Scratch → FCOS3D → ViDAR: Pretraining Effects on BEVFormer

Yicheng Zou[1], Zhifeng Zheng[2], Lucy Tan[2]

***Abstract*— Pretraining plays a critical role in 3D object detection because 3D data are difficult to acquire and time consuming to label [1]. To evaluate how initialization influences the performance of BEVFormer on the nuScenes mini dataset, we examine three pretraining strategies while keeping the training schedule fixed at 22 epochs: training from scratch, FCOS3D pretraining and ViDAR pretraining. The results clearly highlight the importance of task-specific pretraining.Training from scratch performs poorly (mAP: 0.03%), showing that BEVFormer needs strong initialization to learn meaningful 3D features. The FCOS3D pretraining significantly increases performance (mAP: 19.22%) and ViDAR achieves the best results (mAP: 20.45%) due to its stronger geometric supervision. We make available our open-source implementation at https://github.com/chahyon-ku/gsplatam.**

I. Introduction

Recently, visual autonomous driving has advanced rapidly, with monocular or multi-view camera inputs [2]. Existing models have demonstrated strong capability in building BEV representations [3] and enabling high-quality perception [4], prediction [5], and planning [6]. However, despite these remarkable improvements, such models still depend heavily on accurate 3D supervision—such as occupancy labels [7], 3D bounding boxes [8], and motion trajectories [9]—which which demand significant labeling effort in both time and cost. To mitigate the reliance on costly annotations, pretraining [10] has become an essential strategy for scaling up downstream tasks. By designing pretext objectives that exploit large, easily accessible unlabeled data, pretraining enables models to learn useful representations and improve downstream performance even when labeled data is limited.

This project investigates how three different initialization schemes—**training from scratch**, **FCOS3D pretraining**, and **ViDAR pretraining**—affect BEVFormer's performance on the nuScenes-mini dataset. BEVFormer is inherently geometry-dependent, relying on multi-view lifting and spatiotemporal attention to form BEV features. Therefore, pretraining methods that include depth cues, 3D geometry, or motion forecasting can significantly enhance their downstream detection accuracy.

Our experiments show that training BEVFormer from scratch results in almost no meaningful performance, indicating that strong priors are necessary to learn 3D structure. Initializing with FCOS3D offers clear gains by introducing monocular 3D cues, while ViDAR achieves the best results thanks to its self-supervised modeling of semantics, geometry, and temporal dynamics. Overall, our study provides a systematic comparison of pretraining strategies and highlights the importance of geometry-aware initialization for BEV-based perception.

*This work was not supported by any organization

[1]Yicheng Zou is with the Department of Electrical and Computer Engineering, University of Michigan yichzou@umich.edu

[2]Lucy Tan, and Zhifeng Zheng are with the Department of Mechanical Engineering, University of Michigan zhifeng@umich.edu lucytan@umich.edu

II. Related Works

A. Depth-Aware and Geometry-Aware Pretraining

Depth prediction is a powerful form of geometric supervision for autonomous driving. FCOS3D [11]improves monocular 3D estimation by projecting 3D centers into 2D space and learning depth, orientation, and dimensions jointly. It provides strong geometric cues that transfer well to BEV perception and is commonly used as the default initialization for BEVFormer.

Other depth-related pretraining methods similarly demonstrate the benefits of geometric supervision. DD3D [12], [13] shows that large-scale, self-supervised depth pretraining dramatically improves downstream 3D detection by teaching the backbone consistent single-image geometry. BEVDepth [14]further emphasizes depth reliability by combining camera intrinsics with depth-aware feature lifting, improving BEV detectors through structured geometric constraints.

Together, these works highlight a central theme: **adding explicit or implicit depth supervision during pretraining consistently enhances downstream 3D understanding**. Our evaluation of FCOS3D pretraining follows this direction by testing how monocular geometric supervision affects BEVFormer's 3D detection.

B. Self-Supervised and Cross-Modal 3D Representation Learning

Self-supervised 3D learning techniques aim to reduce the need for annotated 3D data. PointContrast [15]uses contrastive learning on point clouds to learn geometry-aware features transferable to downstream tasks. STRL [16] improves on this by incorporating cross-view occlusion reasoning and volumetric consistency.

In addition, several works examine cross-modal pretraining where vision models are trained to predict LiDAR-like signals. "See the Point" [15] and similar methods demonstrate that supervising image encoders using point-cloud reconstruction improves 3D scene understanding. These approaches share conceptual similarities with ViDAR, which also uses cross-modal forecasting to infuse geometric priors into the visual encoder.

Our study is aligned with this line of work: it investigates whether **geometry- and motion-rich self-supervised pre-**

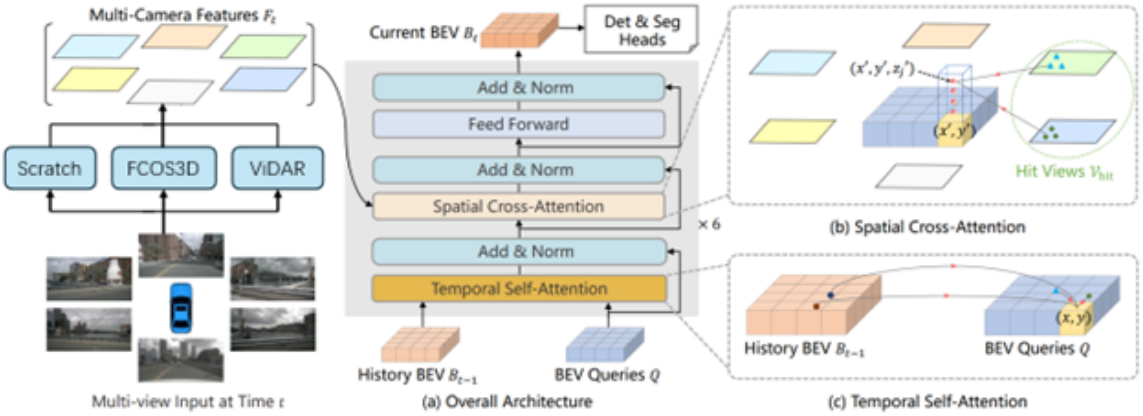

Fig. 1: Framework Overview of BEVFormer Under Three Initialization Settings

training can better support downstream BEV detection than traditional monocular 3D pretraining.

III. Methodology

A. Overview

This project investigates how different pretraining strategies affect the performance of BEVFormer on the nuScenes mini dataset. BEVFormer is a multi-camera transformer-based 3D perception framework that converts perspective-view images into a unified bird's-eye-view (BEV) representation using spatial and temporal attention [14]. Since the model operates entirely in 3D space, pretraining methods that contain 3D information are expected to provide a more suitable initialization than generic 2D feature learning. Therefore, we focus on two pretraining approaches that incorporate 3D geometry and depth-related signals [17] and compare three initialization strategies: (1) training from scratch with random initialization, (2) FCOS3D pretraining (the default pretraining model used by BEVFormer), and (3) ViDAR pretraining. This setup allows us to study how different initializations affect BEVFormer's ability to learn effective BEV representations and perform 3D object detection.

B. Initilization Strategies

Training from Scratch (Random Initialization) . initializes all network weights randomly following standard initialization schemes. This provides a reference point to measure the benefits of any pretraining strategy. Without pretrained features, the model must learn all visual representations from the limited nuScenes training data.

FCOS3D Pretraining [11]**.** is a fully convolutional single-stage monocular 3D object detector that extends the 2D FCOS framework to predict full 3D bounding box attributes from image features. The method reformulates the 7-DoF 3D box parameters into image-plane representations by projecting the 3D center and decoupling 2D and 3D attributes. FCOS3D assigns objects to feature levels based on their projected scale and trains the network to predict depth, orientation, dimensions, and the projected 3D center using a Gaussian-based centerness formulation.

FCOS3D provides strong geometric supervision that aligns well with the requirements of BEVFormer, whose BEV generation pipeline relies on the backbone's ability to encode depth cues and geometric structure when lifting perspective-view features into a bird's-eye-view space. A backbone pretrained with FCOS3D already contains many of these properties: it learns depth-sensitive features and perspective cues from monocular supervision, develops an understanding of how 2D image patterns relate to 3D positions, and acquires representations of object size, scale, and orientation. It also builds spatial structure awareness that supports 3D reasoning. With these forms of geometric knowledge already embedded in the backbone, BEVFormer has a much stronger starting point for BEV feature construction.

ViDAR pretraining [18]**.** ViDAR is a self-supervised pretraining framework designed for end-to-end autonomous driving systems. By forcing the model to predict the future from past frames, ViDAR naturally learns scene flow and object motion, which are crucial for temporal modeling and future estimation. At the same time, it involves the reconstruction of point clouds from images, which supervises the multi-view geometry and semantic modeling. So ViDAR captures the three essential aspects for autonomous driving pre-training: semantics,3D geometry, and temporal dynamics. These learned representations can then help improve downstream tasks.

As shown in Fig.2 , ViDAR contains three parts: history encoder, latent rendering and future decoder.

History Encoder in ViDAR can be instantiated with any BEV visual encoder, including BEVFormer. This makes ViDAR naturally compatible as a pretraining pipeline for BEVFormer. By encoding multi-frame images, this module learns to generate temporally consistent BEV features, which provides high-quality initialization parameters for downstream BEV-based perception tasks.

The Latent Rendering module maps BEV features into a 3D geometric latent space, pushing the model to learn multi-view geometry, depth relationships, and coherent spatial structure. By requiring features from different camera views to align in 3D, this stage encourages geometry-aware

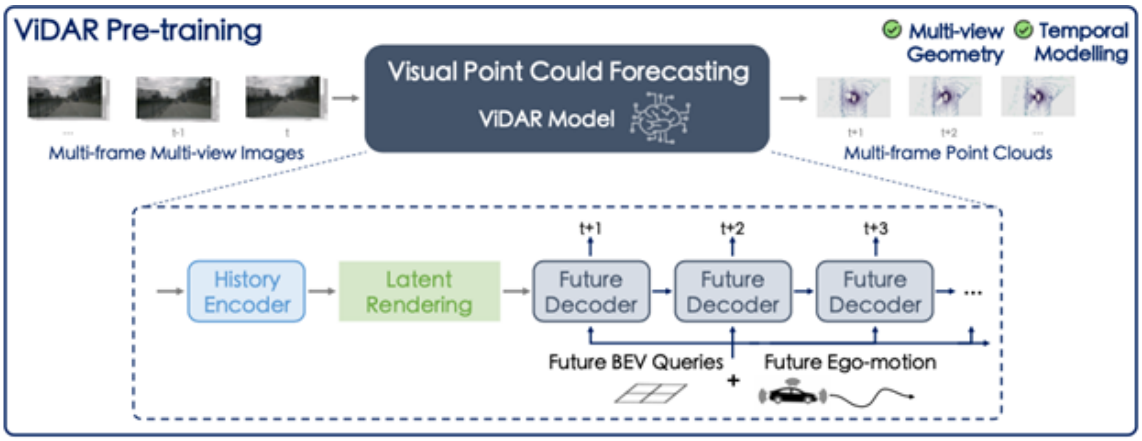

Fig. 2: ViDAR architecture

representations rather than purely image-plane cues. Such supervision aligns well with BEVFormer's need for accurate multi-view geometry when forming BEV queries.

The Future Decoder is an autoregressive transformer that predicts future BEV features and point clouds. By forecasting future frames, the model learns temporal dynamics such as object motion, scene flow, and ego-motion. This temporal reasoning directly benefits BEVFormer's temporal self-attention, which must capture motion patterns and maintain consistency across frames for downstream 3D detection tasks.

IV. Results

In this section, we present quantitative and qualitative results on the nuScenes-mini datase together with the CAN bus expansion data [19]. We first study the effect of different pretraining strategies on BEVFormer, and then provide qualitative visualizations of the ViDAR-pretrained model. All BEVFormer-Base models are trained for 22 epochs under identical settings, varying only in the initialization methods.

A. Effect of Pretraining Strategies on BEVFormer Performance

Table I summarizes the 3D detection performance of BEVFormer-Base on the nuScenes-mini dataset with CAN bus extension under three initialization strategies: random initialization ("Scratch"), FCOS3D monocular 3D-detection pretraining, and ViDAR visual point-cloud forecasting pretraining. All models share the same ResNet101-DCN [20] backbone and BEVFormer encoder architecture, which aggregates multi-camera images into spatiotemporal BEV features via spatial cross-attention and temporal self-attention.

Overall Quantitative Results. Training BEVFormer from scratch performs very poorly on this small dataset: mAP is only **0.03%** and NDS is around **2%**, while the error metrics all remain close to 1.0. This confirms that BEVFormer strongly relies on a good initialization when only limited labeled data are available. In particular, without pretraining the model fails to learn a meaningful 3D geometry and motion representation from nuScenes-mini alone. Both task-specific pretraining strategies bring large gains over scratch. FCOS3D pretraining increases mAP to **19.22%** and NDS to **13.97%**, and substantially reduces translation and scale errors (mATE **0.948**, mASE **0.594**). ViDAR pretraining further improves on FCOS3D for most metrics. Initializing from ViDAR raises mAP from **19.22%** to **20.45%** (+1.23 absolute) and NDS from **13.97%** to **15.02%** (+1.05), indicating overall better 3D detection quality. Localization accuracy also benefits. ViDAR achieves the best orientation and attribute quality (mAOE **1.23**, mAAE **0.614**), suggesting that its visual point-cloud forecasting objective helps the BEV encoder capture richer 3D structure and object dynamics than monocular detection alone.

Metric	Scratch	FCOS3D Pretrain	ViDAR Pretrain	Best
mAP	0.03	19.22	20.45	ViDAR
NDS	2%	13.97	15.02	ViDAR
mATE	1.311	0.946	0.9	ViDAR
mASE	1.259	0.594	0.597	FCOS3D
mAOE	1.000+	1.246	1.23	ViDAR
mAVE	1.000+	1.509	1.759	FCOS3D
mAAE	1.000+	0.641	0.614	ViDAR

TABLE I: Performance of Pretraining

Comparison between FCOS3D and ViDAR pretraining. Although both FCOS3D and ViDAR are task-specific pretraining schemes for autonomous driving, they emphasize different aspects of the representation. FCOS3D is a monocular 3D detector whose training objective directly regresses 3D box centers, sizes and velocities from a single image. As a result, FCOS3D pretraining provides BEVFormer with strong priors on bounding-box extent and motion, which is reflected by its slightly better scale and velocity errors (mASE and mAVE). In contrast, ViDAR is pre-trained with the visual point cloud forecasting task, where the encoder is required to predict future LiDAR point clouds from historical multi-view images [21]. This cross-modal, temporal objective encourages the model to jointly capture scene semantics, fine-grained 3D geometry and object dy-

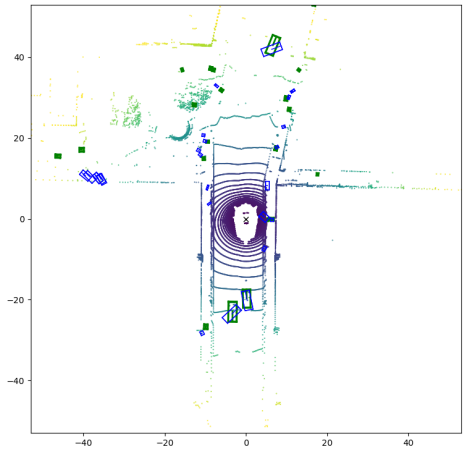

Fig. 3: BEV qualitative results with ViDAR-pretrained BEVFormer

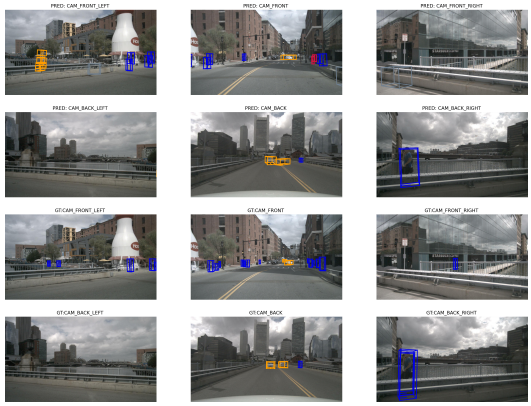

Fig. 4: Multi-view qualitative results with ViDAR-pretrained BEVFormer

namics at the BEV level. Consequently, ViDAR-initialized BEVFormer achieves higher overall detection quality (mAP and NDS) and better localization and orientation accuracy (mATE, mAOE, mAAE), while remaining comparable on scale and velocity. Overall, ViDAR offers a more holistic 3D representation that is particularly beneficial for BEV-based, multi-camera encoders under the low-data regime of nuScenes-mini, making it a stronger pretraining strategy than monocular 3D detection pretraining.

B. Qualitative Results with ViDAR-Pretrained BEVFormer

Qualitative examples in Fig. 3 illustrate how BEVFormer behaves when initialized with ViDAR pretraining. In the BEV visualization, the predicted 3D bounding boxes align well with both the LiDAR point cloud and the HD map priors: vehicles around the ego car are tightly localized and the boxes follow lane geometry and road boundaries. Most residual errors are concentrated on the yaw angle—predicted boxes are sometimes slightly rotated with respect to ground truth while their centers remain well aligned. This behavior is consistent with the inherent ambiguity of estimating object heading from camera-only observations, the discretization of BEV features, and the limited size of the nuScenes-mini training set. It also reflects the fact that ViDAR pretraining mainly encourages accurate 3D occupancy and motion representation, rather than explicitly enforcing highly precise box orientations. The multi-view image results in Fig. 4 further confirm these observations. Across the six surrounding cameras, most nearby vehicles and large pedestrians are correctly detected, and the projected 3D boxes match the image evidence in position and aspect ratio. Objects that are partially occluded by guardrails or captured from extreme viewpoints are occasionally missed or slightly missized, and very distant small targets remain challenging. Overall, however, the detections are spatially coherent across views and consistent with the BEV map, demonstrating that ViDAR pretraining endows BEVFormer with strong cross-view geometric understanding and robust multi-camera 3D perception, with only minor residual inaccuracies in object orientation.

V. Conclusion

In this work, we investigated the impact of pretraining strategies on BEVFormer for multi-camera 3D object detection. We compared three initialization methods—training from scratch, FCOS3D pretraining, and ViDAR pretraining—and found that initialization has a strong impact on performance. Training from scratch nearly fails (0.03% mAP), FCOS3D provides large gains, and ViDAR achieves the best results, benefiting from its video-based modeling of geometry and temporal dynamics. However, it shows that ViDAR's occupancy- and motion-focused objectives lead to less accurate orientation estimation.Our experiments clearly demonstrate this sensitivity: training BEVFormer from scratch yields almost no usable performance, indicating that the model cannot learn meaningful 3D structure from the limited supervision available in nuScenes-mini alone.

Our study is limited by using nuScenes-mini due to computational constraints, and future work should evaluate on the full dataset, explore additional pretraining objectives, and investigate combining complementary pretraining strategies.Future work may extend this analysis to larger datasets, integrate BEV-specific self-supervised objectives, or explore hybrid pretraining strategies that combine monocular geometric cues with temporal forecasting. Overall, our findings highlight the importance of task-aligned, geometry- and temporal-aware pretraining for BEV-based 3D perception.

Our implementation and experiment code are publicly available at:`https://github.com/yich0304/Vidar`

REFERENCES

[1] Z. Zhang, R. Girdhar, A. Joulin, and I. Misra, "Self-supervised pretraining of 3d features on any point-cloud," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 252–10 263.

[2] H. Liu, Y. Teng, T. Lu, H. Wang, and L. Wang, "Sparsebev: High-performance sparse 3d object detection from multi-camera videos," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 18 580–18 590.

[3] L. Chen, C. Sima, Y. Li, Z. Zheng, J. Xu, X. Geng, H. Li, C. He, J. Shi, Y. Qiao, *et al.*, "Persformer: 3d lane detection via perspective transformer and the openlane benchmark," in *European Conference on Computer Vision*. Springer, 2022, pp. 550–567.

[4] L. Fan, F. Wang, N. Wang, and Z.-X. Zhang, "Fully sparse 3d object detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 351–363, 2022.

[5] J. Gu, C. Hu, T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. 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*, 2023, pp. 5496–5506.

[6] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 853–17 862.

[7] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9297–9307.

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[9] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9710–9719.

[10] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, *et al.*, "A cookbook of self-supervised learning," *arXiv preprint arXiv:2304.12210*, 2023.

[11] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 913–922.

[12] J. Yang, J. M. Alvarez, and M. Liu, "Self-supervised learning of depth inference for multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7526–7534.

[13] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, "Monodtr: Monocular 3d object detection with depth-aware transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4012–4021.

[14] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. arxiv 2022," *arXiv preprint arXiv:2203.17270*.

[15] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany, "Pointcontrast: Unsupervised pre-training for 3d point cloud understanding," in *European conference on computer vision*. Springer, 2020, pp. 574–591.

[16] S. Huang, Y. Xie, S.-C. Zhu, and Y. Zhu, "Spatio-temporal self-supervised representation learning for 3d point clouds," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6535–6545.

[17] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 1477–1485.

[18] Z. Yang, L. Chen, Y. Sun, and H. Li, "Visual point cloud forecasting enables scalable autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 673–14 684.

[19] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[20] X. Wang, X. Wang, J. Li, M. Wang, H. Guo, and Y. Zhang, "Dsg-fsod: a few-shot object detection method based on deformable convolutions and attention integration," in *Complex Intelligent Systems*, 2025, pp. Volume 12, article number 21.

[21] Z. Yang, L. Chen, Y. Sun, and H. Li, "Visual point cloud forecasting enables scalable autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 14 673–14 684.