EgoDyn-Bench: Evaluating Ego-Motion Understanding in Vision-Centric Foundation Models for Autonomous Driving

Finn Rasmus Schäfer[1], Yuan Gao[1], Dingrui Wang[1], Thomas Stauner[2], Stephan Günnemann[3], Mattia Piccinini[1], Sebastian Schmidt[2,3], and Johannes Betz[1]

[1] Professorship of Autonomous Vehicle Systems, Technical University of Munich, Munich, Germany
finn.schaefer@tum.de
[2] Bayerische Motoren Werke AG, Munich, Germany
[3] Data Analytics and Machine Learning Group, Technical University of Munich, Munich, Germany

Abstract. While Vision-Language Models (VLMs) have advanced high-level reasoning in autonomous driving, their ability to ground this reasoning in the underlying physics of ego-motion remains poorly understood. We introduce *EgoDyn-Bench*[4], a diagnostic benchmark for evaluating the semantic ego-motion understanding of vision-centric foundation models. By mapping continuous vehicle kinematics to discrete motion concepts via a deterministic oracle, we decouple a model's internal physical logic from its visual perception. Our large-scale empirical audit spanning 20+ models, including closed-source MLLMs, open-source VLMs across multiple scales, and specialized VLAs, identifies a significant **Perception Bottleneck**: while models exhibit logical physical concepts, they *consistently fail to accurately align them with visual observations*, frequently underperforming classical non-learned geometric baselines. This failure persists across model scales and domain-specific training, indicating a structural deficit in how current architectures couple visual perception with physical reasoning.

We demonstrate that providing explicit trajectory encodings substantially restores physical consistency across all evaluated models, revealing a functional **disentanglement between vision and language**: ego-motion logic is derived almost exclusively from the language modality, while visual observations contribute negligible temporal signal. This structural finding provides a standardized diagnostic framework and a practical pathway toward physically aligned embodied AI.

Keywords: Ego-motion · Physical Reasoning · Foundation Models

[4] Project page: TUM-AVS/EgoDyn-Bench-Website. Code: TUM-AVS/EgoDyn-Bench. Dataset: fnc1901/EgoDyn-Bench.

1 Introduction

Classical autonomous driving systems explicitly model ego-motion through estimated physical state variables such as velocity, acceleration, and yaw rate [12, 20, 32]. These representations are not auxiliary but fundamental: they ensure that perception, planning, and control remain consistent with the vehicle's underlying dynamics.

Recent advances in vision-centric foundation models, particularly Vision-Language Models (VLMs), propose an alternative paradigm in which high-level reasoning and planning are performed directly from visual observations [15, 31, 45, 47]. In these approaches, explicit ego-state representations are typically absent, and motion must be inferred implicitly from image sequences.

This shift raises a fundamental question: *Do such models form a physically consistent understanding of ego-motion, or does their visual reasoning remain decoupled from the vehicle's underlying dynamics?*

While current benchmarks primarily assess high-level planning and reasoning tasks [28, 36, 43], none verify whether model outputs are physically consistent with the ego-vehicle's own kinematic state, leaving a critical axis of embodied understanding entirely unevaluated.

To address this gap, we introduce *EgoDyn-Bench*, a benchmark for explicitly evaluating ego-motion understanding in vision-centric models. We formulate this as a semantic video question-answering task grounded in physically derived labels, enabling controlled and interpretable assessment of whether model predictions semantically align with the underlying dynamics. Using this framework, we analyze modern vision-centric models and study the role of explicit dynamic information in their performance. Our results highlight limitations of current alignment based on visual observations and dynamic concepts, and additionally provide a strategy for improving models without retraining.
Our contributions are as follows:

- **Ego-Motion Benchmark:** We introduce *EgoDyn-Bench*, a benchmark that explicitly evaluates ego-motion understanding in vision-centric foundation models, isolating physical consistency from downstream task performance.
- **Physically-Grounded Evaluation Framework:** We propose a semantic abstraction and deterministic oracle-based labeling pipeline that maps continuous ego-dynamics to interpretable motion concepts, enabling reproducible and controlled evaluation.
- **Large-Scale Empirical Analysis:** Through our audit, we show that current VLMs exhibit a fundamental "visual grounding deficit", failing to reliably capture ego-motion from visual input alone despite possessing physically consistent but biased reasoning capabilities.
- **Recovering Grounding via Explicit Dynamics:** We demonstrate that providing textual dynamic state information yields substantial performance gains, providing a pathway to improve physical consistency without the need for expensive retraining.

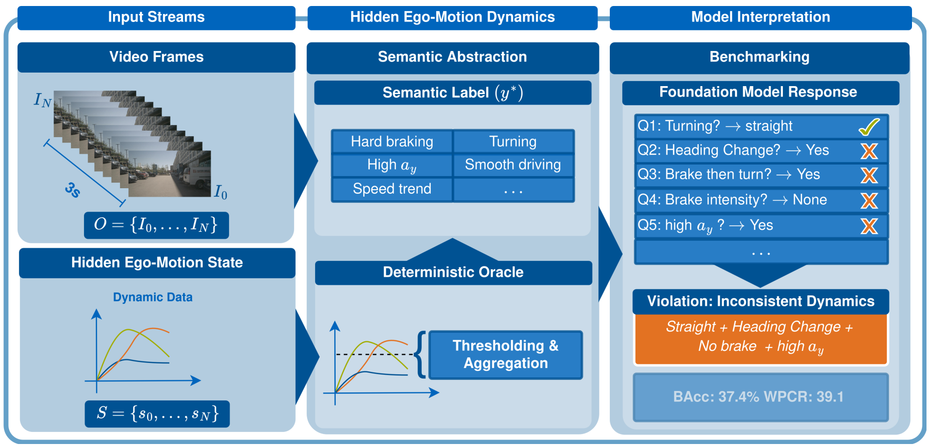


Fig. 1: EgoDyn-Bench Overview. Continuous kinematic states S are mapped to semantic labels via a deterministic oracle to define a VideoQA task over visual observations O. Models are evaluated on their ability to infer motion dynamics through semantic, temporal, and physical consistency ($WPCR$) metrics.

2 Related Work

Existing evaluation frameworks for vision-centric foundation models in the autonomous driving domain can be grouped into four primary paradigms: (i) logical reasoning and decision interpretability, (ii) spatial intelligence, (iii) numerical trajectory forecasting and closed-loop driving, and (iv) physical reliability audits. Our work targets a missing axis across all four: *whether predicted decisions are consistent with the underlying physical concepts of ego-motion over time.*

Standard Autonomous Driving Benchmarks and Metrics Classical autonomous driving evaluation spans both dataset-based and simulator-based protocols. Large-scale datasets such as nuScenes [4] and Waymo [33] support open-loop evaluation of perception and trajectory prediction, typically using displacement-based metrics (*e.g.*, ADE/FDE). In contrast, closed-loop benchmarks and simulators such as nuPlan [17] and CARLA [7] evaluate planning and control through metrics such as route completion and time-to-collision. While these protocols effectively assess task-level completion, they abstract away the underlying physical reasoning. As a result, models might achieve high performance through spurious visual correlations rather than genuine kinematic understanding, posing significant risks to downstream generalization and safety.

Logical Reasoning and Decision Interpretability (LR) Frameworks such as DriveLM [31] and Reason2Drive [24] evaluate reasoning through structured or interpretable decision outputs. These approaches represent behavior as discrete linguistic instructions, focusing on semantic plausibility. Because these linguistic instructions are inherently discrete and lack continuous temporal constraints, they cannot guarantee that a sequence of decisions will result in, or be grounded in, a kinematically feasible maneuver. Our setting addresses this by requiring reasoning grounded in temporally continuous motion rather than isolated semantic descriptions.

Object-Centric Spatial Intelligence. Recent benchmarks like Ego3D-Bench [9] and RADAR [5] evaluate spatial relations and volumetric overlap of external objects. In contrast, ego-motion understanding is fundamentally self-referential, requiring the grounding of the agent's own motion state within the temporal visual stream. *EgoDyn-Bench* addresses this gap by providing a structured diagnostic to evaluate whether a model's high-level semantic interpretation of its own movement is accurately anchored in physical concepts.

Trajectory Forecasting and Control. Standard benchmarks like Argoverse [41], ScenePilot-Bench [40], and EgoTraj-Bench [21] evaluate motion via displacement-based metrics. However, spatial accuracy does not guarantee kinematic feasibility or compliance with underlying physical concepts. Instead of assessing motion generation, *EgoDyn-Bench* provides an isolated diagnostic of the model's intrinsic high-level physical understanding, evaluating whether its semantic interpretations are accurately anchored in continuous kinematic constraints.

General Physics Audits. Benchmarks like DriveBench [43] expose "text-only resilience," where models rely on language priors rather than visual grounding. Physics audits such as QuantiPhy [27] and Morpheus [46] show that models struggle with general conservation laws and external object collisions. In contrast, *EgoDyn-Bench* isolates the understanding of embodied kinematics, testing whether models can infer their own mechanically valid motion states directly from sequential visual streams.

Limitations of Existing Approaches To summarize, current evaluation paradigms fracture the driving problem along a consistent fault line: classical approaches, including optical flow, visual odometry, and displacement-based trajectory metrics, offer rigorous geometric tracking but no semantic understanding, while foundation model benchmarks evaluate high-level reasoning from visual snapshots without any grounding in the vehicle's own kinematic state. *EgoDyn-Bench* bridges this division by formally testing whether vision-centric semantic predictions satisfy the kinematic concepts related to continuous ego-motion. Our analysis reveals that model failures stem from misalignment between visual observations and physical motion concepts, not from an absence of physical reasoning capacity.

3 EgoDyn-Bench

To examine the kinematic motion understanding of foundation models, we introduce *EgoDyn-Bench*. Our benchmark is the first to evaluate whether vision-centric models can infer physically consistent ego-motion concepts from visual observations. A comparison to existing frameworks is shown in Table 1. Unlike prior object-centric or regression-based benchmarks, we isolate *self-referential motion understanding* as a semantic reasoning problem grounded in vehicle kinematics. Our benchmark consists of: (i) a task formulation mapping visual inputs

Table 1: Comparison of VLM evaluation benchmarks in autonomous driving. Existing benchmarks evaluate external environments or abstract actions, but fail to assess the agent's internal kinematics. *EgoDyn-Bench* isolates this missing self-referential axis. **OCS**: Object-Centric Spatial; **CLT**: Closed-Loop & Trajectory; **LR**: Logical Reasoning; **GPA**: General Physics Audits; **KEM**: Kinematic Ego-Motion.

Benchmark	OCS	CLT	LR	GPA	KEM
Classical & Trajectory Benchmarks					
nuScenes-QA [28]	✓	✗	✗	✗	✗
nuPlan [17]	✓	✓	✗	✗	✗
ScenePilot / EgoTraj [40]	✗	✓	✗	✗	✗
Vision-Language & Reasoning Benchmarks					
DriveLM / Reason2Drive [31]	✗	✗	✓	✗	✗
Ego3D / RADAR [9]	✓	✗	✓	✗	✗
DriveBench / QuantiPhy [43]	✗	✗	✗	✓	✗
Proposed Benchmark					
***EgoDyn-Bench* (Ours)**	✗	✗	✗	✗	✓

to semantic motion concepts, (ii) a dataset of real-world and augmented driving sequences, and (iii) a reproducible labeling pipeline deriving ground-truth from physical signals. Figure 1 provides a conceptual overview.

3.1 Problem Formulation

Goal. We formulate visual ego-motion understanding as a semantic question-answer task: given a *visual observation sequence* and a *natural language query* regarding the vehicle's movement, a model must produce a *semantic response* grounded in the underlying motion. By shifting from raw state regression to a linguistic interface, we evaluate a model's ability to extract functional physical concepts from observations rather than simply regressing numerical values.

Input. Formally, the model receives a sequence of visual observations

$$O = \{I_0, \ldots, I_N\},$$

sampled at frequency f_{cam} over a temporal window of duration τ. This is paired with a natural language query P that specifies a motion-related property (*e.g.*, turning direction or braking behavior). Following established protocols for trajectory forecasting and planning [17, 41], we utilize fixed-length segments where $\tau = 3\,\mathrm{s}$. This duration provides sufficient context for characteristic maneuvers to unfold as observable spatio-temporal changes while remaining brief enough to isolate distinct semantic behaviors and avoid confounding scene transitions. While our experiments utilize 3-second clips, the formulation remains flexible across varying temporal horizons.

Kinematics. The ego-vehicle's motion is characterized by a physical state sequence $S = \{s_0, \ldots, s_N\}$, recorded from onboard sensors or simulated ground-truth during data collection and strictly withheld from the model during inference.

$$s_t = [v_t, a_t, j_t, \omega_t, \theta_t]^\top$$

captures the vehicle's speed v_t, longitudinal acceleration a_t, jerk j_t, yaw rate ω_t, and heading θ_t. Rather than regressing these exact numerical states, our formulation probes whether models grasp the functional physics of ego-motion, evaluating the semantic implications of these kinematics rather than precise estimation.

Evaluation. We define a vision-centric foundation model $\mathcal{F}_\theta$ that maps the visual and textual inputs to a predicted semantic response:

$$\mathcal{F}_\theta(O, P) \to \hat{R},$$

where $\hat{R}$ is an answer selected from a predefined semantic space (binary or multiple-choice options). To enable objective evaluation, we define a deterministic **oracle** $\mathcal{G}$:

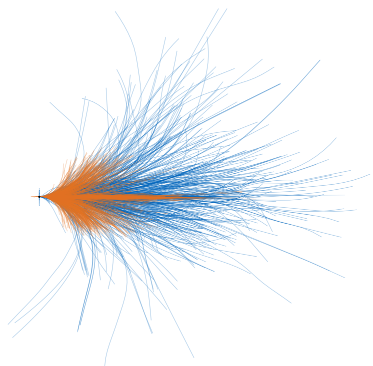
$$\mathcal{G}(S, P) \to R^*,$$

which maps the physical state sequence S and query P to a ground-truth semantic answer R^*. This ensures that labels are derived directly from measurable kinematics rather than subjective human annotation. Given the short horizon τ, sensor drift is negligible, and our sensitivity analysis confirms that model rankings remain robust to variations in the oracle's thresholds (see Supplementary).

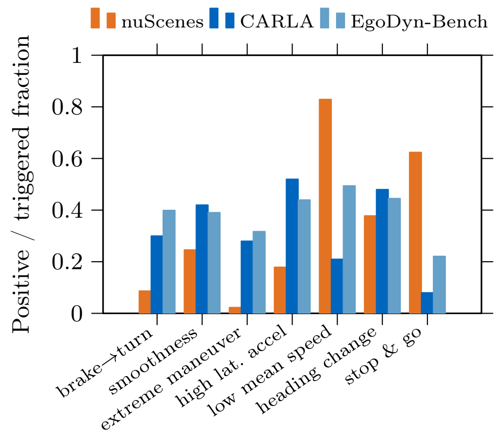
3.2 Dataset Construction

To enable controlled evaluation of ego-motion understanding, our data generation pipeline must satisfy three requirements: broad coverage of ego-dynamics, access to withheld physically grounded motion signals for oracle annotation, and a controllable distribution of motion regimes. To achieve this, *EgoDyn-Bench* employs a hybrid approach: we utilize *nuScenes* [4] for real-world driving sequences and augment underrepresented motion regimes using targeted simulations in CARLA [7], driven by CommonRoad scenarios [19] and an adaptable motion planner [37] according to [8].

Distribution Balancing & Data Curation. Real-world datasets, like *e.g.*, *nuScenes*, provide authentic visual and kinematic synchronization but are inherently biased toward low-dynamic, routine driving. To ensure a physically comprehensive distribution across longitudinal and lateral profiles, we explicitly control the benchmark's statistics through a structured four-stage curation pipeline. First, *dynamic characterization* identifies the natural low-dynamic bias of the real-world logs. Next, *dynamic mining* extracts rare but informative maneuvers directly from *nuScenes*. To balance the remaining underrepresented regions, such as emergency braking or high lateral acceleration, we employ *targeted augmentation*, injecting dynamically diverse trajectories simulated in CARLA. Finally, all sequences undergo human *validation* to verify the extracted motion



(a) Trajectory Distributions



(b) Balancing Effect on Dynamics

Fig. 2: Effect of Dataset Augmentation. (a) Spatial coverage of *nuScenes* (orange) vs. *CARLA-derived* scenarios (blue). CARLA expands the state-space to include complex maneuvers required for robust benchmarking. **(b)** Positive label fractions for representative questions. *EgoDyn-Bench* corrects the low-dynamic bias of nuScenes by injecting dynamically augmented synthetic sequences.

signals and labeling rules. Detailed curation thresholds are provided in the Supplementary Material. Figure 2 illustrates how this mixture effectively corrects spatial and dynamic biases.

Domain Alignment. To mitigate the visual domain gap between simulated and real-world sequences, we apply a photometric style transfer model (NVIDIA Cosmos Transfer 2 [29]) to the CARLA scenarios. Because our benchmark evaluates motion understanding rather than photometric fidelity, this alignment strictly prioritizes the preservation of geometric and kinematic cues over exact visual realism. We verify the recoverability of these essential motion signals across domains via geometric baselines, detailed in Section 5.3.

3.3 Semantic Abstraction and Label Generation

We formalize discrete driving maneuvers by applying a deterministic thresholding scheme to the continuous state S. These thresholds are calibrated on the dataset distribution and cross-verified against standard automotive kinematics literature [3, 16, 44] to ensure physical plausibility. Thresholds are used exclusively by the oracle $\mathcal{G}$ to derive ground-truth labels and are never disclosed to evaluated models, which must infer semantic motion concepts from visual observations alone. A full sensitivity analysis under uniform threshold perturbation by a factor $\alpha \in [0.5, 1.5]$, measured via Kendall's τ [18], confirms stable model rankings ($\tau > 0.9$) across all perturbation levels. This design ensures robustness and generalization. The semantic definitions can be adapted to new domains or specific research requirements by adjusting the threshold parameters.

Semantic Categories. We evaluate 14 distinct question categories within a unified prompt template, spanning two complementary reasoning dimensions: (i) *direct dynamics*, which probe instantaneous or aggregated motion properties such as speed regime, braking intensity, lateral acceleration, and driving smoothness; and (ii) temporal comparative, which require the model to reason about the ordering or co-occurrence of events across the clip.[5]

3.4 Evaluation Metrics

We evaluate ego-motion understanding by treating the model's natural language responses as semantic reasoning over discrete motion concepts. Let $y_i \in \mathcal{Y}$ denote the oracle label for a query P_i on clip i, and $\hat{y}_i$ the predicted label obtained by mapping the model response $\hat{R}_i$ to the label space $\mathcal{Y}$ via a deterministic parser. Our metrics are designed to distinguish between a model's ability to extract information from pixels (correctness) and its ability to maintain logically sound internal physical reasoning (consistency).

Semantic Correctness. While *EgoDyn-Bench* is balanced across question categories to avoid dataset-driven bias, models often exploit internal linguistic priors rather than grounding responses in visual content [10]. To ensure performance reflects genuine physical reasoning, we utilize **Balanced Accuracy** and **Macro-F1**. Balanced Accuracy (BAcc) is defined as the mean of class-wise recalls:

$$\text{Bal. Acc.} = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \frac{1}{N_c} \sum_{i=1}^{N} 1[\hat{y}_i = c \wedge y_i = c], \tag{1}$$

where N_c is the number of ground-truth instances for class c. For temporal queries that compare event ordering, we report **Temporal Accuracy**, the fraction of correctly predicted temporal orderings.

Weighted Physics Consistency Rate (WPCR). Beyond correctness, we introduce WPCR to diagnose the internal coherence of model predictions. We define a set of Boolean physical constraints $\mathcal{R} = \{r_m\}$, as reported in Table 2. Importantly, WPCR is not a measure of accuracy against the ground truth, but a diagnostic of internal physical coherence, assessing whether a model's collective answers for a single clip ($\tau = 3\,\text{s}$) satisfy the physics of motion.

To prevent models from achieving high consistency scores by simply avoiding committed predictions, we weight each clip's consistency contribution by the fraction of rules it triggers. Let $\mathcal{C}$ denote the set of evaluated clips, T_c the number of applicable rules and V_c the number of violations for clip c:

$$\text{WPCR} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left(1[V_c = 0 \wedge T_c > 0] \cdot \frac{T_c}{|\mathcal{R}|} \right). \tag{2}$$

[5] Full question prompts, answer options, and labeling rules with calibrated thresholds are provided in the supplementary material.

Table 2: WPCR Kinematic Constraints. Boolean implication rules ($A \Rightarrow B$) used to compute WPCR, each evaluated on a single clip.

Rule	Boolean Implication ($A \Rightarrow B$)
	Heading / Lateral Dynamics (Hard)
R_1	Heading Change = Yes $\Rightarrow$ Turn Direction $\neq$ Straight
R_2	High Lateral Acceleration = Yes $\Rightarrow$ Turn Direction $\neq$ Straight
R_3	Turn Direction = Straight $\Rightarrow$ No Significant Heading Change
R_4	Turn Direction = Straight $\Rightarrow$ No High Lateral Acceleration
	Speed Regime / Mean Speed (Hard)
R_5	Speed Regime = Highway $\Rightarrow$ Mean Speed is Not Low
R_6	Speed Regime = Stopped $\Rightarrow$ Mean Speed is Low
R_7	Speed Regime = Stopped $\Rightarrow$ Speed Trend $\neq$ Accelerating
	Brake-then-Turn Compound (Hard)
R_8	Brake-then-Turn = Yes $\Rightarrow$ Braking Intensity $\neq$ None
R_9	Brake-then-Turn = Yes $\Rightarrow$ Turn Direction $\neq$ Straight
	Stop-and-Go (Hard)
R_{10}	Stop-and-Go = Yes $\Rightarrow$ Speed Regime $\neq$ Stopped

Hard Boolean implications are a deliberate design choice: ego-motion concepts are physically discrete and mutually exclusive, leaving no meaningful notion of partial correctness. A soft metric would mask systematic reasoning failures behind gradual penalty curves, obscuring the architectural deficits this benchmark is designed to expose.

We additionally report **Physics Coverage (PCov)** as $\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\frac{T_c}{|\mathcal{R}|}$, representing the mean fraction of physical constraints triggered per clip. A low PCov indicates that few rules are triggered per clip, which would make a high WPCR score uninformative. High PCov confirms that the consistency rules are actively exercised across the benchmark.

4 Experiments

We evaluate the ability of vision-centric foundation models to infer ego-motion dynamics from visual observations using *EgoDyn-Bench*. Our experiments analyze (i) differences across model families, (ii) the impact of domain priors, and (iii) the influence of explicit dynamic information on motion understanding.

4.1 Baselines

We evaluate a set of non-foundation model baselines that estimate ego-motion directly from visual input, spanning both classical and learning-based approaches. Specifically, we consider (i) classical optical flow based on motion field formulations [13, 22], (ii) feature-based visual odometry using KLT tracking with essential matrix estimation [11, 23, 30], (iii) learned optical flow using RAFT [35], and (iv) learning-based visual odometry using TartanVO [39]. The baselines share the same visual input and are evaluated against the same oracle-derived ground-truth labels as the foundation models. To produce answers, they apply heuristic mapping rules to their estimated motion signals, for example, optical flow magnitude to speed trend, using the same question categories and answer spaces defined by the oracle. Further details on each baseline are available in the supplementary material.

4.2 Evaluated Model Families

We evaluate representative models from three categories: (i) closed-source multimodal foundation models (*e.g.*, GPT-5.1, Claude), (ii) open-source vision-language models (*e.g.*, Qwen-VL), and (iii) domain-specific architectures for physically grounded reasoning (*e.g.*, RoboTron-Drive). This categorization, detailed fully in Tables 3 and 4, enables a direct comparison between general-purpose models and those with domain-specific inductive biases.

4.3 Evaluation Protocol

EgoDyn-Bench comprises 14,000 QA pairs across 1,000 balanced 3-second driving scenarios (500 real-world, 500 simulated). We evaluate 14 question categories (direct and temporal dynamics) via deterministic binary and multiple-choice templates. Full parsing rules, prompts, and code are available in the supplementary material. As introduced in Section 3.4, we report balanced accuracy, Macro-F1, temporal accuracy, and the introduced WPCR to assess both semantic correctness and physical coherence.

4.4 Input Settings and Evaluation Axes

To analyze the contribution of visual motion cues, we evaluate two input settings: **(i) Vision-only:** Models receive only visual observations. We uniformly sample 10 frames from each 3-second clip ($\approx$ 3.3 FPS). This sampling rate preserves sufficient temporal resolution for macroscopic dynamic reasoning while maintaining computational feasibility across the extensive suite of evaluated models. **(ii) Vision + Dynamics:** Models additionally receive explicit ego-motion signals as structured text. To isolate optimal representation formats, we ablate *four* textual trajectory encodings: a high-level *Summary* (8 scalar statistics covering kinematic means and extrema, *e.g.*, max/mean speed, max lateral acceleration), a dense kinematic *Timeseries* (per-channel v, a, ω, j values at N evenly-spaced timesteps), spatial *Coordinates* (zero-centered x, y waypoints and heading θ at N timesteps), and a *Full* combination of both timeseries and coordinate data. All explicitly provided kinematic information is temporally aligned with the N subsampled images.

Our subsequent analysis is structured along three complementary axes: (i) assessing vision-only performance to test implicit motion extraction, (ii) ablating textual dynamics to isolate reliance on numerical signals, and (iii) comparing visual domains to separate reasoning deficits from dataset artifacts.

5 Results & Discussion

We evaluate the ego-motion reasoning capabilities of vision-centric foundation models across three complementary axes: *(i) vision-only performance* to assess implicit motion extraction from visual input, *(ii) dynamics-informed question answering* to quantify the contribution of explicit trajectory signals, and *(iii) domain invariance* to separate reasoning deficits from dataset artifacts.

5.1 Vision-Centric Study

Our initial analysis investigates the models' ability to infer semantic ego-motion concepts directly from visual observations without auxiliary state information, probing their capacity for zero-shot physical grounding from visual cues alone. The results, summarized in Table 3, reveal three key insights:
(i) Classical Baselines Outperform VLMs. Vision-centric foundation models exhibit a significant performance gap compared to simple non-foundation model approaches. While classical baselines are restricted to the geometrically-answerable subset (6/14 questions), they outperform even the largest closed-source MLLMs on this overlapping subset (Visual Odometry baseline: BAcc 63.8% vs. GPT-5.1: 55.1%, Gemini 3 Pro: 59.6%, Qwen3-VL-8B: 52.3%), underscoring a fundamental struggle in current architectures to extract low-level kinematic representations from visual input. Temporally shuffling the input frames leaves performance unchanged (BAcc 39.0 vs. 38.9 ordered), indicating the deficit lies not in the visual encoder but in downstream temporal integration.
(ii) Scaling and Domain Paradox. Scaling to larger closed-source MLLMs yields only marginal gains over smaller open-source counterparts: the best closed-source model (Gemini 3, BAcc 47.0%) outperforms the best open-source 8B model (Cosmos-Reason 2-8B, BAcc 39.9%) by only 7.1%, a negligible margin given the orders-of-magnitude difference in scale. Domain-specific VLAs perform comparably to or below general open-source models of equivalent size (RoboTronDrive: 38.6%), indicating that neither scale nor in-domain training resolves the underlying architectural failure to ground physical reasoning in visual observation alone. Increasing temporal resolution to 10 FPS likewise yields no improvement (+0.4pp on BACC for Qwen3-VL, see Supplementary), confirming the bottleneck is structural rather than input-limited.
(iii) Predictive Fallback Bias. A consistent disparity between raw and balanced accuracy across VLMs indicates that models default to a single dominant answer when physical reasoning fails, rather than discriminating across classes. Our Visual Odometry baseline nearly eliminates this gap (65.1% vs. 63.8%), suggesting that explicit geometric representations effectively anchor predictions and mitigate response bias.

5.2 Dynamics-Informed Reasoning

We evaluate the impact of providing explicit ego-motion signals as auxiliary input. The results in Table 4 yield the following conclusions:

Table 3: Metrics for vision-only ablation. All metrics are reported as percentages. Note that no explicit dynamic state inputs are provided for all models in this evaluation. Best values are presented **bold**; second-best are underlined. [1] Evaluated on a functional subset of the questions. [2] Evaluated without visual observation ($O = \{\emptyset\}$). [3] Evaluated with static visual observation ($O = \{I_0\}$). [4] Evaluated with shuffled visual observations ($O' = \left\{I_{\sigma(i)}\right\}_{i=0}^{N}$), where the temporal order is randomized

Model	Parsable	Semantic			Temporal		Consistency	
	↑	Acc. ↑	BAcc. ↑	F1 ↑	Acc. ↑	F1 ↑	WPCR ↑	PCov. ↑
Geometric Baselines								
Flow Heuristic[1] [13, 22]	/	58.4	47.0	42.2	48.5	43.9	46.5	79.6
Visual Odometry[1][11, 23, 30]	/	65.1	**63.8**	**63.0**	62.7	61.1	48.0	97.1
RAFT Flow[1][35]	/	**69.8**	59.6	56.8	53.9	58.2	47.5	85.3
TartanVO[1][39]	/	67.8	54.4	50.7	**65.9**	**64.3**	42.4	99.8
Ablated Baselines								
Qwen3-VL-8B[2][2]	100.0	42.3	33.5	19.3	33.3	20.2	20.0	**100.0**
Qwen3-VL-8B[3][2]	100.0	47.5	37.0	30.0	41.0	31.6	97.4	83.9
Qwen3-VL-8B[4][2]	100.0	49.1	39.0	31.7	40.4	30.0	**98.4**	84.8
Closed-Source VLMs/MLLMs								
Claude Sonnet 4.5 [1]	99.7	49.6	38.0	32.9	36.8	34.9	83.4	94.4
GPT-5.1 [26]	100.0	54.3	45.2	40.1	43.0	38.7	96.3	76.9
gemini-2.0-flash [34]	100.0	46.2	36.5	32.4	37.2	35.1	54.1	97.5
gemini-3-pro-preview [34]	94.3	53.6	47.0	44.3	47.3	45.2	59.3	98.4
Open-Source VLMs								
Qwen3-VL-2B [2]	100.0	47.1	37.4	31.2	42.3	37.5	39.1	99.8
Qwen3-VL-4B [2]	100.0	49.7	39.8	34.4	39.7	35.1	82.9	90.0
Qwen3-VL-8B [2]	100.0	49.8	38.9	33.0	39.3	32.7	97.9	84.6
InternVL3-2B [48]	99.9	44.5	32.8	20.4	36.2	26.4	39.4	99.7
InternVL3.5-4B [38]	100.0	46.8	36.0	25.5	42.7	28.7	48.0	99.5
InternVL3.5-8B [38]	100.0	47.0	38.8	30.0	42.7	33.5	48.6	78.0
InternVL3.5-38B [38]	100.0	46.5	37.6	26.8	44.8	27.0	72.6	72.4
Camreasoner-8B [42]	91.1	45.0	36.9	27.4	37.2	28.6	46.3	92.2
Cosmos Reason 2-2B [25]	100.0	46.2	35.1	23.6	39.5	25.8	72.7	**100.0**
Cosmos Reason 2-8B [25]	100.0	48.8	39.9	35.8	41.0	36.7	92.2	80.6
VLA Models								
ImpromptuVLA [6]	100.0	47.3	37.8	28.9	40.2	29.2	41.1	72.8
RoboTron-Drive [14]	99.4	48.0	38.6	32.0	41.3	31.3	69.8	96.8

(i) Explicit Dynamics Consistently Improve Performance. Integrating explicit dynamics improves performance across all models and metrics, confirming that the failures in Table 3 stem from a misalignment between visual observations and physical motion concepts, not from an absence of physical reasoning capacity. When provided with explicit kinematic data, models demonstrate the ability to reason about ego-motion that visual input alone fails to activate. While encoding identical physical information, trajectory-only inputs in different forms yield different balanced accuracy scores, indicating that models must process dynamic representations rather than simply forwarding subtle labels.

(ii) Models Bypass Visual Input for Motion Reasoning. The trajectory-informed experiments expose an asymmetry in modality importance. For Qwen3-

Table 4: Metrics for vision and trajectory ablation. All metrics are reported as percentages. Note that explicit dynamic state inputs are provided for all models in this evaluation. Further advanced embedding ablations are available in the supplementary material. Best values are in **bold**; second-best are <u>underlined.</u> [1] Evaluated without visual observation ($O = \{\emptyset\}$).

Model	Parsable	Semantic			Temporal		Consistency	
	↑	Acc. ↑	BAcc. ↑	F1 ↑	Acc. ↑	F1 ↑	WPCR ↑	PCov. ↑
Baseline (Default: Summary Encoding)								
Qwen3-VL-8B[1] [2]	100.0	65.9	59.6	54.1	53.1	46.3	33.0	**100.0**
InternVL3.5-8B [38]	100.0	61.9	55.7	49.4	55.3	45.4	17.5	**100.0**
Closed-Source VLMs/MLLMs (Default: Summary Encoding)								
Claude Sonnet 4.5 [1]	98.1	71.5	63.5	61.1	58.7	58.1	97.1	91.2
GPT-5.1 [26]	100.0	<u>71.9</u>	<u>66.1</u>	<u>64.2</u>	53.6	49.1	97.8	84.3
gemini-2.0-flash [34]	100.0	65.1	55.8	52.6	59.4	57.1	66.6	98.9
gemini-3-pro-preview [34]	98.8	**75.8**	**68.3**	**67.7**	69.8	70.0	87.7	97.2
Open-Source VLMs (Default: Summary Encoding)								
InternVL3-8B [48]	100.0	53.8	44.8	33.4	46.3	32.1	81.0	96.8
Cosmos Reason 2-8B [25]	100.0	64.3	56.5	54.0	52.8	51.1	71.4	<u>97.6</u>
Open-Source VLA (Default: Summary Encoding)								
ImpromptuVLA [6]	100.0	55.6	52.1	45.6	48.0	39.8	39.7	90.3
Trajectory Encoding Ablation (Qwen3-VL-8B [2])								
- Summary	100.0	63.6	54.7	52.7	43.4	41.1	89.7	90.9
- Timeseries (Kinematics)	100.0	69.7	62.2	61.9	<u>76.7</u>	<u>77.9</u>	92.6	90.6
- Coordinates (Spatial)	100.0	55.3	46.6	43.0	49.9	48.9	**97.9**	62.2
- Full (Times. + Coord.)	100.0	69.0	61.3	60.4	**77.3**	**78.4**	<u>97.8</u>	78.1

VL-8B, replacing visual frames entirely with trajectory text yields a BAcc of 59.6%, a significant +20.7pp surge over the vision-only baseline (38.9%). Reintroducing visual frames to this text-only baseline recovers a negligible +2.6pp gain under the best encoding strategy. Furthermore, with suboptimal encodings, performance regresses *entirely* below the text-only baseline. This reveals a functional decoupling in current architectures: **ego-motion logic is derived almost exclusively from the language modality**, while visual observations serve as redundant or even interfering signals that the reasoning core fails to integrate.

(iii) Consistency Relies on Static Visual Context. WPCR rises sharply from 20.0 with no visual input to 97.4 with a single static frame, but increases negligibly when additional frames are provided. This shows that physical consistency in model predictions is driven by the presence of any visual context, not by temporal reasoning over the frame sequence. Adding explicit trajectory text further improves WPCR, but reduces the contribution of visual input to near zero, confirming that motion reasoning is routed almost exclusively through the language modality.

(iv) The Encoding Advantage. Structured kinematic data consistently outperforms high-level semantic summaries across all models. With optimized encoding (Timeseries), smaller open-source models match or exceed closed-source MLLMs on temporal and consistency metrics, suggesting that representation quality is a more significant performance driver than parameter scale as is. Consequently, our conclusions suggest that simply scaling model size is insufficient for embodied tasks. Instead, future research must prioritize developing stronger physical alignment strategies during pre-training to bridge this vision-language gap.

5.3 Isolating the Domain Gap

To verify that our results are driven by ego-motion complexity rather than a simulation-to-reality gap or style-transfer artifacts, we conduct a controlled ablation across three visual domains: (i) real-world data (*nuScenes*), (ii) raw synthetic data, and (iii) style-transferred synthetic data used in these experiments.

As shown in Table 5, a representative VLM and baseline performance remain consistent across all three domains. This invariance indicates that the observed difficulties stem from a fundamental deficit in ego-motion reasoning rather than visual domain shifts. Notably, the stable performance of our geometric baselines across real and style-transferred sequences further supports the use of our synthetic data for assessing real-world ego-motion understanding.

Natural Visual-Artifact Ablation. 80 of 500 CARLA-transferred clips (16%) contain spatial artifacts from upstream CARLA rendering. As these are temporally stable within each clip, optical flow is preserved while photometric quality is degraded. Per-clip accuracy on this subset differs from that of the other 420 by ≤ 3 pp across leaderboard models, with mixed direction, further confirming the perception bottleneck: photometric quality is not meaningfully exploited. The subset list is released for downstream studies.

Table 5: Comparison of balanced accuracy for VLM and Baselines across different data sources. Best values are in **bold**; second-best are <u>underlined.</u> [1] Representative VLM comparison point.

Model	Real ↑	Sim ↑	Transf. ↑	Δ max ↓
Geometric Baselines				
Flow Heuristic [13, 22]	49.3	39.0	44.4	10.3
Visual Odometry [11, 23, 30]	<u>63.5</u>	**62.4**	**64.0**	<u>1.6</u>
RAFT Flow [35]	**65.5**	<u>60.5</u>	54.6	10.9
TartanVO [39]	51.3	51.0	<u>56.0</u>	5.0
VLM Baselines				
Qwen3-VL-8B[1][2]	40.0	39.7	41.0	**1.3**

6 Conclusion

We introduced *EgoDyn-Bench* to evaluate physical ego-motion understanding in vision-centric foundation models. Our audit of 20+ models reveals a consistent and severe **Perception Bottleneck**: despite possessing physically consistent internal reasoning, current VLMs and VLAs fail to ground it in visual observations, frequently lagging behind classical non-learned geometric baselines, a deficit that persists independently of model scale, domain-specific training, and visual domain. When explicit kinematic encodings are provided, performance recovers substantially across all models. This, however, exposes a **structural asymmetry**: ego-motion understanding is derived almost exclusively from the language modality, with visual observations contributing negligible temporal signal. This functional disentanglement between vision and language is the central architectural failure *EgoDyn-Bench* diagnoses. Resolving it, through native alignment between dynamic representations and visual perception during pre-training, is the critical open challenge for physically grounded embodied AI.

Limitations.*EgoDyn-Bench* targets ego-motion understanding over short horizons (3 s), where individual maneuvers are cleanly separable and attributable, extending to long-horizon and multi-agent scene-level reasoning is a natural next step. As a grounding diagnostic, it isolates whether models align physical concepts with visual observation rather than measuring closed-loop driving performance.

Future Work. To address the pure reliance on language modality to understand the motion kinematics, we aim to examine specific kinematic encoding paired with explicit alignment strategies for enhancement of vision-centric foundation models.

References

1. Anthropic: Claude sonnet 4.5 model card. `https://www.anthropic.com/news/claude-sonnet-4-5` (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), `https://arxiv.org/abs/2511.21631`
3. Bergasa, L.M., Almería, D., Almazán, J., Yebes, J.J., Arroyo, R.: Drivesafe: An app for alerting inattentive drivers and scoring driving behaviors. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. pp. 240–245 (2014). `https://doi.org/10.1109/IVS.2014.6856461`
4. Caesar, H., et al.: nuscenes: A multimodal dataset for autonomous driving (2020), `https://arxiv.org/abs/1903.11027`

5. Chen, Y., Zhan, Z., Lin, X., Song, Z., Liu, H., Lyu, Q., Zu, Y., Chen, X., Liu, Z., Pu, T., Chen, T., Wang, K., Lin, L., Wang, G.: Radar: Benchmarking vision-language-action generalization via real-world dynamics, spatial-physical intelligence, and autonomous evaluation (2026), https://arxiv.org/abs/2602.10980
6. Chi, H., ang Gao, H., Liu, Z., Liu, J., Liu, C., Li, J., Yang, K., Yu, Y., Wang, Z., Li, W., Wang, L., Hu, X., Sun, H., Zhao, H., Zhao, H.: Impromptu vla: Open weights and open data for driving vision-language-action models (2025), https://arxiv.org/abs/2505.23757
7. Dosovitskiy, A., et al.: Carla: An open urban driving simulator (2017), https://arxiv.org/abs/1711.03938
8. Gao, Y., Hua, D., Piccinini, M., Schäfer, F.R., Moller, K., Li, L., Betz, J.: Stylevla: Driving style-aware vision language action model for autonomous driving (2026), https://arxiv.org/abs/2603.09482
9. Gholami, M., Rezaei, A., Weimin, Z., Mao, S., Zhou, S., Zhang, Y., Akbari, M.: Spatial reasoning with vision-language models in ego-centric multi-view scenes (2025), https://arxiv.org/abs/2509.06266
10. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering (2017), https://arxiv.org/abs/1612.00837
11. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, 2 edn. (2004)
12. Heilmeier, A., Wischnewski, A., Hermansdorfer, L., Betz, J., Lienkamp, M., Lohmann, B.: Minimum curvature trajectory planning and control for an autonomous race car. Vehicle System Dynamics **58**(10), 1497–1527 (2020). https://doi.org/10.1080/00423114.2019.1631455, https://doi.org/10.1080/00423114.2019.1631455
13. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial Intelligence **17**(1), 185–203 (1981). https://doi.org/https://doi.org/10.1016/0004-3702(81)90024-2, https://www.sciencedirect.com/science/article/pii/0004370281900242
14. Huang, Z., Feng, C., Yan, F., Xiao, B., Jie, Z., Zhong, Y., Liang, X., Ma, L.: Robotron-drive: All-in-one large multimodal model for autonomous driving (2025), https://arxiv.org/abs/2412.07689
15. Jiang, S., Huang, Z., Qian, K., Luo, Z., Zhu, T., Zhong, Y., Tang, Y., Kong, M., Wang, Y., Jiao, S., Ye, H., Sheng, Z., Zhao, X., Wen, T., Fu, Z., Chen, S., Jiang, K., Yang, D., Choi, S., Sun, L.: A survey on vision-language-action models for autonomous driving (2025), https://arxiv.org/abs/2506.24044
16. Johnson, D.A., Trivedi, M.M.: Driving style recognition using a smartphone as a sensor platform. In: 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC). pp. 1609–1615 (2011). https://doi.org/10.1109/ITSC.2011.6083078
17. Karnchanachari, N., Geromichalos, D., Tan, K.S., Li, N., Eriksen, C., Yaghoubi, S., Mehdipour, N., Bernasconi, G., Fong, W.K., Guo, Y., Caesar, H.: Towards learning-based planning:the nuplan benchmark for real-world autonomous driving (2024), https://arxiv.org/abs/2403.04133
18. KENDALL, M.G.: A new measure of rank correlation. Biometrika **30**(1-2), 81–93 (06 1938). https://doi.org/10.1093/biomet/30.1-2.81, https://doi.org/10.1093/biomet/30.1-2.81
19. Klischat, M., Althoff, M.: Generating critical test scenarios for automated vehicles with evolutionary algorithms. In: Proc. of the IEEE Intelligent Vehicles Symposium. pp. 2352 – 2358 (2019). https://doi.org/10.1109/ivs.2019.8814230

20. Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Kolter, J.Z., Langer, D., Pink, O., Pratt, V., Sokolsky, M., Stanek, G., Stavens, D., Teichman, A., Werling, M., Thrun, S.: Towards fully autonomous driving: Systems and algorithms. In: 2011 IEEE Intelligent Vehicles Symposium (IV). pp. 163–168 (2011). `https://doi.org/10.1109/IVS.2011.5940562`
21. Liu, J., Zhou, J., Ye, K., Lin, K.Y., Wang, A., Liang, J.: Egotraj-bench: Towards robust trajectory prediction under ego-view noisy observations (2025), `https://arxiv.org/abs/2510.00405`
22. Longuet-Higgins, H.C.: A computer algorithm for reconstructing a scene from two projections. Nature **293**(5828), 133–135 (1981)
23. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI'81: 7th international joint conference on Artificial intelligence. vol. 2, pp. 674–679 (1981)
24. Nie, M., Peng, R., Wang, C., Cai, X., Han, J., Xu, H., Zhang, L.: Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving (2024), `https://arxiv.org/abs/2312.03661`
25. NVIDIA, :, Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., Dworakowski, D., Fan, J., Fenzi, M., Ferroni, F., Fidler, S., Fox, D., Ge, S., Ge, Y., Gu, J., Gururani, S., He, E., Huang, J., Huffman, J., Jannaty, P., Jin, J., Kim, S.W., Klár, G., Lam, G., Lan, S., Leal-Taixe, L., Li, A., Li, Z., Lin, C.H., Lin, T.Y., Ling, H., Liu, M.Y., Liu, X., Luo, A., Ma, Q., Mao, H., Mo, K., Mousavian, A., Nah, S., Niverty, S., Page, D., Paschalidou, D., Patel, Z., Pavao, L., Ramezanali, M., Reda, F., Ren, X., Sabavat, V.R.N., Schmerling, E., Shi, S., Stefaniak, B., Tang, S., Tchapmi, L., Tredak, P., Tseng, W.C., Varghese, J., Wang, H., Wang, H., Wang, H., Wang, T.C., Wei, F., Wei, X., Wu, J.Z., Xu, J., Yang, W., Yen-Chen, L., Zeng, X., Zeng, Y., Zhang, J., Zhang, Q., Zhang, Y., Zhao, Q., Zolkowski, A.: Cosmos world foundation model platform for physical ai (2025), `https://arxiv.org/abs/2501.03575`
26. OpenAI: Gpt-5.1. `https://openai.com/gpt-5` (2025)
27. Puyin, L., Xiang, T., Mao, E., Wei, S., Chen, X., Masood, A., Fei-fei, L., Adeli, E.: Quantiphy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models (2025), `https://arxiv.org/abs/2512.19526`
28. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario (2024), `https://arxiv.org/abs/2305.14836`
29. Research, N., et al.: Cosmos-transfer1: Conditional world generation with adaptive multimodal control (2025), `https://arxiv.org/abs/2503.14492`
30. Shi, J., Tomasi: Good features to track. In: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 593–600 (1994). `https://doi.org/10.1109/CVPR.1994.323794`
31. Sima, C., et al.: Drivelm: Driving with graph visual question answering (2025), `https://arxiv.org/abs/2312.14150`
32. Stahl, T., Wischnewski, A., Betz, J., Lienkamp, M.: Multilayer graph-based trajectory planning for race vehicles in dynamic scenarios. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). pp. 3149–3154 (2019). `https://doi.org/10.1109/ITSC.2019.8917032`
33. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhao, S., Cheng, S., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. arXiv (2019)

34. Team, G., et al.: Gemini: A family of highly capable multimodal models (2025), https://arxiv.org/abs/2312.11805
35. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow (2020), https://arxiv.org/abs/2003.12039
36. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models (2024), https://arxiv.org/abs/2402.12289
37. Trauth, R., Moller, K., Würsching, G., Betz, J.: Frenetix: A high-performance and modular motion planning framework for autonomous driving. IEEE Access pp. 1–1 (2024). https://doi.org/10.1109/ACCESS.2024.3436835
38. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), https://arxiv.org/abs/2508.18265
39. Wang, W., Hu, Y., Scherer, S.: Tartanvo: A generalizable learning-based vo (2020), https://arxiv.org/abs/2011.00359
40. Wang, Y., Zheng, Y., Fan, W., Wang, T., Chu, H., Tian, D., Gao, B., Wang, J., Chen, H.: Scenepilot-bench: A large-scale dataset and benchmark for evaluation of vision-language models in autonomous driving (2026), https://arxiv.org/abs/2601.19582
41. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., Ramanan, D., Carr, P., Hays, J.: Argoverse 2: Next generation datasets for self-driving perception and forecasting (2023), https://arxiv.org/abs/2301.00493
42. Wu, H., Cai, Y., Li, Z., Ge, H., Sun, B., Yuan, J., Wang, Y.: Camreasoner: Reinforcing camera movement understanding via structured spatial reasoning (2026), https://arxiv.org/abs/2602.00181
43. Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q.A., Liu, Z., Pan, L.: Are vlms ready for autonomous driving? an empirical study from the reliability, data and metric perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6585–6597 (October 2025)
44. Xing, Y., Lv, C., Cao, D.: Personalized vehicle trajectory prediction based on joint time-series modeling for connected vehicles. IEEE Transactions on Vehicular Technology **69**(2), 1341–1352 (2020). https://doi.org/10.1109/TVT.2019.2960110
45. Xu, Z., et al.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model (2024), https://arxiv.org/abs/2310.01412
46. Zhang, C., Cherniavskii, D., Tragoudaras, A., Vozikis, A., Nijdam, T., Prinzhorn, D.W.E., Bodracska, M., Sebe, N., Zadaianchuk, A., Gavves, E.: Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments (2025), https://arxiv.org/abs/2504.02918
47. Zhou, X., Liu, M., Yurtsever, E., Zagar, B.L., Zimmer, W., Cao, H., Knoll, A.C.: Vision language models in autonomous driving: A survey and outlook. IEEE Transactions on Intelligent Vehicles pp. 1–20 (2024). https://doi.org/10.1109/TIV.2024.3402136

48. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., Gao, Z., Cui, E., Wang, X., Cao, Y., Liu, Y., Wei, X., Zhang, H., Wang, H., Xu, W., Li, H., Wang, J., Deng, N., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Lv, H., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models (2025), https://arxiv.org/abs/2504.10479