SAME STATE, DIFFERENT DECISIONS: DISENTANGLING FORMATION AND CONTEXT-CONDITIONED EXECUTION MECHANISMS IN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) often achieve higher accuracy on textual inputs than on semantically matched visual inputs describing the same task instance, with both modalities ultimately processed by the same decoder. We investigate this performance gap by causally delving into the decision process from modality-specific computation, through a cross-modal relay, to the downstream computation that continues to shape the final prediction. Our analysis identifies a relay state at which decision states transfer bidirectionally between visual and textual continuations. Yet identical relay states can still produce different predictions under different retained contexts, revealing that cross-modal executability can emerge before final decision commitment. This exposes a factorized decision process with distinct causal contributions from state formation and context-conditioned execution, while further localization attributes the residual execution effect primarily to post-relay attention reading distributed prefix information. Guided by these findings, we introduce Factorized Counterfactual Relay Distillation (FCRD), which distills Formation and Execution effects into a lightweight corrector that jointly updates the relay state and its surrounding context. Across four VLMs and five paired tasks, our method closes over 60% of the text-vision performance gap on average, demonstrating the consistency of the identified decision structure across model architectures and task types. Our analysis dissects the internal mechanisms underlying this gap and provides principled targets for correcting visual errors.

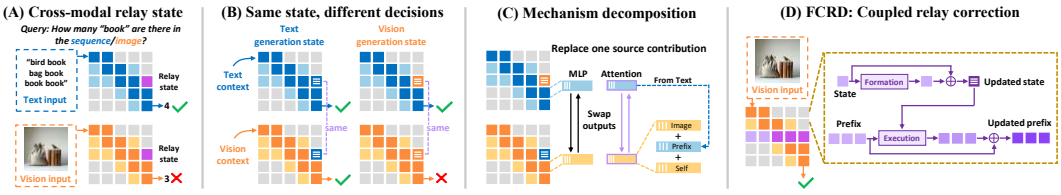


Figure 1: Overview of our causal analysis of cross-modal decisions and coupled relay correction.

1 INTRODUCTION

Recent years have witnessed remarkable progress in building vision-language models (VLMs) that support reasoning across both textual and visual modalities (Li et al., 2025; Wu et al., 2026). However, a pervasive performance asymmetry exists between semantically matched visual and textual tasks, with visual inputs typically being more difficult (Cohen & Mooney, 2025; Wang et al., 2026; Yan et al., 2026). For example, a model may correctly solve a counting or arithmetic problem from a textual description while failing on its visual counterpart. Notably, this discrepancy remains even when the paired inputs express the same task and are ultimately processed by the same model toward a common output space, which suggests that the gap reflects a deeper difference in how modality-specific information is converted into decisions (Huang et al., 2025). To address this asymmetry, one must explain how modality-specific computation is progressively transformed into a final decision within a shared model.

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.

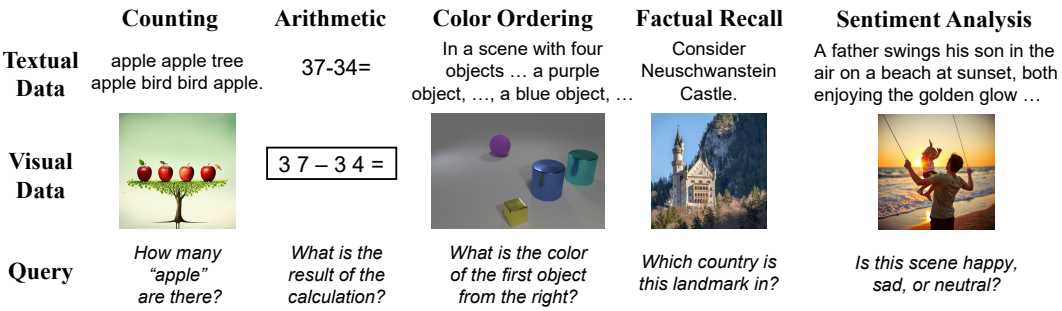


Figure 2: Five paired tasks. Each column shows a Text input (top), its Vision counterpart (middle), and the corresponding query (bottom).

Recent mechanistic studies have shown that analogous visual and textual tasks can recruit substantially different internal circuits and have identified modality-specific data processing as a major source of their performance gap (Nikankin et al., 2025). Related work has further separated visual evidence extraction from subsequent reasoning over the extracted evidence (Avogaro et al., 2026; Gou et al., 2026). These studies identify where visual and textual computations diverge, but leave unexplained how such differences propagate and are ultimately translated into different predictions. We therefore investigate the computational transition from modality-specific processing to decision making (Figure 1).

Across four VLMs and five paired visual–textual tasks, we use layerwise bidirectional causal interchange to identify a cross-modal relay before the final output, where a decision-position state produced under either modality can be executed by the downstream computation of the other. Correct and erroneous decisions that arise naturally in the model transfer through the same relay, establishing its role in native decision making. Strikingly, an identical relay state can preserve a visual error under one modality-conditioned context and recover the correct answer under another. The frequency of such reversals varies sharply across models and tasks, which reveals a continuum from strongly context-sensitive to nearly state-dominated relays and separates cross-modal executability from decision commitment.

To explain this variation, we construct a 2×2 state–context factorization that separates *Formation*, the effect of changing the state reaching the relay under fixed context, from *Execution*, the effect of changing the context acting on the same state. Their relative contributions differ substantially across models and tasks and induce distinct repair structures among natural errors. Localizing Execution through component interchange shows that post-relay Attention carries markedly stronger bidirectional effects than matched MLP computation. Decomposing Attention by source position further attributes most of this effect to broader prefix representations, which identifies context reading as a key mechanism that continues to shape the decision after the relay.

Guided by these findings, we introduce Factorized Counterfactual Relay Distillation (FCRD), which distills counterfactual Formation and Execution effects into a lightweight corrector that uses only native visual computation at inference and improves performance consistently across multiple models and tasks. Our results provide a causal account of the internal transition from modality-specific processing to shared decision making, and verify that understanding this transition exposes effective targets for correcting visual errors.

2 Paired Tasks and Decision-State Interventions

For task instance i, let (x_i^V, x_i^T, y_i) denote paired Vision and Text inputs with target y_i. After zero-based decoder block l, write the residual array as $R_{l,i}^M = (C_{l,i}^M, S_{l,i}^M)$ for $M \in \{V, T\}$, where $S_{l,i}^M \in \mathbb{R}^d$ is the *decision-position state* at the final prompt position and $C_{l,i}^M \in \mathbb{R}^{(L_i^M - 1) \times d}$ contains the preceding positions of a length-L_i^M prompt. Native denotes unmodified model computation. A state interchange (Geiger et al., 2021) replaces only S while preserving the recipient's context, sequence length, positional configuration, and attention mask. The frozen suffix resumes at $l + 1$

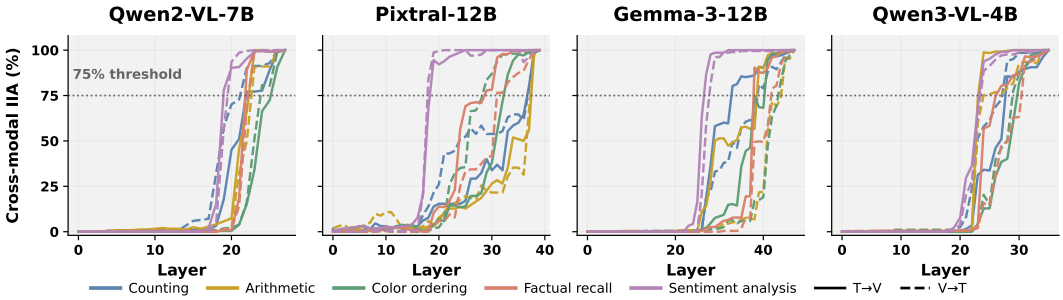


Figure 3: Layerwise donor-answer IIA on retained native-clean target-donor pairs across four VLMs and five tasks. Solid and dashed curves show Text-to-Vision and Vision-to-Text interchange, respectively. The dotted line marks the fixed 75% threshold.

and produces candidate logits $f_l^M(C, S)$ and decision $\pi_l^M(C, S)$. Exchanging one d-dimensional vector accommodates different prefix lengths.

Tasks and datasets. As shown in Figure 2, we apply the five paired tasks introduced by Nikankin et al. (2025). **(1) Object Counting** queries the number of instances of a specified object type. **(2) Arithmetic** evaluates a two-operand expression. **(3) Color Ordering** queries an object's color at a specified position in a CLEVR (Johnson et al., 2017) scene. **(4) Factual Recall** queries a property of a named or pictured entity. **(5) Sentiment Analysis** classifies a scene as happy, sad, or neutral. Each task pairs an image with a textual representation of the same instance. We evaluate the first answer token within the task's candidate set after merging duplicate token IDs, so Arithmetic accuracy concerns the first result token. Appendices A.1 and A.2 detail prompts and task construction. We also evaluate FCRD on IsoBench (Fu et al., 2024) and TabMWP (Lu et al., 2022) to assess its effectiveness on public reasoning benchmarks (Appendix I).

Models and populations. We study four popular transformer-based VLMs: Qwen2-VL-7B-Instruct (Wang et al., 2024), Pixtral-12B (Agrawal et al., 2024), Gemma-3-12B-IT (Team et al., 2025), and Qwen3-VL-4B-Instruct (Bai et al., 2025), all of which use an "adapter" architecture Liu et al. (2023). We use each model's native input processing and chat template and keep its pretrained weights frozen throughout the intervention analyses and corrector training. Our analysis covers all 20 model-task settings and focuses on understanding the common accuracy advantage that textual inputs exhibit over their visual counterparts in most settings. Appendix A.4 provides the details of the analysis populations and layer boundaries along , while Appendix A.3 describes the intervention protocols. All experiments were conducted on a single NVIDIA A800 GPU with 80 GB of memory.

3 An Executable Relay Precedes Decision Commitment

3.1 Cross-Modal Executability Across Depth

To distinguish state formation from subsequent state use, we locate a *cross-modal relay* where either modality's answer-specific state can guide the other's continuation. Let $\mathcal{P}$ contain retained pairs (i, j) of target i and donor j from the fixed upstream all split, with different answers and both examples natively correct in Vision and Text. Matching preserves the query's semantic group and controls nuisance variation through task-specific distances where available (Appendix B). For each pair, we replace the target's decision-position state with the donor's at each layer and measure interchange intervention accuracy (IIA) from source modality M' to recipient modality M in both directions as

$$\mathrm{IIA}_{M'\to M}(l) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \mathbf{1}\left[\pi_l^M(C_{l,i}^M, S_{l,j}^{M'}) = y_j\right]. \tag{1}$$

For each model–task setting, the relay layer H is the earliest layer where both Text-to-Vision and Vision-to-Text IIA reach at least 75%, and remains fixed in subsequent analyses.

Figure 3 traces Text-to-Vision and Vision-to-Text IIA across layers and shows that every setting reaches the bilateral 75% threshold before the final layer. Table 1 reports the selected relay layer

Table 1: **Relay layers and corresponding IIA.** Each cell reports H with IIA percentages for $\mathrm{T}\rightarrow\mathrm{V}/\mathrm{V}\rightarrow\mathrm{T}$ in parentheses.

Task	Qwen2-VL-7B	Pixtral-12B	Gemma-3-12B	Qwen3-VL-4B
Counting	22 (77.2% / 90.2%)	37 (80.0% / 78.5%)	39 (90.9% / 81.2%)	28 (85.6% / 94.1%)
Arithmetic	23 (100.0% / 90.6%)	38 (100.0% / 98.0%)	45 (99.6% / 96.0%)	24 (98.8% / 75.3%)
Color Ordering	26 (90.8% / 99.1%)	33 (86.5% / 97.6%)	44 (99.6% / 83.1%)	30 (81.8% / 86.6%)
Factual Recall	23 (99.6% / 99.1%)	31 (96.1% / 76.5%)	42 (96.1% / 78.8%)	31 (96.6% / 93.1%)
Sentiment Analysis	20 (90.2% / 93.5%)	19 (94.5% / 98.8%)	28 (97.6% / 81.4%)	24 (93.8% / 88.5%)

H and its corresponding IIA in both directions, showing that cross-modal executability emerges at task- and model-specific depths. Transfer between correctly solved instances establishes an answer-specific interface, while connecting this interface to the native Text–Vision gap requires examining the conflicting decisions produced for the same instance. We therefore test whether the correct Text state repairs the visual error and the erroneous Vision state reproduces that error in Text.

3.2 Native Correct and Erroneous Decisions Transfer

For each paired example where native Text is correct and native Vision predicts an error $w_i \neq y_i$, we transfer the correct Text state into the Vision context and the erroneous Vision state into the Text context at H. The correct-state rescue rate (CRR) measures exact recovery of y_i in the first direction, while the error transfer rate (ETR) measures exact reproduction of the original w_i in the second. This analysis retains examples with a feasible, semantically matched Text-clean control donor whose answer differs from both y_i and w_i; its denominator N_C therefore includes the control-feasibility filter (Appendix C).

CRR ranges from 82.6% to 100% and ETR from 8.3% to 98.7% (Figure 4), showing that the relay carries both correct answers and specific native errors. The variation in error transfer motivates isolating the recipient context's role by comparing continuations initialized with the same relay state.

3.3 Identical States Can Lead to Different Decisions

To isolate context dependence, we extend the comparison to all N_D eligible Text-correct/Vision-wrong examples and initialize both continuations with the same native Vision state $S^V_{H,i}$ at H. This population omits the control-donor requirement and can therefore exceed N_C. *Same-state divergence* measures how often $\pi^T_H(C^T_{H,i}, S^V_{H,i})$ differs from the native Vision decision $\pi^V_H(C^V_{H,i}, S^V_{H,i}) = w_i$. *Exact recovery* measures how often the Text-context continuation returns y_i, distinguishing error correction from changes to another wrong answer.

Figure 5 shows divergence rates of 1.3%–91.7% when both continuations receive the same native Vision state at H. Pixtral Sentiment reaches 91.7% divergence and 90.3% exact recovery, so nearly all context-induced prediction changes repair the visual error. Pixtral Arithmetic exhibits only 1.3% divergence and no recovery, preserving the error under almost every context swap. This contrast shows that context sensitivity varies substantially across tasks. On Counting, divergence exceeds recovery across three of the four models, indicating that context also redirects some predictions to

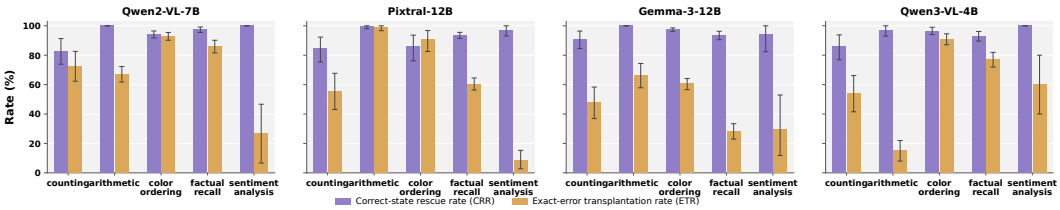


Figure 4: Native decision transfer at the relay on the control-eligible Text-correct/Vision-wrong population. CRR is exact ground-truth rescue in Vision; ETR is exact transfer of the original Vision error into Text. Whiskers denote 95% bootstrap intervals.

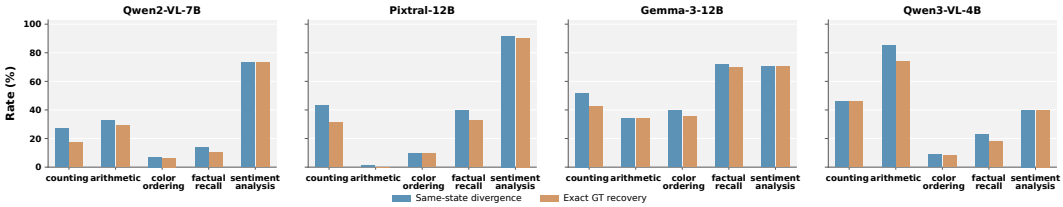


Figure 5: Same-state divergence and exact ground-truth recovery when the native Vision relay vector is held identical under Vision and Text contexts. Bars show point estimates on the eligible Text-correct/Vision-wrong population from the upstream native audit.

another incorrect answer. Together with the state-transfer results, these findings establish that both state and context shape the final decision. We therefore cross the two states and contexts to quantify their contributions to the accuracy gap and how state effects depend on context.

4 Separating Formation and Execution

At the relay, we combine each eligible pair's native states and contexts. Context-first, state-second subscripts define the candidate-logit vector $\ell_{ab,i} = f_H^a(C_{H,i}^a, S_{H,i}^b)$, its prediction $\hat{y}_{ab,i} = \arg\max_c \ell_{ab,i,c}$, and accuracy A_{ab} for $a, b \in \{V, T\}$. Conditions VV and TT are native; VT changes the state and TV changes the context relative to VV.

$$F = \tfrac{1}{2}\big[(A_{VT} - A_{VV}) + (A_{TT} - A_{TV})\big], \quad E = \tfrac{1}{2}\big[(A_{TV} - A_{VV}) + (A_{TT} - A_{VT})\big],$$
$$G = A_{TT} - A_{VV} = F + E, \qquad I = A_{TT} - A_{VT} - A_{TV} + A_{VV}. \tag{2}$$

Formation and Execution average state and context effects over the two backgrounds, respectively, while I measures their dependence. The identities $F = (A_{VT} - A_{VV}) + I/2$ and $E = (A_{TV} - A_{VV}) + I/2$ show that the decomposition allocates the interaction equally and recovers the native accuracy gap exactly as $G = F + E$.

Formation has the larger absolute point effect in all 20 settings (Figure 6), including the two Pixtral settings where the native Text–Vision gap and Formation effect are negative. Execution reaches 12.9% on Pixtral Sentiment, 6.8% on Pixtral Factual Recall, and 6.3% on Gemma Factual Recall, so the larger Formation effect coexists with a consequential context contribution. Interaction is negative throughout, from -25.8 to -0.2 points: the increment from changing one coordinate decreases after the other coordinate has changed.

Individual errors further distinguish the two repair routes. Among factorial rows with $\hat{y}_{VV} \neq y$ and $\hat{y}_{TT} = y$, an error is *Formation-only* when only VT recovers y, *Execution-only* when only TV recovers it, *dual-rescuable* when both hybrids recover it, and *interaction-dependent* when both hybrids remain wrong. As shown in Table 7, Formation-only errors account for 88.4% of this Qwen2 Color Ordering cohort and 148/149 Pixtral Arithmetic errors, whereas 12/17 Gemma Sentiment errors are dual-rescuable. The population contrasts and error taxonomy therefore describe complementary scales of the same intervention: state formation has the larger average effect, while context can supply an alternative route to recovery for particular errors. We next localize the post-relay computation that enables this context-driven recovery.

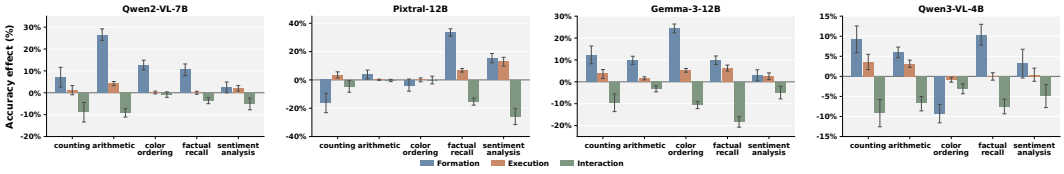


Figure 6: Formation, Execution, and interaction contrasts in candidate accuracy across the eligible factorial populations reported in percentage points. Whiskers implies 95% bootstrap intervals.

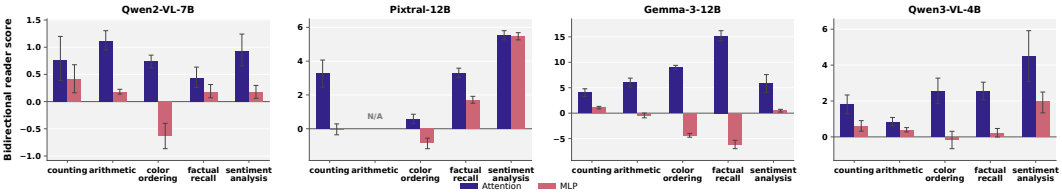


Figure 7: Bidirectional reader scores for Attention and MLP output interchange over the same setting-specific frozen window on the exact-recovery contradiction cohort. Whiskers show 95% bootstrap intervals. Pixtral Arithmetic is N/A because its exact-recovery contradiction cohort is empty.

5 Attention Reads Context After the Relay

5.1 A Matched Comparison of Attention and MLP

To locate the computation responsible for context-driven recovery, we use the fixed *same-state contradiction cohort* where S_H^V yields w under Vision context and y under Text context. We fix this cohort and a post-relay *reader window* from upstream analyses, then cache decision-position Attention and MLP outputs from both same-state continuations before component intervention. Swapping each component class bidirectionally over the same layers and position compares their effects under matched coverage.

Let V and T denote the two same-state continuations. For $X \in \{\mathrm{Attn}, \mathrm{MLP}\}$, we measure effects with the logit margin $m = \ell_y - \ell_w$ (Zhang & Nanda, 2024) against the original Vision error. Let $V \leftarrow T_X$ denote Vision-context continuation with cached Text-context outputs of class X, and define $T \leftarrow V_X$ analogously. The bidirectional reader score is

$$B_X = \tfrac{1}{2}\big[\underbrace{m(V \leftarrow T_X) - m(V)}_{\text{Text-to-Vision effect}} - \underbrace{\{m(T \leftarrow V_X) - m(T)\}}_{\text{Vision-to-Text effect}}\big]. \tag{3}$$

Forward shifts toward y and reverse shifts toward w both increase B_X, so $B_{\mathrm{Attn}} - B_{\mathrm{MLP}}$ compares how strongly the components carry the context-dependent decision difference. Each intervention uses fixed cached outputs whose effects propagate through the recipient's remaining computation.

Attention produces positive reader scores above MLP in every evaluable setting (Figure 7), with Pixtral Sentiment nearly tied at 5.53 and 5.47. The matched comparison identifies Attention as the stronger carrier of context-conditioned Execution and motivates tracing which prefix positions supply its weighted-value output.

5.2 Source Positions Supplying the Reader

To determine where post-relay Attention obtains the information that shapes the decision, we partition its weighted-value output by source position. For head h at each reader layer, attention weights a_{qj}^h combine source values v_j^h at decision position q as

$$z^h = \sum_j a_{qj}^h v_j^h = z^h_{\mathrm{visual}} + z^h_{\mathrm{nonvisual}} + z^h_{\mathrm{self}}, \tag{4}$$

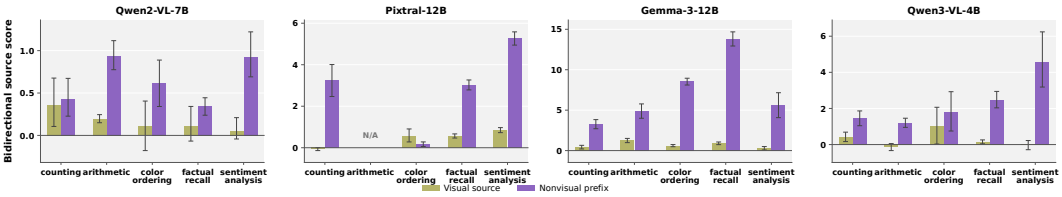


Figure 8: Bidirectional source scores for visual and nonvisual-prefix weighted-value contributions on the reader cohort. Whiskers show 95% bootstrap intervals; N/A denotes the empty Pixtral Arithmetic cohort.

where the disjoint groups contain image positions, non-image prefix positions, and the decision position. Before output projection, we replace one current recipient contribution with its cached counterpart from the other unmodified same-state continuation. Since Text has no image positions, visual-source interchange removes the current visual contribution from Vision and injects the cached Vision contribution into Text. Nonvisual-source interchange swaps the aggregate prefix contribution in both directions. Other current contributions remain unchanged without attention renormalization, and Equation 3 scores the final margin changes. Appendix H details the operations and windows.

As presented in Figure 8, it is evident that the nonvisual prefix produces the stronger source score in 18 of 19 settings. These positions can already encode visual information, locating the later read in distributed contextual representations. Although source outputs add linearly, their effects can interact through the nonlinear suffix. This motivates updating context alongside the relay state so that post-relay Attention reads their corrected representations.

6 FCRD: Correcting Visual Decisions at the Relay

6.1 Coupled Formation and Execution Updates

FCRD turns the state–context decomposition into coupled low-rank updates at H, allowing the frozen suffix to process a corrected state and context together. Here, S_H and C_H denote native Vision residuals. The Formation branch uses S_H and its change from $H-1$. With affine-free layer normalization LN and $\phi = \mathrm{GELU}$, its update is

$$\delta S_H = U_F \phi(V_S \mathrm{LN}(S_H) + V_D \mathrm{LN}(S_H - S_{H-1})) \,. \tag{5}$$

For context position p, let $t(p) \in \{\mathrm{vis}, \mathrm{pre}\}$ identify visual or nonvisual-prefix membership. The Execution update uses elementwise interaction $\odot$ in a shared latent space and source-specific output maps,

$$u_p = \phi(W_C \mathrm{LN}(C_{H,p})) \odot \phi(W_S \mathrm{LN}(S_{\mathrm{cond}})) \,, \qquad \delta C_{H,p} = U_{t(p)} u_p. \tag{6}$$

All down-projections map d dimensions to rank r and all up-projections map r to d. FCRD first computes $S'_H = S_H + \delta S_H$, then conditions Execution on $S_{\mathrm{cond}} = S'_H$ to obtain $C'_H = C_H + \delta C_H$. Both updates occur at H, and the frozen decoder resumes at $H+1$. For branch evaluations, F-only updates S_H while retaining the native context, whereas E-only updates C_H with $S_{\mathrm{cond}} = S_H$ and leaves the state unchanged.

The state and context substitutions VT and TV provide branch-specific teacher effects relative to VV. Let $\mathcal{C}(a) = a - \mathrm{mean}_c(a_c)\mathbf{1}$ center candidate logits, where $\mathbf{1}$ is the all-ones vector. Binary gates g_F, g_E indicate whether the corresponding hybrid reduces target-label negative log-likelihood relative to native Vision. The gated targets are

$$\Delta^\star_F = g_F \mathcal{C}(\ell_{VT} - \ell_{VV}), \qquad \Delta^\star_E = g_E \mathcal{C}(\ell_{TV} - \ell_{VV}). \tag{7}$$

Let z_F, z_E, and z_J denote the logits of F-only, E-only, and FCRD, respectively, and let z_0 denote native Vision replay logits. We optimize

$$\mathcal{L} = \mathrm{CE}(z_J, y) + \lambda_R \mathcal{R} + \lambda_{\mathrm{eff}} \Big[\mathrm{MSE}(\mathcal{C}(z_F - z_0), \Delta^\star_F) + \mathrm{MSE}(\mathcal{C}(z_E - z_0), \Delta^\star_E) \Big]. \tag{8}$$

Here, λ_{eff} weights effect supervision and λ_R weights relative update energy. Let $\mathcal{I}_{\mathrm{tr}}$ index the current training examples and ms denote mean squared entries. Embedding δC_H in the full residual array with a zero decision-position row gives δR_C, and

$$\mathcal{R} = \mathbb{E}_{i \sim \mathrm{Unif}(\mathcal{I}_{\mathrm{tr}})} \left[\frac{\mathrm{ms}(\delta S_{H,i})}{\max\{\mathrm{ms}(S_{H,i}), \varepsilon_R\}} + \frac{\mathrm{ms}(\delta R_{C,i})}{\max\{\mathrm{ms}(R_{H,i}), \varepsilon_R\}} \right]. \tag{9}$$

The fixed floor $\varepsilon_R > 0$ bounds the detached denominators. Labels supervise the gates and task loss, with closed gates assigning zero effect targets. Inference uses native Vision activations and the learned corrector at the fixed relay (Appendix I).

6.2 Performance and Branch Effects

For the controlled tasks, we cross-fit corrector parameters over five folds while fixing H from Section 3.1. Each fold evaluates F-only, E-only, and FCRD from one corrector trained on the other four folds. We pool held-out predictions within settings and average gains equally across settings.

Table 2: Pooled out-of-fold performance. Vision and Text report original VLM accuracies (%), while FCRD reports visual accuracy with our corrector. ΔF, ΔE, and $\Delta\mathrm{Acc}$ denote gains over Vision in percentage points with F-only, E-only, and both branches, respectively. GC measures the percentage of the positive Text–Vision gap closed by FCRD.

Model-Task	Vision	Text	ΔF	ΔE	FCRD	ΔAcc	GC
Qwen2-Counting	70.94%	79.32%	0.79%	3.40%	76.18%	$+\mathbf{5.24}\% \pm \mathbf{3.06}\%$	$\mathbf{62.50}\%$
Qwen2-Arithmetic	67.60%	98.30%	9.40%	2.50%	80.10%	$+\mathbf{12.50}\% \pm \mathbf{3.72}\%$	$\mathbf{40.72}\%$
Qwen2-Color Ordering	74.18%	86.96%	−0.10%	−0.21%	74.18%	$+\mathbf{0.00}\% \pm \mathbf{0.52}\%$	$\mathbf{0.00}\%$
Qwen2-Factual Recall	68.12%	78.64%	2.85%	0.08%	71.04%	$+\mathbf{2.93}\% \pm \mathbf{2.19}\%$	$\mathbf{27.82}\%$
Qwen2-Sentiment	93.03%	97.54%	4.10%	2.05%	96.72%	$+\mathbf{3.69}\% \pm \mathbf{2.65}\%$	$\mathbf{81.82}\%$
Pixtral-Counting	66.67%	53.75%	7.51%	2.40%	74.77%	$+\mathbf{8.11}\% \pm \mathbf{2.25}\%$	–
Pixtral-Arithmetic	21.70%	25.50%	6.50%	0.00%	29.80%	$+\mathbf{8.10}\% \pm \mathbf{1.75}\%$	$\mathbf{213.16}\%$
Pixtral-Color Ordering	81.05%	77.42%	0.40%	0.00%	81.45%	$+\mathbf{0.40}\% \pm \mathbf{0.55}\%$	–
Pixtral-Factual Recall	45.02%	85.13%	7.28%	0.87%	53.48%	$+\mathbf{8.47}\% \pm \mathbf{2.87}\%$	$\mathbf{21.10}\%$
Pixtral-Sentiment	70.08%	98.36%	24.59%	19.67%	95.08%	$+\mathbf{25.00}\% \pm \mathbf{1.85}\%$	$\mathbf{88.41}\%$
Gemma3-Counting	73.04%	89.01%	5.76%	3.93%	82.46%	$+\mathbf{9.42}\% \pm \mathbf{1.99}\%$	$\mathbf{59.02}\%$
Gemma3-Arithmetic	87.60%	99.00%	8.00%	3.00%	95.70%	$+\mathbf{8.10}\% \pm \mathbf{0.74}\%$	$\mathbf{71.05}\%$
Gemma3-Color Ordering	46.39%	76.10%	4.99%	3.32%	52.31%	$+\mathbf{5.92}\% \pm \mathbf{1.06}\%$	$\mathbf{19.93}\%$
Gemma3-Factual Recall	66.53%	82.67%	2.69%	2.61%	70.57%	$+\mathbf{4.03}\% \pm \mathbf{1.96}\%$	$\mathbf{25.00}\%$
Gemma3-Sentiment	93.03%	98.36%	2.05%	2.87%	97.54%	$+\mathbf{4.51}\% \pm \mathbf{0.97}\%$	$\mathbf{84.62}\%$
Qwen3-Counting	80.37%	93.19%	4.45%	4.45%	85.60%	$+\mathbf{5.24}\% \pm \mathbf{2.09}\%$	$\mathbf{40.82}\%$
Qwen3-Arithmetic	88.80%	97.90%	5.80%	0.90%	95.40%	$+\mathbf{6.60}\% \pm \mathbf{1.39}\%$	$\mathbf{72.53}\%$
Qwen3-Color Ordering	79.69%	69.56%	0.21%	0.31%	79.95%	$+\mathbf{0.26}\% \pm \mathbf{0.52}\%$	–
Qwen3-Factual Recall	62.26%	72.55%	1.66%	0.08%	65.43%	$+\mathbf{3.16}\% \pm \mathbf{1.34}\%$	$\mathbf{30.77}\%$
Qwen3-Sentiment	90.98%	94.67%	4.10%	0.82%	95.90%	$+\mathbf{4.92}\% \pm \mathbf{1.82}\%$	$\mathbf{133.33}\%$
Average	71.35%	82.70%	5.15%	2.65%	77.68%	$+\mathbf{6.33}\% \pm \mathbf{0.44}\%$	$\mathbf{63.09}\%$

As shown in Table 2, FCRD substantially narrows the modality gap, improving visual accuracy by 6.33% on average and closing 63.09% of the positive Text–Vision gap. Gap closure exceeds 80% on Sentiment for all four models, demonstrating that lightweight relay updates recover much of the textual performance advantage while keeping the VLM weights frozen. The branch comparison highlights the complementary benefits of state and context correction. On Gemma Counting, FCRD improves accuracy by 9.42%, compared with 5.76% for F-only and 3.93% for E-only, showing that context updates strengthen state correction. FCRD also achieves 76.46% macro-average repair precision, the proportion of correctness-changing cases that repair native visual errors (Appendix I).

To further evaluate FCRD beyond the controlled tasks, we apply the same correction architecture to IsoBench and TabMWP. As presented in Table 3, accuracy improves for all four VLMs with gains reaching 12.17% for Pixtral on IsoBench and 5.03% for Gemma on TabMWP. As shown in Table 3, FCRD achieves these gains with an extremely lightweight corrector whose trainable parameters account for only 0.0018%–0.0032% of the total model parameters. Appendix J presents detailed parameter counts. These results verify the superb generalization and effectiveness of FCRD with negligibly small additional trainable parameters.

6.3 Learned Effects Retain Factor Specialization

To relate performance gains to the intended branch specialization, we compare learned and intervention effects on held-out examples from the controlled tasks. We use the ground-truth margin

Table 3: Visual accuracy on the evaluated IsoBench and TabMWP subsets. Parentheses report gains over the original VLM in percentage points. The last row reports FCRD's trainable parameter ratios, with the corresponding parameter counts in parentheses.

Dataset	Setting	Qwen2-VL-7B	Pixtral-12B	Gemma-3-12B	Qwen3-VL-4B
IsoBench	Baseline	64.54%	58.67%	62.08%	71.25%
	FCRD	65.67% (+1.13%)	70.84% (+12.17%)	63.49% (+1.41%)	77.50% (+6.25%)
TabMWP	Baseline	68.37%	75.65%	71.23%	72.64%
	FCRD	71.79% (+3.42%)	77.66% (+2.01%)	76.26% (+5.03%)	75.40% (+2.76%)
Trainable Params		0.0024% (0.20M)	0.0023% (0.28M)	0.0018% (0.22M)	0.0032% (0.14M)

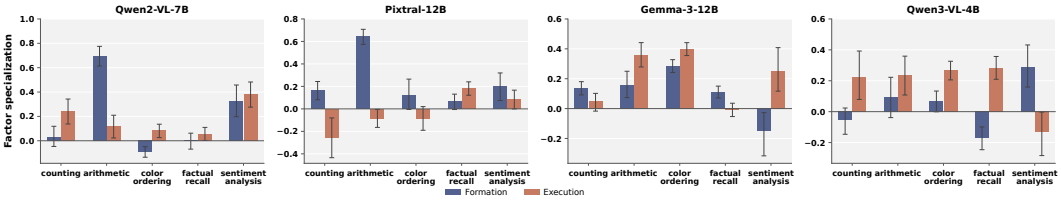


Figure 9: Formation and Execution specialization on held-out examples. Each panel shows one model with paired bars reporting D_F and D_E for each task. Positive values indicate stronger alignment with the corresponding factor. Whiskers denote 95% paired bootstrap intervals.

$q(\ell) = \ell_y - \max_{c \neq y} \ell_c$, recomputing the strongest competitor for each logit vector, and define intervention gains $O_F = q(\ell_{VT}) - q(\ell_{VV})$ and $O_E = q(\ell_{TV}) - q(\ell_{VV})$ alongside learned gains $S_F = q(z_F) - q(\ell_{VV})$ and $S_E = q(z_E) - q(\ell_{VV})$. Using Pearson correlation ρ, we compare how closely the two branches track each intervention effect through

$$D_F = \rho(O_F, S_F) - \rho(O_F, S_E),\ D_E = \rho(O_E, S_E) - \rho(O_E, S_F),\ S = \tfrac{1}{2}(D_F + D_E). \quad (10)$$

Positive D_F and D_E show that each branch better tracks its corresponding intervention effect. Figure 9 shows mean specialization scores of $D_F = 0.146$ and $D_E = 0.131$, with positive combined specialization in 18 of 20 settings. Arithmetic produces particularly strong Formation specialization in Qwen2 and Pixtral, while Gemma exhibits stronger Execution than Formation specialization on Arithmetic and Color Ordering. These patterns show that the learned branches track distinct intervention effects with task-dependent strengths. Together with the visual accuracy gains, this predictive alignment connects effective correction on unseen examples to the state and context contributions identified by the causal analysis.

7 Related Work

Mechanistic Interpretability of VLMs. Mechanistic studies trace visual information through hidden representations (Gandelsman et al., 2024; Neo et al., 2025) and identify task-specific layers and components (Shi et al., 2026; Li et al., 2026) alongside architecture-dependent communication routes (Serra et al., 2026). However, models can attend to relevant evidence while answering incorrectly (Liu et al., 2026), raising the question of whether a transferable decision state already fixes the answer. We address this question by holding the relay state identical across contexts, then use component and source interventions to identify the post-relay Attention computation that continues to shape the decision.

Cross-Modal Representations and Transfer. Shared representational geometry (Huh et al., 2024) and successful linear mappings (Merullo et al., 2022) support a common representational basis across modalities. However, geometric separation (Liang et al., 2022) and weak patch-level alignment (Jiang et al., 2026) persist alongside accuracy gaps on paired inputs (Fu et al., 2024; Nikankin et al., 2025). Such input-level comparisons change both state and context, motivating our Formation–Execution intervention to separate their contributions and quantify their interaction. Within the representation-finetuning paradigm (Wu et al., 2024), FCRD turns these counterfactual effects into supervision for coupled state and context corrections.

8 Conclusion

We investigate the Text-Vision accuracy gap by distinguishing state Formation from context-conditioned Execution in VLMs. We find that cross-modal executability emerges before decision commitment, as identical relay states can yield different answers under different contexts. Post-relay Attention mediates this context dependence by reading distributed prefix information. These findings guide FCRD, which distills the factorized intervention effects into coupled state and context corrections that improve visual accuracy and narrow the modality gap across models and tasks. Our findings establish Formation and Execution as complementary targets for correcting visual errors. Their task-dependent contributions motivate the development of adaptive correction methods that balance state and context updates to improve the performance of VLMs.

AI USE STATEMENT

We used generative AI tools to improve the clarity and readability of the manuscript. We manually reviewed the passages revised with AI assistance and checked their factual accuracy and consistency with our intended meaning. We also reviewed the entire manuscript and take full responsibility for its content.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide an anonymous supplementary package with experiment implementations for four VLMs and trained FCRD checkpoints for Qwen3-VL-4B-Instruct on Counting, IsoBench, and TabMWP. We include evaluation metadata, archived intervention records, and code for checkpoint inspection, inference, and result analysis. These materials allow readers to examine the supplied correctors and reconstruct the intervention analyses from saved experiment records.

We ran the Qwen3 experiments in float32 on a single NVIDIA A800 GPU with 80 GB of memory under Ubuntu 22.04, Python 3.10, PyTorch 2.7.1 with CUDA 12.8, and Transformers 5.17.0. We specify exact Python dependencies in `requirements.txt` and provide instructions for environment setup, model and dataset downloads, and checkpoint loading in `README.md`. The same README describes the supplied artifacts and gives commands for inference, scoring, and analysis.

We separate prediction generation from scoring against the dataset answer keys so that readers can inspect model outputs before calculating evaluation metrics. We also provide archived records and analysis scripts for checking the measured intervention effects and their summary statistics. Together, the checkpoints, experiment records, and documented evaluation workflow support independent examination of model behavior and verification of the corresponding results.

REFERENCES

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

Niccolo Avogaro, Nayanika Debnath, Li Mi, Thomas Frick, Junling Wang, Zexue He, Hang Hua, Konrad Schindler, and Mattia Rigotti. Sparc: Separating perception and reasoning circuits for test-time scaling of vlms. *arXiv preprint arXiv:2602.06566*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Vanya Cohen and Raymond Mooney. Met-bench: Multimodal entity tracking for evaluating the limitations of vision-language and reasoning models. *arXiv preprint arXiv:2502.10886*, 2025.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. Isobench: Benchmarking multimodal foundation models on isomorphic representations. *arXiv preprint arXiv:2404.01266*, 2024.

Yossi Gandelsman, Alexei Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *International conference on learning representations*, volume 2024, pp. 18395–18416, 2024.

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

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Xin Jin, Zhenguo Li, James Kwok, and Yu Zhang. Reasoning-aligned perception decoupling for scalable multi-modal reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 140359–140415, 2026.

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

Kung-Hsiang Huang, Can Qin, Haoyi Qiu, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. Why vision language models struggle with visual arithmetic? towards enhanced chart and geometry understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4830–4843, 2025.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jiachen Jiang, Jinxin Zhou, Bo Peng, Xia Ning, and Zhihui Zhu. Analyzing fine-grained alignment and enhancing vision understanding in multimodal language models. *Advances in Neural Information Processing Systems*, 38:874–899, 2026.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Black Forest Labs. Flux. *https://github. com/black-forest-labs/flux*, 2024.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31645–31653, 2026.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1578–1597. IEEE, 2025.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations*, volume 2026, pp. 72849–72876, 2026.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 66352–66385. Curran Associates, Inc., 2025. doi: 10.52202/085713-2228. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5fcd540792da599adf1b932624e98f1f-Paper-Conference.pdf.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alessandro Serra, Francesco Ortu, Emanuele Panizon, Lucrezia Valeriani, Lorenzo Basile, Alessio Ansuini, Diego Doimo, and Alberto Cazzaniga. The narrow gate: Localized image-text communication in native multimodal models. *Advances in Neural Information Processing Systems*, 38: 74684–74712, 2026.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision function layer in multimodal llms. *Advances in neural information processing systems*, 38:27510–27537, 2026.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xingrui Wang, Jiang Liu, Chao Huang, Xiaodong Yu, Ze Wang, Ximeng Sun, Jialian Wu, Alan Yuille, Emad Barsoum, and Zicheng Liu. Xmodbench: Benchmarking cross-modal capabilities and consistency in omni-language models. In *International Conference on Learning Representations*, volume 2026, pp. 19508–19521, 2026.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *Advances in Neural Information Processing Systems*, 38:143297–143330, 2026.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962, 2024.

Hao Yan, Xingchen Liu, Hao Wang, Zhenbiao Cao, Handong Zheng, Liang Yin, Xinxing Su, Zihao Chen, Jihao Wu, Minghui Liao, et al. Visuriddles: Fine-grained perception is a primary bottleneck for multimodal large language models in abstract visual reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 136213–136234, 2026.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

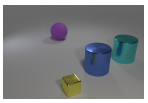
A Experimental and Technical Details

A.1 Prompt Examples

Table 4 presents the prompt examples of each task in both text and visual modalities.

Table 5 reports the separate modality inventories for Qwen2, Pixtral, Gemma3, and Qwen3. The paired populations used in the controlled experiments are given for all four models in Table 6; each retained Vision example has a corresponding Text input. The amount of prompts for the same task and modality varies between models due to tokenization considerations (e.g. all object colors in the color ordering task must be tokenized to a single token for positional alignment purposes—prompts that contain multi-token object colors are filtered out).

Table 4: Representative Text and Vision prompts for the five paired tasks. Each Vision input contains the displayed image followed by its query. Model-specific chat markers are omitted.

Task	Text prompt	Vision prompt
Object Counting	Sequence: apple apple tree apple bird bird apple. How many "apple" are in the sequence? Answer in a single number.	 How many "apple" are in the image? Answer in a single number.
Two-Operand Arithmetic	Question: 37-34. What is the result of the given arithmetic calculation? Answer in a single number.	3 7 - 3 4 = What is the result of the given arithmetic calculation? Answer in a single number.
Color Ordering	In a scene with four objects arranged horizontally, there is a purple object, a yellow object, a blue object and a cyan object. What is the color of the first object from the right? Answer in a single word.	 What is the color of the first object from the right? Answer in a single word.
Factual Recall	Consider Neuschwanstein Castle. Which country is this landmark in? Answer in a single word.	 Which country is this landmark in? Answer in a single word.
Sentiment Analysis	"""A father swings his son in the air on a beach at sunset, both enjoying the golden glow . . . """ Is this scene happy, sad, or neutral? Answer in a single word.	 Is this scene happy, sad, or neutral? Answer in a single word.

A.2 Task Construction Details

Our experiments use the paired task dataset introduced by Nikankin et al. (2025), and this section summarizes the construction procedures for its textual and visual inputs, including task-specific representation choices and filtering rules.

Object Counting. The counting examples were assembled from a pool of 30 object categories, including bananas, forks, and books, with each textual sequence listing seven randomly sampled instances from no more than four categories. The original study selected this combination of sequence length and category diversity to provide a meaningful counting challenge while preserving high model accuracy. Corresponding images were generated with SD3-XL (Podell et al., 2024) using the object quantities specified in the textual sequences. Since generated scenes could deviate

Table 5: **Text and Vision prompt counts for four VLMs.**

Task	Qwen2-VL-7B		Pixtral-12B		Gemma-3-12B		Qwen3-VL-4B	
	L	V	L	V	L	V	L	V
Counting	1524	383	1524	334	1524	383	382	382
Arithmetic	1000	1000	1000	1000	1000	1000	1000	1000
Spatial Ordering	1865	1925	472	497	1865	1925	1925	1925
Factual Recall	416	1265	454	1265	453	1265	1264	1264
Sentiment Analysis	232	245	237	245	222	245	244	244

from the requested quantities, the dataset authors inspected every image and retained those whose object counts agreed with the associated specification.

Two-Operand Arithmetic. Each arithmetic instance contains a pair of two-digit operands, which fixes the locations of corresponding digit tokens across textual prompts. This restriction supports positional alignment under the digit-level tokenization used by all models in the source study, so that, for example, the units digit of the first operand consistently occupies the second position in the expression. The visual counterpart presents the same calculation on a white canvas measuring 75×338 pixels, with large black characters centered in the image. This uniform rendering keeps the visual presentation focused on the expression itself.

Color Ordering. The visual examples originate from CLEVR (Johnson et al., 2017) and are restricted to scenes containing four colored objects. The dataset authors adopted this scene size after observing that three-object examples offered little difficulty, whereas five-object examples substantially reduced model accuracy. Each selected scene was paired with a textual account of its four objects that specified their colors and omitted shape attributes. This representation maintains a common positional structure across prompts, since shape names can span different numbers of tokens and consequently shift the positions of subsequent content.

Sentiment Analysis. The initial collection comprised 120 scene descriptions for each of the labels happy, sad, and neutral, generated with ChatGPT (Hurst et al., 2024). The dataset authors reviewed these descriptions and discarded those that insufficiently expressed their assigned sentiment. For the textual inputs, descriptions containing at least 20 tokens were restricted to their first 20 tokens and followed by an ellipsis, while shorter descriptions were excluded. The source study chose this length because most generated descriptions exceeded the cutoff and the retained content was judged sufficient to communicate the intended sentiment. Flux1-Schnell (Labs, 2024) produced the corresponding images from the scene descriptions, after which a second manual screening removed images with inconsistent scene content or inadequate expression of the target sentiment.

Factual Recall. This task uses three relation-query templates adopted from Hernandez et al. (2024), covering the sport associated with an athlete, the country containing a landmark, and the instrument played by a person, with each query requesting a one-word answer. Textual inputs identify the entity by name through the introductory span `Consider [X]`, where `[X]` is replaced by an entity such as Michael Jordan. Eligible names make this entire span exactly five tokens long, which preserves positional alignment and, according to the source study, accommodates the largest number of entities in the underlying collection.

The visual counterparts were collected from publicly licensed images on Wikimedia Commons. For each entity, ten candidate images were ranked by CLIP (Radford et al., 2021) scores, and the highest-ranked image was selected before resizing and center-cropping to 256×256 pixels. The dataset authors then screened the images for cues that directly disclosed the queried property, such as a basketball player holding a basketball, and removed these examples to focus the task on factual recall. For both textual and visual inputs, counterfactual examples were selected within the same query template, preserving the type of factual relation under comparison.

A.3 Models and intervention coordinates

We evaluate Qwen2-VL-7B-Instruct, Pixtral-12B, Gemma-3-12B-IT, and Qwen3-VL-4B-Instruct with frozen pretrained weights. After zero-based decoder block l, the residual array is written as $R_l = (C_l, S_l)$, where S_l is the final prompt-position vector and C_l contains every preceding position. A state intervention copies one donor S_l into the recipient array, preserves the recipient context, prompt length, positional configuration, and attention mask, and resumes the frozen suffix at block $l + 1$. The models share a fixed hidden width within each architecture, so this one-vector coordinate can move between paired prompts without aligning their prefix positions.

The reader intervention acts at the final prompt position after the relay: it caches Attention or MLP outputs from two continuations that share the native Vision relay vector, exchanges the named output at every block in a fixed window, and retains all other recipient computation. The source intervention refines the Attention operation before its output projection by partitioning the current

Table 6: Relay layers and question-specific denominators for all four models. Layer indices are zero-based, and reader-window endpoints are inclusive. The full paired population is N; N_R, N_C, N_D, and N_B follow the eligibility definitions in this section. An empty $N_B = 0$ leaves the reader window undefined.

Model	Task	H	N	N_R	N_C	N_D	N_B	Reader window
Qwen2	Counting	22	382	184	69	70	12	23–26
Qwen2	Arithmetic	23	1000	245	314	314	92	24–26
Qwen2	Color ord.	26	1925	229	397	397	25	27–27
Qwen2	Factual	23	1264	226	244	244	25	24–26
Qwen2	Sentiment	20	244	215	15	15	11	21–24
Pixtral	Counting	37	333	65	65	67	21	38–39
Pixtral	Arithmetic	38	1000	102	149	149	0	–
Pixtral	Color ord.	33	496	208	63	63	6	34–39
Pixtral	Factual	31	1264	230	561	561	183	32–39
Pixtral	Sentiment	19	244	165	72	72	65	20–39
Gemma	Counting	39	382	208	84	85	36	40–43
Gemma	Arithmetic	45	1000	249	121	121	41	46–47
Gemma	Color ord.	44	1925	236	636	636	224	45–47
Gemma	Factual	42	1264	231	296	296	206	43–46
Gemma	Sentiment	28	244	210	17	17	12	29–32
Qwen3	Counting	28	382	202	65	65	30	29–32
Qwen3	Arithmetic	24	1000	247	100	100	74	25–28
Qwen3	Color ord.	30	1925	187	202	202	16	31–34
Qwen3	Factual	31	1264	204	260	260	47	32–35
Qwen3	Sentiment	24	244	209	20	20	8	25–28

decision-position weighted-value sum into visual positions, earlier nonvisual-prefix positions, and the decision position itself. Replacing one aggregate source contribution leaves the other recipient contributions, output projection, and bias in place, and the Attention weights retain their original normalization after this replacement.

A.4 Populations and fixed layer boundaries

The experiments answer different conditional questions and consequently use different denominators. The full paired population N supports the factorial and out-of-fold correction analyses. The relay population N_R contains native-clean target–donor pairs for which both target and donor are correct in both modalities. The natural-transfer population N_C contains Text-correct/Vision-wrong targets that also admit a semantically matched third-answer control donor. The same-state population N_D contains all eligible Text-correct/Vision-wrong examples and does not impose that donor-feasibility condition. Finally, N_B freezes the subset in which an identical native Vision relay state yields the original error under Vision context and the target under Text context; this exact-recovery cohort defines the reader and source experiments.

The relay H is selected once per model–task setting by the bilateral criterion in Equation 1, and subsequent experiments reuse that boundary. Reader windows start at $H + 1$ and end at the setting-specific late block shown in Table 6. Pixtral Arithmetic has $N_B = 0$, so its reader and source effects are scientifically undefined while the other experiments retain their respective Pixtral Arithmetic populations.

A.5 Estimands and uncertainty

Binary outcomes are reported as percentages, differences between binary means as percentage points, candidate-logit effects in their native logit units, and negative log likelihood on its usual scale. Relay, transfer, same-state, reader, and source intervals use within-setting row bootstrap resampling of their named population; paired contrasts draw the same row indices for both conditions. Taxonomy proportions use Wilson intervals when an interval is shown. The specialization analysis aligns all four oracle–learned quantities on one finite row set and uses 5000 shared resamples per

setting and recomputes all four correlations and their three derived differences within each draw; Spearman resamples recompute average ranks after resampling. The reported 95% percentile intervals are pointwise descriptions of setting-level estimands and do not apply a multiplicity correction.

B Relay Selection Details and Controls

The relay analysis asks when a final-position residual state becomes executable by either modality's suffix and whether that transfer retains answer identity. Each target and its primary donor are native-correct in both modalities, share a task-specific semantic group, and have different answer tokens. Counting matches the queried object and minimizes

$$d_{\text{count}}(i,j) = \sum_{o \neq o_q} |c_i(o) - c_j(o)|\,,$$

where $c_i(o)$ is the multiplicity of each nonqueried object. Arithmetic matches the operator and minimizes

$$d_{\text{arith}}(i,j) = |a_{1,i} - a_{1,j}| + |a_{2,i} - a_{2,j}|.$$

Color Ordering matches the ordinal and queried side, Factual Recall matches the question template, and Sentiment uses its shared classification template. Compatible donors in these last three tasks have zero distance.

Fixed pairs and controls. The upstream `all` pool combines the answer-balanced discovery and evaluation partitions used to construct the native-clean paired inputs. We retain every target for which the required donor and controls can be formed, with the resulting N_R reported in Table 6. Answer differences are checked on the evaluated first-token IDs, and the donor answer must belong to the target's candidate set. A pseudorandom choice among minimum-distance donors uses seed 42 combined with a stable hash of the target key. The same donor may serve several targets.

Each target also receives five native-clean control records from the same semantic group whose answer-token IDs differ from both the target and the primary donor. Selection cycles through answer identities and orders the records within each identity by distance after a seeded shuffle. When distinct eligible records are exhausted, the remaining slots are sampled with replacement from this control pool. Pair and control identities are fixed before the layer scan and remain identical across directions and layers. The scan covers every decoder block, from 0–27 for Qwen2, 0–39 for Pixtral, 0–47 for Gemma, and 0–35 for Qwen3. The first bilateral crossing of the 75% IIA point-estimate threshold fixes H before corrector training. Relay selection precedes parameter fitting and draws on the same underlying task data as the later full-population analyses. Factorial evaluation and cross-fitting use their complete paired populations N.

Control readout and answer specificity. At each layer, the cross-modal condition and the recipient-modality donor control substitute the primary donor's state from the corresponding modality. Each of the five third-answer controls supplies a real state from the cross-modal source. We run its recipient suffix separately and then average the five outcomes with equal weight, so a repeated record contributes once for each occupied slot. All IIA curves count predictions of the *primary donor's* answer token. The dotted control curve therefore measures how often a state carrying a third answer produces that primary answer.

For target answer token y_i and primary donor token y_j, the answer-specificity margin is $m_{ij}(\ell) = \ell_{y_j} - \ell_{y_i}$. Table B.1 reports the mean across targets of

$$m_{ij}(\ell_{\text{cross}}) - \frac{1}{5}\sum_{k=1}^{5} m_{ij}(\ell_{\text{control},k}),$$

with the control average taken after the five separate continuations. This contrast compares the same two answer identities under the same recipient context.

The depth profiles connect the rise in cross-modal transfer to the answer carried by the substituted state. At H, the primary donor produces a larger donor-versus-target margin than the matched third-answer controls in all 40 setting–direction comparisons, with mean contrasts from 3.08 to 33.09

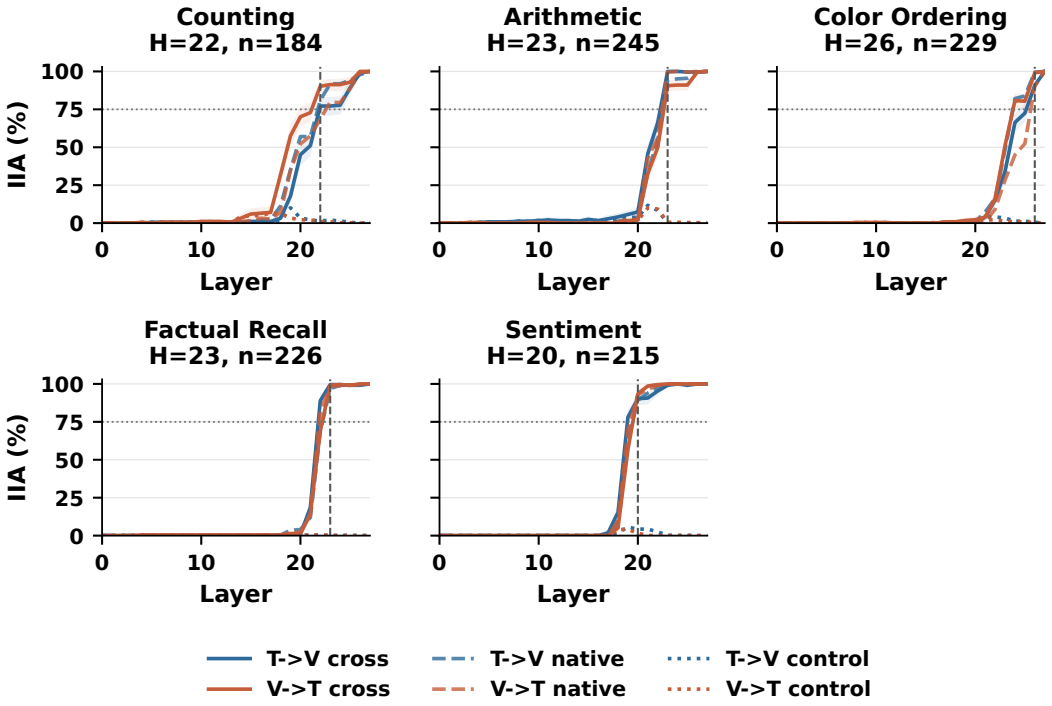


Figure 10: Qwen2 layerwise primary-donor-answer IIA. Blue and orange curves use Vision and Text as recipient, respectively. Solid curves show cross-modal transfer, dashed curves use a donor from the recipient modality, and dotted curves average the five matched third-answer controls. Shading gives the saved 95% bootstrap intervals for cross-modal IIA. Horizontal and vertical lines mark the 75% selection threshold and the frozen relay H; $n = N_R$.

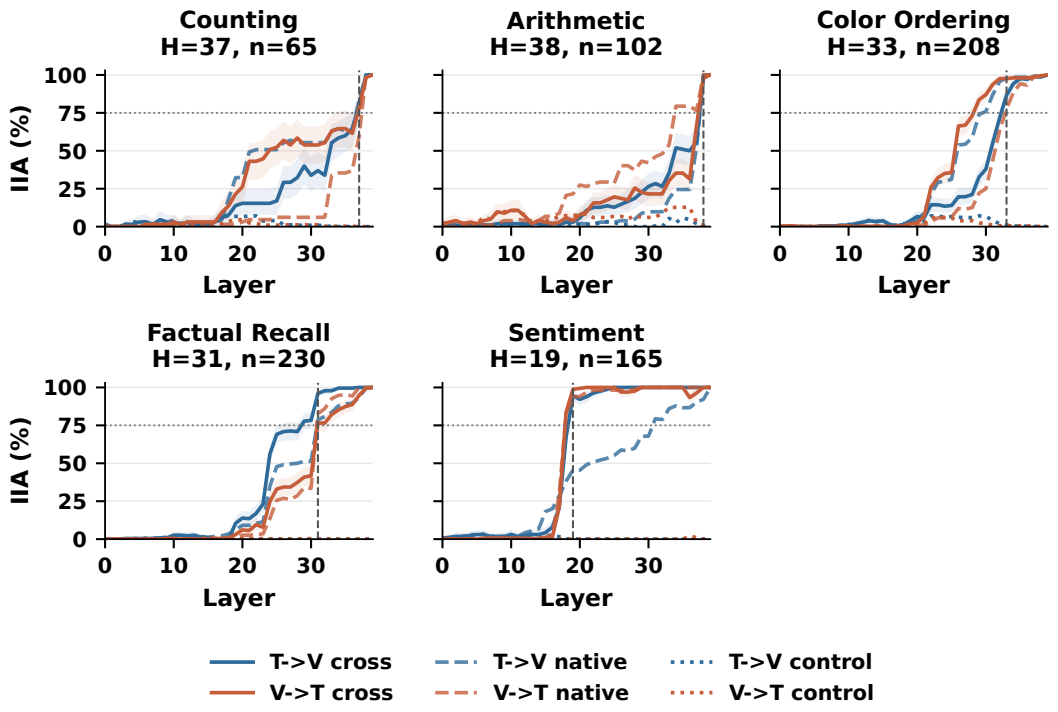


Figure 11: Pixtral layerwise primary-donor-answer IIA. The conditions, colors, and intervals follow Figure 10. Each task uses its own fixed target–donor population and relay H.

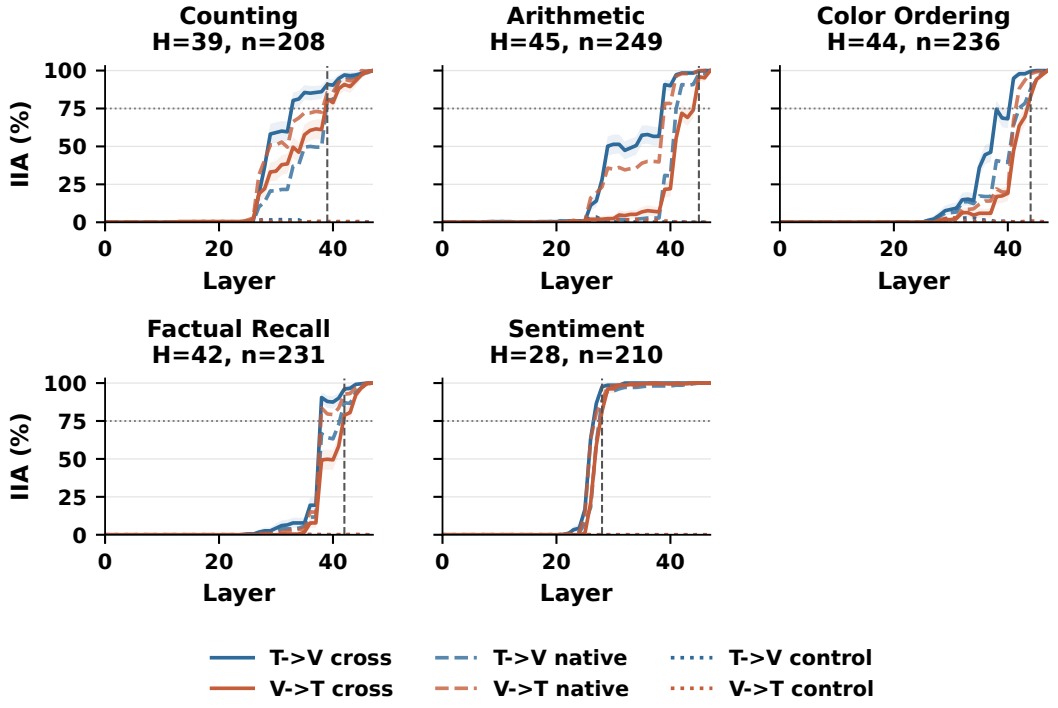


Figure 12: Gemma layerwise primary-donor-answer IIA. The conditions, colors, and intervals follow Figure 10. Each task uses its own fixed target–donor population and relay H.

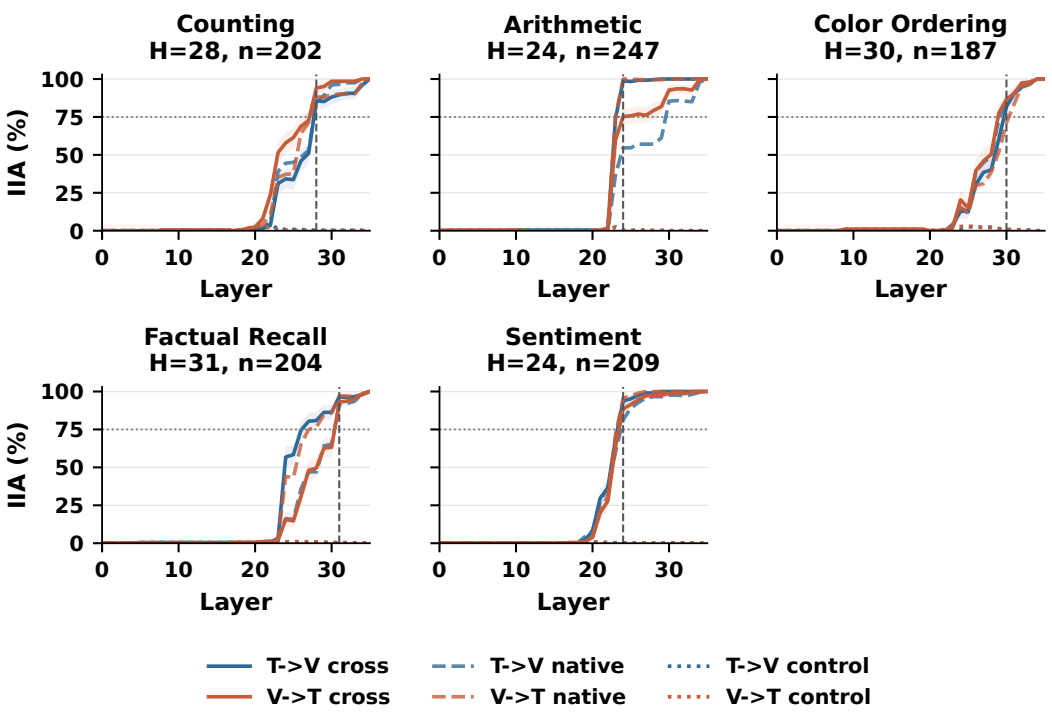


Figure 13: Qwen3 layerwise primary-donor-answer IIA with the same bilateral transfer and matched controls. All 36 measured decoder blocks are retained. Shading denotes the saved 95% bootstrap intervals for the cross-modal curves, and each panel reports its frozen H and N_R.

logit units (Table B.1). Because these comparisons retain the target and recipient context, their separation identifies answer-specific transfer through the relay. The recipient-modality controls and full depth profiles further show how this transfer develops within each architecture. The matched controls hold the target, recipient context, and insertion position fixed while changing the answer carried by the donor state. Their low primary-answer IIA and smaller primary-versus-target margin separate answer transfer from a response to state replacement alone. The two directions also preserve their own depth profiles: Qwen3 Arithmetic reaches 98.8% Text-to-Vision IIA and 75.3% Vision-to-Text IIA at $H = 24$. Their common threshold crossing fixes one usable boundary while the direction-specific values describe how strongly each suffix already follows the donor. Subsequent native-transfer experiments reuse this boundary to test decisions formed for the same instance.

Table B.1: Answer specificity at the frozen relay. Entries are the mean cross-minus-control difference in the primary-donor-minus-target logit margin, followed by the saved 95% paired bootstrap interval. The five control continuations are averaged within each of the N_R target rows before the contrast. Scores retain each model's native logit scale.

Model	Task	N_R	$T \to V$ [95% CI]	$V \to T$ [95% CI]
Qwen2	Counting	184	3.08 [2.72, 3.45]	4.11 [3.76, 4.44]
Qwen2	Arithmetic	245	4.82 [4.58, 5.06]	3.09 [2.90, 3.28]
Qwen2	Color ord.	229	6.23 [5.89, 6.57]	9.34 [8.89, 9.81]
Qwen2	Factual	226	9.58 [9.16, 10.00]	6.75 [6.39, 7.08]
Qwen2	Sentiment	215	5.51 [4.98, 6.05]	4.41 [4.01, 4.81]
Pixtral	Counting	65	7.65 [6.28, 9.01]	6.62 [5.50, 7.73]
Pixtral	Arithmetic	102	3.99 [3.42, 4.60]	4.13 [3.60, 4.66]
Pixtral	Color ord.	208	7.18 [6.79, 7.54]	6.50 [6.12, 6.87]
Pixtral	Factual	230	11.38 [11.01, 11.72]	6.14 [5.80, 6.48]
Pixtral	Sentiment	165	9.03 [8.23, 9.83]	6.32 [6.07, 6.58]
Gemma	Counting	208	14.74 [13.91, 15.56]	14.38 [13.54, 15.17]
Gemma	Arithmetic	249	15.42 [15.01, 15.82]	11.90 [11.44, 12.36]
Gemma	Color ord.	236	21.76 [20.80, 22.69]	16.10 [15.26, 16.97]
Gemma	Factual	231	33.09 [32.01, 34.21]	32.13 [30.86, 33.38]
Gemma	Sentiment	210	19.57 [18.93, 20.21]	22.94 [21.79, 24.10]
Qwen3	Counting	202	16.89 [15.53, 18.27]	18.83 [17.66, 20.02]
Qwen3	Arithmetic	247	19.65 [19.16, 20.11]	10.89 [10.18, 11.61]
Qwen3	Color ord.	187	13.84 [13.16, 14.48]	13.24 [12.57, 13.92]
Qwen3	Factual	204	17.73 [16.57, 18.85]	13.66 [12.76, 14.55]
Qwen3	Sentiment	209	20.37 [19.79, 20.94]	18.62 [17.54, 19.64]

C Additional Native Decision Transfer Results

For a Text-correct/Vision-wrong target, let y be the shared target and $w \neq y$ the native Vision error. The control donor is Text-correct and shares the target's semantic group. Its answer token u belongs to the target's candidate set and differs from both y and w, which defines the control-feasible population N_C. At H, the native Text state enters the Vision context and the native Vision state enters the Text context. Both third-answer controls use the same donor *Text* vector $S_H^T(u)$, inserted into the target's Vision and Text contexts in separate continuations. Every condition uses the same N_C target rows and candidate space.

Figure 14 partitions each final prediction into y, w, u, and *other*, which contains every remaining candidate token. The Text-state/Vision-context intervention recovers y on an equal-setting mean of 94.2% of rows, while the Vision-state/Text-context intervention reproduces the particular error w on 59.4%. The third-answer controls in Figure 15 follow u in 91.0% of Vision continuations and 87.2% of Text continuations. These comparisons connect the relay to naturally formed correct and erroneous decisions and show that a matched state can redirect both recipient contexts toward its own answer identity. CRR and ETR occupy separate intervention bars with their respective outcome partitions.

The control changes the answer identity available at the relay while retaining the recipient instance and its candidate space. Following u therefore tests a specific consequence of the substituted state, and the remaining y, w, and other outcomes show which decisions persist when transfer is incomplete. This comparison complements the native transfer in two ways. Recovery of y establishes that a naturally correct state can guide the visual suffix, whereas reproduction of the exact w establishes that the suffix can also execute the particular error formed by Vision. The conditional population contains conflicting native decisions, so these rates characterize transfer among those conflicts. Together, the two directions and the third-answer controls connect relay executability to the identity of a naturally formed decision.

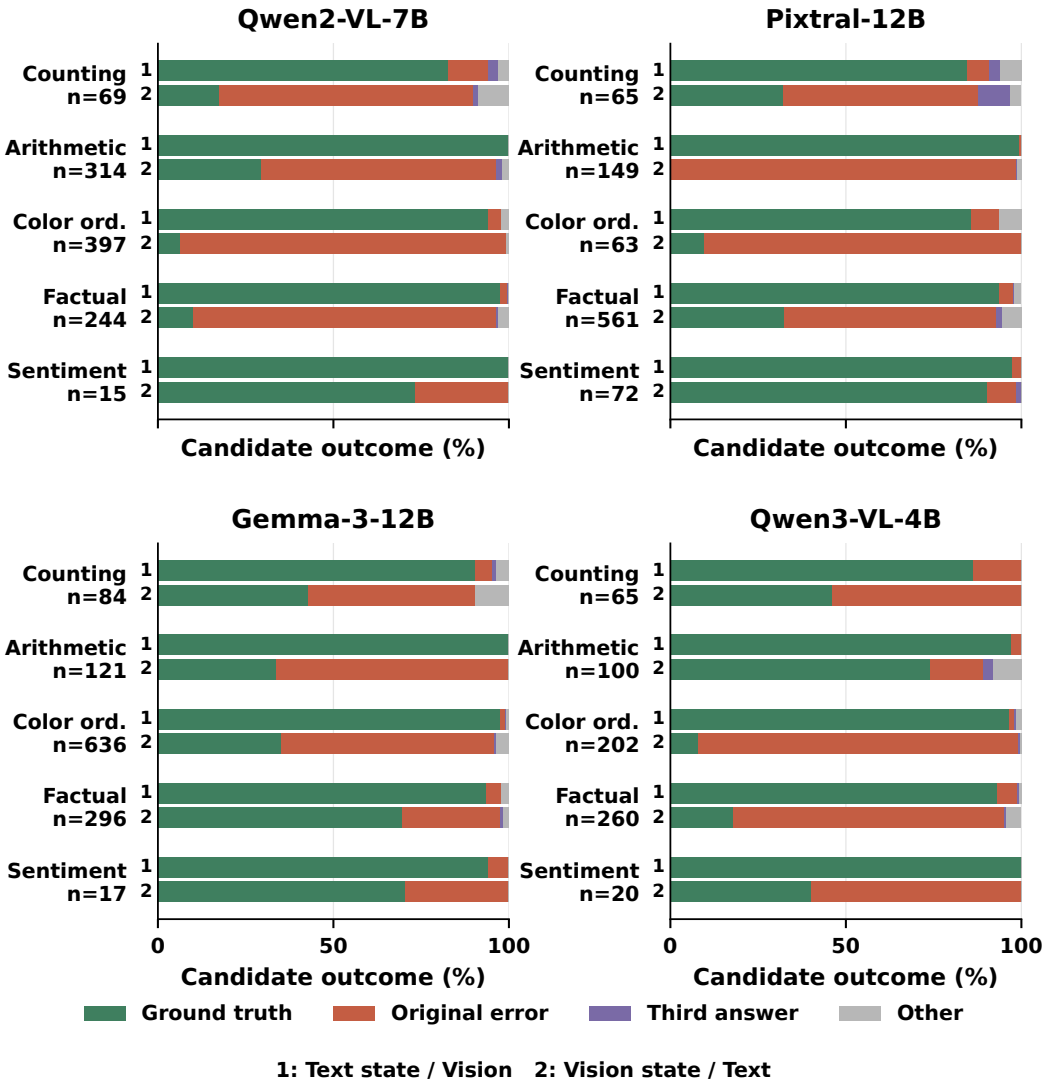


Figure 14: Candidate outcomes for natural correct-state and error-state transfer at the frozen H. For each task, bar 1 inserts the native Text state into Vision and bar 2 inserts the native Vision state into Text. Each bar partitions the same N_C target rows into ground truth y, original error w, control answer u, and other candidates. Each panel retains one model, and every task label gives its own denominator.

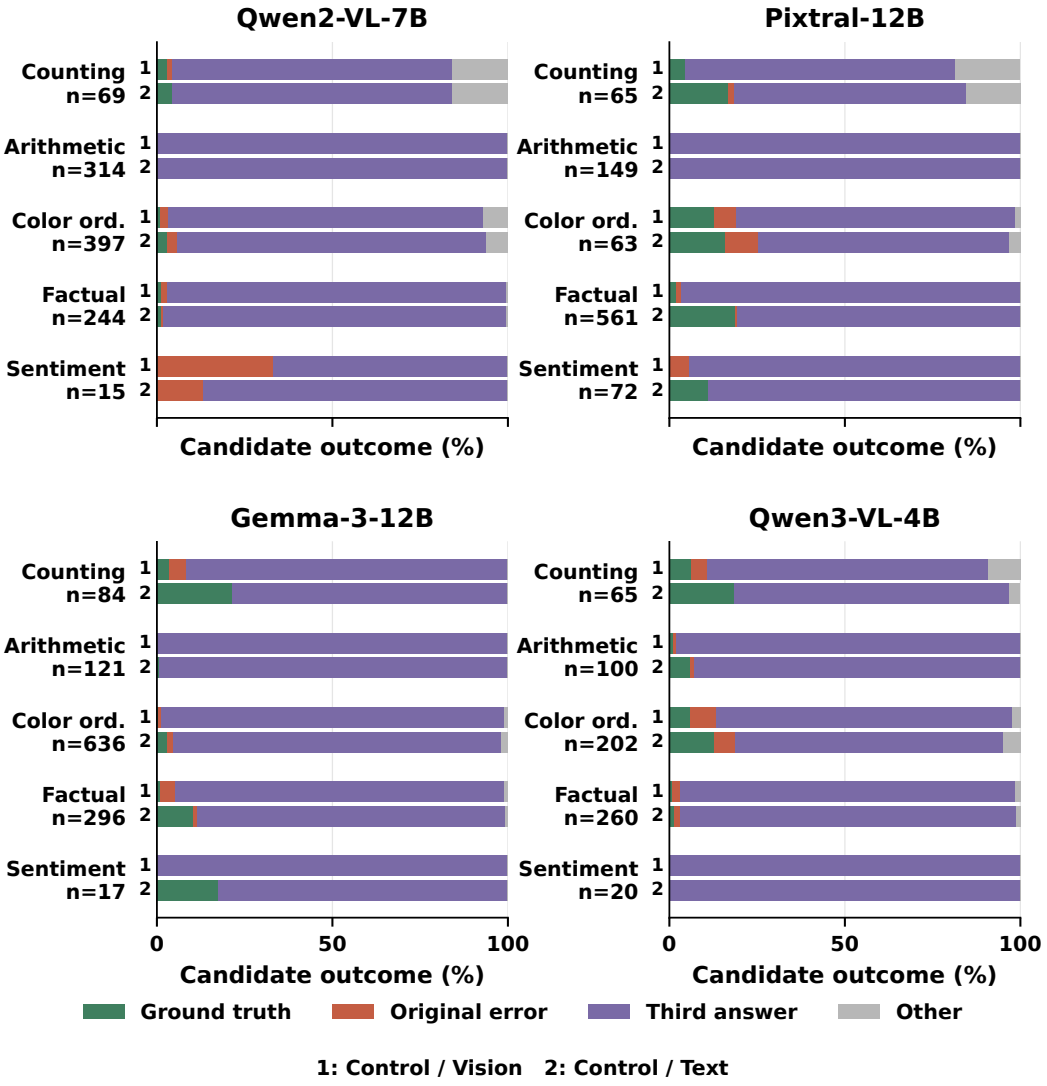


Figure 15: Candidate outcomes for the matched third-answer control. The same donor Text state carrying u enters Vision in bar 1 and Text in bar 2. Colors and populations match Figure 14, so each complete bar sums to 100% within its own intervention.

D ADDITIONAL SAME-STATE RESULTS

The same-state experiment uses all N_D eligible Text-correct/Vision-wrong examples, including those without a feasible third-answer donor. Its primary direction copies the native Vision relay vector S_H^V into both the Vision and Text residual arrays, while the complementary direction copies S_H^T into both arrays. Within each comparison, the two suffixes receive numerically identical decision-position vectors at H, then evolve under their retained contexts, sequence lengths, positional configurations, and masks. Equality is imposed once at this boundary.

The paired bars in Figure 16 separate the two fixed-state comparisons by answer identity. With the Vision state fixed, the baseline prediction is w, so divergence consists of recovery of y and changes to other incorrect candidates. With the Text state fixed, the baseline is y, so divergence consists of transfer to w and changes to the remaining candidates. Text context recovers y from the fixed Vision state on an equal-setting mean of 36.9% of eligible rows, while Vision context retains y from the fixed Text state on 94.2%. The asymmetric outcomes show that the retained context can redirect an executable state, which motivates crossing the two intervention coordinates.

Equality at H fixes the decision-position input to the remaining computation. The comparison then measures how its use changes with the recipient context and the recipient's native continuation conditions. This control gives a direct interpretation to exact recovery: the target answer emerges after the same Vision vector that sustains the original error under Vision context is continued under Text context. For Qwen3 Counting, 30 of 65 rows recover the target, and all 30 decision changes are recoveries. Other settings also contain changes to a different incorrect candidate, which the gray segments retain separately. The complementary Text-state comparison establishes which correctly formed states keep their target answer when the recipient context changes. These conditional outcomes distinguish state transfer from the context-sensitive execution that follows it.

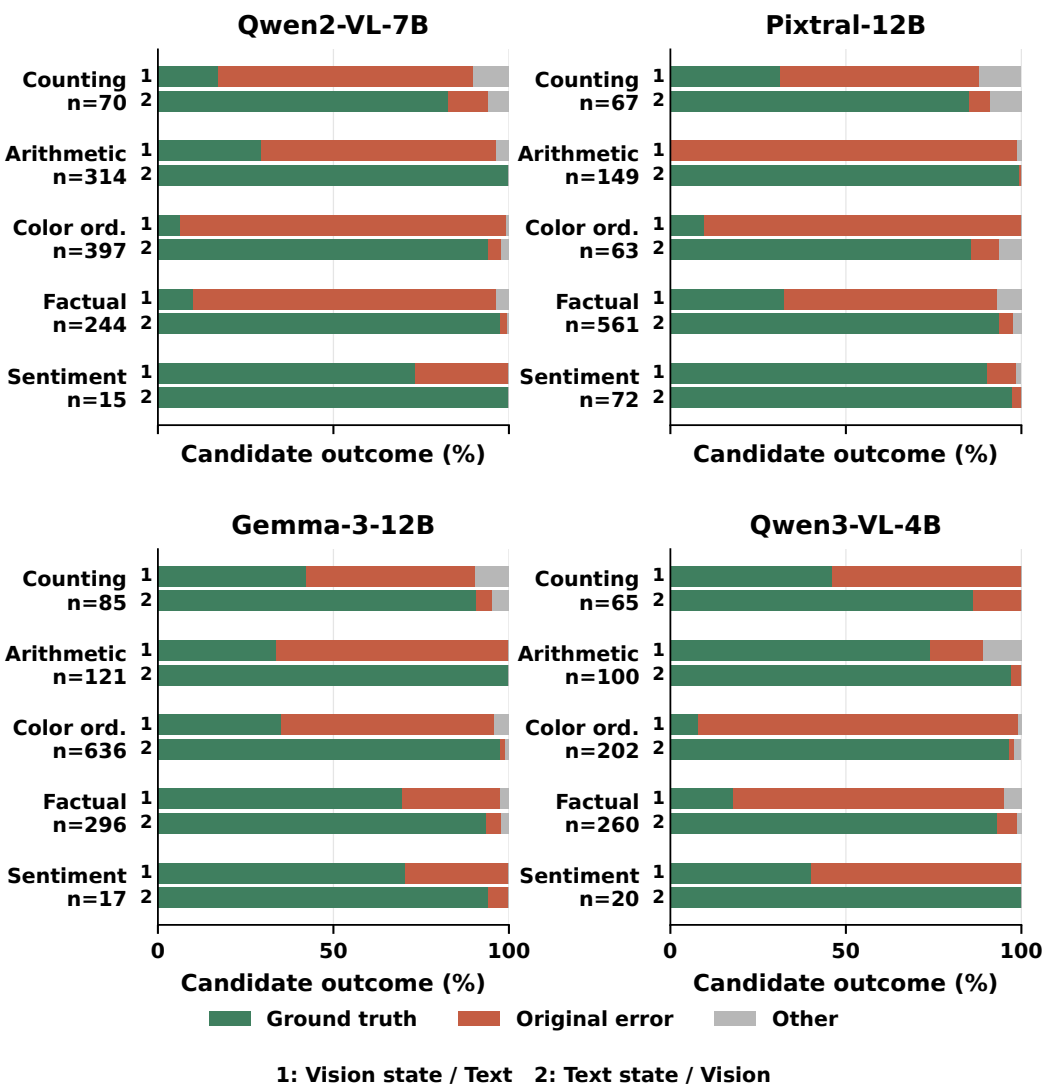


Figure 16: Candidate outcomes under the two fixed-state comparisons at H. Bar 1 uses the fixed Vision state with Text context, and bar 2 uses the fixed Text state with Vision context. Green, orange, and gray segments denote the ground truth, original Vision error, and remaining candidates. Each label reports N_D, and each bar sums to 100%.

E Detailed Formation–Execution Results

The factorial experiment evaluates the complete paired population N under VV, VT, TV, and TT, with context first and state second. The four outcomes are aligned by example identity, and every retained example contributes to all four cells. Conditions VV and TT use the full native Vision and Text forward passes from this factorial evaluation. Conditions VT and TV resume the recipient suffix after replacing the state at H.

For each cell, the ground-truth margin is

$$q(\ell) = \ell_y - \max_{c \neq y} \ell_c,$$

where the strongest competing candidate is recomputed separately for that cell. We apply Equation 2 to candidate correctness and to this margin within each example, then average over the common population. Accuracy contrasts are percentage points and margin contrasts retain logit units. The

Table E.1: Complete four-cell candidate accuracies and their symmetric contrasts for all 20 settings. The first four columns are percentages; F, E, I, G are percentage points. Subscripts give context first and state second. Every row uses the same N examples in all four cells.

Model	Task	N	VV	VT	TV	TT	F	E	I	G
Qwen2	Counting	382	70.94	82.46	76.44	79.32	7.20	1.18	-8.64	8.38
Qwen2	Arithmetic	1000	67.60	98.70	76.40	98.30	26.50	4.20	-9.20	30.70
Qwen2	Color ord.	1925	74.18	87.27	74.70	86.96	12.68	0.10	-0.83	12.78
Qwen2	Factual	1264	68.12	80.46	69.94	78.64	10.52	0.00	-3.64	10.52
Qwen2	Sentiment	244	93.03	97.95	97.54	97.54	2.46	2.05	-4.92	4.51
Pixtral	Counting	333	66.67	52.55	72.37	53.75	-16.37	3.45	-4.50	-12.91
Pixtral	Arithmetic	1000	21.70	25.60	22.00	25.50	3.70	0.10	-0.40	3.80
Pixtral	Color ord.	496	81.05	77.42	81.25	77.42	-3.73	0.10	-0.20	-3.63
Pixtral	Factual	1264	45.02	86.00	59.41	85.13	33.35	6.76	-15.27	40.11
Pixtral	Sentiment	244	70.08	98.36	95.90	98.36	15.37	12.91	-25.82	28.28
Gemma	Counting	382	73.04	90.05	81.41	89.01	12.30	3.66	-9.42	15.97
Gemma	Arithmetic	1000	87.60	99.00	90.80	99.00	9.80	1.60	-3.20	11.40
Gemma	Color ord.	1925	46.39	76.10	56.94	76.10	24.44	5.27	-10.55	29.71
Gemma	Factual	1264	66.53	85.52	81.96	82.67	9.85	6.29	-18.28	16.14
Gemma	Sentiment	244	93.03	98.36	97.95	98.36	2.87	2.46	-4.92	5.33
Qwen3	Counting	382	80.37	94.24	88.48	93.19	9.29	3.53	-9.16	12.83
Qwen3	Arithmetic	1000	88.80	98.10	95.30	97.90	5.95	3.15	-6.70	9.10
Qwen3	Color ord.	1925	79.69	71.95	80.36	69.56	-9.27	-0.86	-3.06	-10.13
Qwen3	Factual	1264	62.26	76.34	65.98	72.55	10.32	-0.04	-7.52	10.28
Qwen3	Sentiment	244	90.98	96.72	93.85	94.67	3.28	0.41	-4.92	3.69

Table E.2: Complete four-cell ground-truth margins and corresponding F, E, I, G. All values are means in native logit units over the aligned N examples. The strongest competing candidate is recomputed for each cell; positive F or E favors the target on average.

Model	Task	N	VV	VT	TV	TT	F	E	I	G
Qwen2	Counting	382	1.66	1.24	1.61	1.01	-0.51	-0.14	-0.18	-0.65
Qwen2	Arithmetic	1000	0.32	2.07	0.95	2.20	1.50	0.38	-0.50	1.88
Qwen2	Color ord.	1925	3.71	2.34	3.58	2.11	-1.42	-0.18	-0.10	-1.60
Qwen2	Factual	1264	1.72	3.57	1.69	3.45	1.81	-0.08	-0.10	1.73
Qwen2	Sentiment	244	3.07	3.67	3.41	4.03	0.61	0.35	0.03	0.97
Pixtral	Counting	333	2.05	-0.03	2.44	0.07	-2.22	0.24	-0.29	-1.98
Pixtral	Arithmetic	1000	-1.90	-1.87	-1.86	-1.83	0.03	0.04	0.00	0.07
Pixtral	Color ord.	496	2.97	3.34	2.82	3.64	0.59	0.07	0.46	0.66
Pixtral	Factual	1264	-0.40	5.73	0.91	6.33	5.77	0.96	-0.71	6.73
Pixtral	Sentiment	244	3.00	7.94	5.75	10.26	4.72	2.54	-0.44	7.26
Gemma	Counting	382	6.23	8.89	8.52	9.18	1.66	1.29	-2.00	2.95
Gemma	Arithmetic	1000	7.76	13.43	9.33	14.42	5.38	1.28	-0.59	6.66
Gemma	Color ord.	1925	-0.48	7.81	2.78	10.17	7.84	2.81	-0.91	10.65
Gemma	Factual	1264	9.59	15.93	14.99	18.56	4.96	4.01	-2.78	8.97
Gemma	Sentiment	244	17.34	18.92	21.58	22.23	1.11	3.77	-0.93	4.89
Qwen3	Counting	382	9.30	10.59	10.25	10.92	0.98	0.64	-0.62	1.62
Qwen3	Arithmetic	1000	8.75	14.64	8.85	14.89	5.96	0.18	0.14	6.14
Qwen3	Color ord.	1925	8.29	6.16	7.70	6.28	-1.77	-0.24	0.71	-2.01
Qwen3	Factual	1264	3.47	6.68	3.85	6.31	2.84	0.00	-0.76	2.84
Qwen3	Sentiment	244	15.96	19.20	17.55	20.03	2.86	1.22	-0.76	4.08

interaction I measures how a coordinate's effect changes with the other coordinate, while $G = F+E$ gives the symmetric accounting of the native gap.

Formation has the larger absolute accuracy effect in all 20 settings, including the three with a negative Formation effect. Execution contributes 12.9 percentage points on Pixtral Sentiment and 6.8 points on Pixtral Factual Recall, so substantial context effects coexist with the larger average state

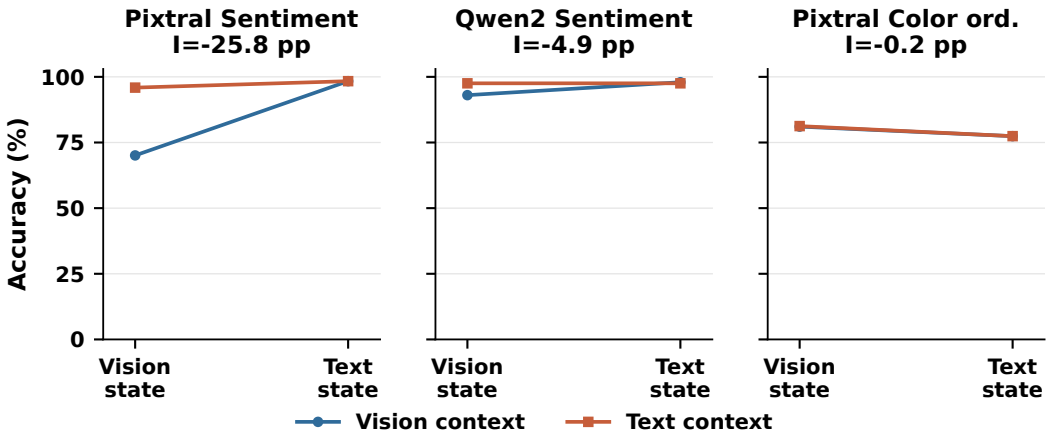


Figure 17: Accuracy interaction structure at the minimum, upper-middle, and maximum I. Settings are sorted by their interaction, and ranks 1, 11, and 20 are displayed. Each line fixes the context while the horizontal axis changes the state, with a common 0–100% accuracy scale.

contribution. The negative interactions, from -25.8 to -0.2 points, mean that each coordinate's accuracy contrast decreases after the other coordinate changes. Figure 17 displays the minimum, upper-middle, and maximum interactions under a fixed descriptive rule. The unequal slopes show how context changes the effect of state substitution within the measured four-cell design.

Tables E.1 and E.2 distinguish these average conditional effects from the single-background gains that define the two teachers. In Qwen3 Arithmetic, replacing the state under Vision context raises accuracy from 88.80% to 98.10%, while changing context with the Vision state raises it to 95.30%. The corresponding gains of 9.30 and 6.50 points become symmetric effects of 5.95 and 3.15 points after averaging over the second background. The negative interaction records the reduced increment once the other coordinate has changed. This distinction connects the descriptive accounting of the native gap to the method: the factorial averages compare both backgrounds, whereas FCRD learns the two specific updates relative to native Vision. Margin results retain changes in the target's separation from its strongest competitor even when candidate correctness stays fixed.

F Repair-Route Definitions and Counts

We condition on factorial rows for which VV is wrong and TT is correct, then classify the two single-coordinate outcomes. Formation-only rows have VT correct and TV wrong, and Execution-only rows have VT wrong and TV correct. Dual-rescuable rows have both hybrids correct, while interaction-dependent rows have both hybrids wrong. These four categories partition the conditional population by the sufficiency of its oracle interventions. Their names describe these oracle outcomes, and the learned F-only and E-only evaluations in Section I measure a separate property of the trained branches.

Formation-only rows account for an equal-setting mean of 58.4% of the conditional population and dual-rescuable rows for 35.8%. State substitution is therefore sufficient for many individual errors, while dual-rescuable cases expose an alternative context-based route to the same target. Interaction-dependent rows account for 4.8% and require both substitutions within this design. Execution-only rows account for 1.0%, which distinguishes context-specific recovery from the larger group that either coordinate can repair. Table 7 retains all category counts, including rare routes.

Execution can supply a useful repair route in both the Execution-only and dual-rescuable groups. The rarity of the former therefore describes how often context substitution is uniquely sufficient; the latter records where context and state each provide a successful alternative. Qwen3 Arithmetic illustrates this overlap: 72 of its 100 conditional rows are dual-rescuable, while 25 are Formation-only, two are Execution-only, and one requires both substitutions. A learner that updates both coordinates can consequently draw on two intervention targets across a substantial shared set of errors. The cate-

Table 7: Counts in the four mutually exclusive repair-route categories, conditional on VV wrong and TT correct. Each row sums to its own denominator n. The VT-only and TV-only columns denote Formation-only and Execution-only oracle routes, respectively.

Model	Task	VT only	TV only	Dual	Interaction	n
Qwen2	Counting	50	4	8	8	70
Qwen2	Arithmetic	222	0	92	0	314
Qwen2	Color ord.	351	2	23	21	397
Qwen2	Factual	213	0	25	6	244
Qwen2	Sentiment	4	0	11	0	15
Pixtral	Counting	36	0	21	10	67
Pixtral	Arithmetic	148	0	0	1	149
Pixtral	Color ord.	48	0	5	9	62
Pixtral	Factual	343	1	182	35	561
Pixtral	Sentiment	7	2	62	0	71
Gemma	Counting	42	1	35	7	85
Gemma	Arithmetic	80	0	41	0	121
Gemma	Color ord.	397	1	223	15	636
Gemma	Factual	75	4	202	15	296
Gemma	Sentiment	4	0	12	1	17
Qwen3	Counting	29	3	27	6	65
Qwen3	Arithmetic	25	2	72	1	100
Qwen3	Color ord.	181	2	14	5	202
Qwen3	Factual	195	0	47	18	260
Qwen3	Sentiment	12	0	8	0	20

gory definition remains fixed by the four oracle outcomes when learned branches are evaluated later. This separation lets the specialization analysis ask whether branch behavior reflects the identified routes on held-out examples.

For Pixtral Color Ordering and Sentiment, the native Vision evaluation associated with the factorial cells classifies one example as correct that was incorrect in the same-state population. The respective taxonomy denominators are consequently 62 and 71, while N_D is 63 and 72. The four-cell taxonomy uses its own displayed denominators throughout.

G Additional Attention–MLP Comparisons

The reader cohort N_B contains the same-state rows for which the native Vision relay vector produces w under Vision context and recovers y under Text context. Both cache-building continuations therefore start with the identical vector S_H^V, including the continuation under Text context. We cache their Attention and MLP outputs at the final prompt position over the window from $H+1$ through the inclusive endpoint in Table 6. This fixed window contains the evaluated blocks between the relay and the setting-specific late boundary, and the same rows and window support every component comparison.

For each intervention, a fresh copy of the recipient residual at H receives the same Vision relay vector. We install the selected component replacements at every block in the window and execute one suffix forward pass from $H+1$, so the later blocks receive the effects of the earlier replacements. Donor outputs remain fixed at their cached values, while all unselected components and positions follow the recipient computation. The simultaneous Attention+MLP condition exchanges both components over this same window and has no learned parameters.

The exchanged Attention output has already combined the heads and passed through the output projection, and the MLP output has passed through its final projection. Both are intercepted before their respective residual additions. For Gemma, the exchanged outputs additionally include the corresponding post-Attention or post-MLP RMS normalization. This placement gives each model's actual residual-stream contribution as the object of exchange, which makes the MLP intervention a component control at the same positions and depths as Attention.

Table G.1: Directional reader effects in logit-margin units on the frozen exact-recovery cohort. Each component uses the same N_B rows, relay vector and reader window. Columns show $\Delta^{T\to V}$ and $-\Delta^{V\to T}$; Attention+MLP denotes simultaneous frozen-output exchange. A dash identifies the empty Pixtral Arithmetic cohort.

Model	Task	N_B	Attention		MLP		Attention+MLP	
			$T\to V$	$-V\to T$	$T\to V$	$-V\to T$	$T\to V$	$-V\to T$
Qwen2	Counting	12	0.76	0.77	0.38	0.44	0.76	0.77
	Arithmetic	92	1.10	1.13	0.17	0.19	1.10	1.13
	Color ord.	25	0.74	0.74	-0.61	-0.65	0.74	0.74
	Factual	25	0.43	0.45	0.19	0.18	0.43	0.45
	Sentiment	11	0.96	0.91	0.19	0.16	0.96	0.91
Pixtral	Counting	21	3.27	3.27	-0.06	-0.03	3.27	3.27
	Arithmetic	0	—	—	—	—	—	—
	Color ord.	6	0.57	0.57	-0.94	-0.76	0.57	0.57
	Factual	183	3.30	3.30	1.70	1.72	3.30	3.30
	Sentiment	65	5.53	5.53	5.27	5.67	5.53	5.53
Gemma	Counting	36	3.87	4.23	1.06	1.11	3.87	4.23
	Arithmetic	41	6.02	6.02	-0.82	-0.03	6.02	6.02
	Color ord.	224	8.98	8.98	-4.78	-3.93	8.98	8.98
	Factual	206	15.02	15.36	-5.66	-6.61	15.02	15.36
	Sentiment	12	5.06	6.53	0.44	0.51	5.06	6.53
Qwen3	Counting	30	1.26	2.33	0.49	0.73	1.26	2.33
	Arithmetic	74	0.64	1.01	0.34	0.44	0.64	1.01
	Color ord.	16	2.49	2.55	-0.18	-0.11	2.49	2.55
	Factual	47	2.53	2.53	0.22	0.23	2.53	2.53
	Sentiment	8	4.11	4.83	1.72	2.18	4.11	4.83

Using the fixed margin $m = \ell_y - \ell_w$, forward and reverse changes are

$$\Delta_X^{T\to V} = m(V \leftarrow T_X) - m(V), \qquad \Delta_X^{V\to T} = m(T \leftarrow V_X) - m(T),$$

and the bidirectional reader score is $B_X = \frac{1}{2}(\Delta_X^{T\to V} - \Delta_X^{V\to T})$. Table G.1 displays the reverse change with its sign flipped, so positive entries in either direction indicate movement toward the donor continuation's decision. A positive forward effect increases support for y in Vision, whereas a positive oriented reverse effect increases support for the original error w in Text. The score averages these two directional effects within each row before averaging over N_B.

The mean Attention effect moves Vision toward y and Text toward w in all 19 eligible settings, while its bidirectional score exceeds the MLP point estimate in each setting. The MLP scores vary in sign, which shows why the matched component control matters: exchanging a downstream output can either reinforce or oppose the donor decision, depending on the component and task. For example, Qwen3 has Attention and MLP scores of 2.527 and 0.222 on Factual Recall, and 2.518 and -0.148 on Color Ordering. These comparisons hold the executable relay state fixed while identifying Attention as a substantial carrier of the context-dependent decision.

Table G.2 retains the paired Attention-minus-MLP contrast together with both component scores and the simultaneous control. The paired 95% interval lies above zero in 18 settings; Pixtral Sentiment has a small positive difference of 0.060 with an interval of $[-0.112, 0.229]$. This distinction between the complete pattern of point estimates and the uncertainty of the component separation keeps the interpretation tied to each setting's exact-recovery population. The two directions also establish that the intervention can transfer the erroneous continuation back into Text, which links the observed recovery to the context's decision role under the same relay state.

Table G.2: Bidirectional reader scores and the paired Attention-minus-MLP contrast, in logit-margin units. Here A denotes Attention, M denotes MLP, and A+M denotes their simultaneous exchange. Each cell gives the mean and its 95% within-setting percentile interval. The contrast is computed within each row and then bootstrapped, so the pairing between component conditions is retained. The intervals use 5000 resamples for Qwen2, Pixtral and Gemma, and 2000 for Qwen3. All columns use the displayed N_B; the simultaneous Attention+MLP control exchanges both outputs over the same window.

Model	Task	N_B	B_{A}	B_{M}	$B_{\mathrm{A+M}}$	$B_{\mathrm{A}} - B_{\mathrm{M}}$
Qwen2	Counting	12	0.76 [0.39, 1.20]	0.41 [0.16, 0.68]	0.76 [0.37, 1.20]	0.36 [0.18, 0.54]
	Arithmetic	92	1.12 [0.95, 1.30]	0.18 [0.14, 0.23]	1.12 [0.95, 1.29]	0.93 [0.77, 1.10]
	Color ord.	25	0.74 [0.62, 0.85]	-0.63 [-0.87, -0.40]	0.74 [0.62, 0.86]	1.37 [1.11, 1.64]
	Factual	25	0.44 [0.26, 0.63]	0.18 [0.07, 0.31]	0.44 [0.26, 0.63]	0.26 [0.14, 0.37]
	Sentiment	11	0.93 [0.65, 1.24]	0.18 [0.06, 0.30]	0.93 [0.66, 1.23]	0.76 [0.56, 1.00]
Pixtral	Counting	21	3.27 [2.46, 4.07]	-0.05 [-0.36, 0.29]	3.27 [2.44, 4.05]	3.31 [2.55, 4.10]
	Arithmetic	0	—	—	—	—
	Color ord.	6	0.57 [0.30, 0.86]	-0.85 [-1.18, -0.56]	0.57 [0.30, 0.86]	1.42 [0.87, 2.01]
	Factual	183	3.30 [3.02, 3.58]	1.71 [1.51, 1.92]	3.30 [3.02, 3.59]	1.58 [1.43, 1.75]
	Sentiment	65	5.53 [5.26, 5.80]	5.47 [5.26, 5.68]	5.53 [5.26, 5.81]	0.06 [-0.11, 0.23]
Gemma	Counting	36	4.05 [3.30, 4.80]	1.08 [0.84, 1.33]	4.05 [3.33, 4.80]	2.96 [2.36, 3.60]
	Arithmetic	41	6.02 [5.18, 6.89]	-0.42 [-0.92, 0.02]	6.02 [5.17, 6.91]	6.44 [5.43, 7.52]
	Color ord.	224	8.98 [8.56, 9.40]	-4.35 [-4.73, -4.00]	8.98 [8.59, 9.40]	13.33 [12.70, 13.97]
	Factual	206	15.19 [14.22, 16.21]	-6.14 [-6.99, -5.33]	15.19 [14.21, 16.19]	21.33 [19.79, 22.86]
	Sentiment	12	5.80 [4.09, 7.58]	0.48 [0.22, 0.75]	5.80 [4.10, 7.63]	5.32 [3.79, 7.00]
Qwen3	Counting	30	1.80 [1.29, 2.33]	0.61 [0.32, 0.90]	1.80 [1.30, 2.32]	1.19 [0.93, 1.49]
	Arithmetic	74	0.82 [0.61, 1.08]	0.39 [0.28, 0.52]	0.82 [0.61, 1.08]	0.43 [0.31, 0.56]
	Color ord.	16	2.52 [1.85, 3.27]	-0.15 [-0.66, 0.31]	2.52 [1.84, 3.24]	2.67 [1.82, 3.76]
	Factual	47	2.53 [2.06, 3.05]	0.22 [-0.01, 0.47]	2.53 [2.08, 3.01]	2.31 [1.89, 2.73]
	Sentiment	8	4.47 [3.08, 5.92]	1.95 [1.34, 2.50]	4.47 [3.15, 6.03]	2.52 [1.73, 3.57]

H Attention Source Intervention Details

At each reader block and head, we partition the decision-position weighted-value sum as in Equation 4. The visual mask selects the specified visual key/value positions, self selects the final prompt position, and the nonvisual-prefix mask selects the remaining earlier positions. The Text visual mask is empty. These disjoint groups retain the original causal and visibility weights, so masked positions contribute zero through the existing Attention pattern.

The zero-based, half-open Vision spans are $[15, 51)$ for Qwen2 Arithmetic and $[15, 96)$ for its other four tasks, $[2, 273)$ for the four eligible Pixtral tasks, $[5, 261)$ for Gemma, and $[4, 68)$ for all five Qwen3 tasks. The right-end image marker in Qwen2 and Pixtral lies outside the visual span and belongs to the nonvisual prefix. System, chat, and query positions outside the image span also belong to that group, except for the final position assigned to self. The source labels consequently

identify the positions from which Attention reads representations, and the same exact-recovery rows and reader window support all three source comparisons.

Aggregate replacement. We cache source contributions from the same two unmodified continuations used for the reader comparison, with S_H^V fixed in both contexts. For head h, decision position q, block l, and selected group g, the intervention is

$$\widetilde{z}^h_{l,q} = z^h_{l,q,\mathrm{rec}} - z^h_{l,q,\mathrm{rec},g} + \bar{z}^h_{l,q,\mathrm{src},g}.$$

Here, the recipient terms come from the current forward pass, including the effects of earlier replacements, and the barred donor contribution is fixed by the unmodified same-state continuation. Each contribution sums the original Attention-weighted values over its source positions. Values are expanded to the query-head layout where grouped-query Attention is used, and the exchanged object is one vector per head. The operation therefore accommodates different prefix lengths through group aggregation.

After replacement, the heads pass through the recipient output projection with its original bias applied once, followed by the model's normal downstream operations. The remaining source contributions retain their live recipient values, and the Attention weights retain their original normalization. The Text-to-Vision visual intervention inserts the zero Text contribution, whereas the reverse intervention inserts the cached Vision visual contribution into the Text continuation. Nonvisual-prefix and self interventions exchange their respective aggregates in both directions under the same rule. These controls distinguish the effects of visual positions, the surrounding prefix, and the decision position within the Attention operation already tested above.

Nonvisual-prefix exchange moves the mean Vision margin toward y and the mean Text margin toward w in all 19 eligible settings, and its bidirectional score exceeds the visual-source score in 18. Pixtral Color Ordering supplies the exception on its six-row cohort, where the visual and nonvisual-prefix scores are 0.569 and 0.161. The separation is large for Gemma Factual Recall, where the nonvisual-prefix and visual scores are 13.807 and 0.895, and for Gemma Arithmetic, where they are 4.880 and 1.252. Qwen3 follows the broader pattern across all five tasks; its Factual Recall scores are 2.453 and 0.139, with a paired difference of 2.314 and an interval of $[1.911, 2.768]$.

The directional results show how this separation arises. On Qwen3 Arithmetic, nonvisual-prefix exchange increases the Vision margin by 0.971 and decreases the Text margin by 1.434, while visual exchange shifts the two means in the opposite directions. On Qwen3 Color Ordering, the nonvisual-prefix score remains larger, with a paired difference of 0.791 and an interval of $[-1.115, 2.862]$; the interval reflects the variability within its 16-row cohort. Self-source scores have absolute values below 0.287 across the 19 settings. Together, these comparisons locate substantial context-dependent readout in representations distributed across earlier nonvisual positions, which supports context as a useful correction coordinate at H. Those positions can already carry information propagated from the image, and the nonlinear suffix allows separately exchanged groups to affect subsequent computation differently even though their pre-projection contributions add linearly.

Table H.1: Bidirectional scores for the three Attention source groups and the paired nonvisual-prefix-minus-visual contrast, in logit-margin units. Subscripts v, p, and s denote visual, nonvisual-prefix, and self positions. Cells give means and 95% within-setting percentile intervals on the same N_B rows. The contrast uses rowwise paired differences; the intervals use 5000 resamples for Qwen2, Pixtral and Gemma and 2000 for Qwen3. The three separate interventions each pass through the nonlinear suffix, so their final scores are compared on a common scale without an additive interpretation.

Model	Task	N_B	B_{v}	B_{p}	B_{s}	$B_{\mathrm{p}} - B_{\mathrm{v}}$
Qwen2	Counting	12	0.36 [0.11, 0.68]	0.43 [0.23, 0.67]	-0.17 [-0.28, -0.07]	0.07 [-0.25, 0.39]
	Arithmetic	92	0.20 [0.15, 0.25]	0.94 [0.78, 1.12]	0.01 [-0.01, 0.02]	0.74 [0.57, 0.92]
	Color ord.	25	0.10 [-0.18, 0.41]	0.62 [0.34, 0.89]	0.01 [0.00, 0.02]	0.51 [-0.05, 1.06]
	Factual	25	0.11 [-0.07, 0.34]	0.34 [0.24, 0.44]	-0.01 [-0.03, 0.01]	0.23 [-0.06, 0.47]
	Sentiment	11	0.05 [-0.04, 0.21]	0.93 [0.69, 1.22]	-0.03 [-0.04, -0.01]	0.87 [0.60, 1.22]
Pixtral	Counting	21	-0.07 [-0.14, 0.01]	3.24 [2.47, 4.01]	0.10 [0.05, 0.15]	3.31 [2.52, 4.10]
	Arithmetic	0	—	—	—	—
	Color ord.	6	0.57 [0.28, 0.90]	0.16 [0.06, 0.28]	-0.09 [-0.16, 0.02]	-0.41 [-0.83, -0.04]
	Factual	183	0.56 [0.47, 0.67]	3.01 [2.79, 3.27]	0.02 [-0.02, 0.06]	2.45 [2.22, 2.68]
	Sentiment	65	0.85 [0.73, 0.97]	5.27 [4.95, 5.58]	0.21 [0.13, 0.28]	4.42 [4.03, 4.82]
Gemma	Counting	36	0.42 [0.20, 0.65]	3.26 [2.70, 3.83]	-0.13 [-0.19, -0.07]	2.84 [2.29, 3.40]
	Arithmetic	41	1.25 [0.99, 1.51]	4.88 [4.00, 5.77]	0.10 [0.05, 0.16]	3.63 [2.66, 4.61]
	Color ord.	224	0.61 [0.48, 0.74]	8.54 [8.11, 8.95]	-0.07 [-0.14, -0.01]	7.92 [7.44, 8.38]
	Factual	206	0.90 [0.74, 1.06]	13.81 [12.93, 14.68]	0.29 [0.18, 0.40]	12.91 [12.10, 13.78]
	Sentiment	12	0.28 [0.08, 0.50]	5.61 [4.07, 7.16]	-0.02 [-0.10, 0.05]	5.33 [4.01, 6.76]
Qwen3	Counting	30	0.42 [0.17, 0.69]	1.46 [1.04, 1.86]	-0.04 [-0.08, 0.01]	1.04 [0.68, 1.42]
	Arithmetic	74	-0.15 [-0.33, 0.06]	1.20 [0.95, 1.46]	0.03 [0.01, 0.05]	1.35 [0.97, 1.72]
	Color ord.	16	1.01 [0.03, 2.06]	1.80 [0.75, 2.93]	-0.11 [-0.21, -0.02]	0.79 [-1.11, 2.86]
	Factual	47	0.14 [0.04, 0.26]	2.45 [2.03, 2.94]	-0.20 [-0.32, -0.10]	2.31 [1.91, 2.77]
	Sentiment	8	-0.04 [-0.28, 0.22]	4.58 [3.19, 6.24]	0.00 [-0.14, 0.13]	4.63 [3.24, 6.29]

Table H.2: Directional source effects on the frozen exact-recovery cohort. For each source group, columns show $\Delta^{T\to V}$ and $-\Delta^{V\to T}$ in logit-margin units, with the same orientation as Table G.1. Positive entries move the recipient toward the donor continuation's decision. Visual exchange removes the Vision visual contribution in the forward direction and inserts it into Text in the reverse direction. A dash identifies the empty Pixtral Arithmetic cohort.

Model	Task	N_B	Visual		Nonvisual prefix		Self	
			$T \to V$	$-V \to T$	$T \to V$	$-V \to T$	$T \to V$	$-V \to T$
Qwen2	Counting	12	0.33	0.38	0.41	0.44	-0.14	-0.19
	Arithmetic	92	0.18	0.21	0.99	0.88	0.03	-0.02
	Color ord.	25	-0.11	0.32	0.41	0.83	0.02	0.01
	Factual	25	0.09	0.13	0.35	0.34	0.00	-0.02
	Sentiment	11	0.03	0.08	0.93	0.93	-0.04	-0.02
Pixtral	Counting	21	-0.13	-0.01	3.11	3.36	0.02	0.18
	Arithmetic	0	—	—	—	—	—	—
	Color ord.	6	0.63	0.51	0.23	0.09	-0.07	-0.12
	Factual	183	0.42	0.70	3.18	2.85	0.09	-0.04
	Sentiment	65	0.19	1.51	5.14	5.40	0.42	-0.01
Gemma	Counting	36	-0.03	0.87	3.07	3.45	-0.11	-0.15
	Arithmetic	41	1.07	1.44	4.83	4.93	0.16	0.04
	Color ord.	224	-0.17	1.39	7.86	9.21	-0.03	-0.12
	Factual	206	0.15	1.64	13.30	14.32	0.21	0.36
	Sentiment	12	-0.17	0.72	4.43	6.78	-0.11	0.06
Qwen3	Counting	30	0.02	0.82	1.17	1.75	0.02	-0.09
	Arithmetic	74	-0.18	-0.11	0.97	1.43	0.01	0.05
	Color ord.	16	0.45	1.57	1.71	1.90	-0.09	-0.14
	Factual	47	0.08	0.20	2.44	2.47	-0.20	-0.20
	Sentiment	8	-0.34	0.25	4.13	5.04	0.03	-0.03

I FCRD TRAINING AND EVALUATION DETAILS

The factorized corrector modifies native Vision activations at the same relay boundary used by the interventions. Its Formation branch updates S_H, and its Execution branch updates the surrounding C_H. E-only conditions on native S_H, while FCRD first forms $S'_H = S_H + \delta S_H$ and conditions the Execution update on S'_H. Both updates occur at H, and the frozen decoder resumes at $H + 1$. This single context update supplies the learned counterpart to the downstream context sensitivity measured by the reader interventions.

Text-derived factorial logits provide the Formation and Execution effect targets. Target labels define the lower-NLL teacher gates and also enter the FCRD task cross-entropy. When a gate is zero, its matching target is the zero effect, so the example remains in the matching loss. F-only, E-only, and FCRD activate branches of one jointly trained corrector, which lets their evaluation measure branch sufficiency and complementarity under a common parameter fit.

Table I.1 gives the model-specific dimensions of the two low-rank branches. The Formation branch contains three bias-free matrices and has $3dr$ parameters, while Execution contains four and has $4dr$ parameters, giving $7dr$ trainable parameters in total. All four models use $r = 8$, so the hidden width determines the parameter count. These small branch modules supply the updates at the fixed relay while the pretrained VLM parameters remain frozen.

Loss evaluation and execution. For Qwen2 Counting, Arithmetic, Color Ordering, and Factual Recall, the teacher gates and task loss use negative log-likelihood over the task's candidate-token subset. Logits are centered over those candidates without grouping equivalent tokens. Each effect MSE averages over candidates and then over examples, with microbatch losses weighted by their example counts. The two effect losses have $\lambda_{\text{eff}} = 1$, and the relative-update penalty has $\lambda_R = 10^{-3}$ and $\varepsilon_R = 10^{-6}$. In Equation 9, the state mean square averages over d entries, while the context-update mean square averages over all Ld entries of the full residual array, including the zero decision row. Each ratio uses the corresponding detached native denominator before averaging across examples.

For these four settings, native residuals, native replay logits, and factorial teacher targets are training constants. Each of the F-only, E-only, and FCRD forwards starts from a fresh native residual copy. In the FCRD forward, Execution receives the corrected state S'_H and the native context as its inputs. Gradients pass through the frozen suffix to the corrector, and the branch effect losses and FCRD task loss jointly update its shared parameters. The evaluated corrector is the six-epoch endpoint. A closed teacher gate retains zero-target effect matching, whereas setting $\lambda_{\text{eff}} = 0$ would remove that supervision term.

Cross-fitting population. The five-fold assignment uses the paired example as its unit, keeping its Text and Vision forms together. For Qwen2 Counting, Arithmetic, Color Ordering, and Factual Recall, the assignment stratifies by answer token and native Vision correctness, shuffles within each stratum, and allocates examples cyclically with a continuing fold offset across strata. Each setting has one fixed example-to-fold mapping, and the five disjoint held-out partitions cover the complete population N in Table 6. Pooling their predictions gives one out-of-fold result per example. Relay selection uses the upstream population described in Section B, and H remains fixed while corrector parameters are cross-fitted.

Public-benchmark protocol. Table 3 reports visual accuracy for the original VLM and FCRD on the evaluated IsoBench and TabMWP subsets, separately for each of the four models. Its parenthetical gains are differences in percentage points. The five-fold partition and controlled-task settings above pertain to the five paired tasks, while the public-benchmark table reports a separate evaluation.

Repair precision and correction behavior. For each setting, let e and c denote the numbers of native Vision errors and correct predictions, and let r and d count errors corrected by FCRD and correct predictions made incorrect, respectively, with $N = e + c$. We define

$$\text{Rescue Rate} = \frac{r}{e}, \qquad \text{Damage Rate} = \frac{d}{c}, \qquad \text{Repair Precision} = \frac{r}{r + d}. \tag{11}$$

Table 8: Corrector architecture and controlled-task training settings. Rank, loss weights, learning rate, weight decay, epoch count, and batch sizes apply to all 20 settings. Detailed architecture operations, initialization, and optimizer implementation are specified for Qwen2 Counting, Arithmetic, Color Ordering, and Factual Recall. Dimensions use the model hidden width d and rank $r = 8$.

Item	Specification
Formation branch	Bias-free $d \to r$ maps for state and local depth change, followed by GELU and a bias-free $r \to d$ map.
Execution branch	Bias-free state and context $d \to r$ maps, elementwise latent interaction, and two bias-free $r \to d$ maps assigned to prefix-position groups.
Normalization and initialization	Layer normalization has no affine parameters; down maps use zero-mean normal weights with standard deviation $1/\sqrt{d}$, and up maps start at zero.
Objective weights	Formation effect matching 1; Execution effect matching 1; FCRD task cross-entropy 1; relative-update regularizer 10^{-3}.
Training budget	Learning rate 0.002, weight decay 10^{-4}, six epochs, effective batch 8 accumulated from physical microbatches of size 1, evaluation batch 1, and float32 computation.
Optimizer implementation	AdamW with a constant learning rate for the four Qwen2 settings named in the caption.
Cross-fitting	Five fixed folds. Each run fits four folds and evaluates the remaining fold, after which the held-out rows are concatenated.
Frozen quantities	Pretrained VLM weights, the selected relay H, the candidate set, and the intervention-derived training targets.

Table I.1: Trainable corrector parameters for the four controlled-task models. The hidden width is d, the common rank is $r = 8$, and the counts sum the Formation and Execution matrices. The same architecture sizes apply across the five tasks for each model.

Model	d	r	Formation	Execution	Total
Qwen2	3584	8	86,016	114,688	200,704
Pixtral	5120	8	122,880	163,840	286,720
Gemma	3840	8	92,160	122,880	215,040
Qwen3	2560	8	61,440	81,920	143,360

Rescue rate measures the coverage of native errors, damage rate measures the disruption of native correct predictions, and repair precision measures the fraction of correctness-changing cases that improve the answer. All three rates use pooled out-of-fold predictions within a setting. Changed predictions include every difference between the native and corrected answer tokens, including transitions between two incorrect candidates; repair precision uses only the $r+d$ cases whose correctness changes. Macro averages assign equal weight to the 20 model–task settings.

Table I.2 reports the complete native and learned comparison. F-only and E-only activate individual branches of the same fitted corrector, and the stronger-branch contrast selects the larger pooled branch accuracy once within each setting. This comparison tests whether composing the updates adds predictive value to either branch's separate action under a common parameter fit. The full model exceeds this stronger branch in 17 settings, ties it on Pixtral Color Ordering, and has differences of -0.41 pp on Qwen2 Sentiment and -0.05 pp on Qwen3 Color Ordering. Gemma Counting illustrates the complementarity: the two branches attain 78.80% and 76.96% accuracy, and their composition attains 82.46%. The 3.66-point improvement over the stronger branch connects the effect-matching objectives to useful interaction between state formation and contextual execution during visual inference.

The net accuracy gain in percentage points is

$$\Delta A_{\mathrm{FCRD}} = \frac{100(r-d)}{N}.$$

Table 9 separates rescues, damages, and changes between incorrect answers, while Table I.3 reports their population and conditional rates. Across the 20 settings, repairs contribute an average 8.63 pp and damages subtract 2.30 pp, yielding a 6.33-point mean net gain. Rescues outnumber damages in 19 settings; Qwen2 Color Ordering has 31 of each and retains its original aggregate accuracy. The

Table I.2: Pooled out-of-fold accuracy (%) on all paired examples in each setting. The native Text and Vision conditions use the same examples. F-only, E-only, and FCRD evaluate one jointly fitted corrector with the indicated branches active. $\Delta_{\text{best}} = A_{\text{FCRD}} - \max(A_F, A_E)$ is in percentage points and selects the stronger branch at the setting level. The final row assigns equal weight to all 20 settings.

Model	Task	N	Vision	Text	F-only	E-only	FCRD	Δ_{best}
Qwen2	Counting	382	70.94	79.32	71.73	74.35	76.18	1.83
Qwen2	Arithmetic	1000	67.60	98.30	77.00	70.10	80.10	3.10
Qwen2	Color ord.	1925	74.18	86.96	74.08	73.97	74.18	0.10
Qwen2	Factual	1264	68.12	78.64	70.97	68.20	71.04	0.08
Qwen2	Sentiment	244	93.03	97.54	97.13	95.08	96.72	-0.41
Pixtral	Counting	333	66.67	53.75	74.17	69.07	74.77	0.60
Pixtral	Arithmetic	1000	21.70	25.50	28.20	21.70	29.80	1.60
Pixtral	Color ord.	496	81.05	77.42	81.45	81.05	81.45	0.00
Pixtral	Factual	1264	45.02	85.13	52.29	45.89	53.48	1.19
Pixtral	Sentiment	244	70.08	98.36	94.67	89.75	95.08	0.41
Gemma	Counting	382	73.04	89.01	78.80	76.96	82.46	3.66
Gemma	Arithmetic	1000	87.60	99.00	95.60	90.60	95.70	0.10
Gemma	Color ord.	1925	46.39	76.10	51.38	49.71	52.31	0.94
Gemma	Factual	1264	66.53	82.67	69.22	69.15	70.57	1.34
Gemma	Sentiment	244	93.03	98.36	95.08	95.90	97.54	1.64
Qwen3	Counting	382	80.37	93.19	84.82	84.82	85.60	0.79
Qwen3	Arithmetic	1000	88.80	97.90	94.60	89.70	95.40	0.80
Qwen3	Color ord.	1925	79.69	69.56	79.90	80.00	79.95	-0.05
Qwen3	Factual	1264	62.26	72.55	63.92	62.34	65.43	1.50
Qwen3	Sentiment	244	90.98	94.67	95.08	91.80	95.90	0.82
Macro average			71.35	82.70	76.50	74.01	77.68	1.00

mean changed-prediction rate is 15.10%, which includes changes among incorrect candidates and therefore measures a broader behavioral response to the updates than the accuracy gain does.

The macro-average repair precision is 76.46%, alongside a 38.37% rescue rate and a 3.72% damage rate. Precision and coverage describe complementary aspects of the correction. Pixtral and Gemma Arithmetic both reach 94.51% repair precision, while their rescue rates are 10.98% and 69.35%, respectively, because their native error populations differ. Qwen3 Arithmetic combines 71 rescues with five damages, producing 93.42% precision and a 6.60-point gain. These counts show how the learned update can recover a substantial share of visual errors while retaining originally correct predictions. They also explain why a high repair precision alone does not establish broad recovery: its denominator contains only correctness-changing cases, whereas rescue rate measures coverage of the full native-error population.

Figure 18 retains the five actual held-out fold gains for every setting. The diamonds mark pooled setting gains, which weight each partition by its held-out example count. Individual points show how correction behaves across the disjoint partitions used for cross-fitting; all 100 points remain visible, including folds with zero or negative gain.

For the 17 settings with native Text accuracy above native Vision accuracy, gap closure is computed as $100(A_J - A_V)/(A_T - A_V)$, where A_J is FCRD accuracy, and then averaged equally. Its mean is 63.09%. The ratio retains values above 100% when FCRD exceeds Text, as on Qwen3 Sentiment, where a 4.92-point gain closes 133.33% of the positive native gap. Pixtral Counting, Pixtral Color Ordering, and Qwen3 Color Ordering have nonpositive native gaps and retain their correction results with undefined GC. This conditional denominator allows the method's error recovery to be assessed alongside the amount of textual performance advantage available in each setting.

Table 9: Out-of-fold correction counts for FCRD. $N = e + c$ is the complete population, e and c count native errors and correct predictions, and r and d count rescues and damages. Changed counts every altered predicted token; its final component is a change between two incorrect answers. Thus Changed equals $r + d$ plus the final column.

Model	Task	N	e	c	r	d	Changed	Wrong → wrong
Qwen2	Counting	382	111	271	30	10	41	1
Qwen2	Arithmetic	1000	324	676	172	47	297	78
Qwen2	Color ord.	1925	497	1428	31	31	75	13
Qwen2	Factual	1264	403	861	73	36	170	61
Qwen2	Sentiment	244	17	227	13	4	17	0
Pixtral	Counting	333	111	222	36	9	55	10
Pixtral	Arithmetic	1000	783	217	86	5	280	189
Pixtral	Color ord.	496	94	402	10	8	20	2
Pixtral	Factual	1264	695	569	166	59	400	175
Pixtral	Sentiment	244	73	171	62	1	66	3
Gemma	Counting	382	103	279	50	14	83	19
Gemma	Arithmetic	1000	124	876	86	5	95	4
Gemma	Color ord.	1925	1032	893	243	129	544	172
Gemma	Factual	1264	423	841	90	39	212	83
Gemma	Sentiment	244	17	227	12	1	13	0
Qwen3	Counting	382	75	307	27	7	41	7
Qwen3	Arithmetic	1000	112	888	71	5	88	12
Qwen3	Color ord.	1925	391	1534	39	34	84	11
Qwen3	Factual	1264	477	787	82	42	226	102
Qwen3	Sentiment	244	22	222	14	2	16	0

J Parameter Efficiency of FCRD

To further present the parameter efficiency of FCRD, we report the original VLM size and the trainable parameters of both correction branches in this section. As shown in Table 10, FCRD introduces only 143360 to 286720 trainable parameters, which account for 0.0018% to 0.0032% of the complete model. This extremely lightweight design supports compact corrector storage and reuse of the frozen VLM across tasks. Together with the accuracy improvements in Tables 2 and 3, these results demonstrate that targeted state and context correction delivers substantial performance gains with negligible parameter overhead.

Table I.3: Conditional correction behavior and accuracy contributions. Rescue, damage, repair precision, Changed, and gap closure (GC) are percentages. The repair contribution $100r/N$, damage magnitude $100d/N$, and net gain are percentage points; the damage magnitude is subtracted from the repair contribution. GC is defined only for the 17 settings with positive native Text–Vision gaps. The macro row gives equal-setting means, with GC averaged over those 17 valid settings.

Model	Task	Rescue	Damage	Precision	Changed	$100r/N$	$100d/N$	Gain	GC
Qwen2	Counting	27.03	3.69	75.00	10.73	7.85	2.62	5.24	62.50
Qwen2	Arithmetic	53.09	6.95	78.54	29.70	17.20	4.70	12.50	40.72
Qwen2	Color ord.	6.24	2.17	50.00	3.90	1.61	1.61	0.00	0.00
Qwen2	Factual	18.11	4.18	66.97	13.45	5.78	2.85	2.93	27.82
Qwen2	Sentiment	76.47	1.76	76.47	6.97	5.33	1.64	3.69	81.82
Pixtral	Counting	32.43	4.05	80.00	16.52	10.81	2.70	8.11	–
Pixtral	Arithmetic	10.98	2.30	94.51	28.00	8.60	0.50	8.10	213.16
Pixtral	Color ord.	10.64	1.99	55.56	4.03	2.02	1.61	0.40	–
Pixtral	Factual	23.88	10.37	73.78	31.65	13.13	4.67	8.47	21.10
Pixtral	Sentiment	84.93	0.58	98.41	27.05	25.41	0.41	25.00	88.41
Gemma	Counting	48.54	5.02	78.12	21.73	13.09	3.66	9.42	59.02
Gemma	Arithmetic	69.35	0.57	94.51	9.50	8.60	0.50	8.10	71.05
Gemma	Color ord.	23.55	14.45	65.32	28.26	12.62	6.70	5.92	19.93
Gemma	Factual	21.28	4.64	69.77	16.77	7.12	3.09	4.03	25.00
Gemma	Sentiment	70.59	0.44	92.31	5.33	4.92	0.41	4.51	84.62
Qwen3	Counting	36.00	2.28	79.41	10.73	7.07	1.83	5.24	40.82
Qwen3	Arithmetic	63.39	0.56	93.42	8.80	7.10	0.50	6.60	72.53
Qwen3	Color ord.	9.97	2.22	53.42	4.36	2.03	1.77	0.26	–
Qwen3	Factual	17.19	5.34	66.13	17.88	6.49	3.32	3.16	30.77
Qwen3	Sentiment	63.64	0.90	87.50	6.56	5.74	0.82	4.92	133.33
Macro average		38.37	3.72	76.46	15.10	8.63	2.30	6.33	63.09

Table 10: Parameter counts of the original VLM and the two trainable FCRD branches. The final column reports $100P_{\mathrm{FCRD}}/(P_{\mathrm{VLM}} + P_{\mathrm{FCRD}})$. Counts are identical across the five tasks for each VLM.

VLM	**VLM parameters**	**Formation**	**Execution**	**FCRD**	**Trainable ratio (%)**
Qwen2-VL-7B	8291375616	86016	114688	200704	$\mathbf{0.0024}\%$
Pixtral-12B	12682739712	122880	163840	286720	$\mathbf{0.0023}\%$
Gemma-3-12B	12187325040	92160	122880	215040	$\mathbf{0.0018}\%$
Qwen3-VL-4B	4437815808	61440	81920	143360	$\mathbf{0.0032}\%$

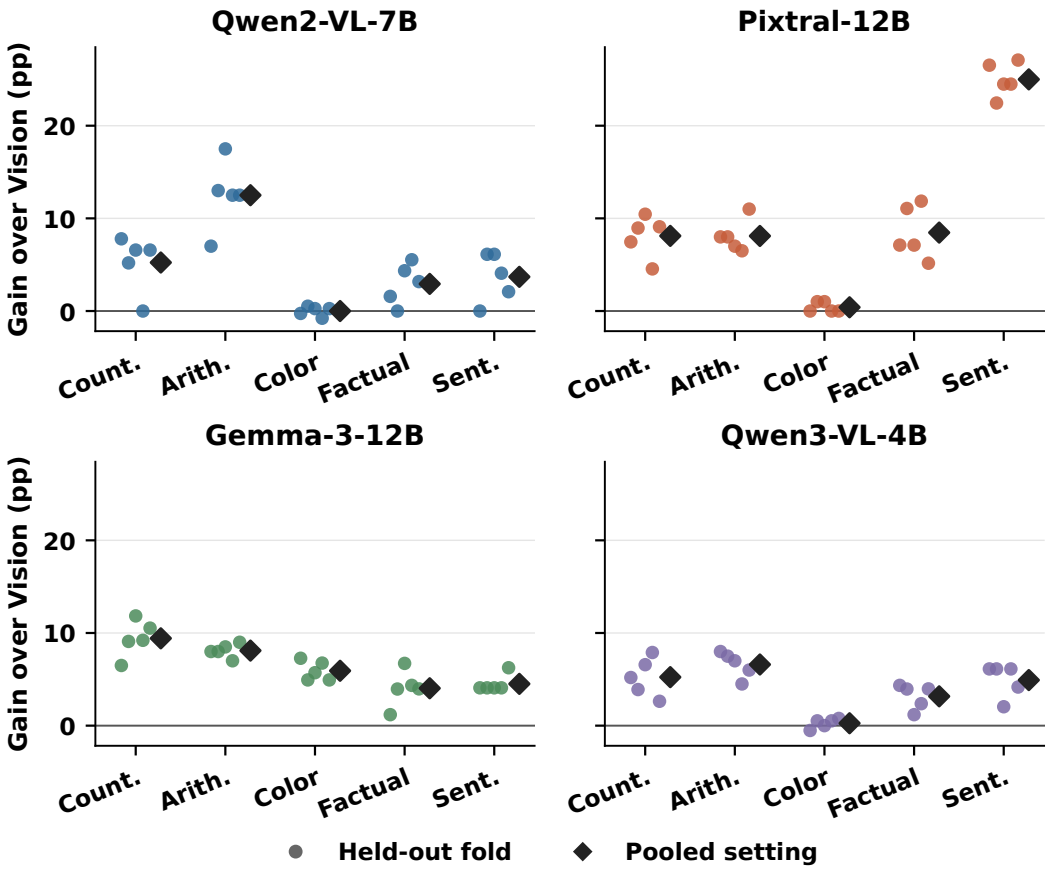


Figure 18: The five held-out fold gains of FCRD over native Vision in each of the 20 settings. Colored circles represent the actual held-out partitions, with small horizontal offsets to separate their positions; black diamonds show the corresponding pooled setting gains. All panels use the same percentage-point scale. Their spread describes differences among the fixed held-out partitions used for cross-fitting.

K Additional Factor Specialization Results

For every out-of-fold row, let $q(\ell) = \ell_y - \max_{c \neq y} \ell_c$, recomputing the strongest competing candidate separately for each logit vector. The native-Vision-background oracle gains and learned branch gains are

$$O_F = q(\ell_{VT}) - q(\ell_{VV}), \quad O_E = q(\ell_{TV}) - q(\ell_{VV}),$$
$$S_F = q(z_F) - q(\ell_{VV}), \quad S_E = q(z_E) - q(\ell_{VV}).$$

These O_F and O_E are ungated row-level hybrid gains for this diagnostic. With ρ computed within one model–task setting, the four aligned correlations are $FF = \rho(O_F, S_F)$, $FE = \rho(O_F, S_E)$, $EE = \rho(O_E, S_E)$, and $EF = \rho(O_E, S_F)$; their paired contrasts are $D_F = FF - FE$, $D_E = EE - EF$, and $S = (D_F + D_E)/2$.

We align the four gains by paired-example identity and held-out fold, and retain a row only when all four gains are finite. Every out-of-fold row satisfies this condition in all 20 settings, and all four gain vectors have nonzero variance. The correlation populations therefore equal the complete N values in Table K.1. Correlations and their contrasts are computed within each setting before taking an equal-setting macro average. Each of the 5,000 bootstrap draws resamples the aligned examples jointly and recomputes all four correlations; Spearman draws recompute average tied ranks. A draw with an undefined correlation is excluded from the corresponding joint contrast intervals. Pixtral Color Ordering retains 4,918 finite draws for these intervals, while the other settings retain 5,000.

The four-column comparison holds each oracle factor fixed while comparing its correspondence with the two learned branches. The native Vision background and example population remain the same, so a positive matched-minus-cross difference identifies the branch that corresponds more closely to that oracle effect on the same examples. Mean FF, FE, EE, and EF are 0.362, 0.216, 0.403, and 0.272, giving mean $D_F = 0.146$, $D_E = 0.131$, and $S = 0.139$. The combined score is positive in 18 settings, including all five Qwen3 tasks, and the complete comparison retains the cross-factor associations produced by the shared frozen decoder. Qwen3 Factual Recall illustrates this distinction: $D_F = -0.171$ and $D_E = 0.280$ yield a positive combined score of 0.054 because the Execution preference exceeds the Formation reversal. These held-out associations complement the upstream intervention evidence by showing which factor distinctions the trained updates preserve; their interpretation remains specific to the measured candidate-margin effects.

The repair-route categories from Section F provide a complementary behavioral view. Table K.2 compares F-only, E-only, and FCRD accuracy within Formation-only and Execution-only oracle categories, whose membership depends on the factorial interventions and remains fixed during learned evaluation. Within the 25 Qwen3 Arithmetic examples in the Formation-only group, F-only corrects 36.0%, E-only corrects 0.0%, and FCRD corrects 52.0%; composition therefore recovers additional answers even within a group defined by one sufficient oracle intervention. This comparison connects continuous effect alignment to actual recovery under a fixed intervention-defined condition. The full table preserves every model–task row and its denominator, including the small Execution-only groups: Qwen3 has three, two, two, zero, and zero such examples across the five tasks. Their conditional accuracies describe these groups precisely, with undefined accuracy for an empty group.

Table K.1: Complete Pearson matched/cross-factor correlations on finite aligned out-of-fold examples. Matched columns FF and EE appear beside their cross-factor controls FE and EF. Each row uses one common population of size N, including native successes and errors.

Model	Task	N	FF	FE	EE	EF
Qwen2	Counting	382	0.195	0.163	0.536	0.292
Qwen2	Arithmetic	1000	0.374	-0.322	0.191	0.074
Qwen2	Color ord.	1925	0.363	0.455	0.501	0.419
Qwen2	Factual	1264	0.047	0.048	0.025	-0.029
Qwen2	Sentiment	244	0.478	0.151	0.539	0.159
Pixtral	Counting	333	0.429	0.269	0.407	0.664
Pixtral	Arithmetic	1000	0.615	-0.028	0.121	0.205
Pixtral	Color ord.	496	0.153	0.031	-0.060	0.029
Pixtral	Factual	1264	0.138	0.074	0.179	-0.002
Pixtral	Sentiment	244	0.639	0.442	0.731	0.649
Gemma	Counting	382	0.662	0.526	0.696	0.649
Gemma	Arithmetic	1000	0.343	0.182	0.571	0.212
Gemma	Color ord.	1925	0.468	0.183	0.666	0.268
Gemma	Factual	1264	0.448	0.336	0.399	0.409
Gemma	Sentiment	244	0.494	0.644	0.666	0.415
Qwen3	Counting	382	0.287	0.342	0.719	0.496
Qwen3	Arithmetic	1000	0.274	0.182	0.281	0.046
Qwen3	Color ord.	1925	0.278	0.213	0.398	0.132
Qwen3	Factual	1264	0.012	0.183	0.210	-0.070
Qwen3	Sentiment	244	0.545	0.254	0.292	0.424

Table 11: Paired specialization estimates with 95% percentile intervals below each point estimate. Pearson is the primary statistic. Spearman S uses ranks recomputed within every resampled population. All four correlations share the same examples within a draw, preserving their dependence when D_F, D_E, and S are formed. The intervals are pointwise within each model–task setting.

Model	Task	Pearson D_F	Pearson D_E	Pearson S	Spearman S
Qwen2	Counting	0.032 [-0.049, 0.117]	0.244 [0.137, 0.343]	0.138 [0.090, 0.183]	0.079 [0.036, 0.123]
Qwen2	Arithmetic	0.696 [0.612, 0.774]	0.117 [0.017, 0.212]	0.406 [0.359, 0.455]	0.541 [0.486, 0.596]
Qwen2	Color ord.	-0.092 [-0.136, -0.049]	0.082 [0.030, 0.139]	-0.005 [-0.029, 0.021]	-0.025 [-0.047, -0.001]
Qwen2	Factual	-0.001 [-0.070, 0.063]	0.054 [0.003, 0.108]	0.026 [-0.006, 0.059]	0.057 [0.018, 0.097]
Qwen2	Sentiment	0.327 [0.193, 0.463]	0.380 [0.274, 0.482]	0.354 [0.289, 0.421]	0.400 [0.323, 0.482]
Pixtral	Counting	0.161 [0.081, 0.243]	-0.258 [-0.433, -0.075]	-0.049 [-0.106, 0.014]	0.001 [-0.067, 0.069]
Pixtral	Arithmetic	0.643 [0.575, 0.712]	-0.084 [-0.164, -0.007]	0.280 [0.234, 0.326]	0.236 [0.187, 0.288]
Pixtral	Color ord.	0.122 [0.000, 0.261]	-0.089 [-0.189, 0.021]	0.017 [-0.048, 0.095]	0.018 [-0.064, 0.101]
Pixtral	Factual	0.064 [-0.007, 0.135]	0.181 [0.121, 0.242]	0.123 [0.093, 0.152]	0.117 [0.084, 0.152]
Pixtral	Sentiment	0.197 [0.077, 0.322]	0.082 [-0.007, 0.167]	0.140 [0.090, 0.190]	0.125 [0.075, 0.176]
Gemma	Counting	0.136 [0.089, 0.181]	0.047 [-0.019, 0.102]	0.092 [0.067, 0.113]	0.048 [0.026, 0.073]
Gemma	Arithmetic	0.161 [0.070, 0.252]	0.358 [0.275, 0.447]	0.259 [0.207, 0.313]	0.247 [0.206, 0.290]
Gemma	Color ord.	0.285 [0.243, 0.328]	0.398 [0.355, 0.442]	0.342 [0.318, 0.367]	0.353 [0.326, 0.379]
Gemma	Factual	0.112 [0.072, 0.151]	-0.009 [-0.052, 0.034]	0.051 [0.040, 0.063]	0.036 [0.023, 0.049]
Gemma	Sentiment	-0.150 [-0.311, -0.024]	0.251 [0.114, 0.408]	0.051 [-0.003, 0.093]	-0.005 [-0.055, 0.047]
Qwen3	Counting	-0.055 [-0.146, 0.024]	0.223 [0.079, 0.392]	0.084 [0.024, 0.156]	0.059 [-0.005, 0.124]
Qwen3	Arithmetic	0.091 [-0.039, 0.222]	0.236 [0.108, 0.359]	0.163 [0.114, 0.217]	0.216 [0.171, 0.261]
Qwen3	Color ord.	0.066 [-0.001, 0.134]	0.266 [0.206, 0.326]	0.166 [0.120, 0.209]	0.190 [0.148, 0.230]
Qwen3	Factual	-0.171 [-0.246, -0.099]	0.280 [0.210, 0.357]	0.054 [0.024, 0.086]	0.120 [0.088, 0.153]
Qwen3	Sentiment	0.291 [0.159, 0.432]	-0.131 [-0.284, -0.003]	0.080 [0.031, 0.125]	-0.005 [-0.069, 0.059]

Table K.2: Learned out-of-fold accuracy (%) within the Formation-only and Execution-only oracle groups. n_F and n_E are the respective group sizes. Each group consists of native Text-correct/Vision-wrong examples whose indicated oracle hybrid alone is correct; membership is independent of the learned predictions. All three learned conditions evaluate the same group. A dash denotes undefined conditional accuracy for $n = 0$.

		Formation-only				Execution-only			
Model	Task	n_F	F-only	E-only	FCRD	n_E	F-only	E-only	FCRD
Qwen2	Counting	50	4.0	0.0	4.0	4	0.0	75.0	100.0
Qwen2	Arithmetic	222	41.9	0.9	52.3	0	–	–	–
Qwen2	Color ord.	351	3.1	0.3	4.8	2	0.0	50.0	50.0
Qwen2	Factual	213	15.5	0.0	18.3	0	–	–	–
Qwen2	Sentiment	4	75.0	50.0	75.0	0	–	–	–
Pixtral	Counting	36	25.0	2.8	33.3	0	–	–	–
Pixtral	Arithmetic	148	23.6	0.0	23.6	0	–	–	–
Pixtral	Color ord.	48	16.7	0.0	16.7	0	–	–	–
Pixtral	Factual	343	19.0	0.3	22.2	1	0.0	0.0	0.0
Pixtral	Sentiment	7	14.3	14.3	14.3	2	100.0	100.0	100.0
Gemma	Counting	42	11.9	0.0	31.0	1	0.0	0.0	0.0
Gemma	Arithmetic	80	70.0	13.8	70.0	0	–	–	–
Gemma	Color ord.	397	5.5	0.0	16.1	1	100.0	100.0	100.0
Gemma	Factual	75	1.3	2.7	14.7	4	75.0	75.0	75.0
Gemma	Sentiment	4	0.0	0.0	25.0	0	–	–	–
Qwen3	Counting	29	20.7	3.4	20.7	3	33.3	0.0	33.3
Qwen3	Arithmetic	25	36.0	0.0	52.0	2	50.0	0.0	100.0
Qwen3	Color ord.	181	5.0	0.0	10.5	2	0.0	50.0	0.0
Qwen3	Factual	195	8.7	0.5	14.4	0	–	–	–
Qwen3	Sentiment	12	58.3	0.0	58.3	0	–	–	–