Extended Abstract Track

Selective Cross-Modal Alignment in Vision-Language Models: Why Position Grounding Succeeds but Physical Reasoning Fails

Abstract

Physical intelligence—the ability to perceive dynamics and anticipate what happens next—is central to general intelligence, yet little is known about whether vision–language models (VLMs) internally represent it. We probe Qwen2.5-VL-3B on 6,000 fully controlled synthetic physics clips and uncover a *selective cross-modal alignment failure*: position is linearly decodable from visual tokens ($R^2 = 0.97$) and perfectly grounded in language (100% success), whereas velocity is encoded only as an implicit, entangled manifold ($R^2 = 0.42$, non-linear probe) that language generation does not use (19% physical-reasoning accuracy; less than 5% causal effect when the velocity-encoding layers are patched). This is not a missing capability: prompt engineering raises physical reasoning by +62% and eliminates a systematic response bias, while leaving the visual representations unchanged (cosine similarity ≈ 1.0). Comparisons with DINOv2 and SigLIP show that language training *created* the velocity encoding (R^2: 0.04/0.07 $\rightarrow$ 0.42), yet the default readout ignores it. Cross-modal alignment in VLMs is therefore selective and prompt-dependent: a model can hold physical knowledge it does not use, and capability must be measured by probing and activation, not by behavior alone.

Keywords: vision-language models, representation geometry, probing, cross-modal alignment, physical reasoning

1. Introduction

Vision–language models (VLMs) ground language in images with impressive accuracy. But grounding is *static* perception—it answers "what is where." General intelligence additionally requires an internal physical model that answers "what happens next": the gap between the two is the gap between a model that can point at a ball and a model that knows where the ball is going.

Whether a VLM possesses physical understanding is usually tested behaviorally, treating the model as a black box (Riochet et al., 2018; Bear et al., 2021; Yi et al., 2020). Such tests cannot answer the mechanistic question: *is physical information encoded in the model's internal representations at all—and if so, why does the model not use it?* Probing has revealed syntax and semantic pipelines in language models (Belinkov and Glass, 2019; Hewitt and Manning, 2019; Tenney et al., 2019), visual concepts in self-supervised encoders (Ranasinghe et al., 2022), and physical properties inside vision networks (Zhan et al., 2024); yet no prior work links kinematic representations in a grounding-capable VLM causally to its language behavior. Vision–language alignment, in turn, is usually treated as a static, all-or-nothing property (Radford et al., 2021; Liu et al., 2023; Peng et al., 2023).

© 2026 .

We investigate Qwen2.5-VL-3B-Instruct (Bai et al., 2025) with layer-wise probing, causal intervention, and prompt engineering, asking three questions: (i) **Representation**: is velocity encoded in the visual tokens? (ii) **Utilization**: does language generation functionally use it? (iii) **Activation**: can prompting activate latent physical reasoning? Our answer, in one sentence: *the model encodes physics it does not use, and a few words of guidance make it use it*—position is explicit and aligned, while velocity is an entangled manifold, unaligned by default but prompt-activatable.

Contributions. (i) A joint probing–intervention–prompting framework for cross-modal alignment analysis; (ii) evidence that velocity lives in an implicit manifold ($R^2 = 0.42$ nonlinear vs. 0.34 linear) while position is explicit ($R^2 = 0.97$ linear); (iii) causal proof that the velocity layers are unused by default (<5% intervention effect) yet activatable (+62% via prompting, bias eliminated, representations unchanged); (iv) the finding that language training *creates* the velocity encoding, which the default readout ignores.

2. Setup and Methods

Synthetic kinematics data. We generate 6,000 short physics clips (2,000 each in *minimal*, *medium*, and *realistic* visual styles) with the pymunk physics engine: 1–3 balls with randomized initial velocities, gravity, elasticity, and friction. Each 8-frame clip ships ground-truth trajectories $(x, y, v_x, v_y, a_x, a_y)$ per frame; splits are 80/20 at the trajectory level.

Layer-wise probing. For each visual-encoder layer l, we train probes to predict velocity from the mean-pooled tokens covering each ball (ground-truth positions avoid detection confounds). A *ridge* probe tests linear decodability; a *two-layer MLP* (hidden width 256) tests whether the information exists non-linearly. A constant-velocity extrapolation baseline ($R^2 \approx 0.90$) and a shuffled-label control calibrate the probes; position probing uses the identical pipeline as a sanity check.

Causal intervention. To test *functional use*, we patch the tokens of the velocity-encoding layers (16 and 24) with their per-token mean and re-run the behavior task (100 clips $\times$ 2 layers = 200 interventions). If language generation reads those layers, accuracy should collapse; a small change indicates the layers are bypassed.

Behavior and prompting. Behavior is measured as multiple-choice direction prediction (A: up / B: down / C: left / D: right; chance 25%) and open-ended equivalents; grounding questions ("where is the ball?") serve as the alignment reference. For activation, we evaluate six prompt formulations (100 clips each): baseline, physical analysis, visual grounding, anti-bias, explicit physics, and chain-of-thought, following the elicitation paradigm of Wei et al. (2022); Brown et al. (2020).

3. Results

3.1. Position is explicit; velocity is an implicit manifold

Position is linearly decodable at $R^2 = 0.97$ (layer 16; above 0.93 for layers 8–31), validating the probing pipeline. Velocity behaves differently (Figure 1): the ridge probe peaks at only $R^2 = 0.34$ (layer 24), while the MLP reaches $R^2 = 0.42$ (layer 16)—far below the position signal and below the constant-velocity baseline ($R^2 = 0.90$).

Extended Abstract Track

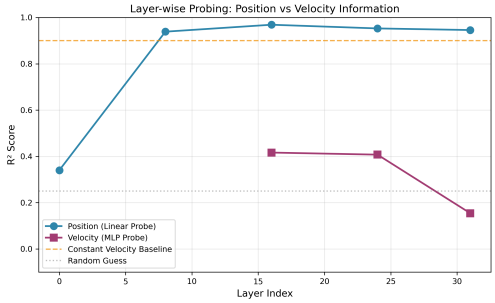


Figure 1: Layer-wise probing of Qwen2.5-VL. Position (linear probe) decodes near-perfectly from mid layers onward; velocity (MLP probe) peaks far lower and below the constant-velocity baseline.

Representation geometry explains *why* velocity is hard to read. Clustering token representations by motion direction yields an average inter-direction cosine similarity of 0.9982 versus an intra-direction similarity of 1.0000: the direction separation is only 0.0018, and a direction classifier reaches 27.6% accuracy, near the 25% chance level. Speed correlates with the representation norm at only 0.037. The first five principal components explain 72% of the variance, yet none correlates with velocity above 0.09: velocity is spread across nearly all 1,280 dimensions as a continuous, entangled manifold rather than isolated into decodable directions. Position, in contrast, occupies an explicit, axis-readable geometry.

3.2. Language does not use velocity by default

Under default prompting, multiple-choice accuracy is 23.6% (118/500), *below* the 25% chance level, with a stark systematic bias: 72.4% of answers are "C" (left) versus 22.2% in the ground truth. Open-ended accuracy is 28.0%: the answer format is not the bottleneck. The most direct contrast comes from paired queries on the same 100 clips: the model describes *where* the ball is with 100% success, yet predicts *where it goes* with only 19% success; 81% of samples are grounding-correct but physics-wrong (Figure 2).

Causal intervention confirms functional non-use. Patching the velocity-encoding layers (16 and 24) with their mean changes accuracy from 20.0% to 17.5%—a 4.5% drop, far below the collapse expected if language generation routed through those layers. The representations hold the velocity information; the language head simply does not read it.

3.3. Prompting activates the latent readout

If the capability were absent, no prompt could recover it. Instead, every prompt formulation improves accuracy, and the best ones—physical analysis and visual grounding—raise it from 21% to 34% (+62%) while reducing the C-bias from 73% to 0%; anti-bias, explicit-physics, and chain-of-thought prompts reach 31%, 30%, and 29%, with the bias likewise eliminated (Figure 3). The capability is *latent, not absent.*

Information-flow analysis confirms where prompting acts: baseline and prompted visual representations are identical (cosine similarity 0.99998), yet all 20 sampled responses

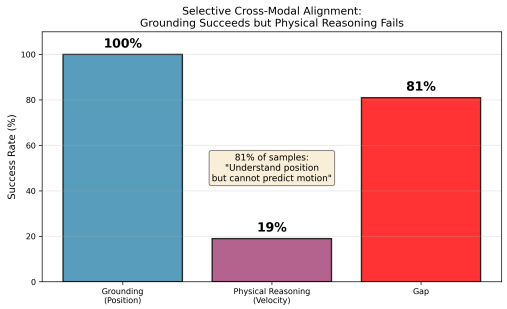


Figure 2: Selective alignment: grounding 100% vs. physics 19% on 100 paired queries.

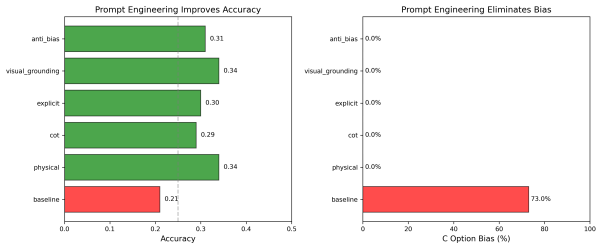


Figure 3: Every prompt strategy improves accuracy (left) and erases the "C" bias (right).

change—prompting alters only the *decoding* (how the language head reads the visual stream), never the *encoding* (what it contains).

3.4. Language training creates the velocity encoding

We probe two vision encoders without strong language training on the identical pipeline. DINOv2 (Oquab et al., 2024) reaches a best velocity R^2 of 0.04 and SigLIP (Zhai et al., 2023) 0.07—essentially no velocity information—while Qwen2.5-VL reaches 0.42, a roughly 10× gain. Language training *taught the visual encoder physics*; the default language readout then ignores it.

4. Discussion and Limitations

Selective alignment is a geometric property. Position is a grounding objective and is both explicitly encoded and perfectly aligned (100%); velocity is not a training objective: it survives only as an entangled manifold and is not aligned (19%), despite being prompt-activatable. Cross-modal alignment is therefore not a global property of a VLM but a *per-information-channel* property, and our evidence suggests a geometric mechanism: information that is encoded explicitly—concentrated, separated, linearly readable—is picked up by the default language readout, whereas information spread as a continuous manifold is bypassed.

Knowledge without use. Capability and behavior can diverge: the same representation yields 21% or 34% accuracy depending on phrasing, and biases appear or vanish accordingly. Benchmarks thus severely underestimate the model; assessment should combine behavior with probing and activation tests. Aligning dynamic properties during training may convert latent knowledge into always-on capability; deployment must assume models silently hold knowledge they do not use.

Limitations. We study one model family at 3B scale on synthetic scenes; generality across architectures, scales, and real videos is open. Probing uses ground-truth positions; the paired contrast and prompt evaluation use 100 clips each, chosen for control. Whether explicit geometry predicts alignability across channels is future work.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1877–1901, 2020.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4129–4138, 2019.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan. Self-supervised vision transformers learn visual concepts in the absence of language. *arXiv preprint arXiv:2203.07766*, 2022.

Ronan Riochet, Mario Y. Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4593–4601, 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023.

Geng Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. A general protocol to probe large vision models for 3D physical understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.