Spatial Retrieval Augmented Autonomous Driving

Xiaosong Jia[1,2*], Chenhe Zhang[1,2*], Yule Jiang[3*], Songbur Wong[3*],
Zhiyuan Zhang[3], Chen Chen[4], Shaofeng Zhang[5], Xuanhe Zhou[3], Xue Yang[3†],
Junchi Yan[3†], Yu-Gang Jiang[1,2]

1. Institute of Trustworthy Embodied AI, Fudan University
2. Shanghai Key Laboratory of Multimodal Embodied AI
3. Shanghai Jiao Tong University
4. Key Laboratory of Target Cognition and Application Technology,
Aerospace Information Research Institute, Chinese Academy of Sciences
5. University of Science and Technology of China

*Equal contributions †Correspondence authors

https://github.com/SpatialRetrievalAD

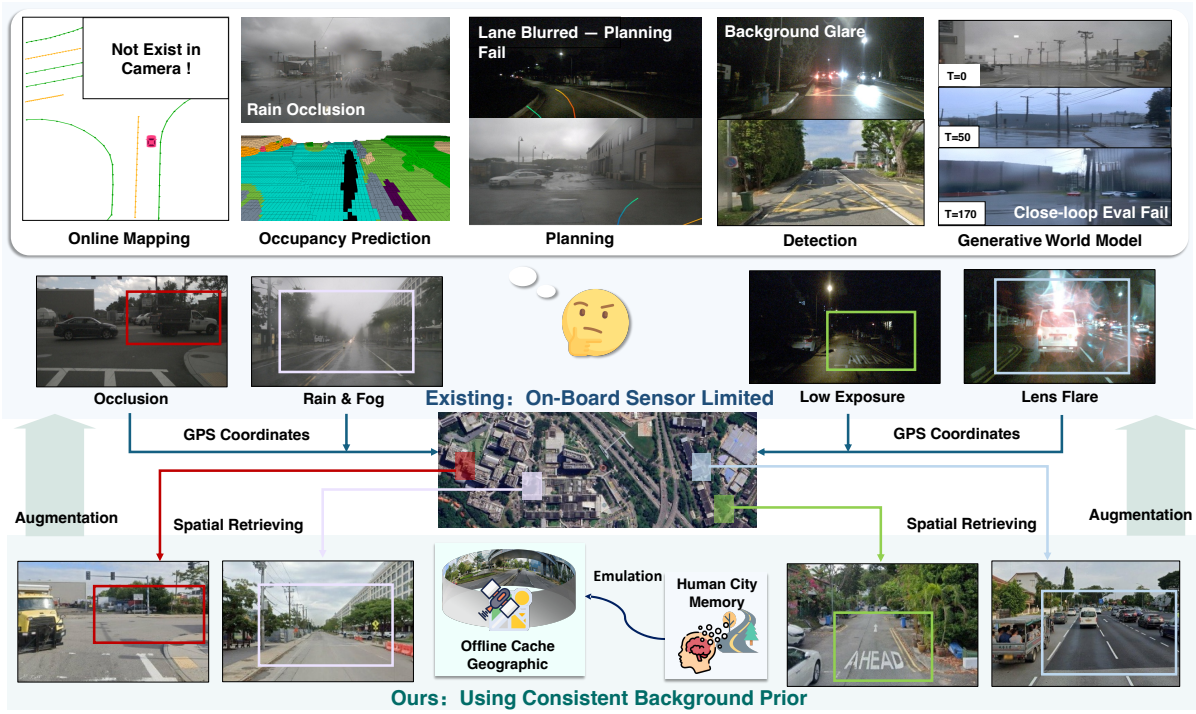

Figure 1. Existing autonomous driving systems (upper) largely rely on onboard sensors, which are vulnerable to perceptual conditions. Inspired by human drivers' ability to recall memory of previously seen roads, we introduce the spatial retrieval paradigm (lower) which utilizes offline cached geographic images as an extra input modality to enhance the performance of multiple AD tasks.

Abstract

Existing autonomous driving systems rely on onboard sensors (cameras, LiDAR, IMU, etc) for environmental perception. However, this paradigm is limited by the drive-time perception horizon and often fails under limited view scope, occlusion or extreme conditions such as darkness and rain. In contrast, human drivers are able to recall road structure even under poor visibility. To endow models with this "recall" ability, we propose the spatial retrieval paradigm, introducing offline retrieved geographic images as an additional input. These images are easy to obtain from offline caches (e.g., Google Maps or stored autonomous driving datasets) without requiring additional sensors, making it a

plug-and-play extension for existing AD tasks.

For experiments, we first extend the nuScenes dataset with geographic images retrieved via Google Maps APIs and align the new data with ego-vehicle trajectories. We establish baselines across five core autonomous driving tasks: object detection, online mapping, occupancy prediction, end-to-end planning, and generative world modeling. Extensive experiments show that the extended modality could enhance the performance of certain tasks. We will open-source dataset curation code, data, and benchmarks for further study of this new autonomous driving paradigm.

1. Introduction

Modern autonomous driving (AD) methods rely on onboard sensors, including cameras, LiDAR, and IMU, to capture environmental information [9, 40]. While this paradigm already achieves strong performance [17, 29, 54], their inputs are **inherently constrained by the limited range and line-of-sight nature of online sensing**. As a result, in visually challenging scenarios such as limited view scope, occlusion, extreme exposure, or adverse weather conditions (rain, snow, fog, etc), the system performance can consequently degrade significantly [28, 66], as in Fig. 1. For instance, tasks like online mapping [34, 39, 63, 64] and occupancy prediction [7, 20, 32, 47, 62] aim to estimate scene structure and limited visibility or occlusions can degrade their recognition of environment and subsequently impair planning [26]. Similarly, recent AD world models [8, 14, 60] struggle to generate novel scenes when the ego-vehicle deviates significantly from recorded logs [48, 60], a limitation tied to the small scope of onboard views, restricting their capability to serve as a simulator for closed-loop evaluation and reinforcement learning.

On the contrary, human drivers recall recent memory of the scene when current visual input is insufficient [42]. In this work, **we aim to enhance AD stacks with broader context beyond the immediate range of onboard sensors by spatial retrieval**. The spatial geographic data could be from Google Maps, where street view and satellite images with latitude and longitude are provided. For autonomous driving companies, their offline cached dataset could also be used. Unlike onboard sensors [1, 3, 16, 23], these geographic data are offline, globally accessible, and unaffected by drive-time disturbances. They provide rich contextual cues from perspectives beyond the ego vehicle, offering a cost-effective way to enrich spatial context without additional sensors or manual annotations.

To systematically investigate this new paradigm, we first build a framework to integrate geographic data into existing autonomous driving datasets. The framework automates data collection and spatial alignment via Google Maps APIs [37] and ego-pose information to fetch and align coordinate systems. Using this framework, we then extend the nuScenes dataset [3] with corresponding geographic images and coordinate-based spatial retrieval APIs. Finally, to study the effects of this new modality, we establish baselines across five key AD tasks: object detection, online mapping, occupancy prediction, end-to-end planning, and generative world modeling. We design a plug-and-play adapter to seamlessly integrate the geographic images into existing models. **Extensive experiments demonstrate this modality improves performance across tasks**.

Our main contributions are summarized as follows:

- We introduce the **spatial retrieval paradigm** for AD, which mitigates the environmental sensitivity of onboard perception and supplies broad, far-horizon context.
- We construct an **extended nuScenes dataset** - nuScenes-Geography with geographic images and spatial retrieval APIs, enabling systematic study of the new paradigm.
- We design a **model-agnostic adapter** and establish baselines across five AD tasks, demonstrating the broad applicability provided by the new modality.
- We will **open-source** data curation pipelines, extended dataset, and all baselines to facilitate future research.

2. Related Work

Autonomous Driving Paradigms. Despite recent advances in end-to-end pipelines [17, 21–25, 46, 53, 57, 58], multi-sensor fusion [4, 9, 11, 24, 33, 40, 45, 65, 67], and temporal modeling [31, 36, 43, 49, 59, 63], these methods are constrained by limited drive-time perception horizon, lacking long-range, location-specific priors. For offline generative world models, existing methods [8, 13, 15, 30, 41, 56] are prone to hallucinate when driving deviation is large due to the lack of spatial grounding. The proposed geographic image modality offers strong spatial information.

Relation to HD Maps. HD maps [10, 12] provide centimeter-level geometric accuracy but are expensive to annotate and maintain, and only encode pre-defined types of information (e.g., lane topology). In contrast, geographic images are easy to collect, provide rich visual details beyond geometry—such as vegetation, building facades, and road textures—and are beneficial for tasks like generative world modeling and occupancy prediction where such visual context is valuable. Our spatial retrieval paradigm complements rather than replaces HD maps.

Retrieval Methods in Autonomous Driving. Retrieval has been used for visual place recognition [51, 52], rule understanding with LLM [5], localization [18], and planning trajectories sampling [6] in autonomous driving. Prior traversal data have also been leveraged for visual odometry [2], online mapping with neural map priors [55], and 3D detection with historical LiDAR [61]. Unlike these task-specific uses of prior data, the proposed spatial retrieval paradigm fetches location-aligned geographic images as

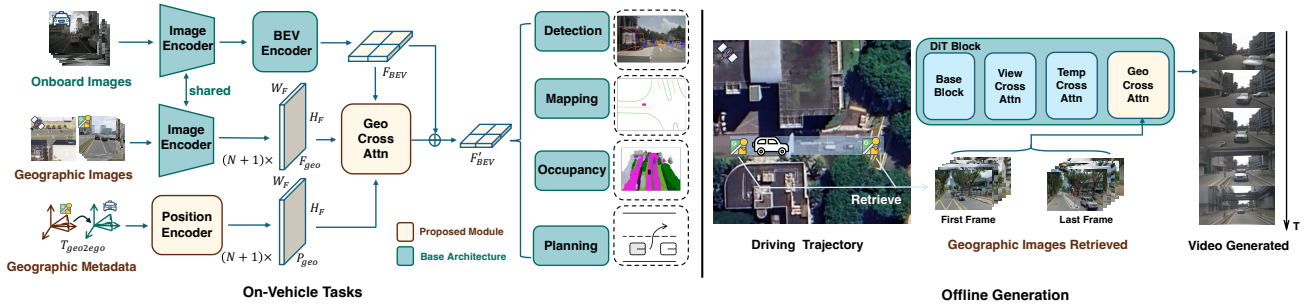

Figure 2. **Spatial Retrieval Adapter.** We adopt cross-attention with standard BEV features ($\mathbf{F}_{\text{BEV}}$) as query and the retrieved geographic features ($\mathbf{F}_{\text{geo}}$) plus corresponding 3D positional encodings ($\mathbf{P}_{\text{geo}}$) as key and value. For generative world model task, we use similar cross-attention architecture, with the noised latents as query. Further, since the whole driving trajectory is known for video diffusion, we retrieve the corresponding geographic images based on the start and end frame positions so that the background becomes consistent.

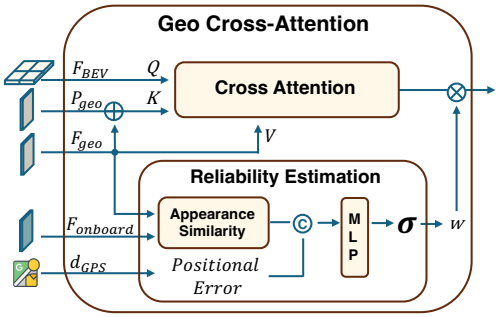

Figure 3. **Reliability Estimation Gate.** We set a Reliability Estimation Gate: when retrieved geographic missing or misaligned, the residual update approximates zero based on difference between pose and image feature.

general-purpose complementary sensory inputs applicable across multiple AD tasks.

3. Methodology

3.1. Spatial Retrieval Paradigm & Task Definition

Suppose a clip of autonomous driving data $\mathcal{D}_{AD}$ is composed of T time-sequenced sensor and pose data $\{(S_t, P_t)\}_{t=1}^{T}$, where each time-step t has onboard sensor data S_t (e.g., multi-view images I_t with camera intrinsics and extrinsics) and the ego-vehicle's pose P_t.

An offline geographic database $\mathcal{D}_{geo}$ is introduced, composed of geographic images I_{geo} and their corresponding metadata (global coordinates and camera parameters) P_{geo}.

We evaluate the effectiveness of the spatial retrieval paradigm across five AD tasks (Table 1). For on-vehicle tasks (3D object detection, online mapping, occupancy prediction, and motion planning), a retrieval function $\mathcal{R}$ is applied at each time-step t, where the function takes current images I_t and ego-pose P_t as inputs and retrieves from $\mathcal{D}_{geo}$ to fetch the most relevant geographic data:

$$\text{RetrievedGeoData}_t = \mathcal{R}(I_t, I_{geo}, P_t, P_{geo}) \quad (1)$$

In this work, for simplicity, we retrieve the nearest geographic images for each camera at each time-step. If the 3D distance is larger than a threshold, the API returns NONE. There could be more advanced ways of retrieval, for example, more retrieved neighborhood images as global context, which we leave for future works to explore.

For the offline task (generative world model), multiple geographic images are retrieved along the desired driving trajectories of the generation, which provides a spatial scaffold to guide long-horizon, globally consistent scene generation with less hallucination.

Table 1. **Evaluated Tasks for Spatial Retrieval Paradigm.**

Task	Usage	Output
Detection	On-Vehicle	3D Bounding Boxes
Mapping	On-Vehicle	Road Lines
Occupancy	On-Vehicle	3D Occupancy Grid
Planning	On-Vehicle	Ego Trajectory
World Model	Offline Server	Future Video Frames

3.2. Spatial Retrieval Adapter

In this section, we introduce a general plug-and-play module, as in Fig. 2 (Left), to incorporate the retrieved geographic data for BEV-based on-vehicle tasks, serving as an intuitive baseline. More advanced modules combining priors of each individual task are left for future works.

Geographic Image and Positional Encoding. The retrieved geographic images are encoded by the same backbone used for onboard cameras to obtain $\mathbf{F}_{\text{geo}}$. To encode the relative spatial relationship between retrieved geographic images and current ego position, we adopt PETR [38] to encode their geographic image patches' 3D positional encoding to obtain $\mathbf{F}_{\text{geo}}^{\text{pos}}$.

Geo-Cross-Attention. Geographic features are integrated into BEV representations via cross-attention with positional encoding, modulated by the reliability score w to handle miss or wrong retrieval (detailed in Section 3.4):

$$\mathbf{F}'_{\text{BEV}} = \mathbf{F}_{\text{BEV}} + w \cdot \text{CrossAttn}(\mathbf{F}_{\text{BEV}}, \mathbf{F}_{\text{geo}} + \mathbf{F}_{\text{geo}}^{\text{pos}}, \mathbf{F}_{\text{geo}}), \quad (2)$$

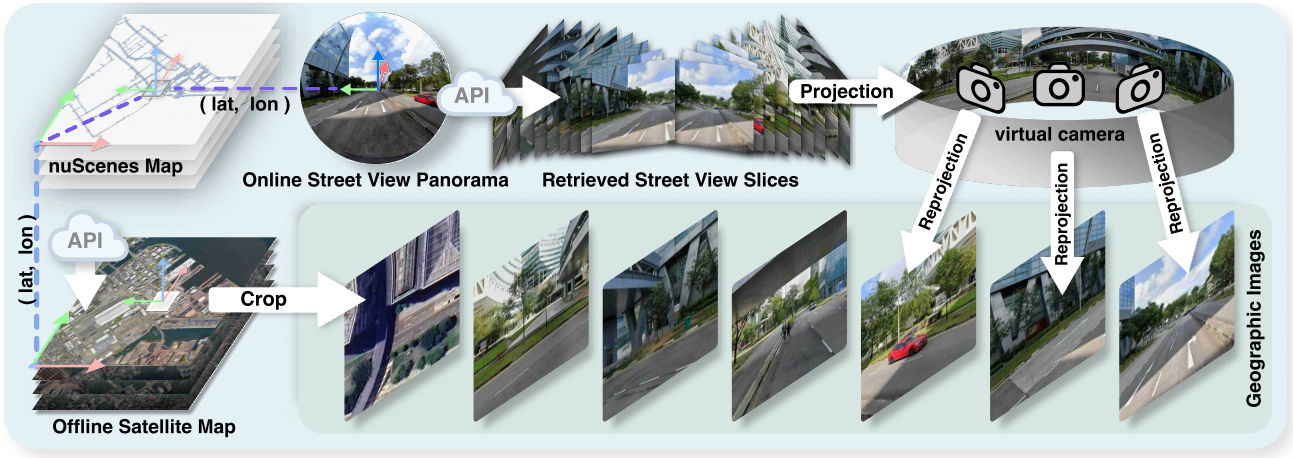

Figure 4. **Geographic Data Curation from Google Maps**. We use the GPS coordinates from nuScenes ego poses to query Google Map. Each unique panorama is downloaded once, decomposed into 18 yaw-sampled views, and projected onto an equirectangular panorama representation. For each camera at each frame in nuScenes, a virtual camera with matched intrinsics/extrinsics reprojects an aligned street view image from equirectangular panorama, effectively reducing redundant downloads and storage.

The augmented BEV features are then fed to original downstream task heads. This plug-and-play design keeps all training objectives and network architectures unchanged.

3.3. Spatial Retrieval For Generative World Model

World models [48, 50, 56, 60] for autonomous driving could serve as data generator, closed-loop evaluator, or RL environment, where they usually run on clusters and servers instead of on-vehicle. As a result, these models have access to future ego trajectory, which allows to prefetch geographic images along upcoming path, similar to Bench2Drive-R [60]. By injecting geographic images at future positions throughout generation, they provide persistent spatial cues that maintain scene consistency.

Geography Extended DiT. Similar to Bench2Drive-R [60], to incorporate geographic data into generation, we inject an extra geographic cross-attention layer into the widely used DiT [44] block after original attention layers:

$$\mathbf{F}' = \mathbf{F} + w \cdot \text{CrossAttn}(\mathbf{F}, \mathbf{F}_{\text{geo}} + \mathbf{F}^{\text{pos}}_{\text{geo}}, \mathbf{F}_{\text{geo}}), \quad (3)$$

where $\mathbf{F}$ denotes noised latent feature and $\mathbf{F}_{\text{geo}}$ denotes retrieved geographic features for the start and end frames of the generated segment. This design allows the model to have geographic context at corresponding future locations.

3.4. Adaptive Fusion with Reliability Estimation

A key challenge in leveraging geographic data is handling missing or misaligned street view images, as in Fig. 10. To reduce the influence of these cases and let the model be robust to unreliable retrieval, we introduce an adaptive fusion mechanism, as in Fig. 3, that dynamically adjusts the contribution of geographic features based on (i) the distance between retrieved location and ego pose; (ii) the similarity of retrieved and current on-board images.

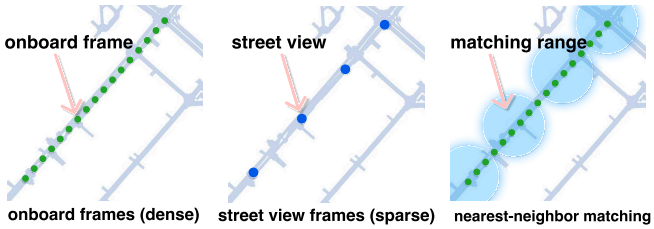

Figure 5. **Correspondence Relation between Frames and Geographic Data.** nuScenes frames exhibit a higher acquisition frequency than street view data along the same road. Each frame is matched to its geographically nearest street view data, where multiple frames may correspond to the same street view data.

Specifically, we set a Reliability Estimation gate module to output reliability score $w \in [0, 1]$:

$$w = \sigma(\text{MLP}([\text{ZNCC}(\mathbf{F}_{\text{onboard}}, \mathbf{F}_{\text{geo}}), d_{\text{GPS}}])), \quad (4)$$

where ZNCC computes the zero-normalized cross-correlation [27] between onboard and geographic features, d_{GPS} is the distance between the street view location and ego position, and σ is the sigmoid function. During training, we supervise w with binary labels (0 for invalid/missing, 1 for valid). At test time, this learned estimator can down-weight unreliable geographic features.

4. nuScenes-Geography: Extended Geographic Data from Google Maps

To systematically investigate the effectiveness of the spatial retrieval paradigm, we introduce **nuScenes-Geography**, extending the widely used nuScenes [3] dataset with geographic data. We collect data using the Google Maps APIs, which provide access to perspective views and satellite images given coordinates and camera parameters, as in Fig. 4.

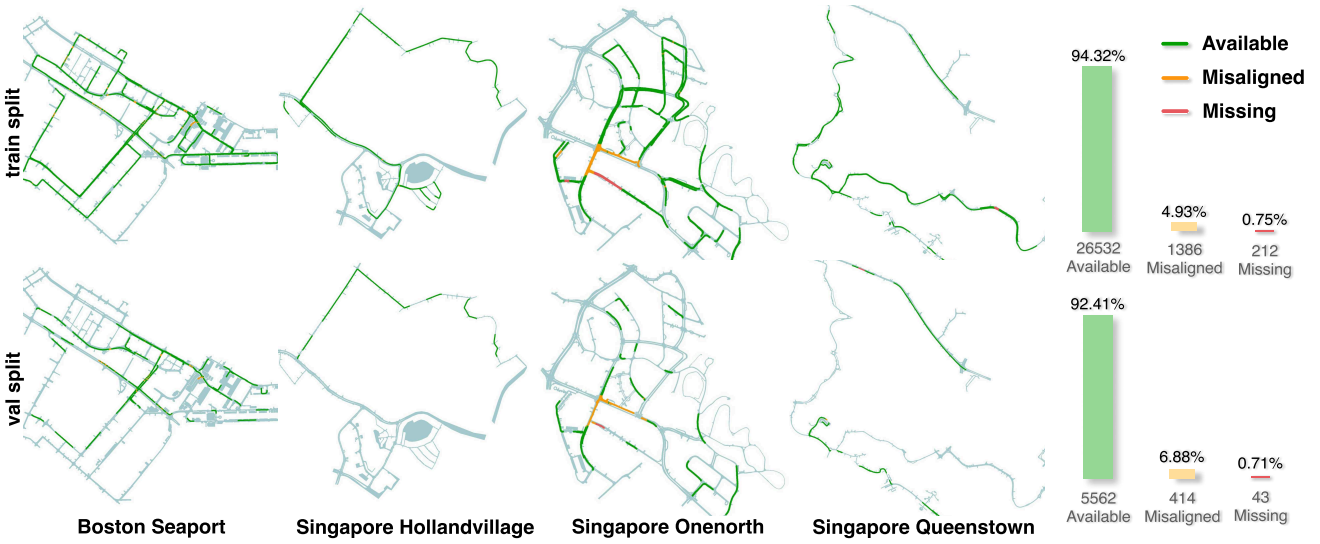

Figure 6. **Coverage of Geographic Data in nuScenes**. Green denotes frames with reliable geographic data; Orange denotes samples where geographic data were retrieved but deemed unreliable and excluded from dataset; Red denotes samples lacking any geographic data.

Coordinate Computation. To associate geographic data with nuScenes frames, we compute the longitude and latitude coordinates of each frame by combining the global origin of the nuScenes map with the ego-vehicle pose. Using these coordinates, we query the Google Maps APIs to obtain street view images and satellite map tiles.

Equirectangular Panorama Representation for Efficient Storage and Retrieval. Since the spatial sampling frequency of street view images is significantly lower than the key frame rate of nuScenes, multiple frames may correspond to the same street view location, as in Fig. 5. To minimize storage overhead, each unique geographic data is retrieved only once and we store a mapping between each geographic data and all its nearest nuScenes frames.

However, different frames in nuScenes require different perspectives for the same street view location. To ensure geometrically correct views while maintaining storage efficiency, we adopt equirectangular panoramic representation:

1. **Data Fetching**: For each street view location, we retrieve 18 perspective images from API with distributed yaw angles (covering 360°) at a fixed pitch of 0°.
2. **Equirectangular Panorama Format**: These images are projected onto a spherical representation and stored in an equirectangular panoramic format.
3. **Virtual Camera Alignment**: For each nuScenes frame and each onboard camera, a virtual camera is instantiated at the corresponding street view location, with intrinsic parameters identical to the nuScenes camera model. The extrinsic transformation is derived from the ego pose and the street view capture point: the rotation follows the original nuScenes camera orientation, while the translation is computed from the latitude–longitude offset between the street view and the ego vehicle. The translation along the z-axis is set to a fixed constant.
4. **Reprojection Retrieval**: Using the virtual camera configuration, we perform perspective projection from the equirectangular panorama to synthesize a street view image, geometrically aligned with the nuScenes frame.

This process ensures spatial consistency and one-to-one correspondence between each onboard frame and its synthesized street view perspectives, while keeping the overall collection pipeline storage-efficient, reducing overall storage by more than 70% compared to directly downloading per-frame street view crops.

Dealing with Missing and Misaligned Geographic Data. As shown in Fig. 10, the Google Maps APIs might return an empty response or misaligned geographic data. As discussed in Sec. 3.4, we design an adaptive fusion mechanism to let the model be able to selectively fuse reliable geographic information by residual gating. During the curation of nuScenes-Geography, we manually check all the downloaded geographic images and identify 1800 misaligned cases, serving as the negative labels of the Reliability Estimation module. Fig. 6 gives an overview of the coverage of geographic data for nuScenes, which is relatively high.

5. Experiments

In this section, we evaluate the proposed spatial retrieval paradigm on the extended nuScenes-Geography dataset across five tasks. We study three potential advantages of adopting spatial retrieval: enhancing static scene understanding, improving planning robustness, and enhancing spatial consistency of generative world model.

Table 2. **Online Mapping Results.** All methods use the ResNet50 backbone and are reproduced. Adding geographic priors brings significant mAP improvements to baselines. The FPS is measured on an NVIDIA 4090 GPU.

Method	Epoch	AP_{ped}(↑)	AP_{div}(↑)	AP_{bound}(↑)	mAP(↑)
MapTR	24	46.3	51.5	53.1	50.3
MapTR+Geo	24	56.5 (+10.2)	67.3 (+15.8)	59.8 (+6.7)	61.2 (+10.9)
MapTR	110	55.9	60.9	61.1	59.3
MapTR+Geo	110	72.5 (+16.6)	75.0 (+14.1)	70.5 (+9.4)	72.7 (+13.4)
MapTRv2	24	59.8	62.4	62.4	61.5
MapTRv2+Geo	24	74.4 (+14.6)	74.7 (+12.3)	71.1 (+8.7)	73.4 (+11.9)
MapTRv2	110	68.1	68.3	69.7	68.7
MapTRv2+Geo	110	79.8 (+11.7)	77.4 (+9.1)	77.3 (+7.6)	78.2 (+9.5)

Table 3. **Occupancy Results on Occ3D-nuScenes.** All methods are reproduced. Geographic priors improve mIoU especially for static terrain, considering it introduces extra information about background.

Method	Overall mIoU (↑)	Per-Class mIoU (↑) (Static Terrain)					
		driveable	other flat	sidewalk	terrain	manmade	vegetation
FBOcc	39.11	80.07	42.76	51.18	55.13	42.19	37.53
FBOcc+Geo	39.74 (+0.63)	82.47 (+2.4)	44.47 (+1.71)	53.26 (+2.08)	57.7 (+2.57)	42.61 (+0.42)	38.57 (+1.04)
FlashOCC:M1	31.92	77.95	36.61	47.17	50.92	36.39	31.08
FlashOCC:M1+Geo	32.35 (+0.43)	78.72 (+0.77)	39.51 (+2.9)	47.88 (+0.71)	51.82 (+0.9)	37.28 (+0.89)	31.94 (+0.86)

Table 4. **Object Detection Results.** The influence is marginal since geographic data mainly helps background.

Method	Backbone	NDS (↑)	mAP (↑)
BEVDet	R50	39.41	30.85
BEVDet+Geo	R50	39.43(+0.02)	30.69(-0.16)
BEVFormer	R101-DCN	51.70	41.60
BEVFormer+Geo	R101-DCN	51.80(+0.10)	41.64(+0.04)

Table 5. **Generative World Model Results.** The empirical results validate the effectiveness of geographic priors.

Method	Frame Size	FVD (↓)	FID (↓)
UVG	35×256×448	36.10	5.82
UVG+Geo	35×256×448	29.97(+6.13)	5.60(+0.22)
MDD	17×224×400	84.43	18.38
MDD+Geo	17×224×400	81.52(+2.91)	18.10(+0.28)

5.1. Scene Understanding

Spatial retrieval provides a stable view of the background, compensating for the vulnerability to extreme visual conditions and limited perception horizon of onboard sensors.

Online Mapping As in Table 2, integrating geographic priors into MapTR [34] and MapTRv2 [35] substantially improves online mapping. The extra background information enables recovery of occluded lanes, as in Fig. 9.

Occupancy Prediction As in Table 3, Extending FB-OCC [32] and FlashOCC [62] shows consistent mIoU gains, especially on static categories. The prior anchors background geometry against sensor noise, as in Fig. 9.

Object Detection As in Table 4, BEVDet [19] and BEVFormer [31] show a negligible improvement with geographic data, which is natural since spatial retrieval mainly brings background information. However, adopting geographic data to distinguish foreground and background and then help object detection is an interesting future direction.

5.2. Planning Robustness

We assess how spatial retrieval facilitates safer planning with VAD [25]. The geographic prior provides consistent road layout information, compensating for sensing instability caused by occlusions or poor lighting. As shown in Table 6, while maintaining comparable trajectory accuracy, our method improves safety margins. Specifically, in challenging night scenes, it reduces the average collision rate from 0.55% to 0.48%, demonstrating the value of geographic priors as a robust guide for safe planning. Fig. 7 visualizes examples.

5.3. Generative World Model Consistency

We further evaluate whether geographic priors help generative world models. Conditioning UniMLVG [8] and MagicDriveDit [14] (For MagicDriveDit, we adjust the test-set sampling stride to 13 to avoid repeatedly sampling near-duplicate segments.) on geographic images yields lower FVD and FID, preventing scene drift and maintaining geometric consistency during rollouts, as in Table 5. This confirms that spatial retrieval acts as a structural scaffold for coherent world modeling.

Table 6. **End-to-end Planning Results.** All methods are reproduced. Introducing geographic information achieves comparable planning accuracy in terms of L2 error. On the challenging night subset, VAD+Geo has lower collision rates, indicating improved safety performance.

Method	Split	L2 (m) ↓				Collision (%) ↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
VAD	Full	**0.34**	**0.61**	0.98	**0.64**	0.16	**0.29**	0.66	0.37
VAD+Geo	Full	0.35(-0.01)	0.62(-0.01)	**0.96**(+0.02)	**0.64**(0.00)	**0.14**(+0.02)	**0.29**(0.00)	**0.64**(+0.02)	**0.36**(+0.01)
VAD	Night	**0.44**	**0.82**	1.35	**0.87**	0.10	0.24	**1.30**	0.55
VAD+Geo	Night	0.46(-0.02)	0.83(-0.01)	**1.33**(+0.02)	**0.87**(0.00)	**0.00**(+0.10)	**0.15**(+0.09)	**1.30**(0.00)	**0.48**(+0.07)

Table 7. **Ablation Study** with Occupancy Task (Static mIoU for non-movable object categories) and Generative World Model Task (FVD).

(a) **Effect of Geo Images**

Variant	S mIoU ↑	FVD ↓
w/o Geo Images	46.66	35.42
w Geo Images	47.86	29.97

(b) **Effect of Positional Encoding**

Variant	S mIoU ↑	FVD ↓
w/o 3DPE	46.22	32.82
w 3DPE	47.86	29.97

(c) **Effect of Reliability Estimation Gate(REG)**

Variant	S mIoU ↑	FVD ↓
w/o REG	47.65	30.95
w REG	47.86	29.97

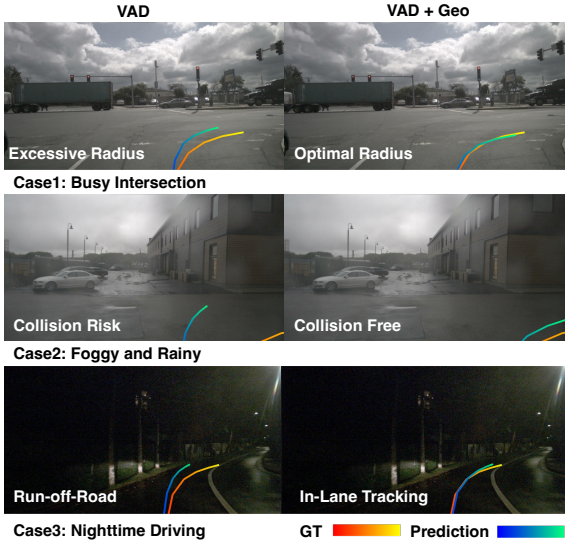

Figure 7. **End-to-End Planning Case Study.** Geographic priors lead to better trajectories in complex intersections, night conditions, and adverse weather.

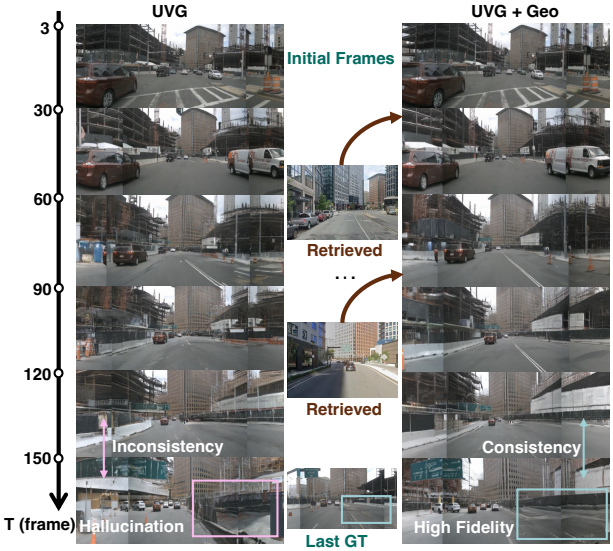

Figure 8. **Generative World Model Case Study.** Geographic data provides spatial clues, enhancing temporal consistency and mitigating hallucinations.

5.4. Visualization of Misaligned Spatial Retrieval

The new paradigm faces the challenge of missing or misaligned retrieval issues, which occur when the offline geographic images are inconsistent with the camera images, as in Fig. 10. This can occasionally happen due to: **(1) Outdated Maps:** Road layouts change due to construction, but the cached map imagery does not accurately reflect this, potentially confusing the model. **(2) GPS/Localization Error:** Inaccurate ego-pose can lead to misalignment between retrieved images and onboard sensor images, which sometimes happen with Google Maps APIs.

5.5. Ablation Study

We conduct ablation studies with occupancy task (FlashOcc [62]) and generative world model task (Unimlvg [8]). As in Table 7, we observe that introducing geographic priors always brings explicit performance gains while positional encoding and reliability estimation gate bring further gains.

5.6. Robustness to Inaccurate Retrieval

To further evaluate the effectiveness of the proposed Reliability Estimation Gate, we assess online mapping method - MapTRv2's [35] robustness under inaccurate retrieval.

We randomly drop the geographic images or replace them with random incorrect images for a certain percentage of frames. Figure 11 shows that the model's performance degrades gracefully as the availability of the prior decreases. Even with 50% of the priors missing or misaligned, the model retains a significant portion of its performance gain over the baseline. This indicates that the proposed Reliability Estimation Gate allows the model to leverage the prior when available but does not lead to catastrophic failure in

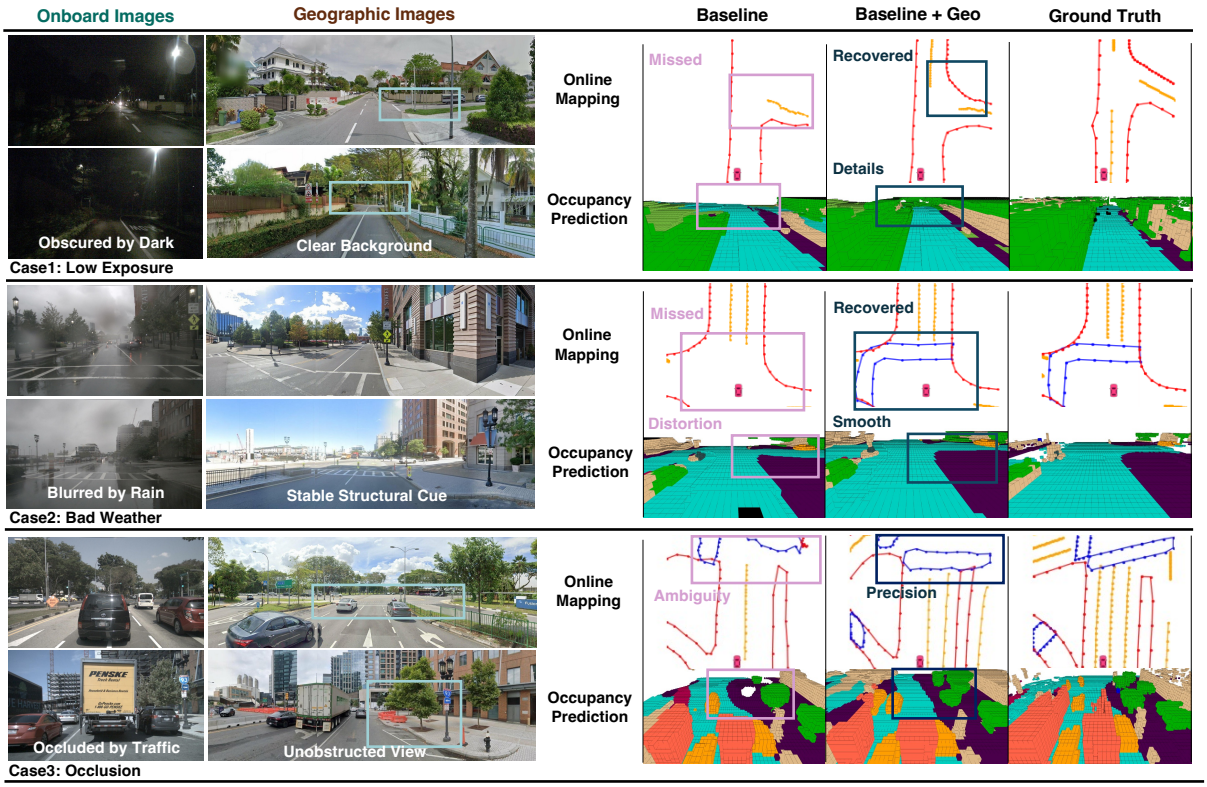

Figure 9. **Geographic Priors Enhance Scene Understanding in Challenging Scenarios.** Our method leverages geographic priors to correct errors in online mapping (top) and occupancy prediction (bottom) under challenging conditions.

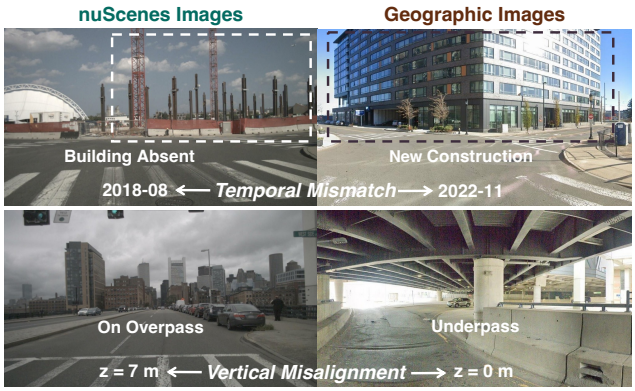

Figure 10. **Visualization of Misaligned Spatial Retrieval.** Examples include a **temporal mismatch** (top) due to scene changes and a **vertical misalignment** (bottom) from altitude errors. Such discrepancies can lead to incorrect maps or occupancy predictions.

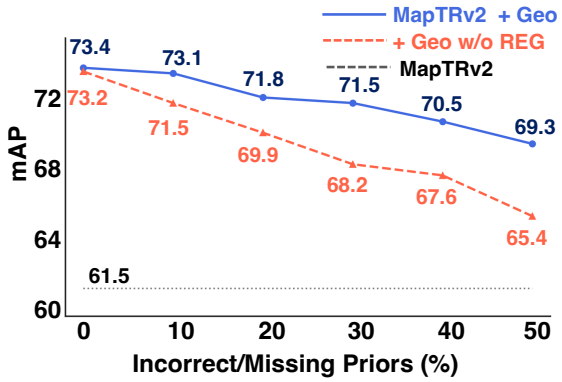

Figure 11. **Robustness to Inaccurate Retrieval.** Performance of MapTRv2+Geo as a function of the percentage of frames where the geographic images are dropped or replaced with wrong images. Performance degrades gracefully, demonstrating the model is not overly reliant on the prior. Notably, without the Reliability Estimation Gate (REG), the performance decline is steeper, indicating reduced robustness to inaccurate priors.

its absence, demonstrating robust real-world applicability.

6. Conclusion

In this work, we present the spatial retrieval paradigm for AD, introducing geographic data as an additional input. We extend nuScenes with geographic data by Google Maps APIs and evaluate five key AD tasks on the extended nuScenes-Geography dataset. We propose a general plug-and-play Spatial Retrieval Adapter module as an intuitive baseline to incorporate geographic data. We propose Reliability Estimation to adaptively fuse geographic information based on the reliability of the retrieved data. Extensive experiments show that the proposed paradigm can enhance the performance of multiple AD tasks, demonstrating the substantial potential of the new paradigm.

Acknowledgements

This work was supported by the Science and Technology Commission of Shanghai Municipality (No. 24511103100) and the New Cornerstone Science Foundation through the XPLORER PRIZE. This work was also in part supported by Scientific Research Innovation Capability Support Project for Young Faculty (U40) of the Ministry of Education of China (SRICSPYF-ZY2025019).

References

[1] Carla simulator. https://github.com/carla-simulator/carla.

[2] Dan Barnes, Will Maddern, Geoffrey Pascoe, and Ingmar Posner. Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1894–1900. IEEE, 2018.

[3] 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.

[4] Hongxiang Cai, Zeyuan Zhang, Zhenyu Zhou, Ziyin Li, Wenbo Ding, and Jiuhua Zhao. Bevfusion4d: Learning lidar-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation. *arXiv preprint arXiv:2303.17099*, 2023.

[5] Tianhui Cai, Yifan Liu, Zewei Zhou, Haoxuan Ma, Seth Z Zhao, Zhiwen Wu, and Jiaqi Ma. Driving with regulation: Interpretable decision-making for autonomous vehicles with retrieval-augmented reasoning via llm. *arXiv preprint arXiv:2410.04759*, 2024.

[6] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14403–14412, 2021.

[7] Chen Chen, Zhirui Wang, Taowei Sheng, Yi Jiang, Yundu Li, Peirui Cheng, Luning Zhang, Kaiqiang Chen, Yanfeng Hu, Xue Yang, et al. Sa-occ: Satellite-assisted 3d occupancy prediction in real world. *arXiv preprint arXiv:2503.16399*, 2025.

[8] Rui Chen, Zehuan Wu, Yichen Liu, Yuxin Guo, Jingcheng Ni, Haifeng Xia, and Siyu Xia. Unimlvg: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving. *arXiv preprint arXiv:2412.04842*, 2024.

[9] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022.

[10] Gamal Elghazaly, Raphaël Frank, Scott Harvey, and Stefan Safko. High-definition maps: Comprehensive survey, challenges, and future perspectives. *IEEE Open Journal of Intelligent Transportation Systems*, 4:527–550, 2023.

[11] Sudeep Fadadu, Shreyash Pandey, Darshan Hegde, Yi Shi, Fang-Chieh Chou, Nemanja Djuric, and Carlos Vallespi-Gonzalez. Multi-view fusion of sensor data for improved perception and prediction in autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2349–2357, 2022.

[12] 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.

[13] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023.

[14] Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrivedit: High-resolution long video generation for autonomous driving with adaptive control. *arXiv e-prints*, pages arXiv–2411, 2024.

[15] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024.

[16] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[17] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[18] Yuekun Hu, Yingfan Liu, and Bin Hui. Combining openstreetmap with satellite imagery to enhance cross-view geolocalization. *Sensors*, 25(1):44, 2024.

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

[20] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023.

[21] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7953–7963, 2023.

[22] Xiaosong Jia, Penghao Wu, Li Chen, Yu Liu, Hongyang Li, and Junchi Yan. Hdgt: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encod-

ing. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):13860–13875, 2023.

[23] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *NeurIPS 2024 Datasets and Benchmarks Track*, 2024.

[24] Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. In *International Conference on Learning Representations (ICLR)*, 2025.

[25] 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. *ICCV*, 2023.

[26] Wei Jiang, Lu Wang, Tianyuan Zhang, Yuwei Chen, Jian Dong, Wei Bao, Zichao Zhang, and Qiang Fu. Robuste2e: Exploring the robustness of end-to-end autonomous driving. *Electronics*, 13(16):3299, 2024.

[27] John P Lewis. Fast normalized cross-correlation. In *Vision interface*, pages 120–123, 1995.

[28] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Jia Zeng, Zhiqi Li, Jiazhi Yang, Hanming Deng, Hao Tian, Enze Xie, Jiangwei Xie, Li Chen, Tianyu Li, Yang Li, Yulu Gao, Xiaosong Jia, Si Liu, Jianping Shi, Dahua Lin, and Yu Qiao. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2151–2170, 2023.

[29] Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2drive: Efficient reinforcement learning by thinking with latent world model for autonomous driving (in carla-v2). In *European Conference on Computer Vision*, pages 142–158. Springer, 2025.

[30] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: layout-guided multi-view driving scenarios video generation with latent diffusion model. In *European Conference on Computer Vision*, pages 469–485. Springer, 2024.

[31] 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. *arXiv preprint arXiv:2203.17270*, 2022.

[32] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023.

[33] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022.

[34] 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. *arXiv preprint arXiv:2208.14437*, 2022.

[35] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *International Journal of Computer Vision*, 133(3):1352–1374, 2025.

[36] 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.

[37] Richard J Lisle. Google earth: a new geological resource. *Geology today*, 22(1):29–32, 2006.

[38] 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.

[39] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *International Conference on Machine Learning*, pages 22352–22369. PMLR, 2023.

[40] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[41] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *European Conference on Computer Vision*, pages 329–345. Springer, 2024.

[42] Steven J Luck and Edward K Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390 (6657):279–281, 1997.

[43] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022.

[44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[45] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7077–7087, 2021.

[46] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8795–8801. IEEE, 2025.

[47] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36:64318–64330, 2023.

[48] Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Freevs: Generative view synthesis on free driving trajectory. *arXiv preprint arXiv:2410.18079*, 2024.

[49] 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.

[50] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.

[51] Yujin Wang, Quanfeng Liu, Jiaqi Fan, Jinlong Hong, Hongqing Chu, Mengjian Tian, Bingzhao Gao, and Hong Chen. Rac3: Retrieval-augmented corner case comprehension for autonomous driving with vision-language models. *arXiv preprint arXiv:2412.11050*, 2024.

[52] Yujin Wang, Quanfeng Liu, Zhengxin Jiang, Tianyi Wang, Junfeng Jiao, Hongqing Chu, Bingzhao Gao, and Hong Chen. Rad: Retrieval-augmented decision-making of meta-actions with vision-language models in autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3838–3848, 2025.

[53] Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15449–15458, 2024.

[54] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. *Advances in Neural Information Processing Systems*, 35:6119–6132, 2022.

[55] Xuan Xiong, Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Neural map prior for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17535–17544, 2023.

[56] Jiazhi Yang, Kashyap Chitta, Shenyuan Gao, Long Chen, Yuqian Shao, Xiaosong Jia, Hongyang Li, Andreas Geiger, Xiangyu Yue, and Li Chen. Resim: Reliable world simulation for autonomous driving. *Advances in Neural Information Processing Systems*, 2025.

[57] Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. Learning substructure invariance for out-of-distribution molecular representations. *Advances in Neural Information Processing Systems*, 35:12964–12978, 2022.

[58] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving, 2023.

[59] Zhenjie Yang, Yilin Chai, Xiaosong Jia, Qifeng Li, Yuqian Shao, Xuekai Zhu, Haisheng Su, and Junchi Yan. Drivemoe: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. *arXiv preprint arXiv:2505.16278*, 2025.

[60] Junqi You, Xiaosong Jia, Zhiyuan Zhang, Yutao Zhu, and Junchi Yan. Bench2drive-r: Turning real world data into reactive closed-loop autonomous driving benchmark by generative model. *arXiv preprint arXiv:2412.09647*, 2024.

[61] Yurong You, Cheng Perng Phoo, Carlos Andres Diaz-Ruiz, Katie Z Luo, Wei-Lun Chao, Mark Campbell, Bharath Hariharan, and Kilian Q Weinberger. Better monocular 3d detectors with lidar from the past. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6634–6641. IEEE, 2024.

[62] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023.

[63] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7356–7365, 2024.

[64] Gongjie Zhang, Jiahao Lin, Shuang Wu, Zhipeng Luo, Yang Xue, Shijian Lu, Zuoguan Wang, et al. Online map vectorization for autonomous driving: A rasterization perspective. *Advances in Neural Information Processing Systems*, 36:31865–31877, 2023.

[65] Peiyuan Zhang, Junwei Luo, Xue Yang, Yi Yu, Qingyun Li, Yue Zhou, Xiaosong Jia, Xudong Lu, Jingdong Chen, Xiang Li, et al. Pointobb-v3: Expanding performance boundaries of single point-supervised oriented object detection. *International Journal of Computer Vision*, 2025.

[66] Yuxiao Zhang, Alexander Carballo, Hanting Yang, and Kazuya Takeda. Perception and sensing for autonomous vehicles under adverse weather conditions: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:146–177, 2023.

[67] Yutao Zhu, Xiaosong Jia, Xinyu Yang, and Junchi Yan. Flatfusion: Delving into details of sparse transformer-based camera-lidar fusion for autonomous driving. *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.