Paper Reading: EMMA: End-to-End Multimodal Model for Autonomous Driving [HXL+25]

Lin Li

November 9, 2025

Overview

Autonomous Driving Challenges

- Traditional modular approach: specialized components for perception, mapping, prediction, planning
- Challenges:
  - Limited inter-module communication
  - Expert-designed interfaces struggle with novel environments
  - Difficult to scale
- End-to-end systems: direct learning from sensors to actions
- Problem: Often specialized for specific tasks, limited generalization

Multimodal LLMs for Autonomous Driving

- **Key advantages of MLLMs:**
  - Trained on vast, internet-scale datasets (rich world knowledge)
  - Superior reasoning capabilities (chain-of-thought)
- **Our approach: EMMA**
  - Built on top of Gemini (MLLM)
  - Treat MLLM as first-class citizen
  - Recast driving tasks as Visual Question Answering (VQA)
  - Leverage pre-trained world knowledge and reasoning

EMMA Overview

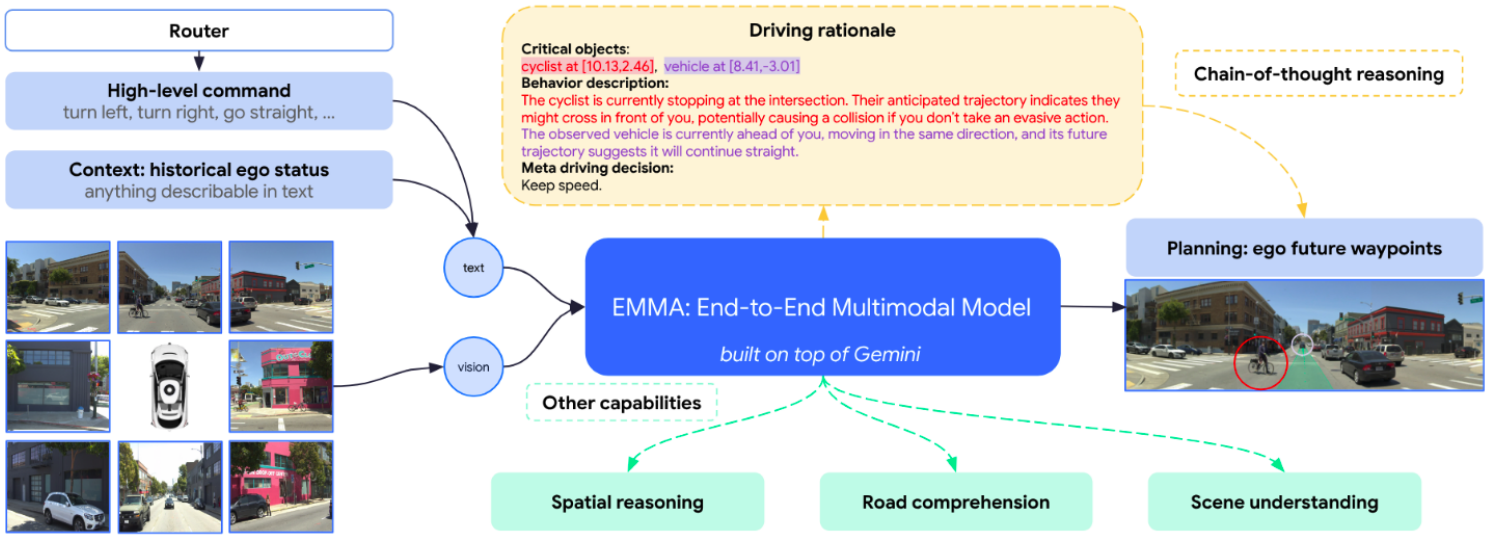


Figure 1: EMMA overview diagram. It takes 3 inputs (**left**): 1) a high-level command from the router, 2) historical status of the ego vehicle, and 3) surround-view camera videos. The model then predicts ego future trajectories (**right**) for motion planning that will be transformed into vehicle driving control signals. Further, we can ask the model to explain its rationale (**top**) before predicting trajectories, which enhances both the performance and explainability of the model through chain-of-thought reasoning. Notably, we incorporate visual grounding into the rationale so that the model also predicts the accurate 3D/BEV location for critical objects. In addition to end-to-end planning, we highlight three additional perception capabilities of our model (**bottom**).

- **Inputs:** Camera videos + text (commands, ego history)
- **Outputs:** Trajectories, 3D objects, road graphs, rationales
- **Key innovation:** All non-sensor data as natural language

Core Formulation

Base Model:

$$\mathcal{O} = \mathcal{G}(\mathbf{T}, \mathbf{V}) \tag{1}$$

Where:

- $\mathcal{G}$: Gemini model
- $\mathbf{T}$: Natural language prompts (text)
- $\mathbf{V}$: Images or videos
- $\mathcal{O}$: Natural language outputs

Key design choice: Represent 3D coordinates as text

- $\mathbf{T}_{\text{coordinates}} = \{(x_i, y_i)\} \approx \text{text}(\{(x_i, y_i)\})$
- Unified language representation space
- Maximizes reuse of pre-trained knowledge

End-to-End Motion Planning

Three key inputs:

1. **Surround-view camera videos** ($\mathbf{V}$)
2. **High-level intent command** ($\mathbf{T}_{\text{intent}}$):
   e.g., "go straight", "turn left", "turn right"
3. **Historical ego status** ($\mathbf{T}_{\text{ego}}$):
   Set of past waypoints $\{(x_t, y_t)\}_{t=-T_h}^{-1}$ in BEV space

Output: Future trajectory waypoints

$$\mathcal{O}_{\text{trajectory}} = \{(x_t, y_t)\}_{t=1}^{T_f} \tag{2}$$

Complete formulation:

$$\mathcal{O}_{\text{trajectory}} = \mathcal{G}(\mathbf{T}_{\text{intent}}, \mathbf{T}_{\text{ego}}, \mathbf{V}) \tag{3}$$

Characteristics of Motion Planning

Three key properties:

1. **Self-supervised:** Only requires future ego locations (no human labels)
2. **Camera-only:** No LiDAR or radar needed
3. **HD map free:** No detailed HD maps beyond high-level routing (e.g., Google Maps)

- Scalable data generation pipeline
- Simplified sensor requirements
- Reduced dependency on expensive mapping infrastructure

Chain-of-Thought Reasoning

Enhanced formulation with reasoning:

$$(\mathcal{O}_{\text{rationale}}, \mathcal{O}_{\text{trajectory}}) = \mathcal{G}(\mathbf{T}_{\text{intent}}, \mathbf{T}_{\text{ego}}, \mathbf{V}) \tag{4}$$

Rationale components ($\mathcal{O}_{\text{rationale}}$):

- **R1 - Scene description:** Weather, time of day, road conditions
- **R2 - Critical objects:** On-road agents with precise 3D/BEV coordinates
  e.g., "pedestrian at [9.01, 3.22], vehicle at [11.58, 0.35]"
- **R3 - Behavior description:** Status and intent of critical objects
- **R4 - Meta driving decision:** High-level driving plan (12 categories)

Key Benefit

Improves performance **and** explainability

EMMA Generalist: Multiple Tasks

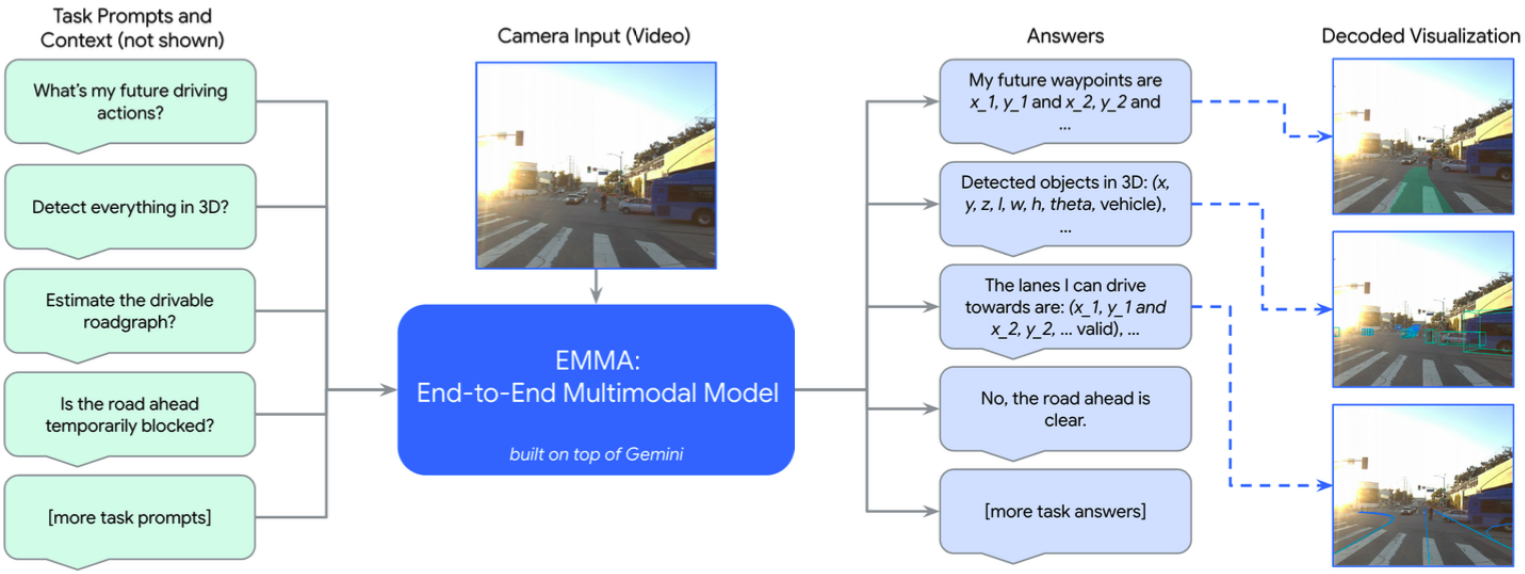


Figure 2: Illustration of EMMA Generalist. Starting with a task prompt (**left**), EMMA generates a corresponding textual prediction (**middle right**), which can then be decoded into a target output format, visualized and overlaid with the input image (**right**). EMMA Generalist is highly versatile, capable of performing a wide range of driving-related tasks, such as end-to-end motion planning, object detection, road graph estimation, and scene understanding Q&A. In the answer text, italicized words represent placeholders that will be dynamically substituted with actual values during task execution.

Task-specific prompts enable multi-task learning:

- Motion planning: "What's my future driving actions?"
- 3D detection: "Detect everything in 3D?"
- Road graph: "Estimate the drivable roadgraph?"

Spatial Reasoning: 3D Object Detection

Output format:

$$\mathcal{O}_{\text{boxes}} = \text{set}\{\text{text}(x, y, z, l, w, h, \theta, \text{cls})\} \tag{5}$$

Where (x, y, z) = center location, (l, w, h) = dimensions, θ = heading

Task formulation:

$$\mathcal{O}_{\text{boxes}} = \mathcal{G}(\mathbf{T}_{\text{detect\_3D}}, \mathbf{V}) \tag{6}$$

Example prompt: *"detect every object in 3D"*

Key finding: Sorting boxes by depth improves detection quality

Road Graph Estimation

Task: Predict drivable lanes as polyline waypoints

$$\mathcal{O}_{\text{roadgraph}} = \mathcal{G}(\mathbf{T}_{\text{estimate\_roadgraph}}, \mathbf{V}) \tag{7}$$

Text encoding format:

- Polylines: `(x1,y1 and... and xn,yn);...`
- Floating-point waypoints (2 decimal precision)
- Semicolon separates polyline instances

Design choices that improve quality:

- Dynamic sampling based on curvature (not fixed points)
- Ego-origin aligned sample intervals
- Shuffled ordering + padding to prevent early termination
- Language-like punctuation structure

Generalist Training Strategy

Training procedure:

- Multiple datasets: $\mathcal{D}_{\text{task}}$ with $|\mathcal{D}_{\text{task}}|$ examples
- Sampling probability $\propto$ dataset size: $|\mathcal{D}_{\text{task}}| / \sum_t |\mathcal{D}_t|$
- Total iterations: $e \times \sum_t |\mathcal{D}_t|$ for e epochs
- Task-specific prompts at inference

Benefits:

- Enhanced knowledge transfer between tasks
- Improved generalization
- Single model handles multiple capabilities
- Better efficiency than training separate models

Datasets Overview

Dataset	Hours	Examples
nuScenes (public)	6	18,686
WOMD (public)	572	487,061
WOD (public)	6	158,081
Internal Planning	**203,117 (355×)**	**24.4M (50×)**
Internal Detection	6,250	11.8M
Internal Roadgraph	64,135	8.3M

- Internal datasets **355× larger** than public benchmarks (hours)
- Enables comprehensive study of data scaling effects

Motion Planning: nuScenes Results

Method	L2 (m) 1s	L2 (m) 2s	L2 (m) 3s
UniAD (supervised)	0.42	0.64	0.91
DriveVLM (supervised)	0.18	0.34	0.68
OmniDrive (supervised)	0.14	0.29	0.55
Ego-MLP (self-supervised)	0.15	0.32	0.59
BEV-Planner (self-supervised)	0.16	0.32	0.57
EMMA (random init)	0.15	0.33	0.63
EMMA (Gemini init)	0.14	0.29	0.54
EMMA+ (pretrained)	**0.13**	**0.27**	**0.48**

- **State-of-the-art** on nuScenes planning benchmark
- 17.1% better than BEV-Planner (self-supervised)
- 12.1% better than OmniDrive (with human labels)

Motion Planning: WOMD Results

Method	L2 (m) 1s	L2 (m) 3s	L2 (m) 5s
MotionLM*	0.045	0.266	0.696
Wayformer*	0.046	0.252	0.628
EMMA† (PaLI)	0.034	0.274	0.797
EMMA+† (PaLI)	0.031	0.239	0.680
EMMA	0.032	0.248	0.681
EMMA (w/ CoT)	0.030	0.241	0.664
EMMA+	0.030	0.225	0.610
EMMA+ (w/ CoT)	**0.027**	**0.203**	**0.543**

- **13.5% improvement** over baselines at 5s horizon
- Works with different MLLMs (Gemini, PaLI)
- Chain-of-thought brings consistent gains

Chain-of-Thought Ablation Study

Scene Desc.	Critical Obj.	Meta Decision	Behavior Desc.	Improvement
✓	✗	✗	✗	+0.0%
✗	✓	✗	✗	+1.5%
✗	✗	✓	✗	+3.0%
✗	✓	✓	✗	+5.7%
✗	✓	✓	✓	**+6.7%**

Key findings:

- Meta decision + critical objects are the main contributors.
- Combined rationale yields a **6.7% improvement**.
- Scene description is neutral on quality but improves explainability.

Data Scaling Results

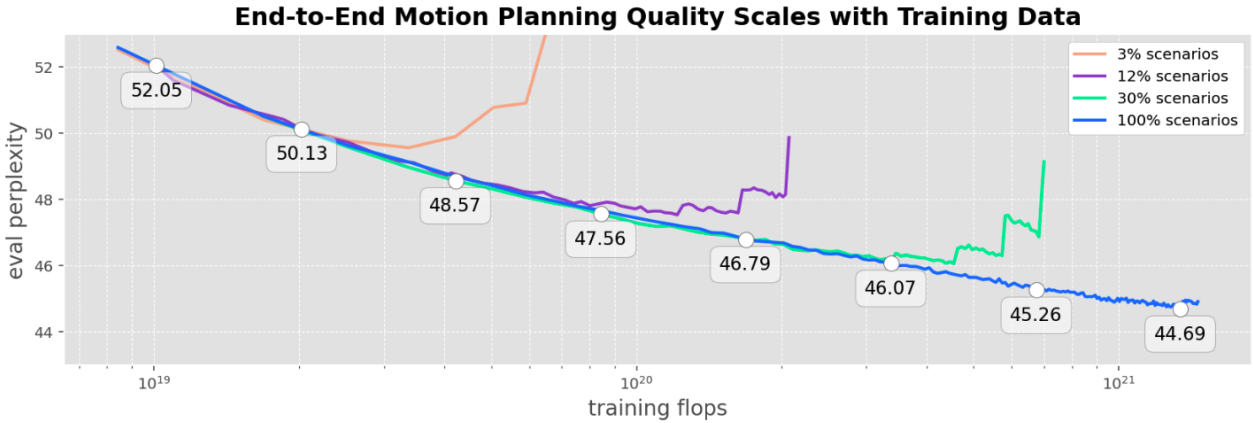


Figure 4: EMMA data scaling experiments on our mega-scale internal dataset. Each curve represents the eval perplexity for end-to-end motion planning as training proceeds with more steps. The x-axis is training compute, measured by floating-point operations (FLOPs) in log scale. The same EMMA model is trained on four sizes of datasets that are sampled with different percentages from 3% to 100% (denoted by different colors). In general, EMMA tends to achieve better quality until overfitting when given more training compute, but it overfits quickly on smaller datasets. We observe the driving quality has not saturated when using the full large-scale dataset.

- Trained on 3%, 12%, 30%, 100% of internal dataset
- Larger datasets consistently achieve lower eval perplexity
- **Key insight:** Performance has **not saturated** even with full mega-scale dataset
- Suggests further gains possible with more data

3D Object Detection: WOD Results

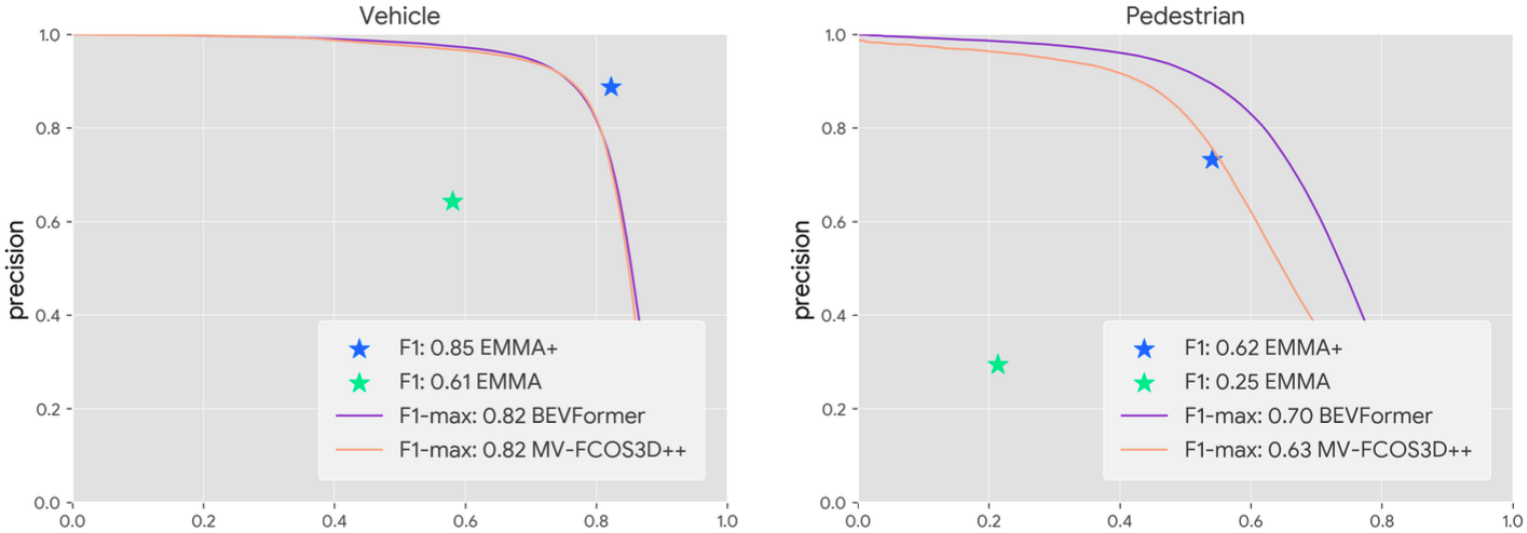


Figure 5: Camera-primary 3D object detection experiments on WOD (Sun et al., 2020) using the standard LET matching (Hung et al., 2024). EMMA+ achieves competitive performance on the detection benchmark in both precision/recall and F1-score metrics. Compared to state-of-the-art methods, it achieves 16.3% relative improvements in vehicle precision at the same recall or 5.5% recall improvement at the same precision.

Camera-primary 3D detection (LET matching):

- **Vehicle:** 16.3% better precision at same recall vs. BEVFormer
- **Pedestrian:** Competitive with MV-FCOS3D++
- Strong performance especially in near range

Road Graph Estimation: Design Ablation

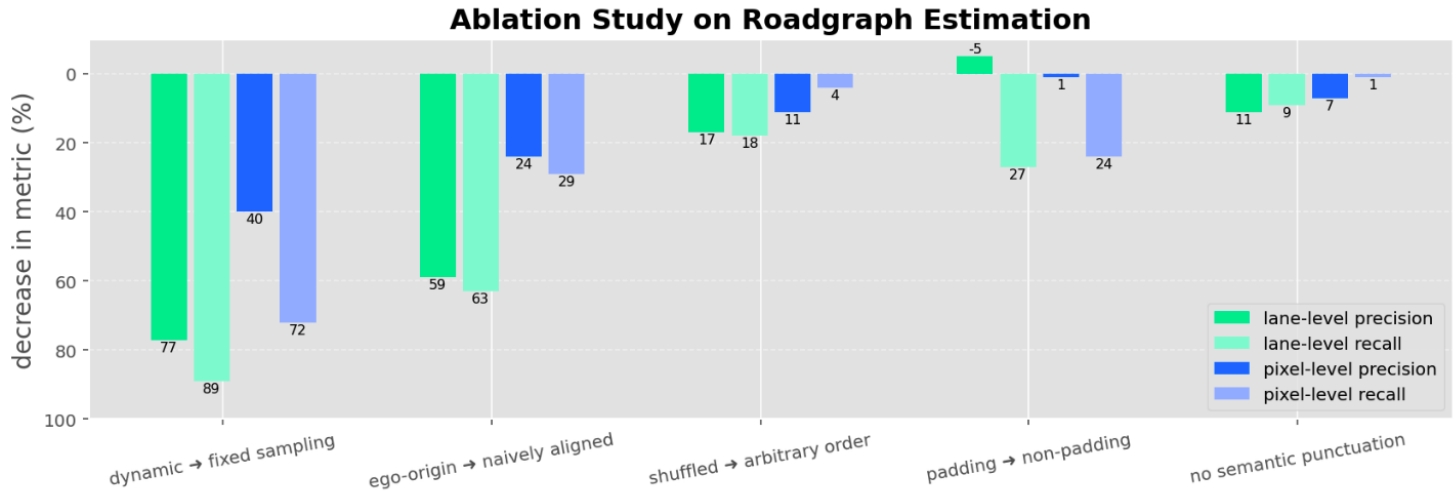


Figure 6: Ablation study on road graph estimation. To evaluate the influence of different components in our road graph estimation model, we ablate each configuration and measure the corresponding impact on quality. Dynamic sampling (**leftmost**) of road graph polylines based on lane curvature and length proves to be the most significant factor, leading to a substantial 70% to 90% change in lane-level precision and recall. In contrast, aligning the model with a language-like representation, *i.e.*, semantic punctuation (**rightmost**), has a more modest effect, contributing to only <10% change in precision and recall of any metric.

Impact of design choices (decrease in metric %):

- Dynamic sampling (vs fixed): **40-90% degradation**
- Ego-origin aligned: 25-60% degradation
- Shuffled ordering: 20-40% degradation
- Padding: 25% degradation

Generalist Co-Training Results

Tasks Trained			Relative Improvement		
Plan	**Det**	**Road**	**Plan**	**Det**	**Road**
✓	✓	-	+1.6%	+2.4%	-
✓	-	✓	+1.4%	-	+5.6%
-	✓	✓	-	+2.4%	+3.5%
✓	✓	✓	**+1.4%**	**+5.5%**	**+2.4%**

Key findings:

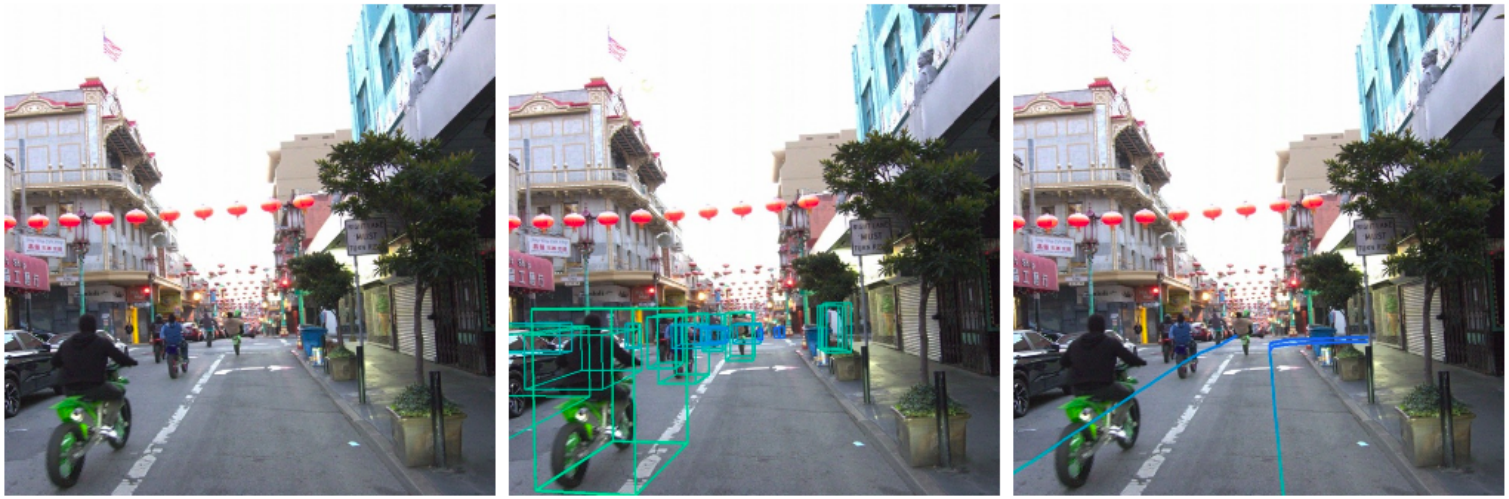
- Co-training **improves** all individual tasks
- Up to **5.5% improvement** for detection
- Demonstrates effective knowledge transfer
- Planning task provides strong complementary signal

Qualitative Results: Success Cases

(a) A garbage bag appears on the freeway, so our predicted trajectory suggests to nudge slightly to the right to avoid it.

(b) A ladder appears on the freeway, and our predicted trajectory suggests to switch to the left lane to bypass it appropriately.

Qualitative Results: More Examples



(k) A fleet of fast-moving motorcyclists pass by. The predicted trajectory suggests pausing to allow them to pass safely. Notably, motorcyclists are accurately identified by our model (middle).

(l) A motorbike is moving on a narrow lane at night, and yields to the right. Our predicted trajectory

Key Innovations

1. **MLLM as first-class citizen**
   - Leverage Gemini's world knowledge and reasoning
   - Unified language representation for all tasks
2. **Self-supervised + camera-only + HD map free**
   - Scalable data generation pipeline
   - Reduced sensor and mapping dependencies
3. **Chain-of-thought reasoning**
   - Improves performance (+6.7%)
   - Enhances explainability
4. **Generalist model capabilities**
   - Single model for multiple tasks
   - Co-training improves individual task performance

Limitations and Future Directions

Current limitations:

- **3D spatial reasoning:** Camera-only, no LiDAR/radar fusion
- **Closed-loop evaluation:** Requires expensive sensor simulation
- **Computational cost:** Higher inference latency than specialized models
- **Memory and video:** Limited to 4 frames, lacks long-term memory

Future research directions:

- Integrate 3D sensing modalities (LiDAR, radar)
- Extend to longer video sequences and memory modules
- Develop efficient inference and model optimization techniques
- Improve verification of predicted driving signals

Key Takeaways

Main Contributions

1. **EMMA**: End-to-end multimodal model built on Gemini for autonomous driving
2. **State-of-the-art planning** on nuScenes, competitive on WOMD
3. **Competitive 3D detection** on camera-primary WOD benchmark
4. **Generalist model** that improves across multiple tasks via co-training

Impact

Demonstrates that multimodal LLMs can serve as powerful foundation models for autonomous driving, opening new research directions for:

- Scalable end-to-end learning
- Explainable decision-making
- Open-world generalization

Thank you for your attention!

Questions?

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan, *Emma: End-to-end multimodal model for autonomous driving*, 2025.