VISION-LANGUAGE MODELS CANNOT TELL WHAT MOVES WITHOUT KNOWING HOW THEY MOVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models for driving are asked whether a nearby vehicle is moving or stationary, and the usual remedy for their errors is to show them more frames. We find that more frames do not help. A BLIP-2 model finetuned on nuScenes scores 62.1% from one frame and 62.9% with two past frames fused through gated cross-attention, and three open vision-language models change by -3.8 to $+2.7$ points when given three frames instead of one. We prove that this failure concerns information and not architecture. The motion state of an object is not identifiable from its positions in the camera frame, because that frame mixes the motion of the object with the apparent motion that the observer causes, and it becomes identifiable once the relative pose of the camera between frames is known. On nuScenes, object displacement separates the two classes with an AUC of 0.557 in the ego frame and 0.935 after ego-motion compensation, and one token that carries the compensated displacement raises the finetuned model to $88.2 \pm 0.3\%$, or $88.1 \pm 0.4\%$ with the image blacked out. How the quantity enters the model matters as much as whether it is present: the same twelve numbers give 66.2% through a subtractive structure that cannot represent rotation and 76.2% with auxiliary supervision of that structure, whereas a module that builds the exact group action in reaches $88.1 \pm 0.2\%$, and stating the compensated displacement in the prompt of an open model without training adds between 0 and 10 points. A model on a moving platform needs its own motion as an input and the compensation as a structure, not a longer clip.

1 INTRODUCTION

A camera on a moving car records every object in the car's own frame. A parked truck slides across that frame when the car drives past it, and a truck that keeps pace with the car stays still. Biology met this problem long ago. Under the reafference principle, an animal predicts the sensory change that its own movement will cause and subtracts the prediction, so that the remainder is attributed to the world (von Holst & Mittelstaedt, 1950; Sperry, 1950; Wolpert et al., 1995). Human observers parse optic flow in this way when they judge whether an object moves while they move themselves (Warren & Rushton, 2009). Geometry states the same fact. The image motion of a rigid scene is a function of the observer's translation and rotation, and the motion of an object is what remains after that function is removed (Longuet-Higgins & Prazdny, 1980; Hartley & Zisserman, 2004).

Vision-language models (VLMs) now answer questions about driving scenes. Models that connect an image encoder to a language model (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; Driess et al., 2023) are trained and evaluated on driving question answering (Qian et al., 2024; Sima et al., 2024; Xu et al., 2024). One of the most basic questions in these benchmarks is whether an object is moving or stationary. A single image leaves this question open, so the natural remedy is temporal context: past frames, or a video encoder (Bertasius et al., 2021). Reports on the visual side of VLMs give reason for doubt. VLMs miss visual patterns that are plain to people (Tong et al., 2024; Fu et al., 2024), and they reason poorly about metric space (Chen et al., 2024; Yang et al., 2025; Gholami et al., 2026). Video-language benchmarks are often solved from one frame (Buch et al., 2022; Lei et al., 2023) or from priors in the question (Goyal et al., 2017; Geirhos et al., 2020). Closest to our setting, Schäfer et al. (2026) show that VLMs do not read the ego-motion of the vehicle from video. In their benchmark, ego-motion answers improve by 20.7 points when the trajectory is given as text. That study concerns the motion of the observer, and it diagnoses without retraining.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

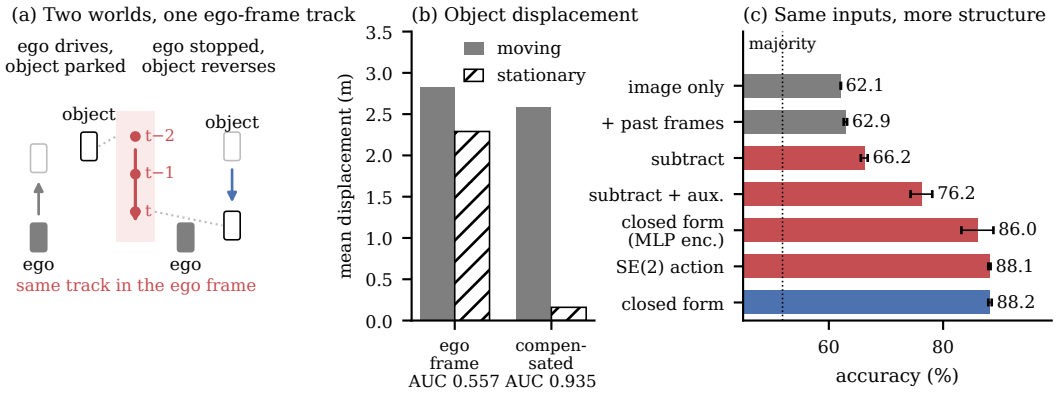


Figure 1: **More frames do not reveal motion; ego-motion does.** (a) A parked object seen from a driving car and a reversing object seen from a stopped car trace the same path in the camera frame, so the path does not determine the motion state (Proposition 1). (b) On nuScenes, a stationary vehicle is displaced by 2.29 m on average in the ego frame and by 0.16 m after ego-motion compensation; displacement separates moving from stationary vehicles with an AUC of 0.557 before compensation and 0.935 after. (c) Accuracy on 372 class-balanced test questions (majority class 51.9%), mean and standard deviation over seeds. Past frames leave BLIP-2 where one frame leaves it (62.1% and 62.9%). Given ego-frame positions and relative ego pose, accuracy depends on the structure through which they enter: 66.2% for a subtraction that ignores rotation, 76.2% with auxiliary supervision of it, $86.0 \pm 2.8\%$ for the closed-form displacement through an MLP encoder, $88.1 \pm 0.2\%$ for the SE(2) action with a learned residual, and $88.2 \pm 0.3\%$ for the closed-form compensation with the residual frozen.

These reports leave one question unanswered. When a VLM fails to tell a moving object from a stationary one, does it lack visual evidence, or does it lack a quantity that the visual evidence about the object does not contain? The two answers call for different research. The first calls for more frames and stronger video encoders. The second calls for a different input.

We begin with the first answer and find that it fails (Figure 1c). We attach a temporal module to BLIP-2: cross-attention over the frames at $t{-}2$ and $t{-}1$ behind a $\tanh$ gate, in the manner of Alayrac et al. (2022). We finetune it with LoRA (Hu et al., 2022) on nuScenes (Caesar et al., 2020). On 372 class-balanced motion questions the temporal model scores 62.9% and the single-frame model 62.1%. Three open VLMs that we prompt without training change by -3.8, -1.1 and $+2.7$ points when they receive three frames instead of one, and none of the four differences excludes zero. The single-frame scores lie above the majority rate of 51.9%, so an image does carry evidence: a vehicle in a parking lane is likely parked, and a vehicle in a traffic lane is likely moving. What the extra frames add is nothing. The evidence covers four models of three families, prompted and finetuned, and in no case does the third frame change the accuracy.

We then prove that the second answer is correct (Section 2). Let $\boldsymbol{x}_t$ be the position of the object in the world, $g_t \in \mathrm{SE}(2)$ the pose of the vehicle, and $\boldsymbol{o}_t = g_t^{-1}\boldsymbol{x}_t$ the position that the camera frame records. Proposition 1 states that every sequence $\boldsymbol{o}_{t-k}, \ldots, \boldsymbol{o}_t$ is consistent both with a world in which the object is stationary and with a world in which it moves. Proposition 2 states that the displacement of the object equals $\|\boldsymbol{o}_t - \Delta g\, \boldsymbol{o}_{t-1}\|$, where $\Delta g = g_t^{-1} g_{t-1}$ is the relative pose over the interval. In other words, a longer sequence of observations does not settle the question, whereas one relative pose per interval does. The global position of the vehicle cancels, so odometry is enough and localization is not required. The data agree (Figure 1b). A stationary vehicle is displaced by 2.29 m on average in the ego frame, against 2.83 m for a moving one, and displacement separates the classes with an AUC of 0.557. After compensation the averages are 0.16 m and 2.59 m, and the AUC is 0.935.

The scope of Proposition 1 is the observation of the object. A full image also shows the static background, from which ego-motion can be estimated. Visual odometry and SLAM do so (Mur-Artal et al., 2015; Teed & Deng, 2020; 2021), and methods that split optical flow into a rigid part

and a residual segment moving objects from a moving camera (Menze & Geiger, 2015; Yin & Shi, 2018; Ranjan et al., 2019; Yang & Ramanan, 2021). Object motion is thus identifiable from images in principle, through ego-motion as an intermediate quantity. Our results indicate that the VLMs we test do not extract that quantity, which matches the finding of Schäfer et al. (2026) that classical visual odometry outperforms VLMs on ego-motion. Temporal 3D detectors for driving take the same view in practice and align past features with the ego pose before they fuse them (Li et al., 2022; Park et al., 2023). VLM pipelines for driving have no counterpart of this step.

The identified quantity is sufficient. We map the compensated displacement to one token and place it beside the image tokens of BLIP-2, as embodied language models do with state vectors (Driess et al., 2023). Accuracy rises from 62.1% to $88.2 \pm 0.3\%$ over three seeds. It stays at $88.1 \pm 0.4\%$ when the image is blacked out, so the token carries the answer. It falls to $74.2 \pm 1.2\%$ when the token carries the ego-frame displacement instead, so the compensation carries the token. A threshold on the compensated displacement, with no language model at all, reaches 88.4%.

How the quantity enters the model matters as much as whether it is present, and this is the second finding of the paper. We give the model the ingredients of Proposition 2, the ego-frame positions at three time steps and the relative poses between them, and vary only the structure through which they are combined. A module that subtracts a pose-predicted ego displacement from the observed one reaches $66.2 \pm 0.6\%$, and $76.2 \pm 1.9\%$ when the recovered displacement is regressed onto the true one during training, a use of privileged information in the sense of Vapnik & Vashist (2009). Corollary 1 explains why this structure is capped: the apparent displacement of a stationary object depends on its position when the vehicle rotates, with a term of norm $2\sin(|\Delta\theta|/2)\,\|\boldsymbol{o}_{t-1}\|$, which a pose-only subtraction cannot represent. On test questions with an ego yaw change above 3° the subtractive module falls to 64.4% while the closed form rises to 93.3%. A module that applies the group action of Proposition 2 and learns only a bounded residual reaches $88.1 \pm 0.2\%$, and once the compensation is exact the encoder matters little: the same quantity through a plain MLP encoder gives $86.0 \pm 2.8\%$. The wrong structure costs 12 to 22 points. Information that a bounded learner cannot use is not usable information (Xu et al., 2020b), and a network learns a computation reliably when its structure matches that computation (Xu et al., 2020a), in the sense in which equivariant networks build a group action into their layers (Cohen & Welling, 2016). The same holds at the input: stating the compensated displacement in the prompt of an open VLM without training adds 5.9 and 9.7 points to two models and nothing to a third, whereas the trained token adds 26.

Our result connects to three lines of work. Work on identifiability shows that factors of variation cannot be separated without side information or inductive bias (Locatello et al., 2019; Khemakhem et al., 2020). Ego-motion is such side information, and here the identifying map is known in closed form. Work on world models separates the change that an agent causes from the change that is exogenous to it (Ha & Schmidhuber, 2018; Efroni et al., 2022; Wang et al., 2022). The apparent motion of a parked car is the agent-caused part of a driving scene. Work on representation learning has used ego-motion as a training signal for visual features (Jayaraman & Grauman, 2015; Agrawal et al., 2015). We use it as an input at test time and show that the question is otherwise undetermined.

Both inputs are available on a vehicle. The relative pose over 0.5 s comes from wheel and inertial odometry, whose error grows with distance travelled and is small over such an interval (Kuutti et al., 2018). Under additive pose noise of 0.5 m and 2° per interval, well above what such odometry produces, the closed form keeps 78.2% and the structured module 76.9% (Section 4.4). Object positions come from camera 3D detectors, whose mean translation error on nuScenes is 0.6 to 0.8 m (Simonelli et al., 2019; Wang et al., 2021a;b; Chen et al., 2022); our experiments use annotated positions, and we discuss this limit in Section 6.

In summary, we show that temporal context does not improve the motion-state answers of a finetuned VLM or of three open VLMs, and we trace the failure to the input and not to the encoder. We prove that motion state is not identifiable from ego-frame observations and is identifiable from them together with relative ego pose, and we confirm the sufficiency with a single token that reaches 88.2%. We show that the structure through which this quantity enters the model decides the outcome, from 66.2% for a structure that Corollary 1 shows to be misspecified to 88.1% for the exact group action. We regard the identifiability claim as settled, since it is a theorem, the temporal claim as established for four models on one dataset, and the structural claim as established for one finetuned backbone.

2 IDENTIFIABILITY OF OBJECT MOTION UNDER EGO-MOTION

We work in the ground plane, which is the setting of the nuScenes annotations; roll and pitch enter only through the projection to the plane and are not needed for the statements below. The object has world position $\boldsymbol{x}_t \in \mathbb{R}^2$, the vehicle has pose $g_t = (\boldsymbol{R}_t, \boldsymbol{c}_t) \in \mathrm{SE}(2)$, and the ego-frame observation is as follows:

$$\boldsymbol{o}_t = g_t^{-1}\boldsymbol{x}_t = \boldsymbol{R}_t^\top(\boldsymbol{x}_t - \boldsymbol{c}_t). \tag{1}$$

The motion state over k steps is $y = \mathbf{1}\left[\|\boldsymbol{x}_t - \boldsymbol{x}_{t-k}\| > \tau\right]$ for a threshold $\tau > 0$.

Proposition 1 (Non-identifiability)**.** *For every observation sequence $(\boldsymbol{o}_{t-k}, \ldots, \boldsymbol{o}_t)$ and every $y \in \{0, 1\}$, there exist poses (g_s) and world positions $(\boldsymbol{x}_s)$ that satisfy equation 1 and have motion state y.*

That is, the ego-frame track of an object fixes nothing about its motion state; what a predictor can extract from the track is the prior over ego-motion in the data. The statement is a worst case over poses, but restricting the poses does not remove the ambiguity: the construction in the proof uses only a translation of norm at most $\|\boldsymbol{o}_t - \boldsymbol{o}_{t-k}\| + \tau$, so the two worlds remain admissible under any kinematic limit that allows the vehicle to travel that far over the interval, which is exactly what a vehicle does whenever the observed track was produced by its own motion. Section 4.2 measures how much of the ambiguity remains in driving data, where the prior over ego-motion is informative.

Proposition 2 (Identifiability from relative pose)**.** *Let $\Delta g = g_t^{-1} g_{t-1} = (\Delta\boldsymbol{R}, \Delta\boldsymbol{c})$. Then the displacement satisfies the following identity:*

$$\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\| = \|\boldsymbol{o}_t - \Delta g\, \boldsymbol{o}_{t-1}\| = \|\boldsymbol{o}_t - \Delta\boldsymbol{R}\, \boldsymbol{o}_{t-1} - \Delta\boldsymbol{c}\|. \tag{2}$$

The right-hand side is invariant to a common rigid transformation of all poses, and for $k > 1$ the relative poses compose, $\Delta g_{t-k\to t} = \Delta g_t \cdots \Delta g_{t-k+1}$, so the identity holds over any number of steps.

Corollary 1 (Subtraction is exact only without rotation)**.** *The apparent displacement of a stationary object is $(\Delta\boldsymbol{R} - \boldsymbol{I})\, \boldsymbol{o}_{t-1} + \Delta\boldsymbol{c}$. It is a function of Δg alone if and only if $\Delta\boldsymbol{R} = \boldsymbol{I}$; the position-dependent term has norm $2\sin(|\Delta\theta|/2)\, \|\boldsymbol{o}_{t-1}\|$.*

For a vehicle at 30 m and a yaw change of 5° the position-dependent term is 2.6 m, the size of the displacement of a moving vehicle over the interval. Corollary 1 therefore predicts that a model which subtracts a pose-only estimate will err on turns, and Section 4.3 tests this prediction. Proofs are in Appendix A.

3 MODELS: THE SAME INPUTS THROUGH DIFFERENT STRUCTURES

Task. The input is a camera image I_t from the camera that the question names, optionally the two previous keyframes I_{t-1}, I_{t-2} at 0.5 s spacing, and a question that refers to one vehicle. The output is one of two answers, *moving* or *stationary*. The label comes from the nuScenes attribute annotation: `vehicle.moving` is *moving*, and `vehicle.stopped` and `vehicle.parked` are *stationary*. The label is thus an annotation of state and not a threshold on displacement; Section 4.2 measures how well displacement predicts it.

Side inputs. Models below the first two rungs also receive the ego-frame positions $\boldsymbol{o}_{t-2}, \boldsymbol{o}_{t-1}, \boldsymbol{o}_t \in \mathbb{R}^2$ of the vehicle in metres and the relative poses $\Delta g_{t-1}, \Delta g_t$, each given as a translation and a yaw change. These twelve numbers are the ingredients of equation 2 and nothing else.

Backbone. All trained models share BLIP-2 with a Flan-T5-XL language model (Li et al., 2023; Chung et al., 2024) and LoRA adapters. Each side-input model maps its inputs to one vector in the token space of the language model and places it beside the image tokens. The rungs differ only in how that vector is computed.

- **Image only** and **Image + past frames.** No side input. The second model attends to past frames through gated cross-attention.

- **Subtract.** The module computes the observed ego-frame displacement $\boldsymbol{o}_t - \boldsymbol{o}_{t-1}$, predicts the ego-induced displacement from the pose alone with a learned map, subtracts, and encodes the difference. By Corollary 1 this structure is exact only when $\Delta\boldsymbol{R} = \boldsymbol{I}$.
- **Subtract + aux.** The same module with an auxiliary loss $\lambda\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|^2$ between the recovered displacement $\hat{\boldsymbol{u}}$ and the true world-frame displacement $\boldsymbol{u}$, with $\lambda = 0.02$. The target $\boldsymbol{u}$ is used in training only.
- **SE(2)-act.** The module applies the group action of equation 2, $\hat{\boldsymbol{u}} = \boldsymbol{o}_t - \boldsymbol{R}(\Delta\theta + \delta_\theta)\,\boldsymbol{o}_{t-1} - (\Delta\boldsymbol{c} + \boldsymbol{\delta}_c)$. The residuals are the bounded output of a small network, $\delta_\theta = 3^\circ \tanh(\cdot)$ and $\boldsymbol{\delta}_c = 0.3\,\mathrm{m}\,\tanh(\cdot)$, they are zero at initialization, their squared norm is penalized, and the residual head is trained at one tenth of the learning rate after one epoch in which it is frozen. The auxiliary loss above is used with $\lambda = 0.5$, chosen on the validation split. The structure is exact at initialization, where the recovered displacement has an RMSE of 0.045 m against the annotation; learning can only correct for sensor error. Appendix D reports a variant without the bound, which leaves the exact solution during training.
- **Closed-form token.** The residual of SE(2)-act is frozen at zero, so the token encodes $\hat{\boldsymbol{u}} = \boldsymbol{o}_t - \Delta g\,\boldsymbol{o}_{t-1}$ with no learned component before the encoder. Two variants change only what is encoded: an **MLP encoder** maps the ego-compensated displacement of the annotated box over the two intervals (six numbers) through a three-layer perceptron, and an ablation encodes the uncompensated ego-frame displacement through the same MLP encoder.

We also report a **threshold rule** that uses no language model: it answers *moving* when $\|\hat{\boldsymbol{u}}\| > \tau$, with $\tau = 0.4$ m chosen on the validation split. This rule bounds what the binary task itself requires. Training details and cost are in Appendix B.

4 Experiments

Protocol. We take the motion-state questions of the DriveLM question set (Sima et al., 2024) on nuScenes (Caesar et al., 2020) keyframes; each question names one vehicle by its 2D box and camera. We split scenes into 482 for training, 84 for validation and 117 for testing, with no scene shared between splits. The test set has 400 class-balanced questions. Every number in this section is computed on the 372 questions for which all models have valid side inputs (193 moving, 179 stationary), so the majority-class rate is 51.9%. The 28 excluded questions failed the match between the 2D box of the question and a 3D annotation (Appendix B). Hyperparameters, λ and τ are selected on the 1,864 motion questions of the validation scenes. Trained models are run with three seeds (five for the MLP-encoded token) and reported as mean $\pm$ standard deviation; the image-only model has no stochastic component beyond initialization and converged to the same checkpoint in three runs, so it is reported once. Intervals are 95% Wilson intervals.

4.1 More frames do not help

Claim: temporal visual context does not improve motion-state answers, in a finetuned model or in open VLMs.

Table 1 compares one frame with three. For finetuned BLIP-2 the gain from two past frames is 0.8 points, three questions in 372. For the three open models the changes are -1.1, $+2.7$ and -3.8 points, and for all four models the paired interval on the difference contains zero. The single-frame scores are not at chance. The finetuned model reaches 62.1% and LLaVA-OneVision 72.8% from one image, which is the evidence that a single image plausibly carries: lane position, orientation to the road, and brake lights are correlated with the motion state. Frames add nothing to that evidence, in agreement with the single-frame bias reported for video-language benchmarks (Buch et al., 2022; Lei et al., 2023). The last column previews Section 4.3: when the compensated displacement is stated in the prompt, two open models gain 5.9 and 9.7 points and one gains nothing, although its predictions change on 28 questions; the same quantity given to the finetuned model as a trained token gains 26, and stating the relative ego pose alone, without the compensation, gains nothing (Appendix C). The two routes differ in training as well as in form, so the comparison does not isolate the form of the input; what it shows is that an open VLM does not use the quantity when it is merely stated, which is the point of Section 4.3. Prompts and the paired intervals are in Appendix C,

Table 1: Temporal context does not improve accuracy (%) on the 372 test questions; majority class 51.9%. Top: BLIP-2 finetuned on nuScenes. Bottom: open VLMs prompted without training, given one frame, three frames, or three frames plus a sentence that states the compensated displacement of the vehicle in metres. Δ: three frames minus one frame; for all four models the paired 95% interval on $n = 372$ contains zero (Appendix C).

Model	1 frame	3 frames	Δ	3 frames + comp. displacement
BLIP-2, finetuned	62.1	62.9 ± 0.3	$+0.8$	88.2 ± 0.3 (trained token)
Qwen2.5-VL (Bai et al., 2025)	50.8	49.7	-1.1	56.7 (as text)
InternVL3 (Zhu et al., 2025)	65.6	68.3	$+2.7$	75.3 (as text)
LLaVA-OneVision (Li et al., 2025)	72.8	69.1	-3.8	69.1 (as text)

Table 2: Displacement of the vehicle over the question interval. The first two rows use the nuScenes annotations of all test questions with a matched 3D box; the last row recomputes the compensated displacement from the twelve side inputs through equation 2 on the 372 common questions. Compensation leaves moving vehicles unchanged and removes the apparent motion of stationary ones.

Displacement	Moving (m)	Stationary (m)	AUC
ego frame, $\|\boldsymbol{o}_t - \boldsymbol{o}_{t-1}\|$	2.83	2.29	0.557
world frame, $\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|$	2.59	0.16	0.935
compensated from side inputs, equation 2	–	–	0.938

which also reports earlier runs in which the temporal gates of the finetuned model converged to $|\tanh| \approx 0.01$.

4.2 Ego-motion decides what the track means

Claim: the ambiguity of Proposition 1 is present in driving data, and the compensation of Proposition 2 removes it.

Table 2 shows that a stationary vehicle moves 2.29 m in the ego frame, 81% of the displacement of a moving one, and that displacement barely separates the classes (AUC 0.557). After compensation the AUC is 0.935. Computing the compensated displacement from the ego-frame positions and the relative poses through equation 2, rather than reading it from world-frame annotations, gives 0.938, which confirms that the twelve side inputs carry the full quantity. The threshold rule on this quantity, with $\tau = 0.4$ m from the validation split, answers 88.4% of the test questions; by construction its errors are vehicles annotated as moving whose displacement is below 0.4 m over the interval, or as stopped or parked while their displacement exceeds it.

4.3 The information is sufficient; the structure decides whether it is used

Claim: one compensated token solves the task; the same inputs through a misspecified structure do not.

Table 3 holds the central result. The closed-form token reaches $88.2 \pm 0.3\%$, and $88.1 \pm 0.4\%$ with the image blacked out: the answer is in the token, and the image adds nothing once the token is present. Encoding the uncompensated ego-frame displacement through the same MLP encoder gives $74.2 \pm 1.2\%$ against $86.0 \pm 2.8\%$ for the compensated one, so the compensation, not the token mechanism, produces the gain. The remaining rows receive every number that the closed form uses and differ only in structure. The subtractive module, whose structure Corollary 1 shows to be exact only without rotation, reaches $66.2 \pm 0.6\%$, and $76.2 \pm 1.9\%$ when its recovered displacement is supervised; a larger auxiliary weight does not close the gap (Appendix D). Figure 2 tests the corollary directly: the accuracy of the supervised subtractive module falls from 77.9% to 64.4% as the ego yaw change grows from below 1° to above 3°, while the closed form does not fall. SE(2)-act, which builds the action in and learns a bounded residual, reaches $88.1 \pm 0.2\%$ with a recovered-

Table 3: Same side inputs, different structures. Accuracy (%) on the 372 test questions, mean ± standard deviation over three seeds (five for the MLP-encoded token). RMSE: per-component error of the recovered displacement $\hat{\boldsymbol{u}}$ before the encoder, where the module exposes one.

Model	Compensation	Accuracy	RMSE of $\hat{\boldsymbol{u}}$ (m)
Image only	–	62.1	–
Image + past frames	–	62.9 ± 0.3	–
Ego-frame displ., MLP encoder (ablation)	none	74.2 ± 1.2	–
Subtract	pose-only	66.2 ± 0.6	–
Subtract + aux	pose-only, supervised	76.2 ± 1.9	–
Closed-form token, MLP encoder	exact	86.0 ± 2.8	0
Closed-form token	exact	88.2 ± 0.3	0.045
same, image blacked out	exact	88.1 ± 0.4	0.045
SE(2)-act	exact + bounded residual	88.1 ± 0.2	0.07
Threshold rule, no language model	exact	88.4	0.045

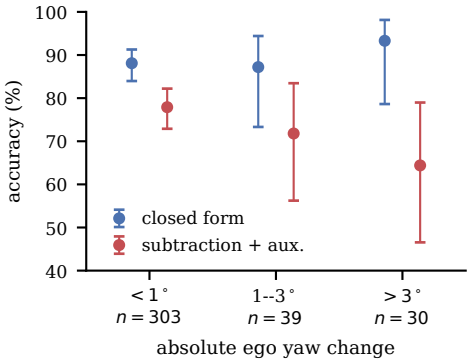


Figure 2: Test accuracy by absolute ego yaw change, with 95% Wilson intervals. The subtractive module errs where Corollary 1 places its error (77.9, 71.8, 64.4%); the closed form does not (88.1, 87.2, 93.3%). The upper bins hold 39 and 30 questions.

Table 4: Additive Gaussian pose noise per 0.5 s interval, translation σ_c and yaw σ_θ, on the relative pose. Closed form: threshold rule with τ reselected on validation data at the same noise level (fixed τ in Appendix E). SE(2)-act is trained and evaluated at that level, mean of three seeds. Accuracy (%), 372 test questions.

σ_c / σ_θ	Closed (τ)	SE(2)-act	Image
0 m / 0°	88.4 (0.40)	88.1	62.1
0.05 m / 0.5°	86.0 (0.50)	85.8	62.1
0.2 m / 1°	80.4 (0.60)	83.3	62.1
0.5 m / 2°	78.2 (1.30)	76.9	62.1

displacement RMSE of 0.07 m; on annotated poses the learned residual has nothing to correct, and the module reproduces the closed form. Once the compensation is exact, the encoder that carries it matters little: the same displacement through a plain MLP encoder gives 86.0 ± 2.8% over five seeds (83.6 to 89.3%), and the seed variance is that of the encoder, not of the answer. A misspecified structure therefore costs 12 to 22 points relative to the exact one; the gap is not a matter of capacity, since the frozen action adds no parameters where the subtractive modules add a learned map, and not a matter of supervision, since the auxiliary loss recovers less than half of it. The threshold rule reaches 88.4% without a language model. We report it because it bounds what the binary task alone requires: on this question set the language model adds nothing beyond the token, and Section 6 states what this does and does not imply. Earlier runs, an unbounded residual that leaves the exact solution, and an earlier implementation of the closed-form token are in Appendix D.

4.4 Sensor noise

Claim: with pose noise well above that of vehicle odometry, the closed form and the structured module stay far above the image-only model.

Relative pose over 0.5 s comes from wheel and inertial odometry, whose drift accumulates over hundreds of metres and is far below a metre over such an interval (Kuutti et al., 2018). Table 4 adds Gaussian noise to the relative pose up to 0.5 m and 2°, well above that level. The closed form keeps 86.0% at the realistic level and 78.2% at the largest, and SE(2)-act keeps 85.8% and 76.9%; both stay

15 to 26 points above the image-only model. The threshold must grow with the noise, from 0.4 m to 1.3 m, because noise on the relative pose passes through equation 2 into the recovered displacement of a stationary vehicle. A threshold fixed on noise-free data falls to 57.8% at the largest noise level for that reason, which is a property of the threshold and not of the compensation (Appendix E). SE(2)-act is retrained at each noise level, so it sees the noise distribution that the closed form only meets through its reselected threshold; even so, the learned residual does not recover what the closed form loses, as expected for zero-mean noise that is independent of the inputs. Object positions in all experiments are annotated; the mean translation error of camera 3D detectors on nuScenes is 0.6 to 0.8 m (Simonelli et al., 2019; Wang et al., 2021a;b; Chen et al., 2022), and we have not evaluated detector outputs (Section 6).

5 Related work

VLM perception and single-frame bias. Multimodal language models miss visual details that people see at once (Tong et al., 2024; Fu et al., 2024) and reason poorly about metric and egocentric space (Chen et al., 2024; Yang et al., 2025; Gholami et al., 2026). Video-language models are often no better than their single-frame versions (Buch et al., 2022; Lei et al., 2023), and question answering is prone to priors in the question (Goyal et al., 2017) and to shortcuts in general (Geirhos et al., 2020). Schäfer et al. (2026) find that VLMs read ego-motion from text and not from video. We add the object-level counterpart of that finding, and we give the reason: without ego-motion the object-level question is not identifiable, so no encoder can answer it from the object alone.

Driving question answering. NuScenes-QA (Qian et al., 2024), DriveLM (Sima et al., 2024) and DriveGPT4 (Xu et al., 2024) pose motion-state questions to VLMs from images alone, and EgoDyn-Bench (Schäfer et al., 2026) and Ego3D-Bench (Gholami et al., 2026) measure ego-motion and egocentric spatial reasoning. These benchmarks report accuracies; we report why one of their question types has a ceiling from images and what input lifts it.

Ego-motion in 3D perception and classical geometry. Temporal 3D detectors align past features to the current ego pose before fusion (Li et al., 2022; Park et al., 2023), and the monocular and multi-view detectors whose error we cite (Simonelli et al., 2019; Wang et al., 2021a;b; Chen et al., 2022) supply the positions such a pipeline would use. The geometry of image motion under observer motion is classical (Longuet-Higgins & Prazdny, 1980; Hartley & Zisserman, 2004), visual odometry and SLAM estimate the observer's motion from the static scene (Mur-Artal et al., 2015; Teed & Deng, 2020; 2021), and scene-flow and motion-segmentation methods separate object motion from camera motion (Menze & Geiger, 2015; Yin & Shi, 2018; Ranjan et al., 2019; Yang & Ramanan, 2021). Representation learning has used ego-motion as a training signal (Jayaraman & Grauman, 2015; Agrawal et al., 2015). None of this has entered VLM pipelines for driving; our contribution is to show what is lost when it is absent and what is gained by one token when it is present.

Identifiability, usable information and structure. Disentangling factors of variation is impossible without side information or inductive bias (Locatello et al., 2019; Khemakhem et al., 2020); ego pose is the side information here, and the identifying map is known. Information that a bounded learner cannot extract is not usable (Xu et al., 2020b), networks learn a computation reliably when their structure aligns with it (Xu et al., 2020a; Cohen & Welling, 2016), and probes measure what a representation exposes rather than what it contains (Belinkov, 2022). Our structural ladder is an instance of these principles with the target computation known exactly.

World models and reafference. World models separate the change that an agent causes from exogenous change (Ha & Schmidhuber, 2018; Efroni et al., 2022; Wang et al., 2022), and the reafference principle describes the same subtraction in animals (von Holst & Mittelstaedt, 1950; Sperry, 1950; Wolpert et al., 1995; Warren & Rushton, 2009). Learning with privileged information (Vapnik & Vashist, 2009) describes our auxiliary supervision, which uses the true displacement in training only.

Table 5: The three claims of the paper and the evidence for each.

Claim	Evidence
More frames do not improve motion-state answers	Finetuned BLIP-2: 62.1 vs. 62.9%; three open VLMs: -3.8 to $+2.7$ points, all intervals contain zero (Table 1)
Motion state is not identifiable from ego-frame observations, and is identifiable given relative ego pose	Propositions 1 and 2; AUC 0.557 before and 0.935 after compensation (Table 2); one compensated token: 88.2%, 88.1% blacked out, 74.2% uncompensated (Table 3)
The structure through which the quantity enters decides whether it is used	Same inputs: 66.2 and 76.2% (subtraction) against $88.1 \pm 0.2\%$ (SE(2) action); errors of subtraction grow with yaw as Corollary 1 predicts (Figure 2); compensated text without training: 0 to 10 points, ego pose as text: 0 (Table 1, Appendix C)

6 Discussion

A threshold matches the model, and that is the point. Table 5 lists the claims and their evidence. A threshold on the compensated displacement (88.4%) matches the best VLM (88.2%), and this is the result the paper predicts rather than a weakness of it. If the missing ingredient were capacity, more capacity would have helped and a threshold would not; because it is a missing quantity, supplying the quantity helps by whatever route. The language model earns its place where the question must first be grounded to an object, where several objects are involved, and where the answer is free-form; the token we add is one vector beside the image tokens and composes with all of these.

Why not estimate ego-motion from the image? Visual odometry reads the observer's motion from the static background, and classical odometry beats VLMs at that task (Schäfer et al., 2026), so an image-only pipeline that first estimates ego-motion and then compensates is possible in principle. It is a longer route to the same quantity: the relative pose it would produce enters equation 2 exactly as the odometry pose does, and the token pipeline accepts either source. We used the sensor that every vehicle carries because it removes one estimation step and its error.

What generalizes. The argument does not depend on driving. Any model on a moving platform, a wrist camera on a robot arm or a drone, observes the world in its own frame, and any question about the motion of something in that world is undetermined until the motion of the platform is supplied. The remedy that worked here is cheap: one relative pose per interval, one closed-form transformation, one token.

Limitations. Object positions come from annotations, and the structural ladder is measured on one finetuned backbone. The compensation is a fixed transformation, so improvements in 3D detection transfer to it without retraining; independent position error of the size of current camera detectors (0.6 to 0.8 m) adds about 1 m to the recovered displacement of a stationary vehicle, the regime of the last row of Table 4, where the threshold rule keeps 78%, and a bias that persists across frames cancels in equation 2. Proposition 1 does not depend on the backbone, and the temporal result holds for four.

7 Conclusion

We asked why a vision-language model cannot tell a moving vehicle from a parked one, and why more frames do not help. The answer is that the question is not identifiable from what the camera records about the object: the ego frame adds the observer's motion to every track, and no number of frames removes it. One relative pose per interval makes the question identifiable, one closed-form transformation answers it, and one token carries the answer into the language model, from 62% to 88%. Given the same inputs, a structure that ignores rotation stays 12 to 22 points below the closed form, and the exact group action reaches it every time. A model on a moving platform needs its own motion as an input and the compensation as a structure; the argument is about the observer, not about driving, and it applies wherever a model is asked what moves in a world it is moving through.

AI USE STATEMENT

Generative AI tools were used solely to assist with language polishing of the manuscript, including improving the grammar, clarity, readability, and phrasing of author-written text. All AI-assisted revisions were carefully reviewed and edited by the authors to ensure that they accurately reflect the intended scientific meaning. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses the publicly available nuScenes dataset under its research licence and involves no human subjects. The finding that current VLMs cannot infer object motion from frames concerns the safety of driving assistants that rely on such models, and we regard its publication as beneficial to that end.

REPRODUCIBILITY STATEMENT

The formal statements are in Section 2 and their proofs in Appendix A. The question construction, scene splits, label rule, the common test subset and the excluded questions are in Section 4 and Appendix B, with the training configuration shared by all rungs, the selection of λ and τ on the validation split, and the seeds. The prompts and the paired test for the open VLMs are in Appendix C; the earlier runs, the ablations of the structured module and the per-seed results are in Appendix D; the noise protocol is in Appendix E. Code, question lists, per-question predictions and the analysis scripts are provided in the supplementary material.

REFERENCES

Pulkit Agrawal, João Carreira, and Jitendra Malik. Learning to see by moving. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning (ICML)*, 2021.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Boyuan Chen, Zhuo Xu, Sean Kirmani, et al. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. EPro-PnP: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning (ICML)*, 2016.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, et al. PaLM-E: An embodied multimodal language model. In *International Conference on Machine Learning (ICML)*, 2023.

Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provably filtering exogenous distractors using multistep inverse dynamics. In *International Conference on Learning Representations (ICLR)*, 2022.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Mohsen Gholami, Ahmad Rezaei, Weimin Zhou, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. In *International Conference on Learning Representations (ICLR)*, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Dinesh Jayaraman and Kristen Grauman. Learning image representations tied to ego-motion. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Sampo Kuutti, Saber Fallah, Konstantinos Katsaros, Mehrdad Dianati, Francis Mccullough, and Alexandros Mouzakitis. A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications. *IEEE Internet of Things Journal*, 5(2):829–846, 2018.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision (ECCV)*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019.

H. Christopher Longuet-Higgins and Kvetoslav Prazdny. The interpretation of a moving retinal image. *Proceedings of the Royal Society of London. Series B*, 208(1173):385–397, 1980.

Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3D object detection. In *International Conference on Learning Representations (ICLR)*, 2023.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. NuScenes-QA: A multimodal visual question answering benchmark for autonomous driving scenario. In *AAAI Conference on Artificial Intelligence*, 2024.

Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Finn Rasmus Schäfer, Yuan Gao, Dingrui Wang, Thomas Stauner, Stephan Günnemann, Mattia Piccinini, Sebastian Schmidt, and Johannes Betz. EgoDyn-Bench: Evaluating ego-motion understanding in vision-centric foundation models for autonomous driving. In *European Conference on Computer Vision (ECCV)*, 2026.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *European Conference on Computer Vision (ECCV)*, 2024.

Andrea Simonelli, Samuel Rota Bulò, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3D object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Roger W. Sperry. Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, 43(6):482–489, 1950.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)*, 2020.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009.

Erich von Holst and Horst Mittelstaedt. Das reafferenzprinzip. *Naturwissenschaften*, 37:464–476, 1950.

Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3D object detection. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021a.

Tongzhou Wang, Simon S. Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised MDPs: Learning world models better than the world itself. In *International Conference on Machine Learning (ICML)*, 2022.

Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In *Conference on Robot Learning (CoRL)*, 2021b.

Paul A. Warren and Simon K. Rushton. Optic flow processing for the assessment of object movement during ego movement. *Current Biology*, 19(18):1555–1560, 2009.

Daniel M. Wolpert, Zoubin Ghahramani, and Michael I. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995.

Keyulu Xu, Jingling Li, Mozhi Zhang, Simon S. Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. What can neural networks reason about? In *International Conference on Learning Representations (ICLR)*, 2020a.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations (ICLR)*, 2020b.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024.

Gengshan Yang and Deva Ramanan. Learning to segment rigid motions from two frames. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A PROOFS

Proof of Proposition 1. Fix the observations $(\boldsymbol{o}_s)_{s=t-k}^{t}$. For $y = 0$, pick any point $\boldsymbol{x}^\star$, set $\boldsymbol{x}_s = \boldsymbol{x}^\star$ for all s, and set $g_s = (\boldsymbol{I}, \boldsymbol{x}^\star - \boldsymbol{o}_s)$. Then $g_s^{-1}\boldsymbol{x}_s = \boldsymbol{x}^\star - (\boldsymbol{x}^\star - \boldsymbol{o}_s) = \boldsymbol{o}_s$, and the displacement is zero. For $y = 1$, set $g_s = (\boldsymbol{I}, \boldsymbol{0})$ and $\boldsymbol{x}_s = \boldsymbol{o}_s$ for $s < t$, choose $\boldsymbol{v}$ with $\|\boldsymbol{o}_t + \boldsymbol{v} - \boldsymbol{o}_{t-k}\| > \tau$, and set $g_t = (\boldsymbol{I}, \boldsymbol{v})$ and $\boldsymbol{x}_t = \boldsymbol{o}_t + \boldsymbol{v}$. Then equation 1 holds at every step and the displacement exceeds τ; $\boldsymbol{v}$ can be taken with $\|\boldsymbol{v}\| \le \|\boldsymbol{o}_t - \boldsymbol{o}_{t-k}\| + \tau$, which gives the bound quoted after the proposition. □

Proof of Proposition 2. From equation 1, $\boldsymbol{x}_s = \boldsymbol{R}_s\boldsymbol{o}_s + \boldsymbol{c}_s$. Hence $\boldsymbol{R}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) = \boldsymbol{o}_t - \boldsymbol{R}_t^\top\boldsymbol{R}_{t-1}\boldsymbol{o}_{t-1} - \boldsymbol{R}_t^\top(\boldsymbol{c}_{t-1} - \boldsymbol{c}_t)$. The composition $g_t^{-1}g_{t-1}$ maps $\boldsymbol{o}$ to $\boldsymbol{R}_t^\top\boldsymbol{R}_{t-1}\boldsymbol{o} + \boldsymbol{R}_t^\top(\boldsymbol{c}_{t-1} - \boldsymbol{c}_t)$, so $\Delta\boldsymbol{R} = \boldsymbol{R}_t^\top\boldsymbol{R}_{t-1}$, $\Delta\boldsymbol{c} = \boldsymbol{R}_t^\top(\boldsymbol{c}_{t-1} - \boldsymbol{c}_t)$, and the right-hand side equals $\boldsymbol{o}_t - \Delta g\,\boldsymbol{o}_{t-1}$. Rotations preserve the norm, which gives equation 2. Replacing every g_s by hg_s for a fixed $h \in \mathrm{SE}(2)$ leaves $g_t^{-1}g_{t-1}$ unchanged, which gives the invariance. □

Proof of Corollary 1. A stationary object has $\boldsymbol{o}_t = \Delta g\,\boldsymbol{o}_{t-1}$, so $\boldsymbol{o}_t - \boldsymbol{o}_{t-1} = (\Delta\boldsymbol{R} - \boldsymbol{I})\boldsymbol{o}_{t-1} + \Delta\boldsymbol{c}$. The map $\boldsymbol{o} \mapsto (\Delta\boldsymbol{R} - \boldsymbol{I})\boldsymbol{o}$ is constant if and only if $\Delta\boldsymbol{R} = \boldsymbol{I}$. For a planar rotation by $\Delta\theta$, $\|(\Delta\boldsymbol{R} - \boldsymbol{I})\boldsymbol{o}\| = 2\sin(|\Delta\theta|/2)\|\boldsymbol{o}\|$. □

Effect of pose noise on the recovered displacement. With pose noise $(\boldsymbol{\epsilon}_c, \epsilon_\theta)$ the recovered displacement of a stationary object is $(\boldsymbol{R}(\epsilon_\theta) - \boldsymbol{I})\,\Delta\boldsymbol{R}\,\boldsymbol{o}_{t-1} + \boldsymbol{\epsilon}_c$ up to first order, whose norm is of order $\sigma_c + \sigma_\theta\|\boldsymbol{o}_{t-1}\|$. At $\sigma_c = 0.5$ m and $\sigma_\theta = 2°$ this is about 0.5 m plus 0.035 per metre of range, which is why the validation-selected threshold in Table 4 rises to 1.3 m.

B DATA, TRAINING AND COST

Table 6: Question set.

Split	Scenes	Questions
Train	482	all motion-state questions of the DriveLM training scenes
Validation	84	1,864 motion questions; used for hyperparameters, λ and τ
Test	117	400 class-balanced; 372 with valid side inputs (193 moving, 179 stationary)

Questions. Questions are the perception-stage motion-state questions of DriveLM-nuScenes v1.1 (Sima et al., 2024), of the form "What is the observed status of object [tag, camera, x, y]?" with the answers *moving* or *stationary*; the DriveLM training split is divided by scene into our training, validation and test scenes.

Side inputs. Ego-frame positions are the centres of the annotated 3D boxes of the vehicle expressed in the ego frame of each keyframe, in metres. Relative poses are computed from the annotated ego poses of consecutive keyframes as a planar translation and a yaw change. The auxiliary target $\boldsymbol{u}$ is the world-frame displacement of the annotated box; the compensation computed from the side inputs through equation 2 differs from it by an RMSE of 0.045 m, which reflects the separate annotation of box and ego pose at each keyframe and is the value reported for the closed-form token in Table 3. A question is excluded from the common subset when the 2D box that the question refers to has no 3D annotation with sufficient overlap; this affected 7 moving and 21 stationary questions, mostly parked vehicles at the image border or partly occluded.

Training. All models finetune BLIP-2 (ViT-g, Q-Former, Flan-T5-XL) with LoRA adapters on the language model, with the same learning rate, schedule and number of steps across rungs; the side-input modules are trained from scratch. Subtract + aux uses $\lambda = 0.02$; SE(2)-act uses $\lambda = 0.5$, which was the best of $\{0.02, 0.1, 0.5, 1.0\}$ on the validation split (61.4, 76.3, 79.9 and 59.4% validation accuracy). The threshold $\tau = 0.4$ m of the threshold rule and the noise-specific thresholds of Table 4 are selected on the validation split at the corresponding noise level. Seeds are 0, 1, 2, and 0 to

4 for the MLP-encoded token. The image-only model has no stochastic component beyond its initialization and reached the same checkpoint, with identical predictions on all 372 questions, in three runs.

C TEMPORAL-CONTEXT DETAILS

Open VLMs. Each model receives the image of the camera named in the question, or the three keyframes $t-2$, $t-1$, t in order, and the question followed by the instruction to answer with one of the two words. In the third condition the prompt additionally states the compensated displacement of the vehicle over the interval, as its two components and its norm in metres. Prompts were fixed on the validation split and run once on the test questions. Table 7 gives the paired comparison of three frames against one on the 372 test questions: the counts of questions that both settings answer correctly, that only one setting answers correctly, and that both miss, with the 95% interval on the difference computed from the discordant pairs. Every interval contains zero. For LLaVA-OneVision the predictions with and without the displacement sentence differ on 28 questions and agree on 344, and the two accuracies coincide. A fourth condition states the two relative ego poses in the prompt (yaw and translation per interval) without the compensated displacement: Qwen2.5-VL 50.0%, InternVL3 69.4%, LLaVA-OneVision 58.1%, against 49.7, 68.3 and 69.1% with three frames alone, so the pose by itself is not used either.

Table 7: Paired comparison of three frames against one frame on the 372 test questions. Counts: questions answered correctly by both settings, by one frame only, by three frames only, and by neither. The interval is the 95% interval on the accuracy difference from the discordant pairs. BLIP-2 is one seed of the finetuned model (the three-seed mean in Table 1 is 62.9%).

Model	1f	3f	Δ	95% interval	both	1f only	3f only	neither
Qwen2.5-VL	50.8	49.7	-1.1	$[-3.0, +0.9]$	180	9	5	178
InternVL3	65.6	68.3	$+2.7$	$[-2.6, +8.0]$	198	46	56	72
LLaVA-OneVision	72.8	69.1	-3.8	$[-8.5, +1.0]$	223	48	34	67
BLIP-2, finetuned	62.1	62.6	$+0.5$	$[-1.0, +2.0]$	228	3	5	136

Earlier runs with a fixed front camera. Our first experiments used the front camera image for every question, including questions that name another camera. We report the observations from those runs here because they concern the mechanism and because all main-text numbers were re-computed after the error was corrected. In eight configurations of the temporal module, differing in learning rate and gate initialization, the gate converged to $|\tanh| \approx 0.01$, so the trained model closed the path to the past frames. Linear and nonlinear probes for motion state at the vision encoder, the Q-Former and the projection layer, on object crops and on full images, showed a gain from three frames over one of zero or below, and a TimeSformer encoder changed the gain by -3.2 to 0.0 points. A model with the temporal module present and absent under identical conditions produced identical predictions. A single-frame model and a three-frame model reached 53.5% and 53.25% on the 400-question test set, and blacking out the image lowered the single-frame model from 43.8% to 42.2% on an unbalanced set. These runs disagree with the corrected runs in the single-frame level (62.1% against 53.5%), which is the effect of showing the model the wrong camera, and agree with them in the absence of a temporal gain.

D STRUCTURAL-LADDER DETAILS

The MLP-encoded closed-form token. This module maps the ego-compensated displacement of the annotated box over the two intervals through a three-layer perceptron to one token; it was our first implementation of the closed-form token. Its five seeds reached 84.1, 89.0, 83.9, 89.3 and 83.6%; the first three (mean $85.7 \pm 2.9\%$) gave 86.3% with the image blacked out. On the fixed-front-camera data the same module reached $89.6 \pm 0.4\%$, 88.7% blacked out, and 67.6% when given the ego-frame displacement, and an MLP over the raw side inputs, with no structure, reached 56.6% there. The main text uses the SE(2) parameterization with the residual frozen, which encodes the same quantity with lower seed variance.

An unbounded residual leaves the exact solution. With annotated poses the optimal residual is zero, and SE(2)-act starts there, with a recovered-displacement RMSE of 0.045 m. Trained without a bound on the residual, at the full learning rate and with $\lambda = 0.02$, two seeds ended at 57.8% and 65.9% and the RMSE grew to between 12 and 19 m: the answer loss moved the module away from the exact solution. On the same data, freezing the residual gave $88.2 \pm 0.3\%$, training the residual at one tenth of the learning rate after one frozen epoch gave $88.2 \pm 0.3\%$, and a two-stage schedule without the reduced learning rate gave $86.9 \pm 1.9\%$, all with RMSE 0.07 m; the main text uses the reduced-learning-rate variant retrained on corrected data ($88.1 \pm 0.2\%$).

Auxiliary weight of the subtractive module. Subtract + aux in the main text uses $\lambda = 0.02$, the value of our first implementation, whereas SE(2)-act uses the validation-selected $\lambda = 0.5$. The asymmetry does not favour our conclusion: with $\lambda \in \{0.1, 0.5, 1.0\}$ the subtractive module reached 74.9 ± 2.2, 73.7 ± 2.2 and $73.9 \pm 1.8\%$ in the fixed-front-camera runs, so its accuracy is limited by the structure and not by the weight of the supervision.

Yaw stratification. Bins of absolute ego yaw change hold 303, 39 and 30 test questions. Subtract + aux: 77.9, 71.8, 64.4%; closed-form token: 88.1, 87.2, 93.3%.

E POSE-NOISE DETAILS

Table 8: Closed form under additive pose noise with the threshold fixed at the noise-free value $\tau = 0.4$ m, against the validation-reselected threshold of Table 4. Accuracy (%) on the 372 test questions.

σ_c (m) / σ_θ ($^\circ$)	Fixed $\tau = 0.4$	Reselected τ
0.05 / 0.5	87.1	86.0 ($\tau = 0.5$)
0.2 / 1	75.8	80.4 ($\tau = 0.6$)
0.5 / 2	57.8	78.2 ($\tau = 1.3$)
0.5 / 5	53.8	–

At the largest noise levels the fixed-threshold rule approaches the rate of answering *moving* for every question (51.9%): noise of 0.5 m on the relative translation alone exceeds τ for most stationary vehicles. At the weakest level the fixed threshold is one point above the reselected one, which reflects the selection on validation rather than test data. SE(2)-act in Table 4 was trained and evaluated with noise of the stated level added to the relative pose and used the bounded residual; a Subtract + aux model trained and evaluated in the same way reached 75.4, 74.9 and 73.8% at the three levels in earlier fixed-front-camera runs.