A Training Details

Stage	Prediction		BEV Feature Attention				
Map Model	LR	Weight Decay	Patch Size	MLP_{dim}	Depth	Heads	Head_{dim}
MapTR [22]	5E-4	1E-4	$(20, 10)$	512	6	16	64
MapTRv2 [23]	3.5E-4	1E-2	$(20, 10)$	64	6	16	64
MapTRv2-Centerline [23]	3.5E-4	1E-2	$(20, 20)$	64	4	12	32
StreamMapNet [37]	3.5E-4	1E-3	$(10, 5)$	128	6	16	64

Table 3: The hyperparameters used when training HiVT [40] with various online mapping models in Sec. 3.1, where agent-lane attention is replaced with agent-BEV attention.

Map Model	LR	Weight Decay	Dropout
MapTR [22]	1.5E-4	0.05	0.2
MapTRv2 [23]	1.5E-4	0.05	0.2
MapTRv2-Centerline [23]	2E-4	0.05	0.2

Table 4: The hyperparameters used when training DenseTNT [12] with various online mapping models in Sec. 3.2, where lane vectors are enhanced with BEV grid features.

Prediction		BEV Feature Attention				
LR	Weight Decay	Patch Size	MLP_{dim}	Depth	Heads	Head_{dim}
5E-4	1E-2	$(10, 5)$	128	6	16	64

Table 5: The hyperparameters used when training DenseTNT [12] with StreamMapNet [37] in Sec. 3.3, where agent information is replaced with temporal BEV feature attention.

A.1 Data Preprocessing

To ensure a fair comparison across different map estimation and prediction models, we unify the orientations of the BEV features and the resulting estimated map. The scene is centered at the ego-vehicle frame, with the positive y-axis aligned with the forward-moving direction, and the positive x-axis aligned with the right side of the ego-vehicle. The BEV features are adjusted accordingly. The perception range $(H \times W)$ is $60m \times 30m$. Due to the limits of AV perception, we only predict for agents within this perception range.

A.2 Model Training

To address the potential variability in convergence rates between different integration approaches and map-prediction combinations, each model is individually adjusted to optimize performance. The BEV dimension is 200×100 for MapTR models [22, 23] and 100×50 for StreamMapNet [37].

During training, online map estimation models are trained first as in [13]. This produces map element polylines with corresponding uncertainties. During inference, we also save the BEV features produced by the trained models, providing the necessary data for all three settings for prediction: Baseline, where only lane vectors are used; Uncertainty, where uncertainty is incorporated; and our approach, where BEV features are incorporated. After we obtain this modified dataset, HiVT [40] and DenseTNT [12] are trained following the different strategies in Sec. 3.

In Sec. 3.1, HiVT's local encoder is modified by replacing agent-lane features with agent-BEV features. As seen in Tab. 3, we reduce the attention module size as the complexity of the mapping model increases. This adjustment is shown via the decrease in MLP layer size and head dimension across the MapTR series. The increase in BEV patch size for MapTRv2-Centerline compared to MapTRv2 also indicates a coarser feature representation. The output dimension of the attention module is adjusted to match the original agent-lane feature dimension, ensuring compatibility with the HiVT's global interaction module.

For the approach in Sec. 3.2, we tune the prediction training hyperparameters to accommodate the additional information provided by BEV features, as seen in Tab. 4. Due to the increased complexity of input data, the learning rate is reduced to the order of 10^{-4} and weight decay is increased to 0.05 from 0.01 to ensure smooth training convergence. Dropout is also increased slightly from 0.1 to 0.2. When encoding lane information in the point-level subgraph of Vectornet, the hidden layer size is doubled to accommodate the extra BEV features after concatenating them with the original raw lane vertices.

The hyperparameter choices for Sec. 3.3 are shown in Tab. 5. Prediction model values are adjusted in the same way as Tab. 4, with a smaller learning rate to ensure convergence. For the BEV attention module, the hyperparameter choices are the same as in the corresponding row of Tab. 3.

B Additional Quantitative Comparisons

B.1 Runtime Comparisons

Below, runtime is measured on an RTX 4090 GPU from when raw RGB camera images are input to when trajectories are produced.

Model Combination	Base (ms)	Ours (ms)
HiVT + MapTR	22.4	9.1
HiVT + MapTRv2	26.7	13
HiVT + StreamMapNet	33.6	29.4

C Additional Visualizations

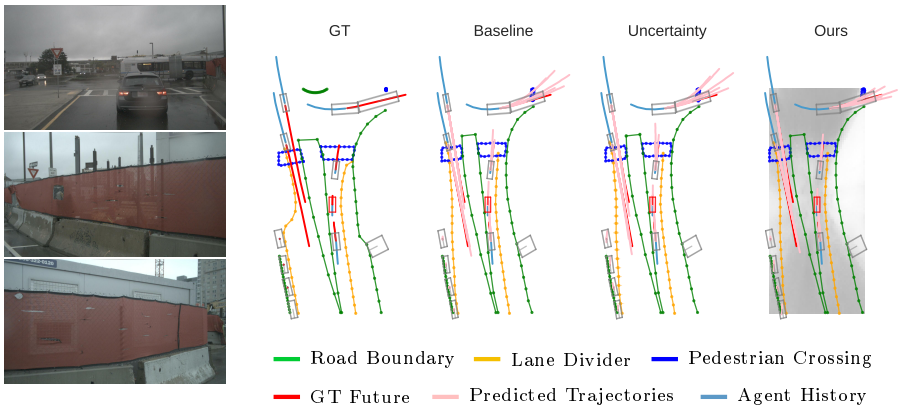

Fig. 7: StreamMapNet [37] and HiVT [40] combined using the strategy in Sec. 3.1. By replacing lane information with temporal BEV features, HiVT is able to better predict stopping behavior, avoiding overshooting the GT (as in the Baseline and Uncertainty-enhanced approach).

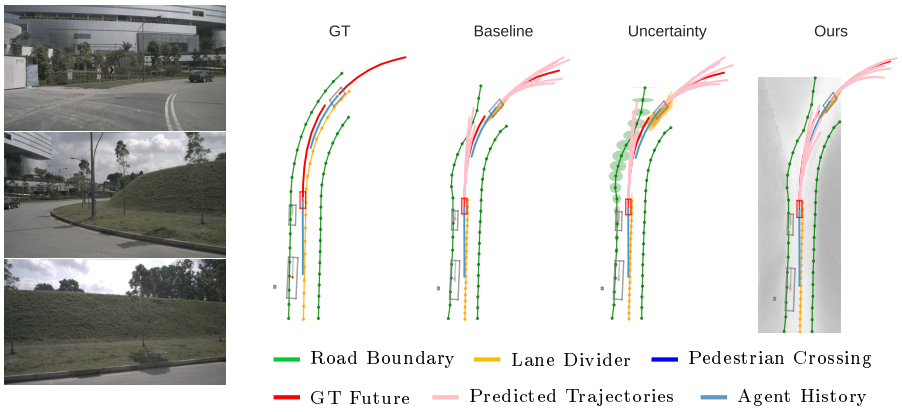

Fig. 8: StreamMapNet [37] and HiVT [40] combined using the strategy in Sec. 3.1. By replacing lane information with temporal BEV features, HiVT's predictions respect boundaries, in contrast to both the Baseline and Uncertainty-enhanced approaches which deviate outside the green road boundary. Further, our approach's predicted trajectories align more closely to the GT.

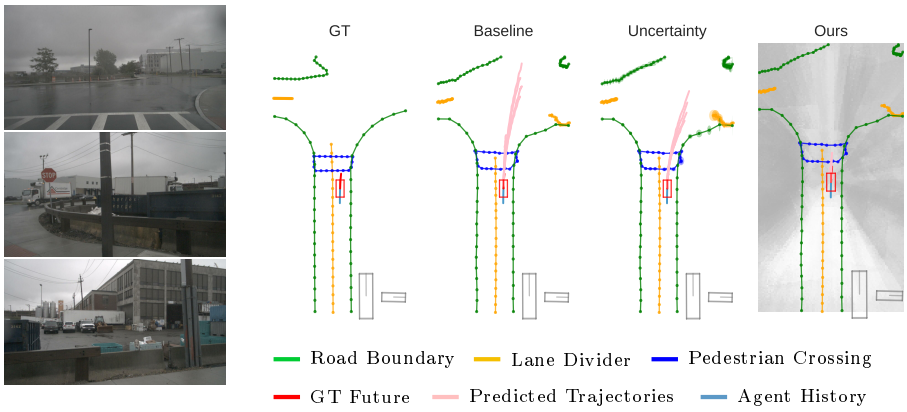

Fig. 9: MapTR [22] and DenseTNT [12] combined via the strategy in Sec. 3.2. Our augmentation of map vertices with BEV features enables DenseTNT to produce accurate trajectories, preventing overshooting at an intersection as seen in the Baseline and Uncertainty-enhanced setups.

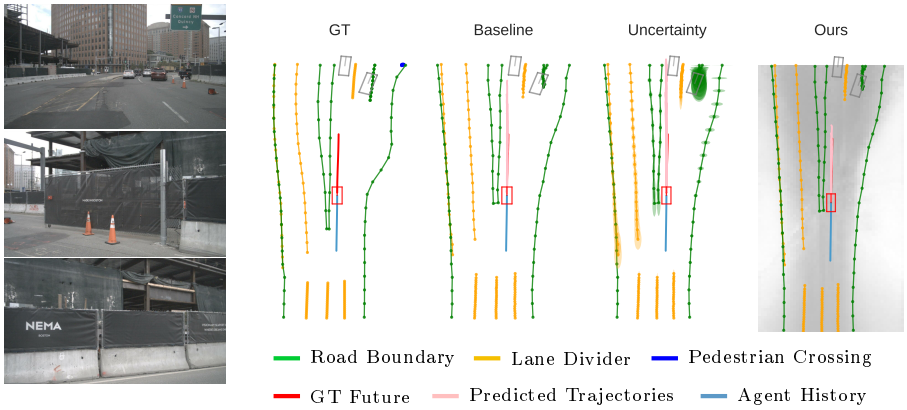

Fig. 10: StreamMapNet [37] and DenseTNT [12] combined using the strategy in Sec. 3.3. By replacing agent trajectory information with BEV features, DenseTNT is able to predict more accurate trajectories, compared to the significant overshooting outputs from the Baseline and Uncertainty-enhanced approaches.