DriveLM

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, Hongyang Li

ECCV 2024 Oral

→ VLGA

Jin Yao, Dhruva Kurra, Tom Lampo, Zezhou Cheng, Danhua Guo, Burhan Yaman

06/26 arXiv preprint

Presenters: Soongeol, Akshay, Jim

Georgia Tech

Outline

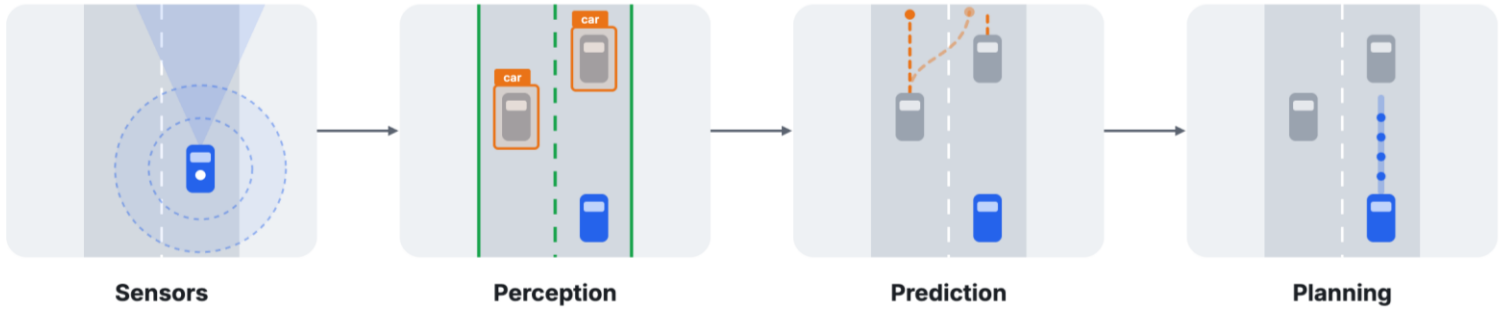
- DriveLM
  - Overview
  - Task
  - model
  - Metrics and experiments
- VLGA

Georgia Tech

Introduction: Typical self-driving stacks

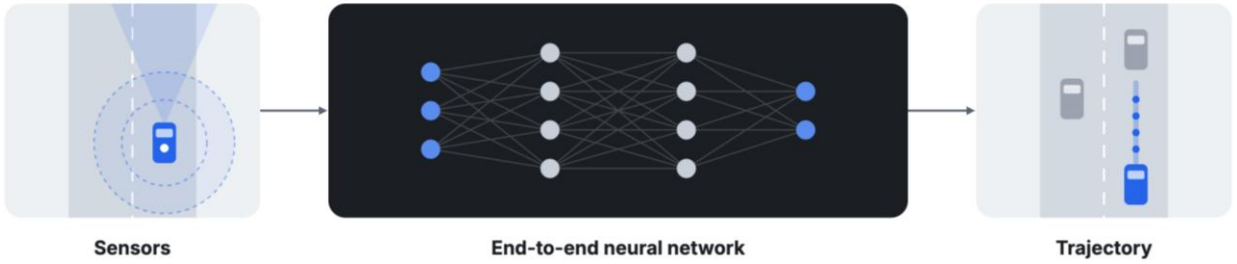
Modular stack

Specialized models for perception, prediction and planning, chained in sequence



End-to-end stack

One network maps raw sensor input directly to a future trajectory



Limitations

- ❌ generalization to unseen scenarios
- ⚠️ human interpretability
- ⚠️ generalization to a new sensor suite

Georgia Tech

Introduction: MLLMs for self-driving

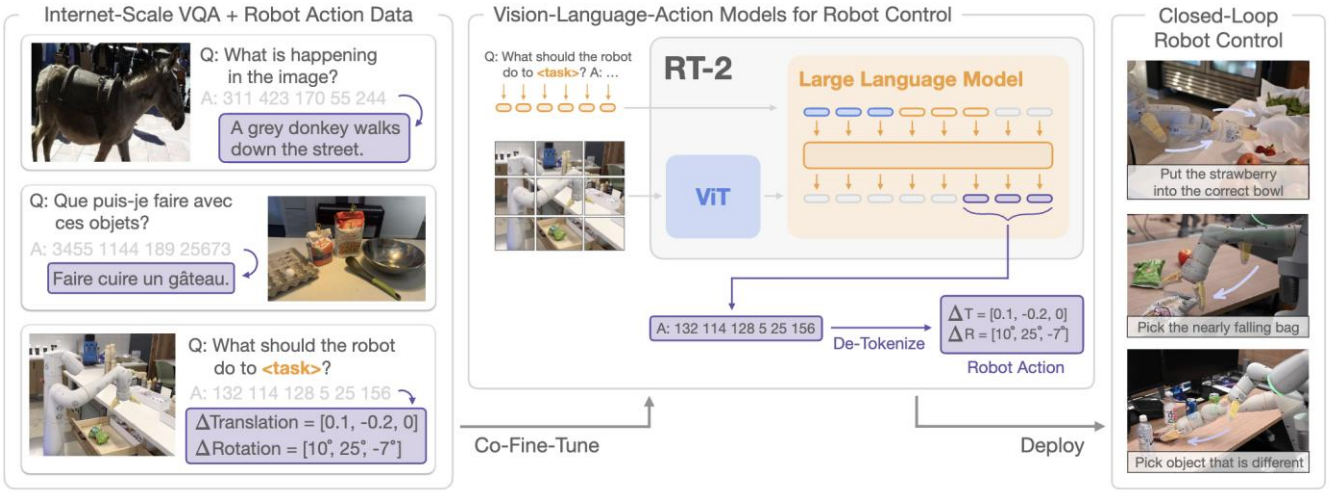
MLLMs exhibit good commonsense reasoning!

This can address some limitations of a self-driving stack

- Generalization to unseen scenarios
  - Commonsense reasoning!
- Generalization to a new sensor suite
  - Trained on large scale image / video datasets, robust to viewpoint changes
- Human interpretability
  - Language modality – talk to your self-driving car!

Related prior work

VLMs for robotics / embodied reasoning

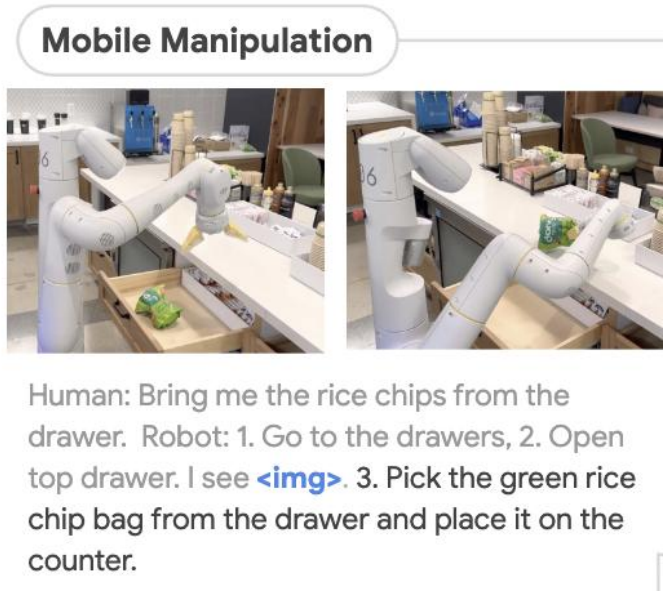


Google DeepMind

https://robotics-transformer2.github.io

2023-8-1

RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

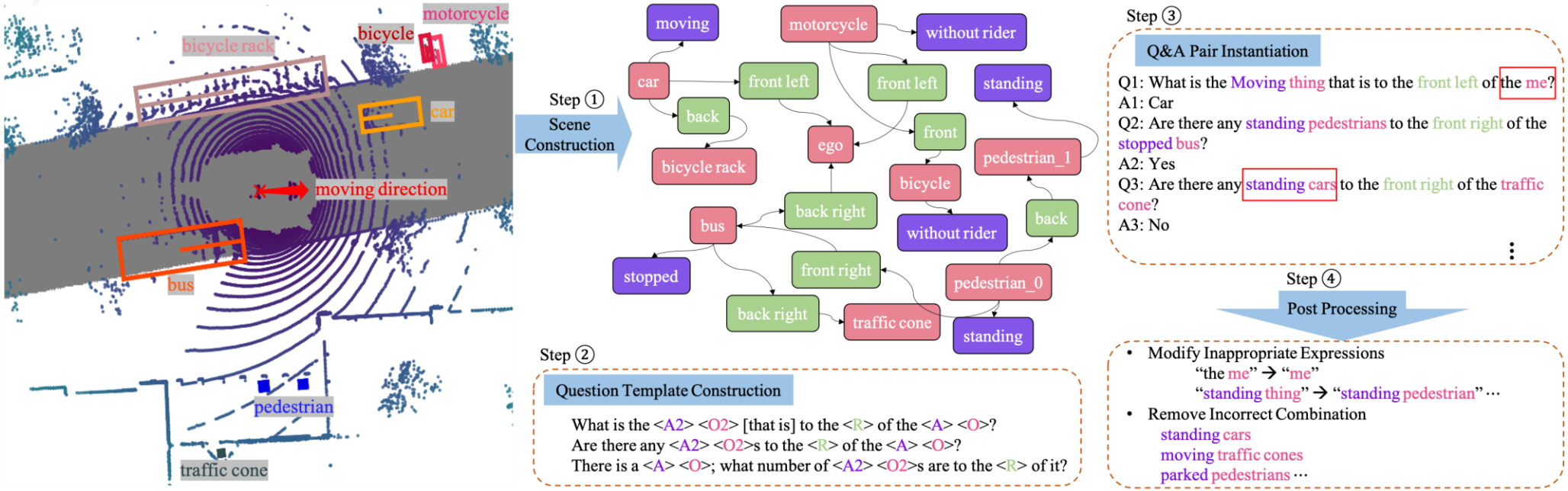


PaLM-E: An Embodied Multimodal Language Model
Driess et al, 2023

Georgia Tech

Related prior work

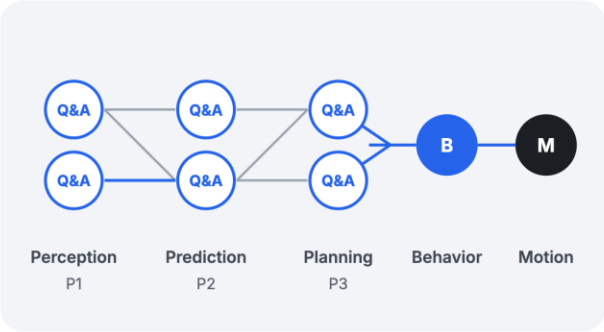
VLMs for self-driving VQA



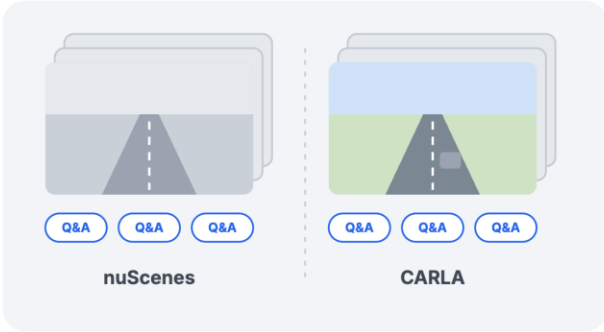
NuScenes-QA

Georgia Tech

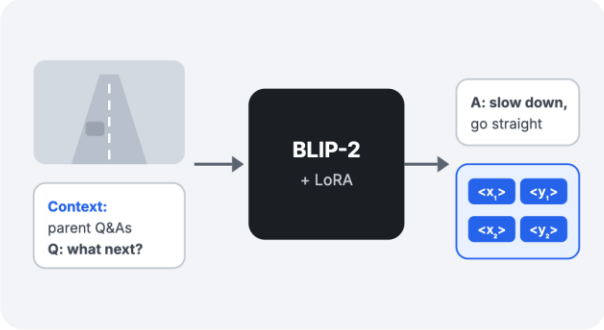
DriveLM overview



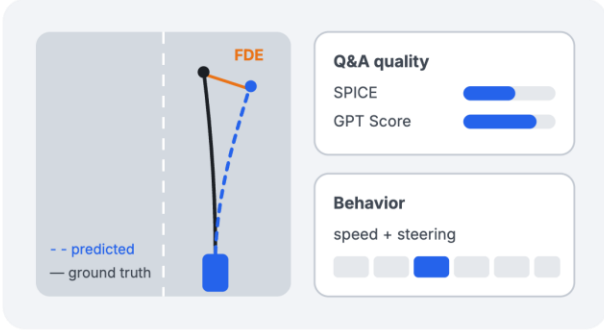
01 Task: Graph VQA



02 Data: DriveLM-Data



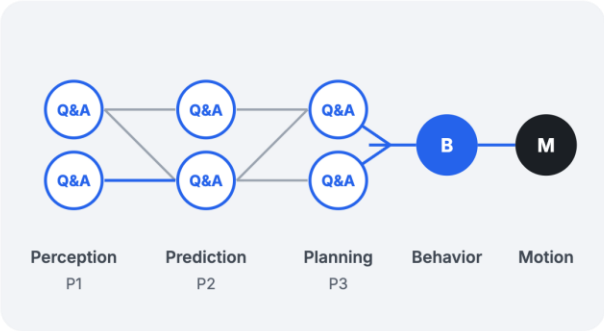
03 Model: DriveLM-Agent



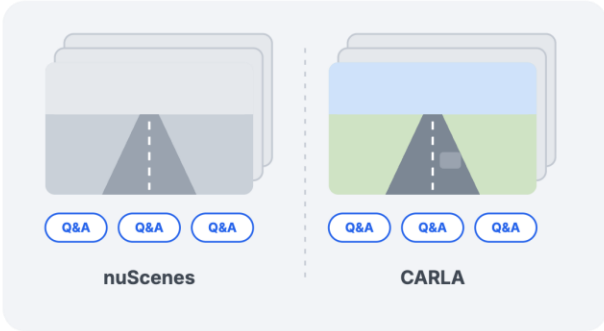
04 Metrics

Georgia Tech

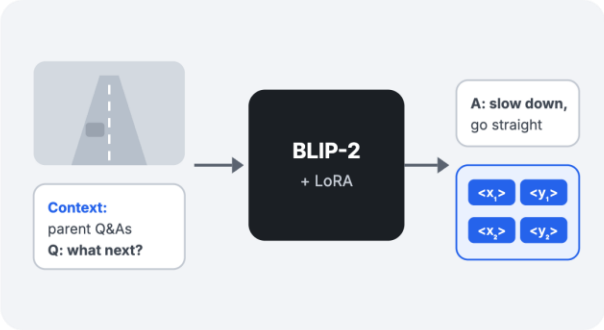
DriveLM overview



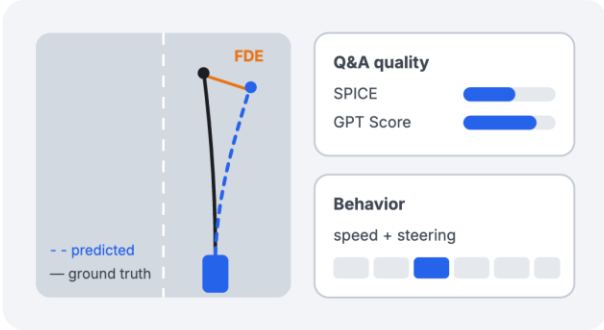
01 **Task: Graph VQA**



02 **Data: DriveLM-Data**



03 **Model: DriveLM-Agent**



04 **Metrics**

Georgia Tech

DriveLM task: Graph-structured VQA

Goal: A sequential reasoning structure that aids the driving task

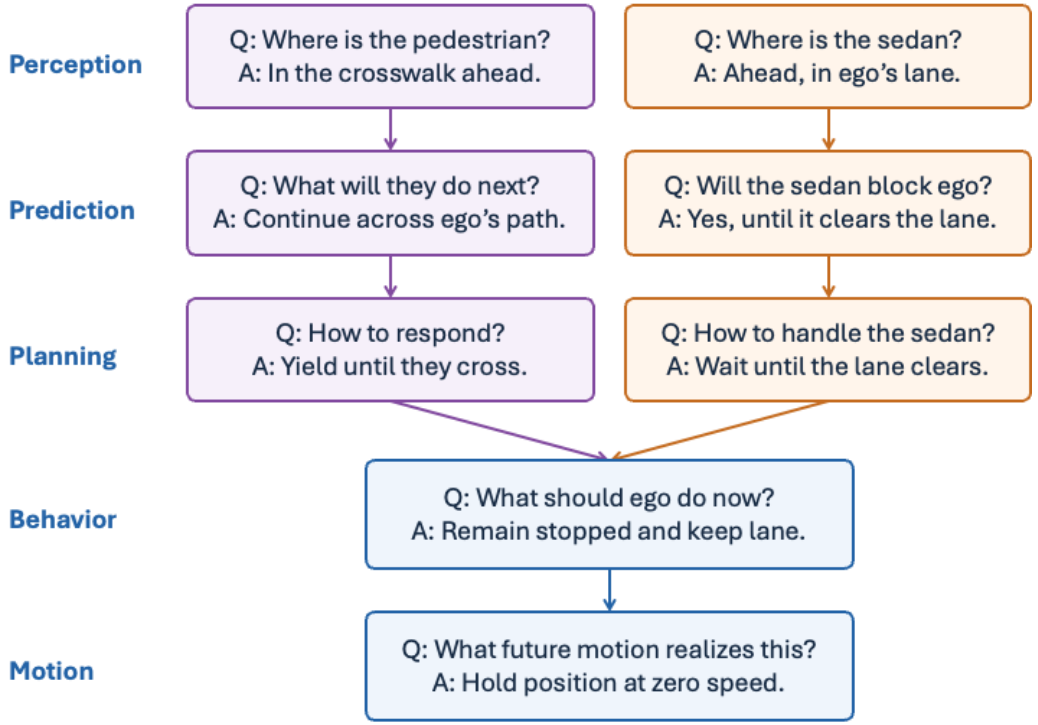
- Mimic human-like reasoning using DA-graph structure

Scenario



Ego is stopped before the crossing.
The pedestrian and sedan both
occupy the intended path.

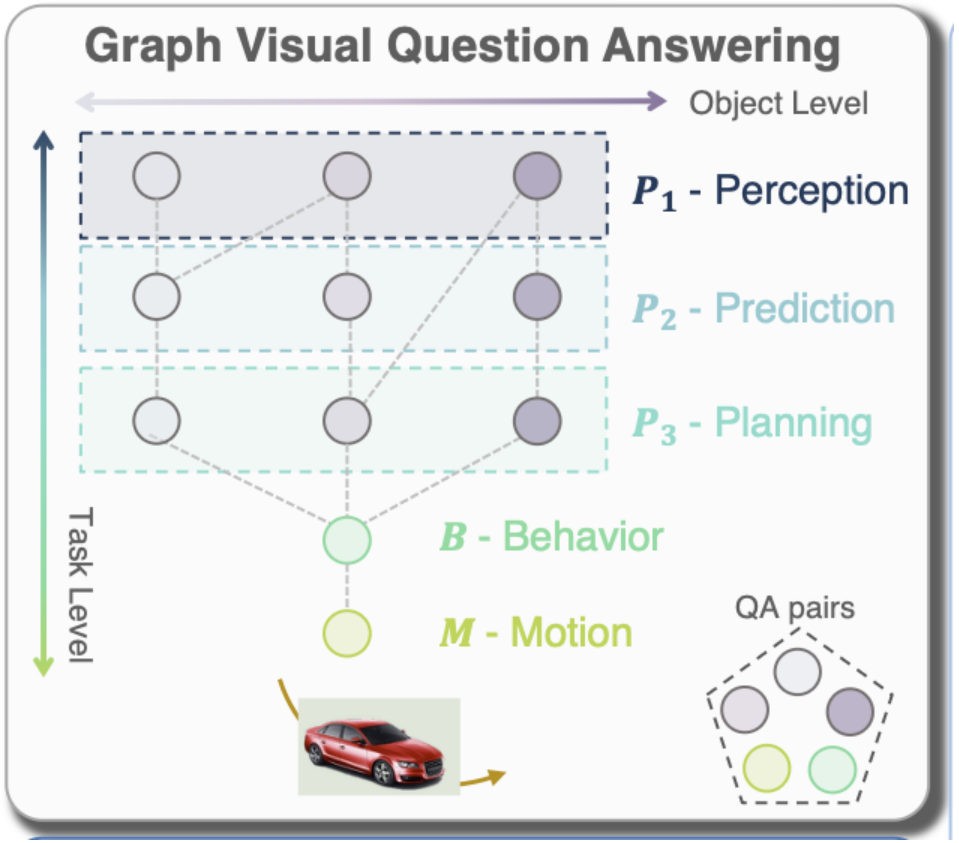
QA nodes and their context dependencies



Node – QA pair
Edge – context dependency

Georgia Tech

Graph VQA connects dependent questions

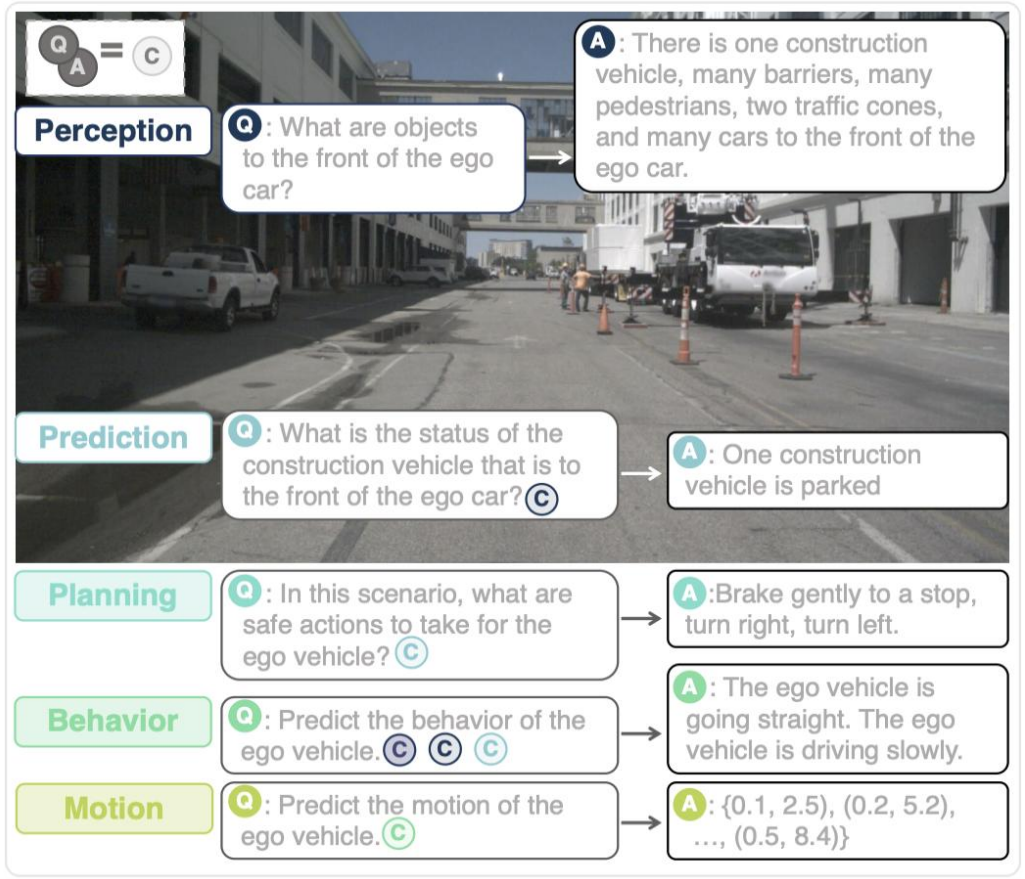


- Two parent QAs
  - The pedestrian may cross
  - The Sedan may enter our path
- One planning Question
  - What should the ego vehicle do considering both objects

- Node : a question & answer pair
- Edge : a logical dependency
- Object level & Task level Link
- For each question, the model recieves the connected parent node (Question and answer as context)

Georgia Tech

Perception, Prediction and Planning



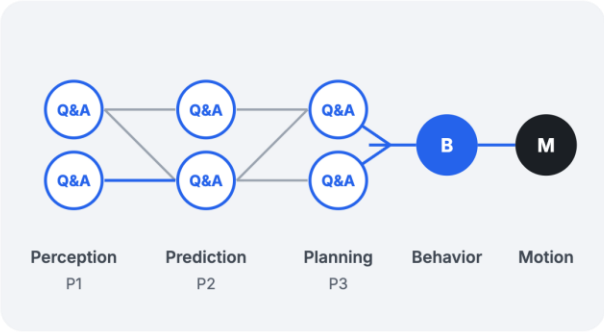
- Perception: identifies objects
  - Identify the construction vehicle
- Prediction : estimate state and movement
  - Estimate its state: parked
- Planning : coniser safe response
  - Brake gently to a stop

Georgia Tech

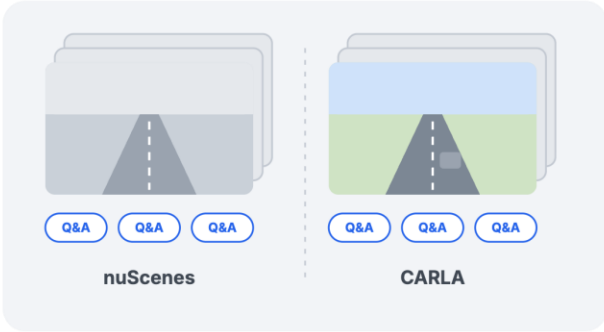
Example – with DriveLM agent inference



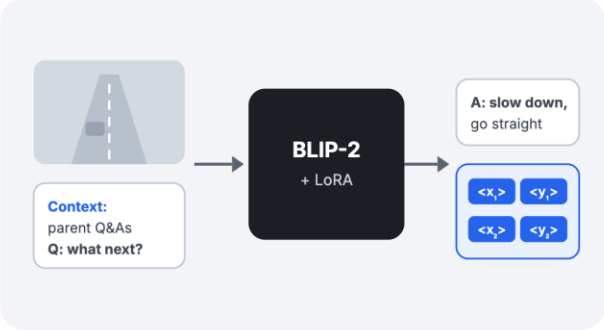
DriveLM overview



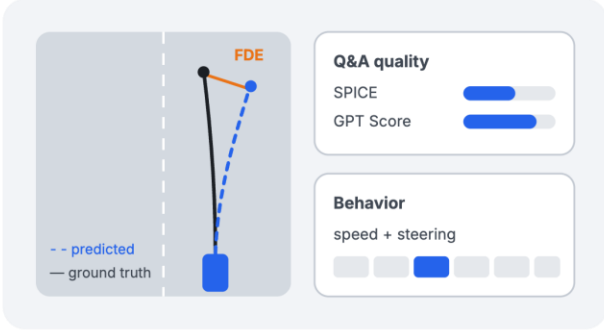
01 Task: Graph VQA



02 Data: DriveLM-Data



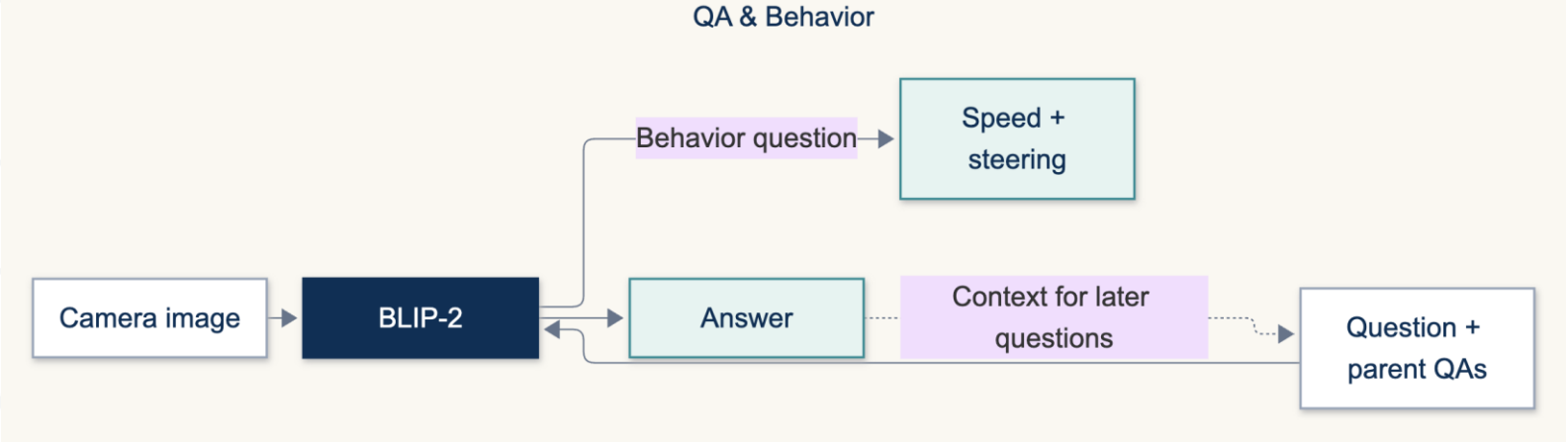
03 Model: DriveLM-Agent

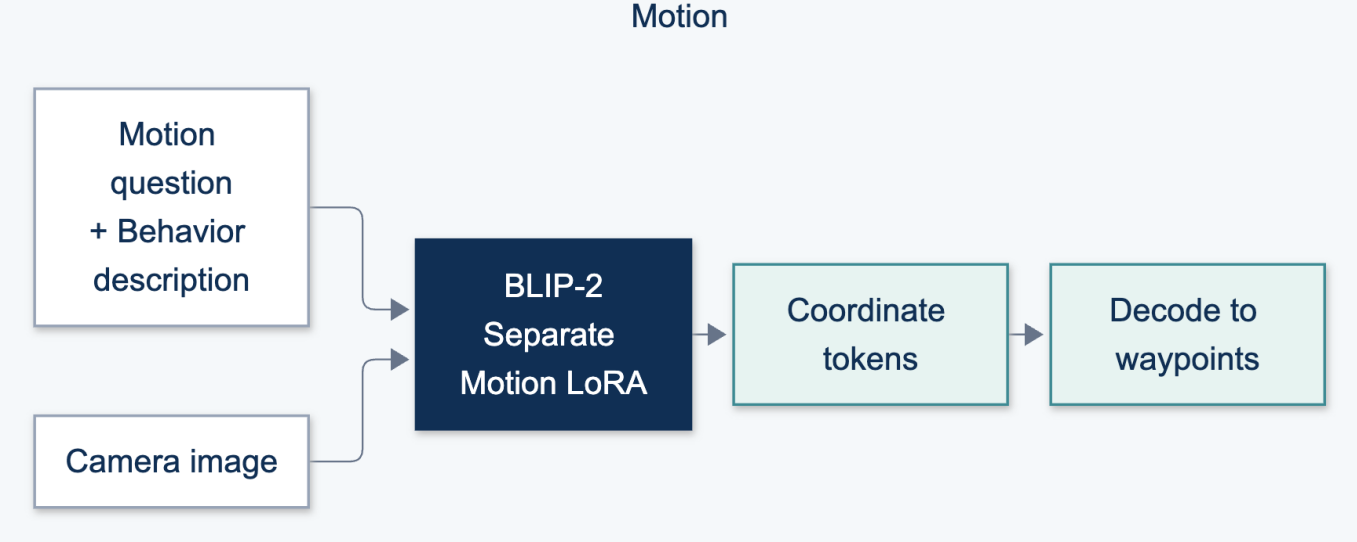


04 Metrics

Georgia Tech

From Visual Question to Driving Actions





Blip-2 takes camera image & motion question + behavior description

Then outputs coordinate tokens which will decoded in waypoints

Georgia Tech

Parent QAs become the next question context

Question with pred-context: **Perception Q:** What are objects to the front of the ego car? **Perception A:**There are many **cars**, one **truck** and one **pedestrian** to the front of ego car. **Q:** What is the status of the **cars** that are to the front of the ego car?

Answer: Many **cars** are moving and one is parked

GT: Three **cars** are moving.

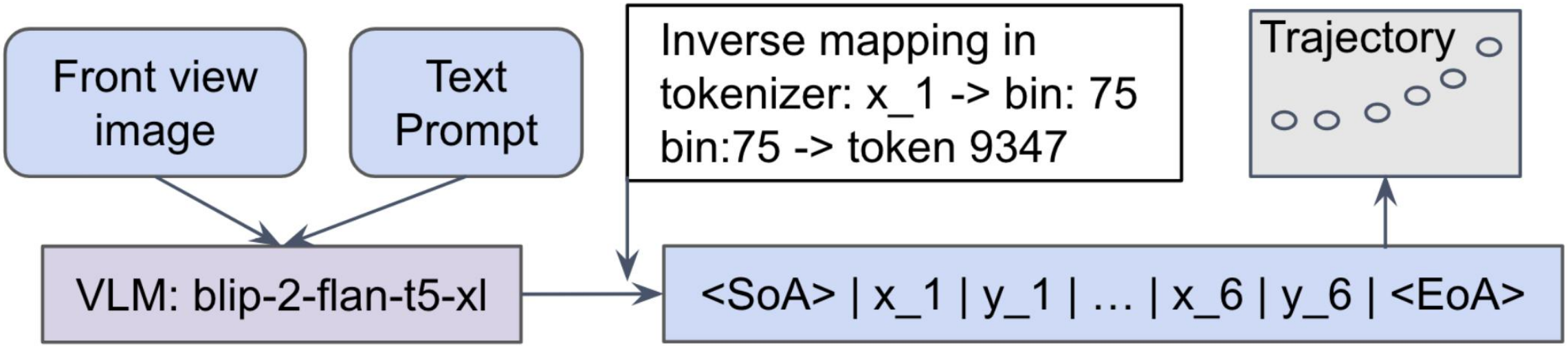
SPICE: 0.667

GPT score: 40

- Read the input
  - Parent perception QA
  - Current question about the truck
- Read the output
  - Prediction : moving
  - Reference: parked
- Blip-2 receives
  - Image
  - Child question
  - Parent QAs

Georgia Tech

Behavior connects QA reasoning to motion



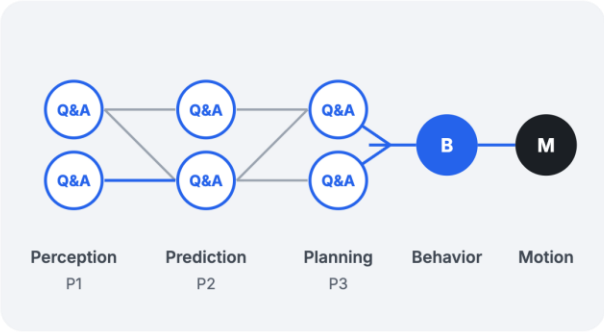
- Behavior
  - 5 speed, 5 steering categories
- Motion
  - Future ground-plane (x,y) waypoints

Each coordinate uses one of 256 bins
Generated tokens map back to coordinates

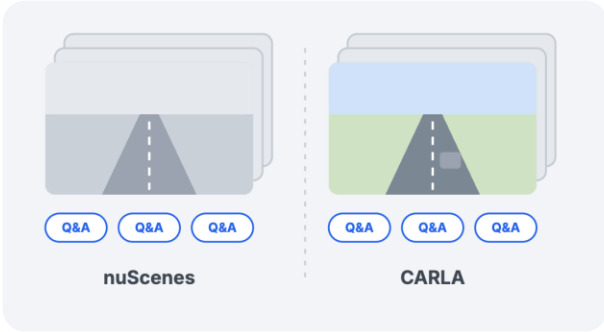
Motion uses the image and behavior with sparately trained LoRA weights

Georgia Tech

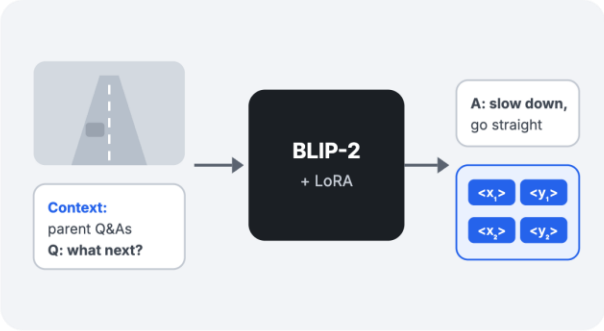
DriveLM overview



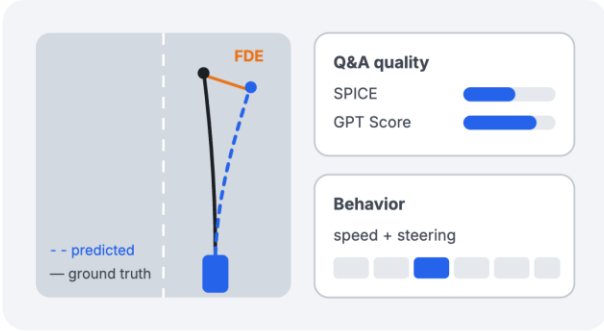
01 Task: Graph VQA



02 Data: DriveLM-Data



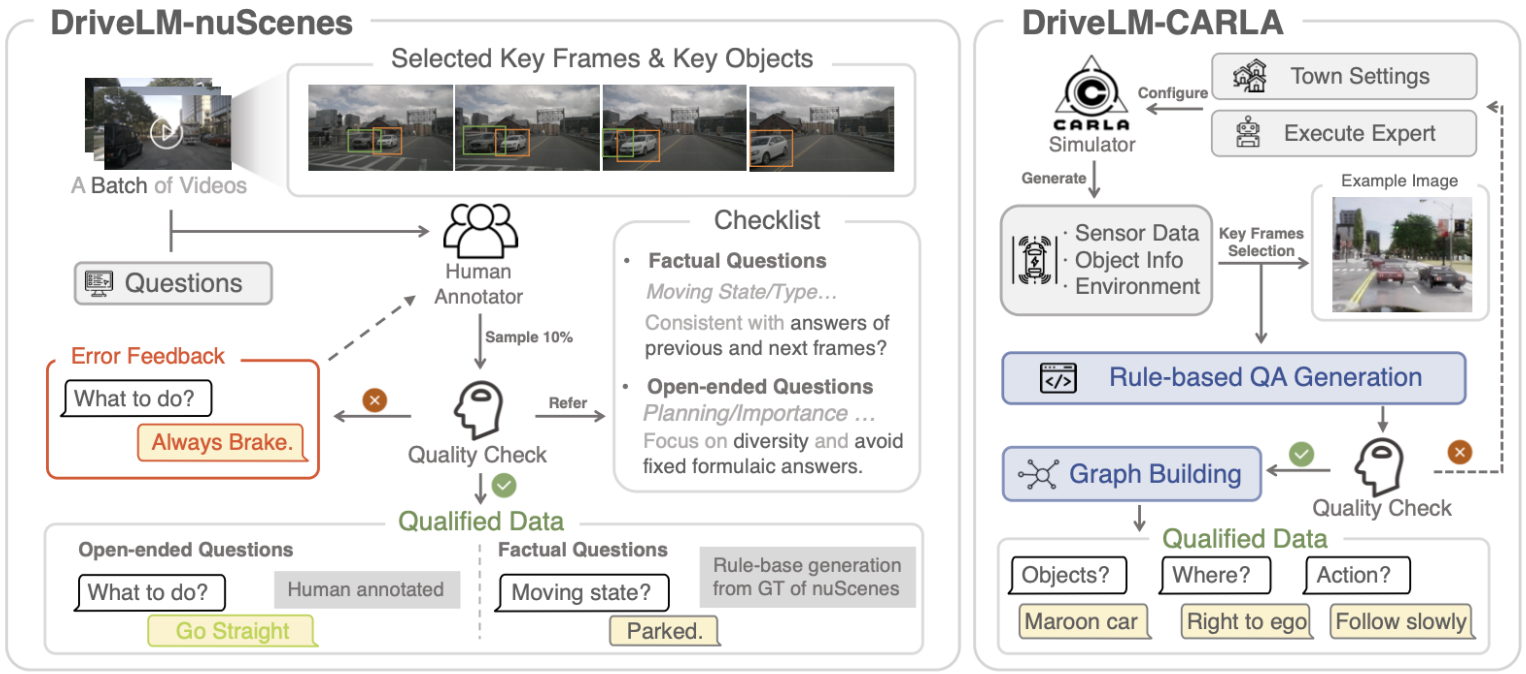
03 Model: DriveLM-Agent



04 Metrics

Georgia Tech

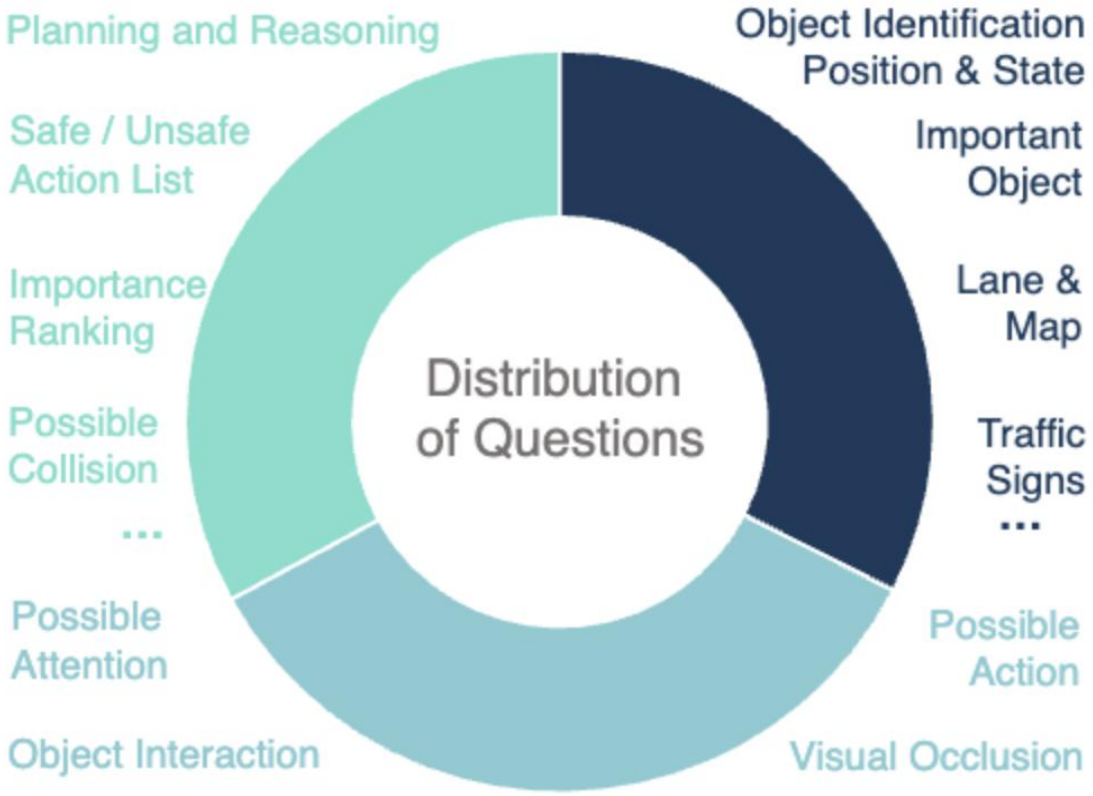
How DriveLM builds QA annotations



- Each sample connects
  - Driving image
  - QA targes
  - Dependencies between QAs

Georgia Tech

DriveLM Data annotations: Question distribution



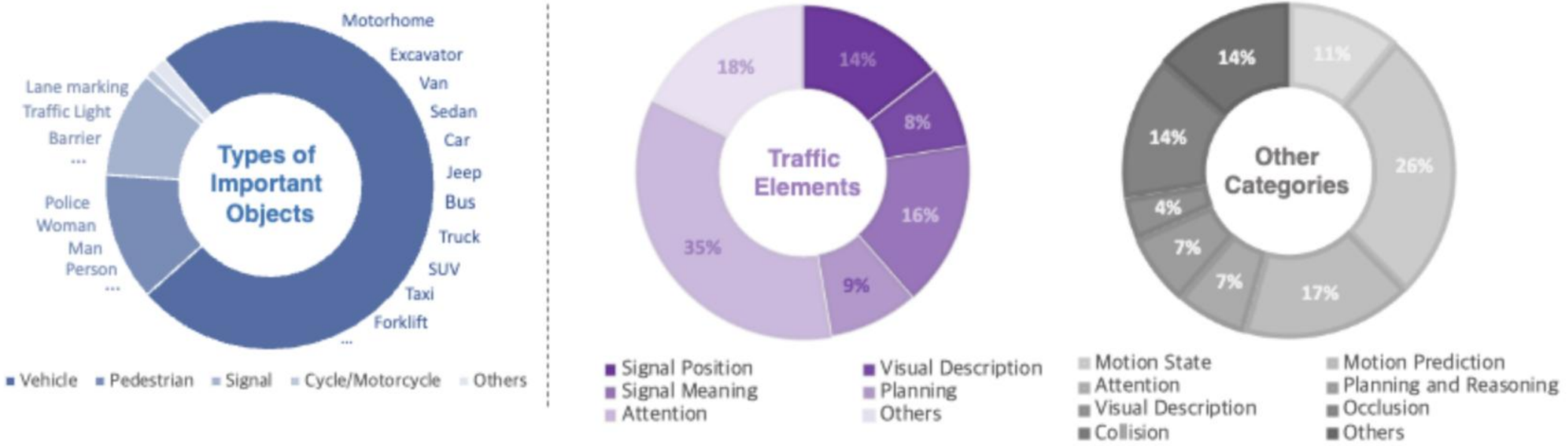
- Perception:
  - Understanding the current driving environment.
- Prediction:
  - Anticipating possible actions and interactions.
- Planning:
  - Reasoning about safe driving decisions.

Most of questions are annotated by human annotators

Answering this question servers as an indirect measure of human like driving reasoning.

Georgia Tech

DriveLM-nuScene Object and Question distribution



- The distribution of key objects (LEFT)
  - Extracted from visual description
- The distribution of question type related to different key Objects (Right)

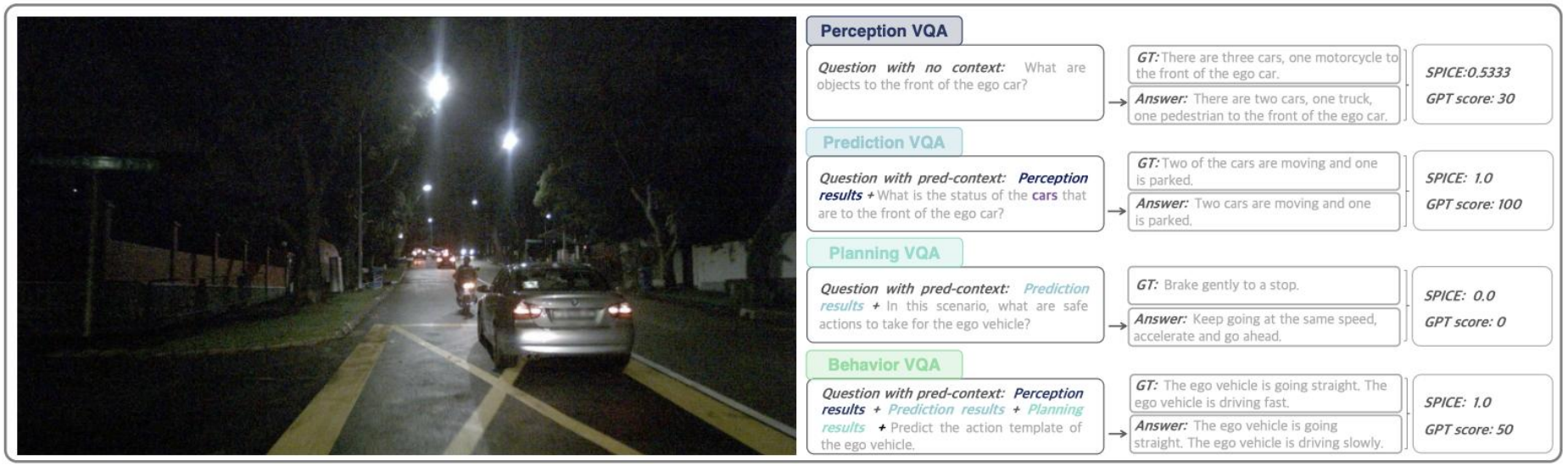
Example QA

- Question: "Which object should the ego vehicle pay attention to?"
- Answer: "The red car approaching from the left."

- Object Category
  - Vehicle → Car
- Question Type
  - Attention

Georgia Tech

Two datasets: Real Scenes & Simulated



DriveLM – nuScenes

Real scenes

4,871 Keyframes

4,072 train / validation
91.4 QAs per frame
Human annotations + existing labels

Perception

Question: What are important objects to the front of the ego car?

Answer: The important objects in the scene are the maroon car to the front left of the ego vehicle and the maroon van to the front of the ego vehicle.

GT: The important objects in the scene are the maroon car to the front left of the ego vehicle and the olive van to the front of the ego vehicle.

Scene Image



DriveLM – CarLA

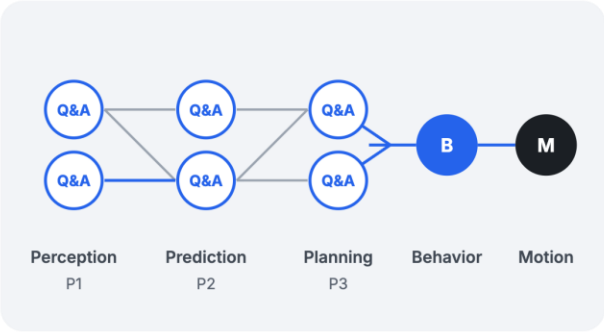
Simulated

64,285 frames

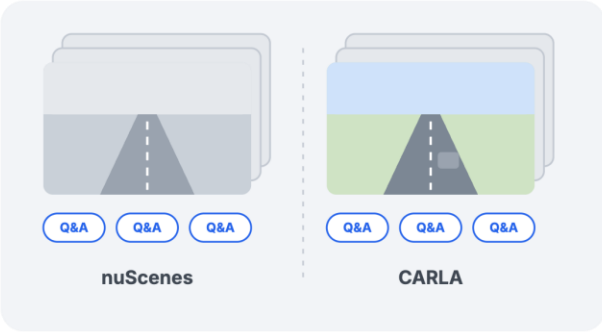
5,721 key frames within the full set
24.4QAs per frame in the full dataset

Georgia Tech

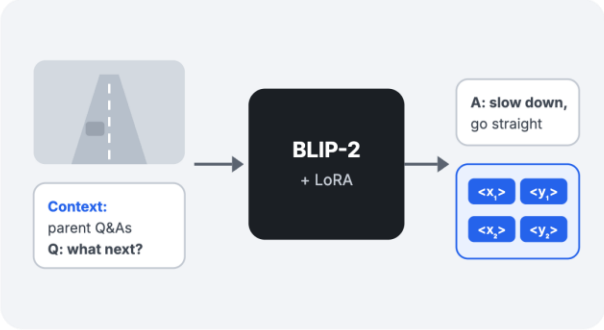
DriveLM overview



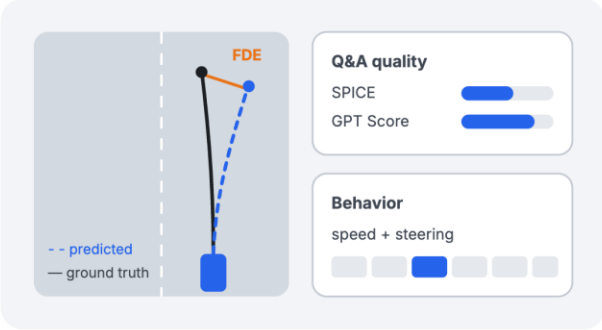
01 **Task: Graph VQA**



02 **Data: DriveLM-Data**

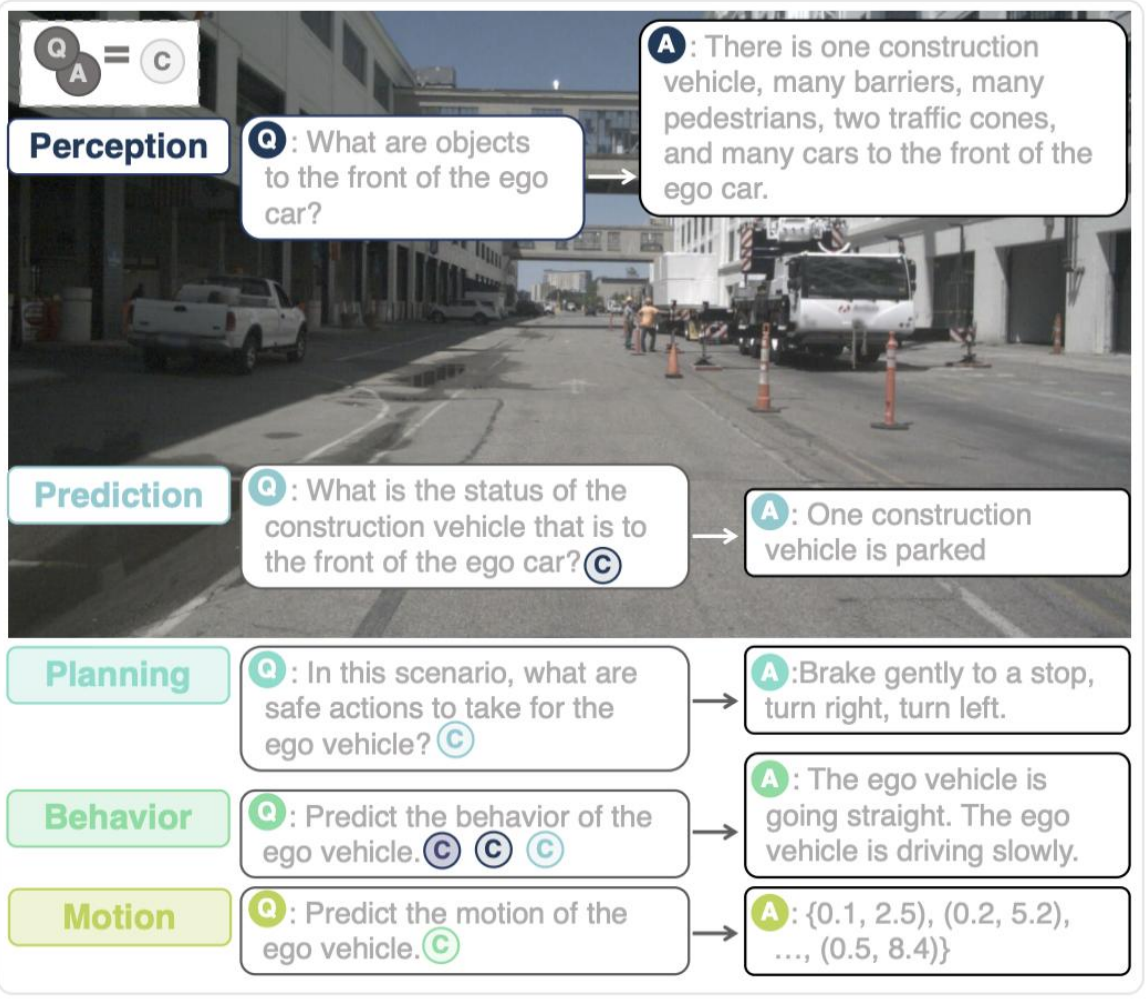


03 **Model: DriveLM-Agent**



04 **Metrics**

Metrics: answers, behavior and motion



- QA answers
  - SPICE/ GPT score :Agreement with reference answers
  - Behavior accuracy: speed and steering

SPICE : similarity of predicted vs GT texts (ignoring semactic meaning)
GPT score: question, ground truth and predicted answer scored by GPT

Motion: waypoint error and collision rate

$$\text{ADE} = \frac{1}{T}\sum_{t=1}^{T} \|\hat{p}_t - p_t\|_2$$

$$\text{FDE} = \|\hat{p}_T - p_T\|_2$$

Average Displacement Error (ADE): averages waypoint error;
Final Displacement Error (FDE): final point error.

Georgia Tech

Experiment & Results:Qualitative results on Waymo



Perception VQA

Question with no context: What are objects to the front of the ego car?

→ *Answer:* There are two cars, one truck, one pedestrian to the front of the ego car.

Prediction VQA

Question with pred-context: ***Perception results*** + What is the status of the **cars** that are to the front of the ego car?

→ *Answer:* Two cars are parked and one is moving.

Planning VQA

Question with pred-context: *Prediction results* + In this scenario, what are safe actions to take for the ego vehicle?

→ *Answer:* Keep going at the same speed, accelerate and go ahead.

Behavior VQA

Question with pred-context: ***Perception results*** + *Prediction results* + *Planning results* + Predict the action template of the ego vehicle.

→ *Answer:* "The ego vehicle is going straight. The ego vehicle is driving fast..

Perception VQA

Question with no context: What are objects to the front of the ego car?

→ *Answer:* There are many cars, one pedestrian to the front of the ego car。

Prediction VQA

Question with pred-context: ***Perception results*** + What is the status of the **pedestrians** that are to the front of the ego car?

→ *Answer:* One pedestrian is moving and one is standing.

Planning VQA

Question with pred-context: *Prediction results* + In this scenario, what are safe actions to take for the ego vehicle?

→ *Answer:* Keep going at the same speed, decelerate gradually without braking..

Behavior VQA

Question with pred-context: ***Perception results*** + *Prediction results* + *Planning results* + Predict the action template of the ego vehicle.

→ *Answer:* The ego vehicle is going straight. The ego vehicle is driving slowly.



- Show casing strong generalization capability.
- Zero shot generalization from nuScene to Waymo

Georgia Tech

Experiment & Results:Failure cases

Perception

Question: Is there a person in the scene?

Answer: Yes, there is a person in the scene.

Prediction

Context: Yes, there is a person in the scene. ***Question:*** What should the ego vehicle do based on the pedestrian that is crossing the road?

Answer: The ego vehicle should stop and give the pedestrian the right of way.

Behavior

Context: The ego vehicle should stop and give the pedestrian the right of way. ***Question:*** Predict the behavior of the ego vehicle.

Answer: The ego vehicle is going straight. The ego vehicle is driving slowly.

Scene Image



Perception

Question: Is there a person in the scene?

Answer: No, there is no person in the scene.

Prediction

Context: No, there is no person in the scene. ***Question:*** What should the ego vehicle do?

Answer: The ego vehicle should follow the road to the next junction.

Behavior

Context: The ego vehicle should follow the road to the next junction. ***Question:*** Predict the behavior of the ego vehicle.

Answer: The ego vehicle is going straight. The ego vehicle is driving fast.

Scene Image



- **Scene 1 — Behavior error.**

The context explicitly says to stop and yield to the pedestrian

but the model predicts that the vehicle will continue straight slowly.

- **Scene 2 — Perception error propagation.**

The model failed to detect the pedestrian and concludes that the vehicle should follow the road.

This error carries over to the behavior prediction: the vehicle will continue straight at high speed.

Georgia Tech

Experiment & Results: Driving performance and zero shot transfer

Model	nuScenes	Waymo transfer
UniAD-Single	1.80	4.16
BLIP-RT-2	2.63	2.78
DriveLM-Agent: None	1.39	2.76
DriveLM-Agent: Graph	1.74	2.63

Motion ADE(m), lower is better

None: behavior then Motion
Graph: p1-p3 QAs & inform behavior

Graph helps on waymo
None has lower error on nuScenes

Georgia Tech

Experiment & Results: Fine tuning and graph context improve QA

SPICE, higher is better. Compare the base BLIP-2 with the fine-tuned Agent.

Dataset	Model	None	Graph	GT
nuScenes	BLIP-2	4.34	7.71	8.19
nuScenes	DriveLM-Agent	42.56	49.54	50.29
CARLA	BLIP-2	10.46	10.30	16.18
CARLA	DriveLM-Agent	72.71	75.26	79.07

NuScenes fine-tuning: 4.34 → 42.56 (None context)
Additional graph-context gain: 42.56 → 49.54 (DriveLM-Agent)

Graph uses predicted parent QAs.
GT uses reference parent QAs at test time.

Georgia Tech

Experiment & Results: Which QA stages help the behavior decision

ID	Perception			Prediction			Planning			Behavior		
	P_{1-1}	P_{1-2}	P_{1-3}	P_{2-1}	P_{2-2}	P_{2-3}	P_{3-1}	P_{3-2}	P_{3-3}	Acc. ↑	Speed ↑	Steer ↑
1	✓	-	-	-	-	-	-	-	-	54.69	66.83	75.22
2	✓	✓	-	-	-	-	-	-	-	55.32	67.33	74.34
3	✓	✓	✓	-	-	-	-	-	-	53.94	65.33	75.72
4	✓	✓	✓	✓	-	-	-	-	-	58.82	71.83	**80.98**
5	✓	✓	✓	✓	✓	-	-	-	-	57.07	71.96	78.97
6	✓	✓	✓	✓	✓	✓	-	-	-	57.70	72.22	79.22
7	✓	✓	✓	✓	✓	✓	✓	-	-	**58.95**	72.59	80.23
8	✓	✓	✓	✓	✓	✓	✓	✓	-	57.95	**72.97**	79.97
9	✓	✓	✓	✓	✓	✓	✓	✓	✓	57.49	69.89	80.63

Train and evaluate with cumulative QA subsets

- Predction adds Usefull context
- Planning adds little in this comparison

Georgia Tech

Better QA does not guarantee better driving

Metric	None	Graph	Observed change
nuScenes QA SPICE ↑	42.56	49.54	Better answers
nuScenes motion ADE ↓	1.39 m	1.74 m	Worse trajectories
Waymo motion ADE ↓	2.76 m	2.63 m	Better transfer

Possible explanations :

- Error in parent QAs and propagated.
- The behavior summary may omit detail from ealier QA answer
- Missmatch objective between QA and motion: QA measures answer quality, motion evaluation measures trajectory errors.

Georgia Tech

More QA context does not always improve behavior

P1: Perception – current scene

P1-1 Identify important objects

P1-2 Describe object X's motion status

P1-3 Describe object X's appearance

P2: Prediction — future interactions

P2-1 Predict object X's future state

P2-2 Will X be in the ego vehicle's path?

P2-3 Which objects need attention next?

P3: Planning — possible actions

P3-1 Choose actions considering X

P3-2 Identify actions risking a collision

P3-3 Identify safe actions in the scene

ID	QA types included	Behavior acc. (%)
3	P1-1, P1-2, P1-3	53.94
4	All P1 + P2-1	58.82
6	All P1 + all P2	57.70
7	All P1 + all P2 + P3-1	58.95
9	All P1 + all P2 + all P3	57.49

More QA does not consistently improve accuracy.

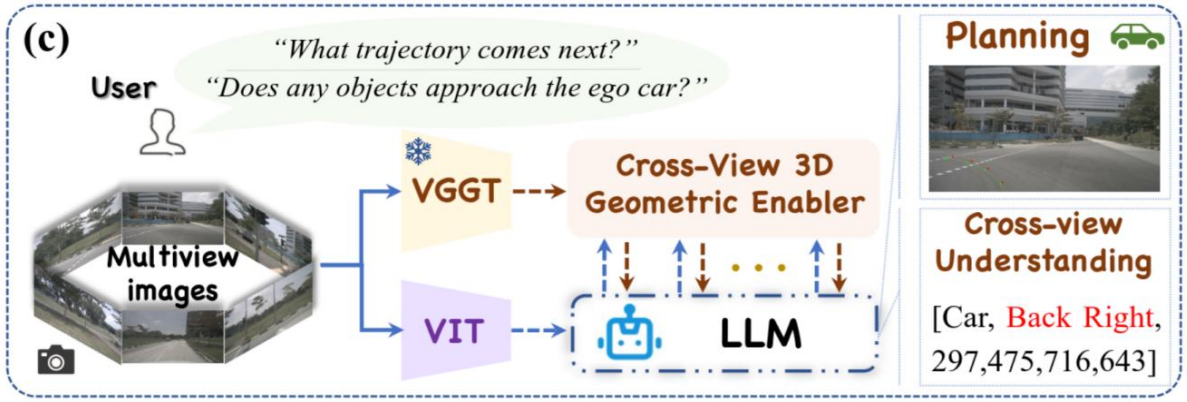
Georgia Tech

VLGA: Vision-Language-Geometry-Action Models for Autonomous Driving

Georgia Tech

Vision-Language-Geometry-Action: Problem

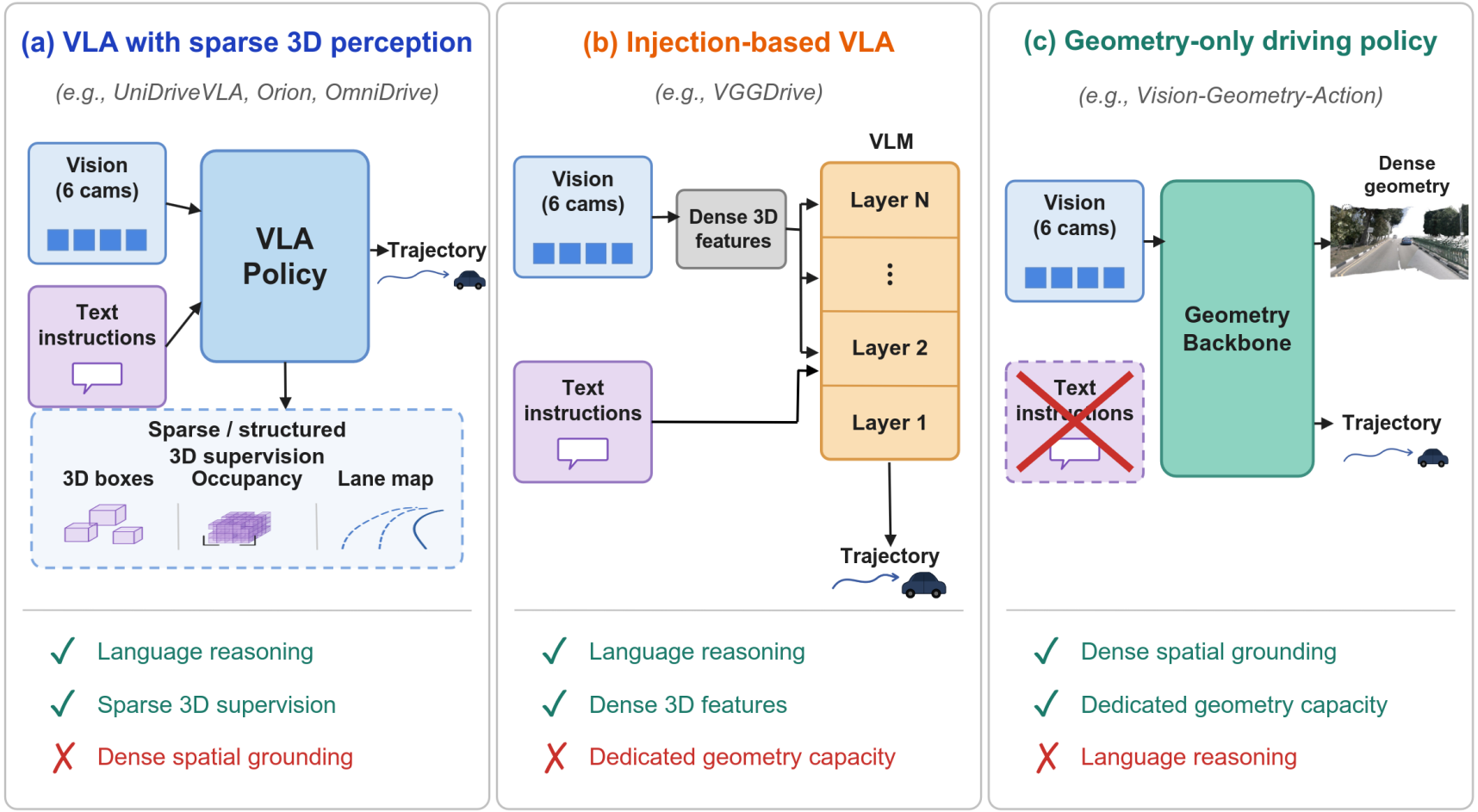
- Dense 3D grounding is essential for tight-clearance safety
  - Squeezing into a tight parking space
- How do we equip a language-enabled driving policy with dense 3D grounding?
  - Without harming the language priors
  - And without relying only on sparse 3D



VGGDrive (No explicit geometry capacity)

VLGA Prior Paradigms: VLA With Sparse 3D perception

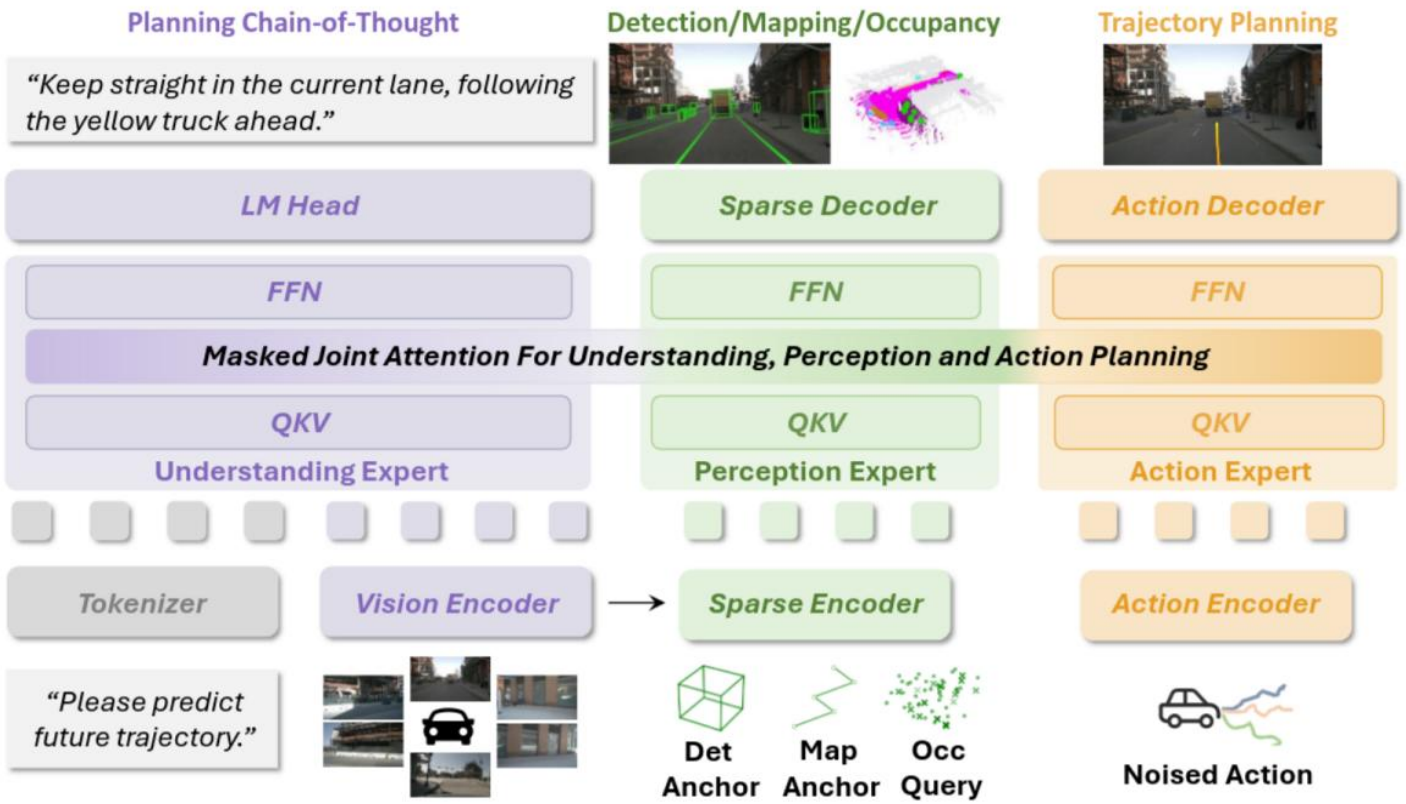
Prior works have different approaches to 3D and language



Georgia Tech

VLGA Prior Work: UniDriveVLA

- UniDriveVLA is a "sparse" style mixture of transformers
- Three experts handle semantics, sparse perception, and trajectory generation
  - Continuous trajectories via flow matching over 6 future steps

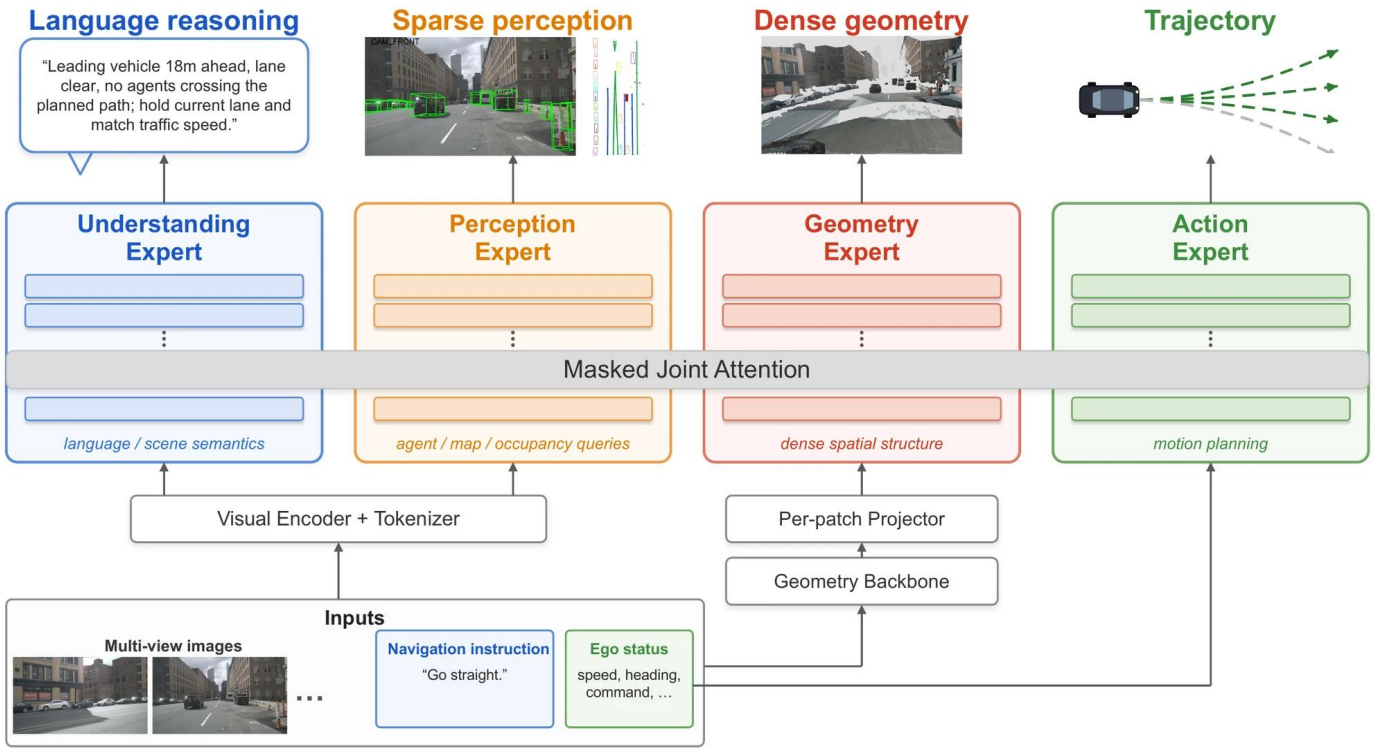


Vision-Language-Geometry-Action Architecture

Mixture-of-Transformers based on UniDrive VLA

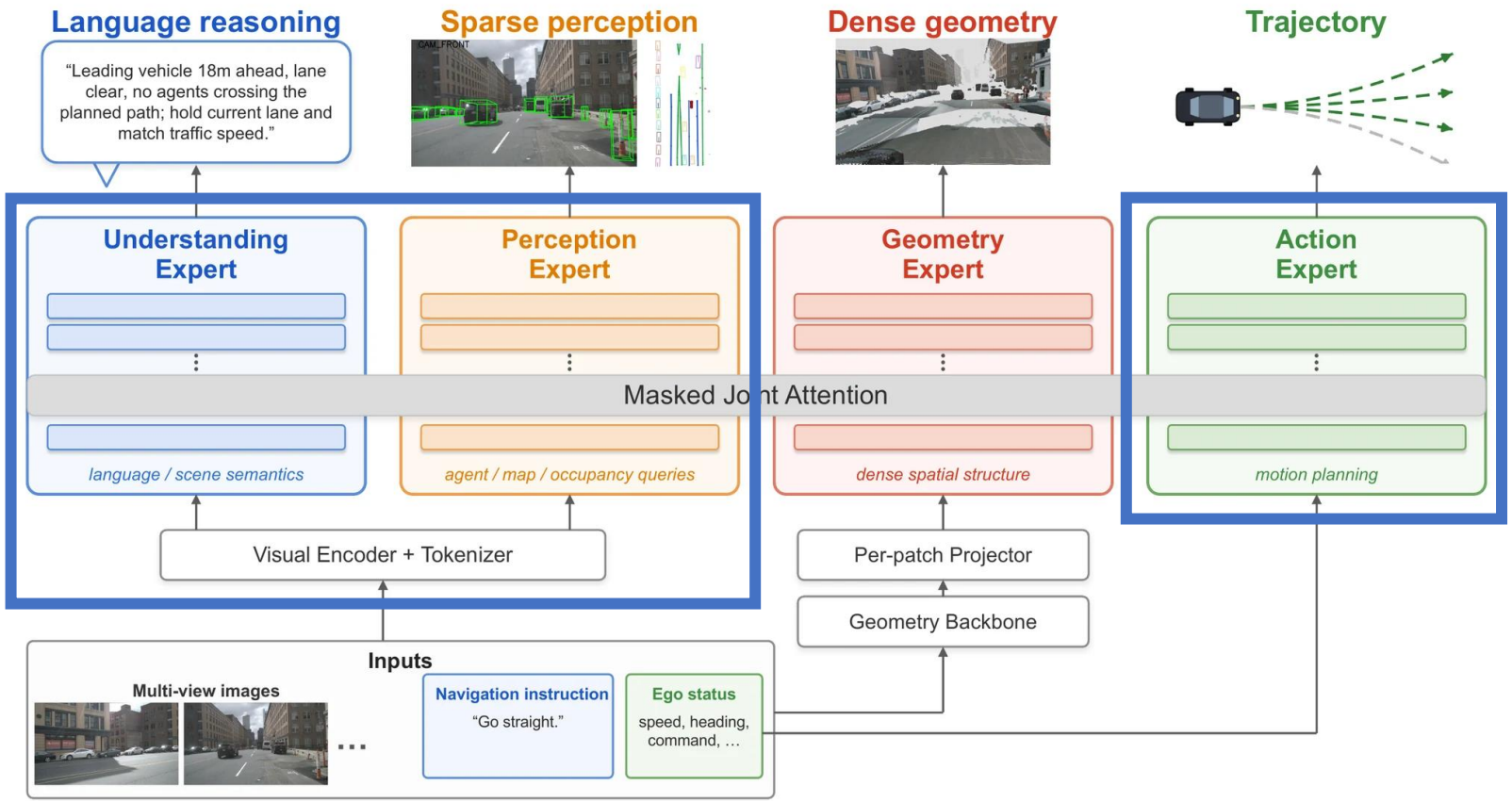
New explicit Geometry Expert reconstructs dense 3D

Understanding & Perception experts are frozen



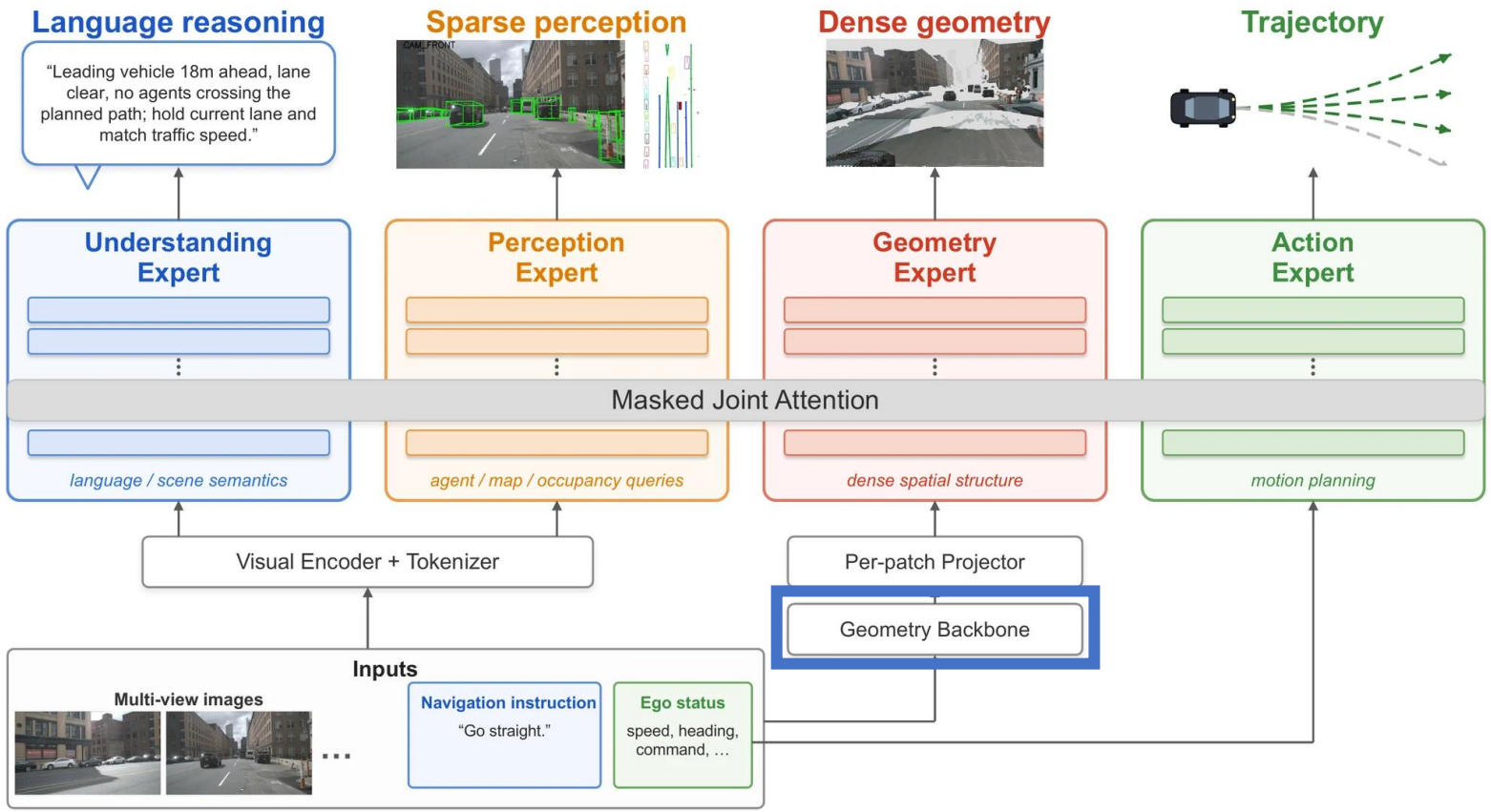
VLGA Training: Initialization

- VLM backbone, Understanding & Perception expert initialized from UniDriveVLA
  - Qwen3-VL-2B and 8B, depending on variant



VLGA Training: Initialization

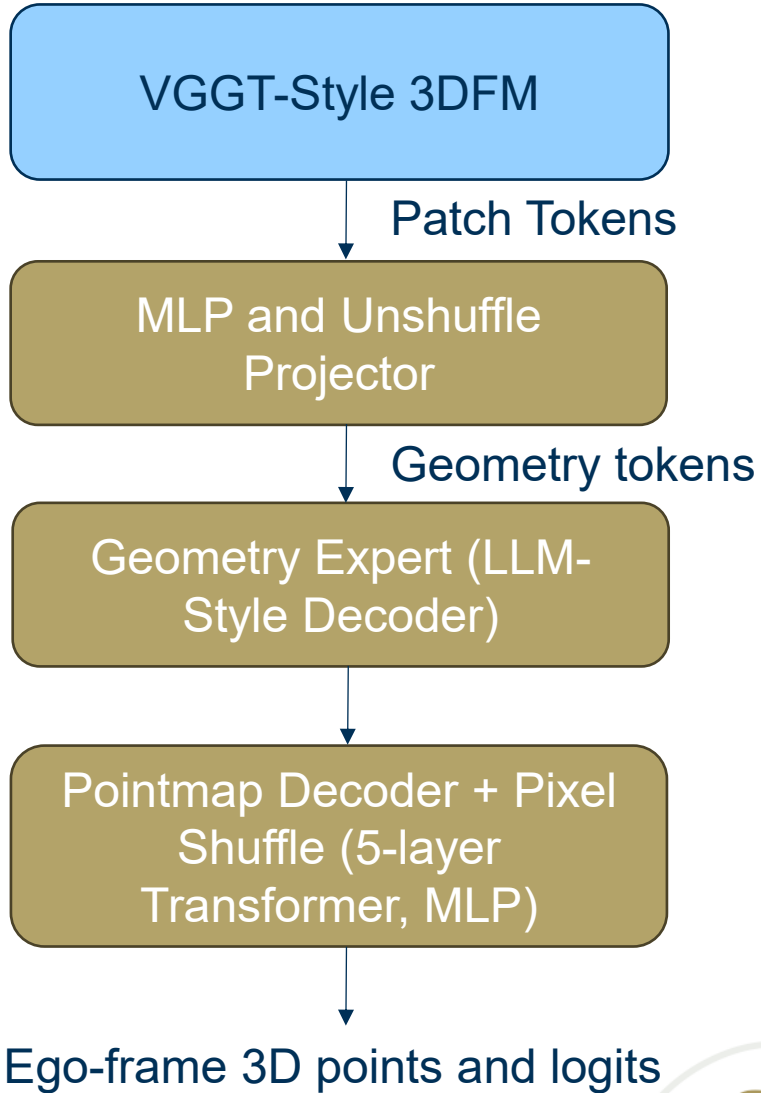
- Geometry backbone initialized from DVGT-2



Geometry Expert

- Geometry expert consists of
  - 3DFM backbone with MLP projector
  - LLM-style decoder with joint attention
  - Pointmap decoder
- Final prediction is a "pointmap"
  - 3D points in the ego-frame
  - Logits for uncertainty

The key idea is that geometry representation will share info with other experts (via masked joint attention)



Georgia Tech

Masked Joint Attention: Expert Communication

Understanding/Perception/Action come from UniDriveVLA

Geometry attends to understanding and perception, not the other way around

Query / Key	Understanding	Perception	Geometry	Action
Understanding	✓	X	X	X
Perception	✓	✓	X	X
Geometry	✓	✓	✓	X
Action	✓	✓	✓	✓

Georgia Tech

VLGA Training Overview

- Two stage training:
  - Geometry warmup: Only pointmap loss, training geometry expert and decoder
  - Adaptation: Both geometry and action experts receive updates
- Datasets:
  - CARLA Bench2Drive: 220 routes in 12 simulated towns
  - NuScenes: 28,000 training keyframes

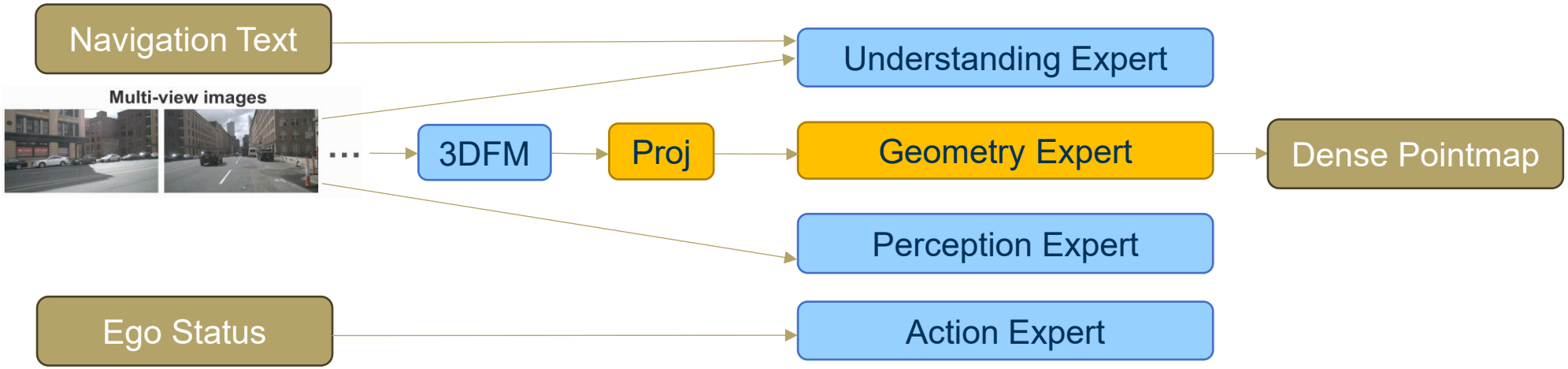
Georgia Tech

VLGA Training: Geometry Warmup Stage

- First stage focuses on geometry
- Only the Geometry expert is trained via pointmap loss

$$\mathcal{L}_{\text{pmap}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \left(\frac{\|\hat{\mathbf{x}}_p - \mathbf{x}_p^{\text{gt}}\|_1}{b_p} + \log b_p \right), \quad b_p = \text{softplus}(c_p),$$

- Everything else is frozen

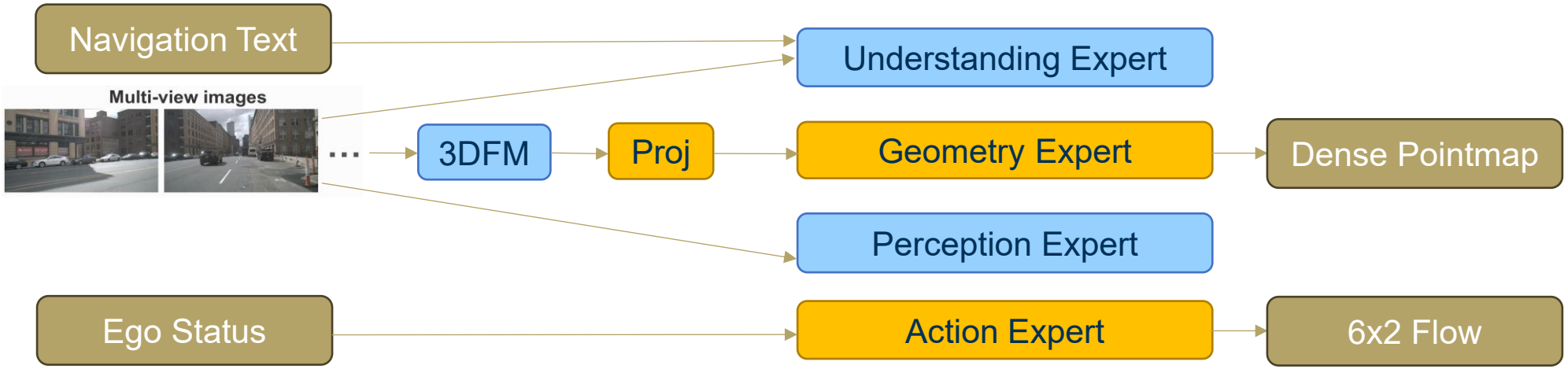


VLGA Training: Joint Stage

- Next we unfreeze the action expert
- Train with pointmap loss and action (flow matching) loss

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{act}} + \lambda_{\text{pmap}} \mathcal{L}_{\text{pmap}}$$

- Idea: Adapt the action expert to the joint attention with geometry



Georgia Tech

VLGA Experiments: nuScenes Open-Loop

Without ego-state, VGLA-Large (8B) achieves the lowest L2 over:

- 6019 validation samples across 150 urban scenes

Upshot: Dense geometry helps (slightly) when odometry is not provided

	L2 @ 1s (m)	L2 @ 2s (m)	L2 @ 3s (m)	Avg
UniDriveVLA-Base	0.28	0.51	0.82	0.54
UniDriveVLA-Large	0.27	0.49	0.77	0.51
VGLA-Base (2B)	**0.26**	0.50	0.81	0.52
VGLA-Large (8B)	**0.26**	**0.48**	**0.76**	**0.50**

Georgia Tech

VLGA Experiments: nuScenes Open-Loop

Without ego-state, VGLA-Large (8B) achieves 18% lower collision rate than UniDrive

Upshot: dense geometry grounding may tangibly improve safety

	Collision @ 1s (%)	Collision @ 2s (%)	Collision @ 3s (%)	Avg
DVGT-2	0.20	0.37	0.66	0.41
UniDriveVLA-Base	0.08	0.13	0.31	0.17
UniDriveVLA-Large	0.03	0.10	0.21	0.11
VGLA-Base (2B)	0.03	0.10	0.28	0.14
VGLA-Large (8B)	**0.02**	**0.06**	**0.18**	**0.09**

Georgia Tech

Ablation: Pointmap Supervision

This notion is also suggested directly by ablating the model config.

Geometry expert alone is not sufficient to fully reduce collisions

Model Config	Average L2 (m)	Average Collision (%)	Empirical Outcome
Baseline (no dense geometry)	0.539	0.169%	Lateral drifts in turns
+ Geometry Expert	0.529	0.149% (-11.8%)	Extra capacity reduces collisions
+ Dense Pointmap Supervision	0.524	0.136% (-19.5%)	Explicit 3D further reduces collisions

Georgia Tech

VLGA Experiments: Bench2Drive Closed-Loop

Closed loop vehicle control in CARLA, which (to some degree) also tests failure recovery

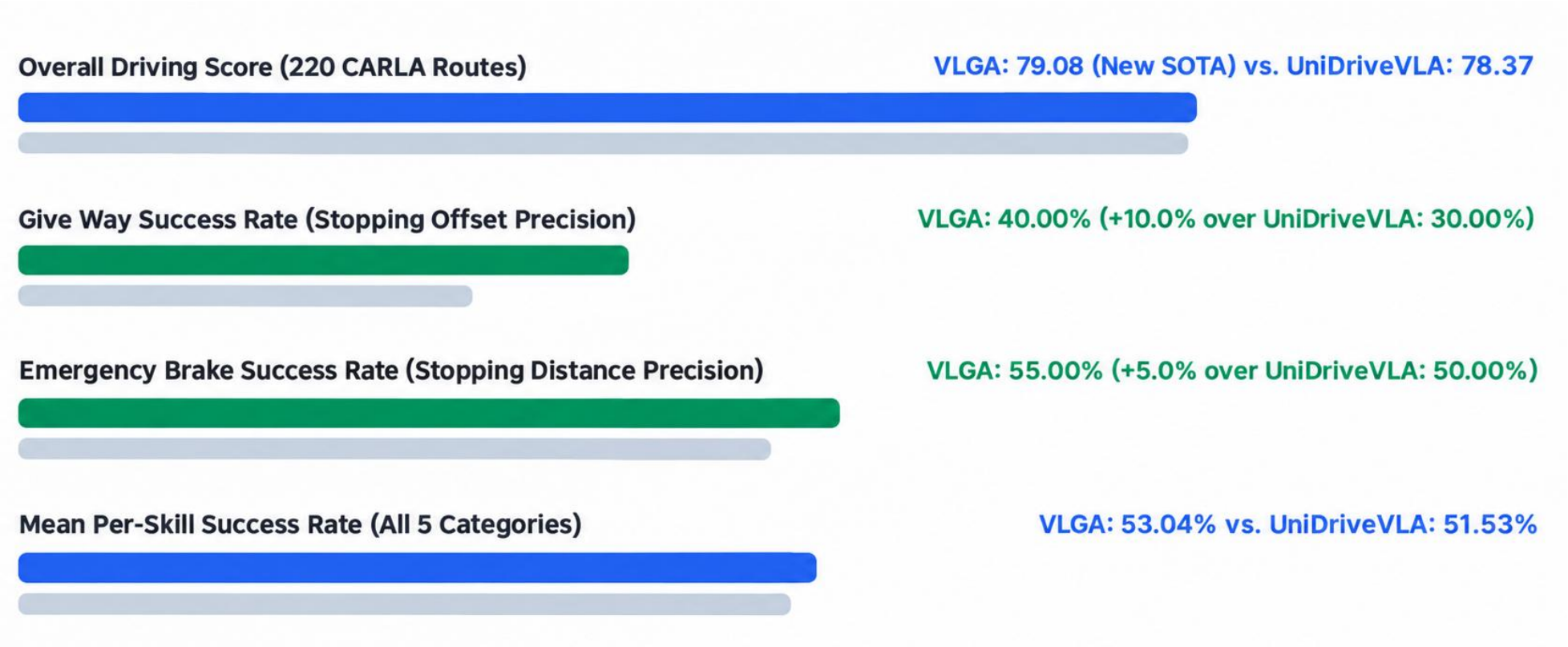
SOTA in overall driving score, but not universally better than UniDrive

	Avg. L2	Driving Score	Success Rate	Efficiency	Comfortness
Drive-MoE	**0.38**	74.22	48.64	175.96	15.31
Orion	0.68	77.74	**54.62**	151.48	17.38
UniDriveVLA	0.72	78.37	51.82	**198.86**	**11.78**
VGLA-Large	0.69	**79.08**	52.73	194.63	13.06

Georgia Tech

Per-Skill Loop Success Rate (CARLA)

Gains are highest on scenarios requiring precise 3D understanding

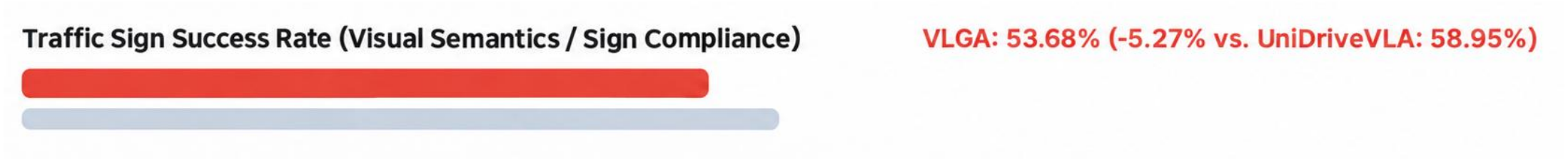


Georgia Tech

Per-Skill Loop Success Rate (CARLA)

However, some scenarios do not require dense 3D and degrade.

Discussion question: is masking out geometry in the understanding expert enough to preserve language?



Georgia Tech

Experiments and Results Summary

- Open-loop Planning on nuScenes:
  - Best on 15/16 metrics among methods without Ego status
  - Lowest L2 average and 2nd lowest collision
- Closed-loop driving on Bench2Drive in CARLA:
  - SOTA driving score of 79.08 vs. UniDriveVLA of 78.37
  - Improved success rate of 52.73%

Qualitative Results

VGLA better fits the ground truth compared to UniDriveVLA

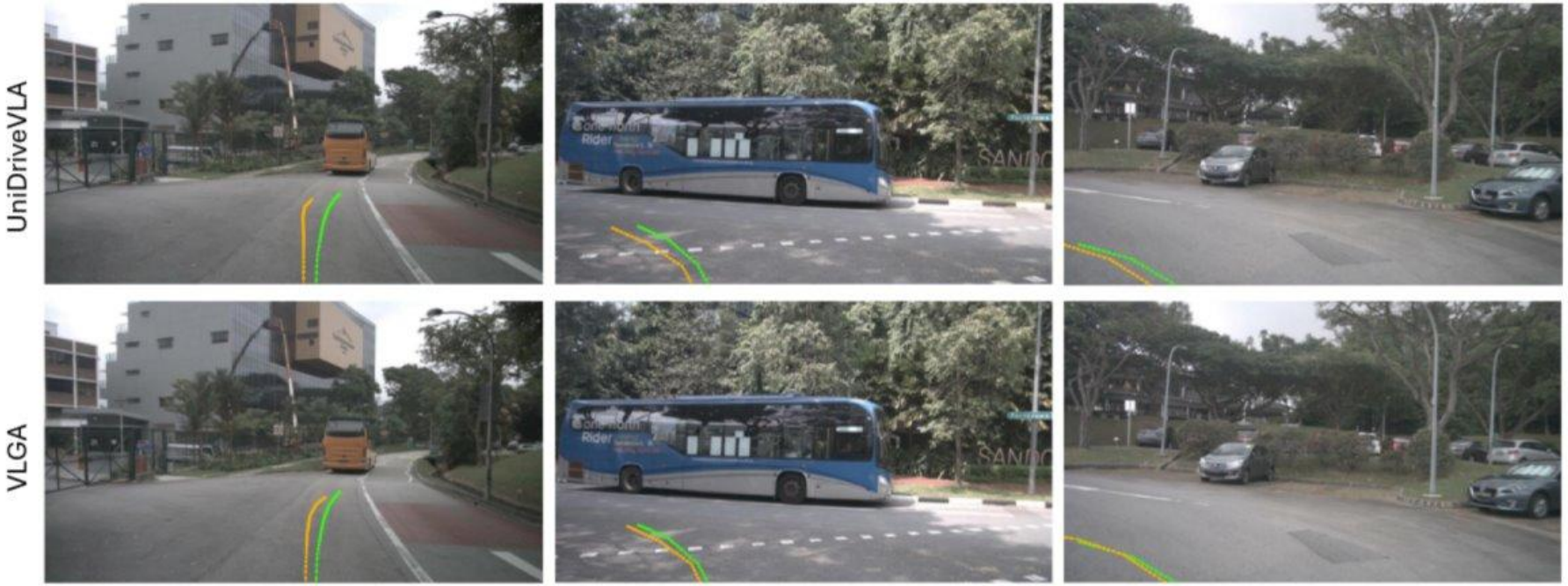


Figure 3 Qualitative planning comparison on the nuScenes validation set. The predicted 3-second trajectory (yellow) and ground truth (green) are projected onto the front camera. VLGA stays closer to the ground truth through turns and around nearby vehicles, while the baseline (UniDriveVLA) drifts laterally.

Georgia Tech

From DriveLM to VLGA: VLMs to VLAs

Fundamentally, DriveLM and VLGA approach the problem of driving differently

- **Paradigm**:
  - Text reasoning via Graph VQA vs. Feed-forward latent fusion
- **Language use**:
  - explicit reasoning as an output vs. primarily for semantics
- **3D Grounding**:
  - No 3D at all vs. Both Sparse and Dense 3D
- **Evaluation**:
  - Similar datasets, but larger scale in VLGA (799 vs 6019 samples in nuScenes)
  - Open-loop vs. Closed-loop in CARLA due to computational feasibility

Limitations

- Dense 3D supervision is not necessarily grounding
  - This alone does not guarantee the action expert actually uses latent 3D features
- Per-frame geometry experts greatly increases inference cost
- Language inputs are contrived
- Necessity of dedicated geometry capacity is never demonstrated
  - Does not strictly ablate vs. Injected geometry

Discussion Questions

- Where and when does language help with self-driving? Is it really necessary as an **input**?
- Can geometry emerge from sufficiently-scaled action learning? Or is explicit supervision necessary?