ForeSight: Multi-View Streaming Joint Object Detection and Trajectory Forecasting

Sandro Papais Letian Wang Brian Cheong Steven L. Waslander
University of Toronto
{sandro.papais, letian.wang, brian.cheong, steven.waslander}@robotics.utias.utoronto.ca

Abstract

We introduce ForeSight, a novel joint detection and forecasting framework for vision-based 3D perception in autonomous vehicles. Traditional approaches treat detection and forecasting as separate sequential tasks, limiting their ability to leverage temporal cues from past forecasts. ForeSight addresses this limitation with a multi-task streaming and bidirectional learning approach, allowing detection and forecasting to share query memory and propagate information seamlessly. The forecast-aware detection transformer enhances spatial reasoning by integrating trajectory predictions from a multiple hypothesis forecast memory queue, while the streaming forecast transformer improves temporal consistency using past forecasts and refined detections. Unlike tracking-based methods, ForeSight eliminates the need for explicit object association, reducing error propagation with a tracking-free model that efficiently scales across multi-frame sequences. Experiments on the nuScenes dataset show that ForeSight achieves state-of-the-art performance, achieving an EPA of 54.9%, surpassing previous methods by 9.3%, while also attaining the highest mAP among multi-view detection models and maintaining competitive motion forecasting accuracy.

1. Introduction

Perception systems in autonomous vehicles (AVs) are critical for understanding the dynamic driving environment. Vision-based 3D detection has gained popularity due to its low deployment cost and ability to interpret visual elements. Despite this progress, vision-based systems struggle in complex driving scenarios, particularly with occluded or partially visible objects which can pose critical safety risks. To address these challenges, temporal learning methods that leverage multi-frame histories have emerged as a solution, improving detection by aggregating past observations and motion cues.

Existing temporal learning methods fall into two categories: *dense-feature* and *sparse-query* learning. Dense-feature methods [17, 22, 27, 29, 41] align historic bird's-eye-view (BEV) features to the current frame for fusion. Sparse-query methods [33–35, 57] represent objects as sparse queries and use attention [51] to selectively fuse temporal features, achieving strong efficiency and performance. However, they often overlook the multi-modal motion of surrounding agents, limiting their ability to fully exploit temporal information. Given advances in motion forecasting [50, 54, 65, 66], neglecting to fully incorporate future movement in detection discards valuable cues.

Traditional motion forecasting models rely on curated ground-truth trajectories, assuming perfect detection and tracking. However, this assumption does not hold in real-world, online perception scenarios, leading to degraded performance [63]. Meanwhile, emerging end-to-end methods integrating perception and forecasting architectures into joint training show promise in the presence of perception errors [11, 16]. These methods rely on tracking to bridge the gap between forecasting and perception. However, they trade off detection accuracy for tracking performance and introduce errors from incorrect trajectories association, bottlenecking learning. Moreover, these forecasting methods recompute trajectories at each timestep, making them computationally inefficient. By integrating forecasting with temporal detection through a streaming memory queue, ForeSight enhances computational efficiency and ensures flexible propagation of historical context.

Our key insight is to re-think the traditional autonomy stack, where motion forecasting is typically used for planning and then discarded. Instead, we propose an integrated system where motion forecasts feed back into itself and detection through a joint streaming memory queue. Therefore, we introduce **ForeSight**, a novel framework that jointly optimizes detection and forecasting through unified process for query initialization, attention, and propagation. Experiments show that bidirectional query flow enhances both tasks, achieving state-of-the-art end-to-end forecasting performance. ForeSight uses a unified transformer architecture for temporal detection and track-free forecasting with a

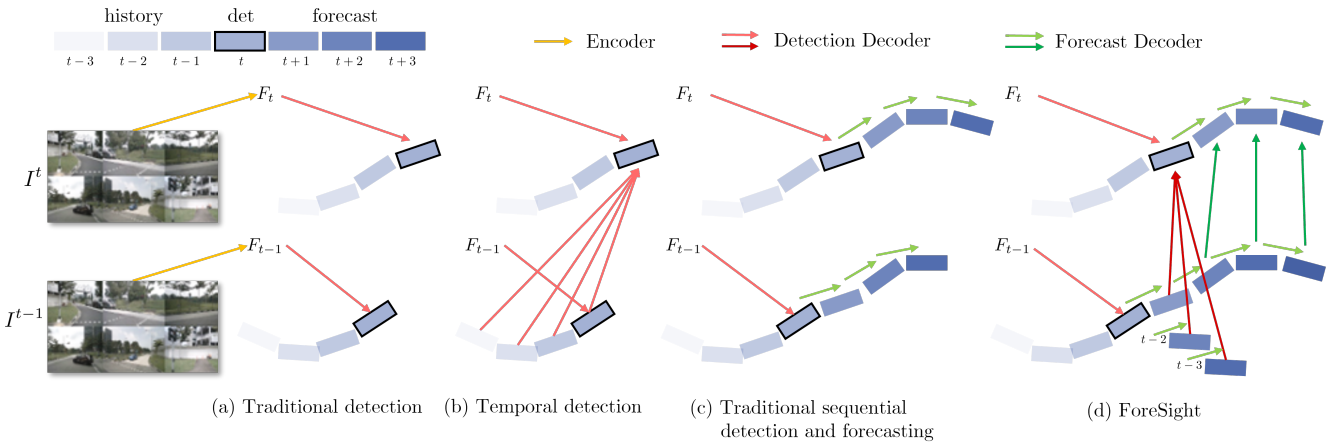

Figure 1. A comparison of temporal learning perception methods related to ForeSight. (a) Traditional object detection methods leverage single-frame sensor data. (b) Temporal sparse detectors typically keep memory of past detection locations without explicit forecasting capabilities. (c) Typical joint detection-forecasting works pass information from detection to forecasting and rely on tracking for temporal information. (d) ForeSight further propagates motion forecast information to future frames to be reused in detection and forecasting.

joint memory system, allowing past detections and forecasts to serve as a prior on the current detections and forecasts. To our knowledge, ForeSight is the first joint perception and forecasting method to propagate all queries bidirectionally with closed-loop feedback. ForeSight represents a promising advance toward improved memory systems for efficient architectures for autonomous systems. To summarize, our key contributions are:

- We introduce ForeSight, a joint streaming detection and forecasting framework for multi-view images. Unlike existing methods that treat these as sequential tasks, ForeSight enables both to share temporal priors via a unified memory, enhancing performance of both tasks. Our bidirectional query propagation allows information to flow not just from detection to forecasting, as in conventional autonomy stacks, but also back from forecasting to both tasks, enriching detection with temporal cues.
- We introduce a tracking-free streaming forecast approach to improve direct integration between detection and forecasting by removing the tracking bottleneck and adding a shared memory access. We find the forecasting model benefits from this approach, leading to improved end-to-end forecasting.
- Extensive experiments on the NuScenes dataset [1] demonstrate that ForeSight outperforms existing perception and forecasting methods, achieving the highest EPA by a large margin of 9.3% over SOTA joint perception and forecasting method UniAD[16]. ForeSight also surpasses the best multi-view 3D detection baseline, StreamPETR, with a 2.1% mAP improvement.

2. Related Works

2.1. Multi-View 3D Object Detection

Multi-view 3D object detection identifies objects in 3D space by transforming 2D image features from multiple viewpoints into a unified 3D representation. BEV-based methods project 2D features onto a BEV plane via depth estimation, as in LSS [42], CaDDN [43], and BEVDet [19]. Sparse query-based methods, including DETR3D [59] and PETR [34], use learnable 3D queries to aggregate multi-view features via attention [51]. ForeSight follows the sparse query-based approach for object detection.

2.2. Temporal Multi-View 3D Object Detection

Temporal detection methods improve single-frame models by aggregating information across multiple timesteps, enhancing velocity estimation and occlusion handling. The temporal fusion approach depends on the representation type: dense BEV-based or sparse query-based. Dense BEV-based methods like BEVDet4D [17] and Fiery [15] warp and concatenate dense BEV features from past frames, improving detection stability and accuracy. SOLOFusion [41] extends this by caching features across a sliding window for long-term fusion. BEVFormer [29] further enhances dynamic scene modeling with deformable attention to improve accuracy in dynamic environments.

Sparse query-based methods, such as StreamPETR [57], PETRv2 [35], and Sparse4Dv2 [33], cache object queries from past frames to efficiently capture temporal dependencies. These approaches attend to past detection and initialize new queries from them to reduce computational overhead compared to dense BEV-based methods while main-

taining strong performance. However, they lack explicit supervision for propagating detections into the future, limiting temporal learning. ForeSight addresses this by jointly learning detection and forecasting, explicitly supervising how past detections influence future timesteps.

2.3. Motion Forecasting

Motion forecasting for AVs predicts future agent states based on the road map, agent history, and semantic type. Early works [4, 7, 44] rasterize scene context into bird-eye-view images for CNN processing. Later methods adopt vector-based encodings with permutation-invariant operators such as pooling [50], graph convolutions [10, 30], and attention mechanisms [9, 24, 58, 65]. Alternative scene representations, including dynamic insertion areas [52] and Frenet frame-based approaches [53, 55, 64], improve generalizability. However, most assume noise-free inputs, leading to vulnerability in real-world settings. Some works address uncertainties in object classes [20], map elements [12], and object tracks [60, 61], but focus solely on forecasting task trained alone on curated ground truth trajectories.

2.4. Joint Detection and Forecasting

Training motion forecasting separately from detection on curated data leads to degraded performance in real-world scenarios[63]. Joint training mitigates this issue by improving the information flow between both tasks. Early works such as FaF [37], IntentNet [2], and PnPNet [31] demonstrate the benefits over standalone conventional models. Recent methods, such as InterFuser [45], VAD [21], ReasonNet [46] and LMDrive [47] have extended this idea to add downstream tasks such as planning and control. End-to-end joint training of the full set of driving tasks has gained traction, with UniAD [16] and ViP3D [11] integrating sparse queries across all tasks. However, most of these recent methods enforce tracking via ground truth query assignments which introduces information bottlenecks [3] and results in poorer detection performance [6].

While end-to-end methods only consider a one-way flow of information from detection to forecasting, recent works explore backward propagation from forecasting to detection. FaF [37] and FrameFusion [25] showed that using motion prediction to propagate past detection boxes and averaging them with current detections improves detection performance. MoDAR [28] explored directly forecasting previous sensor data as additional input to the detection model for temporal prior modeling. However, these methods are not jointly trained together to perform detection and forecasting. ForeSight introduces a novel streaming joint detection and forecasting framework with bidirectional propagation, enabling both tasks to be trained together in closed-loop for the first time.

3. Method

Our ForeSight framework extends query-based temporal 3D object detectors [57] by introducing a forecasting transformer for joint training of detection and prediction tasks. A streaming query propagation mechanism efficiently leverages temporal information, enabling queries to persist within the model over time. ForeSight processes multi-view camera images to detect surrounding objects and generate multi-modal forecasts over future time steps, optionally incorporating high-definition (HD) maps.

An overview of ForeSight is shown in Figure 2. It consists of two primary components: scene encoders for feature extraction and a transformer network for joint streaming detection and forecasting. Section 3.1 discusses the embedding extraction from multi-view images and HD maps. Section 3.2 details the ForeSight transformer architecture for predicting detections and forecasts, and how queries are initialized and propagated. Finally, Section 3.3 describes the optimization approach to jointly train our model end-to-end, including the losses and data augmentation strategy.

3.1. Scene Embedding Extraction

The first stage of ForeSight involves extracting scene embeddings from multi-view images and HD maps. Camera images serve as the primary sensor input for perceiving surroundings. HD maps provide an additional information source essential for understanding and forecasting object motion. The images and HD maps are encoded separately by individual encoders. The ForeSight transformer can then attend to the relevant scene embeddings to perform detection and forecasting.

Image Encoder. At each time step, we pass multi-camera images to a standard image backbone network (e.g., ResNet-50 [14]), and extract the visual scene features of each camera view. Following previous works [34], each image feature is fused with a 3D positional embedding to generate 3D position-aware features. These features are then used as a sequence of image feature tokens $\mathbf{F}_I$ passed to our joint detection and forecasting transformer.

Map Encoder. The HD map, if available, can be encoded to provide road context information, including lanes, traffic signs, and drivable areas, informing the future motion of objects. Following the vectorized representation [10, 30], we represent the HD map as a directed lane graph $\mathcal{G}(V, E)$, where the lanes are broken up into polylines to form the graph nodes V, and lane connectivity constructs the graph edges E. At each time step, we consider the nearby map elements within an inflated region around the ego agent to account for possible map interaction outside the detection range. Following [8], we first encode the lane polyline sequence according to their connectivity using multi-layer gated recurrent units (GRU), then we apply a graph attention network (GAT) on these lane features to further aggre-

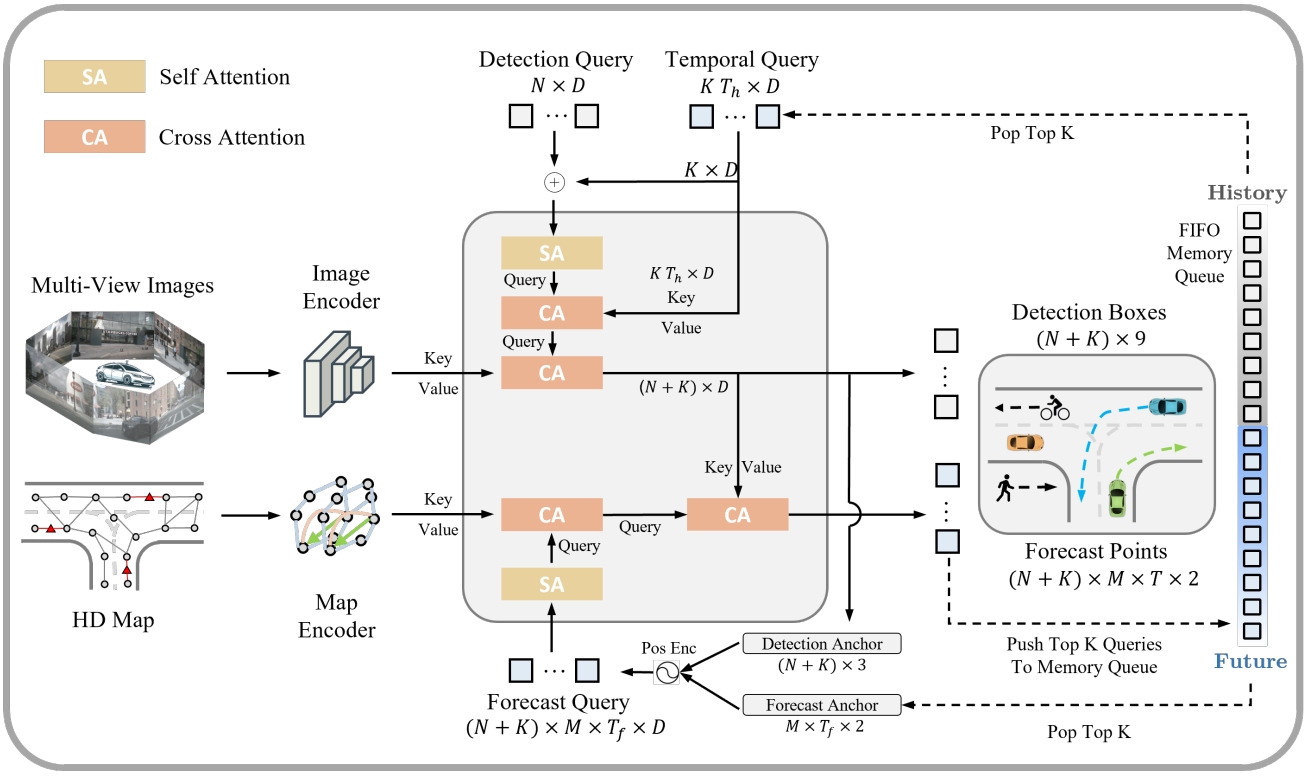

Figure 2. Overall architecture for the proposed ForeSight method. The model takes multi-view images and an optional HD map as inputs and encodes them separately to produce scene embeddings. Detection queries are combined with a subset of temporal queries from a memory bank and attend to themselves, other temporal queries, and the image embeddings to produce detected objects. The detection positions are combined with forecast anchors to produce forecast queries and attend to themselves, map embeddings, and detection queries to produce future forecasts. Finally, the forecast queries are next pushed to the memory bank for future use.

gate local context from the graph structure. We treat lane graph node embedding as a map feature tokens $\mathbf{F}_M$ for later use in our architecture.

3.2. The ForeSight Transformer

At the core of our transformer-based approach are two main concepts: (1) a carefully designed process for query initialization, propagation, and exploitation, and (2) an iterative application of self-attention layers to capture intra-query interactions, along with cross-attention layers to aggregate information from scene embeddings. This architecture enables robust integration of temporal information and facilitates interaction and joint learning of detection and forecasting tasks.

Joint Streaming Memory Queue. A new joint memory management and propagation system was needed to perform joint streaming detection and forecasting. We use a joint memory of shape $\mathbf{q}_{mem} \in \mathcal{R}^{K \times T_h \times T_f \times D}$ to represent the top-K objects across T_h historic frames and T_h future frames with a hidden dimension D. This queue holds the highest confidence past detection and associated forecast queries, allowing the model to leverage multiple future hypotheses as priors. At each step, the queue is shifted along the dimension T_h to match the current time. The corresponding 3D positions, $\mathbf{p}_q$, of the detection and forecast queries are maintained in the ego reference frame and updated across time using the ego vehicle transformation matrix, $\mathbf{T}_{ego}$, by

$$\begin{bmatrix} \mathbf{p}_{q,t} \\ 1 \end{bmatrix} = \mathbf{T}_{ego,t}^{-1} \mathbf{T}_{ego,t-1} \begin{bmatrix} \mathbf{p}_{q,t-1} \\ 1 \end{bmatrix} \tag{1}$$

We then maintain the query memory in a first-in-first-out manner (FIFO) by appending new decoder outputs and popping the oldest features. While introducing a future queue increases memory requirements, the computation to process it is nearly the same since only $K \times T_h$ out of $K \times T_h \times T_f$ are used for inference at the current frame.

Detection Query Initialization. As in many transformer-based detection methods [34, 59], we use N detection queries of dimension D, $\mathbf{q}_{det} \in \mathcal{R}^{N \times D}$, a learned representation, to identify and locate objects within the input data. Specifically, at each time step, the detection queries $\mathbf{q}_{det}$ are initialized as positional encodings (PE) of a set of 3D object

anchor positions $\mathbf{x}_{det} \in \mathcal{R}^{N\times 3}$:

$$\mathbf{q}_{det} = \text{MLP}\left(\text{PE}\left(\mathbf{x}_{det}\right)\right) \tag{2}$$

where $\mathbf{x}_{det}$ is initialized from a random uniform distribution across the normalized detection range in local 3D space.

In addition to detection queries, we also leverage temporal memory queries at each timestep, $\mathbf{q}_{temp} \in \mathcal{R}^{K\times T_h\times D}$, from the historical horizon T_h. Past objects provide valuable cues about the presence and positions of objects in the current frame. Past methods in both temporal detection[13] and tracking [38] rely on only initializing queries from the top-k last frame outputs. In contrast, our approach takes the top-k from all past queries within T_f frames forecasted to the current time. Since our queries can be initialized over multiple frames in the past, this enables a new capability of our model relative to past works in that it can initialize temporal detection queries across occlusions.

Forecast-Aware Detection Transformer. The detection queries $\mathbf{q}_{det}$ and temporal queries $\mathbf{q}_{temp}$ are then fused to aggregate information from image feature tokens $\mathbf{F}_I$ to detect objects in the current frame. Specifically, our approach utilizes temporal queries from the past historical frame, $\mathbf{q}_{temp}^{t-1} \in \mathcal{R}^{K\times D}$, and concatenate them with the detection queries, $\mathbf{q}_{det}$. Here, temporal queries carry forward information on objects detected in the most recent past frame, while detection queries are dedicated to identifying newly appearing objects. These concatenated queries, in the shape of $\mathcal{R}^{(N+K)\times D}$, are then fed into self-attention layers to capture intra-query interactions, and into cross-attention layers to attend to all temporal queries in the full historic horizon $\mathbf{q}_{temp}$. Additional cross-attention layers are then applied, where the queries aggregate information from image feature tokens $\mathbf{F}_I$. Finally, the resulting queries are sent into successive transformer layers for refinement and then decoded as bounding boxes $\mathcal{R}^{(N+K)\times 9}$, which include 3D coordinates, box sizes, heading, classes, and confidence.

Forecast Query Initialization. For each decoded detection, ForeSight forecasts its states over the future horizon T_f. First, we extract the 2D coordinates and heading from the decoded bounding boxes to form detection anchors of shape $\mathcal{R}^{(N+K)\times 3}$, representing the current states of detected objects. Next, to produce multi-modal forecasts, we follow [4, 16, 48] by creating forecast anchors of shape $\mathcal{R}^{M\times T_f\times 2}$. A subset of these anchors are generated via k-means clustering on ground-truth trajectory data to represent trajectory points of distinct modalities. Our work extends this approach further by adding a set of temporal anchors as an additional modality from the previous highest confidence forecast trajectory output. This allows the forecasting model to improve temporal stability and reuse previous computations. Finally, we combine the detection anchors in shape $\mathcal{R}^{(N+K)\times 3}$ with forecast anchors in shape $\mathcal{R}^{M\times T_f\times 2}$, apply positional encodings as in Eq 2, and produce initial forecast queries of shape $\mathcal{R}^{(N+K)\times M\times T_f\times D}$.

Joint Streaming Forecast Transformer. To perform forecasting, the initial forecast queries first self-attend to themselves, and then cross attend to temporal forecast queries. The queries can then optionally cross-attend to the map feature tokens $\mathbf{F}_M$, integrating static scene graph information to enhance trajectory predictions. Specifically, each agent's forecasting queries locally attend to N_{map} elements nearest the agent for efficient agent-map interaction modeling. Lastly, they cross-attend to the processed detection queries, before being decoded into future waypoints of size $\mathcal{R}^{(N+K)\times M\times T_f\times 2}$ and associated confidence after the final layer. By attending to scene embeddings, past temporal detection queries, and past temporal forecast queries, the forecast transformer provides robust trajectory forecasts consistent with the current detections and past forecasts, and accounts for complex spatial patterns. This enables ForeSight to be the first work to jointly stream historical memory from both detection and forecasting by using both detection and forecast memory queues for query initialization and cross attention. Lastly, the queries are decoded and the top K forecasts are pushed into the temporal query bank for use at future times.

3.3. Optimization Approach

The ForeSight framework is trained using a multi-task objective to optimize both detection and forecasting simultaneously. For detection, a combination of focal loss [32] for classification and L1 loss for bounding box regression is used. For forecasting, the model minimizes the trajectory prediction error using an average displacement error (ADE), final displacement error (FDE), and classification loss for the mode closest to the ground truth following [16]. In addition, an auxiliary loss is used to supervise ROI feature extraction which also contains a classification and regression loss on targets in the 2D image space [56]. The joint loss function is formulated as:

$$\mathcal{L}_{\text{ForeSight}} = \lambda_{\text{det}}\,\mathcal{L}_{\text{det}} + \lambda_{\text{forecast}}\,\mathcal{L}_{\text{forecast}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} \tag{3}$$

where λ_{det}, $\lambda_{\text{forecast}}$, and λ_{aux} are balancing weights. The forecasting targets are assigned to detections using the nearest ground truth object up to a threshold of 2 meters. This unified loss and assignment encourages the model to leverage shared scene embeddings effectively while maintaining high performance in both detection and forecasting tasks. By adopting this holistic design, the streaming ForeSight architecture ensures tight integration between detection and forecasting, enabling more accurate and consistent scene understanding over time.

Methods	Backbone	Image Size	Frames	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	FPS↑
BEVDet4D [17]	ResNet50	256 × 704	2	0.322	0.457	0.703	0.278	0.495	0.354	0.206	16.7
PETRv2 [35]	ResNet50	256 × 704	2	0.349	0.456	0.700	0.275	0.580	0.437	0.187	18.9
BEVDepth [27]	ResNet50	256 × 704	2	0.351	0.475	0.639	0.267	0.479	0.428	0.198	15.7
BEVStereo [26]	ResNet50	256 × 704	2	0.372	0.500	0.598	0.270	0.438	0.367	0.190	12.2
BEVPoolv2 [18]	ResNet50	256 × 704	9	0.406	0.526	0.572	0.275	0.463	0.275	0.188	16.6
VideoBEV [13]	ResNet 50	256 × 704	2	0.422	0.535	**0.564**	0.276	0.440	0.286	0.189	-
SOLOFusion [41]	ResNet50	256 × 704	17	0.427	0.534	0.567	0.274	0.511	**0.252**	**0.181**	11.4
StreamPETR* [57]	ResNet50	256 × 704	4	0.445	0.538	0.640	0.267	0.443	0.259	0.206	**31.7**
ForeSight (ours)	ResNet50	256 × 704	4	**0.466**	**0.560**	0.614	**0.266**	**0.370**	0.258	0.201	23.5
BEVDepth [27]	ResNet101	512 × 1408	2	0.412	0.535	0.565	0.266	0.358	0.331	0.190	-
BEVFormer [29]	ResNet101-DCN	900 × 1600	4	0.416	0.517	0.673	0.274	0.372	0.394	0.198	3.0
Sparse4D [33]	ResNet101-DCN	900 × 1600	4	0.436	0.541	0.633	0.279	0.363	0.317	**0.177**	4.3
SOLOFusion [41]	ResNet101	512 × 1408	17	0.483	0.582	**0.503**	0.264	0.381	0.246	0.207	-
StreamPETR* [57]	ResNet101	512 × 1408	4	0.486	0.578	0.600	0.270	**0.343**	0.248	0.192	**6.4**
ForeSight (ours)	ResNet101	512 × 1408	4	**0.502**	**0.589**	0.578	**0.255**	0.346	**0.238**	0.193	5.1

Table 1. Detection performance on NuScenes validation set, against comparable SOTA vision-based methods. ForeSight delivers competitive performance compared to SOTA methods, achieving the highest mAP. Notably, ForeSight surpasses the second-best detection baseline, StreamPETR, with a notable 2.1% improvement in mAP. *Results generated with the same number of queries and frames as ForeSight for a fair comparison.

4. Experiments

4.1. Experiment Setup

Datasets. We evaluated ForeSight on the NuScenes 3D detection dataset [1]. NuScenes consists of 1,000 scenarios; each scenario is roughly 20s long, with a sensor frequency of 20 Hz. Each sample contains images from 6 cameras heading at different angles from the vehicle. Camera parameters including intrinsics and extrinsics are available. NuScenes provides detection annotations every 0.5s; in total there are 28k, 6k, and 6k annotated samples for training, validation, and testing, respectively. There are 10 total classes to evaluate against, including multiple vehicle, pedestrian, and cyclist classes.

The official NuScenes forecasting benchmark only provides forecast ground truth for highly-curated objects, while we aim at detecting and forecasting all objects in the sensor data. Therefore, following past works [11, 16, 39] we construct a NuScenes forecasting dataset for training and evaluation from the labeled sensor data. We leverage the available unique track identities for each detection to construct ground truth future forecasts for every detection box in the dataset. We follow the standard practice of using 12 future frames (6 seconds) for our forecasting data.

Metrics. In our experiments, we focus on key detection metrics such as mean Average Precision (mAP) to assess detection capabilities. Following the official nuScenes evaluation protocol with a 2-meter true positive distance threshold, we also provide mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (AOE), mean Average Velocity Error (mAVE), and mean Average Attribute Error (mAAE). To capture all aspects of the detection task, we report the NuScenes Detection Score (NDS), a consolidated scalar metric. We include the Frames Per Second (FPS) to evaluate the computation efficiency.

Forecasting metrics include minimum Average Displacement Error (minADE), minimum Final Displacement Error (minFDE), Miss Rate (MR), and End-to-End Prediction Accuracy (EPA)[11]. We follow common practice in existing motion forecasting benchmark [5, 49, 62] to compute the metrics over the top 6 different trajectory mode hypotheses by only evaluating the minimum distance trajectory from the ground truth. The maximum distance for the miss rate is taken as 2 meters. To compute the forecasting metrics, we first associate the current detections to the closest ground truth bounding boxes with a maximum distance threshold of 1 meter.

Implementation Details. We conduct experiments with the popular image backbones, ResNet50 and ResNet101 [14] in Table 1, to facilitate comparisons to other methods. Following past works [29, 34, 57] the ResNet model is initialized from pretrained weights trained on NuImages[1]. The map encoder was adapted from PGP [8] and converted from an agent-level encoder to a scene-level encoder. We utilize $N = 300$ detection queries and $K = 128$ temporal queries for a total of 428 queries that are decoded into detection bounding boxes. We use a temporal history of $T_H = 4$ frames and future forecast horizon for $T_f = 12$ frames. The forecast queries are generated with $M = 6$ forecast modes and permuted with the 428 detection queries across 12 future frames for a total of 30,816 forecast queries that are decoded into multimodal forecast waypoint positions.

We use 6 transformer layers to decode detection boxes and 3 transformer layers for the forecast positions. The dimension D of all embeddings is 256. For efficiency,

only the $N_{map} = 16$ map nodes closest to each detection are used for cross-attention with the map embeddings. For the weights in our multi-task objective, we use $\lambda_{det} = 1, \lambda_{forecast} = 2$. We train our model for 20 epochs using the AdamW [36] optimizer with a batch size of 16 and base learning rate of 4×10^{-4}. A cosine annealing learning rate decay is used starting at 10^{-3} with 500 warm-up iterations and a weight decay of 10^{-2} is applied.

4.2. Comparison Against State-of-the-Art

We evaluate our detection and forecasting performance on NuScenes against comparable state-of-the-art (SOTA) vision-based methods, which are shown in Table 1 and Table 2, respectively.

Detection Performance. ForeSight outperforms state-of-the-art (SOTA) methods, achieving the highest mAP and NDS—two pivotal metrics in detection benchmarks. Notably, ForeSight surpasses the best detection baseline, StreamPETR, with a notable 2.1% improvement in mAP. This achievement highlights ForeSight's enhanced precision and robustness in object detection and temporal reasoning. These findings underscore the importance of leveraging feedback from forecasting to improve detection capabilities.

Method	EPA ↑	mAP ↑	minADE ↓	minFDE ↓	MR ↓
VAD [21]	-	0.329	**0.678**	**0.882**	**0.08**
PF-Track [39]	-	0.422	2.82	3.57	-
PnPNet [16, 31]	0.222	0.382	1.15	1.95	0.226
ViP3D [11]	0.226	-	2.05	2.84	0.246
StreamPETRcv*	0.453	<u>0.486</u>	1.64	2.98	0.229
UniAD [16]	<u>0.456</u>	0.382	0.708	1.02	0.13
ForeSight (ours)	**0.549**	**0.502**	<u>0.689</u>	<u>0.938</u>	<u>0.102</u>

Table 2. End-to-end detection and forecasting performance on NuScenes val set. ForeSight achieves competitive performance against comparable SOTA methods. *Detection results with constant velocity forecasting.

Forecast Performance. We compare our forecasting results to other open-sourced forecasting methods that operate directly from perception inputs. For end-to-end prediction accuracy [11] (EPA) our framework achieves the best performance, beating UniAD by a substantial 9.3% due to higher forecast hit rate and lower false positive detection rate. This highlights the broad applicability of ForeSight in end-to-end autonomous driving systems by joint streaming detection and forecasting for superior performance.

We also show that Foresight maintains minADE prediction performance of other top methods but on a much larger set of real trajectories due to our more accurate perception matching more detections to the ground truth. The only method that outperforms ours in minADE forecasting, VAD, also exhibits the worst detection performance. This underscores a key limitation of directly comparing forecasting performance across different perception approaches. Stronger detectors capture more challenging objects, potentially skewing prediction comparisons. As a result, evaluating minADE alone unfairly penalizes improved detection by introducing a more diverse and difficult sample set. Another challenge in forecasting evaluation from perception is the limited number of forecasting-from-sensor methods, each relying on different image backbones and perception networks, which can influence downstream forecasting performance. Establishing a fair and unified benchmark remains an important direction for future research.

4.3. Ablation Study and Analysis

We investigate the design choices made in ForeSight's architecture and present the analysis in Table 3. We start with a baseline detector based on the PETR model [34] with the ResNet-50 backbone and subsequently add the components that make up our approach of bidirectional learning for joint detection and forecasting.

Detect Frames	Forecast		Propogate		HD Map	Metrics	
	TF	DF	DP	FP		mAP ↑	minADE ↓
1						0.361	-
1	✓					0.361	2.83
1		✓				0.361	1.05
4			✓			0.440	-
4		✓		✓		<u>0.463</u>	<u>0.735</u>
4		✓		✓	✓	**0.466**	**0.709**

Table 3. Ablation of design decisions on nuScenes validation set. Starting from a baseline single-frame detector [34] we compare tracking-based forecasting (TF) vs detection-based forecasting (DF), direct propagation (DP) vs forecast-based propagation (FP), and offline HD map usage with the full ForeSight system in the last row.

Tracking-Free vs. Tracking-Aware Forecast. Tracking-based forecasting is a common forecasting approach [8] which is trained on ground truth trajectory histories and therefore requires accurate track data to produce good results. When applying these models to real tracks generated from our detections, we find that performance is significantly degraded due to the lack of robustness to errors in the tracks . Instead, we rely on a detection-based forecasting (DF) approach which substantially improves minADE by 1.78 m compared to a similar tracking-based forecasting (TF) approach trained on the nuScenes prediction dataset. By removing the tracking bottleneck, the forecasting model is significantly more robust to perception errors and shows better results for the challenging driving scenarios contained in the forecasting data we generated from NuScenes. ForeSight's tracking-free design facilitates a more streamlined information flow within the perception-forecasting pipeline.

Impact of HD Map. ForeSight is optionally able to make

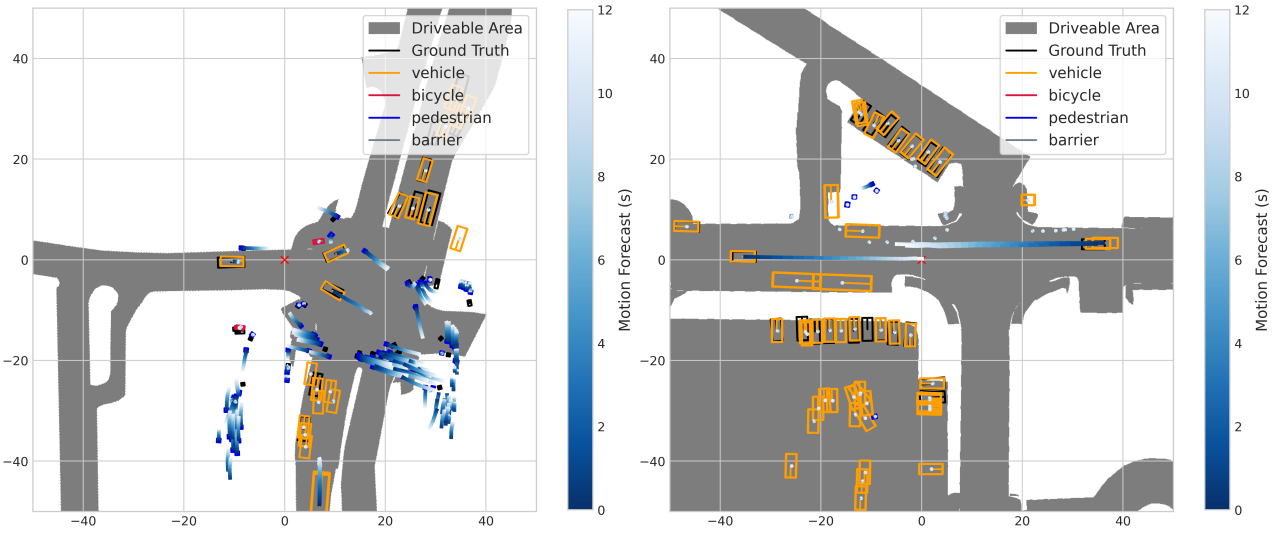

Figure 3. Qualitative case with many pedestrians crossing (left) and occluded parked cars (right). ForeSight detects the crossing pedestrians and forecasts their motion while predicting that the vehicles will stay stationary to yield to the pedestrians. ForeSight can also detect limited visibility vehicles using previous observations.

use of an HD map for forecasting via cross-attention to encoded map queries. While detection does not directly leverage HD maps, it benefits indirectly through improved forecast propagation. Table 3 shows the HD map usage can slightly improve forecasting performance with a reduction of 0.026 minADE, highlighting its role in enhancing forecasting and indirectly improving detection.

Impact of Bidirectional Propagation. We evaluate the effectiveness of ForeSight's bidirectional query propagation, which enhances temporal reasoning and maintains query association across frames. By propagating queries between detection and forecasting, ForeSight outperforms traditional sparse-query models that propagate only forward. Specifically, using naïve direct detection query propagation, following the StreamPETR [57] approach, we achieve a detection mAP of 0.440, a 2.6% mAP improvement over the baseline single-frame baseline detector. We show that we can further improve performance on the detection task to 0.463 mAP by propagating the forecast query back to the detection task. Together these enhancements show a clear benefit to ForeSight's bidirectional query propagation and joint learning for detection and forecasting.

4.4. Qualitative Results

Qualitative results are provided to visualize the outputs of ForeSight and better understand the model's ability to process temporal information and use it effectively. Figure 3 shows a case where many static vehicles are parked. By using past detections and correctly predicting that they will be stationary the model can accurately detect parked cars even when they are heavily occluded. In addition, the model correctly forecasts moving vehicles and stationary vehicles.

Figure 3 also shows complex behavior with a large group of pedestrians. We can see that the model can accurately forecast the social interaction between these clusters of dynamic objects and correctly predicts their motion together. Even though many of these objects are obstructed at the given time, they have been seen in the past and the model correctly detects they are still present in the group. In addition, the correct conditional motion is predicted such that the vehicles in the scene are yielding to the pedestrians and cyclists. These visualizations further demonstrate ForeSight's strength in detecting partially obstructed objects and predicting object trajectory, reinforcing its innovative ability to unlock improved performance.

5. Conclusion

In this work, we present ForeSight, a unified framework for joint detection and forecasting in vision-centric autonomous driving. Our approach bridges the gap between perception and prediction by introducing a bidirectional learning mechanism and joint streaming memory, where object queries propagate seamlessly between detection and forecasting. This design not only enhances detection accuracy in complex driving scenarios but also improves forecasting robustness to real-world perception uncertainties. Extensive evaluations on the nuScenes dataset validate ForeSight's effectiveness, achieving state-of-the-art performance for detection. With bidirectional query propagation and a tracking-free architecture, ForeSight advances autonomous vehicle perception, enhancing efficiency, reliability, and safety in real-world navigation.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 6

[2] Sergio Casas, Wenjie Luo, and Raquel Urtasun. Intentnet: Learning to predict intention from raw sensor data. In *Conference on Robot Learning*, pages 947–956. PMLR, 2018. 3

[3] Sergio Casas, Ben Agro, Jiageng Mao, Thomas Gilles, Alexander Cui, Thomas Li, and Raquel Urtasun. Detra: A unified model for object detection and trajectory forecasting. *arXiv preprint arXiv:2406.04426*, 2024. 3

[4] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*, 2019. 3, 5

[5] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019. 6

[6] Brian Cheong, Jiachen Zhou, and Steven Waslander. Jdt3d: Addressing the gaps in lidar-based tracking-by-attention. *arXiv preprint arXiv:2407.04926*, 2024. 3

[7] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2090–2096. IEEE, 2019. 3

[8] Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In *Conference on Robot Learning*, pages 203–212. PMLR, 2022. 3, 6, 7

[9] Lan Feng, Quanyi Li, Zhenghao Peng, Shuhan Tan, and Bolei Zhou. Trafficgen: Learning to generate diverse and realistic traffic scenarios. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3567–3575. IEEE, 2023. 3

[10] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11525–11533, 2020. 3

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5496–5506, 2023. 1, 3, 6, 7

[12] Xunjiang Gu, Guanyu Song, Igor Gilitschenski, Marco Pavone, and Boris Ivanovic. Accelerating online mapping and behavior prediction via direct bev feature attention. *arXiv preprint arXiv:2407.06683*, 2024. 3

[13] 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*, 2024. 5, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 6

[15] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15273–15282, 2021. 2

[16] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1, 2, 3, 5, 6, 7

[17] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 1, 2, 6

[18] Junjie Huang and Guan Huang. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment. *arXiv preprint arXiv:2211.17111*, 2022. 6

[19] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2

[20] Boris Ivanovic, Kuan-Hui Lee, Pavel Tokmakov, Blake Wulfe, Rowan Mcllister, Adrien Gaidon, and Marco Pavone. Heterogeneous-agent trajectory forecasting incorporating class uncertainty. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12196–12203. IEEE, 2022. 3

[21] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 3, 7

[22] Sanmin Kim, Youngseok Kim, In-Jae Lee, and Dongsuk Kum. Predict to detect: Prediction-guided 3d object detection using sequential images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18057–18066, 2023. 1

[23] Youngwan Lee, Joong-won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and gpu-computation efficient backbone network for real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 12

[24] Quanyi Li, Zhenghao Mark Peng, Lan Feng, Zhizheng Liu, Chenda Duan, Wenjie Mo, and Bolei Zhou. Scenarionet:

Open-source platform for large-scale traffic scenario simulation and modeling. *Advances in neural information processing systems*, 36, 2024. 3

[25] Xirui Li, Feng Wang, Naiyan Wang, and Chao Ma. Frame fusion with vehicle motion prediction for 3d object detection. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4252–4258. IEEE, 2024. 3

[26] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1486–1494, 2023. 6

[27] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 1, 6

[28] Yingwei Li, Charles R Qi, Yin Zhou, Chenxi Liu, and Dragomir Anguelov. Modar: Using motion forecasting for 3d object detection in point cloud sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9329–9339, 2023. 3

[29] 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 *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 2, 6

[30] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 541–556. Springer, 2020. 3

[31] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2020. 3, 7

[32] T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017. 5

[33] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d v2: Recurrent temporal fusion with sparse model. *arXiv preprint arXiv:2305.14018*, 2023. 1, 2, 6

[34] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 2, 3, 4, 6, 7

[35] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. 1, 2, 6

[36] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[37] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018. 3

[38] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 5

[39] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17928–17938, 2023. 6, 7

[40] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3142–3152, 2021. 12

[41] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris M Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 6

[42] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 2

[43] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. 2

[44] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer, 2020. 3

[45] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*, pages 726–737. PMLR, 2023. 3

[46] Hao Shao, Letian Wang, Ruobing Chen, Steven L Waslander, Hongsheng Li, and Yu Liu. Reasonnet: End-to-end driving with temporal and global reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13723–13733, 2023. 3

[47] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024. 3

[48] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and lo-

cal movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022. 5

[49] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 6

[50] 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)*, pages 7814–7821. IEEE, 2022. 1, 3

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1, 2

[52] Letian Wang, Yeping Hu, Liting Sun, Wei Zhan, Masayoshi Tomizuka, and Changliu Liu. Hierarchical adaptable and transferable networks (hatn) for driving behavior prediction. *arXiv preprint arXiv:2111.00788*, 2021. 3

[53] Letian Wang, Liting Sun, Masayoshi Tomizuka, and Wei Zhan. Socially-compatible behavior design of autonomous vehicles with verification on real human data. *IEEE Robotics and Automation Letters*, 6(2):3421–3428, 2021. 3

[54] Letian Wang, Yeping Hu, Liting Sun, Wei Zhan, Masayoshi Tomizuka, and Changliu Liu. Transferable and adaptable driving behavior prediction. *arXiv preprint arXiv:2202.05140*, 2022. 1

[55] Letian Wang, Jie Liu, Hao Shao, Wenshuo Wang, Ruobing Chen, Yu Liu, and Steven L Waslander. Efficient reinforcement learning for autonomous driving with parameterized skills and priors. *arXiv preprint arXiv:2305.04412*, 2023. 3

[56] Shihao Wang, Xiaohui Jiang, and Ying Li. Focal-petr: Embracing foreground for efficient multi-camera 3d object detection. *IEEE Transactions on Intelligent Vehicles*, 9(1): 1481–1489, 2023. 5

[57] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023. 1, 2, 3, 6, 8, 12, 13

[58] Xishun Wang, Tong Su, Fang Da, and Xiaodong Yang. Prophnet: Efficient agent-centric motion forecasting with anchor-informed proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21995–22003, 2023. 3

[59] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2, 4

[60] Xinshuo Weng, Boris Ivanovic, Kris Kitani, and Marco Pavone. Whose track is it anyway? improving robustness to tracking errors with affinity-based trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6573–6582, 2022. 3

[61] Xinshuo Weng, Boris Ivanovic, and Marco Pavone. Mtp: Multi-hypothesis tracking and prediction for reduced error propagation. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1218–1225. IEEE, 2022. 3

[62] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 6

[63] Yihong Xu, Loïck Chambon, Éloi Zablocki, Mickaël Chen, Alexandre Alahi, Matthieu Cord, and Patrick Pérez. Towards motion forecasting with real-world perception inputs: Are end-to-end approaches competitive? In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18428–18435. IEEE, 2024. 1, 3

[64] Luyao Ye, Zikang Zhou, and Jianping Wang. Improving the generalizability of trajectory prediction models with frenet-based domain normalization. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11562–11568. IEEE, 2023. 3

[65] Yang Zhou, Hao Shao, Letian Wang, Steven L Waslander, Hongsheng Li, and Yu Liu. Smartrefine: A scenario-adaptive refinement framework for efficient motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15281–15290, 2024. 1, 3

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

The supplementary material includes four key sections designed to provide a comprehensive evaluation of ForeSight. Section A presents additional experimental results with varied backbones and configurations to demonstrate the model's scalability and robustness. Section B offers a detailed class-wise analysis, highlighting ForeSight's superior performance across diverse object categories, including challenging cases such as trailers and motorcycles. Section C showcases qualitative results, illustrating ForeSight's ability to leverage temporal and spatial information for accurate detection and forecasting in dynamic and occluded environments. Section D discusses the method's limitations, including dependencies on high-quality data, challenges in adverse weather, the need for standardized forecasting benchmarks, and adaptability to dynamic scenes, while proposing avenues for future improvement. These sections collectively validate ForeSight's effectiveness, versatility, and areas for refinement.

The reviewers can access the code here for additional information on our implementation: https://anonymous.4open.science/r/ForeSight. The authors intend to publicly release the code alongside the release of the final manuscript.

A. Detection Performance Analysis

We highlight ForeSight's enhanced performance on challenging scenarios like low-visibility and occluded objects, critical for safe autonomy. While these cases are a small portion of the data, they are crucial long-tail challenges. Leveraging past forecasts, our model can more effectively detect objects with limited sensor visibility and coverage. Quantitative analysis with the ResNet-50 configuration of ForeSight confirms improved performance on low-visibility objects, with $<$40% of the object visible, and achieve a 0.9% higher Average Recall (AR) on these challenging objects relative to the baseline as seen in Table 4.

Methods	**0-0.4**	**0.4-0.6**	**0.6-0.8**	**0.8-1.0**
StreamPETR [59]	0.401	0.455	0.448	0.468
ForeSight (ours)	**0.410**	**0.461**	0.451	0.468

Table 4. Detection performance (AR) for object visibility level.

B. Additional Backbone Results

We conduct additional experiments with the popular image backbone, v2-99 [23] in Table 5, to further explore our results. Following past works [57] the V2-99 model is initialized from a pretrained checkpoint on DD3D [40].

ForeSight delivers improved performance compared to the SOTA StreamPETR method with the larger backbone, achieving a higher mAP and NDS—two pivotal metrics in detection benchmarks. ForeSight surpasses StreamPETR with a notable 1.9% improvement in mAP. This highlights ForeSight's improved temporal reasoning also scales with different types of upstream networks.

C. Class-wise Analysis

In Table 6, we present a detailed comparison of class-wise performance for the V2-99 backbone on the NuScenes validation set. We show results for AP at the 2.0m threshold. ForeSight consistently outperforms the baseline comparison across most object classes, demonstrating its robustness and versatility in detecting a diverse range of objects in dynamic driving scenarios.

Significant improvements are observed in challenging classes such as trailers and motorcycles, where the mAP gains are 15.0% and 1.8%, respectively. These categories often suffer from lower detection rates due to less frequent appearance in the data, irregular shapes, and variability in motion patterns. By leveraging enhanced temporal modeling and spatial context, ForeSight can provide a more accurate and reliable detection for these difficult cases.

Additionally, for high-performance classes such as cars and pedestrians, ForeSight continues to achieve competitive or superior results, maintaining its overall edge in performance. This demonstrates that the proposed method not only excels in challenging scenarios but also scales effectively across well-represented object categories.

These results underscore ForeSight's ability to generalize across object types and validate its superiority in real-world applications requiring precise multi-class detection. The improved performance on underrepresented or challenging object classes further highlights the method's enhanced temporal reasoning and adaptability.

D. Limitations

Limited Comparisons for Forecasting. Due to the lack of standardized forecasting-from-perception benchmarks, our evaluation relies on adapting the NuScenes detection and tracking dataset following the approach of past works. Direct forecasting comparison to other end-to-end methods is challenging due to differing configurations (e.g. backbones, perception models) and a lack of standardized evaluation frameworks. Future work will explore establishing a common benchmark to evaluate similar methods with common upstream models to provide a proper fair comparison.

High-Quality Data Dependancy. The effectiveness of ForeSight may depend on the quality of the input data. Since we rely on tight coupling of temporal information errors in camera calibration, localization, or map inaccuracies can propagate through the pipeline, potentially degrading both detection and forecasting outcomes. Due to the feedback loop in the pipeline, it could also be more susceptible to these errors that deteriorate performance. The robustness

Methods	Backbone	Image Size	Frames	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
StreamPETR* [57]	V2-99	900 × 1600	4	0.479	0.569	0.622	0.263	**0.364**	0.256	0.194
ForeSight (ours)	V2-99	900 × 1600	4	**0.489**	**0.577**	**0.598**	**0.259**	0.376	**0.250**	0.194

Table 5. Detection performance on NuScenes validation set for an additional backbone and image size. ForeSight delivers a higher mAP and NDS against the comparable SOTA vision-based method StreamPETR. ForeSight surpasses the second-best detection baseline, StreamPETR, with a notable 2.0% improvement in mAP. *Results generated with the same number of queries, frames, and resolution as ForeSight for a fair comparison.

Methods	Backbone	Car↑	Pedestrian↑	Bicycle↑	Bus↑	Motorcycle↑	Trailer↑	Truck↑
StreamPETR* [57]	V2-99	0.810	0.729	0.603	0.710	0.627	0.366	**0.628**
ForeSight (ours)	V2-99	**0.812**	**0.731**	**0.608**	**0.750**	**0.638**	**0.421**	0.607

Table 6. Detection performance on NuScenes validation set comparing class-wise performance based on AP with a 2.0 meter distance threshold. We observe that our model improves on nearly all classes of objects and performs significantly better on challenging classes with lower performance such as trailers, motorcycles, and bicycles.

of errors could be explored in future work along with mitigation strategies.

Adverse Weather. A potential limitation of ForeSight could also be sensitivity to adverse weather conditions such as heavy rain, snow, or fog. These conditions can degrade the quality of camera sensor inputs by obscuring object boundaries, reducing contrast, and introducing noise. As ForeSight relies heavily on vision-based perception, any reduction in image quality directly impacts both detection and forecasting accuracy. Future work could explore the integration of additional sensor modalities, such as LiDAR or radar, which are more robust to weather-induced impairments. Robust data augmentation strategies simulating adverse weather conditions during training may also improve the resilience of the model in such challenging environments.

Simplified Scene Representations: The method's reliance on predefined object categories limits its adaptability to environments with novel object classes not represented in training data. The ability to segment or detect map elements online has also been explored in other works and could be integrated into this method instead of using an offline map or no map.