MD4AD: DATASET PROPOSAL AND ANALYSIS

Daniel Yang* **Junhong Zhou*** **Patrick Chen*** **Tianzhi Li ***
{danielya, junhong2, bochunc, tianzhil}@andrew.cmu.edu

1 [4 POINTS] PROBLEM DEFINITION AND DATASET CHOICE

PROBLEM DEFINITION

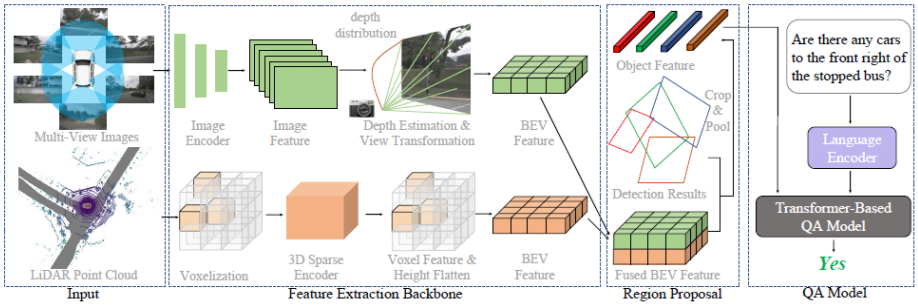

Figure 1: Baseline Framework Qian et al. (2024).

We aim to improve **visual question answering (VQA) in the autonomous driving domain** by enhancing the baseline model of **NuScenes-QA** Qian et al. (2024) with **Open Emma** Xing et al. (2024), an advanced QA model. Our work is based on the paper *NuScenes-QA: A Multi-Modal Visual Question Answering Benchmark for Autonomous Driving Scenario* by Tianwen Qian et al. Qian et al. (2024), which introduces a novel benchmark for VQA in autonomous driving. As above Figure 1 in the paper, our goal is to replace the existing QA model with **Open EMMA** while ensuring that the outputs align with the **ground truth** and **NuScenes-QA question-answering context** through a carefully designed **loss function**.

Additionally, we propose to **replace the feature extraction backbone** and **region proposal module** with **more advanced vision and LiDAR encoders**. We also explore the integration of **knowledge graphs or scene graphs (or both)** to improve spatial reasoning and object relationship understanding. This enhanced framework aims to **achieve higher accuracy and better generalization** in the NuScenes-QA benchmark.

DATASET CHOICE

We utilize **NuScenes-QA**, a large-scale **multi-modal visual question answering dataset** specifically designed for **autonomous driving**. This dataset is built upon **NuScenes**, a widely used **3D perception dataset**, and contains:

- **34,000+ driving scenes**
- **450,000+ question-answer pairs**
- **Multi-modal inputs**: multi-view images, LiDAR point clouds, and text (QA pairs)
- **Diverse question types**: existence, counting, object query, object-status query, and comparison

The NuScenes-QA dataset and related resources can be accessed at `https://github.com/qiantianwen/NuScenes-QA`.

* Everyone Contributed Equally – Alphabetical order

Compared to existing VQA benchmarks, NuScenes-QA introduces new challenges, such as **multi-frame temporal reasoning**, **multi-modal sensor fusion**, and **dynamic scene understanding**. These challenges make it an ideal dataset for evaluating the effectiveness of our proposed modifications.

1.1 [0.5 POINTS] WHAT PHENOMENA OR TASK DOES THIS DATASET HELP ADDRESS?

The NuScenes-QA dataset helps address **visual question answering (VQA) task for autonomous driving**, enabling models to understand **spatial status, object interactions, and temporal events in dynamic traffic environments**. The dataset provides **multi-modal data** (multi-view images, LiDAR, and text QA pairs) to improve perception-based question answering, making it useful for tasks like **scene understanding, object detection, and reasoning in complex urban scenarios**.

1.2 [0.5 POINTS] WHAT ABOUT THIS TASK IS FUNDAMENTALLY MULTIMODAL?

This task is fundamentally multimodal because **it requires integrating multiple sensor modalities**—RGB images and LiDAR point clouds—to sense the environment, and joint learning with language input (text) to **derive meaningful answers**. Questions often require **correlating spatial and depth information** from different perspectives, making it insufficient to rely on a single modality. For example, a question like "What is the distance between the red car and the pedestrian?" requires using **RGB images** for color detection and **LiDAR point clouds** for precise depth estimation. The dataset is designed in such a way that **cross-modal reasoning** is required, where **vision, language, and 3D spatial information** collectively contribute to accurate question answering.

1.3 HYPOTHESES

[1 points] Hypothesis Fusion of camera and LiDAR data improves object recognition in complex driving environments.

We hypothesize that by combining camera images and LiDAR point clouds, we can significantly improve the accuracy of object recognition in the autonomous driving setting. The camera data provides rich visual details, while LiDAR offers precise depth information, especially in challenging environments like night driving or adverse weather conditions. This hypothesis will be tested by measuring the improvement in object detection accuracy when both modalities are used together, compared to using camera data or LiDAR data individually in the NuScenes-QA dataset.

[1 points] Hypothesis Scene graphs enhance object relationship understanding in question answering.

We hypothesize that scene graphs improve the understanding of object relationships by explicitly linking visual data (e.g., images and LiDAR point clouds) with natural language, enhancing the accuracy of question answering in autonomous driving. Scene graphs provide a structured and semantically rich representation of the environment, allowing the model to better grasp spatial and relational connections between objects (e.g., "car is near pedestrian," "traffic light is ahead of vehicle"). This explicit representation helps the model answer complex questions by leveraging object relationships, such as "What is the distance to the nearest traffic sign?" The hypothesis will be tested by comparing question-answering models that utilize scene graphs for reasoning with those that process raw sensor data directly, assessing how scene graphs improve the model's ability to understand and reason about object relationships in the NuScenes-QA dataset.

[1 points] Hypothesis Incorporating knowledge graphs enhances spatial reasoning ability in driving-related question answering.

We hypothesize that by integrating knowledge graphs into the question-answering process, the model will improve its ability to reason about spatial relationships and interactions between objects, which is crucial for One-Hop (H1) questions in the NuScenes-QA dataset. These questions require the model to understand the positioning, movement, or proximity of objects in the scene, such as determining which vehicle is ahead of another or how far apart two objects are. Knowledge graphs can provide structured, relational information about object types, spatial constraints, and dynamic behaviors that are difficult to infer from RGB and LiDAR data alone. This hypothesis will

be tested by measuring the model's performance on One-Hop questions with and without the use of a knowledge graph, focusing on spatial reasoning tasks like identifying object relationships or predicting object trajectories in the NuScenes-QA dataset.

2 [6 POINTS] DATASET ANALYSIS

2.1 [1 POINTS] DATASET PROPERTIES

The **NuScenes-QA** dataset is designed to support visual question answering (VQA) in autonomous driving scenarios. It integrates multiple data modalities, including images, LiDAR point clouds, and rich annotations for comprehensive scene understanding. Key properties of the dataset include:

- **Dataset Size:** Over **34,000 driving scenes** and **450,000 question-answer pairs** under diverse urban environments.
- **Storage Size:** Approximately **1.4 TB** of data, including sensor data, annotations, and metadata.
- **Framerate:** Sensor data captured at **2 Hz** for LiDAR and **12 Hz** for cameras.
- **Physical Hardware Platform:** Collected using a **full-scale autonomous vehicle platform** equipped with **six cameras**, **five radars**, and **one 32-beam LiDAR sensor**.
- **Modalities:** Multi-modal inputs including **multi-view RGB images**, **LiDAR point clouds**, **radar data**, and **text of Question-Answering pairs**.
- **Annotation Types:** Detailed annotations with **3D bounding boxes**, **object categories**, **tracking IDs**, and **semantic maps** to support complex reasoning tasks.
- **Question Types:** A diverse set of question formats such as **existence**, **counting**, **object recognition**, **status analysis**, and **comparison** questions.
- **Class Distribution:** The dataset contains a total of 23 object classes, with the majority being **car** objects, followed by **adult** and **barrier** objects. This distribution highlights a noticeable class imbalance within the dataset.
- **Temporal Sequences:** Multi-frame sequences for analyzing **temporal dynamics** and **object motion**.
- **Scene Complexity:** Includes diverse driving scenarios with varying **weather conditions**, **lighting**, and **traffic densities**, reflecting real-world autonomous driving environments.

2.2 [0.5 POINTS] COMPUTE REQUIREMENTS

1. Files (can fit in RAM?) The nuScenes dataset is relatively large (around 1.4TB in total), which generally exceeds typical RAM capacities (e.g., 32GB or 64GB) of standard workstations. Consequently, most workflows involve storing the dataset on a local or network drive and loading subsets of the data into memory as needed (e.g., mini-batch loading during training). We will store the data in the server and using batch processing to retrieve the data while training.
2. Models (can fit on GCP/AWS GPUs?) Based on our analysis, we plan to use OpenEMMA with Llama-3.2-11B-Vision-Instruct (nearly 44GB) as our target model and fine-tune it using our four NVIDIA A6000 GPUs. Each A6000 provides 48GB of VRAM, giving us a total of 192GB, which should be sufficient to handle the memory demands of an 11-billion-parameter model. This setup allows us to distribute the computational load effectively across the GPUs, ensuring efficient training and fine-tuning. Compared to cloud-based GPU instances on platforms like GCP or AWS, our local configuration offers a cost-effective solution, provided we optimize the fine-tuning process to fully leverage multi-GPU parallelism and manage memory efficiently.

2.3 [2 POINTS] MODALITY ANALYSIS

Using a small sample of the data (e.g. validation splits), generate statistics and plots for three relevant properties of the data.

1. QA Paring Texts Analysis (Figure 2):

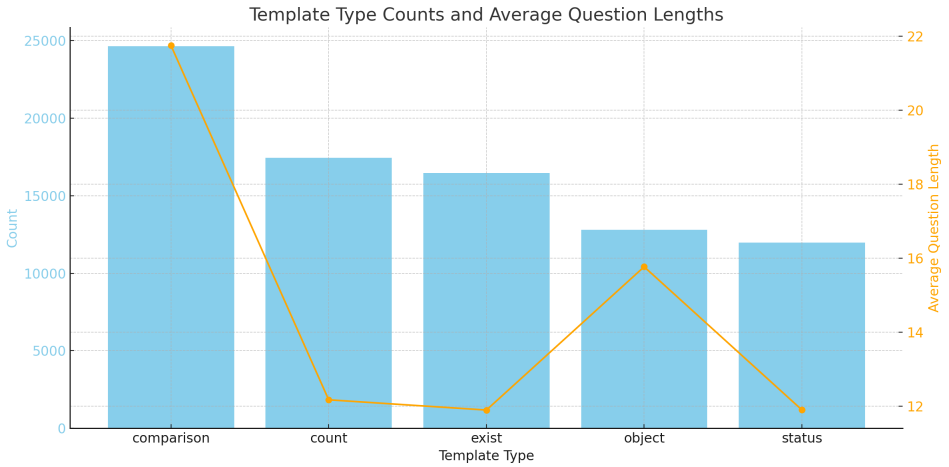

Figure 2: The bar chart represents the count of questions for each template type from NuScenes-QA, shown in blue. The orange line illustrates the average length of questions (in words) for each template type.

2. Category Distribution (Figure 3)

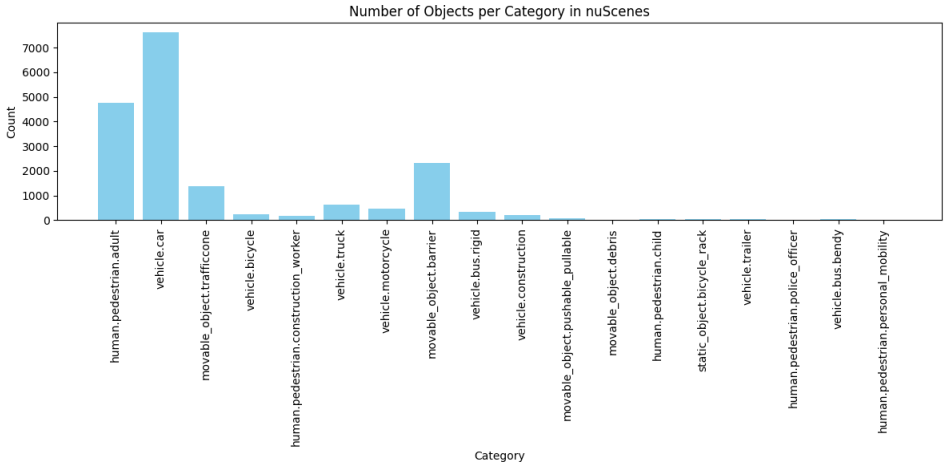

Figure 3: This bar chart illustrates the distribution of object categories in the nuScenes mini dataset. Each bar represents the total number of annotated instances for a specific category, such as vehicles, pedestrians, and movable objects.

3. Distribution of Bounding Boxes per Scene (Figure 4)

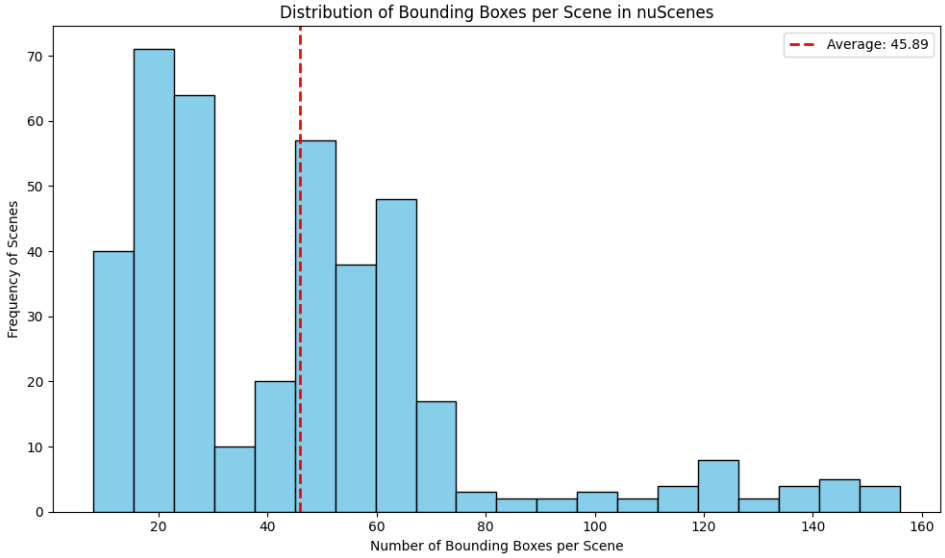

Figure 4: This histogram illustrates the distribution of bounding boxes per scene in the nuScenes mini dataset. The x-axis represents the number of bounding boxes in a scene, while the y-axis indicates the frequency of the scene with that number of bounding boxes. The red dashed line marks the average number of bounding boxes per scene.

4. Number of Objects per Scene (Figure 5)

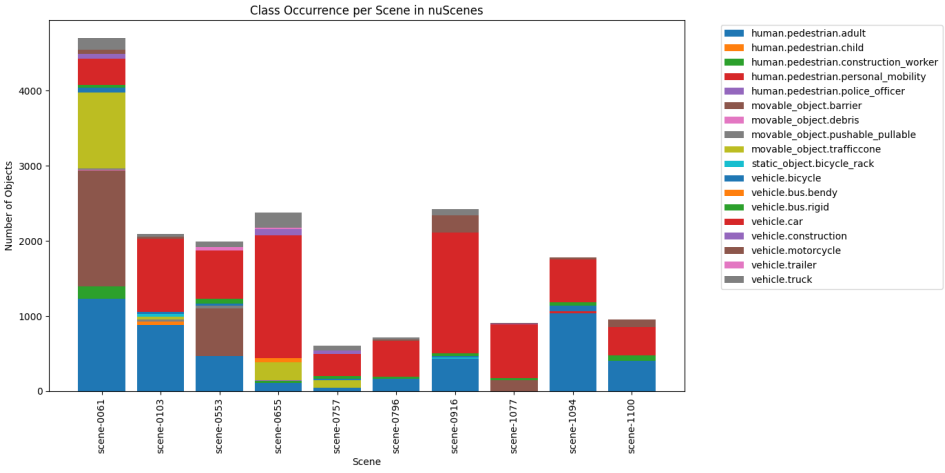

Figure 5: This stacked bar chart illustrates the distribution of object classes across different scenes in the nuScenes mini dataset. Each bar represents a scene, with colored segments indicating the number of occurrences of various object categories, such as vehicles, pedestrians, and movable objects.

5. Location and Date Distribution Visualization (Figure 6)

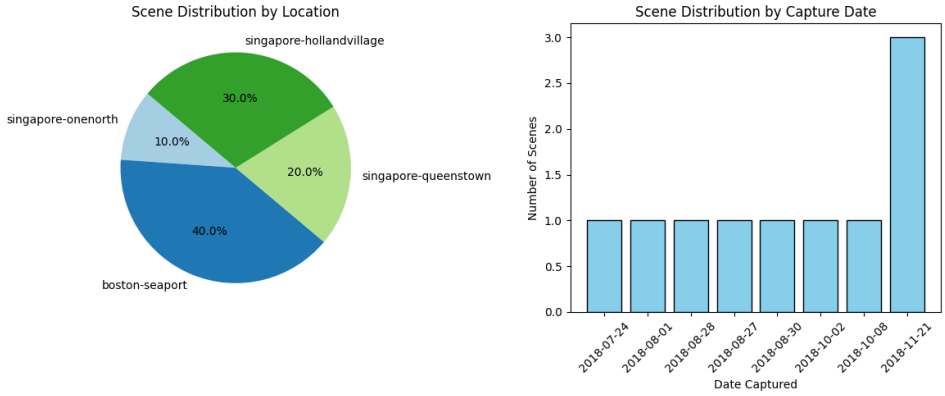

Figure 6: The pie chart illustrates the distribution of scenes across different locations in the nuScenes mini dataset, such as 'Singapore' or 'Boston'. The bar graph shows the number of scenes captured on different dates.

2.4 [0.5 POINTS] METRICS USED

The primary evaluation metric used in NuScenes-QA is **Top-1 Accuracy**, which measures the proportion of correctly predicted answers out of the total number of questions. The accuracy is reported across different **question categories** and **reasoning complexities**, including:

- **Existence (Exist)**: Checks whether specific objects exist in the scene.
- **Counting (Count)**: Evaluates the model's ability to count objects meeting certain criteria.
- **Object Recognition (Object)**: Measures the accuracy of identifying specific objects based on descriptions.
- **Status Recognition (Status)**: Assesses the model's ability to determine the status (e.g., moving, parked) of objects.
- **Comparison (Comparison)**: Evaluates the model's performance in comparing attributes between objects.

Additionally, the dataset distinguishes between:

- **Zero-Hop (H0)**: Questions that require no complex reasoning across objects.
- **One-Hop (H1)**: Questions that involve spatial reasoning or relationships between objects.

The accuracy is calculated for both **H0 and H1 separately** and then averaged to report the overall performance.

2.5 [2 POINTS] BASELINES

The following baselines are all listed in the benchmark in nuScenes-QA paper Qian et al. (2024)

- **Q-Only Baseline**: This baseline model only considers the question text without using any visual input. It helps to understand how much language biases contribute to the performance. Although it performs decently for simple existence questions, it struggles with more complex reasoning tasks due to the absence of visual context.
- **BEVDet + BUTD (Bottom-Up and Top-Down Attention)**: This model combines BEVDet Huang et al. (2021), which processes multi-camera images to generate Bird's Eye View (BEV) features, with the BUTD attention mechanism Anderson et al. (2018). It improves performance by focusing on salient regions in the images. This combination is effective for object recognition tasks but struggles with complex reasoning due to limited LiDAR integration.

- **CenterPoint + MCAN (Modular Co-Attention Network)**: CenterPoint Yin et al. (2021) uses LiDAR data for precise 3D object detection, while MCAN Yu et al. (2019) handles the interaction between visual and textual features through attention mechanisms. This model shows improved performance for spatial reasoning tasks, leveraging LiDAR's strength in capturing structural information.
- **MSMDFusion + MCAN**: This is the strongest baseline, fusing multi-scale features from both LiDAR and camera data (MSMDFusion) Jiao et al. (2023) and utilizing MCAN Yu et al. (2019) for advanced feature interaction. The fusion approach significantly boosts performance by exploiting complementary information from different modalities, especially for tasks like comparison and status recognition.

Models	**Exist**			**Count**			**Object**			**Status**			**Comparison**			**Acc**
	H0	H1	All	H0	H1	All	H0	H1	All	H0	H1	All	H0	H1	All	
Q-Only	81.7	77.9	79.6	17.8	16.5	17.2	59.4	38.9	42.0	57.2	48.3	51.3	79.5	65.7	66.9	53.4
BEVDet+BUTD	87.2	80.6	83.7	21.7	20.0	20.9	69.4	45.2	48.8	55.0	50.5	52.0	76.1	66.8	67.6	57.0
CenterPoint+BUTD	87.7	81.1	84.1	21.9	20.7	21.3	70.2	45.6	49.2	62.8	52.4	55.9	81.6	68.0	69.2	58.1
MSMDFusion+BUTD	89.4	81.4	85.1	25.3	21.3	23.2	73.3	48.7	52.3	67.4	55.4	59.5	81.6	67.2	68.5	59.8
GroundTruth+BUTD	98.9	87.2	92.6	76.8	38.7	57.5	99.7	71.9	76.0	98.8	81.9	87.6	98.1	76.1	78.1	79.2
BEVDet+MCAN	87.2	81.7	84.2	21.8	19.2	20.4	73.0	47.4	51.2	64.1	49.9	54.7	75.1	66.7	67.4	57.9
CenterPoint+MCAN	87.7	82.3	84.8	22.5	19.1	20.8	71.3	49.0	52.3	66.6	56.3	59.8	82.4	68.8	70.0	59.5
MSMDFusion+MCAN	89.0	82.3	85.4	23.4	21.1	22.2	75.3	50.6	54.3	69.0	56.2	60.6	78.8	68.8	69.7	60.4
GroundTruth+MCAN	99.6	95.5	97.4	52.7	39.9	46.2	99.7	86.2	88.2	99.3	95.4	96.7	99.7	90.2	91.0	84.3

Table 1: Top-1 accuracy across different question types in the NuScenes-QA test set. H0 denotes zero-hop and H1 denotes one-hop. C denotes camera, L denotes LiDAR.

3 TEAM

3.1 EXPERTISE

We have the following expertise in the underlying modalities required by this task:

1. Patrick Chen: First-year MSCV student. Research paper in CV, DL framework projects.
2. Junhong Zhou: First-year MSCV student. Research on CV/Autonomous driving.
3. Daniel Yang: First-year MSCV student. Research on motion prediction for Autonomous driving applications; took Deep Learning Systems in Fall 2024.
4. Tianzhi Li: First-year MSCV student. Research on Vision-Language models.

REFERENCES

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. URL https://arxiv.org/abs/1707.07998.

Scott M. Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Benjamin Sapp, C. Qi, Yin Zhou, Zoey Yang, Aurelien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Drago Anguelov. Large scale interactive motion forecasting for autonomous driving : The waymo open motion dataset. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9690–9699, 2021. URL https://api.semanticscholar.org/CorpusID:233307215.

Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. In *arXiv preprint arXiv:2112.11790*, 2021. URL https://arxiv.org/abs/2112.11790.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, James Guo, Drago Anguelov, and Mingxing Tan. Emma: End-to-end multimodal model for autonomous driving. *ArXiv*, abs/2410.23262, 2024. URL https://api.semanticscholar.org/CorpusID:273695673.

Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023. URL https://arxiv.org/abs/2209.03102.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. *arXiv preprint arXiv:2305.14836*, 2024. URL https://github.com/qiantianwen/NuScenes-QA.

Shuo Xing, Chengyuan Qian, Yuping Wang, Hongyuan Hua, Kexin Tian, Yang Zhou, and Zhengzhong Tu. Openemma: Open-source multimodal model for end-to-end autonomous driving. *arXiv*, December 2024. doi: 10.48550/arXiv.2412.15208. URL https://github.com/taco-group/OpenEMMA.

Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Centerpoint: A unified framework for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL https://arxiv.org/abs/2006.11275.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019. URL https://arxiv.org/abs/1906.10770.

A APPENDIX

A.1 EXTRA DATASET

We also evaluate the Waymo Open Motion Dataset (WOMD)Ettinger et al. (2021) as it provides a comprehensive and diverse collection of motion data for autonomous driving research. WOMD includes high-quality sensor data from lidars and cameras, detailed object trajectories, and high-definition 3D maps, covering a wide range of real-world driving scenarios. Its rich annotations and large-scale data make it ideal for tasks such as motion prediction, trajectory forecasting, and behavior modeling of road agents like vehicles, pedestrians, and cyclists.

1. Data Volume (GBs): The dataset comprises over 100,000 segments, each lasting 20 seconds, resulting in more than 570 hours of data. The total storage requirement for the dataset is substantial, though the exact size in gigabytes is not specified.

2. Framerate: Data is recorded at a frequency of 10 Hz, meaning each second of data contains 10 frames

3. Physical Hardware Platform: Data collection was performed using Waymo's autonomous vehicles, which are equipped with an array of sensors, including LiDAR and cameras. These sensors are precisely synchronized and calibrated to ensure high-quality data acquisition.

4. Data Types: The dataset includes:

- **Lidar Data**: The dataset contains data from five lidars - one mid-range lidar (top) and four short-range lidars (front, side left, side right, and rear).
- **Camera Data**: The dataset contains images from five cameras associated with five different directions. They are front, front left, front right, side left, and side right. One camera image is provided for each pair in JPEG format.
- **Object Trajectories**: Detailed 3D bounding boxes tracking the movement of various road agents such as vehicles, pedestrians, and cyclists.
- **High-Definition 3D Maps**: Comprehensive maps providing context for the trajectories.

A.2 Compute Requirements

1. Files (can fit in RAM?) The Waymo Open Motion Dataset (WOMD) comprises over 100,000 segments, each 20 seconds long at a 10 Hz sampling rate, resulting in more than 570 hours of data and exceeding 1 TB in total size. The full dataset is too large to fit into RAM for typical workstations. However, it can be processed in smaller batches or individual segments, depending on the available memory. High-performance machines with large RAM capacities may handle more extensive portions of the dataset in memory, but for most use cases, data streaming or chunk-wise processing will be necessary.
2. Models (can fit on GCP/AWS GPUs?)
   Baseline models provided for WOMD are designed to be compatible with common cloud GPU platforms like GCP and AWS. These models like EmmaHwang et al. (2024) can be trained and deployed on GPUs such as NVIDIA Tesla V100, A100, or T4, which are available on GCP and AWS. The computational demands will vary depending on the model complexity and batch size, but standard GPU instances should suffice for training and inference tasks related to WOMD.

A.3 Modality analysis

Using a small sample of the data (e.g. validation splits), generate statistics and plots for three relevant properties of the data.

1. Percent of Scenes vs Number of Agents
2. Trajectories and Maximum Speed Analysis

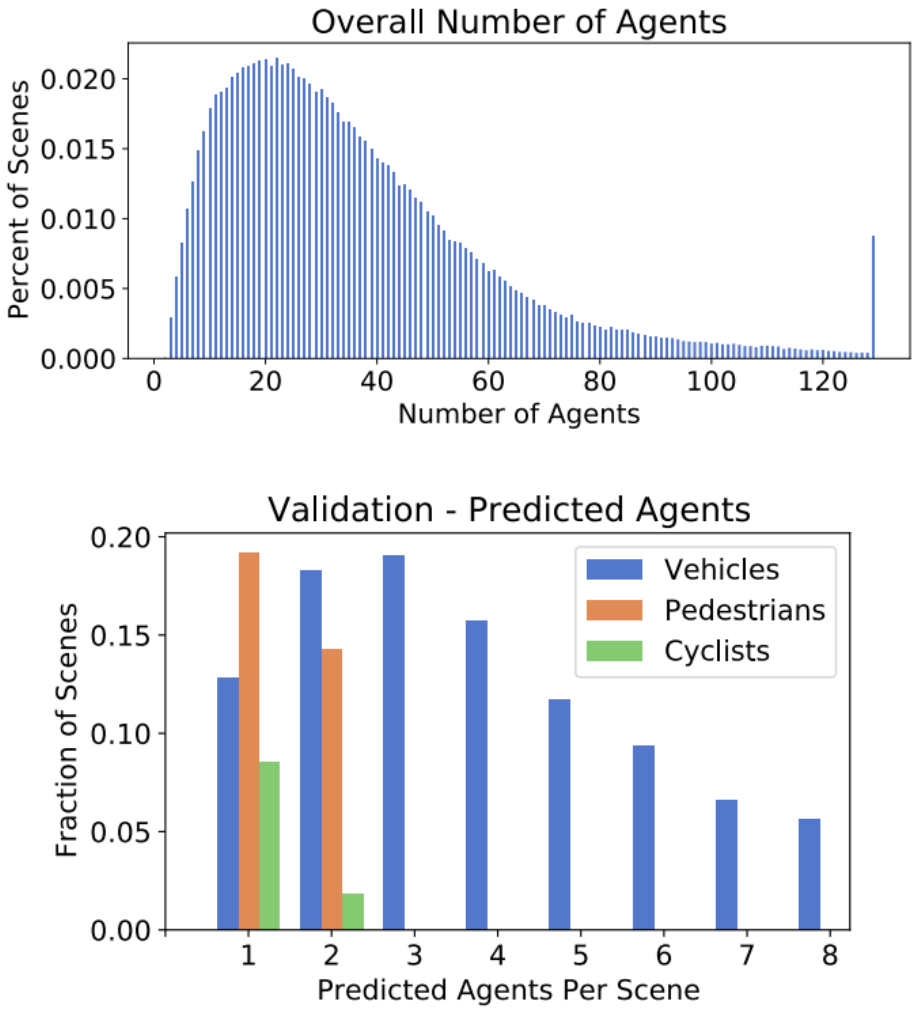

Figure 7: The Waymo Open Motion Dataset(WOMD)Ettinger et al. (2021) contains many agents including pedestrians and cyclists. Top: 46% of scenes have more than 32 agents, and 11% of scenes have more than 64 agents. Bottom: In the standard validation set, 33.5% of scenes require at least one pedestrian to be predicted, and 10.4% of scenes require at least one cyclist to be predicted.

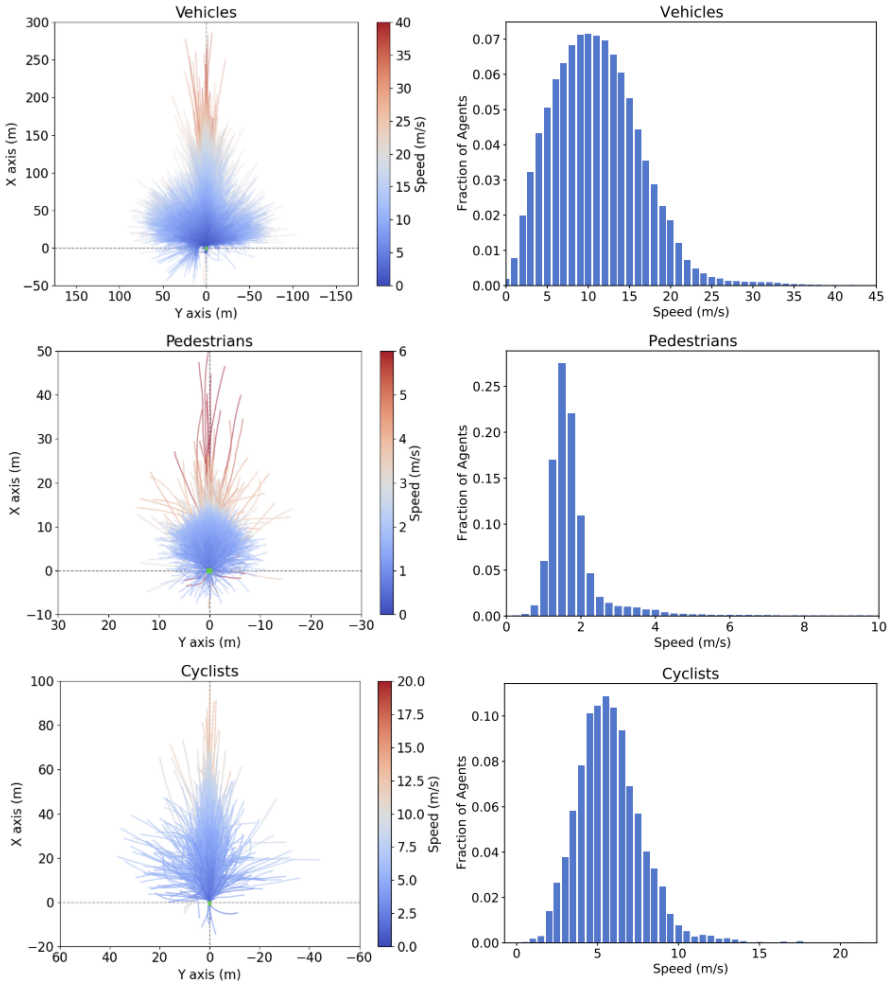

Figure 8: Agents selected to be predicted have diverse trajectories. Left: Ground truth trajectory of each pre- dicted agent in a frame of reference where all agents start at the origin with heading pointing along the positive X axis (pointing up). Right: Distribution of maximum speeds achieved by all of the agents along their 9 second trajectory. Plots depict variety in trajectory shapes and speed profiles.