CARD: Diagnosing Belief-to-Action Routing Failures in Vision–Language Models

Souptik Majumdar, Fabian Kögel and Andreas Bulling
Universität Stuttgart
70569, Stuttgart
Germany
souptik.majumdar, fabian.koegel, andreas.bulling @vis.uni-stuttgart.de

Abstract

Linear probes and activation steering have uncovered that vision-language models (VLMs) internally represent mental states such as agents' beliefs, knowledge, and intentions. However, it is unclear whether and how these representations are used by downstream predictions along these axes. To close this gap, we introduce *Cross-Axis Routing Diagnostic* (CARD), which steers activations along one axis while measuring the response of a *different* axis's prediction. Applied to open-weight VLMs on *Relay Chain* – a new cooperative grid-world benchmark we propose – we diagnose a critical routing failure: models fail to incorporate belief representations into their next action prediction, effectively leaving valuable information about their partners unused.

1 Introduction

Vision-language models (VLMs) are increasingly used in human-AI cooperation, such as embodied robots, AI coaches, or driving co-pilots, that require Theory of Mind (ToM; Premack and Woodruff, 1978) to infer what the interaction partner might see, know or want. To diagnose whether such models actually represent such mental states of their partners, two families of diagnostic tools have emerged: *Linear probing* tests if and where states are encoded in the models, while *activation steering* has primarliy been used to verify whether probed directions causally drive the model's prediction and to control its outputs. However, they leave one question unanswered: Does the model's internal representation of a partner's mental state actually drive its actions?

Existing behavioural benchmarks (Gu et al., 2024; Bortoletto et al., 2025; Shi et al., 2025; Chen et al., 2024a) score outputs along three axes, belief, knowledge and intentions but without inspecting and connecting their internal representations to the model predictions. Thus, they cannot discriminate wrong actions that reflect an absent representation from those that reflect an unused one. Alternatively, probes and prior work in activation steering (Alain and Bengio, 2016; Belinkov, 2022; Hewitt and Liang, 2019; Elazar et al., 2021; Bortoletto et al., 2024; Räuker et al., 2023; Geiger et al., 2021; Wu et al., 2023b) access internal representations but apply both read-out and intervention only to the *same* axis. For example, a probe might find a belief representation, and steering confirms its influence on next belief prediction (same-axis) but does not test its influence on action prediction (cross-axis). Thus, both approaches do not diagnose the routing of mental representations to predictions across axes. This matters especially in cooperative tasks, where the correct action depends on the mental state representation of the partner agent, which the VLM may encode internally, yet doesn't use when predicting its next action.

To address this limitation, we make the following contributions:

1. We introduce *Cross-Axis Routing Diagnostic* (CARD), a novel approach to diagnose how mental state representations influence model predictions across axes by steering representations in one axis and assessing the prediction of another.

2. To evaluate CARD, we propose *Relay Chain*, a new multi-agent cooperative grid-world benchmark that extends the grid-world format of GridToM (Li et al., 2025) with a cooperative action axis.

3. Our experiments demonstrate critical routing failures across four open-weight VLM families, where the belief representation causally influences the belief and knowledge predictions but not the action prediction – despite the belief being linearly decodable at 78–94% from the same layer.

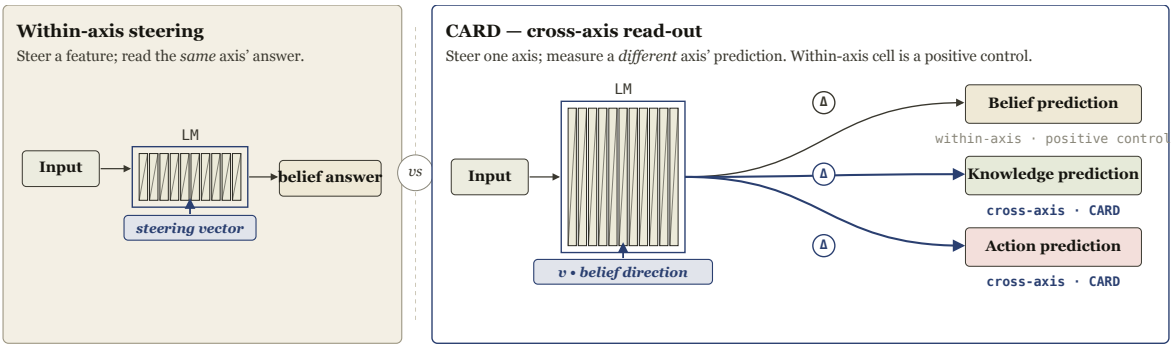

Figure 1: Overview of CARD. For each axis $a \in \{\text{belief, intent, knowledge}\}$ a linear probe on paired True-Belief / False-Belief activations at layer L yields a unit-norm direction v_a. At inference time, we hook the residual stream at the same layer and add αv_a (steering axis a, $\alpha \in \{-10, +10\}$), then read out the answer head for a *different* question axis $b \neq a$. A causally-routed feature moves the cross-axis read-out; an encoded-but-unread feature does not. Within-axis cells ($a = b$, diagonal) serve as positive controls.

2 Related Work

Behavioural ToM benchmarks. A large body of work evaluates ToM in foundation models through mental-state questions (Gu et al., 2024; Bortoletto et al., 2025; Shi et al., 2025; Xu et al., 2024; Wu et al., 2023a; Chen et al., 2024a; Xiao et al.; Chen et al., 2024c; Jin et al., 2024; Kim et al., 2023; Zhang et al., 2025). SimpleToM (Gu et al., 2024) demonstrated a gap between explicit and applied ToM on text LLMs while ToM-SSI (Bortoletto et al., 2025) extended to situated multi-agent interactions and reported a behavioural drop from percept to belief to intention. MindPower (Zhang et al., 2025) showed that ToM-aware prompting improved cooperative success rates in embodied VLM settings. CARD complements these by testing whether a model's internal mental-state representations actually drive its cooperative actions, not just whether its overall predictions are correct.

Probes and activation steering. Probes and standard (same-axis) activation steering to inspect models have a long history in text-only LLMs (Alain and Bengio, 2016; Belinkov, 2022; Elazar et al., 2021; Bortoletto et al., 2024; Räuker et al., 2023): A linear probe verifies that a feature is encoded in the residual stream, and steering amplifies the probe direction to test whether the model's answer to the same feature shifts. GridToM (Li et al., 2025) brought this protocol to multimodal grid-world ToM and was the closest prior multimodal-ToM work pairing probes with activation interventions. CARD extends the protocol from same-axis to a cross-axis matrix that tests whether a probe direction is read by a *different* prediction. Like Bortoletto et al. (2024), we use linear probes with Hewitt–Liang random-label controls (Hewitt and Liang, 2019) to verify that the recovered belief direction is structured rather than a probe-expressivity artefact.

Studying routing in VLMs Mechanistic interpretability tools to study full circuits from inputs to outputs that drive predictions include activation patching and causal mediation (Vig et al., 2020; Meng et al., 2022), activation-addition / steering vectors (Turner et al., 2024; Li et al., 2023; Rimsky et al., 2024; Zou et al., 2023), direct logit attribution (Wang et al., 2022), automated circuit discovery (Conmy et al., 2023), and linear-relation decoding (Geva et al., 2023; Hernandez et al., 2024). Recent work extended these to VLMs (Yang et al., 2026; Liu et al., 2025; Häon et al., 2025). Unlike distributed alignment search Geiger et al., 2021; Wu et al., 2023b, which gradient-searches a rotation aligning a high-level variable to a subspace, CARD uses a fixed pre-trained probe direction and conditions the read-out on a specific prediction. These techniques have largely targeted factual recall, refusal, or bias; CARD applies cross-axis routing diagnostics to mental-state representations in cooperative agents.

3 Method

Our *Cross-Axis Routing Diagnostic* (CARD) takes three steps: (1) identify a steering vector for a mental representation in one axis, e.g. belief; (2) steer the model in its direction and measure the prediction of a *different* axis, e.g. action; (3) repeat the steering in the opposite direction as a control.

Paired-statement probe extraction. For each axis $A \in \{\text{belief, intent, knowledge}\}$ and scenario s, we present the VLM with the keyframe clip and a pair of statements $\langle s_A^+, s_A^- \rangle$: s_A^+ matches the ground-truth axis label (e.g. for belief, "the protagonist saw the traveller crossing" when the protagonist did see it) and s_A^- is its negation. Let $\mathbf{z}^{\ell,h}(s) \in \mathbb{R}^{d_h}$ denote the per-(layer, head) activation at the final input token under each statement. Per (layer, head), an L^2-regularised logistic regression is fit on the $2N$ activations $\{\mathbf{z}^{\ell,h}(s_A^+), \mathbf{z}^{\ell,h}(s_A^-)\}_s$ with binary labels (1 for s_A^+, 0 for s_A^-), yielding a unit-norm probe direction $u_A^{\ell,h} \in \mathbb{R}^{d_h}$ that points from negated toward matched, plus a held-out accuracy. We use the top-K=56 (layer, head) pair ranked by held-out accuracy as the steering pool ($\sim$7% of $L \times H$ heads, following the head-selection convention of Li et al., 2023). $\sigma_A^{\ell,h}$ is the per-head activation standard deviation along $u_A^{\ell,h}$ on the training set. For False-Belief scenarios the matched-class caption is taken from the protagonist's view rather than the world state (see Appendix B.1).

Cross-Axis Routing Diagnostic (CARD). Given the steering-axis probe set $\{(u_A^{\ell,h}, \sigma_A^{\ell,h})\}$ over the top-K pool, we register a forward hook on the self-attention output of every layer that contains at least one selected (layer, head) pair. The hook adds, at the last-token position only and only to the per-head slices of the selected pairs, a scaled multiple of the probe direction:

$$\mathbf{z}_{\ell,h} \leftarrow \mathbf{z}_{\ell,h} + \alpha\, \sigma_A^{\ell,h}\, u_A^{\ell,h}, \quad \alpha \in \{-10,\, 0,\, +10\}, \tag{1}$$

where α is unit-free (the additive magnitude is $|\alpha|\,\sigma$ training-projection standard deviations) and remaining heads are left untouched. All K heads are steered simultaneously, so the intervention is multi-direction by construction; the rest of the forward pass runs unchanged. The VLM then answers a question on a *different* axis $B \neq A$, yielding a TB$-$FB accuracy gap $g_B(\alpha) = \mathrm{acc}_B^{\mathrm{TB}}(\alpha) - \mathrm{acc}_B^{\mathrm{FB}}(\alpha)$. Steering with α=+10 and α=$-$10 should move this gap in opposite directions if axis A's direction causally influences the axis-B prediction; we summarise the effect as the *gap-shift* – the change in g_B between the two steering polarities:

$$M_{A,B} = g_B(+10) - g_B(-10). \tag{2}$$

We additionally report a mean-accuracy shift $M_{A,B}^{\mathrm{acc}} = \max_{\alpha \in \{\pm 10\}} \mathrm{acc}_B(\alpha) - \mathrm{acc}_B(0)$ to catch uniform answer modulations a signed gap-shift would miss. On Relay Chain we populate the 3×3 matrix with $A, B \in \{\text{belief, intent, knowledge}\}$: same-axis entries ($A = B$) serve as positive controls; off-diagonal entries test whether the linearly-decodable A direction is used by the B prediction. A column $\{M_{A,B^*}\}_A \approx 0$ paired with at least one non-trivial row implies the B^* prediction is *decoupled from A's linearly-decodable direction* – the routing failure.

Four-tuple signature. On the unified success+counter partition, for each (VLM, protagonist, question axis) we record the four-tuple $\left(\mathrm{acc}_{\mathrm{succ}}^{\mathrm{TB}}, \mathrm{acc}_{\mathrm{succ}}^{\mathrm{FB}}, \mathrm{acc}_{\mathrm{cnt}}^{\mathrm{TB}}, \mathrm{acc}_{\mathrm{cnt}}^{\mathrm{FB}}\right)$. This signature discriminates the three behaviours (constant-prior, world-state tracker, belief-conditional), which are indistinguishable on success-only data.

Projection-removal control. As a negative-direction test, we subtract a scaled projection of the per-head residual along $u_A^{\ell,h}$:

$$\mathbf{z}_{\ell,h} \leftarrow \mathbf{z}_{\ell,h} - c\, \sigma_A^{\ell,h}\, (u_A^{\ell,h} \cdot \mathbf{z}_{\ell,h})\, u_A^{\ell,h}, \quad c \in \{0.5, 1, 2, 5\}. \tag{3}$$

c=1 is canonical projection-removal; c>1 over-subtracts.

Prompt-resistance test. We ask whether prompt engineering alone can recover belief-conditional action. Twelve variants (cost framing, chain-of-thought, perception-first / rule-based, in-context exemplars; Appendix B.6) are evaluated across 192 cells. A variant passes on a lever-holder (a_1 or a_2) if both $\mathrm{acc}^{\mathrm{TB}} > 50\%$ and $\mathrm{acc}^{\mathrm{FB}} > 50\%$ – correctly releases under TB and correctly holds under FB.

Evaluation protocol. Each axis question is presented as a binary choice and parsed from JSON. Probes, top-K selection, and CARD evaluation share the *unified success+counter* partition (444 scenarios), split 50/50 into a disjoint *probe-set* (222 scenarios; probe training and top-K ranking) and *eval-set* (222 scenarios; CARD, projection-removal, behavioural baselines), stratified by protagonist $\times$ world-state outcome (App. B.1). Accuracy is reported separately for TB vs FB and, where relevant, separated into success vs counter blocks.

4 Dataset and Task

Cooperative Relay Chain dataset. We evaluate CARD (section 3) on *Relay Chain* – a novel three-agent cooperative grid-world environment

(see Figure 2). CARD's diagnostic logic – steer one axis and read out a *different* prediction – requires a testbed in which a partner's mental state causally drives an action prediction that is distinct from the belief prediction that it steers. While GridToM (Li et al., 2025) pairs probes with activation interventions and supplies multi-agent scenarios and belief-attribution questions (first- and second-order beliefs), it does not offer a cooperative action axis: All questions only ask what an agent believes but not what the protagonist (the ego agent whose mental state we probe) should do given the partner's belief. The belief-to-action CARD cell is therefore undefined – not because a partner is absent, but because there is no action prediction that depends on the partner's mental state.

In contrast, on the Relay Chain, two agents (a_1, a_2) each hold a lever that opens one gate; a *traveller* (a_3) must cross both gates in sequence to reach the goal. The cooperative structure is a *relay*: lever-operators must keep their gates open in sequence as the traveller passes through. Crucially, as no single agent can solve this task alone, the traveller's optimal action is well-defined only relative to the partner's perception, turning belief-to-action into a well-defined test of routing.

Formal task description. We treat each scenario as a tuple $s = (w_0, \tau, a_p)$. w_0 is the initial grid layout (agents, levers, gates, goal). $\tau = (w_0, w_1, \ldots, w_T)$ with $T \in [16, 32]$ is the deterministic cooperative trajectory – the joint multi-agent state at each step. $a_p \in \{a_1, a_2, a_3\}$ is the protagonist – the ego agent whose mental state we probe. The VLM does not see τ directly: it receives a k-frame visual rendering ($k \in [4, 8]$) subsampled at relay-phase boundaries (a_1 acquires lever, a_2 acquires lever, a_3 reaches goal) plus midpoints (see Appendix A.2).

Belief variants. Each scenario appears in two layouts that hold τ and agent identities fixed and differ only in a_p's observation of the cooperative event (a_p's partner releases/crosses): True-Belief (s^{TB}, event within a_p's field of view) and False-Belief(s^{FB}, event in fog of war). Let $e \in \{0, 1\}$ indicate whether the event in fact *occurs* in the world. The *counter pair* replaces s with s' where $e = 0$ (the lever is never released, the gate never opens), preserving the same TB/FB perceptual structure.

Questions and labels. For each axis $A \in \{\text{belief}, \text{intent}, \text{knowledge}\}$ we ask one binary question $q_A(a_p)$ whose ground-truth label $y_A(s, b) \in \{0, 1\}$ is a deterministic function of (w_0, τ, e, b), where $b \in \{\mathrm{TB}, \mathrm{FB}\}$ selects the belief variant. *Belief* asks whether the protagonist *perceived* the cooperative event; *knowledge* asks whether they can *justify* that it happened (perception plus integration of frame evidence); *intent* asks the next cooperative action. The VLM M outputs a prediction $\hat{y}_{M,A}(s, b)$ that we score against y_A. Per-axis label logic is in Appendix A.6.

Full construction details (scenario generator, paired layouts, counter-scenarios, per-axis ground-truth labels) can be found in Appendix A.

Counter scenarios. On success-only data (the typical ToM benchmark setup, where the cooperative event always occurs), a model that emits the safe default on every input – "don't release" for lever-operators, "keep waiting" for the traveller – is correct on the majority of cells, so high accuracy is not evidence of belief-conditional reasoning. The counter pair ($e = 0$) breaks this shortcut: the safe default becomes wrong precisely where it was right under success, so constant-prior and belief-conditional models produce visibly different accuracy patterns. Construction mechanics and per-axis counts can be found in Appendix A.5, Table 3. Figure 2 illustrates the overall layout.

Three axes. Building on the belief–desire–intention framework (Bratman, 1987) and the percept–belief–intention (PBI) causal structure used in recent ToM benchmarks (Bortoletto et al., 2025), we instrument each scenario along three axes: *belief* – does the protagonist see the cooperative event (first-order PBI percept; what content the agent holds in mind?); *intent* – in the BDI sense of *present-directed action-recommendation*: should the protagonist take the cooperative action now (PBI intention); *knowledge* – does the protagonist *know whether* the cooperative event occurred separate from the propositional content of belief). Knowledge serves as a control: within counter scenarios (where the world did not change), it is the only axis whose label still flips between TB and FB – with observation alone, while belief and intent labels stay constant. Including knowledge in the probe-extraction pool forces the learned direction to track observation rather than the world-state outcome. Appendix D.1 verifies that the three probe directions are neither collinear nor orthogonal.

Models. Four open-weight VLMs: Qwen2-VL-7B (Yang et al., 2024a), Gemma-4-E4B-it (Gemma

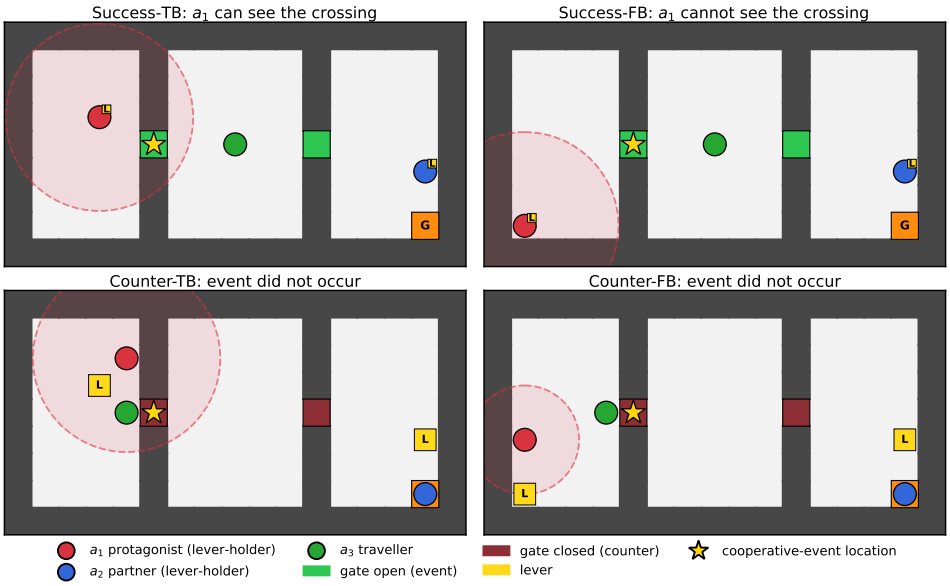

Figure 2: Relay Chain four-tuple discrimination layout. Panel elements: protagonist a_1 (red, lever-operator) with field-of-view disk (radius 5); partner a_2 (blue, lever-operator); traveller a_3 (green); gates (green when open, dark-red when closed); levers L; goal G. Top row: *success* (event occurs). Bottom row: *counter* (lever-operator never releases; gate stays closed). Left column: True-Belief (TB) – a_1's lever close enough to perceive the crossing. Right column: False-Belief (FB) – lever beyond a_1's view.

Team, Google DeepMind, 2026), LLaVA-NeXT-Video-7B-hf (Zhang et al., 2024), and InternVL2.5-8B (Chen et al., 2024b). We used multi-frame inputs (4–8 keyframes per scenario).

5 Results

We conducted a series of experiments that together show a *routing failure* on the cooperative-action axis. First, we verify belief is linearly decodable and used on the belief question (positive control). We then establish a behavioural dissociation: the same model that reads belief on the belief question collapses to a constant per-protagonist prior on the intent question. CARD steering along the linearly-decodable belief direction substantially shifts the belief and knowledge predictions but leaves the cooperative-action prediction essentially unchanged – the mechanistic evidence. Three ablations close the natural alternative explanations: projection-removal rules out amplitude limitation, a prompt-resistance test rules out a prompting artefact, and CLEVR rules out a globally-broken steering hook.

Belief is decodable and used on the belief question. We confirm belief is encoded and behaviourally used on the belief question – a binary read-out of whether the protagonist saw the cooperative event ("Does a_p see the partner release / cross?"). A linear probe on the TB-vs-FB activation difference reaches held-out accuracy well above chance on every VLM (see Table 1). We rule out spurious correlations with a Hewitt–Liang random-label control (Hewitt and Liang, 2019), which collapses probe accuracy to chance (see Table 11 in the Appendix). Behaviourally, every VLM answers the same question near ceiling on the TB layout and much lower on FB (see Table 1) – consistent with the prediction reading event occurrence but not gating by the protagonist's field of view. We thus rule out the possibility that belief is simply not represented: any later failure on the intent question cannot be blamed on missing or unreadable belief.

Model	Probe acc.	TB-acc.	FB-acc.
Qwen2-VL	81.5 ± 1.8	100.0	20.0
Gemma-4	94.5 ± 0.6	79.3	30.7
LLaVA-NeXT	78.1 ± 0.9	80.0	20.0
InternVL2.5	93.3 ± 0.5	89.3	2.0

Table 1: Positive control: linear-probe accuracy (mean $\pm$ std across three seeds; chance $= 50\%$) and behavioural TB/FB accuracy on the belief question.

Belief is used on belief and knowledge but ignored on intent questions. We find a three-way

behavioural dissociation on Relay Chain: the three axis questions, asked of the same model on the same scenarios with the same activations, produce qualitatively distinct accuracy patterns across all four VLMs (see the per-VLM matrix in Table 6 in the Appendix).

On the *belief question* every VLM reproduces the *world-state tracker* signature: across the four (success/counter) × (TB/FB) cells, the answer is correct on three (success-TB, counter-TB, counter-FB) and wrong only on success-FB, where the event occurred but outside the protagonist's field of view (per-VLM breakdown in Table 6 in the Appendix). We read this as the prediction using event occurrence but not gating by observation.

On the *knowledge question* the same models become behaviourally belief-conditional in the success block: TB is answered "knows" and FB is answered "doesn't know", correctly tracking the protagonist's observation rather than the world state alone. We conclude observation is at least partially consulted for this question.

On the *intent question* the same models collapse to a *constant per-protagonist prior* ("don't release" on every cell, incorrect only on success-TB where releasing is the right action). The constant-prior signature is indistinguishable across the four VLMs – a striking convergence given otherwise divergent probe accuracies and architectures, and the strongest sign that intent answers are generated without consulting either event occurrence or observation.

We thus observe three distinct signatures on the same model and the same activations: belief-Q reads world state, knowledge-Q reads observation, and intent-Q reads neither. The intent question alone is behaviourally belief-insensitive – the lonely outlier among the three axes. We next probe this dissociation mechanistically with CARD, asking whether the structural cause is upstream (belief unreachable from the intent prediction) rather than behavioural (belief read but ignored).

CARD intent column is uniformly null (routing failure). We find a clean three-way dissociation under CARD steering. Across the full $4{\times}3{\times}3 = 36$-cell sweep (four VLMs × three protagonists × three probe directions; see Table 7 in the Appendix), the best-α intent accuracy stays essentially at its no-steering baseline on nearly every cell, while the same steering hooks substantially shift the belief and knowledge predictions. Averaged across protagonists (see Figure 3), intent is flat on every VLM and the belief and knowledge axes move; the per-(VLM, protagonist) breakdown (see Figure 6 in the Appendix) shows the intent null holds across all 12 cells, not just the target lever-operator a_1. The gap-shift matrix on a_1 (see Figure 4) tells the same story at the gap-shift level.

Probe	Belief Q (%)	Intent Q (%)	Epist. Q (%)
Qwen2-VL			
belief	15.3±6.2	0.0	29.2±6.9
intent	2.1±2.8	0.0	0.0
knowledge	3.5±3.5	0.0	0.0
Gemma-4			
belief	33.3±7.6	4.9±3.5	29.9±7.6
intent	16.7±6.2	0.0	19.4±6.9
knowledge	0.7±1.4	0.0	6.2±4.9
LLaVA-NeXT			
belief	41.0±8.3	0.0	40.3±8.3
intent	22.2±6.9	0.0	67.4±7.6
knowledge	24.3±7.6	0.0	60.4±8.3
InternVL2.5			
belief	31.9±7.6	5.6±4.2	29.9±7.6
intent	1.4±2.1	0.0	3.5±3.5
knowledge	35.4±8.3	15.3±6.2	20.1±6.9

Table 2: Per-scenario flip rate on a_1: of 144 scenarios, the fraction whose answer differs between $\alpha{=}{+}10$ and $\alpha{=}{-}10$ steering (i.e., whether steering moves the model *at all*, before changes average out). In the intent column (middle), 9 of 12 (VLM, probe) entries are exactly 0; only three are above zero (Gemma-4 and InternVL2.5 on the belief probe, InternVL2.5 on the knowledge probe; $\leq$15.3%), with direction-balanced flips that cancel in net accuracy (Table 7 in the Appendix). The belief and knowledge columns flip many scenarios under the same hook.

We further show the null is per-scenario tight. On a_1 the intent argmax is invariant to $\alpha{=}{\pm}10$ steering on the overwhelming majority of scenarios across all four VLMs (see Table 2); the small fraction that do flip (InternVL2.5 and Gemma-4) are direction-balanced and cancel in net accuracy, while the belief and knowledge questions flip a substantial fraction of scenarios under the same hooks. We rule out the possibility that the intent answer was already pinned to one side before steering: without steering, the intent baseline distribution is mixed, so the null cannot be explained that way. Analysis confirms the null is statistically tight: across 108 intent measurements, the largest improvement in mean intent accuracy is bounded at 18.7 pp at 95% confidence, versus 22.9 pp on the

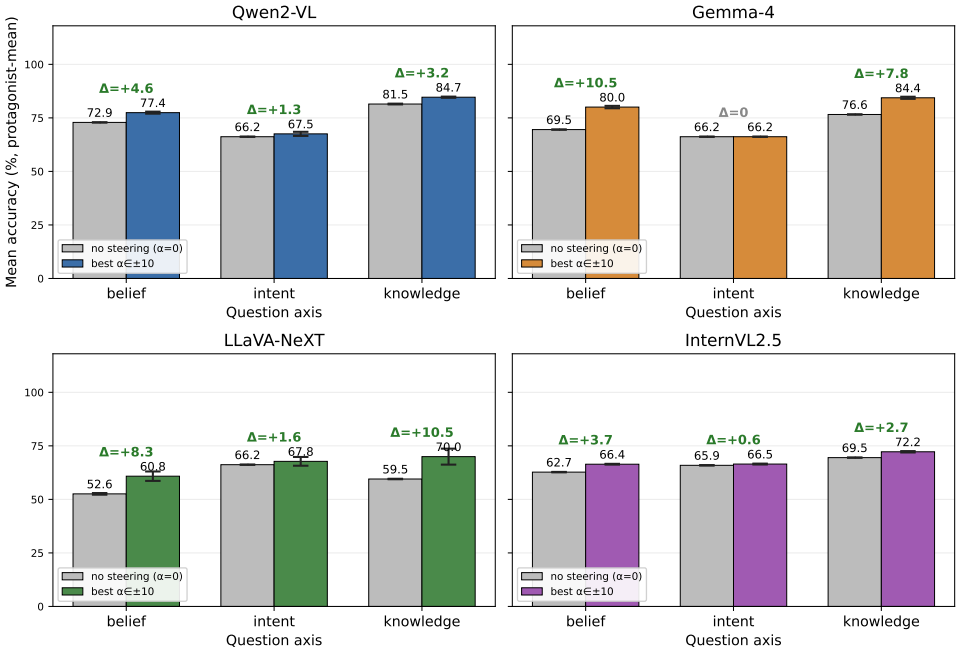

Figure 3: Steering improves accuracy on the belief and knowledge predictions, but never on the intent answer prediction, averaged across the three protagonists (a_1, a_2, a_3). For each (VLM, question axis) we report the mean TB/FB accuracy with no steering (α=0, grey) and with the best-case steering ($\alpha \in \{-10, +10\}$ across all probe directions, in VLM color). Δ on top is best−baseline. Belief and knowledge improve by $+2.7$ to $+10.6$ pp; the intent Δ is essentially 0 on every VLM. The identical intent baseline ($\sim 66\%$) across VLMs reflects the constant-prior signature. Per-protagonist breakdown in Figure 6 in the Appendix.

belief positive control (Appendix C.5).

We interpret this as a structural *routing failure*: the linearly-decodable belief direction is encoded, readable on the belief and knowledge predictions, but not used in the action prediction.

Ablation studies. We finally evaluated our findings against four alternative explanations: amplitude-limited steering (projection-removal control), a prompting artefact, cross-axis steering on a non-ToM task (CLEVR), and a last-token-locus artefact (distributed-token steering).

Projection-removal control

An alternative is that our additive steering simply does not push the belief subspace hard enough to register at the intent prediction, and a stronger or differently-shaped perturbation would. We test the opposite limit: instead of adding belief, we *remove* it from the residual stream entirely by subtracting its $c\,\sigma$-scaled projection from the per-head slice (Eq. 3) at $c \in \{0.5, 1, 2, 5\}$. The intent TB−FB gap stays essentially flat at every c on every VLM with a complete sweep, while the same removal shifts the *belief*-question gap toward chance as expected (positive control). The intent null survives both directions of perturbation, so it is structural, not a consequence of under-driving the belief subspace.

No prompt achieves belief-conditional action

Another possibility is that the intent prediction routes belief only when scaffolded by the right prompt. We test this with a pre-registered sweep of 12 prompt variants (CoT, perception-first, rule-based, high-stakes cost, 1/2-shot in-context) $\times$ four VLMs $\times$ four protagonist conditions (192 total; Figure 11 in the Appendix). No prompt recovers belief-conditional action.

Cross-axis Steering on Non-ToM task

Finally, the hook itself could be at fault – too weak to register anywhere. Applying the same hook to CLEVR on the same models (Table 10 in the Appendix) yields substantial best-α accuracy gains where there is headroom; the few flat cells reflect ceiling effects on shape and colour. The intent null is therefore a target-axis property, not a hook artefact.

Distributed-token steering

A further possibility is that the intent prediction reads belief at a different token position than our last-token hook. We re-run CARD with the steering hook fired at two additional positions (*vision-end*

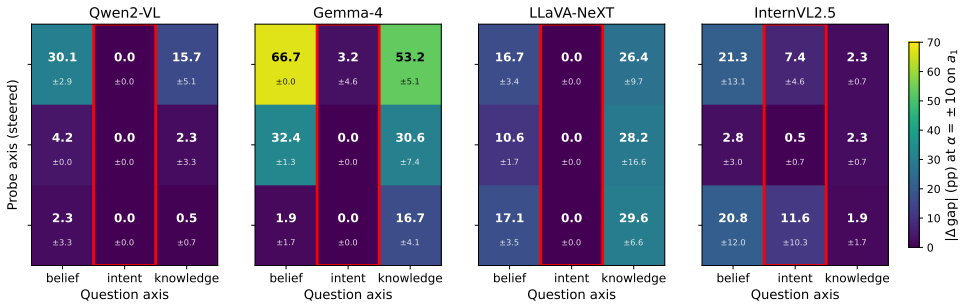

Figure 4: CARD intent column on lever-operator a_1 is uniformly null (red box) across four VLMs. Rows: probe axis being steered. Columns: question axis being asked. Cell value: $|\text{gap}(\alpha{=}+10) - \text{gap}(\alpha{=}-10)|$ in pp, mean $\pm$ std across three seeds. Non-intent columns shift by tens of pp under the same probe; the intent column does not move. Same steering hook, same scenarios, same protagonist – only the intent prediction fails to read from the belief direction.

and a *vision-plus-last* slice covering both) in addition to the *last-token* target, using three probe sources: the paired-statement belief and intent probes plus a per-position condition probe trained at *vision-end*. Sweeping $\alpha \in \{-10, +10\}$ yields 18 (probe-source $\times$ target-position) measurements. Across all, the maximum intent $|\Delta_{\text{acc}}|$ vs the $\alpha{=}0$ baseline is 0.3 pp; maximum shift is around 1 pp (see Figure 13 in the Appendix). The result is therefore not an artefact of the last-token hook position.

Summary. We established that belief is linearly decodable and used on the belief question, then a behavioural dissociation between belief- and intent-question signatures, then the causal dissociation under intervention, and finally closed the three alternative explanations (amplitude, prompt, hook). Belief direction is encoded, readable on the belief and knowledge predictions, and not used by the action prediction – a routing failure, not an absence.

6 Discussion

The simplest account is that the cooperative-action prediction does not use the linear belief direction. Additive steering and projection-removal both shift the belief and knowledge predictions but leave the intent prediction essentially unchanged on every VLM and every protagonist; the same prediction defaults to a constant per-protagonist prior that the visual evidence does not modulate. Two observations rule out the simpler alternative that the prediction is just broken or saturated. First, the prediction *does* vary across inputs — its prior differs by protagonist ("don't release" on a_1, a_2; "go" on a_3) — so it is not globally stuck on one answer. Second, the belief signal *is* present in the activations: the belief and knowledge predictions, exposed to the same residual stream, do use it. The prediction is therefore capable of varying and the information it would need is available; the failure to use belief is a routing fact specific to the intent prediction, not a general inability to respond.

7 Conclusion and Future work

In this work we introduced CARD as a new cross-axis routing diagnostic that disambiguates "feature absent" from "feature encoded but unused by the action prediction". Using this new diagnostic we showed in a cooperative task that current VLMs encode a partner's belief well enough to answer the belief question, yet none usefully use that belief when asked what to do. This represents a structural routing failure visible in current VLMs. One natural extension remains open: nonlinear probes would test whether the action prediction reads a subspace our linear probes miss. As such, our results suggest that verbal-ToM alignment does not imply action-ToM alignment: alignment training that updates only verbalised belief predictions, without an explicit objective on the action prediction, would leave this routing gap unaddressed.

Limitations

We demonstrate the routing failure on one task family – Relay Chain, a 2D grid-world – and four open-weight VLMs. CARD and projection-removal require activation-stream access, so we cannot subject closed-source VLMs to the mechanistic arms; We do not claim generality to larger model scales (40–400B, where applied-ToM behaviour is known to improve (Gu et al., 2024; Bortoletto et al., 2024)), non-cooperative ToM, or real-world video.

Ethical considerations

The work uses publicly available open-weight VLMs and a synthetic grid-world task with no human subjects or personally-identifying data. CARD reports an internal property of the model and surfaces a limitation that deployers should be aware of; we see no direct misuse pathway. Datasets, prompts, and code will be released under permissive licenses.

References

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

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

Matteo Bortoletto, Constantin Ruhdorfer, and Andreas Bulling. 2025. Tom-ssi: Evaluating theory of mind in situated social interactions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32252–32277.

Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. 2024. Brittle minds, fixable activations: Understanding belief representations in language models. *arXiv preprint arXiv:2406.17513*.

Michael E. Bratman. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and 1 others. 2024c. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586.

Gemma Team, Google DeepMind. 2026. Gemma 4: Our most capable open models to date. Technical report, Google DeepMind. Accessed: May 2026.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*.

Bear Häon, Kaylene Stocking, Ian Chuang, and Claire Tomlin. 2025. Mechanistic interpretability for steering vision-language-action models. *arXiv preprint arXiv:2509.00328*.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, volume 2024, pages 10504–10526.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pages 2733–2743.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Xinyang Li, Siqi Liu, Bochao Zou, Jiansheng Chen, and Huimin Ma. 2025. From black boxes to transparent minds: Evaluating and enhancing the theory of mind in multimodal large language models. *arXiv preprint arXiv:2506.14224*.

Yiming Liu, Yuhui Zhang, and Serena Yeung-Levy. 2025. Mechanistic interpretability meets vision language models: Insights and limitations. In *The Fourth Blogpost Track at ICLR 2025*.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, and 1 others. 2025. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 ieee conference on secure and trustworthy machine learning (satml)*, pages 464–483. IEEE.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2025. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1510–1519.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. 2024. Activation addition: Steering language models without optimization.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023a. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. 2023b. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in neural information processing systems*, 36:78205–78226.

Yang Xiao, Jessie Wang, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. Tomvalley: Evaluating the theory of mind reasoning of llms in realistic social context.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 39 others. 2024a. Qwen2 technical report. *ArXiv*, abs/2407.10671.

Jingcheng Yang, Tianhu Xiong, Shengyi Qian, Klara Nahrstedt, and Mingyuan Wu. 2026. Circuit tracing in vision-language models: Understanding the internal mechanisms of multimodal thinking. *arXiv preprint arXiv:2602.20330*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024b. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Ruoxuan Zhang, Qiyun Zheng, Zhiyu Zhou, Ziqi Liao, Siyu Wu, Jian-Yu Jiang-Lin, Bin Wen, Hongxia Xie, Jianlong Fu, and Wen-Huang Cheng. 2025. Mindpower: Enabling theory-of-mind reasoning in vlm-based embodied agents. *arXiv preprint arXiv:2511.23055*.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: A strong zero-shot video understanding model.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Dataset construction

A.1 Dataset breakdown

Relay Chain has 444 unique scenario bases distributed across three protagonist conditions (a_1, a_2 as lever-holders and a_3 as traveller). We render each base in two layouts that preserve the trajectory but differ in the protagonist's lever-to-gate Manhattan distance – True-Belief (≤ 5, lever within a_p's field of view of the gate) and False-Belief (> 5, lever beyond) – yielding 888 evaluation cases per axis (Table 3). Scenarios split into *success* blocks where the cooperative event happens and *counter* blocks where it is omitted, so the same protagonist appears once with and once without the cooperative outcome on identical perceptual access. The full 444-scenario partition is split $50/50$ into a disjoint probe-set and eval-set (App. B.1).

A.2 Relay Chain scenario generator

We render scenarios as 11×11 grids with three agents (a_1, a_2, a_3), two gates, and one goal cell. Each scenario is a 16–32-step trajectory; we subsample 4–8 keyframes covering the cooperative event window. Keyframes anchor on four relay-phase boundaries: (i) episode start; (ii) a_1 acquires lever A; (iii) a_2 acquires lever B; (iv) a_3 reaches the goal. We then insert midpoints between consecutive boundaries to yield up to 7 distinct frames per scenario, following GridToM's keyframe helper. When a phase is never reached, the last available step is used as a fallback so the boundary list stays strictly monotonic. Agents move on Manhattan-distance shortest paths; lever-state transitions are scripted to produce the target cooperative chain. Random seeds are saved with each scenario for reproducibility.

A.3 Perceptual-access minimal pair

For each scenario we generate two layouts that share trajectory and event sequence but differ in protagonist's lever position relative to the gate it controls. TB: lever placed at Manhattan distance ≤ 5 from the gate (within field of view). FB: lever at distance > 5. All other state (agent identities, gate positions, goal, traveller path) is identical. We verify pair matching by hashing the trajectory and confirming bit-exact agreement on the non-protagonist state slice.

A.4 Example scenario layouts

Figure 5 shows one True-Belief and one False-Belief layout for each protagonist condition. For the lever-holders (a_1, a_2), TB places the protagonist's own lever within Manhattan distance 5 of the gate it controls (inside the field-of-view disk); FB places it beyond. For the traveller (a_3), TB places the agent within view of the gate ahead so the gate-state is perceivable; FB places it beyond. All panels show the success block (both gates open), so the contrast across columns is purely perceptual.

A.5 Counter-story generator

We construct the counter pair for each scenario by re-running the scripted trajectory but *omitting* the cooperative event: the lime traveller stalls 3 cells before the gate (it never crosses), or the pink lever-holder fails to release the lever at the cooperative deadline (the gate never opens for the cooperative crossing). We preserve the protagonist's perceptual access (TB/FB layout), rewrite captions with the asymmetric oracle-vs-pov rule (§3), and recompute the binary correct answer by enumeration.

A.6 Axis-specific ground-truth labels

We compute ground-truth labels deterministically from the layout and event-outcome pair. Belief: "does the protagonist see the event". Intent: "should the protagonist take the cooperative action now", mapped to release/keep-holding for lever-holders and continue/wait for the traveller. Knowledge: "does the protagonist have enough perceptual access to determine whether the cooperative event occurred" – for lever-holders this is "can the protagonist see the traveller", for the traveller it is "can the protagonist see at least one lever-holder". The knowledge label depends only on perceptual visibility, not on whether the event in fact occurred; success and counter blocks therefore share the same knowledge label on each layout (TB $\rightarrow$ TRUE, FB $\rightarrow$ FALSE), so a knowledge-conditional answer is forced to track perceptual access rather than the world-state outcome.

B Probe + behavioral protocol details

B.1 Probe extraction

For each axis we extract per-(layer, head) activations at the final token under each paired statement. The difference $\Delta_A = a(s_A^+) - a(s_A^-)$ is the training input for the linear probe; labels are ± 1 for

Protagonist	Role	Success	Counter	Total
a_1	lever-operator	100	44	144
a_2	lever-operator	100	50	150
a_3	traveller	100	50	150
Total		**300**	**144**	**444**

Table 3: Relay Chain dataset breakdown by protagonist and world-state outcome. All three axes (belief, intent, knowledge) share the same scenario base; each entry is a multi-frame trajectory with 4–8 keyframes.

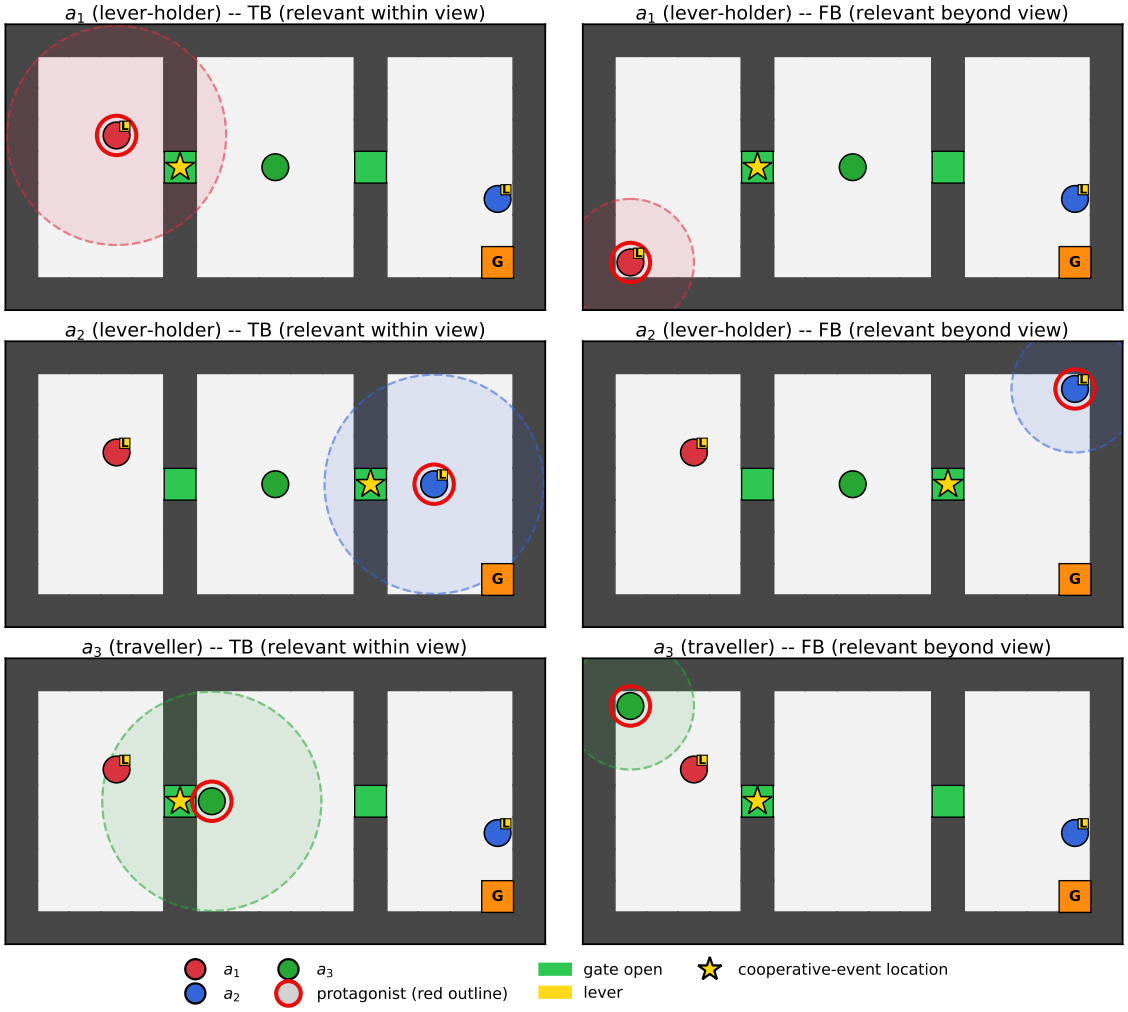

Figure 5: Relay Chain scenario layouts for each protagonist condition (rows: a_1 and a_2 lever-holders, a_3 traveller) under True-Belief and False-Belief layouts (columns). The protagonist is highlighted with a red outline; the shaded disk shows their field of view (radius 5 cells; anything outside is fog of war and not perceptible). In TB the protagonist stands next to its gate-controlling lever or just before the gate ahead; in FB it stands at the far lever or in a corner, so the field-of-view disk shifts with it and the cooperative event falls outside. All panels show the success block (both gates open) so the contrast is purely perceptual.

matched/mismatched statements. We use logistic regression with default L^2 regularisation (C=1).

K-selection / CARD-eval protocol. To prevent K-selection bias from probe training, top-K ranking, and CARD evaluation sharing scenarios, we split the unified success+counter partition (444 scenarios) into a fully disjoint *probe-set* and *eval-set* stratified 50/50 by protagonist × world-state outcome (222 scenarios each). The probe-set is used for probe training (single 75/25 stratified train/val split per (layer, head), seeded for reproducibility) and top-K=56 cell ranking by held-out val accuracy; CARD is evaluated only on the disjoint eval-

set. To quantify probe-training variance we re-run with three seeds $\{42, 43, 44\}$ for the internal $75/25$ split and report the mean across seeds.

B.2 CARD steering hook

We implement the steering hook as a forward hook on each chosen layer's self-attention output: at the last-token position only, $\mathbf{z}_{\ell,h} \leftarrow \mathbf{z}_{\ell,h} + \alpha\, \sigma_A^{\ell,h}\, u_A^{\ell,h}$ (Eq. 1), with $u_A^{\ell,h}$ the unit-norm probe direction at that (layer, head), $\sigma_A^{\ell,h}$ the training-set standard deviation of the scalar projection $u_A^{\ell,h} \cdot \mathbf{z}_{\ell,h}(s_A^+)$, and $\alpha \in \{-10, 0, +10\}$. We steer the top-56 per-(layer, head) cells ranked by held-out probe accuracy (cells selected by $\mathrm{argsort}(\text{val\_acc})$ descending), the same selection used by the canonical probe-extraction sweep. We do *not* normalise by sequence position or apply position-specific scaling.

B.3 Projection-removal

At each (layer, head) we subtract $c\,\sigma_A^{\ell,h}\,(u_A^{\ell,h} \cdot \mathbf{z}_{\ell,h})\,u_A^{\ell,h}$ from $\mathbf{z}_{\ell,h}$ at the last-token position (Eq. 3), with $c \in \{0.5, 1, 2, 5\}$. $\sigma_A^{\ell,h}$ is the per-(layer, head) standard deviation of $u_A^{\ell,h} \cdot \mathbf{z}_{\ell,h}(s_A^+)$ across the unified-partition training set. Under projection-removal LLaVA-NeXT occasionally appends commentary tokens after the JSON answer; we extract the answer with a regex on the *yes/no* field in addition to the canonical JSON-parse path. Recomputed accuracies match the raw answer-token distribution and are reported in §5.

B.4 Temporal probe accuracy (Qwen2-VL)

We report linear-probe held-out accuracy at four trajectory fractions $t \in \{0.25, 0.5, 0.75, 1.0\}$ on Qwen2-VL, for the three cooperative-ToM axes (Table 4). Peak (best layer/head) accuracy is saturated at 100% from the earliest timestep; mean accuracy varies by axis but is stable to within ~ 1 pp. The representation is formed early. We restrict the temporal sweep to Qwen2-VL on compute-budget grounds (each timestep is a full per-(layer, head) probe-extraction pass over 444 scenarios); we treat it as a single-VLM ablation, not as a generality claim.

B.5 Pre-registration record

We pre-registered the T-A falsifier before running the V0–V11 sweep: a cell passes only if both TB and FB accuracy exceed 50% on a lever-holder. The hypothesis space (12 variants $\times$ 4 VLMs $\times$ 4 protagonist conditions) was fixed before any data was collected.

B.6 T-A prompt variants

Table 5 summarises the 12 prompt variants used in the T-A pre-registered sweep, grouped into four families (cost framing, chain-of-thought, perception-first / rule-based, in-context exemplars). Each variant prepends a prefix to the canonical intent question; the verbatim prefix text for every variant is in §B.7 below.

B.7 T-A prompt prefixes (verbatim)

We reproduce here the exact prefix appended before the canonical intent question for each T-A variant. Variants whose prefix differs by protagonist role list both lever-holder and traveller text.

V0 (baseline). No prefix. The canonical intent question is asked unchanged.

V1 (cost: implicit consequences).

> ***Lever-holder:*** *Reminder of cooperative consequences: if the lever-holder releases while the traveller has not yet crossed the gate, the gate closes on the traveller and the team fails the task. If the lever-holder keeps holding after the traveller has crossed, the team is delayed but does not fail.*
>
> ***Traveller:*** *Reminder of cooperative consequences: if the traveller continues toward a closed gate, it gets stuck and the team fails. If the traveller waits when the gate is open, the team is delayed but does not fail.*

V2 (cost: explicit numerical).

> ***Lever-holder:*** *Cost framing: action 'Yes' (release/no-longer-hold) → cost* 2.0 *if the traveller has not yet crossed, else cost* 0*. Action 'No' (keep holding) → cost* 1.0 *if the traveller has already crossed, else cost* 0*. Choose the action that minimises expected cost given what the protagonist can observe.*
>
> ***Traveller:*** *Cost framing: action 'Yes' (continue toward gate) → cost* 2.0 *if the gate is currently closed, else cost* 0*. Action 'No' (stop and wait) → cost* 1.0 *always (delay but no failure). Choose the action that minimises expected cost given what the protagonist can observe.*

V3 (utility framing).

> *Question framing: what action would MAXIMISE the team's chance of completing the cooperative task successfully, given what the protagonist can currently observe?*

V4 (chain-of-thought).

> *Think step by step: (1) FIRST identify what the protagonist can perceive given the protagonist's position and fog-of-war range. (2) THEN consider what the optimal action is conditional on*

Axis	t=0.25	t=0.5	t=0.75	t=1.0	peak
belief	69.9	69.1	69.3	69.3	100.0
intent	95.6	95.1	94.7	94.3	100.0
knowledge	91.9	91.2	90.8	90.4	100.0

Table 4: Linear-probe held-out accuracy (%) as a function of trajectory fraction t on Qwen2-VL, for the three cooperative-ToM axes (belief, intent, knowledge). "Peak" = best (layer, head) probe; the remaining columns are means across all (layer, head) cells. The peak is saturated at 100% from t=0.25 onward. Mean accuracy decreases very slightly (~ 1 pp) with t: while the *best* head remains perfectly recoverable, fewer heads on average carry a clean axis-specific signal as later tokens accumulate downstream context. The relevant headline is the saturated peak: the feature is present from early in the trajectory.

Variant	Family	Description
V0	cost	baseline; canonical intent question, no prefix
V1	cost	implicit consequences: "if released too early, gate closes on traveller; team fails."
V2	cost	explicit numerical cost: release=cost 2.0 if not crossed, hold=cost 1.0 if crossed
V3	cost	utility framing: "what action MAXIMISES team success?"
V4	CoT	step-by-step: perception $\rightarrow$ belief $\rightarrow$ cost $\rightarrow$ action
V5	CoT	high-stakes framing: "a wrong release may kill the team"
V6	CoT	V4 chain-of-thought + V2 numerical costs
V7	perception-rule	"State what the protagonist can see, then answer."
V8	perception-rule	explicit rule: "if you can see the crossing, release; else hold."
V9	perception-rule	V7 perception-first + V2 cost numbers
V10	in-context	one-shot worked example (Manhattan distance $\leq 5 \rightarrow$ release; $> 5 \rightarrow$ hold)
V11	in-context	two-shot: V10 plus a second example of opposite polarity

Table 5: The 12 T-A prompt variants used to evaluate prompt-resistance. Each variant prepends a prefix to the canonical intent question; in-context variants additionally include worked examples.

what the protagonist actually knows. (3) FINALLY produce a single Yes or No answer. The protagonist may be partly ignorant about events outside its fog. Only choose the answer the protagonist would optimally pick given its actual perceptual state.

V5 (high-stakes safety framing).

Safety-critical reminder: this is a cooperative team task in which a wrong action by the lever-holder may TRAP the traveller permanently (team failure), and a wrong action by the traveller may STRAND the agent in a closed-gate room (team failure). Only choose actions that minimise the risk of catastrophic failure given what the protagonist can actually observe.

V6. V4 chain-of-thought prefix concatenated with the V2 explicit-cost prefix (lever-holder or traveller, by role).

V7 (perception-first).

BEFORE answering, identify what the protagonist agent can actually perceive. The protagonist has a fog-of-war view of radius 5 cells from its current position; cells beyond that radius are NOT visible to the protagonist. The lever-holder's view of the gate they control depends on the lever's distance from the gate. If the lever sits next to the gate (Manhattan distance ≤ 5), the lever-holder CAN see who crosses. If the lever sits far from the gate (Manhattan distance > 5), the lever-holder CANNOT see who crosses. Use only what the protagonist actually observes to choose the answer.

V8 (rule-based).

Rule: a lever-holder should release its lever ONLY IF it can directly perceive that the agent it is enabling has finished crossing the gate. Otherwise it must keep holding. A traveller should continue toward a gate ONLY IF it can perceive that the gate is open. Apply this rule given what the protagonist actually observes (fog-of-war radius 5).

V9. V7 perception-first prefix concatenated with the V2 explicit-cost prefix (by role).

V10 (one-shot in-context).

EXAMPLE: a lever-holder agent's lever sits 2 cells away from a gate (within fog-of-war radius 5). The lever-holder CAN see the gate. The lever-holder observes the traveller crossing the gate. Question: should the lever-holder release the lever now? Answer: Yes, because the lever-holder has perceived the traveller crossing and the gate is no longer needed.

Now answer the following with the same belief-conditional reasoning. If the protagonist CAN perceive the relevant event, recommend the action that responds to that perception. If the protagonist CANNOT perceive the event, recommend the conservative action.

V11 (two-shot in-context).

EXAMPLE 1 (TrueBelief case): a lever-holder agent's lever sits 2 cells away from a gate (within fog-of-war radius 5). The lever-holder CAN see

the gate and observes the traveller crossing. Question: should the lever-holder release? Answer: Yes, because the protagonist perceived the crossing.

EXAMPLE 2 (FalseBelief case): a lever-holder agent's lever sits 10 cells away from a gate (beyond fog-of-war radius 5). The lever-holder CANNOT see the gate and does not know whether the traveller has crossed. Question: should the lever-holder release? Answer: No, because the protagonist has NOT perceived the crossing and releasing would risk trapping the traveller.

Now answer the following with the same belief-conditional reasoning. Distinguish: does the protagonist actually perceive the event in question, or not? Choose the action that matches the protagonist's actual perceptual access.

C Full results tables

C.1 Per-VLM, per-protagonist, per-axis behaviour matrix

We report the full per-cell behavioural accuracies that underlie the three-way dissociation in F2. The matrix covers 4 VLMs × 3 axes × 3 protag × 2 outcomes × 2 beliefs = 144 cells (Table 6). Readers should look for the *shape* contrast: belief-Q cells track the world-state outcome, knowledge-Q cells track TB/FB (perceptual access), intent-Q cells stay on the per-protagonist prior.

C.2 Intent Δ_{acc} per (VLM, protag, probe)

For every (VLM × protagonist × probe direction) cell in the $4 \times 3 \times 3 = 36$-cell CARD sweep on the intent question, we report best-α accuracy gain over the no-steering baseline (Table 7). This is the load-bearing evidence for the generalised routing-failure claim in §5.

C.3 CARD per-cell breakdown (Δ_{gap} and Δ_{acc})

For lever-holder a_1, we report both the gap shift and the mean-accuracy shift for every (probe, question, VLM) cell. Across the full 4 VLMs × 3 probes = 12 intent-question cells, 9 are exactly **0.00** on both metrics. The three nonzero cells: Gemma-4 under belief-direction steering ($|\Delta_{\text{gap}}|$=9.7 pp, $|\Delta_{\text{acc}}|$=4.9 pp); InternVL2.5 under belief-direction steering ($|\Delta_{\text{gap}}|$=8.3 pp, $|\Delta_{\text{acc}}|$=5.6 pp); and InternVL2.5 under knowledge-direction steering ($|\Delta_{\text{gap}}|$=25.0 pp, $|\Delta_{\text{acc}}|$=13.9 pp, direction-balanced flips that cancel in net accuracy). This rules out both belief-conditional modulation and uniform answer-flip across the matrix.

C.4 Per-protagonist CARD breakdown

Figure 6 breaks down the protagonist-mean of Figure 3 into a 4×3 grid over (VLM, protagonist). The intent column Δ is essentially 0 in every cell, confirming that the routing failure is not protagonist-specific.

C.5 CIs and Wilcoxon sign test on the intent null

Across all 108 intent measurements (4 VLMs × 3 protagonists × 3 probes × 3 seeds), steering improves intent accuracy by at most 18.7 pp (95% paired-bootstrap upper bound), vs 22.9 pp on the belief control. The intent answer remains invariant to α in $69/108$ of these; the remainder skews toward the constant-prior default (27 vs 12, Wilcoxon $p = 0.009$) – steering pushes intent toward its default, not toward belief-conditional behaviour.

C.6 CARD 3×3 heatmaps per VLM

For each VLM we show one panel per protagonist; each panel is a 3×3 grid in which rows are the direction we steered along (belief / intent / knowledge) and columns are the question we then asked. The cell number is how much the TB−FB gap *moves* between α=+10 and α=−10, in pp.

A large cell value means the gap moved a lot – not that the model got more accurate: a TB-correct / FB-wrong pattern can flip to TB-wrong / FB-correct under steering, giving a large gap shift with mean accuracy unchanged. The intent column (red box) is the case we care about: even where it lights up, the companion accuracy-improvement matrix (Table 7) is essentially zero on every cell. The largest values sit on the diagonal (steer-belief → ask-belief, etc.) – these are the expected positive controls, where steering an axis strongly moves its own answer head.

Figures 7–10 show all four VLMs.

C.7 T-A 192-cell heatmap

For each (variant V0–V11, VLM, protag ∈ {overall, a1, a2, a3}) we report TB accuracy, FB accuracy, and their gap (Figure 11). Readers should look for cells in which a lever-holder column exceeds the pre-registered {TB > 50, FB > 50} threshold – no cell does.

C.8 Projection-removal per-VLM sweep

For each VLM and each $\alpha \in \{0.5, 1, 2, 5\}\sigma$ we report the intent TB−FB gap on the unified partition (Figure 12). The flat curves – nearly horizontal

VLM	Axis	Lever a_1	Lever a_2	Trav. a_3
Qwen2-VL	belief	100/0/100/100	100/50/100/100	100/0/100/100
	intent	0/100/100/100	0/100/100/100	53/47/100/100
	know.	100/100/48/100	100/69/72/100	53/100/44/98
Gemma-4	belief	100/0/100/100	69/52/100/100	100/4/100/100
	intent	0/100/100/100	0/100/100/100	100/0/100/100
	know.	100/100/59/100	100/100/58/100	0/100/44/100
LLaVA-NeXT	belief	100/0/57/52	100/0/10/42	53/47/100/100
	intent	0/100/100/100	0/100/100/100	53/47/100/100
	know.	89/51/48/52	100/52/100/42	53/47/44/56
InternVL2.5	belief	84/0/95/100	75/4/100/100	100/0/94/98
	intent	0/100/100/100	0/100/100/100	75/23/98/100
	know.	99/89/0/93	98/87/0/92	2/100/0/100

Table 6: Full D2 four-tuple discrimination matrix (success-TB / success-FB / counter-TB / counter-FB) for every (VLM, axis, protagonist).

VLM	Protag.	Probe=belief	Probe=intent	Probe=knowledge
Qwen2-VL	a_1	$+0.0$	$+0.0$	$+0.0$
	a_2	$+0.0$	$+0.0$	$+0.0$
	a_3	$+4.0 \pm 2.9$	$+0.0$	$+0.0$
Gemma-4	a_1	$+0.0$	$+0.0$	$+0.0$
	a_2	$+0.0$	$+0.0$	$+0.0$
	a_3	$+0.0$	$+0.0$	$+0.0$
LLaVA-NeXT	a_1	$+0.0$	$+0.0$	$+0.0$
	a_2	$+0.0$	$+4.0 \pm 5.7$	$+0.0$
	a_3	$+0.0$	$+0.7 \pm 0.5$	$+0.0$
InternVL2.5	a_1	$+0.0$	$+0.0$	$+0.0$
	a_2	$+0.0$	$+0.0$	$+0.0$
	a_3	$+1.3 \pm 0.5$	$+1.3 \pm 1.1$	$+1.1 \pm 0.8$

Table 7: Best-α intent accuracy gain (pp) over the no-steering baseline (α=0) for every (VLM, protagonist, probe direction) cell, reported as mean $\pm$ std across three seeds (seeds $\{42, 43, 44\}$; the $\pm$ is dropped when the std rounds to 0.0). Most cells are exactly 0 on every seed; across the full multi-seed sweep, $95/108$ cells are exactly 0 and the largest gain anywhere is $+12.00\,\%$. Compare with belief and knowledge accuracies, which improve by $+2.7$ to $+10.6$ pp on the across-protagonist mean under the same hooks (Figure 3): the intent prediction does not use the belief direction usefully on any cell.

across α on every VLM – are the visual form of the A1 amplitude-invariance argument.

C.9 Vendor architecture and rigidity ordering

The four VLMs differ in text-decoder architecture and in their empirical resistance to prompt perturbation (Table 9); the more rigid decoders are the ones whose intent answers were hardest to move under any T-A prompt in our sweep.

D Additional controls and diagnostics

D.1 Off-diagonal cosine measurements

Probe directions on the top-56 (layer, head) cells are moderately aligned, not orthogonal. On the three steered axes (belief, intent, knowledge), pairwise cosines fall in $0.26 - -0.58$ across the four VLMs. The belief–intent cosine is 0.55 (Qwen2-VL), 0.38 (Gemma-4), 0.57 (LLaVA-NeXT), and 0.28 (InternVL2.5).

This non-orthogonality is what makes the intent column's null informative. Steering along belief is partially also steering along intent, yet the intent prediction does not move. If the two directions were nearly orthogonal, a null cross-axis effect on intent would be trivial; at the observed alignment, the null reflects that the intent prediction does not consume the shared subspace. The non-null sibling columns we observe in CARD (§5) confirm that the linear belief direction *is* causally usable by the belief and knowledge predictions, ruling out an "information-not-present" interpretation.

D.2 CLEVR positive-control breakdown

Table 10 reports the full per-(VLM, question, probe) accuracy-gain matrix for the CLEVR positive control (§5). The sweep covers 100 scenarios per axis and three probe axes (count, shape, colour) on five VLMs: the four Relay Chain

VLM	Probe	Question	Δ_{gap} (pp)	Δ_{acc} (pp)
Qwen2-VL	belief	belief	$+30.56$	-15.28
	belief	**intent**	$+0.00$	$+0.00$
	belief	knowledge	$+22.22$	$+11.11$
	intent	belief	-4.17	$+2.08$
	intent	**intent**	$+0.00$	$+0.00$
	intent	knowledge	$+0.00$	$+0.00$
	knowledge	belief	$+6.94$	-3.47
	knowledge	**intent**	$+0.00$	$+0.00$
	knowledge	knowledge	$+0.00$	$+0.00$
Gemma-4	belief	belief	$+66.67$	$+0.00$
	belief	**intent**	-9.72	$+4.86$
	belief	knowledge	$+59.72$	-20.14
	intent	belief	$+33.33$	-16.67
	intent	**intent**	$+0.00$	$+0.00$
	intent	knowledge	$+22.22$	-19.44
	knowledge	belief	$+1.39$	-0.69
	knowledge	**intent**	$+0.00$	$+0.00$
	knowledge	knowledge	$+12.50$	$+6.25$
LLaVA-NeXT	belief	belief	$+20.83$	-4.86
	belief	**intent**	$+0.00$	$+0.00$
	belief	knowledge	-13.89	$+40.28$
	intent	belief	$+8.33$	-13.89
	intent	**intent**	$+0.00$	$+0.00$
	intent	knowledge	$+9.72$	-14.58
	knowledge	belief	-12.50	$+11.81$
	knowledge	**intent**	$+0.00$	$+0.00$
	knowledge	knowledge	-23.61	$+20.14$
InternVL2.5	belief	belief	-31.94	-20.14
	belief	**intent**	$+8.33$	-5.56
	belief	knowledge	$+2.78$	-26.39
	intent	belief	$+0.00$	$+0.00$
	intent	**intent**	$+0.00$	$+0.00$
	intent	knowledge	-2.78	$+2.78$
	knowledge	belief	$+31.94$	$+20.14$
	knowledge	**intent**	-25.00	$+13.89$
	knowledge	knowledge	-4.17	$+17.36$

Table 8: CARD per-cell breakdown on lever-holder a_1, reporting both the gap shift $\Delta_{\mathrm{gap}} = \mathrm{gap}_B(\alpha{=}{+}10) - \mathrm{gap}_B(\alpha{=}{-}10)$ and the mean-accuracy shift $\Delta_{\mathrm{acc}} = \overline{\mathrm{acc}}_B(\alpha{=}{+}10) - \overline{\mathrm{acc}}_B(\alpha{=}{-}10)$, where $\overline{\mathrm{acc}}_B = \frac{1}{2}(\mathrm{TB}_B + \mathrm{FB}_B)$. Bold rows are the intent-question column: 9 of the 12 cells are $\mathbf{0.00}$ on *both* metrics across the four VLMs (Qwen2-VL and LLaVA-NeXT all three probes; Gemma-4 and InternVL2.5 under intent and knowledge probes), and the remaining three (Gemma-4 and InternVL2.5 under belief steering, InternVL2.5 under knowledge steering) shift by at most $|\Delta_{\mathrm{gap}}|{=}25.0$ pp / $|\Delta_{\mathrm{acc}}|{=}13.9$ pp, ruling out both belief-conditional modulation (Δ_{gap}) and any uniform answer-flip (Δ_{acc}). Sibling columns (belief, knowledge) show non-trivial shifts under the same probe directions.

VLM	layers	heads	head_dim	hidden	rigidity
Qwen2-VL-7B	28	28	128	3584	hardest
LLaVA-NeXT-7B	32	32	128	4096	middle
Gemma-4-E4B	34	8	320	2560	softest
InternVL2.5-8B	32	32	128	4096	intermediate

Table 9: Text-decoder architecture specs and empirical rigidity ordering of the four VLMs. *Rigidity* = resistance of the intent constant prior to prompt perturbation in the T-A sweep: Qwen2-VL is *hardest* (no V0–V11 cell breaks $30/100$ on a_1); LLaVA-NeXT is *intermediate*; Gemma-4 is *softest* (V2 cost-explicit breaks the prior on a_2 to $70/86$, though not toward belief-conditional behaviour). The ordering matches loosely with head dimensionality (Gemma's 320 vs. 128 for the others), but the dataset (4 vendors) is too small to draw architectural conclusions; a quantitative regression of rigidity against pretraining mix and RLHF intensity is left to future work.

backbones plus SmolVLM-500M (Idefics-3 family). We added SmolVLM-500M explicitly because Qwen2-VL and InternVL2.5 sit at ceiling on shape and colour, and we wanted a backbone with abun-

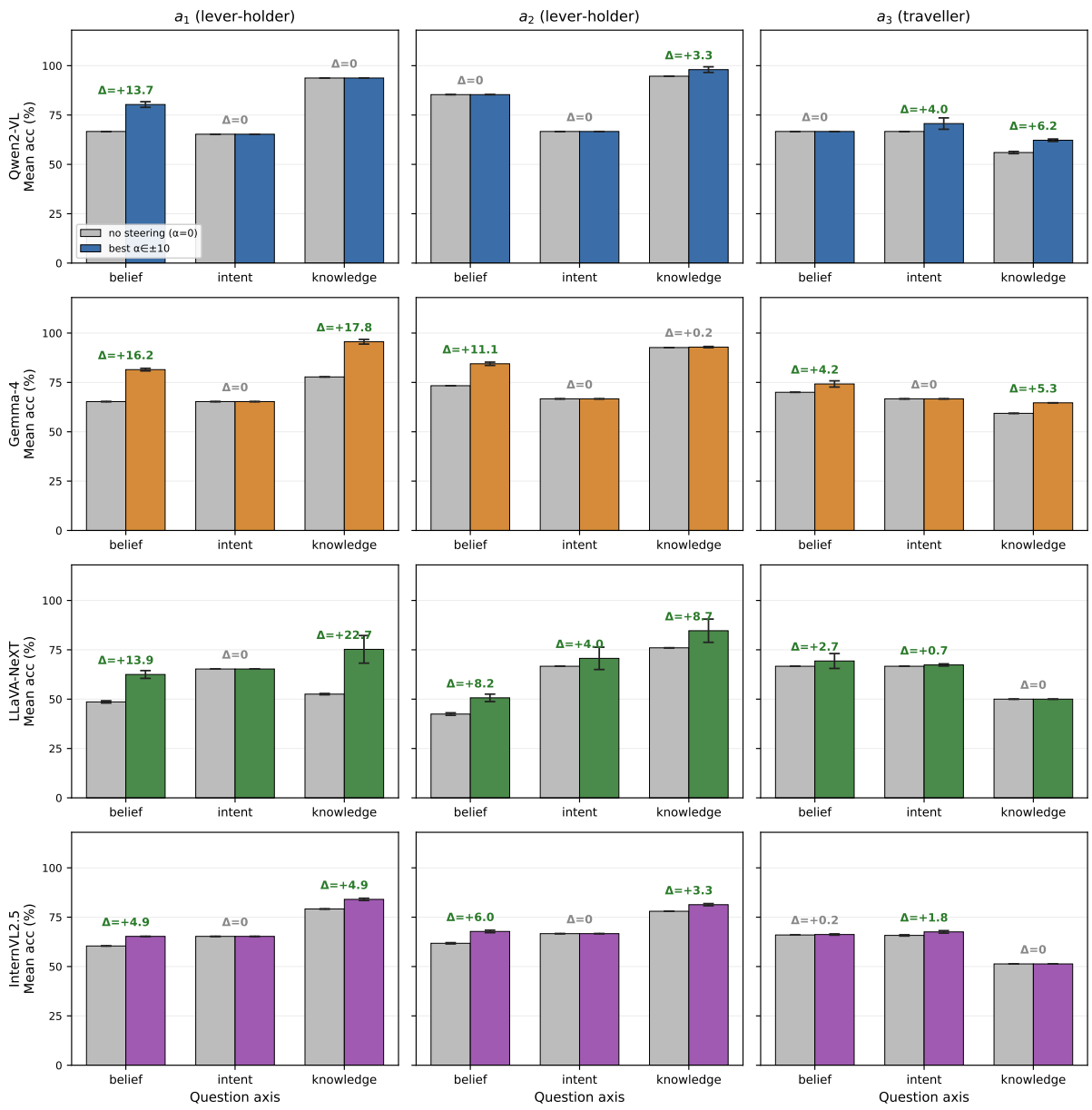

Figure 6: Per-protagonist breakdown of Figure 3. 4×3 grid: rows are VLMs, columns are protagonists (a_1, a_2 lever-holders; a_3 traveller). Each cell shows mean TB/FB accuracy at α=0 (grey) vs the best $\alpha \in \{-10, +10\}$ across probe directions (VLM color); Δ on top is best−baseline. The intent column Δ is essentially 0 in every (VLM, protagonist) cell (12/12), while sibling answer heads (belief, knowledge) move on the protagonists that have headroom. The routing failure is therefore not protagonist-specific.

dant headroom on every axis. CLEVR's True/False variants share the same correct answer by construction, so the gap-shift metric is uninformative here ($\Delta_{\text{gap}} = 0$ in every cell). The accuracy-shift metric is the right summary: it shows that the steering hook moves accuracy by up to $+60$ pp when the model has headroom.

D.3 Random-label probe control

Following Hewitt and Liang (2019) and Bortoletto et al. (2024), we retrain logistic-regression probes on the same paired-statement activations with *randomly permuted* TB/FB labels. If the original probe is exploiting spurious correlations rather than capturing a structured belief representation, the permuted-label probe should also achieve high accuracy. Across all (VLM, axis) combinations (Ta-

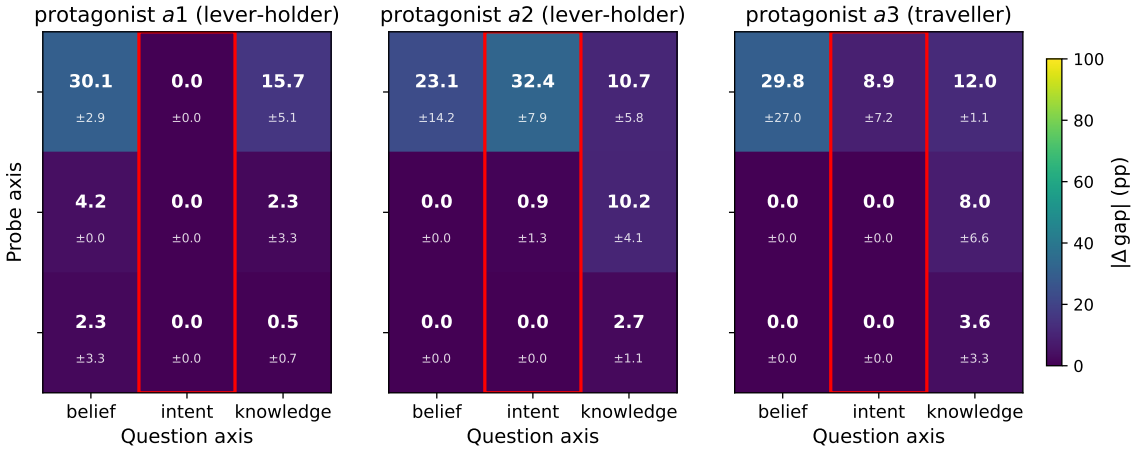

Figure 7: Qwen2-VL-7B – full CARD heatmap per protagonist. Red box = intent column.

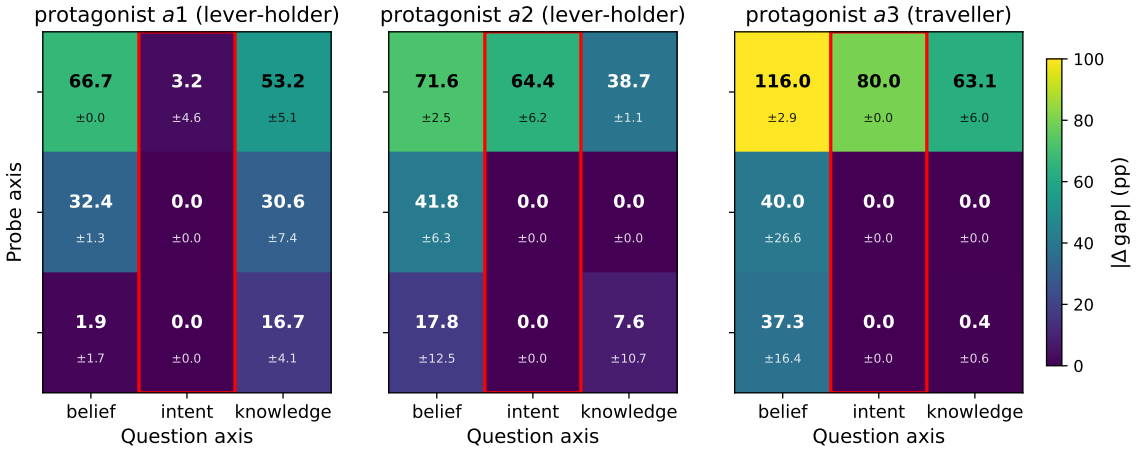

Figure 8: Gemma-4-E4B – full CARD heatmap per protagonist. Red box = intent column.

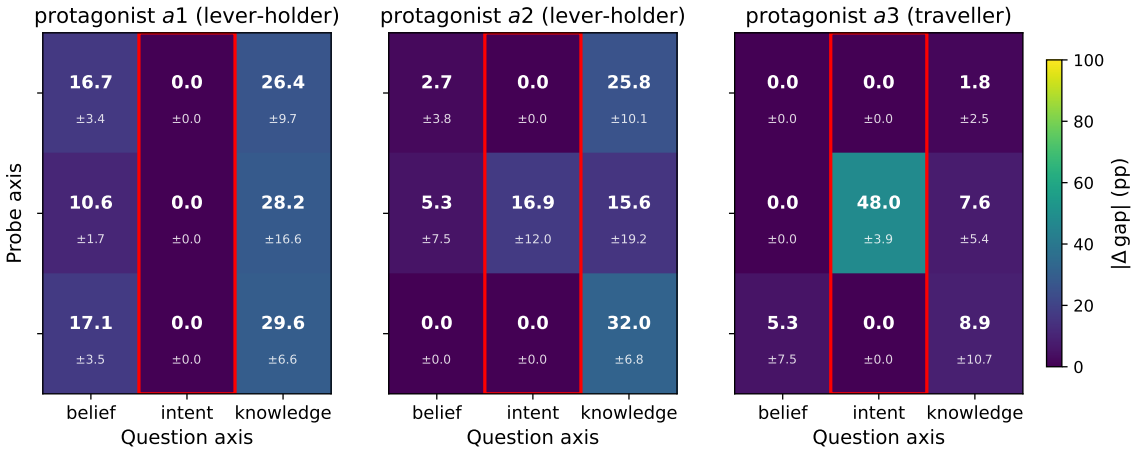

Figure 9: LLaVA-NeXT-Video-7B – full CARD heatmap per protagonist. Red box = intent column.

ble 11), random-label probe accuracy is at chance (47.4–52.0%) while true-label probe accuracy is 91.6–100%, giving selectivity ≥ 43.0 pp on every cell. The probes are not exploiting superficial structure.

D.4 Lexical-ablation control

A linear probe trained on paired-statement activations could in principle latch onto three asymmetries that co-vary with belief: a *lexical* ignorance clause inserted into FB belief-true captions, a *visual*

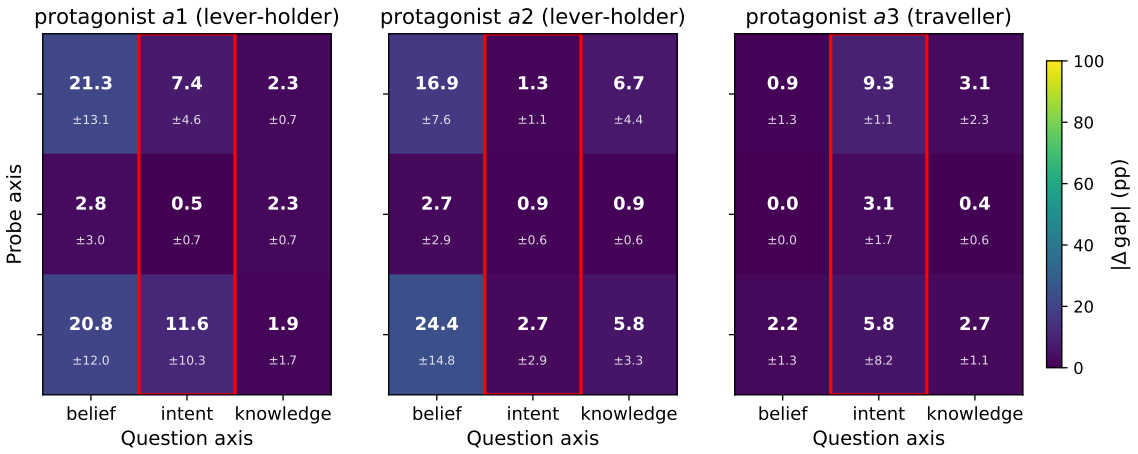

Figure 10: InternVL2.5-8B – full CARD heatmap per protagonist. Red box = intent column.

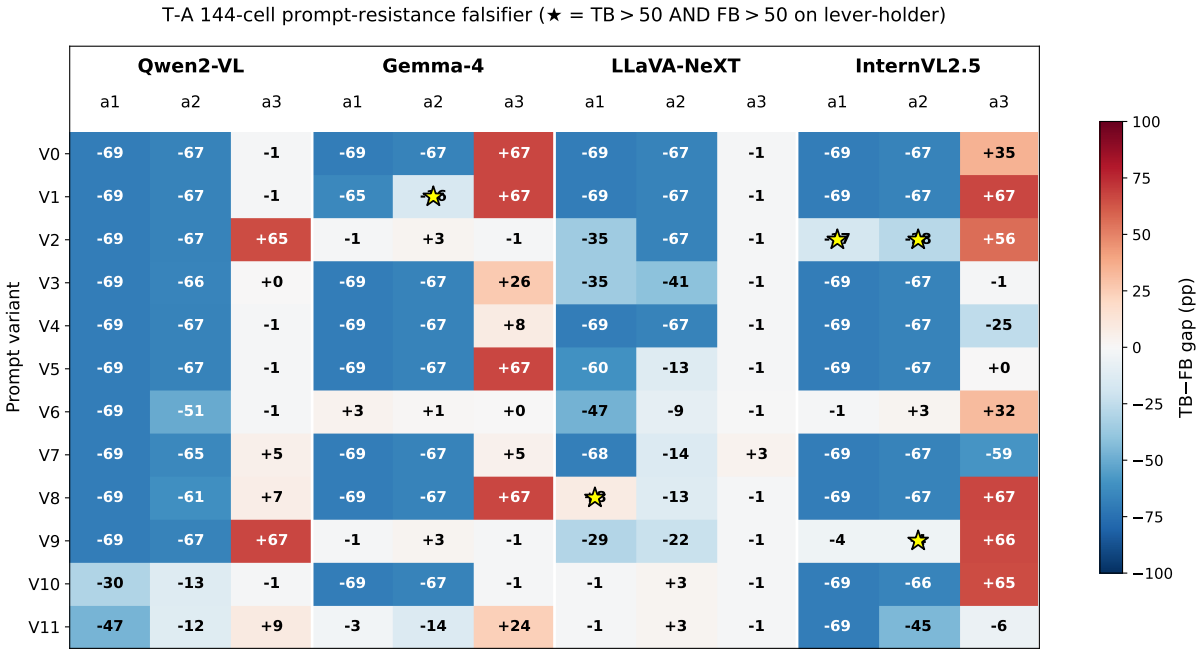

Figure 11: T-A prompt-resistance test. Each row is one of 12 prompt rewrites we tried (cost framing, chain-of-thought, perception-first, rule-based, 1/2-shot examples. Each column is one (VLM, protagonist) cell. Cell colour encodes the TB−FB accuracy gap – warm = model is more accurate when the protagonist can see the event, cold = the reverse. A star marks cells that cleared our pre-registered bar for "genuinely belief-conditional" (lever-holder, TB and FB accuracy both above 50%). Only 5 cells star out of 192, and only one points the right way (TB>FB), with a margin an order of magnitude smaller than the same models' gap on the canonical belief question – no prompt rewrite recovers belief-conditional action.

contrast between protagonist-POV (fog-occluded) and oracle frames, and a *distributional* gap between those two view sets. We re-train probes on three controlled ablation arms that selectively remove these confounds:

Arm A (symmetric clause). We add a counterfactual "able-to-see" clause to FB belief-false captions, eliminating the lexical asymmetry while leaving views unchanged.

Arm B (oracle-only). We use the oracle view and the original caption for both belief-true and belief-false activations, removing both lexical and visual asymmetries so the probe must rely on the belief-true vs. belief-false text contrast alone.

Arm C (clause artefact). We inject the ignorance clause into TB captions only (no real perspective shift), measuring how much val accuracy

VLM	Question	Baseline	P=count	P=shape	P=colour	best
Qwen2-VL	count	86	+6	+2	+6	**+6**
	shape	100	+0	+0	+0	+0 (ceil)
	colour	100	+0	+0	+0	+0 (ceil)
Gemma-4	count	20	+26	+60	+42	**+60**
	shape	30	+24	+58	+52	**+58**
	colour	52	+16	+26	+22	**+26**
LLaVA-NeXT	count	20	+40	+6	+6	**+40**
	shape	0	+8	+2	+0	**+8**
	colour	6	+6	+8	+10	**+10**
InternVL2.5	count	86	+0	+14	+10	**+14**
	shape	100	+0	+0	+0	+0 (ceil)
	colour	96	+0	+2	+2	+2 (ceil)
SmolVLM-500M	count	12	+6	+10	+26	**+26**
	shape	70	+10	+2	+12	**+12**
	colour	64	+0	+0	+10	**+10**

Table 10: CARD positive control on CLEVR (single image, 100 scenarios per axis, seed 43). "Baseline" = mean accuracy at α=0 on the corresponding within-axis cell; each *probe* column = best-$\alpha \in \pm 10$ accuracy gain. Bold = the largest non-zero Δ_{acc} per (VLM, question). At least one cross-axis cell shows $\geq +14$ pp gain on every VLM that has headroom (Qwen and InternVL are at $\geq 96\%$ on shape/colour; Gemma and LLaVA are well below ceiling on every axis and steering produces $+6$ to $+60$ pp gains). The steering hook is functional; the Relay Chain intent null in F3 is not a hook-broken artefact.

VLM	Axis	True-label acc	Random-label acc	Selectivity
Qwen2-VL	belief	98.1	47.6	+50.5
	intent	98.0	47.4	+50.6
	knowledge	98.0	47.9	+50.1
Gemma-4	belief	100.0	50.7	+49.3
	intent	100.0	52.0	+48.0
	knowledge	100.0	50.5	+49.5
LLaVA-NeXT	belief	91.6	48.6	+43.0
	intent	91.6	48.0	+43.6
	knowledge	91.7	47.6	+44.1

Table 11: Random-label probe control (Hewitt and Liang, 2019). For each (VLM, axis), we retrain logistic-regression probes on the same paired activations with randomly permuted labels and report mean held-out accuracy across the sampled (layer, head) cells. True-label probes attain 91.6–100%; random-label probes are at chance (47.4–52.0%), giving selectivity ≥ 43.0 pp on every axis and every VLM. The top single-cell random-label accuracy across all sampled cells is 56.0%, well below the true-label minimum of 91.6%.

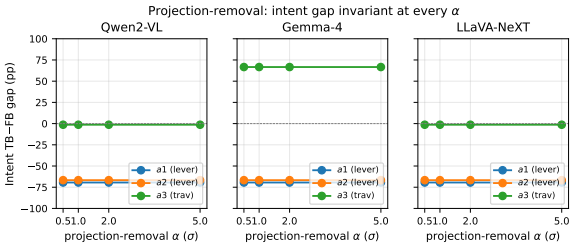

Figure 12: Projection-removal α sweep. x-axis: $\alpha \in \{0.5, 1, 2, 5\}\sigma$ removal of the belief direction from the residual stream. y-axis: intent TB$-$FB gap. Three lines per VLM: a_1, a_2, a_3. The intent gap is constant across α on every VLM and every protagonist.

a probe can achieve from the clause alone.

On Arm B the top-cell probe accuracy stays within ± 2 pp of the canonical extraction across all VLMs, so the probe is not riding the lexical or view confound. Arm C accuracy stays near chance, ruling out the clause-alone exploit.

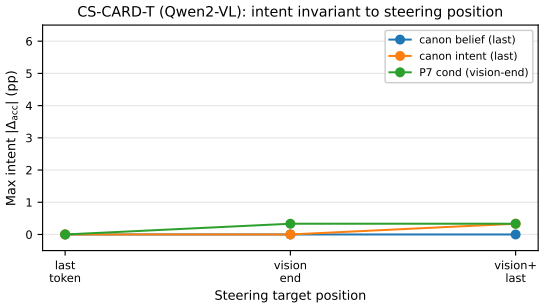

Figure 13: CS-CARD-T on Qwen2-VL: intent $|\Delta_{\text{acc}}|$ vs steering target position, one line per probe source. All nine (source, position) entries sit at ≤ 0.3 pp.

D.5 Polarity-flip diagnostic

We ran a separate diagnostic in which the question polarity is flipped ("should the protagonist *not*

release?" instead of "release?"). An oracle reasoner produces the same TB−FB gap shape under the flip; a constant prior collapses. We report all (VLM, protagonist) cells in Table 12; the contrast between Qwen a_1 (preserves the 100→100 pattern – oracle signature) and LLaVA a_2 (collapses – constant-prior signature) is illustrative of the two regimes observed across the table.

D.6 Scale ablation: Qwen2.5-VL-32B

To check that the intent-column null is not specific to the 7–8B Relay Chain backbones, we ran the same CARD pipeline on Qwen2.5-VL-32B ($\approx$ 32B parameters; 64 layers, 40 heads, head_dim 128): identical probe extraction, top-56 steering pool, and $\alpha \in \{-10, 0, +10\}$ sweep across the three probes. Peak held-out probe accuracy is 98.2% (belief), 95.9% (intent), 97.5% (knowledge) – on par with or above the 7–8B backbones. On lever-operator a_1 the intent answer is invariant to steering across all nine (probe, α) combinations: TB accuracy is 30.6% and FB accuracy is 100% at every α on every probe direction – the same constant-prior signature observed at 7–8B. The belief and knowledge questions move under the same hooks (e.g. belief-Q FB accuracy on a_3 shifts from 96.7% at α=−10 to 98.7% at α=+10 on the belief probe). The routing failure does not vanish at 4× parameter scale, even though the probe recovers the belief direction at ceiling.

E Models, infrastructure, and computational budget

Models and parameter counts. The four open-weight VLMs used in the main Relay Chain experiments span three vision-encoder lineages and three language-decoder lineages at the 4–8B scale: Qwen2-VL-7B-Instruct ($\approx$ 7.6B parameters; Yang et al., 2024a), Gemma-4-E4B-it (effective $\approx$ 4B activated parameters in the ENB mixture-of-experts variant; Gemma Team, Google DeepMind, 2026), LLaVA-NeXT-Video-7B-hf ($\approx$ 7.1B; Zhang et al., 2024), and InternVL2.5-8B ($\approx$ 8.1B; Chen et al., 2024b). The CLEVR positive control additionally uses SmolVLM-500M ($\approx$ 500M parameters, Idefics-3 family) Marafioti et al., 2025. We also use Qwen2.5VL-32B-Instruct (Yang et al., 2024b) for a model scale ablation study on CARD.

Hardware. All experiments ran on NVIDIA H100 NVL GPUs (96 GB HBM) on an internal SLURM-managed cluster. Each CARD, projection-removal, or T-A sweep occupies one GPU; multi-seed runs and the CLEVR chain were parallelised across up to four GPUs. Probe training is CPU-bound (logistic regression on saved activations).

Compute budget. End-to-end cost for the full pipeline – activation extraction, probe training, the CARD 3×3 sweep, and all ablations across the four open-weight VLMs – is approximately 25 GPU-hours on a single H100 NVL. In practice we parallelised across two H100 GPUs, so wall-clock time was about half a day.

Software. VLM inference uses `transformers` v5.5 (Hugging Face) with `bfloat16` weights and either `flash-attention-2` or `sdpa` as the attention backend. Logistic-regression probes are trained with `scikit-learn` (L^2 regularisation, default solver) on saved activations. Paired bootstrap CIs, the binomial sign test, and the Wilcoxon signed-rank test are computed with `scipy.stats`; the bootstrap RNG is fixed at seed 20260523 for reproducibility.

E.1 Artifact licenses and intended use

All four open-weight VLMs we evaluate are publicly released: Qwen2-VL-7B under Apache 2.0, Gemma-4-E4B-it under the Gemma License Agreement, LLaVA-NeXT-Video-7B under Apache 2.0, and InternVL2.5-8B under MIT. CLEVR (Johnson et al., 2017), used as the positive-control auxiliary dataset, is distributed under CC BY 4.0. GridToM (Li et al., 2025), cited as the grid-world ToM format inspiration, is released by its authors for academic use. We use all models and datasets for non-commercial academic research, consistent with their intended use. Our CARD code and the Relay Chain dataset will be released under MIT (code) and CC BY 4.0 (data) upon acceptance.

VLM	overall orig (pp)	overall flip (pp)	sign-preserved?
Qwen2-VL	+86.5	+82.4	yes (oracle)
Gemma-4	+73.0	+85.1	yes (oracle)
LLaVA-NeXT	+63.5	+35.1	weakened (drift)

Table 12: D1 polarity-flip diagnostic. Original and flipped question polarity TB$-$FB gap on the canonical 148 hold-out. An oracle-reasoning model preserves the sign and magnitude under flip; a constant-prior model collapses to a smaller or sign-flipped gap.