Supplementary Material

The supplementary material includes four key sections designed to provide a comprehensive evaluation of ForeSight. Section A presents additional experimental analysis on performance in challenging scenarios.Section B provides additional ablation results for design decisions. Section C offers a detailed class-wise analysis, highlighting ForeSight's superior performance across diverse object categories, including challenging cases such as trailers and motorcycles. 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.

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 using average recall across different object visibility levels.

B. Additional Ablations

Ablation for historical frame count. Table 5 presents additional experiments varying historical frames. Performance improves with more frames, with diminishing gains beyond 4 frames. In addition, as many prior works also use 4 frames, this choice enables fair comparison.

Ablation replacing historical queries with spatial queries. In Table 6, replacing historical queries with spatial ones (first row) while keeping 900 total detection queries results in a reduction of 10.5% mAP and increase of 3.4 m minADE compared to our model (last row), isolating the benefit of temporal information from query count.

Study of the relation between forecasting errors and detection accuracy. We conduct additional experiments varying the query propagation mechanism (Table 6). Starting from a detection memory queue with stationary past queries, adding forecasting decoder (FD) layers improves forecasting accuracy, which in turn enhances detection, highlighting the synergy between tasks.

Frames	**Det. Queries**	**CA Queries**	**mAP ↑**
0	900 (900+0)	0	0.361
1	900 (644+256)	256	0.419
2	900 (644+256)	512	0.448
4	900 (644+256)	1024	0.466
8	900 (644+256)	2048	**0.467**

Table 5. Ablation of historic frames, detection queries (initialized + propagated), and cross attention (CA) historic queries.

Query Propogation	**mAP ↑**	**minADE ↓**
Constant Pos.	0.445	4.13
FD 1 Layer	0.458	0.979
FD 3 Layers	**0.466**	**0.709**

Table 6. Ablation of query propagation approaches used for historical detection queries.

C. Class-wise Analysis

In Table 7, 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.

Methods	Backbone	Car↑	Pedestrian↑	Bicycle↑	Bus↑	Motorcycle↑	Trailer↑	Truck↑
StreamPETR* [60]	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 7. 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.

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 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.