BEVLM: Distilling Semantic Knowledge from LLMs into Bird's-Eye View Representations

Thomas Monninger*[1] · Shaoyuan Xie*[2] · Qi Alfred Chen[2] · Sihao Ding[1]

[1] Mercedes-Benz Research & Development North America, San Jose, USA

[2] University of California, Irvine, USA

ECCV 2026 | MALMÖ | SEPT 8–12

+46.0%
cross-view MCQ accuracy on Ego3D dataset
BEV vs. same-backbone image tokens

10×
smaller visual encoder
44M BEV vs. 400M ViT parameters

+28.2%
NeuroNCAP closed-loop score
on the UniAD end-to-end pipeline

−16.1%
collision rate in safety-critical
scenarios, and it transfers to VAD

Motivation

LLMs bring commonsense reasoning and long-tail understanding to autonomous driving, but today's VLMs tokenize every camera view and every frame **independently**.

That separation destroys **spatial consistency**, hinders 3D reasoning, and makes compute grow linearly with the number of frames.

BEV representations are compact and spatially consistent, yet they are supervised only by **geometric** annotations such as boxes and maps, so they stay semantically poor.

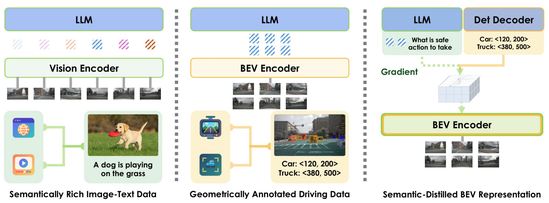

Fig. 1 · Representation comparison. Vision encoders learn from rich image–text data but process views independently (left); BEV encoders are spatially consistent but geometry-only (centre). We distill LLM semantics into BEV (right).

RQ1 Can BEV features be aligned to the language space well enough for an LLM to reason over them?

RQ2 Are BEV representations more effective than perspective-view inputs for spatial reasoning?

RQ3 Can distilling LLM knowledge into BEV encoders improve real driving safety?

Contributions

C1. The first rigorous study comparing multi-view, multi-frame perspective images against joint **BEV** representations for LLM spatial reasoning.

C2. BEVLM, a framework that distills semantic knowledge from an LLM into the BEV encoder while preserving its spatial structure.

C3. Consistent **closed-loop** gains on two end-to-end pipelines, UniAD and VAD, confirming the benefit precisely in safety-critical scenarios.

Align BEV to LLMs

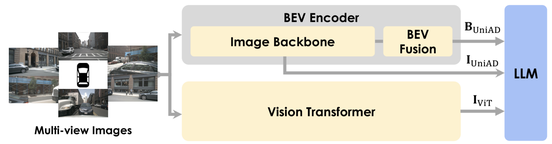

Fig. 2 · Three visual token sources feeding the same LLM: ViT tokens, pre-fusion backbone tokens, and post-fusion BEV tokens.

RQ1 · Aligning BEV with language

A single **MLP projector** maps the max-pooled 50×50 BEV grid (2,500 tokens in total) into the LLM token space; we then answer DriveLM object-existence questions.

BEVLM reaches **90.8%** with InternVL3-1B and **95.3%** with InternVL3-8B, matching, and at 8B surpassing, UniAD's task-specific detection head (92.8%).

Model	Modality	cars	peds.	trucks	cones	barriers	Avg.
Majority class	-	92.7	81.9	65.0	59.6	54.8	78.2
Linear probe	-	94.6	87.2	84.2	83.8	83.9	88.7
$\text{Detection}_{\text{UniAD}}$	-	91.1	90.9	86.7	**99.0**	**99.6**	92.8
$\text{InternVL3}_{\text{1B}}$	B_{UniAD}	94.7	88.9	85.8	90.1	88.6	90.8
$\text{InternVL3}_{\text{8B}}$	B_{UniAD}	97.7	**94.8**	**89.7**	94.0	95.0	**95.3**
$\text{DeepSeek-VL}_{\text{1B}}$	B_{UniAD}	**95.1**	92.0	84.9	92.6	92.2	92.2

Tab. 1 · BEV projector alignment on DriveLM-nuScenes. A simple projector preserves the spatial cues the LLM needs to answer correctly.

RQ2 · BEV vs. perspective views

On the cross-view Ego3D benchmark, questions span objects seen in **different ego cameras**, so the model must reason over the whole scene at once.

Against image tokens from the *same backbone*, BEV gives **+46.0%** MCQ accuracy and **−21.8%** L1 error, and it beats ViT tokens with a **10× smaller** encoder and 45% fewer tokens.

			MCQ Accuracy (%) ↑				L1 (m) ↓
Modality	Enc. Size	#Token	Loc.	Abs-Dist.	Rel-Dist.	Avg.	Dist
I_{ViT}	400M	4,608	13.95	23.81	52.94	28.57	14.15
I_{ViT} w/ ft.	400M	4,608	60.47	**50.00**	**79.41**	**62.19**	7.42
I_{UniAD}	40M	2,250	39.53	26.19	64.71	42.02	9.01
B_{UniAD}	44M	2,500	**67.44**	45.24	73.53	61.34	**7.05**

Tab. 3 · Cross-view reasoning on Ego3D. BEV tokens (44M encoder, 2,500 tokens) beat ViT tokens (400M, 4,608) on every question type.

BEVLM · Semantic Distillation

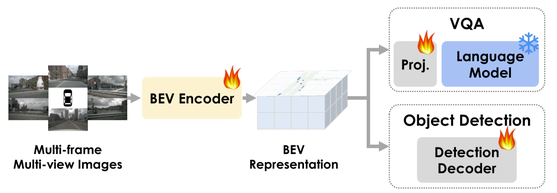

Fig. 3 · BEVLM framework. A frozen LLM supervises the BEV encoder through VQA, while the detection head keeps the grid geometrically faithful.

We **freeze** the language model and train the BEV encoder through VQA, so it must capture safety-relevant semantics that box labels never carry.

Joint **object-detection** training regularizes the grid and prevents forgetting of spatial structure.

DriveLM annotations are rewritten in **ego-centric BEV coordinates** — *"3 m ahead, 1.5 m to the left"* instead of pixel positions.

The projected BEV feature is pulled onto the LLM's ideal semantic embedding, using the frozen LLM's cross-entropy as a differentiable proxy.

End-to-End Driving Results

		Open-Loop (nuScenes)				NeuroNCAP	
Pipeline	Method	L2@1s↓	L2@2s↓	L2@3s↓	Avg.L2↓	Score↑	CR↓
UniAD	Baseline BEV	0.50	0.99	1.67	1.05	2.38±1.52	0.56±0.31
	Rand. aux. head	0.59	1.09	1.79	1.16	1.75±0.96	0.73±0.19
	VLM-AD trans.	0.56	1.02	1.64	1.07	2.02±1.18	0.70±0.23
	VLM-AD CLIP	0.57	1.01	1.64	1.07	2.07±1.13	0.67±0.23
	Distilled BEV_{1B}	**0.46**	**0.91**	**1.55**	**0.97**	2.93±0.69	0.54±0.19
	Distilled BEV_{8B}	0.48	0.94	1.59	1.00	**3.05±1.26**	**0.47±0.30**
VAD	Baseline BEV	0.47	0.79	1.19	0.82	3.24±1.21	0.37±0.24
	Distilled BEV_{8B}	**0.40**	**0.71**	**1.11**	**0.74**	**3.42±0.97**	**0.34±0.20**

Tab. 4 · Open-loop (nuScenes) and closed-loop (NeuroNCAP) planning. The distilled encoder improves both pipelines at every horizon.

UniAD: NeuroNCAP score 2.38 → **3.05** (+28.2%) and collision rate 0.56 → **0.47** (−16.1%).

VAD: the same 8B-distilled encoder lifts the score 3.24 → **3.42** and drops collisions 0.37 → **0.34**. The gain transfers across architectures.

Alternatives fall short: a random auxiliary head *lowers* the score to 1.75, VLM-AD-style CLIP distillation reaches only 2.02–2.07. The gain comes from the **pretrained LLM's semantic space**.

Qualitative: Safety-Critical Scenarios

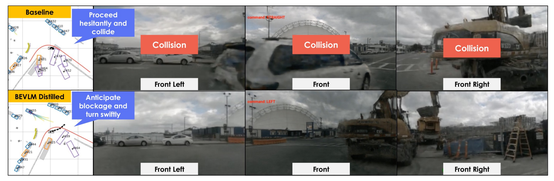

Fig. 4 · Right-turn conflict with a blocked lane. Top row baseline, bottom row BEVLM-distilled; left panel is the BEV plan, the rest are front camera views. The baseline collides; BEVLM turns swiftly.

The distilled model behaves more **adaptively**: it clears blocked intersections instead of stalling and changes lanes early to avoid an oncoming car.

Collision rate barely moves while the score rises, the gain is a real drop in **impact severity**.

Which VQA Data Matters?

Method	Perception	Prediction	Behavior	Planning	NeuroNCAP Score↑	CR@0.0s↓	Imp. Vel.↓
Baseline BEV					2.38±1.52	0.56±0.31	6.11
$\text{Distilled}_{\text{1B}}$ BEV	✓	✓			2.77±1.11	0.54±0.25	4.64
$\text{Distilled}_{\text{1B}}$ BEV			✓	✓	2.75±1.30	0.54±0.26	**3.97**
$\text{Distilled}_{\text{1B}}$ BEV	✓	✓	✓	✓	**2.93±0.69**	**0.54±0.19**	4.08

Tab. 6 · VQA data ablation. *Behavior* and *Planning* questions help more than *Perception* and *Prediction*; all four combined are best.

Every subset beats the undistilled baseline (2.77 / 2.75 vs. 2.38); all four combined give the best score, **2.93**.

Conclusion

BEV is a **better visual representation** for LLMs than perspective images for spatially consistent, far cheaper to encode, and much stronger on cross-view reasoning.

Distilling a frozen LLM into the BEV encoder injects semantics geometry alone cannot, and it **survives to closed-loop safety**.

Semantic distillation is a drop-in upgrade for BEV-based end-to-end driving, across both UniAD and VAD and only at training time.

Project page
qualitative videos

Paper
arXiv:2603.06576