# Dynamics-Informed Vision–Language Models: An Extended Abstract on Dynamics-Aware Reasoning towards next generation Autonomous Systems

Finn Rasmus Schäfer[1] and Prof. Dr.-Ing. Johannes Betz[2]

## I. Introduction & Motivation

With the rise of end-to-end learned architectures in embodied robotics and autonomous driving, recent research has increasingly focused on Vision-Language Models (VLMs) and Vision-Language-Action (VLA) architectures. These approaches belong to the broader family of foundation models, which aim to serve as embodiment-specific general-purpose models. By conditioning primarily on image and text data, such models achieve strong semantic understanding, instruction following, and generalization capabilities.

Despite these successes, current VLM and VLA architectures are vision-dominant by design. In contrast, classical and rule-based robotic and autonomous driving systems explicitly rely on ego-states, system dynamics, and physical constraints. This explicit conditioning enables such systems to enforce safety and dynamic feasibility throughout the decision-making process. While VLA architectures partially address embodiment by coupling perception and action, safety and physical plausibility are typically not enforced at the level of semantic reasoning.

As a consequence, we observe an increasing semantic-dynamic mismatch between what a model *understands* and reasons about, and what a physical system can actually execute. This mismatch becomes particularly critical in safety-critical domains such as autonomous driving and mobile robotics. Motivated by this gap, we argue that the next generation of VLM and VLA architectures must move beyond vision-centric representations and explicitly incorporate ego-dynamics and physical state as aligned modalities within their reasoning process.

## II. Limitations of Vision-Centric Foundation Models

VLM and VLA architectures implicitly assume a direct dependency between perception, reasoning, and action. However, this abstraction typically omits an explicit dependency on the agent's internal physical state. As a result, these models rely on a simplified coupling that can break down, particularly in out-of-distribution or safety-critical scenarios.

Vision-centric representations, largely derived from visual encoder features, struggle to reason about dynamic feasibility. This includes concepts such as braking distance, inertia, stability, and motion-related attributes like aggressive, comfortable, or risky behavior. Moreover, such models are generally unable to account for actuation limits, as they are often trained in an open-loop manner without direct feedback from the system dynamics. While recent work—particularly in general robotics—has begun to explore closed-loop optimization, this remains a non-established and fragmented research direction.

These limitations can lead to actions that are semantically plausible yet dynamically unsafe, as well as biased reasoning that oscillates between overly cautious and overly aggressive behavior depending on the unobserved physical state of the agent. In contrast, classical rule-based systems explicitly model kinematics and dynamics, enforce safety constraints by design, and offer predictable and interpretable behavior. Bridging this gap, by combining semantic reasoning with explicit awareness of physical state and system dynamics, represents a necessary step toward foundation models that are deployable in real-world robotic and autonomous driving applications.

## III. Multi-modal Alignment in Robotics: Current Trends

Recent research has begun to move toward multimodal representations that extend beyond purely vision-centric models, partially addressing the limitations discussed above. In general robotics, several approaches incorporate additional modalities directly into learned representations, enabling more informed control and interaction with the environment [1]. While these methods do not explicitly model system dynamics, they demonstrate the benefits of embedding embodiment-specific information alongside perception.

In autonomous driving, related efforts integrate non-visual sensing modalities into vision-language models. For example, LiDAR information has been incorporated to improve robustness and safety in low-light or visually degraded conditions [2]. Other approaches leverage multi-modal preprocessing pipelines to generate feasible trajectories [3], or employ vision-based priors to localize the agent within a known map [4]. Although differing in scope and application, these works share a common shift away from classical sensor fusion toward representation-level multi-modal alignment.

Together, these trends indicate an emerging consensus that richer, aligned representations are essential for safe and robust robotic behavior. However, ego-dynamics and explicit physical state remain largely absent as first-class

[1]Technical University of Munich, Autonomous Vehicle Systems Lab, Garching b. München, Germany finn.schaefer@tum.de
[2]Technical University of Munich, Autonomous Vehicle Systems Lab, Garching b. München, Germany Johannes.betz@tum.de

 

modalities within current foundation models. This gap highlights the opportunity to extend existing multi-modal alignment paradigms toward dynamics-informed representations that directly support feasibility-aware semantic reasoning in robotics and autonomous driving.

## IV. Dynamics-Informed Vision–Language Models

Dynamics-informed VLM and VLA architectures should explicitly condition semantic reasoning on ego-motion, the agent's dynamic state, and the physical feasibility of intended actions. Achieving this requires incorporating dynamic modalities into a shared representation space aligned with the language model, rather than treating dynamics as a post-hoc filtering mechanism. In this formulation, ego-dynamics act as a conditioning signal during autoregressive reasoning, not as an external safety check applied after inference.

We conceptualize this joint embedding space as

$$\mathcal{S} = \{T, V, D\}, \tag{1}$$

where $T$ denotes text embeddings, $V$ denotes vision embeddings projected into the language model's embedding space, and $D$ denotes dynamic embeddings capturing ego-motion and system state, likewise projected into the same space. The resulting abstract space $\mathcal{S}$ represents a rich, vision-, dynamics-, and prompt-informed representation on which the model can condition its reasoning.

By grounding semantic reasoning in physical state and system dynamics, such models are expected to produce decisions that are not only semantically coherent but also dynamically feasible. This can increase robustness under distribution shift, reduce reliance on rule-based safety overrides, and improve the alignment between intended and executable actions. Importantly, this approach does not replace vision-centric reasoning, but augments it with embodiment-aware context that is essential for real-world robotic and autonomous driving applications.

## V. Implications for Safety, Control, and System Design

Integrating explicit dynamic information into vision-language and vision-language-action models has significant implications for safety, control, and overall system design. By conditioning semantic reasoning on physical state and system constraints, such models can better anticipate infeasible or unsafe actions before execution. This capability emerges from a tighter alignment between intent, physical state, and action execution.

For real-world deployment, this alignment can improve the controllability of embodied agents and reduce reliance on hard intervention layers or externally imposed safety checks. Instead of correcting unsafe behavior post-hoc, safety-aware reasoning becomes an intrinsic part of the decision-making process. Moreover, the ability to explain actions in a physically grounded manner, based on both semantic intent and dynamic state, enhances interpretability and trustworthiness.

Finally, dynamics-informed reasoning can contribute to more robust behavior under partial observability and in rare or edge-case scenarios, where purely vision-centric representations often fail. Together, these implications highlight the potential of dynamics-informed foundation models to bridge the gap between semantic competence and physically reliable behavior in real-world robotic and autonomous driving systems.

## VI. Open Challenges & Future Directions

The development of dynamics-aware VLM and VLA architectures raises two fundamental open challenges: multi-modal alignment and the availability of suitable training and evaluation data.

First, effective multi-modal alignment remains a central difficulty. Aligning vision, language, and dynamic modalities within a shared representation space is non-trivial and prone to shortcut learning or spurious correlations. Without careful supervision and inductive biases, the integration of dynamic information may fail to improve. This may even degrade semantic reasoning. An open research question is how to design alignment strategies that ensure the dynamic state is meaningfully incorporated rather than implicitly ignored. Closely related is the question of evaluation: how should semantic reasoning be assessed under explicit dynamic and physical constraints?

This challenge directly leads to the second issue, namely the lack of appropriate training data and benchmarks. To the best of our knowledge, there currently exist no standardized datasets or benchmarks that explicitly support the training and evaluation of dynamics-informed VLM or VLA architectures. Progress in this direction will require high-quality datasets that jointly capture semantic intent, perception, and ego-dynamic state, as well as evaluation protocols that measure physical feasibility alongside semantic correctness.

Addressing these challenges is essential for advancing foundation models toward reliable embodied intelligence. We argue that dynamics-informed multi-modal representations constitute a necessary step in bridging the gap between semantic competence and physically grounded, deployable robotic behavior.

## References


[1] H. Zhou, C. Ma, and G. H. Lee, "VLA-4D: Embedding 4D awareness into vision–language–action models for spatiotemporally coherent robotic manipulation," *arXiv preprint arXiv:2511.17199*, 2025.

[2] S. Kirchner, N. Purschke, R. Greer, and A. C. Knoll, "DepthVision: Enabling robust vision-language models with GAN-based LiDAR-to-RGB synthesis for autonomous driving," *arXiv preprint arXiv:2509.07463*, 2025.

[3] X. Zhou, X. Han, F. Yang, Y. Ma, V. Tresp, and A. Knoll, "OpenDriveVLA: Towards end-to-end autonomous driving with large vision language action model," *arXiv preprint arXiv:2503.23463*, 2025.

[4] Y. Ji, Y. Wang, Z. Ma, Y. Hu, H. Huang, X. Hu, G. Chen, L. Wu, and X. Chu, "Thinking with map: Reinforced parallel map-augmented agent for geolocalization," *arXiv preprint arXiv:2601.05432*, 2026.