VISTA: VISUAL INFORMATION STRESS TESTING FOR REPRESENTATION SHIFTS AND STATE-DEPENDENT INFLUENCE IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

What changes inside a vision-language model when a visual category remains recoverable under perceptual stress? We study matched color-desaturation and shape-defocus sequences across eight checkpoints, combining frozen readouts, native candidate scoring, and visual-state interventions. We find that severe clean-readout mismatch has a simple correctable component: a training-estimated, class-agnostic translation restores Shape readout accuracy to 0.979–1.000 in the frozen IID evaluation, while representation discrepancies are only partly removed. On a separate, explicitly linked evaluation cohort, all categories remain recoverable by a shared reader, yet Color stress consistently reduces true-class relative evidence and candidate accuracy; Shape changes the evidence in model-dependent directions while endpoint decisions remain near ceiling. Finally, recipient stress strengthens a fixed clean–stress visual contrast in Qwen–Color but weakens it in LLaVA–Shape. These opposite modulations persist across same-family, cross-family, and finite-interval constructions, with natural-direction amplitude advantages over matched controls; other settings are less consistent. The findings characterize translation-correctable readout mismatch, attribute-dependent answer changes, and heterogeneous conditional visual influence rather than a uniform loss of categorical information.

1 INTRODUCTION

Vision-language models (VLMs) must extract visual evidence that remains useful as image conditions change. Evaluations under image corruption show that their multimodal behavior can be sensitive to perceptual variation (Ding et al., 2024; Li et al., 2025a; Qiu et al., 2025; Usama et al., 2025; Saxena et al., 2026; Rehman & Shafique, 2026). An output error, however, does not reveal what changed along the visual pathway. A visual category may remain accessible to an external probe even when the model's own response changes. Understanding such cases requires following both the hidden representation and its downstream effects.

The distinction between decodability and functional relevance is already well established (Ravfogel et al., 2020; Elazar et al., 2021; Huang & Chang, 2025; Liao et al., 2026). The question here is more specific: when perceptual conditions change while category identity remains recoverable, what form does clean-readout incompatibility take, how does native answer evidence respond, and how does the effect of a fixed visual-state contrast vary with its recipient state?

We study these responses using controlled five-way Color and Shape tasks. Matched stimulus families preserve category and scene configuration while varying desaturation or Gaussian defocus. Across eight frozen checkpoints, we compare condition-specific, clean-only, and mixed-condition readers, and test a class-agnostic translation of the pooled visual state. For three representative checkpoints, we explicitly link readout predictions to native candidate scores on the same samples. A separate intervention cohort supports factorial tests in which a fixed visual-token contrast is evaluated under clean and stressed recipient states, followed by cross-family and finite-interval direction analyses.

Three concrete patterns emerge. Severe clean-readout mismatch is largely corrected by translation: Shape readout accuracy reaches 0.979–1.000 after correction in the frozen IID analysis, even

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.

though substantial paired representation differences remain. On the linked native-evidence cohort, Color lowers both relative true-class evidence and candidate accuracy, whereas Shape changes continuous evidence in different directions while endpoint decisions remain near ceiling. Recipient stress also modulates fixed visual contrasts in opposite directions for Qwen–Color and LLaVA–Shape, with agreement across constructions and additional support from direction controls. Other model–attribute combinations show weaker or metric-dependent patterns. Our contributions are:

1. We identify a translation-correctable component of severe clean-readout mismatch, separating decision-rule compatibility from residual representation differences.
2. We characterize attribute-dependent changes in native answer evidence on explicitly matched, recoverable samples, and show when these changes accompany constrained decision degradation.
3. We identify opposite recipient-state modulations that persist across intervention constructions, and evaluate their natural-direction amplitude alongside all weaker and unresolved model–attribute settings.

2 Related Work

2.1 Probing Visual Representations and Functional Relevance

Probing measures information accessible from representations; control tasks and information-theoretic analyses clarify how reader capacity and training affect its interpretation (Hewitt & Liang, 2019; Voita & Titov, 2020; Belinkov, 2022). Amnesic and nullspace-based interventions complement decoding by modifying representations and measuring downstream behavior (Ravfogel et al., 2020; Elazar et al., 2021). Recent visual and multimodal studies likewise distinguish decodability from causal influence and grounding (Huang & Chang, 2025; Liao et al., 2026). Our focus is the form of the change under matched perceptual stress: whether clean-readout incompatibility is translation-correctable, and how native answer evidence changes on the same recoverable examples.

2.2 Activation Interventions and Conditional Visual Influence

Causal tracing and activation patching test how hidden states affect outputs (Meng et al., 2022; Zhang & Nanda, 2024), while representation engineering and activation steering manipulate these states at inference time (Zou et al., 2023; Li et al., 2023; Turner et al., 2023; Rimsky et al., 2024). VLM applications include visual-information localization and multimodal behavior control (Palit et al., 2023; Li et al., 2026; Liu et al., 2025; Li et al., 2025b; Su et al., 2025; Sivakumar et al., 2025; Yin et al., 2026). Rather than selecting an effective steering direction or locating a single influential component, we hold a naturally derived clean–stress contrast fixed and measure how its effect changes with the recipient visual state. Cross-family constructions and norm-matched directions test the scope of the observed modulation.

2.3 Controlled Perturbations and Multimodal Robustness

Image-corruption benchmarks measure robustness under controlled input changes (Hendrycks & Dietterich, 2019); multimodal extensions evaluate behavior under perturbed or naturally corrupted images (Ding et al., 2024; Li et al., 2025a; Qiu et al., 2025; Usama et al., 2025; Saxena et al., 2026; Rehman & Shafique, 2026). VISTA complements behavioral robustness evaluation through matched within-family comparisons: category identity and scene configuration are held fixed while a perceptual parameter changes. This construction allows us to relate clean-readout compatibility, native candidate evidence, and conditional visual influence under the same controlled stress design.

3 Controlled Experimental Framework

3.1 Visual Attributes and Stress Construction

Each clean stimulus belongs to a matched stress family whose members retain the underlying category and scene configuration while varying one perceptual parameter (Figure 1).

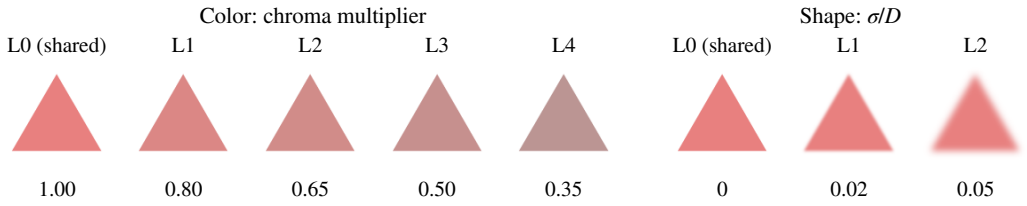

Figure 1: **Matched category-preserving perceptual stress.** Color varies chroma; Shape varies alpha-mask defocus. The sequences share a clean image and preserve the scene configuration. The example is selected by a fixed metadata rule, independently of model outputs.

1. **Color (L0–L4).** The categories are *red*, *yellow*, *green*, *blue*, and *purple*. Four levels of CIELAB chroma reduction progressively desaturate the clean object, with its shape, position, scale, and surrounding scene fixed.
2. **Shape (L0–L2).** The categories are *circle*, *triangle*, *square*, *pentagon*, and *hexagon*. Gaussian defocus is applied to the object's alpha mask before compositing. Its strength is σ/D, with Gaussian scale σ and area-equivalent object diameter D.

High stress, denoted H, is Color L4 or Shape L2. Rendering parameters and grouping rules are given in Appendix A.

3.2 Models, Data Partitions, and Evaluation Protocol

We evaluate InternVL3.5 (2B, 4B, 8B) (Wang et al., 2025), Qwen3.5 (2B, 4B, 9B) (Qwen Team, 2026), and LLaVA-OneVision-1.5 (4B, 8B) (An et al., 2025). Section 4.1 uses all eight checkpoints; Sections 4.2–4.3 use InternVL3.5-2B, Qwen3.5-2B, and LLaVA-OneVision-1.5-4B. All VLM parameters remain frozen.

Reader fitting and evaluation are separated by scaffold group, with all members of a group assigned to the same partition. Primary linear readers use 1,000 fitting representations from 40 groups and 500 validation representations from 20 groups. For Mixed readers, the fitting budget is shared across stress conditions.

The main readout table and native-evidence analysis use the same Confirmation partition, with 1,000 examples from 40 groups per condition. The before/after translation accuracies are reported on the frozen IID test partition, while the complementary representation-discrepancy analysis uses Confirmation. Each comparison retains its own paired clean–stress examples.

The intervention analysis uses a separate cohort of 100 matched clean–stress pairs per model–attribute setting, with 20 pairs per class from 40 scaffold groups. The same-family, cross-family, and local-response analyses share these recipient examples. Table 5 summarizes the evaluation roles and their source identities.

3.3 Representation Probing and Cross-Condition Alignment

We analyze pooled states at H_{enc}, before multimodal projection, and H_{post}, after projection and before the language backbone; main-text results use H_{post}. Three readout protocols characterize different aspects of cross-condition representation behavior:

1. **Condition-Specific (Cond-Spec)** fits one reader per condition to measure within-condition categorical recoverability.
2. **Fixed-L0 (Fixed)** fits on clean L0 and applies the same reader to stressed representations, measuring compatibility with the clean-condition decision rule.
3. **Matched-Mixed (Mixed)** fits one reader across matched conditions, measuring whether a shared cross-condition readout is available.

We evaluate linear and two-layer MLP readers using balanced accuracy (BA; chance level 0.2).

We further examine whether part of the Fixed-L0 mismatch has a simple cross-condition structure. In particular, stressed representations may undergo a systematic shift relative to their matched clean states while remaining categorically separable. We therefore estimate, for each model, interface, axis, and stress level, a class-agnostic translation from matched training pairs:

$$\hat{u}_s = \frac{1}{N_{\text{train}}} \sum_i \left(h_i^{(s)} - h_i^{(0)} \right), \qquad T_s(h) = h - \hat{u}_s. \tag{1}$$

Here, $h_i^{(0)}$ and $h_i^{(s)}$ are the pooled representations of a matched training pair, and N_{train} is the number of such pairs. The same offset is applied to every class, and the clean-trained reader remains fixed. This tests how much clean-readout compatibility can be recovered by removing a condition-level mean shift. We quantify the original and residual mismatch as

$$G_{\text{mis}}^{(s)} = \text{BA}_{\text{Cond}}^{(s)} - \text{BA}_{\text{Fixed}}^{(s)}, \qquad G_{\text{res}}^{(s)} = \text{BA}_{\text{Cond}}^{(s)} - \text{BA}_{\text{Trans}}^{(s)}. \tag{2}$$

We additionally measure the reduction in paired representation discrepancy:

$$R_{\text{disc}} = 1 - \frac{\mathcal{D}_{\text{trans}}}{\mathcal{D}_{\text{raw}}}. \tag{3}$$

Here, $\mathcal{D}_{\text{raw}}$ and $\mathcal{D}_{\text{trans}}$ are the group-averaged mean absolute coordinate differences between matched clean and stressed representations, before and after translation. Pooling, reader training, and discrepancy computation are detailed in Appendix B.

3.4 Native Answer-Evidence Analysis

To ensure preserved categorical recoverability, we analyze matched pairs from the same stress family for which the same frozen linear Mixed reader at H_{post} correctly predicts the category at both the lowest- and highest-stress endpoints. Stress-level trajectories additionally compare each intermediate condition with the L0 image from the same family.

For each image, the frozen VLM scores the same five candidate class labels, keeping the answer space fixed across perceptual conditions. Let $s_i^{(l)}(k)$ denote the sequence-mean log-probability of candidate k for sample i at stress level l, and let y_i be its ground-truth class. We define relative true-class evidence as

$$M_{\text{mean},i}^{(l)} = s_i^{(l)}(y_i) - \frac{1}{K-1} \sum_{k \neq y_i} s_i^{(l)}(k), \qquad K = 5. \tag{4}$$

We use ΔM_{mean} to measure its paired change from L0 and additionally report the balanced accuracy of the constrained decision $\arg\max_k s_i^{(l)}(k)$. To characterize changes across the full candidate set, we further construct the class-conditioned candidate-score shift matrix

$$\Delta S_{c,k}^{(l)} = \mathbb{E}\Big[s_i^{(l)}(k) - s_i^{(0)}(k) \mid y_i = c \Big], \qquad D_{\text{redis}}^{(l)} = \|\widetilde{\Delta S}^{(l)}\|_F, \tag{5}$$

where $\widetilde{\Delta S}^{(l)}$ is obtained by subtracting the mean of each row of $\Delta S^{(l)}$, removing the common score shift shared by all candidates for a given ground-truth class and thereby isolating changes in their relative evidence.

The total displacement $D = \|\Delta S\|_F$ and the complete score-shift matrices are reported in Appendix F. Table 2 uses the original candidate-scoring contract for all three models; the separate InternVL prompt remeasurement is described in Appendix C.

3.5 Factorial Tests of Recipient-State Dependence

The final analysis asks whether the effect of a fixed visual-state contrast is the same in different surrounding visual states. To disentangle the target-state effect from concurrent changes in the surrounding visual state, we construct a four-corner activation intervention that varies the two independently (Meng et al., 2022; Zhang & Nanda, 2024). At the intervention interface, we write the

visual state as $H = (Z, C)$. Here, Z contains the visual tokens selected by a frozen rectangular ROI associated with the scene scaffold, and C contains the remaining visual tokens. The spatial selection is fixed across the matched stress conditions. Each intervention replaces the selected token activations in full while retaining the complement from the chosen recipient state.

For a same-family clean–stress pair, the four states $(Z_0, C_0), (Z_H, C_0), (Z_0, C_H), (Z_H, C_H)$ yield $M_{00}, M_{H0}, M_{0H}, M_{HH}$. For a fixed recipient state C, we measure the effect of the same clean-versus-stressed target contrast as

$$\tau(C) = M(Z_0, C) - M(Z_H, C), \qquad \Delta\tau = \tau(C_H) - \tau(C_0). \tag{6}$$

Thus, $\Delta\tau$ measures the signed change in the target-state effect under recipient stress: positive values indicate an increase, and negative values a decrease.

The four corners also give

$$\begin{aligned} \Delta M_{\text{native}} &= M_{HH} - M_{00} = A + B + I, \\ A &= M_{H0} - M_{00}, \quad B = M_{0H} - M_{00}, \quad I = -\Delta\tau. \end{aligned} \tag{7}$$

This decomposition separates the target-state contribution A, the recipient-state contribution B, and their non-additive interaction $I = -\Delta\tau$. We compute all terms from the same candidate-score metric, using both M_{mean} and $M_{\text{best}} = s(y) - \max_{k \neq y} s(k)$. Native endpoint replay and self-write checks validate the intervention path (Appendix D).

We evaluate all six model–attribute settings with two complementary constructions. In the cross-family analysis, D_0 and D_H denote the clean and stressed token states from a same-label donor family. The same raw donor states are written into clean and stressed recipients from another family, yielding the endpoint contrast $\Delta\tau_D$.

For the local-response analysis, we vary the written state along $D_H + \alpha(D_0 - D_H)$, using $\alpha \in \{-0.10, -0.05, 0, 0.05, 0.10\}$. We fit a separate OLS slope for each donor–recipient pair under each recipient state. Their difference, $\Delta\hat{g}$, measures how the finite-interval response changes with the recipient.

To compare the natural direction with generic perturbations, we use five random directions and one orthogonal direction with the same anchor, step schedule, and direction-norm contract. The random-direction amplitude advantage for pair i is

$$a_{i,\text{rand}} = |\Delta\hat{g}_{i,\text{natural}}| - \frac{1}{5}\sum_{r=1}^{5} |\Delta\hat{g}_{i,\text{random},r}| \,. \tag{8}$$

The orthogonal advantage replaces the random-direction average with $|\Delta\hat{g}_{i,\text{orthogonal}}|$. Reported advantages average these paired differences across dyads.

Uncertainty is estimated by scaffold-group bootstrap, keeping paired conditions together. Same-family analyses resample recipient groups; cross-family and local analyses report sensitivity envelopes of the donor- and recipient-group intervals. All intervals are pointwise. Full estimands and statistical procedures are given in Appendices D and E.

4 Results

4.1 Clean-Readout Mismatch Is Largely Translation-Correctable

At maximum stress, linear Cond-Spec and Mixed readers achieve 1.000 BA across all eight checkpoints, while clean-trained readers can fail severely (Table 1). On Shape, Qwen Fixed-L0 BA falls to 0.200–0.400 and InternVL to 0.714–0.798; LLaVA's linear readout remains at ceiling. The MLP readers do not consistently remove the mismatch, despite near-ceiling condition-specific and mixed-condition performance.

An upstream comparison shows the same readout separation for Qwen and InternVL at H_{enc}. On Shape L2, their representative checkpoints obtain Fixed-L0 BA of 0.583 and 0.806, respectively, while Cond-Spec and Mixed remain at 1.000; LLaVA remains at 1.000 under all three protocols (Table 7). The magnitude of clean-readout mismatch therefore varies across the two visual interfaces.

Table 1: **Recoverability versus clean-readout compatibility at maximum stress.** Balanced accuracy on H_{post} for Color L4 and Shape L2 on the Confirmation test split (1,000 examples, 40 groups per condition). MLP values average five random seeds.

		Color L4			Shape L2		
Model	**Reader**	**Cond.**	**Fixed**	**Mixed**	**Cond.**	**Fixed**	**Mixed**
InternVL3.5-2B	Linear	1.000	0.933	1.000	1.000	0.750	1.000
	MLP	1.000	0.798	0.992	1.000	0.531	1.000
InternVL3.5-4B	Linear	1.000	1.000	1.000	1.000	0.798	1.000
	MLP	1.000	0.834	0.996	1.000	0.502	1.000
InternVL3.5-8B	Linear	1.000	1.000	1.000	1.000	0.714	1.000
	MLP	1.000	0.832	0.976	1.000	0.714	1.000
LLaVA-OV1.5-4B	Linear	1.000	1.000	1.000	1.000	1.000	1.000
	MLP	1.000	0.727	0.998	1.000	0.978	1.000
LLaVA-OV1.5-8B	Linear	1.000	1.000	1.000	1.000	1.000	1.000
	MLP	1.000	0.838	0.983	1.000	0.857	1.000
Qwen3.5-2B	Linear	1.000	0.976	1.000	1.000	0.200	1.000
	MLP	1.000	0.802	0.998	1.000	0.222	1.000
Qwen3.5-4B	Linear	1.000	0.998	1.000	1.000	0.400	1.000
	MLP	1.000	0.785	0.997	1.000	0.362	1.000
Qwen3.5-9B	Linear	1.000	1.000	1.000	1.000	0.400	1.000
	MLP	1.000	0.869	0.995	1.000	0.400	1.000

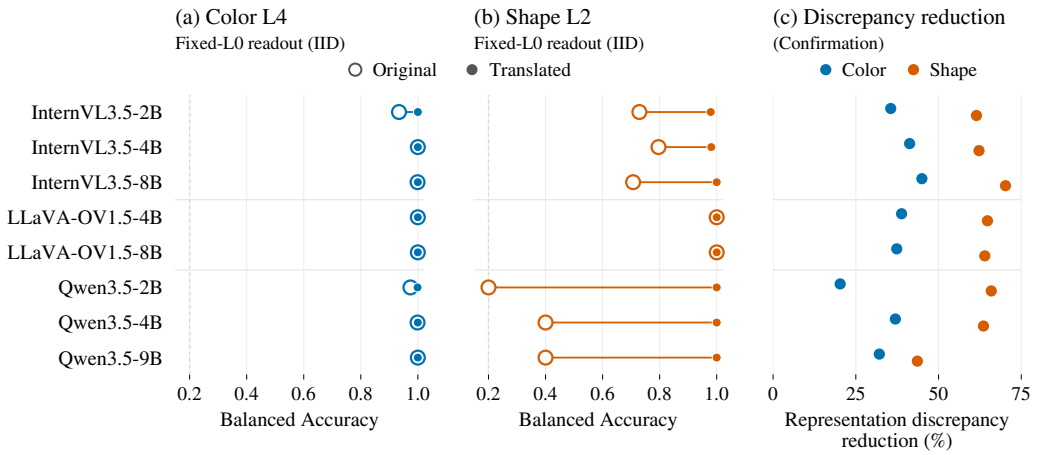

Figure 2: **Translation restores readout compatibility with residual representation displacement.** (a,b) Fixed-L0 linear-reader BA before and after class-agnostic translation on IID; segments pair results within each checkpoint. (c) Paired H_{post} discrepancy reduction on Confirmation. Translations are estimated from training data.

A class-agnostic translation captures a large fraction of this cross-condition incompatibility (Figure 2(a,b)). In the frozen IID evaluation, the three Qwen Shape-L2 Fixed-L0 accuracies are 0.200, 0.400, and 0.400 before correction and 1.000 after translation. Across all eight checkpoints, translated Shape-L2 accuracy reaches 0.979–1.000.

The complementary discrepancy analysis uses the Confirmation partition (Figure 2(c)). Translation reduces paired H_{post} discrepancy by 20.3%–45.0% for Color and 43.6%–70.2% for Shape. Substantial clean–stress coordinate differences therefore remain after removing the estimated condition-level offset.

Table 2: **Native evidence and decisions on recoverable samples.** Confirmation results at L0 and high stress (Color L4; Shape L2). All 1,000 pairs per setting are correctly decoded at both endpoints by the same linear Mixed reader. All measures use the original scoring protocol; brackets are pointwise 95% paired group-bootstrap intervals.

		Candidate BA			
Model	**Axis**	**L0**	**High**	$\Delta M_{\mathbf{mean}}$	$\mathbf{D}_{\mathrm{redis}}$
Qwen3.5-2B	Color	0.956	0.709	$-4.747\ [-4.805, -4.688]$	11.735
Qwen3.5-2B	Shape	0.999	1.000	$+0.121\ [+0.062, +0.192]$	2.425
LLaVA-OV1.5-4B	Color	0.984	0.635	$-3.441\ [-3.465, -3.416]$	9.993
LLaVA-OV1.5-4B	Shape	1.000	1.000	$-1.002\ [-1.019, -0.985]$	3.502
InternVL3.5-2B	Color	1.000	0.600	$-3.876\ [-3.887, -3.865]$	12.660
InternVL3.5-2B	Shape	1.000	1.000	$-0.389\ [-0.406, -0.372]$	2.408

Taken together, these results identify a translation-correctable component of severe clean-readout mismatch. A simple class-agnostic shift can largely restore compatibility with the clean decision rule while substantial clean–stress representation differences remain, separating decision-rule compatibility from residual representational change.

4.2 Attribute-Dependent Changes in Native Evidence and Decisions

All 1,000 matched examples in each of the six representative settings are correctly classified at both endpoints by the same frozen Mixed reader. On these identical examples, Color reduces the mean margin in all three models, while Shape increases it slightly in Qwen and decreases it in LLaVA and InternVL (Table 2). Thus, preserved readout success coexists with a structured, attribute-dependent response in the model's own candidate evidence.

The candidate decisions distinguish two consequences of the evidence changes. Color BA falls from 0.956 to 0.709 for Qwen, from 0.984 to 0.635 for LLaVA, and from 1.000 to 0.600 for InternVL. In contrast, Shape endpoint BA remains at or near 1.000 despite changed margins. Continuous evidence change is therefore not synonymous with a changed class decision. Together with Table 1, these results also show contrasting readout/behavior patterns: Qwen–Shape has severe Fixed-L0 failure but near-perfect native decisions, whereas LLaVA–Color preserves perfect linear Fixed-L0 performance while its native candidate decisions degrade. These comparisons use the same Confirmation split.

The trajectories in Figure 3 distinguish the progression of these changes. For Color, the mean margin becomes more negative and D_{redis} increases from L1 to L4 in all three models. For Shape, D_{redis} also increases from L1 to L2, while candidate accuracy at the clean and maximum-stress endpoints remains near ceiling. InternVL additionally exhibits an intermediate accuracy dip to 0.938 at L1 before returning to 1.000 at L2 (Table 11).

Qwen's Shape scores further distinguish total displacement from relative redistribution: D decreases from 2.808 at L1 to 2.640 at L2, whereas D_{redis} increases from 2.215 to 2.425. The stress response therefore involves attribute-dependent changes in relative candidate evidence, with different consequences for the constrained class decision. Complete score-shift matrices and level-wise diagnostics are reported in Appendices C and F.

4.3 Recipient Stress Modulates Fixed Visual Contrasts in Opposite Directions

The same-family contrasts identify opposite recipient modulations for Qwen–Color and LLaVA–Shape under both margins (Table 3). For M_{best}, the clean-versus-stressed target effect grows from 3.596 to 4.274 under recipient stress in Qwen–Color, but falls from 0.390 to 0.190 in LLaVA–Shape. Both target effects remain positive; it is their modulation that has opposite signs. LLaVA–Color has a smaller positive interaction under both margins. For Qwen–Shape, the M_{mean} interval excludes zero, whereas the M_{best} interval includes zero. Both InternVL settings show negative M_{best} interactions, with M_{mean} intervals that include zero.

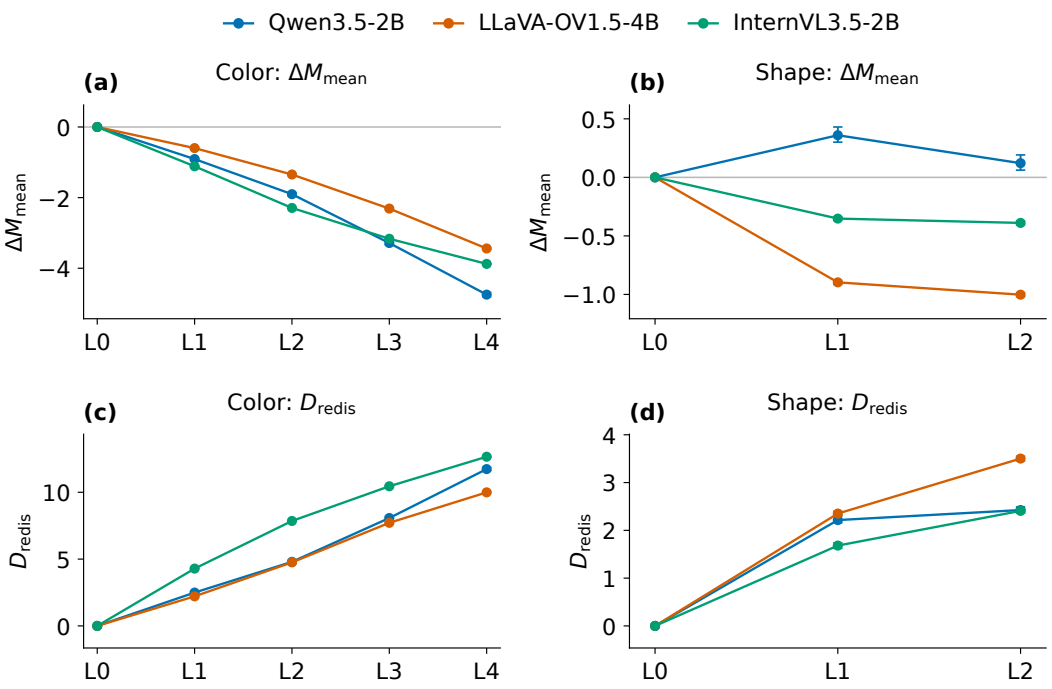

Figure 3: **Stress-indexed native answer-evidence change.** Paired margin change ΔM_{mean} (top) and relative candidate-space displacement D_{redis} (bottom) from L0. The original Confirmation scoring contract is retained for all three models.

Table 3: **Same-family recipient-state dependence.** $\Delta\tau$ compares a fixed target-state contrast across recipient states. Each setting contains 100 pairs from 40 scaffold groups in the intervention cohort, distinct from Table 2. Brackets are pointwise 95% paired group-bootstrap intervals.

Model	**Axis**	$\boldsymbol{\Delta\tau_{M_{\text{best}}}}$	$\boldsymbol{\Delta\tau_{M_{\text{mean}}}}$
Qwen3.5-2B	Color	$+0.678$ $[+0.476, +0.881]$	$+0.780$ $[+0.530, +1.039]$
Qwen3.5-2B	Shape	$+0.091$ $[-0.031, +0.223]$	$+0.380$ $[+0.183, +0.608]$
LLaVA-OV1.5-4B	Color	$+0.057$ $[+0.015, +0.101]$	$+0.103$ $[+0.066, +0.142]$
LLaVA-OV1.5-4B	Shape	-0.200 $[-0.233, -0.170]$	-0.296 $[-0.354, -0.241]$
InternVL3.5-2B	Color	-0.066 $[-0.114, -0.016]$	-0.028 $[-0.064, +0.008]$
InternVL3.5-2B	Shape	-0.085 $[-0.139, -0.032]$	-0.023 $[-0.065, +0.019]$

The native-change decomposition gives the scale and sign of these effects within the intervention cohort (Figure 4). For Qwen–Color, the target-state term A is the largest signed component, with the interaction adding a smaller negative contribution to the native clean-to-stress change. In LLaVA–Shape and InternVL–Shape, terms oppose one another. InternVL–Shape has a small net M_{mean} change because negative target and positive recipient terms largely cancel.

The opposite Qwen–Color and LLaVA–Shape signs persist under cross-family endpoints and finite-interval local responses, including both evidence margins (Figure 5(a,b)). For the local M_{best} analysis, natural-direction amplitude advantages over the random and orthogonal controls are $+0.553$ and $+0.565$ for Qwen–Color, and $+0.096$ and $+0.127$ for LLaVA–Shape. All four reported sensitivity envelopes have positive lower bounds (Figure 5(c)).

Thus, these settings show both consistent signed modulation and larger absolute modulation along the natural direction than along the matched controls.

For LLaVA–Shape, all 100 dyads retain the same strongest wrong candidate within the cross-family endpoint and local-response comparisons. Both effects remain negative for all five named fixed-wrong comparators (Appendix F.6).

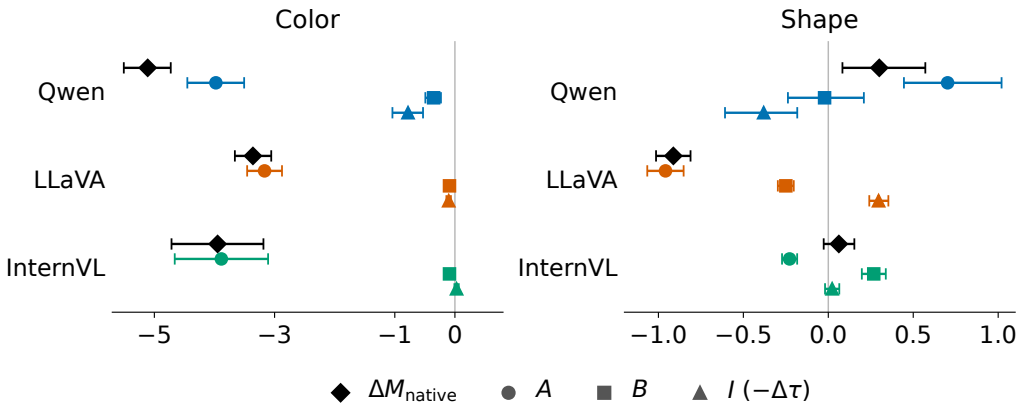

Figure 4: **Decomposition of the intervention cohort's native change.** Target contribution A, recipient contribution B, and interaction I sum to ΔM_{native}. Points use M_{mean}; error bars are pointwise 95% paired group-bootstrap intervals.

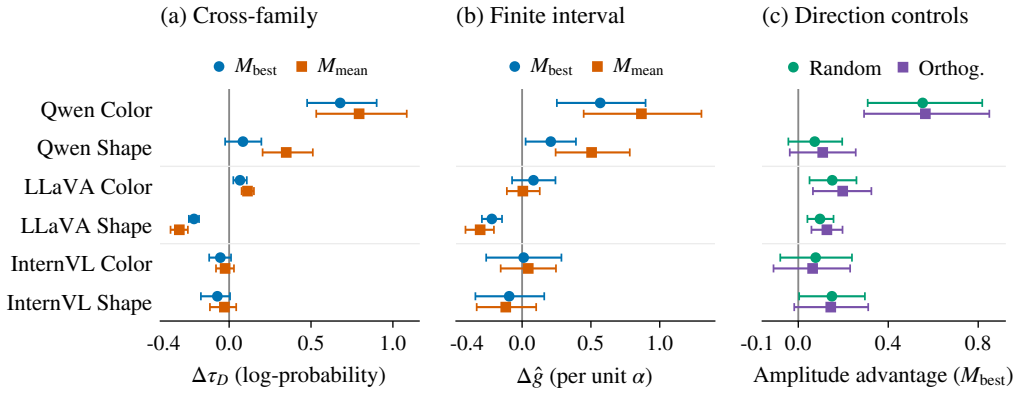

Figure 5: **Cross-construction consistency and direction controls.** (a) Cross-family endpoint contrast $\Delta\tau_D$. (b) Finite-interval response difference $\Delta\hat{g}$. (c) Natural-direction amplitude advantages over random and orthogonal controls under M_{best}. Error bars are sensitivity envelopes of pointwise 95% donor- and recipient-group bootstrap intervals (Appendix E).

The remaining settings separate these properties. Qwen–Shape has positive local responses under both margins; LLaVA–Color has positive amplitude advantages but unresolved signed local modulation. InternVL's cross-family and local intervals include zero, despite its same-family M_{best} effects.

The evidence supports specific, heterogeneous conditional effects rather than a common signed response across all models and attributes.

5 Conclusion

We studied how category-preserving visual stress changes frozen VLM representations and their downstream effects. Severe clean-readout mismatch is largely translation-correctable, while native evidence exhibits attribute-dependent changes that need not alter endpoint decisions. Fixed visual contrasts also show opposite recipient-state modulations in Qwen–Color and LLaVA–Shape, with cross-construction consistency and natural-direction amplitude advantages. The study covers two synthetic attribute axes and fixed spatial token partitions; generalization to naturalistic scenes, open-ended responses, and alternative intervention regions remains to be tested. A specific downstream circuit responsible for the observed modulation remains unidentified.

AI USE STATEMENT

Generative AI tools assisted manuscript organization, drafting and revision, and preparation of code and experiment specifications. The numerical results reported here are taken from recorded experimental outputs, rather than generated as substitute evidence. The authors are responsible for verifying the final text, claims, code, and artifacts.

REPRODUCIBILITY STATEMENT

Appendices A–F specify the stimulus construction, data roles, readout training, candidate scoring, frozen intervention interfaces, statistical estimands, and auxiliary results. The machine-readable exports retain the model, reader, cohort, and source identities needed to distinguish the analyses.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. URL https://arxiv.org/abs/2509.23661.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Peng Ding, Jingyu Wu, Jun Kuang, Dan Ma, Xuezhi Cao, Xunliang Cai, Shi Chen, Jiajun Chen, and Shujian Huang. Hallu-PI: Evaluating hallucination in multi-modal large language models within perturbed inputs. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10707–10715. ACM, 2024. doi: 10.1145/3664647.3681251. URL https://dl.acm.org/doi/10.1145/3664647.3681251.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Lianghuan Huang and Yingshan Chang. Causality $\neq$ decodability, and vice versa: Lessons from interpreting counting ViTs. *arXiv preprint arXiv:2510.09794*, 2025. doi: 10.48550/arXiv.2510.09794. URL https://arxiv.org/abs/2510.09794.

Chunyi Li, Jianbo Zhang, Zicheng Zhang, Haoning Wu, Yuan Tian, Wei Sun, Guo Lu, Xiongkuo Min, Xiaohong Liu, Weisi Lin, Xiao-Ping Zhang, and Guangtao Zhai. R-Bench: Are your large multimodal model robust to real-world corruptions? *IEEE Journal of Selected Topics in Signal Processing*, 19(7):1349–1361, 2025a. doi: 10.1109/JSTSP.2025.3558652. URL https://ieeexplore.ieee.org/document/11006523/.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (37):31645–31653, 2026. doi: 10.1609/aaai.v40i37.40431. URL https://ojs.aaai.org/index.php/AAAI/article/view/40431.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819. PMLR, 2025b. URL https://proceedings.mlr.press/v267/li25ca.html.

Chih-Ting Liao, Fei Shen, Xin Cao, and Tat-Seng Chua. Decodable is not grounded: A vision-ablation arbiter for VLM spatial reasoning. *arXiv preprint arXiv:2606.31257*, 2026. doi: 10.48550/arXiv.2606.31257. URL https://arxiv.org/abs/2606.31257.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, pp. 72402–72419, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b4008025c2182bfe16fcc8566ee14d64-Abstract-Conference.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for BLIP. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2848–2853. IEEE, 2023. doi: 10.1109/ICCVW60793.2023.00307. URL https://doi.org/10.1109/ICCVW60793.2023.00307.

Xinkuan Qiu, Meina Kan, Yongbin Zhou, and Shiguang Shan. Benchmarking multimodal large language models against image corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9014–9023. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00843. URL https://openaccess.thecvf.com/content/ICCV2025/html/Qiu_Benchmarking_Multimodal_Large_Language_Models_Against_Image_Corruptions_ICCV_2025_paper.html.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Saim Rehman and Muhammad Shafique. BRUCE: Benchmarking robustness under corruption escalation for scientific vision-language reasoning. *arXiv preprint arXiv:2608.07742*, 2026. doi: 10.48550/arXiv.2608.07742. URL https://arxiv.org/abs/2608.07742.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Rohit Saxena, Alessandro Suglia, and Pasquale Minervini. VLM-RobustBench: A comprehensive benchmark for robustness of vision-language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/64996.

Anushka Sivakumar, Andrew Zhang, Zaber Ibn Abdul Hakim, and Chris Thomas. SteerVLM: Robust model control through lightweight activation steering for vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23640–23665. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1285. URL https://aclanthology.org/2025.findings-emnlp.1285/.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12964–12974. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.634. URL https://aclanthology.org/2025.acl-long.634/.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023. doi: 10.48550/arXiv.2308.10248. URL https://arxiv.org/abs/2308.10248v1.

Muhammad Usama, Syeda Aishah Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. Analysing the robustness of vision-language-models to common corruptions. *arXiv preprint arXiv:2504.13690*, 2025. doi: 10.48550/arXiv.2504.13690. URL https://arxiv.org/abs/2504.13690.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. In *International Conference on Learning Representations*, pp. 2092–2109, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0480adaf62a918405a5e3b1031e0c056-Abstract-Conference.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. doi: 10.48550/arXiv.2310.01405. URL https://arxiv.org/abs/2310.01405.

A Stimulus Construction and Data Splits

A.1 Visual categories and perceptual stress

All experiments use the controlled Color–Shape stimulus construction in Section 3.1. Readout and native-evidence analyses use one dataset, while the intervention analyses use a separately generated cohort under the same construction. The tasks contain two five-way category spaces,

$$\mathcal{Y}_{\text{Color}} = \{\texttt{red}, \texttt{yellow}, \texttt{green}, \texttt{blue}, \texttt{purple}\}, \tag{9}$$

and

$$\mathcal{Y}_{\text{Shape}} = \{\texttt{circle}, \texttt{triangle}, \texttt{square}, \texttt{pentagon}, \texttt{hexagon}\}. \tag{10}$$

Color stress is generated in CIELAB space with fixed lightness $L^* = 65$ and category-specific hue $25°, 95°, 145°, 270°, 315°$ for red, yellow, green, blue, and purple, respectively. The clean chroma is $C^* = 44$, and stress scales it by the frozen multiplier ρ_C. Shape stress applies Gaussian defocus to the target object's alpha mask before compositing. Its normalized strength is σ/D, where $D = 2\sqrt{A_0/\pi}$ and A_0 is the canonical alpha-mask area. Images are rendered at 256×256 with $4\times$ supersampling. Only Shape L0–L2 are used in the present analyses.

Table 4: **Frozen visual stress axes used in this paper.**

Axis	Level	Operation	Frozen parameter
Color	L0	CIELAB chroma scaling	$\rho_C = 1.00$
Color	L1	CIELAB chroma scaling	$\rho_C = 0.80$
Color	L2	CIELAB chroma scaling	$\rho_C = 0.65$
Color	L3	CIELAB chroma scaling	$\rho_C = 0.50$
Color	L4	CIELAB chroma scaling	$\rho_C = 0.35$
Shape	L0	clean alpha mask	$\sigma/D = 0$
Shape	L1	Gaussian alpha-mask defocus	$\sigma/D = 0.02$
Shape	L2	Gaussian alpha-mask defocus	$\sigma/D = 0.05$

A.2 Matched families and scaffold grouping

Each underlying example is rendered as a *matched stress family*. Members of a family preserve the task category and scene configuration while varying only the designated stress level. Clean–stress comparisons are therefore paired by construction.

The frozen release additionally provides a `scaffold_group_id`, which is used as the grouping key for data partitioning and uncertainty estimation. For ROI-based analyses, the target region is a fixed rectangular region associated with the scaffold rather than a category-contour mask.

A.3 Data partitions

Reader fitting and evaluation are separated at the scaffold-group level. Each primary linear reader uses 1,000 fitting vectors from 40 groups and 500 validation vectors from 20 groups. The same total budgets apply to Matched-Mixed readers, with fitting examples distributed across the selected stress conditions.

Table 1 uses the Confirmation partition, with 1,000 examples from 40 groups per condition. Fixed-L0 and Mixed MLP results use the corresponding fitted readers; Condition-Specific MLP results come from a separate supplementary fit described in Appendix B. The translation accuracy range 0.979–1.000 is evaluated on the IID partition. Table 9 provides the within-IID before/after values. OOD and Confirmation transformation evaluations are retained separately from IID.

The native-evidence analysis uses the original Confirmation candidate scores, linked by sample identity to predictions of the frozen linear Mixed readers. Each stressed example is paired with its clean counterpart. All 1,000 samples per endpoint condition satisfy the recoverability criterion, so

the full-cohort and recoverable-subset estimates coincide. This linkage evaluates pooled readout predictions and native candidate scores; off-diagonal intervention states are analyzed separately.

The intervention cohort contains 200 donor–recipient dyads, 100 per axis, shared across the three representative models. Each cell contains 40 donor groups and 40 recipient groups with paired group topology. The follow-up same-family analysis uses the same 100 recipient examples per axis. These recipients are distinct from the native-evidence cohort. The three checkpoints yield 600 model–dyad observations over 200 visual dyads; the same-family follow-up reuses the already evaluated intervention cohort.

Table 5: **Data roles and evaluation identities.** Sizes are per relevant model–axis condition or fit.

Use	Rows / pairs	Groups	Split / source
Probe fitting	1,000 total vectors	40	Training
Probe validation	500 total vectors	20	Validation
Main probe evaluation	1,000 / condition	40	Confirmation
Translation summary	1,000 / condition	40	IID evaluation
Native evidence	1,000 / condition	40	Confirmation
Cross-family / local	100 dyads / cell	40 per role	Intervention cohort
Same-family four-corner	100 pairs / cell	40	Shared intervention recipients

Development and Confirmation denote the original data roles; analyses that reuse previously inspected data are not treated as new blind tests.

Representative stimuli. Figure 1 shows a red triangle on the first Confirmation scaffold in the fixed metadata ordering. Selection is independent of model outputs. The Color and Shape sequences share the same clean image and scene configuration.

B Representation Extraction and Probe Training

B.1 Representation interfaces and pooling

We analyze two frozen visual interfaces. H_{enc} is captured from the vision encoder before multimodal projection/merging, and H_{post} from the projected visual state before the language backbone. For each captured module output, the first tensor is detached, converted to float, reshaped to $(-1, d)$, and averaged over its first dimension:

$$h_i = \frac{1}{N_i} \sum_{n=1}^{N_i} X_i[n, :]. \tag{11}$$

No ROI weighting, explicit CLS removal, or L2 normalization is applied during this pooling step. If a captured output contains a CLS-like token, it remains part of the mean. Probe fitting subsequently standardizes each feature using the fitting-subset mean and population standard deviation, with a standard deviation floor of 10^{-6}; the same statistics are reused at evaluation.

The exact per-image pre-pooling token ranks/counts were not recorded in the inspected frozen cache metadata. Table 6 therefore reports the frozen hook paths and pooled feature widths, which are sufficient to specify the extraction code path without inferring unrecorded token counts.

A separate Shape ROI sanity analysis uses a fixed rectangular fractional-token pooling rule. For Qwen3.5-2B Shape L2, the global Fixed-L0 reader reaches balanced accuracy 0.2, whereas the Condition-Specific and Matched-Mixed readers both reach 1.0; the ROI readout shows the same qualitative separation.

Table 6: **Representation-extraction interfaces.** Pooled feature widths are reported separately for the encoder and post-projection states.

Model	Interface	Output hook	Width
InternVL3.5-2B	H_{enc}	`vision_model`	1024
	H_{post}	`mlp1`	2048
InternVL3.5-4B	H_{enc}	`model.vision_tower`	1024
	H_{post}	`model.multi_modal_projector`	2560
InternVL3.5-8B	H_{enc}	`vision_model`	1024
	H_{post}	`mlp1`	4096
LLaVA-OV1.5-4B	H_{enc}	`model.visual.blocks.23`	1024
	H_{post}	`model.visual.merger`	2560
LLaVA-OV1.5-8B	H_{enc}	`model.visual.blocks.23`	1024
	H_{post}	`model.visual.merger`	4096
Qwen3.5-2B	H_{enc}	`model.visual.blocks.23`	1024
	H_{post}	`model.visual.merger`	2048
Qwen3.5-4B	H_{enc}	`model.visual.blocks.23`	1024
	H_{post}	`model.visual.merger`	2560
Qwen3.5-9B	H_{enc}	`model.visual.blocks.26`	1152
	H_{post}	`model.visual.merger`	4096

Table 7: **Encoder-level readout results for the representative checkpoints.** Linear-reader balanced accuracy at Shape L2 on the Confirmation partition, with 1,000 examples from 40 scaffold groups per model. These results use H_{enc}, before multimodal projection.

Model	Cond-Spec	Fixed-L0	Mixed
Qwen3.5-2B	1.000	0.583	1.000
LLaVA-OV1.5-4B	1.000	1.000	1.000
InternVL3.5-2B	1.000	0.806	1.000

B.2 Linear readers

Linear readers are trained separately for each model, representation interface, visual axis, and readout protocol. The regularization grid is

$$C \in \{0.01, 0.1, 1, 10\},$$

with selection by validation balanced accuracy. The fitting set contains 1,000 vectors from 40 groups and the validation set 500 vectors from 20 groups. The three primary protocols are Condition-Specific, Fixed-L0, and Matched-Mixed, as defined in Section 3.3.

B.3 MLP readers

The MLP reader has one hidden layer of width 64 with ReLU activation and a linear output layer. It is optimized with softmax cross-entropy using AdamW (learning rate 3×10^{-3}, weight decay 10^{-4}) for 40 full-batch epochs, without early stopping or a learning-rate schedule. Inputs are standardized using the fitting-subset mean and population standard deviation computed in float64 and stored in float32, with a 10^{-6} standard deviation floor and no additional L2 normalization.

Weights are Gaussian-initialized with scale proportional to the inverse square root of the input width and biases are initialized to zero. Five frozen seeds are used:

$$20260906,\ 20260907,\ 20260908,\ 20260909,\ 20260910.$$

The implementation maintains independently initialized parameter sets for the five seeds and optimizes the mean of their cross-entropy losses. Main-text MLP results report the average across these seeded readers. Condition-Specific MLP readers were fitted in a supplementary analysis using the same architecture, training procedure, and five-seed protocol; their results are reported separately from the original Fixed-L0 and Mixed reader fits.

B.4 BUDGET CONTROLS

The reader sweep additionally evaluates fitting subsets with 5, 10, 20, and 40 scaffold groups, together with shuffled-label controls and IID/OOD adaptation conditions. Primary fits use the total budgets in Appendix A.3.

Table 8 reports five clean-trained, label-shuffled linear readers on the maximum-stress Confirmation conditions. Seed variation is descriptive and is not a confidence interval over visual examples.

Table 8: **Shuffled-label readout control.** Confirmation BA of Fixed-L0 label-shuffled linear readers at H_{post}, mean $\pm$ sample standard deviation across five frozen null seeds. Chance BA is 0.2. These are not Cond-Spec null readers.

Model	Color L4	Shape L2
InternVL3.5-2B	0.251 ± 0.133	0.215 ± 0.108
InternVL3.5-4B	0.236 ± 0.060	0.262 ± 0.090
InternVL3.5-8B	0.185 ± 0.084	0.153 ± 0.087
LLaVA-OV1.5-4B	0.215 ± 0.039	0.231 ± 0.063
LLaVA-OV1.5-8B	0.251 ± 0.112	0.188 ± 0.021
Qwen3.5-2B	0.279 ± 0.118	0.168 ± 0.047
Qwen3.5-4B	0.218 ± 0.083	0.213 ± 0.148
Qwen3.5-9B	0.246 ± 0.097	0.263 ± 0.087

B.5 CROSS-CONDITION TRANSLATION AND DISCREPANCY

For each stress level s, the class-agnostic translation is estimated from paired fitting representations:

$$\hat{u}_s = \frac{1}{N_{\text{train}}} \sum_i \left(h_i^{(s)} - h_i^{(0)} \right), \qquad T_s(h^{(s)}) = h^{(s)} - \hat{u}_s. \tag{12}$$

Equivalently, the implementation shifts stressed vectors from the fitting stress mean to the fitting clean mean. A separate vector is fitted for each model, interface, axis, and stress level; no class-specific translation is used, and the clean-trained reader remains fixed.

For each paired example, the raw and translated discrepancies are the mean absolute coordinate differences in the pooled representation space:

$$d_{i,\text{raw}} = \frac{1}{d} \sum_{j=1}^{d} \left| h_{ij}^{(s)} - h_{ij}^{(0)} \right|, \qquad d_{i,\text{trans}} = \frac{1}{d} \sum_{j=1}^{d} \left| T_s(h_i^{(s)})_j - h_{ij}^{(0)} \right|. \tag{13}$$

These distances are computed in the raw pooled space, without probe standardization. The frozen summarizer first averages within scaffold group and then weights groups equally:

$$\mathcal{D}_q = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \frac{1}{n_g} \sum_{i \in g} d_{i,q}, \qquad q \in \{\text{raw}, \text{trans}\}. \tag{14}$$

The reported discrepancy reduction is

$$R_{\text{disc}} = 1 - \frac{\mathcal{D}_{\text{trans}}}{\mathcal{D}_{\text{raw}}}. \tag{15}$$

Figure 2 reports $100 R_{\text{disc}}$ in percent.

IID translation results. Table 9 reports within-IID H_{post} Shape L2 accuracies before and after translation. These values are evaluated separately from the Confirmation readout results in Table 1.

Offset correction versus relative geometry. A common translation preserves within-condition pairwise differences: $T_s(h_i) - T_s(h_j) = h_i - h_j$. For an affine reader, after absorbing its fixed feature standardization into W, b,

$$W(h - \hat{u}_s) + b = Wh + (b - W\hat{u}_s). \tag{16}$$

Thus, readout recovery admits an offset/decision-bias interpretation. The discrepancy analysis measures paired cross-condition displacement; it does not by itself measure deformation of category-relative geometry.

Table 9: **Within-IID linear-readout correction at Shape L2.** The same clean reader is used before and after class-agnostic translation. All numbers are from the frozen transformation analysis.

Model	Fixed-L0 BA	Translated BA
InternVL3.5-2B	0.729	0.979
InternVL3.5-4B	0.796	0.981
InternVL3.5-8B	0.707	1.000
LLaVA-OV1.5-4B	1.000	1.000
LLaVA-OV1.5-8B	1.000	1.000
Qwen3.5-2B	0.200	1.000
Qwen3.5-4B	0.400	1.000
Qwen3.5-9B	0.400	1.000

C Candidate Scoring and Prompt Configuration

C.1 Candidate sets

All candidate-scoring analyses use the same five labels as the corresponding classification task:

$$\mathcal{Y}_{\text{Color}} = \{\texttt{red}, \texttt{yellow}, \texttt{green}, \texttt{blue}, \texttt{purple}\},$$

and

$$\mathcal{Y}_{\text{Shape}} = \{\texttt{circle}, \texttt{triangle}, \texttt{square}, \texttt{pentagon}, \texttt{hexagon}\}.$$

The frozen candidate tokenizer produces the token counts shown in Table 10 for the representative checkpoints.

Table 10: **Candidate answer token counts in the frozen scoring setup.**

Axis	Candidate	Tokens
Color	red	1
Color	yellow	1
Color	green	1
Color	blue	1
Color	purple	1
Shape	circle	1
Shape	triangle	1
Shape	square	1
Shape	pentagon	2
Shape	hexagon	2

C.2 Sequence scoring

For candidate k with autoregressive candidate tokens $t_1, \ldots, t_{L_k}$, the candidate score is the sequence-mean log-probability

$$s(k) = \frac{1}{L_k} \sum_{j=1}^{L_k} \log p(t_j \mid x_{\text{img}}, x_{\text{prompt}}, t_{<j}) \,. \tag{17}$$

Candidate top-1 accuracy is the constrained five-way decision

$$\hat{y}_{\text{cand}} = \arg\max_k s(k). \tag{18}$$

The frozen raw records retain mean score, sum score, token count, and candidate token IDs.

C.3 EVIDENCE MARGINS

The two scalar evidence margins are

$$M_{\text{mean}} = s(y) - \frac{1}{K-1}\sum_{k\neq y} s(k), \qquad M_{\text{best}} = s(y) - \max_{k\neq y} s(k). \tag{19}$$

For $K = 5$, the row-centered true-candidate score is linearly related to M_{mean}:

$$M_{\text{mean}} = \frac{5}{4}\left(s(y) - \frac{1}{5}\sum_{k} s(k)\right). \tag{20}$$

C.4 PROMPT AND CHAT-TEMPLATE CONTRACTS

Original scoring protocol. Table 2 and Figure 3 use the original Confirmation candidate scores for all three representative models, including InternVL. Candidate BA, endpoint margins, and redistribution are computed under this same protocol. The dedicated-prompt remeasurement is reported separately.

Request text and Qwen/LLaVA rendering. The original request contains one user message with the image followed by the question. The exact questions are:

```
What is the color of the target object?  Answer with exactly
one of:  red, yellow, green, blue, purple.

What is the shape of the target object?  Answer with exactly
one of:  circle, triangle, square, pentagon, hexagon.
```

No explicit system message is supplied by the request. The frozen LLaVA chat template inserts `You are a helpful assistant.`; the frozen Qwen template does not insert a default system message. Candidate strings are the lowercase labels exactly, with no added punctuation or `Answer:` prefix.

The image placeholder is `<|vision_start|><|image_pad|><|vision_end|>` before processor expansion. The user frame is `<|im_start|>user` followed by a newline, the image placeholder, the question, and `<|im_end|>`. The LLaVA assistant prefix is `<|im_start|>assistant` followed by a newline. For Qwen, `enable_thinking=False` yields the assistant prefix `<|im_start|>assistant` followed by an empty thinking block `<think>...</think>` and then the answer continuation. Candidate scoring inserts the candidate into the assistant message using `continue_final_message=True`; only candidate-token log-probabilities enter Eq. 17.

For the corresponding free-generation diagnostic, generation is capped at 8 new tokens. The parser applies `strip().lower()` and requires exact membership in the five frozen labels; candidate scoring does not depend on this free-generation parse.

Separate InternVL remeasurement. The dedicated-prompt InternVL remeasurement uses the following system prompt:

```
You are a visual classification function.  Follow the
required output format exactly.
```

and the user prompt

```
Classify the {axis} of the target object in the image.
Return exactly one lowercase label from this list:
{candidates}.  Do not add any other words or punctuation.
Answer:
```

Its free-generation diagnostic uses greedy decoding with temperature 0, top-$p = 1$, `max_new_tokens=24`, and seed 0; the conservative parser applies NFKC normalization, whitespace stripping, and case folding.

The main InternVL candidate-evidence and intervention analyses use the original question wording, whereas the dedicated-prompt remeasurement uses the alternative wording above. The main native-evidence and intervention cohorts remain separate. The different signs of the InternVL–Shape native changes in the evidence table and factorial decomposition are therefore not attributed to the dedicated prompt. The remeasurement also changes generation settings and parsing, so its differences from the original protocol are not attributed to parsing alone.

C.5 Candidate-accuracy diagnostics

Candidate top-1 BA measures whether continuous score changes alter the constrained five-way decision. Table 11 reports the complete original-contract trajectories for the representative checkpoints. Every cell contains 1,000 Confirmation examples from 40 groups. Qwen Shape L0 is 0.999, whereas its L2 value is 1.000. InternVL Shape L1 is 0.938 under the original contract, so its ceiling endpoint values do not imply a flat accuracy trajectory at every stress level.

Table 11: **Candidate BA by stress level under the original scoring protocol.** The endpoint values supply Table 2. The dedicated-prompt remeasurement is reported separately.

Axis	Level	Qwen3.5-2B	LLaVA-OV1.5-4B	InternVL3.5-2B
Color	L0	0.956	0.984	1.000
Color	L1	0.901	0.912	0.954
Color	L2	0.825	0.779	0.800
Color	L3	0.803	0.740	0.632
Color	L4	0.709	0.635	0.600
Shape	L0	0.999	1.000	1.000
Shape	L1	1.000	1.000	0.938
Shape	L2	1.000	1.000	1.000

Under the separate dedicated-prompt contract, InternVL3.5-2B obtains candidate BA of 1.000 at Color L0 and 0.600 at L4, with 1.000 at all three Shape levels. Its greedy-generation classification diagnostic shows the same endpoint pattern: parsed BA decreases from 1.000 to 0.600 for Color and remains 1.000 across Shape levels. These results use the dedicated prompt, generation settings, and conservative parser specified above, and are reported separately from the original-contract candidate-evidence analysis.

Agreement at selected endpoints does not make the two scoring contracts interchangeable. The dedicated-prompt series is retained as a diagnostic, not substituted into the original-contract margin or redistribution analyses.

C.6 Margin nonlinearity and competitor diagnostics

M_{mean} is a linear combination of candidate scores, whereas M_{best} contains a maximum. A change in the strongest wrong candidate can therefore change the latter contrast even when individual candidate-score effects have a simpler additive form. Both margins are reported without substituting one for the other after observing results.

The competitor-sensitivity analysis evaluates cross-family endpoints and local responses using stable-best and switching-best subsets and fixed wrong-candidate margins. Each named comparator k is held fixed, excluding examples with $y = k$. This differs from selecting one fixed rival separately for each example. Appendix F.6 reports the resulting diagnostics. A corresponding fixed-rival analysis of the later same-family four-corner states is not reported.

D Intervention Interfaces and Controls

D.1 Frozen visual intervention interface

The intervention state is written as $H = (Z, C)$, where Z is the frozen set of target visual-token indices and C its complement at the same interface. Table 12 reports the interfaces used for the same-

family, cross-family, and local-response analyses across all six model–attribute settings. Writes use native BF16 activations.

Table 12: **Intervention interfaces and target-token ranges.**

Model	Output hook	Captured state	Target-token count
Qwen3.5-2B	`model.visual.merger`	$(64, 2048)$	42–56
LLaVA-OV1.5-4B	`model.visual.merger`	$(81, 2560)$	49–72
InternVL3.5-2B	`mlp1`	$(1, 256, 2048)$	156–196

The ROI is defined by the normalized `fixed_target_centered_roi` rectangle stored in the frozen pair registry. Pixel boundaries use floor for left/top and ceil for right/bottom. Post-interface tokens with positive fractional overlap are eligible; donor and recipient tokens are matched by nearest relative token-center coordinates. Matches with normalized error above 0.12 are discarded, with a minimum retained proportion of 0.60. The final intervention uses the frozen pair-specific `recipient_indices`. Fractional overlap is used for eligibility and for the separate ROI reader, but selected activations are replaced in full rather than fractionally blended.

D.2 Same-family four-corner construction

For each matched family, the native clean and high-stress states are (Z_0, C_0) and (Z_H, C_H). The off-diagonal states are (Z_H, C_0) and (Z_0, C_H), formed by replacing only the frozen target slice. Candidate-scoring prompt, candidate set, model revision, precision, ROI, and all non-target activations are held fixed.

```
for matched_pair in evaluation_pairs:
    H0 = load_native_visual_state(clean)
    HH = load_native_visual_state(high_stress)
    Z0, C0 = split(H0, target_indices)
    ZH, CH = split(HH, target_indices)
    states = {
        "00": compose(Z0, C0),
        "H0": compose(ZH, C0),
        "0H": compose(Z0, CH),
        "HH": compose(ZH, CH),
    }
    for corner, H_write in states.items():
        score_all_five_candidates(model, visual_state=H_write)
```

The same-family analysis contains 100 matched pairs per model–attribute setting and reuses the recipient states from the cross-family analysis. These examples are distinct from the native-evidence samples in Table 2. The pooled-readout/native-score linkage therefore does not constitute a recoverability audit of the off-diagonal intervention states.

D.3 Endpoint replay and self-write checks

Before off-diagonal scoring, the native clean and high-stress endpoints are replayed through the intervention path. Across all six cells, 1,200 endpoint replays recover stored candidate mean/sum log-probabilities with maximum observed discrepancy zero under the frozen 10^{-8} tolerance. Candidate token counts and native tensor identities also match. The implementation also performs 60 self-write checks, for which selected and non-selected indices, source/destination hashes, shapes, and dtypes are recorded.

D.4 Cross-family donor construction and endpoint estimand

For a same-label donor–recipient dyad, let D_0, D_H denote the clean and high-stress donor states and R_0, R_H the corresponding states of a different recipient family. Let $W_i(D, R)$ denote the full recipient interface state obtained by replacing only the frozen target indices with the mapped donor

tokens D, leaving the recipient complement R unchanged. Raw donor activations are used without recipient-specific norm matching. For evidence measure $M \in \{M_{\text{best}}, M_{\text{mean}}\}$,

$$\tau^M_{D,i}(R) = M_i(W_i(D_0, R)) - M_i(W_i(D_H, R)), \tag{21}$$

$$\begin{aligned}\Delta\tau^M_{D,i} = &[M_i(W_i(D_0, R_H)) - M_i(W_i(D_H, R_H))] \\ &- [M_i(W_i(D_0, R_0)) - M_i(W_i(D_H, R_0))]\,.\end{aligned} \tag{22}$$

The reported estimate is the mean dyad contrast. High stress is Color L4 or Shape L2. Figure 5 and Table 15 use all 100 dyads per cell, with 40 donor groups and 40 recipient groups from the intervention cohort.

D.5 Finite-interval local response and direction controls

The frozen natural donor direction is

$$\delta_i = D_{0,i} - D_{H,i}, \qquad \alpha \in \{-0.10, -0.05, 0, 0.05, 0.10\}, \tag{23}$$

with perturbed state

$$h_{i,\alpha}(R) = W_i(D_{H,i} + \alpha\delta_i, R). \tag{24}$$

The addition is performed in float32 and cast back to the anchor's BF16 dtype before writing. $\alpha = 0$ is the raw stressed-donor endpoint. For each dyad and recipient state, the local response is the OLS slope with intercept,

$$\widehat{g}^M_i(R) = \frac{\sum_\alpha (\alpha - \bar{\alpha})[M_i(h_{i,\alpha}(R)) - \overline{M}_i(R)]}{\sum_\alpha (\alpha - \bar{\alpha})^2}, \qquad \bar{\alpha} = 0, \tag{25}$$

followed by

$$\Delta\widehat{g}^M_i = \widehat{g}^M_i(R_H) - \widehat{g}^M_i(R_0), \qquad \Delta\widehat{g}^M = \frac{1}{N}\sum_i \Delta\widehat{g}^M_i. \tag{26}$$

The slope is therefore estimated per dyad before aggregation.

Each model–attribute setting includes five random directions, one orthogonal direction, and the sign-reversed natural direction, all using the same anchor, alpha schedule, and target-norm protocol. Direction-seed repetitions are averaged within dyad. The sign-reversal sum is numerically zero to machine precision in all six settings.

D.6 Interpretation boundary

Endpoint replay, self-write, and write-identity checks establish that the specified frozen states were manipulated. They do not establish that every off-diagonal state is naturally occurring or identify a unique downstream attention, normalization, gating, or MLP mechanism. The main claim is therefore conditional visual-state influence at the frozen interface.

E Statistical Protocol

E.1 Independent units

The statistical unit is the frozen family/scaffold grouping associated with each analysis. Individual images, stress views of the same family, candidate answers, repeated forwards, reader seeds, and direction seeds are not treated as independent samples. Native-evidence conditions contain 1,000 samples in 40 scaffold groups; each same-family four-corner cell contains 100 matched pairs in 40 groups.

E.2 Paired cluster bootstrap

Unless otherwise specified, confidence intervals use 10,000 bootstrap resamples. Same-family four-corner resampling is over `scaffold_group_id`; each sampled group contributes all associated pairs, all four states, and all candidate scores. Point estimates use sample-weighted means and

Table 13: **Earlier-cohort permutation sensitivity analysis.** Holm-corrected conservative p-values from the separate 50-dyad cohort; these tests do not apply to Table 3.

Model	Axis	Endpoint	Full curve	Candidate weights
Qwen3.5-2B	Color	0.001550	0.034080	0.119049
Qwen3.5-2B	Shape	0.421436	0.095499	1.000000
LLaVA-OV1.5-4B	Color	0.243298	1.000000	0.110219
LLaVA-OV1.5-4B	Shape	0.000300	0.365356	1.000000
InternVL3.5-2B	Color	0.592144	0.365356	1.000000
InternVL3.5-2B	Shape	0.149599	1.000000	1.000000

intervals are the 2.5th–97.5th bootstrap percentiles. The same resampling weights are reused for related scalar quantities so that $\Delta M_{\text{native}} = A + B + I$ and $\Delta\tau = -I$ remain paired within each bootstrap draw. Stress-indexed native-evidence resampling preserves the corresponding paired clean–stress structure. Main-table intervals are pointwise and are not simultaneously adjusted across all model–attribute cells, margins, and controls.

E.3 Cross-family and local-response intervals

Cross-family and local-response analyses compute separate 10,000-resample donor-group and recipient-group bootstrap intervals. The sensitivity envelope uses the smaller lower endpoint and larger upper endpoint of the two intervals. Donor and recipient groups are paired one-to-one in this cohort. These envelopes are not asserted to provide the coverage of a general joint two-way cluster bootstrap. The same-family analysis resamples recipient scaffold groups directly, keeping all four corners paired.

E.4 Reader seeds

Linear fits are deterministic conditional on the fitting subset. MLP results average the five frozen seeded readers described in Appendix B; optimization seeds do not increase the number of independent visual groups.

E.5 Metric sensitivity and earlier-cohort permutation analysis

Both M_{best} and M_{mean} are reported for all six intervention cells. The frozen cross-family primary measure remains M_{best}; M_{mean} is complementary. Interval-crossing cells are retained and no evidence margin is substituted post hoc. The later same-family construction is a follow-up analysis, not an additional preregistered independent confirmation.

A supplementary permutation analysis uses an earlier development cohort with 50 dyads and 20 donor/recipient groups per cell, separate from the 100-pair intervention cohort used in the main tables. It performs 100,000 group-level sign permutations for donor and recipient clusters separately and uses $p_{\text{cons}} = \max(p_{\text{donor}}, p_{\text{recipient}})$. Within each analysis family, the six cell-wise p-values are Holm-corrected. These supplementary tests are not significance tests for Table 3.

The corresponding global conservative uncorrected p-values are 1.399986×10^{-4} for the endpoint statistic and 3.229968×10^{-3} for the full-curve statistic. Candidate-weight analysis uses a softmax normalization only as a diagnostic transformation; it is not interpreted as a calibrated categorical probability.

E.6 Negative results and stopping

Weak, opposite-sign, and interval-crossing results are retained. The final same-family experiment was not extended by changing sample size, evidence margin, intervention layer, or visual axis after inspecting its results. Repeated forwards and aliases of identical written tensors are not counted as additional observations.

F Additional Results and Analyses

F.1 Full stress-level readout and translation results

The machine-readable supplement contains all three readout protocols for the eight checkpoints at both H_{enc} and H_{post}, including balanced accuracy, per-class recall, group-bootstrap intervals, probe margins, confusion matrices, and all five MLP seeds. Figure 6 shows the complete H_{post} stress-level translation analysis corresponding to the maximum-stress summary in Figure 2.

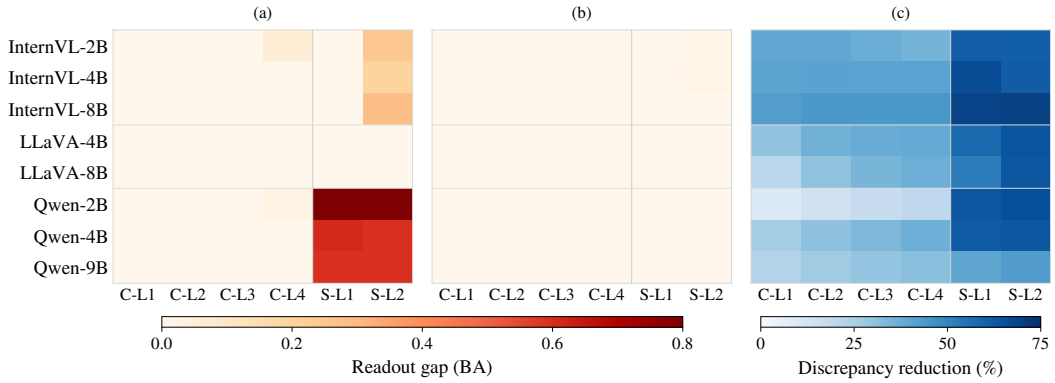

Figure 6: **Cross-condition representation shift across all stress levels.** (a) Original clean-readout mismatch G_{mis}, (b) residual mismatch G_{res} after translation, and (c) paired representation discrepancy reduction for Color L1–L4 and Shape L1–L2 at H_{post}.

The most extreme reader mismatch is Qwen3.5-2B Shape L2. Its global Fixed-L0 reader has per-class recalls $(0, 0, 1, 0, 0)$ for (circle, triangle, square, pentagon, hexagon), yielding balanced accuracy 0.2, while the Condition-Specific and Matched-Mixed readers both reach 1.0. Thus, exact chance-level macro accuracy can correspond to systematic class collapse rather than stochastic guessing.

F.2 Candidate-space matrices

For each representative model and stress level, the supplement stores the 5×5 matrix

$$\Delta S_{c,k} = \mathbb{E}[s^{(l)}(k) - s^{(0)}(k) \mid y = c],$$

together with its row-centered counterpart $\widetilde{\Delta S}$, total candidate-space displacement D, relative redistribution D_{redis}, and candidate top-1 confusion tables. Figure 7 shows the maximum-stress raw ΔS matrices. Row-centered matrices are used to compute D_{redis}.

Table 14: **Total and relative candidate-score displacement.** D and D_{redis} on the Confirmation endpoints under the original candidate-scoring protocol.

Model	**Axis**	D	D_{redis}
Qwen3.5-2B	Color	22.460	11.735
Qwen3.5-2B	Shape	2.640	2.425
LLaVA-OV1.5-4B	Color	13.425	9.993
LLaVA-OV1.5-4B	Shape	5.682	3.502
InternVL3.5-2B	Color	14.727	12.660
InternVL3.5-2B	Shape	2.659	2.408

F.3 Cross-family endpoint results

Table 15 reports the complete six-cell complementary cross-family endpoint analysis. Its estimand is distinct from the same-family $\Delta\tau$ in Table 3.

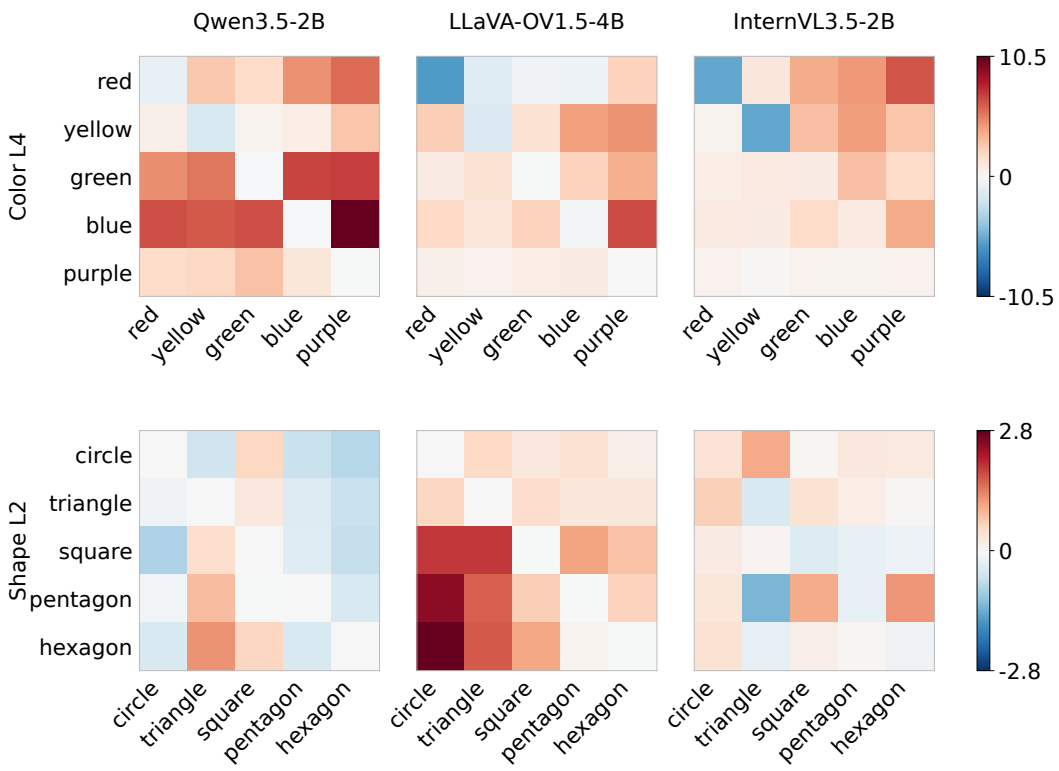

Figure 7: **Raw candidate-answer score changes at maximum stress.** Class-conditioned $\Delta S_{c,k}$ from paired L0, for all three representative models and both attributes. Common row shifts remain in these raw matrices and are removed for D_{redis}.

Table 15: **Cross-family recipient-state dependence.** Values are point estimates and sensitivity envelopes of the two pointwise 95% one-way group-bootstrap intervals (Appendix E).

Model	**Axis**	M_{best}	M_{mean}
Qwen3.5-2B	Color	$+0.678\,[+0.476, +0.901]$	$+0.794\,[+0.532, +1.085]$
Qwen3.5-2B	Shape	$+0.084\,[-0.024, +0.197]$	$+0.349\,[+0.204, +0.511]$
LLaVA-OV1.5-4B	Color	$+0.066\,[+0.025, +0.108]$	$+0.112\,[+0.076, +0.151]$
LLaVA-OV1.5-4B	Shape	$-0.215\,[-0.246, -0.184]$	$-0.304\,[-0.358, -0.252]$
InternVL3.5-2B	Color	$-0.054\,[-0.122, +0.012]$	$-0.025\,[-0.080, +0.029]$
InternVL3.5-2B	Shape	$-0.072\,[-0.172, +0.006]$	$-0.029\,[-0.117, +0.043]$

Qwen–Color and LLaVA–Shape retain opposite signs under both evidence margins. Qwen–Shape is resolved under M_{mean} but not M_{best}, while both InternVL intervals cross zero.

F.4 Finite-interval local responses

The Qwen–Color positive and LLaVA–Shape negative patterns persist under the local construction. Qwen–Shape is also positive under both margins, whereas LLaVA–Color and both InternVL cells remain unresolved under the reported intervals.

F.5 Direction controls

Table 17 reports the frozen amplitude-comparison controls for the local response using M_{best}. The random advantage is the dyad-wise quantity $|\Delta\hat{g}_{\text{natural}}| - \text{mean}_r|\Delta\hat{g}_{\text{random},r}|$; the orthogonal advantage is $|\Delta\hat{g}_{\text{natural}}| - |\Delta\hat{g}_{\text{orthogonal}}|$.

Table 16: **Finite-interval recipient-state response difference.** $\Delta\hat{g} = \hat{g}(R_H) - \hat{g}(R_0)$. Intervals retain the one-way bootstrap sensitivity envelopes described in Appendix E.

Model	Axis	M_{best}	M_{mean}
Qwen3.5-2B	Color	$+0.567\,[+0.252, +0.897]$	$+0.867\,[+0.447, +1.303]$
Qwen3.5-2B	Shape	$+0.208\,[+0.025, +0.391]$	$+0.505\,[+0.243, +0.782]$
LLaVA-OV1.5-4B	Color	$+0.083\,[-0.073, +0.242]$	$+0.005\,[-0.111, +0.128]$
LLaVA-OV1.5-4B	Shape	$-0.220\,[-0.292, -0.146]$	$-0.305\,[-0.411, -0.205]$
InternVL3.5-2B	Color	$+0.011\,[-0.261, +0.286]$	$+0.045\,[-0.156, +0.246]$
InternVL3.5-2B	Shape	$-0.094\,[-0.339, +0.161]$	$-0.119\,[-0.330, +0.102]$

Table 17: **Natural-direction amplitude advantage over control directions.** Intervals retain sensitivity envelopes of the frozen donor/recipient pointwise 95% group-bootstrap intervals. Panel (c) of Figure 5 visualizes these values.

Model	Axis	Random advantage	Orthogonal advantage
Qwen3.5-2B	Color	$+0.553\,[+0.309, +0.818]$	$+0.565\,[+0.292, +0.849]$
Qwen3.5-2B	Shape	$+0.073\,[-0.044, +0.195]$	$+0.109\,[-0.038, +0.255]$
LLaVA-OV1.5-4B	Color	$+0.151\,[+0.050, +0.259]$	$+0.198\,[+0.065, +0.325]$
LLaVA-OV1.5-4B	Shape	$+0.096\,[+0.041, +0.157]$	$+0.127\,[+0.058, +0.197]$
InternVL3.5-2B	Color	$+0.077\,[-0.080, +0.239]$	$+0.064\,[-0.110, +0.231]$
InternVL3.5-2B	Shape	$+0.150\,[+0.004, +0.296]$	$+0.144\,[-0.018, +0.311]$

The absolute values of the mean sign-reversal sums are below 10^{-15} in all six cells. Amplitude advantage is distinct from the signed recipient-state effect: for example, LLaVA–Color shows positive amplitude advantages while its signed local effect remains unresolved.

F.6 Competitor sensitivity of cross-family effects

The competitor diagnostics use the intervention cohort and distinguish stable from switching strongest wrong candidates. For LLaVA–Shape, all 100 dyads have a stable strongest wrong candidate for both the cross-family endpoint and local-response analyses. Their M_{best} effects remain negative, and all five named fixed-wrong comparisons also have negative reported intervals. Each fixed-wrong comparison excludes the 20 examples whose true label equals that comparator, leaving 80 dyads. This limits a competitor-switching explanation for the cross-family and local LLaVA–Shape pattern, without providing a new same-family audit.

InternVL–Color illustrates why the diagnostics are not interchangeable with the full-cohort result. Its stable-best endpoint subset (58 dyads) is $+0.103\,[+0.024, +0.184]$, whereas the switching-best subset (42 dyads) is $-0.271\,[-0.387, -0.161]$; the full-cohort interval crosses zero. InternVL–Shape has 100 stable-best dyads, yet both its endpoint and local full-cohort intervals remain unresolved. These subsets are conditional descriptions selected by observed competitor identities, not independent confirmations or evidence that a full-cohort effect is zero.

These fixed-wrong and stable-best diagnostics use cross-family and local-response scores. A per-example fixed-rival analysis and a new multiplicity-adjusted test family for the same-family four-corner results are not reported.

F.7 Heterogeneous and unresolved findings

The following outcomes are retained rather than collapsed into a binary "confirmed" label:

- Qwen–Shape same-family interaction is resolved under M_{mean} but not M_{best};
- InternVL–Color and InternVL–Shape same-family interactions are resolved under M_{best}, while their M_{mean} intervals cross zero;
- InternVL cross-family and local-response intervals cross zero under the reported estimators;

- LLaVA–Color has a small same-family effect while its signed local finite-interval response is unresolved.

These outcomes motivate the model–attribute-heterogeneity conclusion rather than a universal same-sign or same-magnitude interpretation.

F.8 Negative mechanism-localization results

We additionally attempted to localize recipient-state dependence to specific downstream submodules. These analyses did not establish a stable specific mechanism and are not part of the paper's primary three findings.

For Qwen3.5-2B and LLaVA-OneVision-1.5-4B, the later localization study replayed the frozen local-gain effects, tracked finite-interval tangent propagation through language blocks, and intervened on candidate token-mixing, normalization, MLP, and late-MLP components. Although geometric divergence was observed, no tested component simultaneously satisfied the required early-divergence, bidirectional-transfer, and direction-control criteria.

For InternVL3.5-2B, a Development analysis identified an early stable geometric-divergence candidate around block 7, but downstream transfer interventions did not establish robust bidirectional causation. The q-norm and k-norm inner interventions had intervals crossing zero, and the natural direction generally lacked the required specificity over generic controls.

These negative localization results constrain the interpretation of Section 4.3: the observed recipient-state dependence is a conditional-effect finding at the frozen visual interface, not identification of a unique downstream circuit.