GAINS STAY WHERE THE LOSS WENT: A CHANNEL-BY-CONDITION MAP OF PHYSICAL-PLAUSIBILITY TRAINING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A vision-language model shown a shadow that points toward the light calls the picture physically possible. Fine-tuning on plausibility repairs this, and prior work reports the repair on one readout, usually an accuracy. On ImPhysible, a rendered testbed of matched normal and violated images, we read one model's judgment on four channels, a 0 to 100 score, the first-token logit, the generated verdict, and a probe on the hidden state. We vary which channel and which condition the training loss passed through. Every (channel, condition) cell the loss passed through rose, in every variant we trained. The cells it did not pass through rose, stayed at chance, or moved the other way, and no variant made them rise reliably. A score-order reward lifts score pair-rank accuracy from 0.469 to 0.576 and pushes the logit of the same answers from 0.504 to 0.456. An alignment loss that holds out one domain raises that domain's representation cosine to between 0.556 and 0.729 while its free answer stays between 0.463 and 0.500. Decide which channel you will read, put the loss through it, and test untrained conditions separately. One model, rendered stimuli, and an earlier data version behind part of the channel-axis evidence bound the claim.

1 INTRODUCTION

Show a vision-language model a rendered scene in which the shadow falls toward the light and it will tell you the picture is fine. A person sees the fault at once. Earlier work asked where the failure lives, whether the evidence is missing from what the model encodes or is present and lost on the way to the answer (Nooralahzadeh et al., 2026; Fu et al., 2025; Mallen et al., 2024; Orgad et al., 2025). We started from the same question and found that it has no single answer for a trained model, because the answer depends on where you read the model and on what the training loss touched.

Training a model to judge plausibility does raise its judgments, and the papers that do so report the rise on one readout. Video and image models fine-tuned on implausibility report classification accuracy (Motamed et al., 2025; Hu et al., 2025). Pair-based objectives report task accuracy (Zhang et al., 2025; Wu et al., 2025a) or a rate of visual reliance (Luo et al., 2025). None of these reports asks whether a second readout of the same trained model moved with the first.

We ask exactly that. We call each place where the model's judgment can be read a channel, and we read four of them on every image. They are the score it writes as a number, the first-token logit of *impossible* against *possible*, the verdict word it generates, and a linear probe or cosine on its hidden state. We call the domain, camera viewpoint, and relation direction of a test pair its condition. A cell is one (channel, condition) pair. For each training variant we record which cells the loss passed through, meaning the channel whose output the loss read and the conditions whose pairs it saw. Then we measure every cell.

Every cell the loss passed through rose, in every variant we trained. The other output channels of the same answers and the untrained conditions rose, stayed at chance, or moved the other way, and no variant made them rise reliably (Figure 1). We do not claim a theorem. A lower training objective

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.

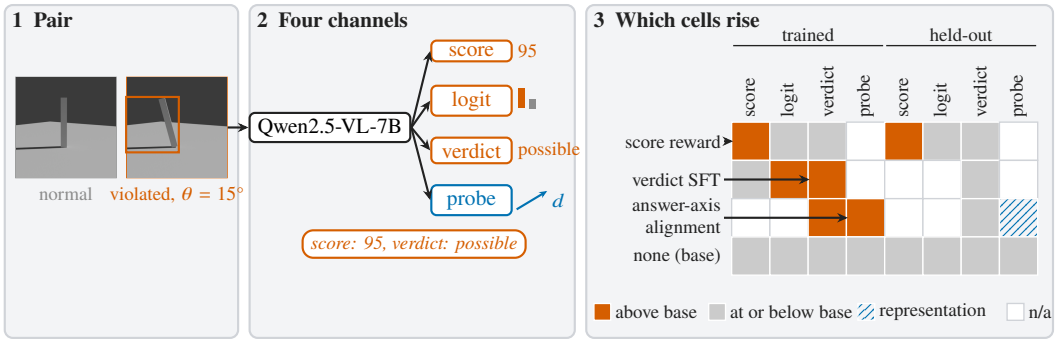

Figure 1: Training on physical plausibility reliably raises only the cells the loss passed through. Left, one normal counterpart and its violated image from the tower domain go through one model and are read on four channels. Right, each row is a training loss, each column a (channel, condition) cell, and colour marks a rise above the base model over three seeds where available. Cells the loss did not pass through are grey when they stayed near base or chance, hatched when only the representation channel moved, and white where that channel was not measured. Exact values are in Tables 2 to 5.

does not by itself guarantee a gain even in a supervised cell (Section 3), and the map is a report of what happened across the variants we trained.

The paper makes three points. (1) A score-order reward raises the score channel and leaves the logit and verdict channels of the same answers below the base model, in nine and eight of twelve domains (Section 4). (2) On a held-out domain, an alignment loss moves the representation toward the answer axis while the free answer stays at chance, and the same holds for a held-out camera viewpoint and a held-out relation direction (Section 5). (3) The loss path predicts the cells that rise, which gives a one-line rule for practitioners and a one-line warning for benchmarks that report a single readout (Section 3).

2 ImPhysible: minimal pairs with a known violation size

ImPhysible is a set of rendered pairs. Every violated image has a normal counterpart rendered from the same scene through the same path, and the renderer knows the violation and its size θ. The two images differ only in the violation. No person labels the data, and no rendering artifact separates the pair, because both images come out of the same pipeline with one parameter changed. We chose rendering over photographs for this reason. A photograph of an impossible scene has no normal counterpart, and a manipulated photograph carries the manipulation as a second signal.

The physical set has twelve domains with 400 pairs each, 4,800 pairs in all, drawn from 7,256 rendered pairs. Each domain has ten levels of θ. Pairs are split by scene into 3,840 training, 480 validation, and 480 test pairs, so no test scene appears in training. Three control domains share the pipeline but carry no physical violation. They change the object count with painted area held equal, the relative bearing of two objects, or the object size with total screen area held equal. Table 1 lists the domains, and Figure 2 shows one pair per domain with the base model's raw answer.

We audit shortcuts before reading any model value. A frozen DINOv2 feature with a linear head separates every physical domain, from 0.780 on shadow length to 1.000 on support, thermal, water, and tower. Every model value in this paper is therefore read as the share above that frozen baseline, and the four domains at 1.000 are marked in every table.

We read judgments as pair-rank accuracy. For a pair, the model is correct when its readout ranks the violated image above the normal counterpart, a tie counts one half, and chance is 0.5. Any constant added to both images of a pair cancels, so a scene offset cannot move this number. Missing answers are never filled in. When a channel returns no parseable value we drop the pair from that channel and report the response rate next to the accuracy. Without training, Qwen2.5-VL-7B (Bai et al., 2025a) sits at chance, 0.531 from its verdicts and 0.522 from its logits on the physical set.

Table 1: Twelve physical domains and three control domains share one design, a normal counterpart and a violated image from the same scene with a violation of size θ. The last column is the pair-rank accuracy of a frozen DINOv2 feature with a linear head, the shortcut baseline every model value is read against. † marks the four domains the frozen feature separates at 1.000. 4,800 physical pairs from 7,256 rendered pairs in total.

Domain	What is violated	θ unit	θ range	Scenes	DINOv2
shadow	shading and cast shadow disagree on the light	deg	2 to 90	10	0.865
support†	object floats while its shadow touches the floor	render param	5 to 100	10	1.000
reflection	mirror image is off its symmetric position	deg	1 to 45	10	0.925
thermal†	smoke cone does not rise from the heat source	deg	5 to 120	40	1.000
water†	water reflection is displaced from the object	cube width	0.03 to 1.0	40	1.000
refraction	pattern behind glass bends too little	IOR drop	0.02 to 0.5	40	0.963
liquid	liquid surface in a tilted glass is not level	deg	5 to 70	40	0.993
perspective	far object is too large for its distance	scale minus 1	0.1 to 2.5	40	0.985
shadow length	shadow length and shading disagree on sun height	deg	2 to 30	10	0.780
highlight	specular highlight sits on the wrong side	deg	10 to 180	40	0.927
tower†	column leans past its tipping angle and stands	deg	8 to 42	40	1.000
rotation	two boxes face different ways	deg	5 to 180	40	0.938
counting (control)	sphere count differs, painted area equal	ratio minus 1	0.08 to 2.0	25	0.500
spatial (control)	bearing of B from A differs	deg	5 to 150	40	0.595
size (control)	object size differs, screen area equal	ratio minus 1	0.01 to 1.0	9	0.978

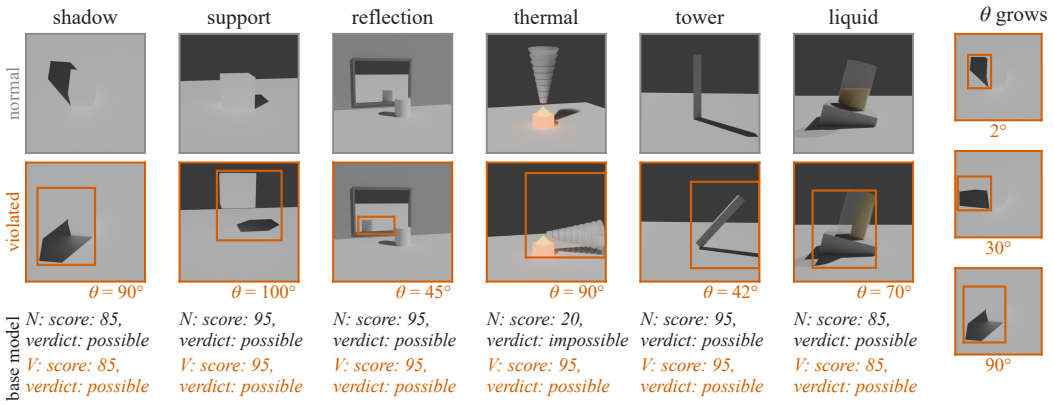

Figure 2: The two images of a pair differ only in one physical violation, and the base model calls the violated image possible in every column. Each column is a domain, top the normal counterpart, bottom the violated image at the largest test θ, and below each pair the base model's raw answer. The right column shows the shadow domain at three sizes of θ. The remaining six domains are in Figure 7.

3 Channels, conditions, and the cells a loss passes through

We read one model's judgment in four places and call each a channel. The score channel is the number the model writes when asked to rate how physically possible the image is from 0 to 100,

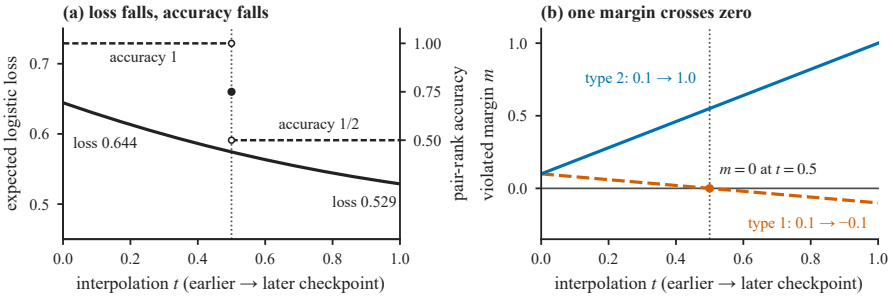

Figure 3: A lower training loss does not guarantee a higher pair-rank accuracy, because an update can sacrifice one near-boundary pair while improving another. Two pair types of equal weight have normal readouts at zero, and their violated margins move linearly from $(0.1, 0.1)$ at the earlier checkpoint ($t = 0$) to $(-0.1, 1.0)$ at the later one ($t = 1$). (a) Expected logistic loss $\ell(z) = \log(1+e^{-z})$ (solid, left axis) and pair-rank accuracy with a zero margin counted as one half (dashed step, right axis). The loss falls from 0.644 to 0.529 while accuracy drops from 1 to 1/2 at the dotted line. (b) The violated margins of type 1 (dashed) and type 2 (solid) with the zero line. Type 1 crosses zero at $t = 0.5$, the same point as the drop in (a).

read so that a higher number means a larger violation. The logit channel is the difference of first-token logits for *impossible* and *possible* on a yes-or-no question. The verdict channel is the word the model generates on the same question, parsed to violated or normal, with unparseable answers left missing. The representation channel is read on the hidden state. It is either a linear probe fit on held-out scenes or the cosine between the mean difference of violated and normal hidden states and a fixed answer axis, written $\cos(\Delta\mu, d)$. The first three are output channels and keep their readout fixed at test time. The representation channel refits or reprojects at test time. These are different measurements, and we do not read a rise in one as a rise in another.

The second axis is the condition, a property of a test pair that the training data either contained or left out. Here that property is the domain, the camera viewpoint, or the direction of a relation, for instance whether the light or the cast shadow was the thing rotated. Crossing the two axes gives a cell, one channel under one condition. The loss passed through a cell if the objective read that channel's output during training and the pairs it saw came from that condition.

Four kinds of variant start from the same base model and see the same training pairs. They differ in where the loss reads the model. A score-order reward uses group-relative policy optimization (Shao et al., 2024) on the number the model writes, so it passes through the score channel. A verdict loss fine-tunes the answer token with a supervised objective. The logit and the verdict sit on that one token, so this loss passes through both at once. An alignment loss pushes the mean difference between violated and normal hidden states toward the answer axis and never touches an output, so it passes through the representation channel alone. A head adapter fits a pairwise ranking head on the hidden state with the verdict token as target. Every variant is a LoRA (Hu et al., 2022) of rank 64 on the language model, vision encoder frozen. Appendix B lists each variant.

What follows is a comparison, not a theorem. On the pairs of a condition valid at both checkpoints, with prompt, parser, decoding rule, and score polarity held fixed, write the pair difference of a channel as $m_i = r(x_{v,i}) - r(x_{n,i})$, so that pair-rank accuracy is the mean of $q(m_i)$ with $q(z) = \mathbf{1}[z > 0] + \frac{1}{2}\mathbf{1}[z = 0]$. Three facts about this number organise the map (proofs in Appendix C). First, if no pair difference decreases, the accuracy cannot fall, and it rises exactly when some pair moves upward between negative, zero, and positive. A lower training objective does not give this condition, even in a supervised cell. With two pair types of equal weight and normal readouts at zero, moving the violated margins from $(0.1, 0.1)$ to $(-0.1, 1.0)$ lowers the mean logistic loss from 0.644 to 0.529 while accuracy falls from 1 to 0.5 (Figure 3).

Second, for two channels a and b read on the same pairs, let e_t be the share of pairs whose differences have different signs on a and b at checkpoint t, with zero as a third category. The gains of the two channels then obey

$$\left|(\mathrm{A}_b^1 - \mathrm{A}_b^0) - (\mathrm{A}_a^1 - \mathrm{A}_a^0)\right| \le e_0 + e_1. \tag{1}$$

A gain on a larger than $e_0 + e_1$ carries to b, and channels that agree in sign at both checkpoints gain by the same amount. A shared LoRA update does not by itself make two readouts agree in sign. Third, on complete pairs a binary readout has pair-rank accuracy $(1 + \mathrm{TPR} - \mathrm{FPR})/2$, so a constant verdict gives exactly 0.5 while the continuous logit under it can still rank. A constant shared by both images of a pair leaves every m_i unchanged, which is why scene offsets never enter our numbers. So we report continuous logits, their signs, and generated verdicts separately on the same valid pairs. These facts say when a gain transfers. They do not say that gains occur only in supervised cells, and that restriction is what the rest of the paper tests. Everything else in the map is measured, and the one-sided summary we draw is that supervised cells rose in every variant we trained while unsupervised cells did all three things.

4 A SCORE LOSS RAISES THE SCORE CELL AND LEAVES THE LOGIT AND VERDICT CELLS BELOW BASE

Under the score-order reward, the pair-rank accuracy of the score channel goes from 0.469 to 0.576 $\pm$ 0.049 over three seeds. The logit and verdict channels of the same answers fall from 0.504 to 0.456 $\pm$ 0.063 and from 0.519 to 0.457 $\pm$ 0.021 (Table 2, Figure 4a, physical set, earlier data version). The score rises in eleven of twelve domains. The logit falls in nine and the verdict in eight. On tower the model reaches 0.863 on the score and 0.233 on the logit, and a typical raw answer reads `score: 95, verdict: possible`. The same tokens carry a confident violation on one channel and a confident normal on the next.

The sign of the change on the unsupervised channels differs by domain. On liquid the logit rises from 0.450 to 0.658 and on highlight from 0.325 to 0.525, while on water it falls from 0.575 to 0.192. This is why the claim is one-sided. The supervised cell rose in every seed. The unsupervised cells rose, stayed, or fell, and the data show all three. Across the ten physical domains the domain-level rise in score and fall in logit have a Spearman correlation of 0.841 (Figure 4b). Dropping the four domains where the frozen feature already separates the pair at 1.000 leaves 0.630. We do not read this as one channel suppressing the other. It says the two did not move together.

The same adapter tells the same story on a task it never saw. On the size control, 160 test pairs of a non-physical size change, the score-trained adapter reaches 0.645 $\pm$ 0.130 on the score channel from a base of 0.447. Its logit falls to 0.169 $\pm$ 0.019 from 0.303 (Table 3, Figure 4c). The verdict sits at 0.526 $\pm$ 0.040 against 0.500. The kind of loss picks the channel that rises, and it picks it on an untrained task too.

A verdict-loss variant gives the mirror image at the extreme. On the same size task the head adapter raises the continuous logit from 0.828 to 0.988. At the same time its generated verdict answers on only 24.7 percent of the images, and every one of the 79 readable answers says normal. On the earlier data version the same adapter answered every image and called all 320 normal, for a verdict accuracy of exactly 0.500. We record this as a collapse of the verdict channel, not as a score. Every logit sits above zero, so the sign readout is pinned at 0.5 while the continuous logit still ranks, the identity of Section 3, with values in Appendix E.

5 UNTRAINED CONDITIONS: THE REPRESENTATION MOVES, THE ANSWER STAYS AT CHANCE

Holding one domain out of an answer-axis alignment loss raises that domain's $\cos(\Delta\mu, d)$ to 0.556, 0.563, and 0.729 for perspective, liquid, and tower, while the free-answer pair-rank accuracy on the same domain stays at 0.463, 0.500, and 0.500 (Table 4, Figure 5). Where that value is exactly 0.500 the trained model calls every held-out image violated, which gives 0.5 by the identity of Section 3, and its first-token logit sits above zero on 96.9 to 100 percent of those rows, so the sign readout is pinned there as well. A verdict-loss control trained on the same nine domains leaves the held-out cosine near zero, between -0.046 and 0.050, and the base model sits between -0.022 and 0.064. The same channel that stays at chance on the held-out domain rises to 0.807, 0.746, and 0.757 on the nine trained domains. The continuous logit on the held-out domain rises on two of the three and falls on the third, which is the pattern the map predicts for a cell the loss did not pass through.

Table 2: A score-order reward raises the score channel in eleven of twelve domains and leaves the logit and verdict channels of the same answers below the base model in nine and eight. Pair-rank accuracy on the test pairs of each domain, base model and score-reward variant, mean ± SD over three training seeds, chance 0.5. The physical row pools the ten domains without water and rotation. Earlier data version. † marks domains the frozen DINOv2 feature separates at 1.000.

	Score		Logit		Verdict	
Domain	base	score reward	base	score reward	base	score reward
shadow	0.500	0.521 ± 0.059	0.438	0.508 ± 0.101	0.500	0.508 ± 0.007
support†	0.338	0.558 ± 0.056	0.463	0.271 ± 0.126	0.588	0.471 ± 0.058
reflection	0.500	0.521 ± 0.036	0.725	0.558 ± 0.063	0.500	0.479 ± 0.036
thermal†	0.487	0.688 ± 0.213	0.525	0.300 ± 0.195	0.600	0.417 ± 0.076
water†	0.412	0.704 ± 0.213	0.575	0.192 ± 0.126	0.675	0.329 ± 0.151
refraction	0.325	0.496 ± 0.106	0.637	0.554 ± 0.026	0.650	0.533 ± 0.058
liquid	0.537	0.521 ± 0.031	0.450	0.658 ± 0.134	0.450	0.496 ± 0.007
perspective	0.537	0.542 ± 0.007	0.525	0.521 ± 0.038	0.550	0.446 ± 0.040
shadow length	0.512	0.554 ± 0.047	0.500	0.429 ± 0.089	0.525	0.487 ± 0.022
highlight	0.475	0.496 ± 0.031	0.325	0.525 ± 0.057	0.400	0.521 ± 0.026
tower†	0.475	0.863 ± 0.107	0.450	0.233 ± 0.138	0.425	0.212 ± 0.141
rotation	0.588	0.704 ± 0.088	0.512	0.425 ± 0.076	0.500	0.500 ± 0.000
physical (400 pairs)	0.469	0.576 ± 0.049	0.504	0.456 ± 0.063	0.519	0.457 ± 0.021
domains above base		11 of 12		3 of 12		3 of 12

Table 3: On an untrained size task the score-reward variant again raises only the score channel, and the verdict-loss head raises the logit while its generated verdict collapses. Pair-rank accuracy on 160 size-control test pairs, chance 0.5. The score-reward row is mean ± SD over three seeds. Response rate is the share of images with a parseable verdict. Current data version except where marked.

Model	Score	Logit	Verdict	Response rate
base	0.447	0.303	0.500	1.000
score reward (three seeds)	0.645 ± 0.130	0.169 ± 0.019	0.526 ± 0.040	1.000
base, verdict prompt		0.828	0.706	1.000
verdict-loss head		0.988	no readable pair	0.247
verdict-loss head, earlier data		0.906	0.500	1.000

The representation channel is the one channel that moves off the trained conditions, and it moves without any loss through the output. That is why the map is sharpest on output channels, and why we never count a rising probe or cosine as a rising answer. Appendix D collects the other cases in which the representation channel rose with no loss through it.

A held-out camera viewpoint gives the same shape on one channel. A verdict-loss variant trained on shadow violations at the original viewpoint reaches 0.900 on unseen scenes at that viewpoint. Turning the camera by 45 degrees leaves 0.579 and 0.629, and 90 degrees leaves 0.508, within the base model's 0.47 to 0.52 (Table 5, Figure 6a). We report the three angles as measured and do not fit a curve to them.

A held-out relation direction gives it too. In the shadow domain the violation can rotate the light source or the cast shadow. An variant trained with only the light rotated reaches 0.904 on light-only test pairs and 0.492 on shadow-only pairs. An variant trained with only the shadow rotated reaches 0.900 on shadow-only pairs and 0.496 on light-only pairs. The base model gives 0.512 and 0.483. Each variant rises where it was trained and stays at base on the other direction. Both variants also stop raising a false alarm when light and shadow rotate together, 0.008 and 0.017 against 0.367 for the base model, so the trained direction is the violation and not the rotation.

Put next to each other, the three conditions say one thing. The output cell rises only on the conditions the loss saw. The representation cell can rise on conditions the loss never saw. Whether the model can be made to answer on those conditions from that representation is not something these variants tested.

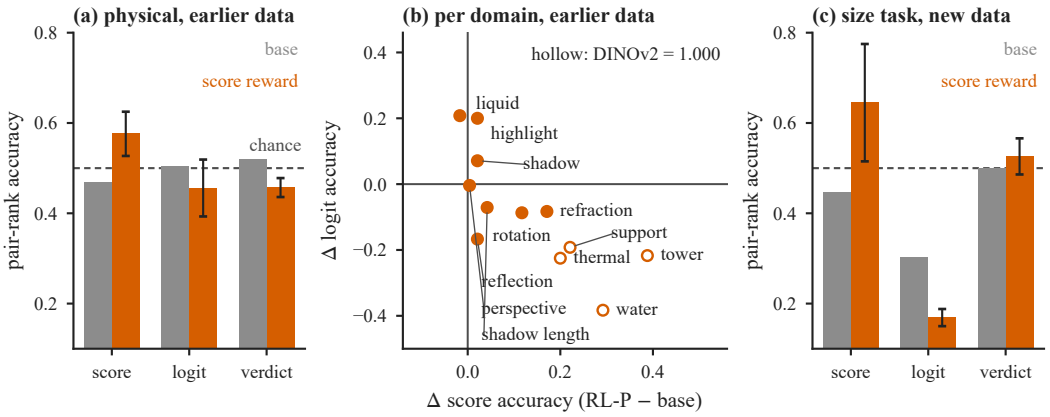

Figure 4: A score-order reward moves the score channel up and the logit and verdict channels down. (a) Pair-rank accuracy of the three output channels before and after training, mean ± SD over three seeds, dashed line chance 0.5, earlier data version. (b) Per-domain change in score against change in logit, hollow markers are the four domains a frozen DINOv2 separates at 1.000. Spearman between score gain and logit drop is 0.841 over the ten physical domains and 0.630 after removing the four hollow domains. (c) The same adapter on the untrained size task, 160 pairs.

Table 4: On a domain held out of the alignment loss the representation aligns with the answer axis while the free answer stays at chance. Rows are the held-out domain. Left block, $\cos(\Delta\mu, d)$ on the held-out test pairs, range -1 to 1. Right block, pair-rank accuracy of the generated answer on the same 40 pairs, chance 0.5, and a value of exactly 0.500 is a constant verdict (Section 3). The alignment variant's free answer on its nine trained domains is 0.807, 0.746, and 0.757 for the three rows. One seed.

	Representation, $\cos(\Delta\mu, d)$			Free answer, pair-rank		
Held-out domain	alignment	verdict SFT	base	alignment	verdict SFT	base
perspective	0.556	0.050	−0.022	0.463	0.500	0.500
liquid	0.563	−0.046	0.064	0.500	0.500	0.562
tower†	0.729	0.048	0.008	0.500	0.500	0.475

6 Related Work

Physical plausibility as a test for vision-language models. That vision-language models miss physical violations a person sees at once is established. Benchmarks built from designer-made commonsense violations (Bitton-Guetta et al., 2023), physical-world question sets (Chow et al., 2025), and possible versus impossible synthetic videos (Bordes et al., 2025; Bai et al., 2025b) all report models near chance where people are near ceiling. Static image sets label violations of physical common sense (Wu et al., 2025b), and image-editing evaluations grade the optical realism of generated pictures (Pu et al., 2025). Image generators get shadows and projective geometry wrong too (Sarkar et al., 2023), and forensics has used that same geometry as ground truth for a decade (Kee et al., 2013). In each of these the failure is measured with one readout, most often accuracy on a choice question. Two things about our testbed matter for what follows. A violated image always comes with a normal counterpart from the same renderer and the same path, as in minimal-change pairs (Awal et al., 2024). And the renderer knows the size θ of the violation it made (Section 2).

Training models to judge plausibility. Fine-tuning does raise plausibility judgments. What gets reported is one channel. TRAVL fine-tunes video-language models on physics implausibility and gives classification accuracy (Motamed et al., 2025), and HCM-GRPO runs group-relative policy optimization on a compact model for the same kind of judgment (Hu et al., 2025). Objectives built on counterfactual or minimal-contrast pairs report task accuracy (Zhang et al., 2025; Wu et al., 2025a). Preference training on prior-versus-vision pairs reports how much more the model relies on the image (Luo et al., 2025). Whether a second readout of the same trained model moved along with

Table 5: Gains follow the trained camera viewpoint and the trained relation direction and fall to chance off them. Top, free-answer pair-rank accuracy of a verdict-loss variant trained on shadow violations at the original viewpoint, evaluated as the camera turns. Bottom, variants trained on one relation direction evaluated on both directions, unseen scenes. Chance 0.5, one seed.

Viewpoint	$-45°$	$0°$ (trained)	$+45°$	$+90°$
verdict SFT variant	0.629	0.900	0.579	0.508
base		0.47 to 0.52		

Relation direction	light-only pairs	shadow-only pairs
trained light-only	0.904	0.492
trained shadow-only	0.496	0.900
base	0.512	0.483

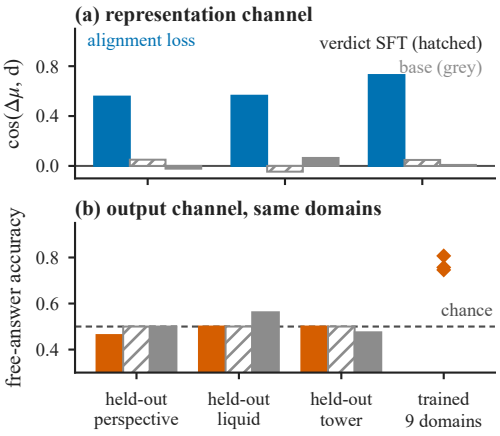

Figure 5: On a held-out domain the representation aligns with the answer axis while the free answer stays at chance. Top, $\cos(\Delta\mu, d)$ on the held-out domain for the alignment variant, a verdict-SFT control, and the base model. Bottom, free-answer pair-rank accuracy on the same domain, dashed line chance 0.5, and the alignment variant's accuracy on its nine trained domains as diamonds. A bar at exactly 0.500 is a constant verdict (Section 3). One seed.

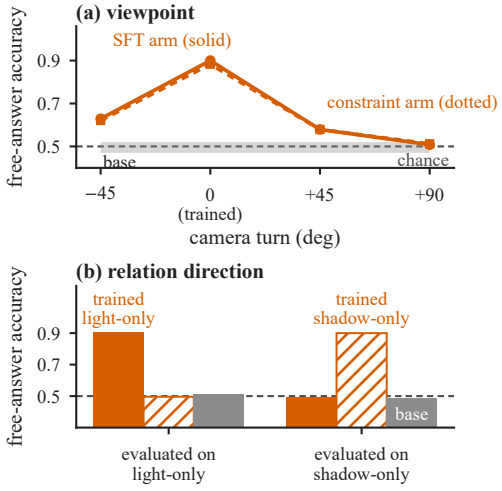

Figure 6: Gains follow the trained viewpoint and the trained relation direction and fall to chance off them. Top, free-answer pair-rank accuracy of the shadow SFT variant (solid), the feature-constraint variant of Appendix F (dotted), and the base model (band) as the camera turns 0, 45, and 90 degrees. Bottom, variants trained on one relation direction evaluated on both directions. Dashed line chance 0.5, one seed.

the reported one is a question none of them asks. We train three kinds of loss on the same pairs and read four channels every time.

The same model gives different answers on different channels. Readouts already disagree before anyone trains anything. The first-token probability of a model and the text it then writes can point to different answers (Wang et al., 2024). Change the choice format and the answer changes (Zheng et al., 2023). A VLM judge will rank two pictures it cannot score (Kumar et al., 2026), and asked for a continuous geometric quantity it anchors instead of grading (Zhang et al., 2026). On the inside, a probe can read out what the output leaves unsaid (Burns et al., 2023; Mallen et al., 2024), and for text models there is now a map of when the probe wins over the answer (Orgad et al., 2025; Villuri et al., 2026). In VLMs the visual representation holds evidence the language output passes over (Fu et al., 2025; Rahmanzadehgervi et al., 2024), even though frozen visual features encode much of the 3D scene (El Banani et al., 2024). One recent account calls this an arbitration failure, not perceptual blindness (Nooralahzadeh et al., 2026), and another shows models encoding what they never express (Pramono et al., 2026). Decodable is not the same as grounded (Liao

et al., 2026), relational violations can sit past what a linear probe reaches (Vompa et al., 2025), and switching on a reasoning channel can make spatial answers worse (Kancheti et al., 2026). All of this describes the gap at rest. We record where it opens after training. The channel a loss passed through rises. The others are under no obligation to follow.

Gains that stay where the loss was. Gains from post-training do not travel for free. Fine-tuning can warp pretrained features and then lose out of distribution (Kumar et al., 2022), and supervised fine-tuning and reinforcement learning do not generalize the same way (Chu et al., 2025; Kirk et al., 2024). RL with verifiable rewards rarely gets past what the base model could already sample (Yue et al., 2025), though it forgets less (Shenfeld et al., 2025). It touches a small subnetwork (Mukherjee et al., 2025), and it can improve under a reward that has nothing to do with correctness (Shao et al., 2025). Rule-based rewards on VLMs invite reward hacking (Shen et al., 2025). In all of these the condition axis is a task or a data distribution. Ours is a physical condition held out on the same renderer, a domain, a camera viewpoint, or the direction of a relation, and we place representation and output channels side by side under each. The condition-side failures are known. The joint map of channel and condition, with the loss path as the predictor, is what this paper adds.

7 LIMITATIONS

The channel-axis evidence in Table 2 comes from an earlier data version of the testbed. The size-task rows, the held-out domain, viewpoint, and relation results come from the current version, and the retraining of the score-reward variant on the current version was not finished at submission. Every result outside Table 2 and Table 3 is one training seed.

The map is a report on one model, Qwen2.5-VL-7B, and on rendered stimuli. A second model we tried exposes only a head and a logit channel, so it cannot fill the matrix, and we make no claim about photographs or about people.

We claim no guarantee. The observation that supervised cells rose in every variant is not a theorem, and a two-pair construction in Appendix C shows a pairwise logistic loss falling while pair-rank accuracy falls with it. No argument we can defend predicts from the geometry of a readout whether an unsupervised channel rises or falls, so we report the sign per domain and stop there. Equation equation 1 gives a sufficient condition for a gain to carry from one channel to another, that the two agree in the sign of the pair difference on most pairs before and after training. We did not measure that agreement on our pairs, so the per-domain signs in Table 2 are reported without it. The correspondence between a camera rotation and a rotation in the representation is likewise not established, and the three viewpoint angles are reported as points.

After verdict-token training the score channel drops to 0.193 on pair-rank, and the row-level rank correlation between score and logit moves from -0.069 at base to -0.803. This is consistent with a change of polarity in how the model uses the number. It does not establish that the two channels carry the same information, so we keep this cell out of the map and describe it in Appendix G.

The representation channel can rise with no loss through it, so the map is sharpest on output channels. Why an output cell stays at chance, whether the evidence stops at the threshold or before the generation, is not something these variants separate. The prescription follows from the observation, and two reward-swapped variants from one checkpoint agree with it on one seed.

8 CONCLUSION

On ImPhysible, every (channel, condition) cell the training loss passed through rose, in every variant we trained. The other output channels and the untrained conditions rose, stayed at chance, or moved the other way, and none of our variants made them rise reliably. If you will read the score, put the loss through the score. If you will read the verdict, train the verdict token. Then test the conditions you did not train on, separately, because no variant of ours made an untrained cell rise reliably. For a benchmark the same fact is a warning. One accuracy reported after fine-tuning tells you which cell rose and nothing about the cell next to it. Two measurements come next, a reward that passes through two channels at the same time, and this map redone on the current data version with three seeds.

AI USE STATEMENT

In this work, we used generative AI tools (Claude, GPT and Gemini models) to implement methods, including evaluation, probing and training scripts and the job files that ran them, to clean and reformat dataset manifests and splits, to give feedback on and help refine the experimental design and the hypotheses it tests, to help organise the conceptual framing of the study (the distinction between a missing capability and a failure of elicitation), to help interpret results by recomputing metrics from raw result rows and summarising them, and to assist with translation between Korean and English. We have not used generative AI tools to generate synthetic data sets: all images are produced by a procedural renderer from scene parameters, and every label comes from the renderer, not from a model. We did not use them to formulate or prove mathematical claims or for qualitative data analysis, and these tasks are not applicable to this work. Additionally, we used generative AI tools to search for and summarise related work, to draft parts of the paper, to edit the text for readability, and to prepare figures and tables. The vision-language models evaluated in this paper are objects of study and not assistants. We have reviewed all AI-assisted work. Code written with AI assistance was reviewed by an author before being merged and is covered by automated tests. Every reported number was recomputed by script from the stored raw model outputs. Every cited work was checked against its original text by an author, and AI-suggested references we could not verify were removed. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All stimuli are procedurally rendered scenes of inanimate objects. No human subjects, personal data, or scraped images were used. The testbed can be used to evaluate detectors of manipulated imagery, and we release it for that purpose.

REPRODUCIBILITY STATEMENT

Section 2 and Appendix A describe the renderer, the domains, and the scene-level split. Appendix B lists every training variant with its configuration, and Appendix H gives per-seed values. The supplementary material contains the code, the pair manifests, the raw model outputs behind every table, and the scripts that recompute each number from them.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. VisMin: Visual minimal-change understanding. *arXiv preprint arXiv:2407.16772*, 2024.

Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025a.

Zechen Bai et al. Impossible videos. In *ICML*, 2025b. arXiv:2503.14378.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Nitzan Bitton-Guetta et al. Breaking common sense: WHOOPS! a vision-and-language benchmark of synthetic and compositional images. In *ICCV*, 2023. arXiv:2303.07274.

Florian Bordes et al. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025.

Collin Burns et al. Discovering latent knowledge in language models without supervision. In *ICLR*, 2023. arXiv:2212.03827.

Wei Chow et al. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *ICLR*, 2025. arXiv:2501.16411.

Tianzhe Chu et al. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *ICML*, 2025. arXiv:2501.17161.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, et al. Probing the 3D awareness of visual foundation models. In *CVPR*, 2024. arXiv:2404.08636.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025.

Robert Geirhos et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. arXiv:2004.07780.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *EMNLP*, 2019. arXiv:1909.03368.

Hu et al. Physical plausibility reasoning via HCM-GRPO: Empowering compact model for superior performance. *arXiv preprint arXiv:2511.10055*, 2025.

Edward J. Hu et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. arXiv:2106.09685.

Kancheti et al. Chain-of-thought degrades visual spatial reasoning capabilities of multimodal LLMs. *arXiv preprint arXiv:2604.16060*, 2026.

Eric Kee, James F. O'Brien, and Hany Farid. Exposing photo manipulation with inconsistent shadows. *ACM Transactions on Graphics*, 32(3), 2013. doi: 10.1145/2487228.2487236.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, et al. Understanding the effects of RLHF on LLM generalisation and diversity. In *ICLR*, 2024. arXiv:2310.06452.

Kumar, Tayebati, Naik, Krishnan, and Trivedi. VLM judges can rank but cannot score: Task-dependent uncertainty in multimodal evaluation. *arXiv preprint arXiv:2604.25235*, 2026.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022. arXiv:2202.10054.

Liao, Shen, Cao, and Chua. Decodable is not grounded: A vision-ablation arbiter for VLM spatial reasoning. *arXiv preprint arXiv:2606.31257*, 2026.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, et al. Probing visual language priors in VLMs. In *ICML*, 2025. arXiv:2501.00569.

Alex Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. Eliciting latent knowledge from quirky language models. In *ICLR*, 2024. arXiv:2312.01037.

Saman Motamed et al. TRAVL: A recipe for making video-language models better judges of physics implausibility. *arXiv preprint arXiv:2510.07550*, 2025.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tur, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models. *arXiv preprint arXiv:2505.11711*, 2025.

Farhad Nooralahzadeh et al. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts. *arXiv preprint arXiv:2604.09364*, 2026.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, et al. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *ICLR*, 2025. arXiv:2410.02707.

Pramono et al. TRAPSBench: Vision-language models encode but fail to express epistemic restraint. In *COLM*, 2026. arXiv:2608.13167.

Yuandong Pu, Le Zhuo, Songhao Han, Jinbo Xing, et al. PICABench: How far are we from physically realistic image editing? *arXiv preprint arXiv:2510.17681*, 2025.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words. *arXiv preprint arXiv:2407.06581*, 2024.

Ayush Sarkar et al. Shadows don't lie and lines can't bend! generative models don't know projective geometry...for now. *arXiv preprint arXiv:2311.17138*, 2023.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, et al. Spurious rewards: Rethinking training signals in RLVR. *arXiv preprint arXiv:2506.10947*, 2025.

Zhihong Shao et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhan Shen et al. VLM-R1: A stable and generalizable R1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Idan Shenfeld et al. RL's razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*, 2025.

Villuri, Shaik, and Doboli. When do internal probes beat reading the answer? miscalibrated readouts and behavior-concealed knowledge in language models. *arXiv preprint arXiv:2609.04582*, 2026.

Vompa, Tammet, and Vaishnav. Beyond the linear separability ceiling: Aligning representations in VLMs. *arXiv preprint arXiv:2507.07574*, 2025.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My answer is C": First-token probabilities do not match text answers in instruction-tuned language models. *arXiv preprint arXiv:2402.14499*, 2024.

Wu, Sun, Wen, and Haber. Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. *arXiv preprint arXiv:2502.13928*, 2025a.

Wu, Wang, Luo, Chen, and Zheng. Detecting violations of physical common sense in images: A challenge dataset and effective model. In *ACM Multimedia*, 2025b. doi: 10.1145/3746027.3755124.

Yang Yue et al. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Zhang, Cai, Fan, Wang, and Wang. CF-VLM: Counterfactual vision-language fine-tuning. *arXiv preprint arXiv:2506.17267*, 2025.

Zhang et al. Anchored, not graded: Vision-language models fail at slant-from-texture perception. *arXiv preprint arXiv:2606.06714*, 2026.

Chujie Zheng et al. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882*, 2023.

A TESTBED DETAILS

Each physical domain defines one violation with a continuous size θ in the units of Table 1 and ten levels of θ between the minimum and the maximum listed there. A scene is a fixed arrangement of camera, light, and objects. For each scene and each level the renderer produces the normal image and the violated image through the same pipeline, changing only the violation parameter. Pairs are split by scene with a fixed seed so that the 480 test pairs come from 36 scenes no training pair uses. Shadow, support, and reflection reuse images from an earlier rendering of those domains, selecting ten scenes and ten levels. The current data version re-rendered 1,249 pairs and passed an automatic check of file hashes, crops, and left-right balance.

Figure 7 shows one pair from every domain. The three control domains carry no physical violation and serve two purposes. Counting, with painted area held equal and half the labels flipped, checks

that a low-level pixel cue cannot drive the frozen feature, and the frozen DINOv2 feature indeed sits at 0.500 there. Size, with total screen area held equal, is the untrained task of Table 3.

The shortcut audit fits a linear head on frozen DINOv2 features under scene-level cross-validation, on the whole image, on a 16-pixel background strip, and on the violation region. Whole-image values are in Table 1. Background-only values reach 0.990 on thermal, 0.944 on support, and 0.917 on highlight, which is why model values on those domains are read with care.

Verdicts are parsed by one shared parser into violated, normal, or missing. Missing answers are never imputed. A pair is scored on a channel only when both images have a value on that channel, and the response rate is reported alongside.

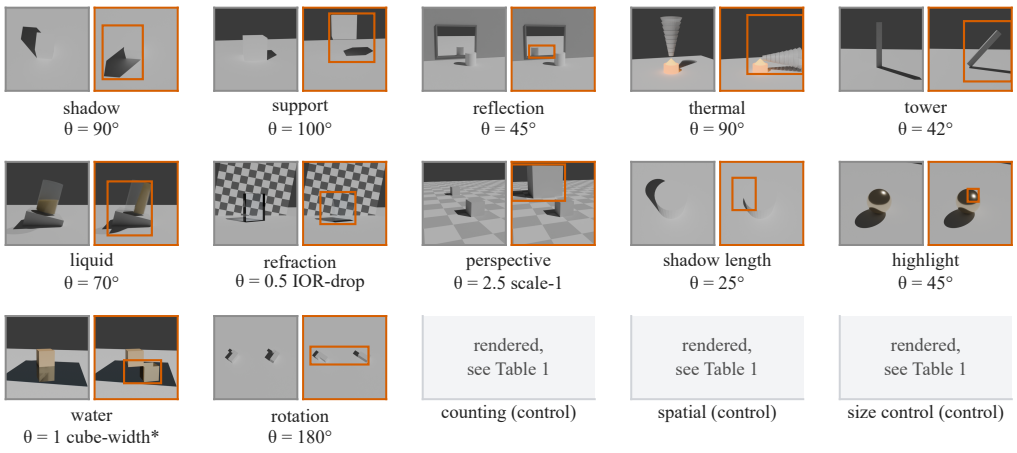

Figure 7: Twelve physical domains and three control domains share one design, a normal counterpart and a violated image from the same scene. Each cell shows the normal counterpart on the left and the violated image at the largest test θ on the right with the violation outlined. Water pairs are flagged for review in the rendering audit (asterisk). The three control domains are listed in Table 1.

B TRAINING VARIANTS AND READOUTS

Table 6 lists the variants. All variants start from Qwen2.5-VL-7B and train LoRA adapters of rank 64 on the language model with the vision encoder frozen. The score-reward variant samples several answers per image, scores each by whether its written number orders the violated image above its counterpart, and applies the group-relative advantage (Shao et al., 2024). The verdict-loss variants minimize cross-entropy on the answer token. The alignment variant minimizes one minus the cosine between the mean hidden-state difference of violated and normal training images and a fixed answer axis d, which is the difference of the unembedding vectors of the two answer tokens. The head adapter trains a pairwise ranking head on the pooled hidden state. Three training seeds are reported where the variant was trained three times.

Table 6: Training variants, the channel and conditions their loss passed through, and the readouts we measure on each.

Variant	Loss passes through	Trained conditions	Seeds
score reward (GRPO)	score	12 domains, original view	3
verdict SFT, all domains	logit, verdict	12 domains	1
verdict SFT, shadow	logit, verdict	shadow, one view, one direction	1
alignment, leave one out	representation	9 of 12 domains	1 per held-out
head adapter	logit, head	12 domains	1

Readouts. Score is the integer the model writes after a prompt asking it to rate physical possibility from 0 to 100, read as 100 minus the number so that higher means more violated. Logit is the first-

token log-probability difference of *impossible* minus *possible* on a yes-or-no prompt. Verdict is the generated answer to the same prompt parsed to violated or normal. Probe is a logistic regression on the pooled hidden state of a fixed layer, fit on training scenes and read on test scenes. Cosine is $\cos(\Delta\mu, d)$ with $\Delta\mu$ the mean hidden-state difference over test pairs of a condition.

C PAIR-RANK ACCURACY: THREE PROPERTIES, THEIR PROOFS, AND WHAT THEY DO NOT GIVE

Fixed comparisons. Use equal weights on $n > 0$ pairs. Superscripts 0 and 1 denote the before and after checkpoints, and $\mathrm{sgn}(0) = 0$. Channels compared in one expression use the same pairs. For a checkpoint curve, keep the intersection of valid pairs across all displayed checkpoints and report each channel's original coverage alongside this subset. This does not replace the full evaluation and does not establish population generalization. With stochastic decoding the statements apply to the saved realizations, and they hold in expectation under a fixed coupling and a common validity event. The score is the parsed generated number, not an expectation over digit tokens. The logit channel uses a fixed position and token contrast. Verdicts are parsed as violated, 1, or normal, 0. None of the properties equates a fitted probe with an optimal probe or with a cosine of mean representations. A loss passes through a channel when that channel's output enters a training target or reward. It says nothing about which parameters moved.

Property 1, gains within a cell. If every pair difference is nondecreasing, A cannot decrease, and it increases exactly when at least one pair moves upward between negative, zero, and positive. Proof. Since q is nondecreasing, $m_i^1 \geq m_i^0$ gives $q(m_i^1) \geq q(m_i^0)$ pair by pair, and averaging proves the nondecrease. The average increases precisely when some summand does, which happens when a pair moves from negative to zero or positive, or from zero to positive. Without the premise the exact change is

$$\mathrm{A}^1 - \mathrm{A}^0 = \frac{1}{2n} \sum_{i=1}^{n} \big[\mathrm{sgn}(m_i^1) - \mathrm{sgn}(m_i^0)\big], \tag{2}$$

which handles discrete scores and verdict ties. An increase in the mean difference says nothing about the sign of this sum. For the two-type counterexample of Section 3, normal readouts are zero and the two types each have weight one half. The expected loss $\ell(z) = \log(1 + e^{-z})$ falls from 0.6443966601 to 0.5288291738 while accuracy falls from 1 to 1/2. There is no sampling and no distribution change. The example needs a heterogeneous update that sacrifices one near-boundary pair while improving the other enough. A pointwise loss decrease under this strictly decreasing ℓ would rule it out, and an average decrease does not. This is a counterexample about a surrogate loss, not a claim that SFT or GRPO minimizes this exact loss. Calibration bounds concern excess risks and do not assert monotone accuracy between arbitrary checkpoints (Bartlett et al., 2006).

Property 2, transfer between channels. For two channels a and b on the same pairs, with e_t the share of pairs whose differences differ in sign at checkpoint t, the gains obey equation equation 1. Proof. If two pair differences have the same sign, their q values agree. Otherwise the two q values differ by at most one. Hence $|\mathrm{A}_b^t - \mathrm{A}_a^t| \leq e_t$, and the triangle inequality gives the bound on the gains. The same step gives $\mathrm{A}_b^1 - \mathrm{A}_b^0 \geq \mathrm{A}_a^1 - \mathrm{A}_a^0 - e_0 - e_1$, which is the transfer condition of Section 3. The condition is sufficient and not necessary, and large disagreement permits either transfer or its failure. A strictly increasing relation between the two readouts at each checkpoint is enough for $e_0 = e_1 = 0$. A Spearman correlation near 1 or -1 is not that relation. Reversing score polarity after inspecting test results changes the channel being evaluated. Nothing in this proof couples different domains, and even within one domain shared parameters permit opposite changes. For one pair, take normal outputs zero and a shared scalar parameter u with margins $m_a(u) = u$ and $m_b(u) = -u$. Moving u from -1 to 1 improves a and worsens b. For a held-out domain, the same channel can have margin u in the training domain and $-u$ in the held-out domain. These constructions rule out a deduction from shared parameters alone. They do not assert that a particular trained LoRA model realizes these functions. The positive statements above need only the observed pair signs, so they apply to any realized SFT or GRPO update.

Identity 3, binary verdicts. On complete pairs, binary pair-rank accuracy equals $(1 + \mathrm{TPR} - \mathrm{FPR})/2$, and a constant output gives 1/2. Proof. For $v_v, v_n \in \{0, 1\}$, the four combinations

give $q(v_v - v_n) = (1 + v_v - v_n)/2$. Average over pairs, with $\mathrm{TPR} = n^{-1}\sum_i v_{v,i}$ and $\mathrm{FPR} = n^{-1}\sum_i v_{n,i}$ on those same pairs. A constant binary output makes the two rates equal. For thresholded logits $H = \mathbf{1}[r_L > 0]$ and generated verdicts V, the same identity and the triangle inequality give $|\mathrm{A}_V - \mathrm{A}_H| \le d$, where d is the share of the $2n$ image-level decisions on which they disagree. A small d implies similar accuracies, and the converse fails. For two pairs let (H_n, H_v) be $(0, 1)$ and $(1, 0)$ and set $V = 1 - H$. Both accuracies are 1/2 although the two disagree on every image. Even increasing every logit pair difference need not improve thresholded verdicts. A pair with logits $(-1, 1)$ moving to $(1, 4)$ has a difference rising from 2 to 3, its continuous pair accuracy stays 1, and its thresholded accuracy falls from 1 to 1/2. A common within-pair additive offset leaves the continuous order unchanged and can change the thresholded order. That a pair shares such an offset is an algebraic assumption, not a consequence of using the same rendering pipeline.

What the assumptions do not give. A positive mean projection of an untrained condition's violation direction onto a readout does not imply pair-rank accuracy above chance. Margins $-1/2$ and $3/2$ with equal weight have a positive mean and accuracy 1/2. A margin distribution symmetric about zero has accuracy 1/2, ties included, because $q(z) + q(-z) = 1$, and the converse is false. No correspondence between a camera angle and an angle in the representation is assumed or established anywhere in the paper. A constant reward removes the GRPO advantage term only when zero-variance advantages are set to zero, and the KL term, other losses, and optimizer state can still move the parameters. The three properties use measured endpoints and need no zero-update assumption. They make no improvement claim for probes or for cosines of mean vectors. What the map carries is therefore the observations of Sections 4 and 5, and the one-sided summary that supervised cells rose in every variant we trained while unsupervised cells did all three things.

D THE REPRESENTATION CHANNEL RISES WITHOUT A LOSS THROUGH IT

Three other measurements show the representation channel moving where no output loss passed. A probe on the connector output of the base model gives 0.817 on the pooled physical set, and the same probe on four trained variants gives 0.869 to 0.900. A probe fit on shadow pairs and read on the other eleven domains rises from 0.580 on the base model to 0.780 on the head adapter. And on the held-out domain of the alignment variant, $\cos(\Delta\mu, d)$ rises to between 0.556 and 0.729 (Table 4). Each of these is a representation readout that refits or reprojects at test time, and none is counted as an output gain. Probe controls follow Hewitt & Liang (2019) and Alain & Bengio (2016), and the background baseline follows the shortcut audit of Appendix A (Geirhos et al., 2020). Relational violations can sit beyond a linear probe (Vompa et al., 2025), and decodable is not grounded (Liao et al., 2026). That is a second reason we do not read these values as answers.

E CONTINUOUS LOGITS, THEIR SIGNS, AND GENERATED VERDICTS

We compare continuous logits, their binary signs, and generated verdicts on the same valid pairs. On 480 physical test pairs of the head adapter, the continuous logit ranks at 0.885, its sign at 0.676, and the generated verdict at 0.670. The gap from the continuous logit to the sign, 0.209, is the share lost when a continuous value is collapsed to a sign. The gap from the sign to the verdict is 0.006, and on 10 of 960 rows the generated verdict disagrees with the logit sign. Three other adapters give the same shape, and the base model gives -0.011 and -0.006 with 85 of 880 rows disagreeing. We read these as three readouts of the same answer token measured at three resolutions. We do not conclude from the small gap that the generation mechanism is a threshold on the logit, because a small difference of accuracies does not bound the row-level disagreement. On the size task of Table 3 every logit is positive for every adapter and the base model, so the sign readout is pinned at 0.5 and only the continuous logit ranks.

F LOSSES THAT DO NOT PASS THROUGH AN OUTPUT CELL

Two variants put a loss near an output cell without passing through it, and neither raised the cell. A verdict-SFT variant followed by a GRPO stage whose reward was flat within each group moved the logit by $+0.03$ and the verdict by -0.01. With the standard deviation of a constant reward handled

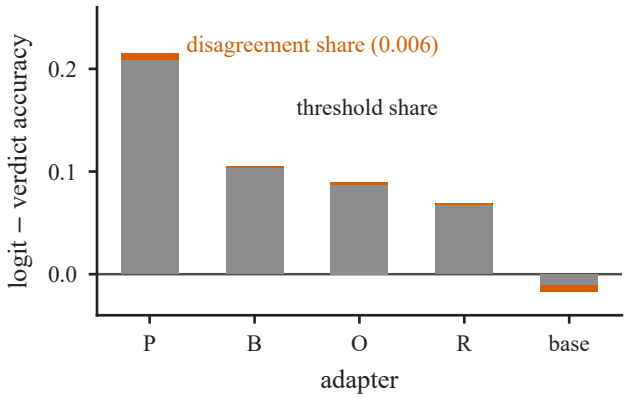

Figure 8: The gap between the continuous logit and the generated verdict is mostly the share lost at the sign. Stacked bars per adapter, sign share in grey, sign-to-verdict share in orange, zero line marked.

safely, the advantage term of that stage is zero, and any remaining update would have to come from the KL term or the optimizer state, which we did not isolate. A feature-constraint variant on the shadow viewpoint task, trained to keep the hidden-state difference fixed under camera rotation, gives 0.883 at the trained viewpoint and 0.621, 0.579, and 0.512 at -45, $+45$, and $+90$ degrees, the same shape as the verdict-SFT variant in Table 5. A symbol pre-training variant that first taught the answer vocabulary reached 0.603 against 0.680 for direct verdict SFT. None of these numbers is a claim about the mechanism. They are the cells where the loss did not go.

G SCORE-CHANNEL POLARITY AFTER VERDICT TRAINING

Asked the score question after verdict-token training on all twelve domains, the verdict-SFT variant gives a score pair-rank of 0.193 from a base of 0.471, with the logit at 0.820 from 0.501 and the verdict at 0.719 from 0.527, with every image answered. The score falls below base on eleven of twelve domains. The row-level Spearman correlation between the written score and the logit difference is -0.069 on the ten physical domains at base and -0.803 after training, negative on all ten domains between -0.136 and -0.912. The negative rank association is consistent with a polarity change, but does not establish information equivalence or an exact monotone reversal. We therefore keep this cell out of the main map and do not count it as evidence for or against channel dissociation.

The same checkpoint is the starting point of two further variants that differ only in the reward (Table 7). A score-order reward takes the score channel from 0.193 to 0.843 on the ten physical domains, above the starting point on all ten. A verdict reward leaves it at 0.206, above the starting point on four domains and below on four. The logit and verdict channels, which the verdict loss had already raised to 0.820 and 0.719, stay between 0.760 and 0.859 under both rewards. Both variants are one seed and start from the earlier-version checkpoint, so we read them as a check on the prescription of Section 8 and not as a row of the map. Water and rotation are listed for completeness and are not in the physical ten of Table 2.

A score-reward variant trained on the current data version from a supervised checkpoint gives, on one seed, the same picture on the size-controlled pairs as Table

A second one-seed check on the current data version asks whether the generated verdict of the score-reward variant reads its own score. On the ten physical domains its verdict pair-rank accuracy is 0.816. Replacing only the score digits in the prompt with the values produced by the supervised checkpoint, and leaving everything else unchanged, lowers it to 0.740, close to the 0.719 of that checkpoint, and lowers the accuracy on pairs whose two answers differ from 0.981 to 0.878. The verdict therefore moves with the number it is given, which is the mechanism the map assumes when it treats the score and verdict channels as coupled through the text. Both checks are preliminary and will be repeated over seeds. eftab:size: the score channel moves from 0.509 to 0.950, the logit

channel from 0.438 to 0.359, and the generated verdict from 0.778 to 0.566. This run finished after the results in the main text were fixed and is reported here as a preliminary check that the sign pattern holds on the current data.

Table 7: Two variants trained from one verdict-trained checkpoint with the reward swapped. Each cell is pair-rank accuracy on score / logit / verdict, 40 test pairs per domain, one seed, earlier data version. The score-order reward raises the score channel on every physical domain and the verdict reward does not, while the logit and verdict channels stay near where the verdict loss left them.

Domain	Base	Start (verdict SFT)	Score reward	Verdict reward
shadow	0.500 / 0.438 / 0.500	0.412 / 0.537 / 0.550	0.650 / 0.700 / 0.662	0.487 / 0.713 / 0.500
support	0.362 / 0.463 / 0.588	0.113 / 0.975 / 0.925	0.950 / 1.000 / 0.900	0.062 / 1.000 / 0.925
reflection	0.500 / 0.725 / 0.500	0.412 / 0.725 / 0.537	0.700 / 0.825 / 0.713	0.400 / 0.825 / 0.537
thermal	0.463 / 0.500 / 0.600	0.013 / 1.000 / 0.988	0.988 / 0.975 / 0.988	0.013 / 1.000 / 0.988
refraction	0.263 / 0.613 / 0.700	0.212 / 0.850 / 0.500	0.963 / 0.887 / 0.787	0.287 / 0.850 / 0.675
liquid	0.537 / 0.450 / 0.450	0.025 / 0.950 / 0.600	0.925 / 0.950 / 0.925	0.037 / 0.950 / 0.850
perspective	0.537 / 0.525 / 0.550	0.125 / 0.975 / 0.825	0.938 / 0.950 / 0.912	0.050 / 0.963 / 0.925
shadow length	0.500 / 0.438 / 0.525	0.425 / 0.575 / 0.588	0.625 / 0.525 / 0.588	0.375 / 0.562 / 0.588
highlight	0.575 / 0.412 / 0.438	0.188 / 0.613 / 0.675	0.688 / 0.650 / 0.688	0.350 / 0.725 / 0.613
tower	0.475 / 0.450 / 0.425	0.000 / 1.000 / 1.000	1.000 / 1.000 / 1.000	0.000 / 1.000 / 1.000
physical ten	0.471 / 0.501 / 0.527	0.193 / 0.820 / 0.719	0.843 / 0.846 / 0.816	0.206 / 0.859 / 0.760
water	0.362 / 0.275 / 0.625	0.000 / 1.000 / 1.000	1.000 / 1.000 / 1.000	0.000 / 1.000 / 1.000
rotation	0.588 / 0.512 / 0.500	0.362 / 0.425 / 0.787	0.838 / 0.463 / 0.775	0.625 / 0.575 / 0.787

H FULL TABLES

Per-seed values behind Table 2: the Spearman correlation between the domain-level score change and logit change is -0.361, -0.310, and -0.844 for the three seeds on the ten physical domains, and -0.270, -0.179, and -0.569 after dropping the four domains the frozen feature separates at 1.000. The seed-averaged values quoted in Section 4 are -0.841 and -0.630. Per-seed size-task values behind Table 3: score 0.681, 0.500, and 0.753, logit 0.150, 0.188, and 0.169, verdict 0.572, 0.506, and 0.500. The held-out continuous logit of the alignment variant in Section 5 is 0.425, 0.750, and 1.000 for perspective, liquid, and tower against 0.350, 0.500, and 0.438 for the base model. A second reward variant that trains the logit rather than the score, and a shared-start comparison of a score reward and a verdict reward, were running at submission and are not reported.