t, and finally predict the BEV feature and BEV differential feature at the timestep t as

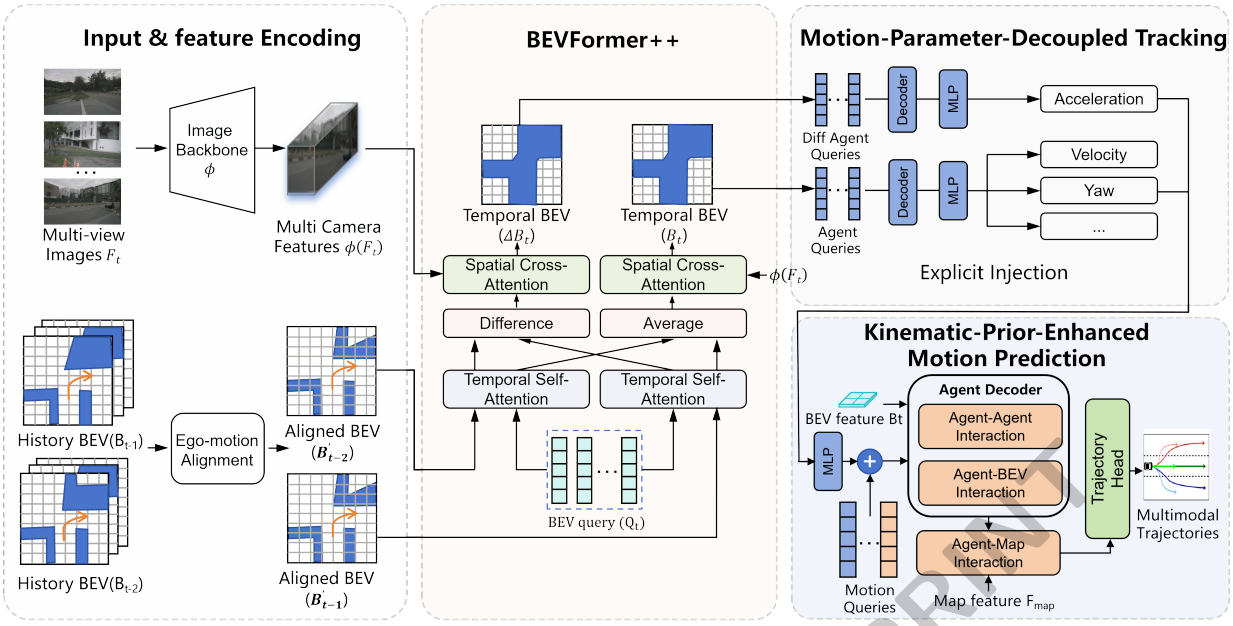

Figure 2: Overall architecture of the proposed method. BEVFormer++ constructs enhanced BEV representations $(B_t, \Delta B_t)$ by aggregating multi-timestep historical features. Based on these representations, the motion-parameter-decoupled tracking module explicitly estimates fine-grained motion states—including acceleration, velocity, and heading angle—which are then injected into the motion prediction module to guide trajectory prediction.

follows:

$$B_t = \omega(avg(\mathbf{TSA}(Q_p, Q, B'_{t-1})), \mathbf{TSA}(Q_p, Q, B'_{t-2}))),$$
$$\Delta B_t = \omega(dif(\mathbf{TSA}(Q_p, Q, B'_{t-1})), \mathbf{TSA}(Q_p, Q, B'_{t-2}))), \tag{6}$$

where $avg(,)$ and $dif(,)$ represent the average and difference operations between two elements, respectively, and ω represent the subsequent series operations in BEVFormer including spatial cross-attention, Normalization and Feed Forward.

Beyond the tradition BEVFormer, BEVFormer++ has the following advantages: First, it takes into account the bev features over a longer historical period, enabling it to more accurately capture changes in the environment and targets, and more effectively mitigate the issue of target occlusion. Second, it considers the difference in BEV (Bird's Eye View) features between adjacent moments, enabling it to more accurately capture the motion trend of the target and provide feature basis for subsequent acceleration evaluation.

3.3 Motion-parameter-decoupled Tracking

Unlike conventional tracking models that infer 3D tracking results solely from BEV features at timestep t, the proposed motion-parameter-decoupled tracking module jointly exploits both BEV features and BEV differential features to produce 3D tracking outputs. Meanwhile, it explicitly estimates each agent's motion parameters, including velocity $\hat{\mathbf{v}} \in \mathbb{R}^{2\times N_a}$, acceleration $\hat{\mathbf{a}} \in \mathbb{R}^{2\times N_a}$, and heading angle $\hat{\theta} \in \mathbb{R}^{2\times N_a}$, where N_a records the number of potential agents. By decoupling motion parameter estimation from spatial localization, the proposed design enables more precise modeling of fine-grained agent dynamics.

To this end, we initialize two sets of learnable queries, $\mathbf{Q}_a \in \mathbb{R}^{N_a\times C}$ and $\mathbf{Q}_{\Delta a} \in \mathbb{R}^{N_a\times C}$, and employ a dual-branch Transformer decoder [Hornik *et al.*, 1989] to decouple the modeling of agent motion states. Specifically, the first branch leverages $\mathbf{Q}_a$ to estimate agents' instantaneous states, including velocity ($\hat{\mathbf{v}}$, heading angle $\hat{\theta}$, and 3D bounding boxes, from the BEV feature B_t, while the second branch utilizes $\mathbf{Q}_{\Delta a}$ to infer higher-order motion states, namely acceleration $\hat{a}$, from the BEV differential feature ΔB_t. In both branches, the learnable queries are first processed by self-attention to capture inter-agent dependencies. They are then used as queries (Q) in a cross-attention module, where the corresponding BEV features serve as keys (K) and values (V) to aggregate rich spatiotemporal context. Finally, the fused representations are decoded into decoupled motion components through multi-layer perceptrons (MLPs [Hornik *et al.*, 1989]). The overall process can be formulated as:

$$\hat{\mathbf{v}}, \hat{\theta}, det = \mathrm{MLP}_{det}(\mathrm{TransformerDecoder}(\mathbf{Q}_a, \mathbf{B}_t)),$$
$$\hat{\mathbf{a}} = \mathrm{MLP}_{dif}(\mathrm{TransformerDecoder}(\mathbf{Q}_{\Delta a}, \Delta\mathbf{B}_t)), \tag{7}$$

The decoupled motion state representations produced by Eq. (7) serve as explicit physical priors for subsequent motion prediction. To ensure the reliability of these priors, we supervise their learning by optimizing a dynamic loss $\mathcal{L}_{\text{dyn}}$, which is formulated using the L_1 norm against the corresponding ground-truth motion labels:

$$\mathcal{L}_{\text{dyn}} = \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} \left(|\hat{\mathbf{v}}_i - \mathbf{v}_i|_1 + |\hat{\mathbf{a}}_i - \mathbf{a}_i|_1\right), \tag{8}$$

where N_{pos} denotes the number of successfully matched pos-

itive proxies. The ground-truth acceleration $\mathbf{a}_i$ is obtained by computing the change in displacement between adjacent future timestamps in the training set. Furthermore, to address the phase wrap-around issue—where angles such as 1° and 359° are physically close but numerically distant, leading to gradient instability in direct regression—we adopt a phase-aware rotation loss to supervise the learning of the yaw angle:

$$\mathcal{L}_{\text{rot}} = \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} |\text{atan2}(\sin(\hat{\theta}_i - \theta_i), \cos(\hat{\theta}_i - \theta_i))|. \quad (9)$$

The total motion state loss is then expressed as:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{dyn}} + (1 - \lambda)\mathcal{L}_{\text{rot}}, \quad (10)$$

where λ is a weighting hyperparameter.

3.4 Kinematic-Prior-Enhanced Motion Prediction

Existing BEV-based motion prediction methods typically rely on BEV features and agent position information to predict future motion states, which may lead to predictions that violate real-world physical laws. To address this, our motion prediction module explicitly integrates agents' motion-parameter-decoupled physical priors, ensuring that the predictions are both more accurate and consistent with physical principles.

Specifically, our motion prediction module first encodes the motion-parameter-decoupled priors using an MLP to obtain the motion prior features $\mathbf{F}_p \in \mathbb{R}^{N_a \times C}$, expressed as:

$$\mathbf{F}_p = \text{MLP}([p_{(x,y)}, \hat{v}_{(x,y)}, \hat{a}_{(x,y)}, \hat{\theta}_{(sin,cos)}]^\top) \quad (11)$$

where (x, y) denotes the BEV coordinates. The motion state prior feature $\mathbf{F_p}$ is then fused additively with the initialized motion queries $\mathbf{Q}_{\text{init}} \in \mathbb{R}^{N_a \times C}$ to produce the physical priors-enhanced motion queries $\mathbf{Q}_m \in \mathbb{R}^{N_a \times C}$:

$$\mathbf{Q}_m = \mathbf{Q}_{init} + \mathbf{F}_p. \quad (12)$$

We then input $\mathbf{Q}_m$ as queries and $\mathbf{B}_t$ as keys and values into the Transformer decoder to aggregate agent and environmental context from the BEV features $\mathbf{B}_t$, yielding the refined latent representation $\mathbf{H}_m \in \mathbb{R}^{N_a \times N_{mode} \times C}$, expressed as:

$$\mathbf{H}_m = \text{TransformerDecoder}(\mathbf{Q}_m.\mathbf{B}_t), \quad (13)$$

where N_{mode} denotes the number of distinct trajectories. To further incorporate topological constraints from the road network, the model performs agent-map interaction. This interaction is formulated as:

$$\mathbf{H}'_m = \text{MHCA}(\mathbf{H}_m, \mathbf{F}_{map}), \quad (14)$$

where $\mathbf{F}_{map}$ represents instance-level map features extracted by a LaneNet [Neven *et al.*, 2018] through hierarchical MLP layers and max-pooling over predicted vectorized lane points. By aggregating environmental context through this cross-attention mechanism, the refined latent representation $\mathbf{H}'_m$ explicitly captures the physical constraints of the surrounding traffic infrastructure.

Finally, we decode $\mathbf{H}'_m$ to generate the agents' motion trajectories $\hat{\mathbf{Y}}_t$:

$$\hat{Y}_t = \text{MLP}_{traj}(\mathbf{H}'_m) \in \mathbb{R}^{N_a \times Nmode \times T \times 2}, \quad (15)$$

where T denotes the future timestamps. Here, N_{mode} represents the number of motion modes, accounting for the inherent multimodality of agent behavior in complex traffic scenarios. By generating N_{mode} distinct trajectories for each agent, the model can represent multiple plausible future outcomes (e.g., turning left vs. going straight), thereby providing a probabilistic distribution of future states rather than a single deterministic path.

4 Experiments

4.1 Dataset

We conduct experiments on the challenging public nuScenes dataset [Caesar *et al.*, 2020], which comprises 1,000 driving scenes, each lasting approximately 20 seconds. Following the official data split, 700 scenes are used for training, 150 for validation, and 150 for testing. Consistent with prior work [Jiang *et al.*, 2023; Hu *et al.*, 2023], trajectory prediction performance is evaluated using Minimum Average Displacement Error (minADE), Minimum Final Displacement Error (minFDE), and Miss Rate (MR). In addition, mean Average Precision (mAP) and nuScenes Detection Score (NDS) [Caesar *et al.*, 2020] are adopted to assess the quality of 3D object detection.

4.2 Implementation Details

We select two vision-based methods VAD [Jiang *et al.*, 2023] and UniAD [Hu *et al.*, 2023] as our baselines, which both use ResNet101 [He *et al.*, 2016] as the image backbone. Our approach replaces their BEVFormer with our BEVFormer++. To make a fair comparison, we follow the settings of the BEVFormer. Specifically, the size of BEV queries is 200×200 with $C = 256$. We adopt learnable positional embedding for BEV queries. The BEV encoder contains 6 encoder layers and constantly refines the BEV queries in each layer. For each local query, we set $N_{head} = 8$, $N_{point} = 4$. In our Motion-parameter-decoupled tracking, we set the weight $= 0.5$ in Eq (10). We use AdamW [Loshchilov and Hutter, 2019] optimizer and Cosine Annealing [Loshchilov and Hutter, 2017] scheduler to train our model with weight decay 0.01 and initial learning rate 2×10^{-4}. We trained our model for 60 epochs on 6 NVIDIA RTX A6000 GPUs with batch size 1 per GPU.

4.3 Comparison with the State-of-the-art Methods

We compare our approach against several state-of-the-art methods, including vision-based approaches such as ViP3D, VAD, and UniAD, as well as trajectory-based methods such as DeMo, LAformer, and FiM. The quantitative results are summarized in Table 1. Our method consistently achieves the best performance across all evaluation metrics, outperforming both vision-based and trajectory-based trajectory prediction approaches. As a vision-based framework, our method significantly enhances temporal modeling capability by explicitly incorporating dynamic information. By introducing motion parameter prediction, our approach effectively bridges the strengths of vision-based and trajectory-based methods, leading to superior overall performance. Furthermore, compared with the two strong baselines, VAD and

Method	minADE$_5$ ↓	minFDE$_5$ ↓	MR$_5$ ↓	FPS
DeMo [Zhang *et al.*, 2024]	0.918	1.613	0.287	/
FiM [Pei *et al.*, 2025a]	0.880	/	0.310	/
LAformer [Liu *et al.*, 2024a]	0.921	1.643	0.27	8.7
Para-Drive [Weng *et al.*, 2024]	0.720	1.090	0.142	5.0
ViP3D [Gu *et al.*, 2023]	1.210	2..210	0.239	/
GoIRL [Pei *et al.*, 2025b]	0.750	/	0.210	/
VAD [Jiang *et al.*, 2023]	0.674	0.919	0.096	4.5
UniAD [Hu *et al.*, 2023]	0.716	1.029	0.154	1.8
Our approach$_{UniAD}$	0.697	0.937	0.128	1.7
Our approach$_{VAD}$	**0.615**	**0.883**	**0.089**	4.4

Table 1: Trajectory Prediction Results on the nuScenes Validate Set Compared with State-of-the-Art Methods.

Method	mAP ↑	NDS ↑
BEVFormer [Li *et al.*, 2022]	0.481	0.569
StreamPETR [Wang *et al.*, 2023]	0.396	0.490
BEVDet4D [Huang and Huang, 2022]	0.451	0.569
BEVFormer++ (Det)	**0.498**	**0.587**
BEVFormer(UniAD) [Hu *et al.*, 2023]	0.310	0.374
BEVFormer(VAD) [Jiang *et al.*, 2023]	0.321	0.455
BEVFormer++ (UniAD)	0.386	0.488
BEVFormer++ (VAD)	**0.401**	**0.502**

Table 2: 3D Detection Performance on nuScenes Validate Set.

BEVFormer++	MPDTraj	minADE$_5$↓	minFDE$_5$↓	MR$_5$↓
×	×	0.674	0.919	0.096
✓	×	0.675	0.915	0.095
×	✓	0.652	0.905	0.092
✓	✓	**0.615**	**0.883**	**0.089**

Table 3: Ablation study of the Proposed Components Tested on the nuScenes Validate Set.

UniAD, our method yields substantial and consistent performance gains. These improvements can be attributed to the enhanced dynamic feature modeling enabled by BEVFormer++, as well as the explicit incorporation of motion parameters into the trajectory prediction module. Regarding computational cost, our method incurs only a minimal overhead. Our approach$_{VAD}$ achieves 4.4 FPS (vs. 4.5 FPS for VAD), and Our approach$_{UniAD}$ reaches 1.7 FPS (vs. 1.8 FPS for UniAD). The slight reduction in frame rate is attributed to the additional motion constraint calculations; however, given the substantial reduction in error rates, this trade-off is well-justified. Together, these components allow the model to more accurately capture agents' positions, velocities, and motion directions, resulting in more precise trajectory predictions.

4.4 Evaluation of BEVFormer++

We evaluate 3D object detection performance under two settings: pure 3D detection and trajectory prediction. As shown in Table 2, we compare BEVFormer++ with three representative BEV feature extraction methods—BEVFormer, StreamPETR, and BEVDet4D—on pure 3D detection tasks. The results indicate that BEVFormer++ achieves the best detection performance among all compared methods. This improvement primarily stems from its enhanced temporal modeling capability, which aggregates BEV features from multiple historical timesteps and explicitly captures temporal variations through BEV differential features. As a result, BEVFormer++ can more robustly perceive dynamic and occluded objects, leading to more accurate and stable 3D detection results. In the trajectory prediction setting, we observe a notable degradation in 3D detection performance, suggesting that the model tends to prioritize trajectory correctness over precise spatial localization. In such cases, the BEV representation alone may be insufficient for accurate 3D positioning, and the resulting localization errors can propagate to downstream trajectory prediction tasks. As evidenced by the experimental results, BEVFormer++ effectively mitigates this issue by introducing explicit supervision on both static and dynamic target parameters. This additional supervision enables more accurate estimation of agents' positions and motion states, thereby reducing error propagation and improving overall trajectory prediction performance.

4.5 Ablation Studies

We perform comprehensive ablation studies on the nuScenes validation set to verify the effectiveness of our proposed BEVFormer++, the Motion-Parameter-Decoupled Tracking (MPDTraj), and the motion state loss formulation. The quantitative results, summarized in Table 3 and Table 4, demonstrate that each component is essential for achieving optimal trajectory prediction performance.

Impact of BEVFormer++. As shown in the first two rows of Table 3, introducing BEVFormer++ (which integrates B_{t-2} and computes explicit differential features) yields improvements in long-term prediction accuracy (minFDE decreases from 0.919 to 0.915) and miss rate. While the standalone gain is moderate, BEVFormer++ is critical for the full system; it provides the high-fidelity differential BEV features (ΔB_t) required by the acceleration branch of the tracking module. Its full potential is unlocked when combined with MPDTraj, where the synergistic integration leads to the state-of-the-art performance (Row 4).

Impact of MPDTraj. The introduction of the MPDTraj module (Row 3) significantly reduces minADE from 0.674 to 0.652 compared to the baseline. This confirms that explicitly decoupling motion parameters (velocity, acceleration, and heading angle) allows the network to capture fine-grained agent dynamics more effectively than implicit latent modeling. Furthermore, the combination of MPDTraj and BEVFormer++ (Row 4) achieves the lowest errors across all metrics (**0.615** minADE). This substantial improvement validates that accurate acceleration estimation relies heavily on the temporal differential cues provided by BEVFormer++.

Effectiveness of the Motion State Loss. To investigate the trade-off between kinematic magnitude and directional constraints, we evaluate the impact of the hyperparameter λ in our joint loss function $\mathcal{L}_{\text{total}}$, as reported in Table 4. The results exhibit a clear optimum at $\lambda = 0.5$. Setting $\lambda = 0.0$ (ignoring dynamics) leads to significant degradation in displacement errors (minADE 0.696), as the model fails to constrain speed and acceleration. Conversely, $\lambda = 1.0$ (ignoring rotation) hampers performance by neglecting the yaw angle, which is critical for trajectory shape. The balanced setting ($\lambda = 0.5$) effectively harmonizes the scalar L_1 constraints

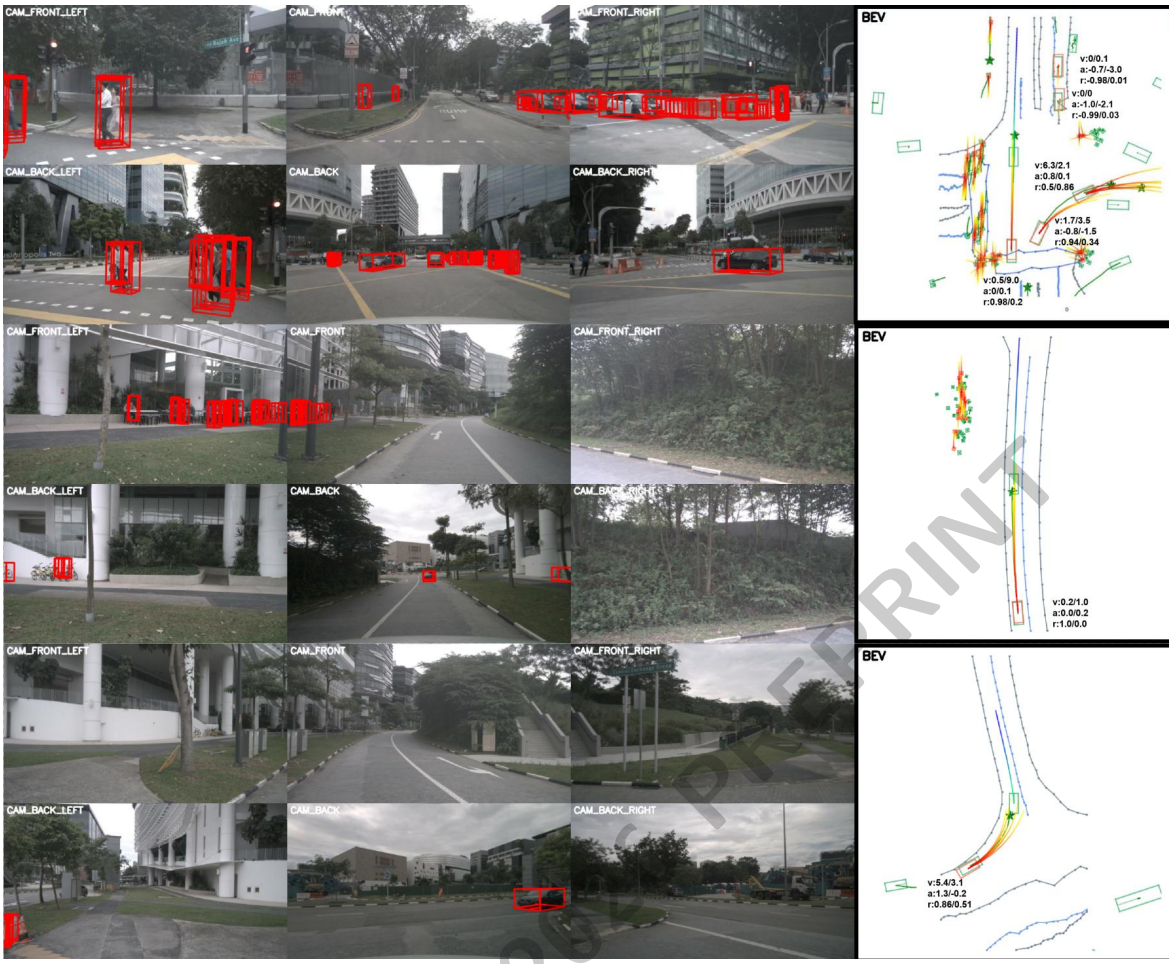

Figure 3: **Qualitative results of BEVFormer++ across complex intersections, straight segments, and turning maneuvers.** Multi-view camera inputs (left) are paired with Bird's-Eye-View (BEV) predictions (right). Red boxes represent 3D detection, and yellow-to-red curves denote multimodal future trajectories. The parameters v, a, and r in the figure represent velocity, acceleration and heading angle respectively. The results demonstrate that explicit motion parameter supervision produces smooth, physically consistent paths that maintain lane alignment and momentum accuracy even during complex maneuvers or sensor occlusions.

λ	**minADE**$_5$ $\downarrow$	**minFDE**$_5$ $\downarrow$	**MR** $\downarrow$
0.0	0.696	1.092	0.117
0.2	0.637	0.971	0.098
0.5	**0.615**	**0.883**	**0.089**
0.8	0.628	0.929	0.093
1.0	0.654	1.019	0.103

Table 4: Performacne Change by Using the Loss Function $\mathcal{L}_{\text{total}}$ under Different Weights λ Tested on the nuScenes Validation Set.

on velocity/acceleration with the phase-aware supervision of heading, yielding the most accurate trajectory predictions.

5 Qualitative Results

We visualize BEVFormer++ performance across complex intersections, straight segments, and turning maneuvers in Fig. 3. By leveraging multi-timestep aggregation, our model maintains robust 3D detection despite occlusions or camera truncations. The Bird's-Eye-View (BEV) results demonstrate smooth, kinematically feasible trajectories (yellow-to-red curves) that remain strictly aligned with agent momentum and road topology. These results confirm that explicit motion parameter supervision effectively eliminates kinematically unfeasible predictions in complex urban environments.

6 Conclusion

In this paper, we present BEVFormer++, a temporally enhanced framework that integrates multi-timestep BEV aggregation and explicit motion parameter prediction for robust trajectory forecasting. By leveraging BEV differential features and a decoupled tracking module, our method effectively captures fine-grained agent dynamics including velocity, acceleration, and heading angle. Extensive experiments on the nuScenes benchmark demonstrate that our approach consistently outperforms state-of-the-art methods and provides physically consistent predictions in complex driving scenarios.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China (No.62272157).

References

[Caesar *et al.*, 2020] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[Ding *et al.*, 2024] Shuxiao Ding, Lucas Maag, Frederic Kettelhoit, Florian Hölzl, and Jürgen Gall. ADA-Track: End-to-end multi-camera 3D multi-object tracking with alternating detection and association. In *CVPR*, 2024.

[Gu *et al.*, 2023] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. ViP3D: End-to-end visual trajectory prediction via 3D agent queries. In *CVPR*, 2023.

[Han *et al.*, 2024] Chunrui Han, Jinrong Yang, Jianjian Sun, Zheng Ge, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. Exploring recurrent long-term temporal fusion for multi-view 3D perception. *IEEE Robotics and Automation Letters*, 9(7):6544–6551, 2024.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Hornik *et al.*, 1989] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[Hu *et al.*, 2023] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving (UniAD). In *CVPR*, 2023.

[Huang and Huang, 2022] Junjie Huang and Guan Huang. BEVDet4D: Exploit temporal cues in multi-camera 3D object detection. *arXiv preprint arXiv:2203.17054*, 2022.

[Huang *et al.*, 2021] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. BEVDet: High-performance multi-camera 3D object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

[Jiang *et al.*, 2023] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized scene representation for efficient autonomous driving. In *ICCV*, 2023.

[Li *et al.*, 2022] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022.

[Li *et al.*, 2024] Tianyu Li, Peijin Jia, Bangjun Wang, Li Chen, Kun Jiang, Junchi Yan, and Hongyang Li. LaneSegNet: Map learning with lane segment perception for autonomous driving. In *ICLR*, 2024.

[Liao *et al.*, 2023] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. MapTR: Structured modeling and learning for online vectorized HD map construction. In *ICLR*, 2023.

[Liu *et al.*, 2022] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3D object detection. In *ECCV*, 2022.

[Liu *et al.*, 2023a] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. VectorMapNet: End-to-end vectorized HD map learning. In *ICML*, 2023.

[Liu *et al.*, 2023b] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*, 2023.

[Liu *et al.*, 2024a] Mengmeng Liu, Hao Cheng, Lin Chen, Hellward Broszio, Jiangtao Li, Runjiang Zhao, Monika Sester, and Michael Ying Yang. LAformer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In *CVPR*, 2024.

[Liu *et al.*, 2024b] Xiaolu Liu, Song Wang, Wentong Li, Ruizi Yang, Junbo Chen, and Jianke Zhu. MGMap: Mask-guided learning for online vectorized HD map construction. In *CVPR*, 2024.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[Madjid *et al.*, 2025] Nadya Abdel Madjid, Abdulrahman Ahmad, Murad Mebrahtu, Yousef Babaa, Abdelmoamen Nasser, Sumbal Malik, Bilal Hassan, Naoufel Werghi, Jorge Dias, and Majid Khonji. Trajectory prediction for autonomous driving: Progress, limitations, and future directions. *arXiv preprint arXiv:2503.03262*, 2025. Accepted for Information Fusion.

[Neven *et al.*, 2018] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: An instance segmentation approach. In *IEEE Intelligent Vehicles Symposium (IV)*, 2018.

[Ngiam *et al.*, 2022] Jiquan Ngiam, Vijay Vasudevan, Benjamin Caine, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, David Weiss, Ben Sapp, Zhifeng Chen, and Jonathon Shlens. Scene transformer: A unified architecture for predicting multiple agent trajectories. In *ICLR*, 2022.

[Pang *et al.*, 2023] Ziqi Pang, Jie Li, Pavel Tokmakov, Dian Chen, Sergey Zagoruyko, and Yu-Xiong Wang. Standing between past and future: Spatio-temporal modeling for multi-camera 3D multi-object tracking. In *CVPR*, 2023.

[Pei *et al.*, 2025a] Muleilan Pei, Shaoshuai Shi, Xuesong Chen, Xu Liu, and Shaojie Shen. Foresight in motion: Reinforcing trajectory prediction with reward heuristics. In *ICCV*, 2025.

[Pei *et al.*, 2025b] Muleilan Pei, Shaoshuai Shi, Lu Zhang, Peiliang Li, and Shaojie Shen. GoIRL: Graph-oriented inverse reinforcement learning for multimodal trajectory prediction. In *ICML*, 2025.

[Philion and Fidler, 2020] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *ECCV*, 2020.

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

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[Wang *et al.*, 2021] Yue Wang, Vitor C Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In *CoRL*, 2021.

[Wang *et al.*, 2023] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3D object detection. In *ICCV*, 2023.

[Weng *et al.*, 2024] Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. PARA-Drive: Parallelized architecture for real-time autonomous driving. In *CVPR*, 2024.

[Xu *et al.*, 2025] Yi Xu, Ruining Yang, Hangcheng Cao, Mohan Wang, and Yifu Zhang. Trajectory prediction meets large language models: A survey. *arXiv preprint arXiv:2506.03408*, 2025.

[Zhang *et al.*, 2024] Bozhou Zhang, Nan Song, and Li Zhang. Demo: Decoupling motion forecasting into directional intentions and dynamic states. In *NeurIPS*, 2024.

[Zhang *et al.*, 2025a] Bozhou Zhang, Nan Song, Xin Jin, and Li Zhang. Bridging past and future: End-to-end autonomous driving with historical prediction and planning. In *CVPR*, 2025.

[Zhang *et al.*, 2025b] Bozhou Zhang, Nan Song, Xiatian Zhu, and Li Zhang. Demo++: Motion decoupling for autonomous driving. *arXiv preprint arXiv:2507.17342*, 2025.