

Percept-WAM: Perception-Enhanced World-Awareness-Action Model for Robust End-to-End Autonomous Driving

Supplementary Material

A. More Method Details

A.1. The Attention Mask for Perception Tasks

To illustrate how Percep-WAM leverages World-PV and World-BEV tokens in the unified VLM backbone, Figure 10 visualizes the attention masks between input and output tokens for the PV- and BEV-detection tasks, respectively.

Specifically for the PV detection task, the attention mask is designed according to three principles: (i) World-PV tokens are fully visible to each other to better fuse PV features; (ii) for each grid-based prediction, the text tokens, grid tokens and output tokens follow the standard causal attention used in LLMs; (iii) to support grid-based parallel AR decoding, grid tokens and output tokens from different grid-based predictions are mutually masked. For BEV detection, two design choices apply: (i) similar to World-PV tokens, World-BEV tokens are fully mutually visible; (ii) to better capture the PV-to-BEV transformation and mitigate overfitting, each output token is constrained to attend only to World-BEV tokens and its corresponding grid token.

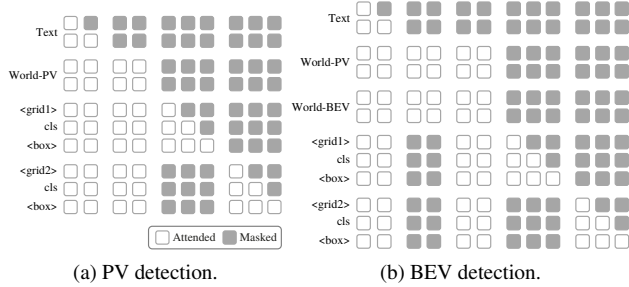


Figure 10. Illustration of attention mask for input tokens for (a) PV and (b) BEV detection tasks.

A.2. Streaming Inference

To meet the demands of real-world applications, we investigate the streaming inference capability of the Percept-WAM model for handling infinitely long conversational visual inputs. Inspired by streaming inference research in current mainstream VLMs [103, 104], we adopt a streaming strategy illustrated in Figure 11. Specifically, as shown in Figure 11a, the trajectory prediction tokens at time step T attend to the tokens of the two most recent frames (*i.e.*, frame T and frame $T-1$). Considering the high computational cost of processing visual tokens in the prefill phase, we reuse the KV cache from previous computations, with the specific strategy available for reference in Figure 11b.

However, inconsistencies between training and inference

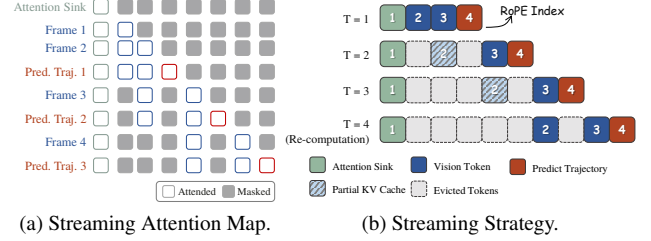


Figure 11. **Streaming inference strategy.** Each trajectory prediction attend previous two frames. The KV cache of historical frames (blue diagonally striped blocks in Fig.(b)) is reused to improve inference efficiency. A dual-recomputation strategy is proposed to mitigate the effects of discontinuous cache positions and error accumulation.

paradigms inevitably lead to *distribution drift*: the cross-attention mechanism between frames theoretically enables historical tokens to implicitly encode historical information of infinite length. This differs significantly from standard video-clip-based training, which relies on fixed-length historical information. Thus, adopting a longer-clip training scheme is crucial for enhancing the model’s generalization capability on extended historical sequences.

To stabilize the computation process while gradually discarding historical visual tokens, we retain the attention sink [105] (corresponding to the green grids in Figure 11b). However, this design causes deviations in positional embedding within the KV cache, arising from the discontinuity between the attention sink and visual frames. To address this issue, a dual-recomputation strategy is proposed with both local attention refinement and global cache recomputation: we recompute rotary positional embeddings (corresponding to the white numbers in Figure 11b) and adopt the specific token recomputation method [106] to correct cross-attention results. Meanwhile, we mitigate error accumulation from increasing inference lengths by periodically caching ViT tokens and recomputing the complete KV cache, safeguarding long-sequence stability.

As shown in Table 7, we achieve 16% and 40% reduction in inference latency for the AR and the Query decoders, respectively, with an accuracy loss of less than 0.01.

A.3. End-to-End Trajectory Prediction

Attention Mask for Query-based Trajectory Decoding. As described in Section 3.3, we introduce several sets of point-level queries and enforce modality-specific alignment by adjusting the attention mask. The attention mask is visualized in Figure 12: the first three query sets attend only

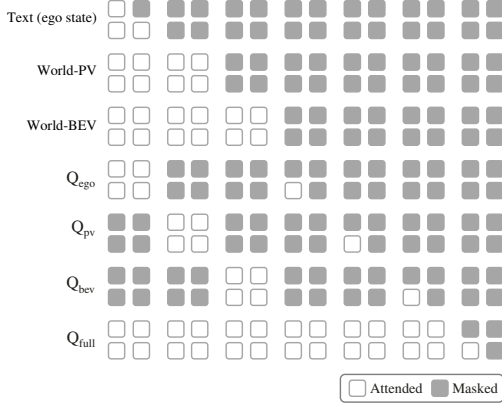


Figure 12. **Attention mask for our query-based trajectory decoding.** Q_{ego} , Q_{pv} , and Q_{bev} are aligned with their corresponding modality tokens, while Q_{full} accesses all features to decode the final trajectory.

to their corresponding modality tokens for trajectory decoding, while the final query set attends to all input tokens to generate the final trajectory.

Detail Settings of Percept-WAM as the Trajectory Selector. As mentioned in Section 4.2, training Percept-WAM solely to replicate ground-truth trajectories is insufficient for achieving strong closed-loop performance, as trajectory supervision often misaligns with real-world evaluation metrics [107]. To mitigate this gap, we introduce a query-based trajectory scoring and selecting approach inspired by Hydra-MDP [97] and GTRS [108], evaluating it on the NAVSIM v1 benchmark.

As showed in Figure 13, instead of direct trajectory replication, we train the model to score a super-dense, pre-clustered trajectory vocabulary \mathcal{V}_{XL} . Half of \mathcal{V}_{XL} is randomly dropped during training to improve robustness. At inference, a reduced vocabulary \mathcal{V}_L is used. Each trajectory in \mathcal{V}_L is first embedded through an MLP and then encoded as \mathcal{V}'_L via a stack of transformer layers:

$$\mathcal{V}'_L = \text{Transformer}(Q, K, V = \text{MLP}(\mathcal{V}_L)). \quad (1)$$

Percept-WAM further integrates contextual cues by querying features from the World-PV Tokens T_{WPV} , World-BEV Tokens T_{WBEV} and Text Tokens T_T , generating World-Action Tokens T_{WA} for trajectory scoring:

$$T_{WA} = \text{Percept-WAM}(Q = \mathcal{V}'_L, K, V = T_{WPV}, T_{WBEV}, T_T). \quad (2)$$

These enriched features are passed through a set of prediction heads to compute trajectory scores. Using a binary cross-entropy objective, we distill rule-based driving priors into our model. At inference time, our model selects the trajectory with the highest composite score, reflecting optimal performance for the given scenario.

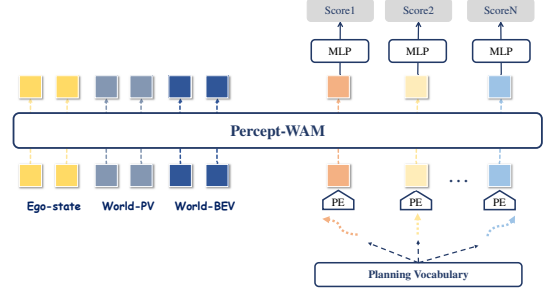


Figure 13. **Query-based trajectory scoring and selecting.** The embedded trajectory vocabularies are trained to be aligned with the World-PV, World-BEV, and Text Tokens, then decoded to generate the composite scores. The trajectory with the maximal score is selected as the planning result.

B. More Experiment Settings

B.1. Two-Stage Training Details.

Training Setting Details. To effectively optimize Percept-WAM for both perception and planning, we adopt a two-stage training scheme. The first stage focuses on enhancing the VLM’s overall 2D and 3D spatial perception, assisted by autonomous-driving general QA tasks, while the second stage trains end-to-end trajectory prediction on top of this perception-enhanced base model. The detailed training hyperparameters are summarized in Table 8.

Data Construction. As shown in Table 2 of the main paper, we employ task-specific training datasets for PV, BEV, and trajectory prediction. To preserve the model’s general capabilities, we further incorporate autonomous-driving QA data during training. For each task family, we specify the data mixture as follows: (i) for PV tasks, the sampling ratio among 2D Detection, Mono 3D Detection, Instance Segmentation, Semantic Segmentation, and Grounding is set to 1:1:2:1:1; (ii) for BEV tasks, the ratio between 3D Detection and BEV Segmentation is 1:1; (iii) when jointly training multiple main tasks (*i.e.*, PV, BEV, trajectory prediction, and Driving QA), we sample each task with equal probability for simplicity.

C. More Experiment Results

C.1. Main Results

Comparison of Visual Grounding Performance. In this section, we evaluate the model’s ability to leverage fine-grained world features for complex visual grounding. Visual grounding is a critical task that associates textual descriptions with corresponding image regions or objects. This task can be further divided into referring expression comprehension (REC) and referring expression segmentation (RES), with output formats of bounding boxes or masks. We present comprehensive comparison results for both tasks in Table 9 and Table 10, respectively. As

Table 8. Hyperparameter configurations for Percept-WAM across training stages. Note that within each stage, jointly trained tasks share the same base hyperparameters. All experiments are trained on 8 nodes.

Training parameter	Stage 1			Stage 2
	PV Perception	BEV Perception	Driving QA	Trajectory Prediction
Learning rate	0.0002	0.0002	0.0002	0.0002
Warmup iterations	1000	1000	1000	500
Training iterations	100000	100000	100000	3000
Batch size	64	64	64	64
Image resolution	1344 × 896	796 × 448	796 × 448	796 × 448
Grid number	10 × 10	40 × 40 for Det, 10 × 10 for Seg	NA	NA
Optimizer	AdamW	AdamW	AdamW	AdamW
Weight decay	0.01	0.01	0.01	0.01
Schedule	Cosine Annealing	Cosine Annealing	Cosine Annealing	Cosine Annealing

Table 9. Comparison of referring expression comprehension (REC) performance, reported using P@0.5. VGM and MLLM refer to the vision generalist model and multimodal large language model, respectively.

Method	Type	RefCOCO			RefCOCO+			RefCOCOg		Avg
		val	testA	testB	val	testA	testB	val	test	
MDETR [109]	VGM	86.8	89.6	81.4	79.5	84.1	70.6	81.6	80.9	81.8
Grounding DINO T [110]		89.2	91.9	86.0	81.1	87.4	74.7	84.2	84.9	84.9
Shikra-13B [111]	MLLM	87.8	91.1	81.8	82.9	87.8	74.4	82.6	83.2	84.0
MiniGPT-v2-7B [112]		88.1	91.3	84.3	79.6	85.5	73.3	84.2	84.3	83.8
VistaLLM-7B [113]		88.1	91.5	83.0	82.9	89.8	74.8	83.6	84.4	84.8
Percept-WAM		89.9	90.8	89.3	85.4	88.2	81.7	86.5	87.0	87.4

Table 10. Comparison of referring expression segmentation (RES) performance, reported using cumulative IoU (cIoU). VGM and MLLM refer to the vision generalist model and multimodal large language model, respectively.

Method	Type	RefCOCO			RefCOCO+			RefCOCOg		Avg
		val	testA	testB	val	testA	testB	val	test	
GLEE-Pro [114]	VGM	80.0	–	–	69.6	–	–	72.9	–	74.2
UNINEXT-H [115]		82.2	83.4	81.3	72.5	76.4	66.2	74.7	76.4	76.6
LISA-7B [116]	MLLM	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5	67.9
VistaLLM-13B [113]		77.2	78.7	73.9	71.8	74.4	65.6	69.8	71.9	72.9
GLaMM-7B [117]		79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	75.6
HiMTok-8B [118]		81.1	81.2	79.2	77.1	78.8	71.5	75.8	76.7	77.7
Percept-WAM		86.5	87.4	86.6	79.9	83.6	75.2	81.3	81.9	82.8

Table 11. Ablation studies of different trajectory decoding and selecting methods on NAVSIM [17] v1 benchmarks. ↓ indicates lower is better, ↑ indicates higher is better. Query-based trajectory scoring and selecting surpasses other trajectory planning mechanisms on the NAVSIM v1 benchmark.

Trajectory Planning Mechanism	NAVSIM v1					
	NC↑	DAC↑	TTC↑	Comf.↑	EP↑	PDMS↑
AR-based trajectory generation	96.4	94.5	92.0	98.7	78.5	84.1
Query-based trajectory generation	96.5	90.3	91.0	99.7	75.6	80.4
Query-based trajectory scoring and selection	98.8	98.6	94.4	99.5	84.8	90.2

shown in Table 9, Percept-WAM achieves top-tier performance on the RefCOCO [84], RefCOCO+ [85], and Ref-

Table 12. BEV segmentation results on nuScenes [18] with geographical train/val splits [107].

Modality	Method	Dri.	Ped.	Lane.	Veh.
C	EAFT [107]	58.06	—	—	—
	Percept-WAM	56.41	33.21	22.03	34.47
C + L	Percept-WAM	67.19	22.06	23.99	56.76

COCog [86] benchmarks among MLLMs, outperforming Grounding DINO [74], a representative VGM, by an average of 2.5 P@0.5. Table 10 demonstrates Percept-WAM’s exceptional pixel-level segmentation performance among VGMs and MLLMs, achieving an average cIoU of 82.8. We demonstrate Percept-WAM’s visual grounding performance on the same images with different descriptions, as shown in Figure 14, highlighting its robust ability to distinguish specific objects with unique attributes from visually similar instances.

BEV Segmentation Results on Geographical Train/Val Splits. For the BEV segmentation task, we additionally report results on nuScenes with the geographical train/val split (following the split protocol in EAFT [107]), as shown in Table 12. The results indicate that our Percept-WAM achieves performance comparable to EAFT, and that incorporating LiDAR inputs consistently provides stable performance gains.

C.2. Ablation Studies

Ablation of Query-based Trajectory Decoding Methods. We ablate different query-mask configurations for trajectory decoding, as summarized in Table 13. Compared to the “full” mode, which uses a single query set Q_{full} , parallel decoding with multiple query sets reduces the trajectory error by approximately 8%.

Ablation of Clustered-Action Design. The clustered-action method discretizes trajectories using the K-Disk clustering algorithm with a maximum vocabulary size of 2048. A greedy clustering approach is applied based on a 0.05-meter distance threshold to group similar trajectories. To improve long-horizon stability, trajectories are segmented into s -frame intervals, where $s \in \{1, 2, 3, 6\}$. As shown in Table 14, a segment length of 2 frames with a 0.05-meter threshold yields the best performance on nuScenes, achieving an L2_avg of 0.3919.

Comparison of E2E Performance. We ablate different trajectory planning strategies in the closed-loop setting on NAVSIM benchmark. Specifically, we compare three methods: (i) AR-based trajectory generation, (ii) query-based trajectory generation, (iii) query-based trajectory scoring and selection. The results are summarized in Table 11.

The comparison between AR-based and query-based methods in Table 7 appears inconsistent with the abla-

Table 13. Query mask ablation results on nuScenes trajectory prediction task. Lower is better. ‘Full’ means using a single query set which can attend to all the input tokens. ‘Parallel’ refers to our approach, where all query sets are decoded in parallel.

Mask mode	L2-avg↓
Full	0.4151
Parallel	0.3821

Table 14. Ablation of AR (cluster) configurations with different frame interval s on nuScenes trajectory prediction task. Lower(↓) is better.

Method	L2-1s↓	L2-2s↓	L2-3s↓	L2-avg↓
AR	0.160	0.356	0.674	0.3970
AR(cluster) $_{s=1}$	0.434	0.744	1.129	0.7692
AR(cluster) $_{s=2}$	0.181	0.356	0.638	0.3919
AR(cluster) $_{s=3}$	0.196	0.390	0.692	0.4260
AR(cluster) $_{s=6}$	0.236	0.494	0.892	0.5406

tion results in Table 11. This discrepancy arises from the misalignment between open-loop and closed-loop metrics. While the query-based method reduces the L2 distance between planned trajectories and ground truth (from 1.1 m to 0.8 m on the NAVSIM navtrain validation split), improved trajectory replication does not always lead to better closed-loop performance, as evidenced by PDMS on the NAVSIM [17] benchmark. Therefore, we adopt query-based trajectory scoring and selection which combines the imitation strength of the query-based approach with the closed-loop metrics, and achieves the best overall performance.

C.3. Illustrations

Trajectory Prediction Results. As shown in Figure 15, our model demonstrates strong trajectory planning in various challenging scenarios. It effectively handles decisions such as yielding to vehicles and pedestrians, navigating hazards in adverse weather, and accounting for occluded objects in low-visibility conditions. These examples highlight the model’s robustness and adaptability, even in long-tail cases.

D. Limitations

Limitations & Future Work. We plan to explore *more efficient and higher-accuracy architectures* for multi-task joint training (perception & trajectory prediction), such as Mixture-of-Experts (MoE), where different tasks will be routed to specialized experts. In addition, for both perception and end-to-end trajectory prediction, we aim to employ *a unified reinforcement learning framework with a multi-objective reward design* to jointly optimize these tasks.

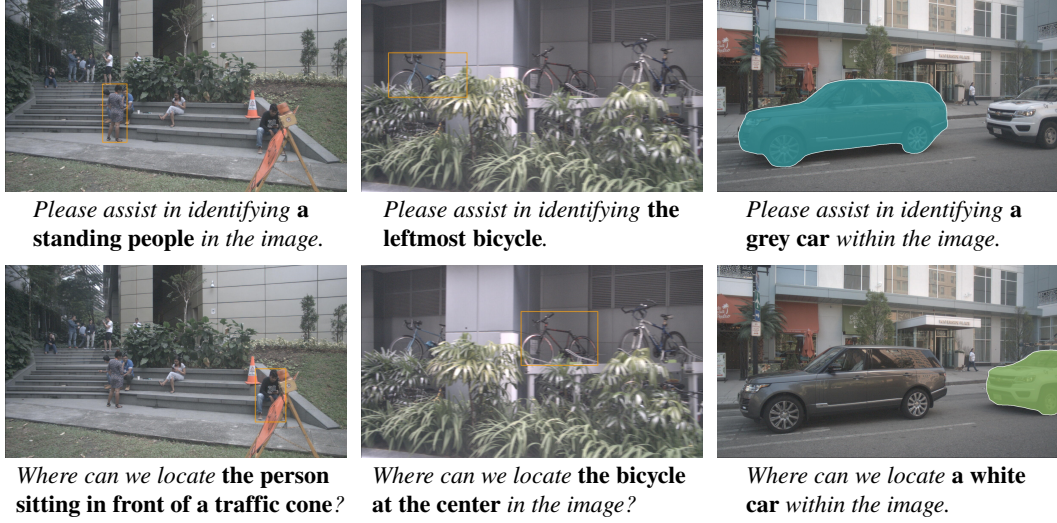


Figure 14. **Illustration of Percept-WAM on the visual grounding task.** Percept-WAM accurately localizes referred objects and exhibits robust understanding of diverse visual attributes.

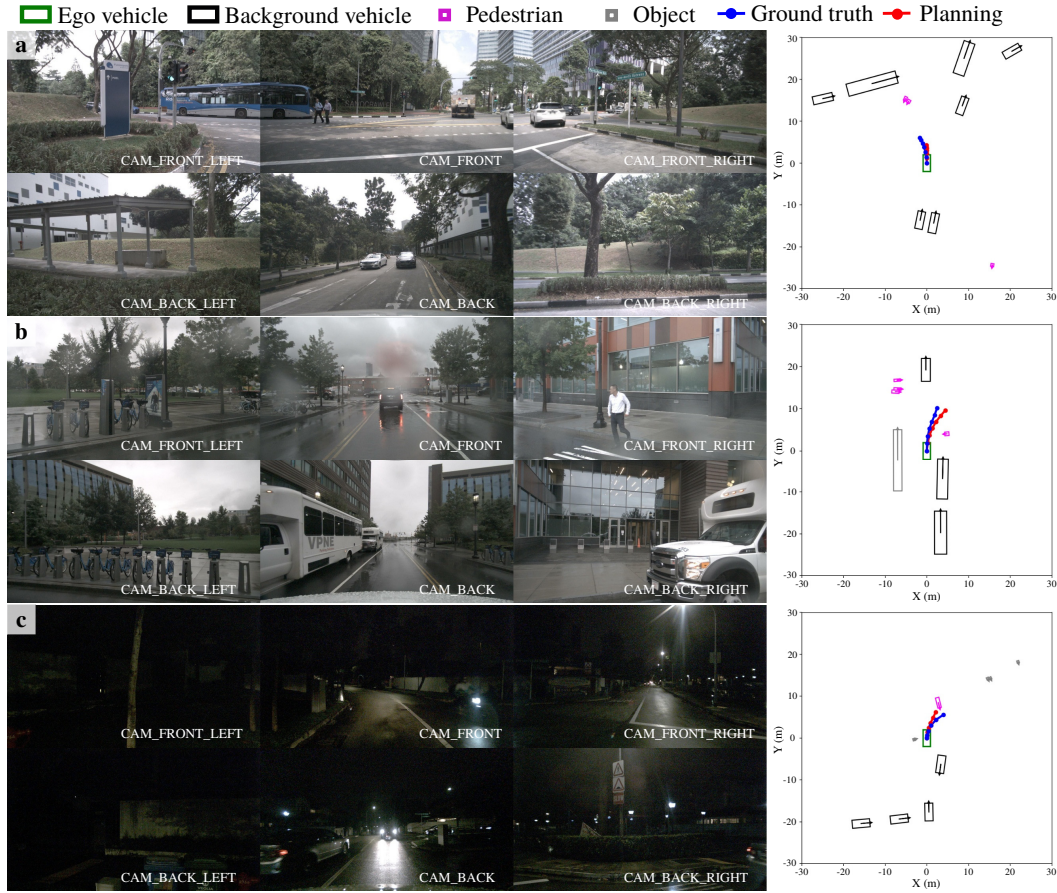


Figure 15. **More illustration of trajectory planning capabilities.** The ego vehicle is shown as a green box, background vehicles as black boxes, pedestrians as purple boxes, and other detected objects as grey boxes. Ground truth trajectories are shown as blue dots, and planned trajectories as red dots. (a) Percept-WAM successfully navigates the ego vehicle turning left, yielding to pedestrians crossing the road. (b) Under rainy conditions, the model detects a jaywalking pedestrian and safely avoids a potential collision. (c) While turning right at night, the model accounts for a cyclist obstructed in the front camera view, ensuring enough space for safe passage. Overall, our model demonstrates strong planning performance, even in challenging and long-tail scenarios.

References

- [1] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024. 1
- [2] Yinan Zheng, Ruiming Liang, Kexin Zheng, Jinliang Zheng, Liyuan Mao, Jianxiong Li, Weihao Gu, Rui Ai, Shengbo Eben Li, Xianyu Zhan, et al. Diffusion-based planning for autonomous driving with flexible guidance. *arXiv preprint arXiv:2501.15564*, 2025. 1
- [3] Rui Fan, Sicen Guo, and Mohammud Junaid Bocus. Autonomous driving perception. Cham, Switzerland: Springer, 2023. 1
- [4] Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. *arXiv preprint arXiv:2504.03164*, 2025. 1
- [5] Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Chenming Zhang, Shuai Liu, and Long Chen. Drivem-llm: A benchmark for spatial understanding with multimodal large language models in autonomous driving. *arXiv e-prints*, pages arXiv–2411, 2024. 1
- [6] Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Weihao Ding, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025. 1
- [7] Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkan Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025. 1
- [8] Yi Xu, Yuxin Hu, Zaiwei Zhang, Gregory P Meyer, Siva Karthik Mustikovela, Siddhartha Srinivasa, Eric M Wolff, and Xin Huang. Vlm-ad: End-to-end autonomous driving through vision-language model supervision. *arXiv preprint arXiv:2412.14446*, 2024. 1
- [9] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025. 1
- [10] 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. 1
- [11] Mingjie Pan, Jiyao Zhang, Tianshu Wu, Yinghao Zhao, Wenlong Gao, and Hao Dong. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17359–17369, 2025. 1
- [12] Max Argus, Jelena Bratulic, Houman Masnavi, Maxim Velikanov, Nick Heppert, Abhinav Valada, and Thomas Brox. cvla: Towards efficient camera-space vlans. *arXiv preprint arXiv:2507.02190*, 2025. 1
- [13] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 1
- [14] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7395–7408, 2025. 1
- [15] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024. 1, 2
- [16] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844, 2024. 1, 3
- [17] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024. 1, 2, 3, 6, 7, 17, 18
- [18] 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. 1, 2, 6, 18
- [19] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresight-drive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025. 1
- [20] Haohan Chi, Huan-ang Gao, Ziming Liu, Jianing Liu, Chenyu Liu, Jinwei Li, Kaisen Yang, Yangcheng Yu, Zeda Wang, Wenyi Li, et al. Impromptu vla: Open weights and open data for driving vision-language-action models. *arXiv preprint arXiv:2505.23757*, 2025. 1
- [21] Tianqi Wang, Enze Xie, Ruihang Chu, Zhenguo Li, and Ping Luo. Drivecot: Integrating chain-of-thought reasoning with end-to-end driving. *arXiv preprint arXiv:2403.16996*, 2024. 1, 2, 3
- [22] Wenhao Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36:61501–61513, 2023. 1, 6

- [23] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8469–8488, 2023. [1](#)
- [24] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12037–12047, 2025. [1](#), [2](#), [7](#)
- [25] Rui Zhao, Yuze Fan, Ziguang Chen, Fei Gao, and Zhenhai Gao. Diffe2e: Rethinking end-to-end driving with a hybrid action diffusion and supervised policy. *arXiv preprint arXiv:2505.19516*, 2025. [1](#)
- [26] Shu Liu, Wenlin Chen, Weihao Li, Zheng Wang, Lijin Yang, Jianing Huang, Yipin Zhang, Zhongzhan Huang, Ze Cheng, and Hao Yang. Bridgedrive: Diffusion bridge policy for closed-loop trajectory planning in autonomous driving. *arXiv preprint arXiv:2509.23589*, 2025. [1](#)
- [27] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. [2](#)
- [28] Qiming Zhang, Meixin Zhu, and Hao Frank Yang. Think-driver: From driving-scene understanding to decision-making with vision language models. In *European Conference on Computer Vision Workshop*, 2024. [2](#)
- [29] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhui Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023. [2](#), [7](#)
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. [2](#), [6](#)
- [31] Jincheng Li, Chunyu Xie, Ji Ao, Dawei Leng, and Yuhui Yin. Lmm-det: Make large multimodal models excel in object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 308–318, 2025. [2](#), [6](#)
- [32] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. [2](#), [3](#), [5](#), [7](#)
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. [2](#)
- [34] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. [2](#)
- [35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. [2](#)
- [36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. [2](#)
- [37] Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsafaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*, 2025. [2](#)
- [38] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *URL https://arxiv.org/abs/2307.15818*, 2024. [2](#)
- [39] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. [2](#)
- [40] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. [2](#), [7](#)
- [41] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer, 2024. [2](#), [3](#), [6](#)
- [42] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pages 194–210. Springer, 2020. [2](#)
- [43] 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*, volume 37, pages 1477–1485, 2023. [2](#)
- [44] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. Bevsegformer: Bird’s eye view semantic segmentation from arbitrary camera rigs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5935–5943, 2023. [2](#)
- [45] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. [2](#)
- [46] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-

task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 2, 4, 7

- [47] 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. 2
- [48] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [49] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 3
- [50] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenescs-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4542–4550, 2024. 3
- [51] Yiheng Li, Cunxin Fan, Chongjian Ge, Zhihao Zhao, Chenran Li, Chenfeng Xu, Huaxiu Yao, Masayoshi Tomizuka, Bolei Zhou, Chen Tang, et al. Womd-reasoning: A large-scale dataset for interaction reasoning in driving. *arXiv preprint arXiv:2407.04281*, 2024. 3
- [52] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1043–1052, 2023. 3
- [53] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 3
- [54] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pages 292–308. Springer, 2024. 3
- [55] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3, 4, 6
- [56] Hao Tang, Chenwei Xie, Haiyang Wang, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Ufo: A unified approach to fine-grained visual perception via open-ended language interface. *arXiv preprint arXiv:2503.01342*, 2025. 4, 6
- [57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4
- [58] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 4, 5
- [59] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 4
- [60] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer, 2022. 4
- [61] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jia-peng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 4
- [62] Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. *arXiv preprint arXiv:2510.12798*, 2025. 4
- [63] Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. Calibrating the confidence of large language models by eliciting fidelity. *arXiv preprint arXiv:2404.02655*, 2024. 4
- [64] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 784–799, 2018. 4
- [65] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in neural information processing systems*, 33:21002–21012, 2020. 4
- [66] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4, 5
- [67] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 4, 5
- [68] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 4, 7
- [69] 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. 4, 7
- [70] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan

- Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 5
- [71] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6
- [72] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022. 6
- [73] Haiyang Wang, Hao Tang, Li Jiang, Shaoshuai Shi, Muhammad Ferjad Naeem, Hongsheng Li, Bernt Schiele, and Liwei Wang. Git: Towards generalist vision transformer through universal language interface. In *European Conference on Computer Vision*, pages 55–73. Springer, 2024. 6
- [74] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 6, 18
- [75] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 913–922, 2021. 6
- [76] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 6
- [77] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 6
- [78] Yufei Zhan, Hongyin Zhao, Yousong Zhu, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon-g: Bridging vision-language and vision-centric tasks via large multimodal models. *arXiv preprint arXiv:2410.16163*, 2024. 6
- [79] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer, 2024. 6
- [80] Peng Liu, Haozhan Shen, Chunxin Fang, Zhicheng Sun, Jiajia Liao, and Tiancheng Zhao. Vlm-fo1: Bridging the gap between high-level reasoning and fine-grained perception in vlms. *arXiv preprint arXiv:2509.25916*, 2025. 6
- [81] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 6
- [82] 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
- [83] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 6
- [84] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 6, 17
- [85] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer, 2016. 6, 17
- [86] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 6, 18
- [87] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, and Oleg Sinavski. Lingoqa: Visual question answering for autonomous driving. *arXiv preprint arXiv:2312.14115*, 2023. 6
- [88] Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024. 6
- [89] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, pages 403–420. Springer, 2024. 6
- [90] Yuhang Lu, Yichen Yao, Jiadong Tu, Jiangnan Shao, Yuexin Ma, and Xinge Zhu. Can llms obtain a driver’s license? a benchmark towards reliable agi for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5838–5846, 2025. 6
- [91] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21819–21830, 2024. 6
- [92] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024. 6
- [93] 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. 7

- [94] Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024. 7
- [95] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahao Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14864–14873, 2024. 7
- [96] Chengran Yuan, Zhanqi Zhang, Jiawei Sun, Shuo Sun, Zefan Huang, Christina Dao Wen Lee, Dongen Li, Yuhang Han, Anthony Wong, Keng Peng Tee, et al. Drama: An efficient end-to-end motion planner for autonomous driving with mamba. *arXiv preprint arXiv:2408.03601*, 2024. 7
- [97] Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024. 7, 16
- [98] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [99] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017. 6
- [100] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016. 6
- [101] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 7
- [102] Zewei Zhou, Tianhui Cai, Seth Z Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*, 2025. 8
- [103] Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. *arXiv preprint arXiv:2510.09608*, 2025. 15
- [104] Zhenyu Ning, Guangda Liu, Qihao Jin, Wenchao Ding, Minyi Guo, and Jieru Zhao. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval. *arXiv preprint arXiv:2505.15269*, 2025. 15
- [105] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 15
- [106] Jiayi Yao, Hanchen Li, Yuhao Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. Cacheblend: Fast large language model serving for rag with cached knowledge fusion. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 94–109, 2025. 15
- [107] Christian Witte, Jens Behley, Cyrill Stachniss, and Marvin Raaijmakers. Epipolar attention field transformers for bird’s eye view semantic segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8660–8669. IEEE, 2025. 16, 18
- [108] Zhenxin Li, Wenhao Yao, Zi Wang, Xinglong Sun, Joshua Chen, Nadine Chang, Maying Shen, Zuxuan Wu, Shiyi Lan, and Jose M Alvarez. Generalized trajectory scoring for end-to-end multimodal planning. *arXiv preprint arXiv:2506.06664*, 2025. 16
- [109] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetrm: modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021. 17
- [110] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 17
- [111] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 17
- [112] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yonyang Xiong, and Mohamed Elhoseiny. Minigt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 17
- [113] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14076–14088, 2024. 17
- [114] Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3783–3795, 2024. 17
- [115] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023. 17
- [116] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 17
- [117] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 17

- [118] Tao Wang, Changxu Cheng, Lingfeng Wang, Senda Chen, and Wuyue Zhao. Himtok: Learning hierarchical mask tokens for image segmentation with large multimodal model. *arXiv preprint arXiv:2503.13026*, 2025. 17