QUANTIFYING CROSS-MODAL INTERACTION IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models map images, video and audio into a joint representation processed alongside text. Yet a modality affecting the output does not establish that the model uses it as the instruction requires. We therefore investigate when modality information is *functionally usable* by a language model. To this end, we introduce a functional geometry that compares inputs by the model's response rather than their representation, making it invariant to how the representation is parameterized. Within this geometry, we define the Functional Interaction Profile (FIP), which measures (i) whether task-relevant evidence influence the response, (ii) whether its effect depends on the question, and (iii) whether meaning impacts it more than nuisance. Experiments on three open models across four modalities confirm the profile's predicted invariances. Applied to these models, the profile reveals which evidence does not reach the response: no evaluated model perceives how events unfold in time, although the same questions are solved from a text description. A larger model still fails these questions, but its profile shows this evidence beginning to interact with the question before its accuracy changes. Accuracy signals whether a model fails, whereas the profile shows *why it fails* and *whether it is improving*. We thus propose that MLLM releases report the profile alongside accuracy, and release the controlled suite and evaluation harness to compute it. By shifting the evaluation from static outcomes to dynamic behavior, FIP can support the development of more transparent and robust multimodal models.

1 INTRODUCTION

Multimodal large language models (MLLMs) map images, video and audio into representations processed alongside text instructions(Liu et al., 2023; Li et al., 2024; Xu et al., 2025a; Cui et al., 2026). But does the language model actually use these representations to carry out the instruction?

In this paper, we introduce a functional geometry for studying when modality evidence is usable by a language model. The proposed functional geometry compares modal inputs not by their position in representation space but by how differently the model responds to them on a chosen observable, such as an answer margin, and is therefore invariant to invertible reparameterizations of the representation. Our perspective leads to three complementary properties that form the core of our Functional Interaction Profile (FIP): **accessibility** (whether task-relevant changes to the modality reach the response), **conditionality** (whether their effect depends on the question) and **selectivity** (whether answer-changing edits move the response more than answer-preserving nuisance). Four further coordinates, from text descriptions, interventions, layer readouts and repeated execution, show where evidence becomes usable and how consistently it affects the response.

The central requirement of question-dependent evidence use is captured by conditionality. We cross two modality states U_0, U_1 with two text states E_0, E_1 and evaluate a declared observable f:

$$\Gamma_f(U_1, E_1; U_0, E_0) = f(U_1, E_1) - f(U_1, E_0) - f(U_0, E_1) + f(U_0, E_0). \tag{1}$$

This classical factorial contrast (Fisher et al., 1966) vanishes for additive responses. Consider a red or blue object paired with "Is it red?" or "Is it blue?" (Fig. 1): neither input alone determines the answer. For the signed answer margin s, solving all four cells correctly with margin at least $m > 0$ requires $\Gamma_s \geq 4m$. Under smoothness assumptions, this contrast equals an integrated mixed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

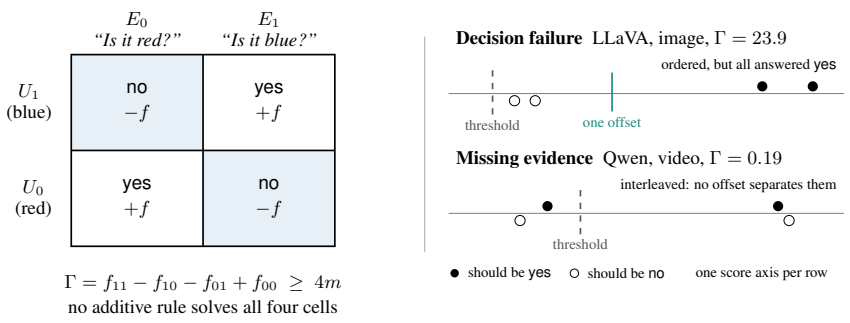

Figure 1: **The four-cell contrast and two kinds of failure.** Left: neither the modality nor the question alone determines the answer, so solving all four cells with margin m requires $\Gamma \geq 4m$ (Theorem 1). Right: two measured failures. In a decision failure the scores are correctly ordered but misplaced, and one offset fixes them; when evidence is missing, no offset can (Sec. 6.2). While accuracy counts both as the same failure, the four scores can tell them apart.

second derivative, so layer-wise sensitivity measures the same quantity locally. Conditionality is thus necessary but not sufficient: the four scores also separate decision failures, which a constant offset repairs, from missing evidence, where no threshold separates the required answers.

After validating the profile, we apply it to 5,012 evaluations of 3 open models on four-cell families spanning 4 modalities and 22 semantic dimensions. Conditionality becomes readable late, at 74–81% of depth, across models and modalities. At the output, modality rather than model determines accessibility and conditionality, while models differ in selectivity. Failure patterns also differ: for LLaVA-OneVision-7B (Li et al., 2024), one answer offset repairs 252 of 386 image failures, whereas for Qwen2.5-Omni-7B (Xu et al., 2025a), it repairs none of 334 video failures. Video and audio reveal a further limitation: event order is solved in 75% of families, yet none of 1,268 evaluations requiring events to unfold over time succeeds. Text descriptions of the same scenes solve all tested families, locating the limit in media perception. The larger Qwen3-Omni-30B-A3B (Xu et al., 2025b) achieves only 0–5% four-cell success, yet its conditionality on speed, repetition, simultaneity rises 9-18×, progress obscured by near-zero accuracy. In summary, our contributions are:

- **A functional geometry of multimodal computation.** We compare inputs through declared observables, establish reparameterization invariance, and connect cross-modal interaction to local response geometry and decision failure.
- **A diagnostic framework and reporting profile.** We introduce the seven-coordinate Functional Interaction Profile (FIP), centered on accessibility, conditionality and selectivity, and propose reporting it alongside accuracy for each modality and semantic dimension.
- **A controlled cross-model and cross-modal analysis.** We characterize how MLLMs use modality evidence, distinguish their failure modes, and identify semantic properties they fail to resolve on the controlled suite.

2 Related Work

The four-cell contrast is the interaction term of classical factorial designs (Fisher et al., 1966), generalized by functional ANOVA and variance-based sensitivity analysis (Hoeffding, 1992; Sobol', 1990); we do not claim this algebra as new. Multimodal interaction measures such as EMAP (Hessel & Lee, 2020), partial information decomposition (PID) (Williams & Beer, 2010; Liang et al., 2023; Xiu et al., 2026; Fang et al., 2026), and InterSHAP (Wenderoth et al., 2024) are distributional, whereas conditionality is defined on the response surface at declared inputs and can vary independently of PID (Proposition 14). Likewise, modality-contribution scores (Gat et al., 2021; Parcalabescu & Frank, 2023) can be large when conditionality is zero (Proposition 2). Our geometry is functional rather than representational (Klabunde et al., 2023); CKA (Kornblith et al., 2019), layer readouts (Belrose et al., 2023), and activation patching (Meng et al., 2022; Heimersheim & Nanda, 2024) enter only after fixing the behavioral contrast. Beyond aggregate accuracy (Ribeiro et al., 2020; Liang et al., 2022; Lee et al., 2024), POPE studies affirmative bias (Li et al., 2023) and TempCompass temporal weaknesses (Liu et al., 2024). We distinguish bias from missing evidence

using the same four scores and show that temporal failures persist at larger scale despite successful text descriptions. None of these lines of work reports together whether modality evidence reaches the answer, whether its effect depends on the question, and whether it follows meaning rather than nuisance; the FIP reports all three, and each line of work is discussed in Appendix A.

3 Functional Geometry of Multimodal Computation

We present the functional geometry of multimodal computation. We first define the observable of the model's response and the geometry it induces on inputs. We then introduce conditionality, the four-cell contrast that measures cross-modal interaction, highlight its key properties, explain how it separates misplaced decisions from missing evidence and how it differs from a distributional decomposition. We conclude by combining conditionality with accessibility and selectivity into FIP.

Model, Inputs, Observable. Let $q = (w_1, \dots, w_N)$ be a text sequence with embedding map $E_\theta(q) = (e_\theta(w_1), \dots, e_\theta(w_N)) \in (\mathbb{R}^d)^N$. For a non-text modality input $x \in \mathcal{X}$, an encoder and connector produce $U_\phi(x) = (u_1, \dots, u_{N_U}) \in (\mathbb{R}^d)^{N_U}$, and the language model receives $X = M(U, E)$, where M prefixes, interleaves, or merges the two sequences. Let $Z_{\Theta,\xi}(X)$ denote the outputs and internal traces computed from X under execution variability ξ, where $\Theta = (\phi, \theta)$. The computation induces a stochastic kernel $K_\Theta(\cdot \mid X) = \mathrm{Law}_\xi[Z_{\Theta,\xi}(X)]$, of which deterministic inference is the point-mass special case. For a bounded or integrable observable $\psi : \mathcal{Z} \to \mathbb{R}^k$,

$$f_{\Theta,\psi}(X) = \mathbb{E}_{Z \sim K_\Theta(\cdot|X)}[\psi(Z)] \tag{2}$$

is the the response. Since M is fixed, we write $f(U, E) = f(M(U, E))$. The observable may be an answer-probability vector, a signed answer margin, which is the observable of our experiments, or a residual-stream state. Every identity below holds componentwise for a vector-valued observable; scalar summaries and the unbiased four-cell estimator under stochastic execution are given in Appendices G.3 and G.4. We suppress fixed subscripts. Every geometric and conditionality claim below is relative to this declared observable and to the modality and text inputs being compared.

3.1 Functional Geometry at Two Scales

Definition 1 (Functional pseudodistance)**.** *Let D be a metric on output distributions, and let $W \succeq 0$ define $\|v\|_W^2 = v^\top W v$ on the observable of Eq. 2. Their pullbacks through the model are*

$$d_{\Theta,D}(X, X') = D(K_\Theta(\cdot \mid X), K_\Theta(\cdot \mid X')) \quad \textit{and} \quad d_{f,W}(X, X') = \|f(X) - f(X')\|_W. \tag{3}$$

This pseudodistance compares inputs through their induced model behavior rather than their representation-space location. It is a pseudometric because distinct inputs may induce identical responses, and a metric on the corresponding quotient space (Proposition 5, Appendix B.1). Its degeneracy thus defines the behaviorally effective input space.

Proposition 1 (Input-coordinate invariance)**.** *Let $h : \mathcal{X} \to \widetilde{\mathcal{X}}$ be bijective and define $\widetilde{K}(\cdot \mid \widetilde{X}) = K(\cdot \mid h^{-1}(\widetilde{X}))$. Then $\widetilde{d}_{\Theta,D}(h(X), h(X')) = d_{\Theta,D}(X, X')$, and the same statement holds for every response-surface quantity below after composing f with h^{-1}.*

See Appendix B.2 for the proof. This invariance allows quantities defined on the response, including conditionality, to be read as a property of the model and not of the chosen input coordinates as distance to a vocabulary-embedding table, which is instead a coordinate-dependent description.

Local form. On a fixed-dimensional smooth chart, flatten the merged input to $x \in \mathbb{R}^p$ and suppose $f : \mathbb{R}^p \to \mathbb{R}^k$ is differentiable with Jacobian $J_f(x)$. For $W \succ 0$, the pullback metric

$$G_f(x) = J_f(x)^\top W J_f(x) \tag{4}$$

satisfies $\|f(x + \delta) - f(x)\|_W^2 = \delta^\top G_f(x)\delta + o(\|\delta\|^2)$. The pullback metric allows us to measure, for each direction in which the input can move, how strongly that direction changes the response. Directions in its kernel change nothing to first order (Proposition 6, Appendix B.3), and its information-geometric form, the pullback Fisher metric, is developed in Appendix G.1. Both scales are defined by the response alone and therefore share the invariance of Proposition 1. Neither the global distance nor the local metric establishes that the effect of the modality depends on the text; that requires a mixed coordinate, which we call conditionality.

3.2 CONDITIONALITY: THE MIXED FUNCTIONAL COORDINATE

Fix two modality states U_0, U_1 and two text states E_0, E_1, and take (U_0, E_0) as the anchor cell.

Definition 2 (Conditionality)**.** *The conditionality of f is the rectangular contrast*

$$\Gamma_f = [f(U_1, E_1) - f(U_0, E_1)] - [f(U_1, E_0) - f(U_0, E_0)] \tag{5}$$

Cross-modal interaction is the dependence of a modality's effect on text, measured by the conditionality Γ_f. It vanishes when the modality has the same effect under both texts, regardless of magnitude, and is nonzero only when that effect depends on the text. Algebraically, Γ_f is the classical interaction contrast of a 2×2 factorial design (Fisher et al., 1966), applied to the model's response surface. This is precisely the structure required by the checkerboard task: diagonal pairs (U_0, E_0) and (U_1, E_1) require yes," whereas off-diagonal pairs require no" (Fig. 1, left).

Coding states 0 as -1 and 1 as $+1$, the four responses $f_{ue} = f(U_u, E_e)$ decompose as

$$f_{ue} = \mu + a x_u + b x_e + d x_u x_e, \qquad d = \Gamma_f / 4, \tag{6}$$

where μ, a, b, and d are the baseline, modality effect, text effect, and interaction, respectively. These terms form a lossless parameterization of the response surface, while Γ_f isolates its departure from additivity. Up to scale and orientation, it is the unique linear contrast eliminating the baseline and both main effects (Proposition 8) and is the second-order term of the higher-order factorial decomposition in Appendix G.2. Proofs are in Appendix C.

Proposition 2 (Additive separability, bias invariance)**.** *Fix anchors U_0, E_0. (i) $\Gamma_f(U, E; U_0, E_0) = 0$ for all U, E if and only if $f(U, E) = \alpha(U) + \beta(E) + c$ for some functions α, β and constant c. (ii) For any such α, β, c, the response $\widetilde{f} = f + \alpha(U) + \beta(E) + c$ has $\Gamma_{\widetilde{f}} = \Gamma_f$.*

By (i), strong modality influence need not imply conditionality: an additive modality effect may be arbitrarily large while remaining independent of the instruction. Thus, showing that a modality affects the output does not establish task-relevant use. Moreover, if $\|\Gamma_f\| \leq \eta$ on an anchored domain, then f lies within η of an additive surface there (Proposition 9, Appendix C.4), so small Γ_f implies near-additivity. By (ii), additive answer-margin offsets cancel from Γ_f, even if they alter thresholded decisions. This guarantee does not extend to nonlinear or cell-specific recalibration.

Conditionality also joins the two scales of Sec. 3.1. If f is twice continuously differentiable on the rectangle joining the anchors, then with $\Delta U = U_1 - U_0$ and $\Delta E = E_1 - E_0$,

$$\Gamma_f = \int_0^1 \int_0^1 D^2_{UE} f(U_0 + \sigma \Delta U, E_0 + \tau \Delta E)[\Delta U, \Delta E] \, d\tau \, d\sigma \tag{7}$$

(Theorem 2, Appendix C.8). Finite conditionality and the mixed curvature of the response surface are therefore one quantity at two scales, so layer-wise mixed sensitivity is a local account of the same contrast rather than a separate diagnostic. The pullback metric of Eq. 4 records only first-order sensitivity: with J_U and J_E its Jacobian blocks for the modality and text inputs, the cross-block $J_U^\top W J_E$ is not a substitute for this second-order quantity, because it can be nonzero for additive responses (Proposition 10, Appendix C.9). Let $s(U, E)$ be a signed answer margin, positive for "yes," on a checkerboard task.

Theorem 1 (Positive-margin matching requires conditionality)**.** *If all four cells are correct with margin at least $m > 0$, that is, $s(U_0, E_0), s(U_1, E_1) \geq m$ and $s(U_0, E_1), s(U_1, E_0) \leq -m$, then $\Gamma_s(U_1, E_1; U_0, E_0) \geq 4m$.*

By Proposition 2, no additive response can solve such a family with positive margin. The converse is false: large conditionality does not guarantee correct signs, calibration, or robustness (see Sec. 3.3).

3.3 CONDITIONALITY AND DECISION FAILURE

Suppose the cells that require "yes" receive the larger scores, but all four scores are positive. Every decision is then "yes," although the scores separate the two answer groups. In the factorial form of Eq. 6 for the answer margin s, this is a large answer offset μ shared by all four cells, while a and b are the modality and question main effects and $d = \Gamma_s / 4$ is the conditionality term.

Proposition 3 (Conditionality and threshold feasibility)**.** *For a fixed checkerboard rectangle, a constant threshold yields strictly correct decisions in every cell exactly when* $\min(s_{00}, s_{11}) > \max(s_{01}, s_{10})$. *The largest worst-cell margin that a constant offset can achieve is* $d - \max(|a|, |b|)$.

See Appendix C.7 for the proof. Failure therefore reflects either insufficient conditionality to overcome the main effects, or a misplaced threshold despite sufficient separation. The former is a response failure; the latter is a decision failure correctable by an offset without changing Γ_s (Proposition 2(ii)). This is the distinction of Fig. 1.

Conditionality is not a distributional decomposition. Eq. 5 is a property of a response surface over declared interventions, whereas partial information decomposition (PID) (Williams & Beer, 2010; Liang et al., 2023; Fang et al., 2026) is a property of a joint distribution, target variable, redundancy definition and estimator. Neither determines the other: on a matching task, $\Gamma_s = 4$ under every input distribution while the synergy of minimum-mutual-information PID vanishes as one input becomes nearly deterministic, and rescaling the score changes Γ_s but no PID atom (Proposition 14). The two are complementary: PID describes how information is allocated under a distribution, whereas conditionality describes the counterfactual non-additivity of a chosen model response.

3.4 Accessibility, Conditionality, Selectivity, and the Reported Profile

We measure three properties for reliable modality use: accessibility, conditionality and selectivity.

Accessibility verifies whether task-relevant evidence reaches the response at all. For an answer-changing transformation T_{sem},

$$a_i = \text{JS}(P(\cdot \mid U(x_i), E(q_i)),\ P(\cdot \mid U(T_{\text{sem}} x_i), E(q_i))), \tag{8}$$

where P is the model's distribution over the candidate answers and JS is the Jensen–Shannon divergence. Saturated distributions can make a_i small even when the scores remain ordered, so accessibility is reported with the scores themselves; its local form is developed in Appendix D.1.

Conditionality is $\Gamma_s = s_{11} - s_{10} - s_{01} + s_{00}$ on the four yes-minus-no scores, reported with its sign and the four scores: the magnitude measures the strength of the contrast, while its direction and the main effects determine whether the response supports the task (its effect depends on the question).

Selectivity assesses whether sensitivity follows meaning rather than merely increasing. Defining n_i as in Eq. 8 for an answer-preserving transformation T_{nuis}, with semantic state and question matched,

$$R_{\text{sel}} = \frac{\text{median}_i\, a_i}{\text{median}_i\, n_i + \varepsilon}, \tag{9}$$

reported with its numerator, denominator and floor, since a near-zero nuisance response can produce a large ratio without strong semantic sensitivity.

Proposition 4 (The three coordinates do not collapse)**.** *Accessibility can be arbitrarily large while conditionality is zero; conditionality can be nonzero while selectivity is arbitrarily poor; and selectivity can be arbitrarily large while conditionality is zero.*

See Appendix D.2 for constructions. A scalar measure cannot capture how a model uses a modality.

Functional Interaction Profile (FIP). A profile reports these three coordinates together with four that extend the same comparison, $\text{FIP} = (A_{\text{sem}}, C_{\text{ctx}}, R_{\text{sel}}, G_{\text{surr}}, C_{\text{med}}, L_{\text{onset}}, V_{\text{exec}})$, gap to a task-sufficient text surrogate, recovery under internal patching, the depth at which a fixed readout exposes conditionality, and variation under repeated execution. Accessibility and conditionality enter as $A_{\text{sem}} = \text{median}_i\, a_i / S_f$ and $C_{\text{ctx}} = \text{median}_i\, |\Gamma_i| / S_f$, where S_f is the median absolute clean answer margin, with each normalized value reported alongside its raw value. The profile has no aggregate score, because these properties need not share an ordering (Proposition 17). The reporting contract, the role of each coordinate, and required qualifications are given in Appendix G.5.

4 Experimental Design

We instantiate the theory with controlled four-cell families and compute FIP on the same inputs.

Controlled Families. Each family crosses two modality states with two question predicates: diagonal cells require "yes" and off-diagonal cells "no", so neither factor alone determines the answer.

For example, clips showing an object once or twice are crossed with questions asking whether it appears exactly once or exactly twice. Images use CLEVR-based renders (Johnson et al., 2016), videos use eight rendered frames without resampling, environmental audio uses ESC-50 mixtures (Piczak, 2015), and speech uses transcript-certified LibriSpeech recordings (Panayotov et al., 2015). The suite spans 22 semantic dimensions, reported separately; each family includes answer-changing semantic variants for accessibility, answer-preserving nuisance variants for selectivity, task-sufficient text surrogates, paraphrases and modality nulls (App. H).

Models and Panel. We evaluate frozen LLaVA-OneVision-7B (Li et al., 2024) on images and Qwen2.5-Omni-7B (Xu et al., 2025a) and MiniCPM-o 4.5 (Cui et al., 2026) on all four modalities: nine model–modality slices and 5,012 evaluations covering 4,000 distinct families. Cross-model comparisons use shared families (Tab. 2). Held-out Qwen3-Omni-30B-A3B (Xu et al., 2025b) and Phi-4-multimodal (Abouelenin et al., 2025) test generalization, with solvable live controls checking suite compatibility (App. I.13). We also evaluate six natural benchmarks (App. I.11).

Observable and Outcomes. Answers use length-normalized teacher-forced log probabilities:

$$S(a \mid X) = \frac{1}{T} \sum_{t=1}^{T} \log p_\theta(a_t \mid X, a_{<t}), \qquad s = S(\text{yes}) - S(\text{no}). \tag{10}$$

The yes-minus-no orientation is fixed across cells. Conditionality is $\Gamma = \Gamma_s = s_{11} - s_{10} - s_{01} + s_{00}$; both s and Γ are in nats. Accessibility and selectivity use the candidate softmax of S. A family is solved only when all four canonical answers are correct. A failed family is a *decision failure* if some constant score offset makes all four correct (Proposition 3). Separately, we test repair with a shared offset per slice, fitted on calibration families and evaluated on disjoint families (App. I.9).

Internal Measurements. The same score is also read inside the network. The layer readout applies the model's own final normalization and output projection to the hidden state after each layer, as if the network ended there, and so gives Γ at every depth. At a given layer, the four hidden states of a family, one per cell, decompose exactly into a mean, a modality effect, a question effect and a conditionality component (Proposition 15). To test whether the conditionality at a layer drives the answer, an intervention erases this component, or transplants another family's, while leaving the other three unchanged, at layer 14, about halfway through the network.

Inference. The semantic family is the sampling unit: cells and variants stay together across splits and are not independent samples. Intervals use family-clustered bootstrap resampling; fitted quantities are evaluated on disjoint families (App. H). Held-out prediction criteria and the layer-14 intervention were specified before their analyses (Appendices I.13 and I.6).

5 Validating the Functional Interaction Profile

We first establish that the Functional Interaction Profile behaves as Sec. 3 predicts: whether conditionality is invariant to reparameterizations of the representation and to answer biases (Propositions 1 and 2), whether finite comparisons separate meaning from nuisance where the local metric of Eq. 4 does not, whether the coordinates carry information beyond accuracy and distributional synergy (Proposition 14), and whether the measurements are stable.

Conditionality is invariant where the representation is not. Subtracting the frozen answer-bias offset from all 5,012 evaluations changes 840 four-cell decisions but changes Γ by at most 10^{-15} (App. I.3), as Proposition 2(ii) requires: a shift applied equally to all four cells cancels from the contrast. Representation summaries behave the opposite way: centering, whitening or norm-matching the same modality tokens changes their distance to the nearest text-token embedding by a median factor of 3,118 (Tab. 9) without changing a single response, as Proposition 1 permits. The response, not the representation, is therefore the object to measure.

Gradients cannot tell meaning from nuisance; finite comparisons can. With the response as the object, the next question is how to probe it: through its gradient or through finite comparisons. On 136 families, the local metric of Eq. 4, which weighs each input direction by how strongly it moves the response, separates both variants from random displacements, yet ranks the answer-changing variant above the answer-preserving one in only 39% of them, below chance (Tab. 8). Selectivity, the ratio of the semantic to the nuisance response change (Eq. 9), orders them correctly in

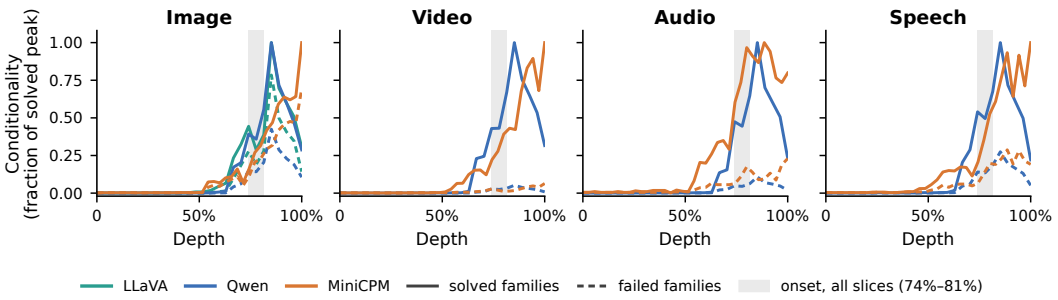

Figure 2: **Interaction forms late, at the same depth for every modality.** Median layer-wise $|\Gamma|$ under a fixed readout, for solved (solid) and failed (dashed) families, as a fraction of the solved families' peak in each slice. The band marks where median conditionality first reaches a quarter of its peak: 74–81% of depth in all nine slices. Failed image families reach much of the conditionality that solved ones do; failed video families reach almost none.

every slice: the median answer-changing variant moves the response more than the median answer-preserving one (Tab. 19). Telling meaning from nuisance therefore requires a finite comparison, which a gradient-based probe over modality tokens cannot provide.

The profile separates models that accuracy ties, and predicts success where synergy does not. Finite comparisons are the right probe, but they must also say more than accuracy and distributional synergy already do. MiniCPM solves 48.4% of both its image and its speech families, yet image has twice the raw conditionality and speech a far larger surrogate gap, the response difference from a text description of the same content (Tab. 17). On 147 families rescored with reworded questions, those with above-median $|\Gamma|$ succeed 36 percentage points more often than the rest in all five slices tested, as Theorem 1 anticipates, since a correct family requires $\Gamma_s \geq 4m$. Synergy from the learned PID estimator of Liang et al. (2023) makes a one-point difference (Tab. 10), and accuracy on the original questions remains the strongest single predictor. The profile thus separates models that accuracy ties, and conditionality carries predictive information that synergy lacks.

The measurements survive rewording, answer reordering and repeated execution. These model differences are meaningful only if the coordinates do not move incidentally, so we vary question wording, answer order, and execution. Across 784 source–paraphrase pairs, signed conditionality keeps a rank correlation of 0.99 under rewording, 0.81 in video, the least stable modality (Tab. 13), and swapping 'yes" and 'no" preserves every sign (App. I.5). Repeated executions show no variability, so V_{exec} sits at its numerical floor (Tab. 15), and mediation is reported with corruption baselines (Tab. 16). The final layer readout reproduces output conditionality with correlation above 0.999 (App. I.5), so the measurements are stable and conditionality can be followed through the network.

6 What Functional Geometry Reveals About MLLMs

We use the profile to characterize MLLMs: where cross-modal interaction forms inside the network, how the modality shapes the evidence that reaches the answer and how models fail, and which properties no evaluated model perceives. The findings describe this panel and suite; counts are exact, and every estimated median, ratio and correlation has a family-bootstrap interval in Tab. 7.

6.1 Where Cross-Modal Interaction Forms

Interaction forms late, and at the same depth for every modality. The layer readout gives conditionality at every depth. Median conditionality first reaches a quarter of its peak, the depth we call its onset, at 74–81% of depth in all nine slices (Fig. 2; Tab. 15), and neither the modality nor the model moves this depth, in agreement with layer-wise PID analyses of vision–language and omni-modal models (Xiu et al., 2026; Fang et al., 2026). Because the readout passes through the model's own answer head, which under-reads early layers (Belrose et al., 2023), onset marks where interaction becomes readable as an answer, not necessarily where it is first computed. The curves locate interaction rather than its cause: the layer-14 intervention changes no decisions (Tab. 14), and later interventions change decisions without selectivity (Fig. 7).

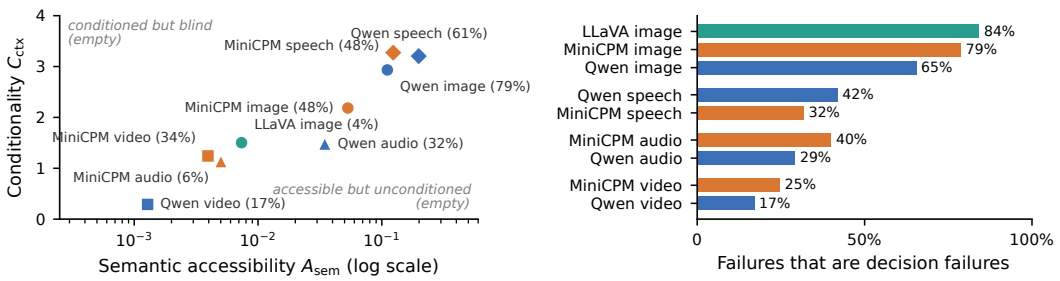

Figure 3: **The modality sets both how evidence is used and how models fail.** Left: Accessibility against conditionality for every slice, with four-cell success in parentheses. The two omni-modal models occupy nearly the same position for each modality, and neither empty quadrant is populated. Right: Share of failed families that some constant offset would make fully correct. Image failures are mostly decision failures whereas video failures mostly miss the required evidence.

The modality changes how much interaction forms, not where. With the depth shared, we compare how much conditionality failed families build relative to solved ones. Failed image families reach 42–78% of the solved families' peak, failed video families only 5–6%, with speech and audio in between (Tab. 15). A failed image family therefore reaches the decision with its question-dependent evidence in place, whereas a failed video family never assembles it.

6.2 How the Modality Shapes Evidence Use

The modality, not the model, determines how evidence is used. At the output, Fig. 3 (left) places each slice by its accessibility, how far an answer-changing variant moves the answer distribution, and its conditionality. In both omni-modal models, speech and image are accessible and strongly question-dependent, while video and environmental audio sit well below them. Within a modality the two models differ in accessibility by at most a factor of 7, whereas across modalities accessibility spans a factor of 150 (App. I.7), though part of this spread reflects each modality's own tasks. The slices therefore group by modality, not by model, and the image–video split seen inside the network reappears at the output. The two coordinates also move together: although they are logically independent (Proposition 4), no slice is question-dependent without being accessible, and their rank correlation across the nine slices is 0.95.

Within a modality, models differ in how selectively they read evidence. If the modality fixes how much evidence reaches the answer, what distinguishes the models is selectivity. Selectivity follows the modality ordering over five orders of magnitude, from 10^4–10^5 for image to 1.7 for Qwen video, which barely tells meaning from nuisance (Tab. 19). Within video, audio and speech, however, the two omni-modal models differ in selectivity by 17 to 52 times, although accessibility differs by at most 7 times: they receive the same evidence but read it with different precision.

Misplaced decisions dominate image failures and decline through speech and audio to video. Proposition 3 separates failures into missing conditionality and misplaced thresholds. Overall, 40% of 3,094 failures are threshold failures (App. I.9), decreasing from 65–84% for image to 32–42% for speech, 29–40% for audio, and 17–25% for video (Fig. 3, right). Accordingly, a calibrated offset rescues 252 of 386 LLaVA image failures but none of 334 Qwen video failures (Tabs. 20–23), without changing conditionality. Thus, image failures often reflect miscalibration, whereas video failures primarily reflect missing evidence.

6.3 What Current MLLMs Cannot Perceive

How events unfold in time is never perceived. These are the dimensions where evidence is missing in every model. Speed, duration, repetition, rotation, and whether two events are simultaneous or sequential are never solved: 0 of 1,268 evaluations over 1,133 distinct families succeed, in any video or audio slice (Fig. 4, left). Yet in the same modalities, event presence is solved in 64% of families and event order in 75%, so success forms a cliff rather than a gradient: 20 of the 55 slice–dimension cells are at exactly zero and 15 exceed 75% (Tab. 26).

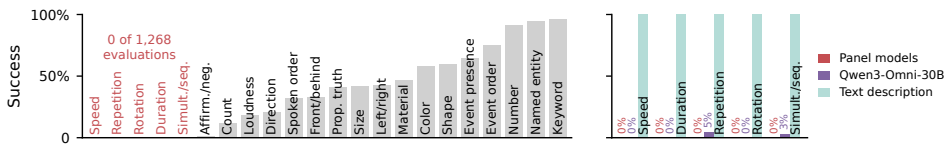

Figure 4: **Models track what happens and in which order, but not how it unfolds in time.** Left: Four-cell success for all 22 semantic dimensions, pooled over the slices in which each is evaluated. The five red dimensions are solved in none of 1,268 video and audio evaluations. Right: On those five, the panel models and the held-out Qwen3-Omni-30B-A3B reach at most 5% from the modality whereas a text description of the same scene is solved every time.

Almost no family with conditionality below one is solved. Theorem 1 explains why: solving a family requires conditionality of at least 4x its margin, and the median $|\Gamma|$ of the five empty dimensions lies between 0.4 and 1.2. Across the whole panel, low conditionality is close to disqualifying: of 1,132 evaluations with $|\Gamma| < 1$, one is solved, and the smallest conditionality among the 1,918 successes is 0.94 (Tab. 27), the empirical form of this bound. The one exception confirms the distinction between the two kinds of failure: MiniCPM solves no video repetition family yet keeps a median $|\Gamma|$ of 9.7 there, a decision failure, whereas its video rotation stays at 0.8 (Tab. 22).

The limit is perceiving the modality, not the question, the frames or model size. Low conditionality could stem from an unanswerable question, an input artifact or insufficient scale; we test each in turn. Replacing the modality with its text surrogate, a description of the same scene, with questions, labels and scoring fixed, solves 192 of 192 families for both Qwen2.5-Omni and Qwen3-Omni (Tab. 29). MiniCPM-o sees each of the eight frames as a separate image, and audio has no frames, yet both fail (App. H.2). The held-out Qwen3-Omni-30B-A3B, with 4.3 times the parameters, stays at 0–5% on the five dimensions while solving its live controls in 72–100% of families (Fig. 4, right). The limit lies in *perceiving* the media. Expanding tests beyond the Qwen family requires a model that passes live controls and Phi-4-multimodal cannot yet do it (Tab. 30).

A larger model begins to perceive these properties before its accuracy shows it. The larger model is not unchanged, however. On speed, repetition and simultaneity, Qwen2.5-Omni and Qwen3-Omni both stay at or near zero accuracy, while median $|\Gamma|$ rises from 0.3 to 2.9, from 0.6 to 4.8 and from 0.4 to 7.9 (Tab. 29). These dimensions have moved from missing evidence towards evidence the decision rule does not yet use, and our prediction, fixed in advance, that conditionality would stay below 2 failed on exactly these three (App. I.13). Conditionality thus records progress that a benchmark tracking these tasks across releases would miss.

6.4 What an MLLM Release Should Report

None of these findings is visible in accuracy, and across 36,000 model–example evaluations on six natural benchmarks, aggregate accuracy hides the same kind of change: on Clotho-AQA, mean-token audio leaves Qwen's accuracy at 31% but changes 46% of its answers (Tab. 25). We therefore propose that MLLM releases report the FIP for each supported modality and semantic dimension of a shared controlled suite, together with four-cell success and the share of decision failures, which the same four scores determine (Proposition 3). Accessibility, signed conditionality and selectivity need only the candidate scores a release already computes, and conditionality needs no gold labels, so an unsolved checkerboard task can be flagged before annotation; our harness produces them for 400 model–family evaluations in under fifteen minutes on one H100 (App. I.14). Open-weight releases can add onset and mediation.

7 Conclusion

We characterized when modality information is usable by a language model through the function the model computes. Applied to current MLLMs, the four-cell contrast shows that cross-modal interaction forms late and at the same depth in every modality; that the modality, far more than the model, sets how much forms and whether a failure is a misplaced decision or missing evidence, while the models differ in how selectively they read that evidence; and that how events unfold in time remains beyond these models, although text makes it solvable and a newer model's responses begin to encode it before its accuracy moves. Because accuracy shows none of this, we propose that MLLM releases report the Functional Interaction Profile alongside it (App. I.15).

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT and Claude, to assist with language editing, grammar correction, and improving the clarity and presentation of the manuscript. We also used generative AI tools to assist with some experimental implementation and execution after verifying the proposed procedures and generated code. All AI-assisted experimental components were reviewed and tested by the authors, and the resulting outputs were checked for correctness. Generative AI tools were not used to determine the scientific conclusions of the work. We take responsibility for the final content of this work, including text, claims, code, experiments, and artifacts produced with the aid of generative AI.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Wanlong Fang, Tianle Zhang, Wen Tao, and Alvin Chan. Towards understanding modality interaction in multimodal language models via partial information decomposition. *arXiv preprint arXiv:2606.00959*, 2026.

Ronald Aylmer Fisher, Ronald Aylmer Fisher, Statistiker Genetiker, Ronald Aylmer Fisher, Statistician Genetician, Great Britain, Ronald Aylmer Fisher, and Statisticien Généticien. *The design of experiments*, volume 21. Oliver and Boyd Edinburgh, 1966.

Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34:21630–21643, 2021.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 861–877, 2020.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 308–334. Springer, 1992.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. *arXiv preprint arXiv:1902.09506*, 2019.

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. *arXiv preprint arXiv:1612.06890*, 2016.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *arXiv preprint arXiv:2305.06329*, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Tony Lee, Haoqin Tu, Chi H Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin S Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, et al. Vhelm: A holistic evaluation of vision language models. *Advances in Neural Information Processing Systems*, 37:140632–140666, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. *arXiv preprint arXiv:1804.00320*, 2018.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard J. Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27351–27393, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-AQA: A crowdsourced dataset for audio question answering. *arXiv preprint arXiv:2204.09634*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *arXiv preprint arXiv:2202.05262*, 2022.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Letitia Parcalabescu and Anette Frank. Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4032–4059, 2023.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, 2015.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4902–4912, 2020.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

Il'ya Meerovich Sobol'. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118, 1990.

Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Measuring cross-modal interactions in multimodal models. *arXiv preprint arXiv:2412.15828*, 2024.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. *arXiv preprint arXiv:2105.08276*, 2021.

Lixin Xiu, Xufang Luo, and Hideki Nakayama. A comprehensive information-decomposition analysis of large vision-language models. *arXiv preprint arXiv:2603.29676*, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023.

A Extended Related Work

Quantifying multimodal interaction. The four-cell contrast is the interaction term of classical factorial designs (Fisher et al., 1966), which functional ANOVA and variance-based sensitivity analysis generalize (Hoeffding, 1992; Sobol', 1990); we do not claim the algebraic contrast as new. For multimodal models, EMAP tests for cross-modal interaction by projecting predictions onto additive functions (Hessel & Lee, 2020), partial information decomposition separates redundant, unique and synergistic information (Williams & Beer, 2010; Liang et al., 2023), including layer by layer in vision–language and omni-modal models (Xiu et al., 2026; Fang et al., 2026), and InterSHAP applies Shapley interaction indices (Wenderoth et al., 2024). These quantities are defined over a data distribution. Ours is defined on the response surface at declared inputs, and Proposition 14 shows that it and PID can vary independently. Modality-contribution scores (Gat et al., 2021; Parcalabescu & Frank, 2023) can be large while conditionality is zero (Proposition 2).

Representations and internal analyses. CKA and related methods compare internal representations (Kornblith et al., 2019), a survey distinguishes representational from functional similarity (Klabunde et al., 2023), and information geometry ties local structure to a predictive distribution (Amari, 1998). Layer readouts expose intermediate predictions but under-read early layers (Belrose et al., 2023), and activation patching provides interventional evidence whose interpretation depends on the corruption, metric and controls (Meng et al., 2022; Zhang & Nanda, 2023; Heimersheim & Nanda, 2024). Our geometry is functional in the sense of Klabunde et al. (2023): it is invariant to reparameterizations of the representation, and layer readouts and patching enter only after the behavioral contrast is fixed.

Evaluation beyond aggregate accuracy. Behavioral testing and holistic benchmarks broaden evaluation beyond a single score (Ribeiro et al., 2020; Liang et al., 2022; Lee et al., 2024), and controlled suites supply the semantic structure that makes interventions interpretable (Johnson et al., 2016; Liu et al., 2024). Two findings bear directly on ours. Yes/no prompting induces a strong affirmative bias, which POPE diagnoses for object hallucination (Li et al., 2023), and video language models struggle with temporal properties (Liu et al., 2024). We separate such a bias from missing evidence using the same four scores, and show that the temporal failures persist at larger scale even though a text description of the same scene solves them.

Explaining individual predictions. LIME and SHAP attribute a single prediction to input features through a locally fitted surrogate or additive feature values (Ribeiro et al., 2016; Lundberg & Lee, 2017). We instead compare four responses produced by a declared set of interventions, so the quantity of interest is fixed by the experiment rather than by an attribution rule.

Positioning. Information measures quantify dependence over a data distribution, contribution scores measure influence, representation measures compare internal states, activation patching localizes mediation, and benchmarks score outcomes. None reports together whether modality evidence reaches the answer, whether its effect depends on the question, and whether it follows meaning rather than nuisance. The Functional Interaction Profile reports all three per modality and semantic dimension, and the same four scores tell a misplaced decision from missing evidence.

B Proofs for Functional Geometry

B.1 Functional quotient

Proposition 5 (Functional quotient)**.** *The distance $d_{\Theta,D}$ of Eq. 3 is a pseudometric on input space. If $X \sim X'$ exactly when their induced output distributions agree, then it is a metric on the quotient $\mathcal{X}/\sim$. The analogous statement holds for $d_{f,W}$ after quotienting by its zero-distance relation.*

Proof of Proposition 5. Non-negativity, symmetry, and the triangle inequality of $d_{\Theta,D}$ follow immediately from the corresponding properties of D. Identity of indiscernibles can fail on raw inputs because two distinct inputs may induce the same output distribution. Equality of induced distributions is an equivalence relation. If $[X]$ denotes the equivalence class of X, define

$\bar{d}([X],[X']) = d(X,X')$. This is well defined: replacing either representative by one with the same induced distribution leaves D unchanged. Because D is a metric, $\bar{d}([X],[X']) = 0$ if and only if $[X] = [X']$. The remaining metric axioms are inherited from D.

For $d_{f,W}$, quotient by $X \sim_W X'$ iff $\|f(X) - f(X')\|_W = 0$. A positive-semidefinite quadratic form is a seminorm, so the same argument gives a metric on its zero-distance quotient. □

B.2 Input-coordinate invariance

Proof of Proposition 1. By construction,

$$\widetilde{K}(\cdot \mid h(X)) = K(\cdot \mid h^{-1}(h(X))) = K(\cdot \mid X). \tag{11}$$

The same holds for X'. Substitution into Eq. 3 proves $\widetilde{d}_{\Theta,D}(h(X),h(X')) = d_{\Theta,D}(X,X')$. Every finite response-surface quantity is preserved because $\widetilde{f}(h(X)) = f(h^{-1}(h(X))) = f(X)$ at each evaluated cell.

If h is a smooth diffeomorphism, the local metric obeys the tensor transformation rule

$$\widetilde{G}(\widetilde{x}) = J_{h^{-1}}(\widetilde{x})^\top G(x) J_{h^{-1}}(\widetilde{x}), \tag{12}$$

so corresponding tangent perturbations have the same quadratic functional energy. □

B.3 Local functional quotient

Proposition 6 (Local functional quotient)**.** $\ker G_f(x) = \ker J_f(x)$. *Hence the first-order functional tangent space is* $T_x\mathcal{X}/\ker J_f(x)$.

Proof of Proposition 6. For every perturbation δ,

$$\delta^\top G_f(x)\delta = (J_f(x)\delta)^\top W (J_f(x)\delta). \tag{13}$$

Since $W \succ 0$, the right-hand side is zero if and only if $J_f(x)\delta = 0$. Therefore $\ker G_f(x) = \ker J_f(x)$. Perturbations that differ by an element of this kernel induce the same first-order response, which gives the stated tangent quotient. □

Since $\operatorname{rank} J_f \le k$, a k-dimensional observable registers at most k directions of an input block of any size.

C Proofs for Conditionality

C.1 Exact anchored decomposition

Proposition 7 (Exact anchored decomposition)**.** *For arbitrary anchors* U_0, E_0 *and any* U, E*,*

$$f(U,E) = f(U_0,E_0) + \big[f(U,E_0) - f(U_0,E_0)\big] + \big[f(U_0,E) - f(U_0,E_0)\big] + \Gamma_f(U,E;U_0,E_0). \tag{14}$$

On a 2×2 rectangle this is Eq. 6 with the anchor cell, rather than the grand mean, as baseline.

Proof of Proposition 7. Insert the definition of Γ_f into the right-hand side of Eq. 14. The two copies of $f(U,E_0)$ and $f(U_0,E)$ cancel, as do the appropriate copies of $f(U_0,E_0)$, leaving $f(U,E)$. □

C.2 Uniqueness of the mixed contrast

Proposition 8 (Uniqueness of the mixed linear contrast)**.** *Let* $L : \mathbb{R}^4 \to \mathbb{R}$ *be linear under cell order* $(00, 01, 10, 11)$*. If* L *vanishes on every additive surface* $s_{ue} = a_u + b_e + c$ *and satisfies* $L(1,-1,-1,1) = 4$*, then* $L(s) = \Gamma_s$ *for every surface* s*.*

Proof of Proposition 8. Under cell order $(00, 01, 10, 11)$, additive surfaces form the span of

$$(1, 1, 1, 1), \qquad (0, 0, 1, 1), \qquad (0, 1, 0, 1). \tag{15}$$

These three vectors are linearly independent, so their annihilator in the dual of $\mathbb{R}^4$ is one dimensional. The vector $q = (1, -1, -1, 1)$ is orthogonal to all three and therefore spans that annihilator. Thus $L(s) = \alpha q^\top s$ for some α. The normalization $L(q) = 4$ and $q^\top q = 4$ imply $\alpha = 1$. Hence $L(s) = s_{00} - s_{01} - s_{10} + s_{11} = \Gamma_s$. □

C.3 Additive separability and additive-bias invariance

Proof of Proposition 2(i). If $f(U, E) = \alpha(U) + \beta(E) + c$, direct substitution into Eq. 5 cancels every term and gives $\Gamma_f = 0$. Conversely, suppose $\Gamma_f(U, E; U_0, E_0) = 0$ for every U, E. Define

$$\alpha(U) = f(U, E_0) - f(U_0, E_0), \tag{16}$$
$$\beta(E) = f(U_0, E) - f(U_0, E_0), \tag{17}$$
$$c = f(U_0, E_0). \tag{18}$$

Rearranging $\Gamma_f = 0$, or applying Proposition 7, yields $f(U, E) = \alpha(U) + \beta(E) + c$. □

C.4 Uniform approximate separability

Proposition 9 (Uniform approximate separability)**.** *Fix anchors and a norm on the observable space. If $\|\Gamma_f(U, E; U_0, E_0)\| \le \eta$ for every point in an anchored domain, then there is an additive function $\widetilde{f}(U, E) = \alpha(U) + \beta(E) + c$ satisfying*

$$\|f(U, E) - \widetilde{f}(U, E)\| \le \eta \tag{19}$$

uniformly on that domain.

Proof of Proposition 9. Define the anchored additive approximation

$$\widetilde{f}(U, E) = f(U, E_0) + f(U_0, E) - f(U_0, E_0). \tag{20}$$

This has the form $\alpha(U) + \beta(E) + c$. Rearranging Eq. 14 gives the exact residual identity

$$f(U, E) - \widetilde{f}(U, E) = \Gamma_f(U, E; U_0, E_0). \tag{21}$$

Taking norms and applying the assumed uniform bound proves Eq. 19. □

C.5 Additive response biases

Proof of Proposition 2(ii). The contrast is linear, so

$$\Gamma_{f+\alpha+\beta+c} = \Gamma_f + \Gamma_\alpha + \Gamma_\beta + \Gamma_c. \tag{22}$$

Each of $\alpha(U)$, $\beta(E)$, and c is additively separable and has zero conditionality by part (i). Therefore $\Gamma_{\widetilde{f}} = \Gamma_f$. □

C.6 Positive-margin matching

Proof of Theorem 1. By definition and the four margin assumptions of Theorem 1,

$$\Gamma_s = s_{11} - s_{10} - s_{01} + s_{00} \ge m - (-m) - (-m) + m = 4m. \tag{23}$$

□

The constant 4 is sharp: equality holds when both positive cells equal m and both negative cells equal $-m$.

C.7 THRESHOLD FEASIBILITY

Proof of Proposition 3. Let $p = \min(s_{00}, s_{11})$ and $n = \max(s_{01}, s_{10})$. After subtracting threshold t, the worst gold margin is $\min(p - t, t - n)$. It is maximized at $t = (p + n)/2$, with value $(p-n)/2$, which is positive exactly when $p > n$. In the factorial representation, $p = \mu + d - |a+b|$ and $n = \mu - d + |a - b|$. Thus

$$\frac{p-n}{2} = d - \frac{|a+b| + |a-b|}{2} = d - \max(|a|, |b|). \tag{24}$$

The final identity holds for all real a, b. □

The factor 4 in Theorem 1 is sharp: the table $(m, -m, -m, m)$ attains equality. To see directly why a large interaction does not guarantee success, add a sufficiently large positive constant to all four entries. Their signs then agree although $\Gamma_s = 4m$ remains unchanged. Alternatively, sufficiently large main effects can make a rectangle infeasible for any threshold even when $d > 0$.

C.8 FINITE-LOCAL CORRESPONDENCE

Theorem 2 (Finite-local correspondence)**.** *Suppose f is twice continuously differentiable on the rectangle joining the anchors, and write $\Delta U = U_1 - U_0$ and $\Delta E = E_1 - E_0$. Then Eq. 7 holds, and consequently, as both displacements vanish,*

$$\Gamma_f = D^2_{UE} f(U_0, E_0)[\Delta U, \Delta E] + o(\|\Delta U\| \, \|\Delta E\|). \tag{25}$$

If, in addition, $D^2_{UE} f$ has operator norm at most L_{UE} on the rectangle, then $\|\Gamma_f\| \leq L_{UE} \|\Delta U\| \|\Delta E\|$.

Proof of Theorem 2. Define

$$g(\sigma, \tau) = f(U_0 + \sigma \Delta U, E_0 + \tau \Delta E). \tag{26}$$

The fundamental theorem of calculus, applied once in each coordinate, gives

$$g(1,1) - g(1,0) - g(0,1) + g(0,0) = \int_0^1 \int_0^1 \frac{\partial^2 g}{\partial \sigma \partial \tau}(\sigma, \tau) \, d\tau \, d\sigma. \tag{27}$$

The left-hand side is Γ_f. By the chain rule,

$$\frac{\partial^2 g}{\partial \sigma \partial \tau}(\sigma, \tau) = D^2_{UE} f(U_0 + \sigma \Delta U, E_0 + \tau \Delta E)[\Delta U, \Delta E], \tag{28}$$

which proves Eq. 7. Continuity of $D^2_{UE} f$ implies that the integrand equals $D^2_{UE} f(U_0, E_0)[\Delta U, \Delta E] + o(\|\Delta U\| \|\Delta E\|)$ uniformly as the rectangle shrinks, proving Eq. 25. Finally, the operator-norm bound holds pointwise inside the integral; integrating it over the unit square gives $\|\Gamma_f\| \leq L_{UE} \|\Delta U\| \|\Delta E\|$. □

C.9 WHY THE PULLBACK CROSS-BLOCK IS DIFFERENT

Proposition 10 (First-order alignment is not interaction)**.** *There exist additive functions for which $J_U^\top W J_E \neq 0$ while $\Gamma_f = 0$ and $D^2_{UE} f = 0$.*

The cross-block measures alignment of first-order output effects, whereas Γ_f measures context-dependent change in those effects.

Proof of Proposition 10. Take scalar U, E, let $f(U, E) = U + E$, and set $W = 1$. The response is additive, so $\Gamma_f = 0$ and $D^2_{UE} f = 0$. Nevertheless $J_U = J_E = 1$, hence $J_U^\top W J_E = 1$. □

D Proofs for Usability and Diagnostic Consequences

D.1 Local forms of accessibility and selectivity

For fixed text E, define the finite modality effect

$$A_f(U, U_0 \mid E) = f(U, E) - f(U_0, E). \tag{29}$$

Locally, a modality displacement δ_U has functional energy

$$\mathcal{E}_f(\delta_U; U, E) = \delta_U^\top G_{UU}(U, E)\delta_U, \qquad G_{UU} = J_U^\top W J_U. \tag{30}$$

A displacement in $\ker J_U$ is inaccessible to first order, and Proposition 11 bounds how much of a displacement the response can register in terms of the spectrum of G_{UU}. Nonzero accessibility is necessary when the modality contains required evidence, but it does not say whether that evidence is used for the right question.

Proposition 11 (Spectral accessibility bounds)**.** *Let $G \succeq 0$ have eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$, and let S_k be the span of its first k eigenvectors. Then every displacement δ satisfies*

$$\lambda_k \|P_{S_k}\delta\|^2 + \lambda_d \|P_{S_k^\perp}\delta\|^2 \leq \delta^\top G \delta, \tag{31}$$

$$\delta^\top G \delta \leq \lambda_1 \|P_{S_k}\delta\|^2 + \lambda_{k+1} \|P_{S_k^\perp}\delta\|^2. \tag{32}$$

Proof. Expand δ in an orthonormal eigenbasis of G. Coefficients in S_k are weighted between λ_k and λ_1, while coefficients in its orthogonal complement are weighted between λ_d and λ_{k+1}. □

Let δ_{sem} be an answer-changing modality displacement and δ_{nuis} an answer-preserving nuisance displacement. We define

$$R_{\text{sel}}^{\text{loc}} = \frac{\mathbb{E}[\mathcal{E}_f(\delta_{\text{sem}})]}{\mathbb{E}[\mathcal{E}_f(\delta_{\text{nuis}})] + \varepsilon}. \tag{33}$$

This local form requires differentiability and is measured only in the local study of Appendix I.3; the finite form R_{sel} of Eq. 9 is the coordinate reported throughout, and the two should not be read as the same number.

D.2 Logical independence of the three coordinates

Proof of Proposition 4. Use scalar binary semantic state u, text state e, and nuisance state n, with anchors at zero and unit semantic/nuisance interventions.

First, $f(u, e, n) = Cu$ has semantic accessibility C^2 under the squared Euclidean output metric, which can be made arbitrarily large, but it is additive and has $\Gamma_f = 0$.

Second, let $f(u, e, n) = ue + Cn$. The semantic modality–text rectangle has $\Gamma_f = 1$. At $e = 1$, the semantic energy is 1 while the nuisance energy is C^2, so $R_{\text{sel}}^{\text{loc}} = 1/(C^2 + \varepsilon)$ can be made arbitrarily small without removing interaction.

Third, return to $f(u, e, n) = Cu$. Its nuisance energy is zero and its semantic energy is C^2, so $R_{\text{sel}}^{\text{loc}} = C^2/\varepsilon$ can be arbitrarily large, while $\Gamma_f = 0$. These constructions establish all three separations. □

D.3 Semantic counterfactual separation

Proposition 12 (Semantic counterfactual separation)**.** *Let p_0, p_1 be answer distributions for semantic counterfactuals with different correct labels $y_0 \neq y_1$. If each is correct against the other label with probability margin at least $m > 0$, then*

$$\mathrm{TV}(p_0, p_1) \geq m. \tag{34}$$

Proof of Proposition 12. Define $r(p) = p(y_0) - p(y_1)$. The two margin assumptions give $r(p_0) \geq m$ and $r(p_1) \leq -m$, hence

$$2m \leq |r(p_0) - r(p_1)|. \tag{35}$$

By the triangle inequality,

$$|r(p_0) - r(p_1)| \le |p_0(y_0) - p_1(y_0)| + |p_0(y_1) - p_1(y_1)| \tag{36}$$
$$\le \|p_0 - p_1\|_1. \tag{37}$$

Therefore $\mathrm{TV}(p_0, p_1) = \frac{1}{2}\|p_0 - p_1\|_1 \ge m$. □

D.4 Null-prior persistence

Proposition 13 (Null-prior persistence). *If* $\mathrm{TV}(P_{\mathrm{full}}, P_{\mathrm{null}}) \le \varepsilon$*, then for every answer event* H*,*

$$P_{\mathrm{full}}(H) \ge P_{\mathrm{null}}(H) - \varepsilon. \tag{38}$$

Proof of Proposition 13. The variational characterization of total variation gives, for every measurable event H,

$$|P_{\mathrm{full}}(H) - P_{\mathrm{null}}(H)| \le \mathrm{TV}(P_{\mathrm{full}}, P_{\mathrm{null}}) \le \varepsilon. \tag{39}$$

Rearranging yields Eq. 38. □

E Functional Interaction Versus Distributional PID

The following proposition makes exact the distinction drawn in Sec. 3.3.

Proposition 14 (Non-equivalence with MMI-PID). *Let* $U \sim \mathrm{Bernoulli}(1/2)$ *and* $E \sim \mathrm{Bernoulli}(\epsilon)$ *be independent, with* $0 < \epsilon \le 1/2$*, let* $Y = \mathbf{1}\{U = E\}$*, and use the score* $s = 2Y - 1$*. Then*

$$\Gamma_s = 4 \text{ for every } \epsilon, \qquad \mathrm{Syn}_{\mathrm{MMI}}(U, E; Y) = h_2(\epsilon) \to 0 \text{ as } \epsilon \downarrow 0, \tag{40}$$

where h_2 *is binary entropy. Conversely, replacing* s *by* $s_c = c(2Y - 1)$ *with* $c > 0$ *leaves every PID atom unchanged but gives* $\Gamma_{s_c} = 4c$*.*

Proof of Proposition 14. For every ϵ, the deterministic matching rule has response table

$$(s_{00}, s_{01}, s_{10}, s_{11}) = (1, -1, -1, 1), \tag{41}$$

so $\Gamma_s = 1 - (-1) - (-1) + 1 = 4$.

Because U is balanced and independent of E, the target $Y = \mathbf{1}\{U = E\}$ is balanced, so $H(Y) = 1$. Conditional on E, Y is also balanced; hence $I(E; Y) = 0$. Conditional on U, the target has Bernoulli entropy $h_2(\epsilon)$, giving

$$I(U; Y) = 1 - h_2(\epsilon). \tag{42}$$

Since Y is deterministic given (U, E), $I(U, E; Y) = H(Y) = 1$. Under two-source MMI-PID,

$$\mathrm{Red}_{\mathrm{MMI}} = \min\{I(U; Y), I(E; Y)\} = 0, \tag{43}$$
$$\mathrm{Unq}_U = 1 - h_2(\epsilon), \tag{44}$$
$$\mathrm{Unq}_E = 0, \tag{45}$$
$$\mathrm{Syn}_{\mathrm{MMI}} = I(U, E; Y) - \mathrm{Red}_{\mathrm{MMI}} - \mathrm{Unq}_U - \mathrm{Unq}_E = h_2(\epsilon). \tag{46}$$

Thus MMI synergy approaches zero although the response surface and Γ_s are fixed.

For the converse direction, $s_c = c(2Y - 1)$ is a bijective recoding of Y for every $c > 0$. Mutual informations, and therefore all MMI-PID atoms, are invariant to this bijection, while linearity gives $\Gamma_{s_c} = 4c$. This proves both non-determination statements. □

Observable relativity. At $\epsilon = 1/2$, the matching target has one bit of MMI synergy and the task score has $\Gamma_s = 4$. On the same sampled inputs, the U-only observable $r(U, E) = 2U - 1$ has $\Gamma_r = 0$. Thus a zero mixed difference for one selected readout does not imply that every variable in the system lacks distributional synergy; conversely, synergy of a target does not fix the scale or geometry of an arbitrary model observable.

F Internal Factorial Coordinates and Causal Scope

F.1 Exact activation decomposition

Proposition 15 (Exact factorial activation component)**.** *Let $x_u = 2u - 1$ and $x_e = 2e - 1$. Any four activation vectors have the unique decomposition*

$$h_{ue} = \mu + x_u a + x_e b + x_u x_e d, \qquad d = \frac{h_{11} - h_{10} - h_{01} + h_{00}}{4}. \tag{47}$$

Replacing d while keeping (μ, a, b) fixed preserves the activation grand mean and both factorial main effects exactly.

Proof of Proposition 15. Define

$$\mu = \frac{1}{4}\sum_{u,e} h_{ue}, \qquad a = \frac{1}{4}\sum_{u,e} x_u h_{ue}, \tag{48}$$

$$b = \frac{1}{4}\sum_{u,e} x_e h_{ue}, \qquad d = \frac{1}{4}\sum_{u,e} x_u x_e h_{ue}. \tag{49}$$

The four sign patterns $1, x_u, x_e, x_u x_e$ are mutually orthogonal over the 2×2 design and each has squared norm four. They therefore form a basis for functions on the four cells, giving $h_{ue} = \mu + x_u a + x_e b + x_u x_e d$ uniquely. Expanding d gives

$$d = \frac{h_{11} - h_{10} - h_{01} + h_{00}}{4} = \frac{\Gamma_h}{4}. \tag{50}$$

Replacing d changes only the coefficient of the orthogonal interaction basis vector; the coefficients μ, a, b are unchanged. Hence the grand mean and both main effects are preserved exactly. □

For a recipient family, interaction erasure uses $h_{ue}^{\text{erase}} = h_{ue} - x_u x_e d$, while transplantation from a donor with interaction d' uses $h_{ue}^{\text{transplant}} = h_{ue} + x_u x_e (d' - d)$. These statements concern the intervened activation table. Downstream nonlinearities may change every output-level factorial coordinate, which is why off-target output controls are required.

F.2 Global additive partitions

Proposition 16 (What nonzero output interaction excludes)**.** *If the selected output can be written as the sum of a modality-only computation and a text-only computation, then $\Gamma_f = 0$. Hence $\Gamma_f \neq 0$ rules out that global additive partition for the selected observable.*

Proof of Proposition 16. Suppose the complete selected response can be written $f(U, E) = \alpha(U) + \beta(E) + c$. Proposition 2 gives $\Gamma_f = 0$. Taking the contrapositive shows that $\Gamma_f \neq 0$ rules out this global additive partition. The argument does not specify where or how the non-additivity is implemented, so no unique localization conclusion follows. □

G Supporting Extensions

G.1 The pullback Fisher metric

For a regular predictive distribution $p_\Theta(y \mid x)$, replacing the output quadratic form of Eq. 4 by local KL divergence yields the pullback Fisher metric (Amari, 1998)

$$G_F(x) = \mathbb{E}_y\left[\nabla_x \log p_\Theta(y \mid x) \nabla_x \log p_\Theta(y \mid x)^\top\right]. \tag{51}$$

This is the information-geometric special case of the same pullback construction: it weights an input displacement by the KL divergence it induces on the predictive distribution rather than by a declared output norm, and it is the form to use when the observable is a full predictive distribution. Our observables are answer scores over declared candidates, so the measurements in this paper use Eq. 4 directly.

G.2 Higher-order interactions by Möbius inversion

The two-factor contrast is the order-two member of a general factorial decomposition. Let $f : \{0,1\}^d \to V$ for a vector space V. For each $S \subseteq [d]$, define the anchored effect

$$\Delta_S f = \sum_{T \subseteq S} (-1)^{|S|-|T|} f(\mathbf{1}_T), \tag{52}$$

where $\mathbf{1}_T$ activates the factors in T and holds the others at zero. Möbius inversion on the Boolean lattice gives

$$f(\mathbf{1}_A) = \sum_{S \subseteq A} \Delta_S f. \tag{53}$$

The decomposition is unique. For $d = 2$, $\Delta_{\{U,E\}} f = \Gamma_f$. Higher-order terms can represent, for example, modality $\times$ instruction $\times$ conversation-history interaction, but a complete design requires 2^d cells. We therefore evaluate the identifiable two-factor coordinate in this paper rather than estimating sparse higher-order terms post hoc.

G.3 Vector observables and scalar reports

All finite interaction identities hold componentwise for $f : \mathcal{U} \times \mathcal{E} \to \mathbb{R}^k$. A reported scalar must declare its scalarization. Scientifically distinct choices include

$$\|\Gamma_f\|_W, \qquad v^\top \Gamma_f, \qquad \Gamma_s \text{ for a declared signed answer margin } s. \tag{54}$$

The signed projection is required for directional statements such as Theorem 1; a norm is appropriate when no output direction is privileged but discards orientation. In particular, the candidate-answer distribution contrast

$$\Gamma_P = P_{11} - P_{10} - P_{01} + P_{00} \tag{55}$$

checks whether a conclusion survives beyond one yes/no margin. Scalar and vector observables must use the same frozen prompts, candidates, calibration, and cell definitions when compared with PID or another baseline.

G.4 Estimation under stochastic execution

Suppose cell (u, e) is evaluated independently n_{ue} times and $\widehat{f}_{ue}$ is its sample mean. The estimator

$$\widehat{\Gamma} = \widehat{f}_{11} - \widehat{f}_{10} - \widehat{f}_{01} + \widehat{f}_{00} \tag{56}$$

is unbiased for Γ_f by linearity of expectation. If the four cell means are independent with per-execution variances σ^2_{ue}, then

$$\mathrm{Var}(\widehat{\Gamma}) = \sum_{u,e} \frac{\sigma^2_{ue}}{n_{ue}}. \tag{57}$$

In the experiments, cells, paraphrases, nuisance variants, and semantic edits within one family share latent content and are not independent row-level units. Inference therefore clusters and splits by semantic family. The formula above describes repeated execution of fixed cells; it does not authorize treating dependent transformations as additional samples.

G.5 Functional Interaction Profile contract

Roles of the seven coordinates. The first three coordinates describe evidence use. The remaining four extend the comparison: would a text description produce the same response (G_{surr}), can an internal patch restore it (C_{med}), at what depth does a fixed readout expose interaction (L_{onset}), and how much does identical evaluation vary (V_{exec})? Tab. 1 gives the meaning and necessary controls for each coordinate. Surrogate comparisons require justified shared information; mediation and onset require internal access. Unavailable measurements have explicit reasons in the same seven-coordinate schema. There is no aggregate ability score: these properties need not share an ordering from better to worse. The four remaining coordinates are independent of the first three in the same sense, and Proposition 17 in Appendix D constructs models that agree on accuracy, accessibility, interaction, and selectivity while differing on surrogate gap, mediation, onset, or execution variability. Seven is a reporting choice rather than a mathematically minimal or exhaustive set.

Table 1: **Reading the seven-component profile.** The coordinates describe evidence use, response agreement and internal behavior, and measurement reliability. Each answers a specific question; availability depends on suitable interventions, surrogates, and model access.

Coordinate	Question answered	Necessary companion
Evidence use		
Semantic accessibility	Do answer-changing inputs produce different responses?	Correctness, raw response distance, and score scale.
Conditionality	Does the effect of the modality depend on the question?	Signed Γ, all four scores, and semantic orientation.
Semantic selectivity	Is the response more sensitive to relevant than irrelevant changes?	Matched references, nuisance strength, numerator and denominator.
Response agreement and internal behavior		
Text-surrogate gap	Does a task-sufficient text description induce the same response? Proposition 18 gives the sufficiency criterion under which the gap is zero.	Independent justification of shared task information.
Causal mediation	Can a specified internal patch restore a response?	Signed layer curves, corruption baseline, and off-target controls.
Interaction onset	At what depth does a fixed readout expose interaction?	Readout definition, layer coverage, and execution floor.
Measurement reliability		
Execution variability	How much does repeated identical evaluation vary?	Repetition count and numerical precision.

Modality utility. A companion measurement compares the gold-answer margin with and without the original modality input, using a specified null such as blank or resampled media. It measures benefit under that intervention.

Normalization. Let $S_f > 0$ be a scale frozen on development data, such as the median absolute clean answer margin. The normalized accessibility and conditionality coordinates are

$$A_{\text{sem}} = \frac{\mathbb{E}D(K(U(s(x)), E(q)), K(U(x), E(q)))}{S_f + \varepsilon}, \tag{58}$$

$$C_{\text{ctx}} = \frac{\mathbb{E}\|\Gamma_f\|_W}{S_f + \varepsilon}, \tag{59}$$

and the surrogate gap uses the same division. For JS quantities the division is a reporting convention rather than a probability divergence. Because it can change orderings between slices, every normalized value is reported with its raw value and scale, and the report records the model revision, preprocessing, prompts, candidates, score convention, semantic and nuisance edits, null, aggregation and scale. Additive calibration preserves raw Γ, and normalized interaction whenever S_f is unchanged. Modality- and dimension-specific results are the primary comparisons; Appendix I.8 specifies the aggregation and internal measurements.

For a declared patching experiment, let $G \in \{U, Q, A\}$ denote modality, question, and pre-answer token groups and let $\rho_{i,\ell,G}$ be signed recovery. One compact path summary is

$$C_G = \mathbb{E}_i \left[\frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} [\rho_{i,\ell,G}]_+ \right], \tag{60}$$

$$C_{\text{med}} = \frac{3}{(C_U + \varepsilon)^{-1} + (C_Q + \varepsilon)^{-1} + (C_A + \varepsilon)^{-1}}. \tag{61}$$

The harmonic form is small when one stage of the declared path is weak. It is only a summary: negative recovery, sign reversals, over-recovery, token groups, and full layer curves must remain available.

Let τ_ℓ be a development-frozen execution threshold and $\Gamma^{\max}$ the peak of the median layer-wise interaction. The normalized onset coordinate is the earliest layer fraction at which interaction exceeds

$\max(\tau_\ell, \Gamma^{\max}/4)$ for a predeclared number K of consecutive measured layers. The peak-relative term matters when repeated executions have no jitter, since τ_ℓ alone then marks the first layer above numerical rounding. A missing component remains missing with a machine-readable reason; it is never imputed into a scalar profile score.

Proposition 17 (Secondary profile coordinates do not collapse)**.** *Models can have the same output accuracy and the same accessibility, interaction, and selectivity while differing in surrogate gap, causal mediation, interaction onset, or execution variability.*

Proof. Two computations can implement the same four output values while one routes the mixed effect through question-token states and the other introduces it only at the answer computation; their output axes match while their patching paths and onset depths differ. Two channels can produce the same selected answer margin but different full answer distributions or internal observables, giving different surrogate gaps. Finally, two stochastic kernels can have the same mean response and accuracy but different within-input dispersion. None of these secondary coordinates is therefore determined by accuracy or by the three primary response summaries. □

Reporting rules. A profile is reported together with the evaluation design that produced it. Two profiles are directly comparable when they use matched semantic families, questions and candidates, preprocessing, controls and aggregation; a changed condition defines a new comparison whose interpretation must be stated. Each report retains the seven coordinate slots, task success, model revision, family identifiers, scoring interface, raw values, scale, uncertainty and probe coverage, and for interaction all four consistently oriented scores, so that the contrast and threshold feasibility can be recomputed. A coordinate is reported in one of three states: a supported estimate under the stated controls, a computed estimate with restricted interpretation, or an unavailable measurement with its reason. Missing internal access or an unjustified text surrogate leaves a slot unavailable rather than zero. Some computed values need their companions to be interpretable: selectivity its numerator, denominator, reference state and floor; mediation its signed curves, corruption and denominator; onset its readout, layer coverage and threshold; and zero execution variability its precision and repeat count. Score-based coordinates require access to candidate scores and internal coordinates require readout or patch access, so a partially available profile reports its supported coordinates with coverage made explicit. The exact identities establish properties of the contrast and the response table; the empirical evidence concerns the tested models, tasks and transformations and does not establish equal stability for all seven coordinates.

G.6 Task-conditioned textual-surrogate equivalence

For a modality input x, textual surrogate $a(x)$, and question q, define

$$G_{\mathrm{surr}}(x, a, q) = D(K_\Theta(\cdot \mid U(x), E(q)), K_\Theta(\cdot \mid E(a(x)), E(q))) \,. \tag{62}$$

Proposition 18 (Sufficient-surrogate criterion)**.** *Suppose the response to both channels factors through the same task statistic z_q and kernel H_q:*

$$K_\Theta(\cdot \mid U(x), E(q)) = H_q(\cdot \mid z_q(x)), \tag{63}$$
$$K_\Theta(\cdot \mid E(a(x)), E(q)) = H_q(\cdot \mid z_q(a(x))). \tag{64}$$

If $z_q(a(x)) = z_q(x)$, then $G_{\mathrm{surr}}(x, a, q) = 0$.

Proof. Both conditions induce the identical distribution $H_q(\cdot \mid z_q(x))$, so their distance is zero. □

A nonzero surrogate gap is not uniquely identifiable: the surrogate may omit relevant information, the modality may contain legitimate extra information, the model may respond to nuisance detail, or the two interfaces may engage different downstream computations. Surrogate analyses must therefore be restricted to tasks for which shared semantics are independently justified.

G.7 Execution variability as an identifiability floor

For a declared distance D, define within-input dispersion and an input-intervention effect by

$$V_{\text{exec}}(X) = \mathbb{E}_{\xi,\xi'} D(Z_{\Theta,\xi}(X), Z_{\Theta,\xi'}(X)), \tag{65}$$

$$V_{\text{input}}(X, X') = \mathbb{E}_{\xi} D(Z_{\Theta,\xi}(X), Z_{\Theta,\xi}(X')). \tag{66}$$

A practical signal-to-execution ratio is

$$R_{\text{exec}} = \frac{V_{\text{input}}(X, X')}{V_{\text{exec}}(X) + \varepsilon}. \tag{67}$$

Effects that do not clear repeated-execution uncertainty are not identifiable as input effects under the selected implementation. Teacher-forced answer scores reduce decoding-branch variability and are therefore the primary observable.

H Experimental Protocol and Reporting Conventions

Table 2: **Evaluation panel.** A family is one semantic intervention rectangle, not one of its cells or variants. Dashes are unsupported interfaces, not missing runs. MiniCPM uses a reduced per-modality budget; its families are drawn from the same pools as Qwen's.

Model	Image	Video	Audio	Speech	Evaluations
LLaVA-OneVision-7B	500	—	—	—	500
Qwen2.5-Omni-7B	500	500	2,000	1,000	4,000
MiniCPM-o 4.5	128	128	128	128	512
Total	1,128	628	2,128	1,128	5,012

H.1 Checkpoints and execution

Table 3 specifies the scientific checkpoints. LLaVA pairs a SigLIP vision encoder with a Qwen2-7B language model; Qwen and MiniCPM include audio encoders. LLaVA's and MiniCPM's evaluated families are subsets of Qwen's, giving 4,000 distinct families across the panel. Runs use bfloat16, scaled dot-product attention, and no quantization. LLaVA uses its original implementation. Qwen uses the Thinker text-output interface; audio generation is outside scope. MiniCPM uses its understanding interface and its frame/image-token path for video.

Table 3: **Scientific checkpoint revisions.**

Checkpoint	Immutable revision
`lmms-lab/llava-onevision-qwen2-7b-ov`	`0b07bf7565e244cf4f39982249eafe8cd799d6dd`
`Qwen/Qwen2.5-Omni-7B`	`ae9e1690543ffd5c0221dc27f79834d0294cba00`
`openbmb/MiniCPM-o-4_5`	`1f761131fa83f5ed3cd6f2f22b225c4501d154fa`

H.2 Semantic families and controls

Image dimensions are color, material, shape, size, left/right, front/behind, and count. Video dimensions are direction, speed, event order, repetition, rotation, and duration. Environmental audio covers event order, event presence, repetition, relative loudness, and simultaneous/sequential events. Speech covers keyword, named-entity, and number identity, proposition truth, spoken order, and affirmation/negation.

Every video is eight 256×256 frames rendered without compression and passed to each model without resampling, so each property is fixed by which frames contain the object and where. Qwen2.5-Omni's encoder merges consecutive frame pairs into one temporal patch, giving four temporal positions in which every property remains distinguishable; MiniCPM-o receives each frame as a separate

Table 4: **The controlled semantic dimensions.** Twenty-four modality–dimension combinations over 22 distinct dimensions: event order and repetition each appear in two modalities, and the results pool them by name. One canonical question is shown per row; the crossed question E_1 is the matching predicate over the other state. These names are used unchanged in the results.

Modality	Dimension	Example canonical question
Image	Colour	Is the small cube brown?
	Shape	Is the small green object a cube?
	Material	Is the purple cube metal?
	Size	Is the green cube small?
	Count	Are there exactly 2 brown objects?
	Left / right	Is the brown cylinder on the left side of the image?
	Front / behind	Is the brown sphere toward the front of the scene?
Video	Event order	Does the yellow circle appear before the blue triangle?
	Direction	Does the object move from left to right?
	Speed	Does the object move slowly (not reaching the far edge)?
	Duration	Is the object visible for a short duration?
	Repetition	Does the object appear exactly once?
	Rotation	Does the pointer rotate clockwise?
Audio	Event presence	Is a train sound present?
	Event order	Does the door-knock sound occur before the fireworks sound?
	Relative loudness	Is the car-horn sound louder than the glass-breaking sound?
	Repetition	Does the footsteps sound occur exactly once?
	Simultaneous / sequential	Do the breathing and pig sounds occur at the same time?
Speech	Keyword identity	Does the speaker say the word “water”?
	Named-entity identity	Does the speaker say the name “elizabeth”?
	Number identity	Does the speaker say the number word “one”?
	Spoken ordering	Does the speaker say “four” before “one”?
	Proposition truth	Does the recording contain both “two” and “four”?
	Affirmation / negation	Does the speaker use an explicit negation word?

image. Environmental-audio families place 2-second ESC-50 clips in a 4.3-second mixture, which involves no frame sampling.

Each family includes an answer-changing semantic variant for accessibility, 12 to 18 answer-preserving nuisance variants for selectivity, a text surrogate stating the task-relevant content, and reworded question paraphrases. Speech families use transcript-certified LibriSpeech recordings (Panayotov et al., 2015).

Each scientific family retains four canonical cells, media provenance, semantic labels, a family identifier, and transformation metadata. The image results use the CLEVR-based scientific build; the earlier two-dimensional synthetic image renderer is infrastructure-only. Video uses eight-frame temporal pairs. Audio mixtures retain source ESC-50 identifiers, event spans, gains, and seeds. Speech states are duration-matched with silence padding; same-target recordings support the declared speaker-change controls.

The three nulls are a blank input (a blank image, silence, or empty frames), the mean projected modality token, and a modality resampled from another semantic family. Where available, resampling matches sequence length, duration or resolution, norm bin, and data stratum. The blank null is the least natural of the three, because its distribution can differ substantially from ordinary inputs. Nuisance transformations include appearance, compression, timing, speaker, and channel changes. Their role depends on preserving the answer relative to the exact reference used in scoring; Appendix I.8 gives the reference convention used for selectivity.

Table 5 reports the three controlled null comparisons for every slice. Eight slices have positive median utility under all three nulls; MiniCPM video changes sign. These are medians of paired differences at the reference cell, not differences of marginal medians or a test of complete four-cell success. The comparison establishes which null was used to assess modality dependence.

Table 5: **Null choice changes the measured utility.** Medians of paired full-minus-null gold margins at U_0E_0, using the same families within a slice. The three nulls define different interventions; these are descriptive estimates, not a claim of null invariance.

Slice	Blank	Mean token	Resampled
LLaVA image	5.438	4.406	5.594
Qwen image	7.75	7.062	9.969
Qwen video	2.188	2.5	1.312
Qwen audio	1.312	0.875	1.375
Qwen speech	2.812	5.023	4.281
MiniCPM image	3.58	2.867	13.31
MiniCPM video	7.325	−6.641	2.361
MiniCPM audio	2.328	2.172	1.109
MiniCPM speech	5.188	3.578	5.742

H.3 Scoring, units, and ties

All four canonical cells retain the yes-minus-no score in Equation 10, using frozen candidate tokenization and the official chat format. No chain-of-thought response is requested. The candidate softmax uses temperature one. Semantic and surrogate JS values use base-two logarithms, so raw JS lies between zero and one. For multiple-choice natural tasks, the gold margin is

$$g(X) = S(a_{\text{gold}} \mid X) - \max_{a \neq a_{\text{gold}}} S(a \mid X). \tag{68}$$

The canonical checkerboard success convention is $s_{00}, s_{11} > 0$ and $s_{01}, s_{10} \leq 0$. The strict additive-bias audit instead requires all gold margins to be positive. This difference matters at ties; counts from the two analyses are not silently combined. The theoretical margin claims use $m > 0$ and therefore avoid ties entirely.

H.4 Splits, uncertainty, and descriptive analyses

The controlled datasets expose a development label. When an analysis needs train/development/test roles, it uses deterministic family partitions rather than an official upstream test split. Source media and derived variants remain together. PID joins use model revision × modality × family ID to distinguish repeated evaluations of the same underlying family by different models. Aggregate cards are summaries, not extra replications.

Profile intervals use 2,000 family-clustered bootstrap draws (Sec. 4), including the scale estimate. Cells, nulls, variants and paraphrases remain grouped with their semantic family. Calibration-disjoint counts, per-dimension counts and oracle threshold feasibility are retrospective summaries of saved outputs and carry no additional significance tests. Cross-model pooling accounts for shared family content and the unequal modality budgets.

I Additional results

This appendix gives the complete measurements behind Sec. 5 and Sec. 6, in the order in which the main text cites them, with the intervals, per-slice counts and qualifications omitted there.

I.1 Panel Overview

Table 6 summarizes every slice before the detailed results: four-cell success, the minimum gold-oriented margin, absolute interaction, and semantic response distance.

I.2 Intervals for the Main-Text Estimates

Table 7 gives a family-bootstrap interval for every median, ratio and correlation quoted in Sec. 5 and Sec. 6. Each is recomputed from the stored family records and checked against the value in the text; counts and proportions of families are exact.

Table 6: **Accuracy and response structure across the controlled suite.** Success requires all four cells to be correct. The remaining columns are family medians: minimum gold margin, absolute interaction, and raw semantic JS. Score magnitudes depend on the model's scoring scale. These columns are complementary measurements, not a cross-model scalar ranking.

Model	Modality	n	Success (%)	Min. margin	$\lvert\Gamma\rvert$	Sem. JS
LLaVA	Image	500	3.8	-2.50	13.41	0.0657
Qwen	Image	500	78.6	3.06	24.20	0.9145
	Video	500	16.6	-2.00	0.69	0.0030
	Audio	2,000	32.0	-0.63	1.94	0.0454
	Speech	1,000	60.9	0.63	7.81	0.4843
MiniCPM	Image	128	48.4	-0.13	22.63	0.5507
	Video	128	33.6	-3.41	5.69	0.0180
	Audio	128	6.3	-3.38	3.23	0.0143
	Speech	128	48.4	-0.34	11.71	0.4425

I.3 Invariance and Local Geometry

Coordinate dependence of representation summaries. Proposition 1 guarantees that responses, d_f and matched finite interactions are unchanged by an invertible reparameterization of the representation that is inverted downstream. Familiar representation summaries carry no such guarantee. On 576 raw-geometry records, 64 per slice, we recompute four summaries under the raw, mean-centered, diagonally whitened and norm-matched versions of the same modality tokens (Tab. 9). The ratio between the largest and smallest nearest-vocabulary Euclidean distance has median 3,118 across slices, with slice medians from 324 to 8,347; for LLaVA image the median distance is 73.03 raw, 6,254 whitened and 0.7493 norm-matched. Pooled CKA moves by a median of 0.093 and by up to 0.284, and pooled cosine by 0.025 to 0.081; only nearest-vocabulary cosine stays within 0.022. The finite contrast is invariant in the implementation as well as by construction: subtracting the frozen answer-bias offset from all 5,012 evaluations changes 840 four-cell decisions while the largest change in Γ is 8.88×10^{-16} (Proposition 2). These transformations change representation descriptors rather than implementing downstream-compensated reparameterizations. They show that a reported CKA or embedding distance jointly describes the model and an often unstated preprocessing convention, not that such analyses are uninformative.

Local sensitivity separates structure from noise, but not meaning from nuisance. The local metric of Eq. 4 is what a gradient-based probe of a modality representation measures. On 136 local families, eight for LLaVA and 16 per omni-modal slice, we compare the displacement-normalized energy of the answer-changing semantic edit, the answer-preserving nuisance edit and a random orthogonal displacement (Tab. 8). Random displacements carry three to four orders of magnitude less energy than either edit in the Qwen slices, so the metric separates structured edits from unstructured ones. It does not separate meaning from nuisance: the semantic edit carries more energy in only 53 of 136 families, significantly below chance (two-sided binomial $p = 0.0126$), and median nuisance energy exceeds median semantic energy in seven of nine slices, the exceptions being LLaVA image and Qwen speech. With 8 to 16 families per slice we report orderings rather than effect sizes. A gradient-based attribution over modality tokens therefore weighs answer-preserving edits at least as heavily as answer-changing ones, and telling them apart requires a finite controlled comparison. The effective rank of the Jacobian lies between 1.47 and 1.90 across slices, below its structural bound of two for the answer score, so the response registers at most a two-dimensional subspace of a modality block of 43 to 1,836 tokens.

Local numerical experiments. The mixed-derivative panel completes 120 of 136 planned evaluations; LLaVA's eight use the finite contrast because second-order differentiation exceeds the available memory. A coarse path quadrature of these strongly non-quadratic response surfaces reproduces the finite contrast within 10% in only 3 of 112 derivative-based rows (Fig. 5). The identity of Theorem 2 is exact, so the gap is numerical, and no finite measurement in the paper depends on it.

Table 7: **Intervals for the estimates quoted in Sections 5 and 6.** Point estimates with 95% percentile intervals from 2,000 draws that resample families within each slice (5,000 for the reworded contrasts). Counts and proportions of families quoted in the main text are exact and are not repeated here.

Quantity	Estimate	95% interval
Section 5		
Representation reparameterization factor (median over slices)	3,118	[3,052, 3,186]
Paraphrase rank correlation of signed Γ, pooled	0.99	[0.99, 0.99]
Paraphrase rank correlation of signed Γ, video	0.81	[0.71, 0.87]
Reworded success, high vs. low $\lvert\Gamma\rvert$ (percentage points)	36	[26, 47]
Reworded success, high vs. low synergy (percentage points)	-1	[-13, 10]
Section 6.1: onset depth and failed-to-solved peak ratio		
Onset, LLaVA image	81%	[74%, 81%]
Onset, Qwen image	74%	[74%, 74%]
Onset, MiniCPM image	77%	[77%, 80%]
Onset, Qwen speech	74%	[70%, 74%]
Onset, MiniCPM speech	77%	[77%, 77%]
Onset, Qwen audio	74%	[74%, 74%]
Onset, MiniCPM audio	74%	[63%, 77%]
Onset, Qwen video	74%	[74%, 74%]
Onset, MiniCPM video	77%	[69%, 83%]
Peak ratio, LLaVA image	78%	[73%, 83%]
Peak ratio, Qwen image	42%	[34%, 47%]
Peak ratio, MiniCPM image	69%	[61%, 80%]
Peak ratio, Qwen speech	27%	[23%, 34%]
Peak ratio, MiniCPM speech	29%	[20%, 49%]
Peak ratio, Qwen audio	11%	[9%, 12%]
Peak ratio, MiniCPM audio	23%	[13%, 35%]
Peak ratio, Qwen video	5%	[4%, 6%]
Peak ratio, MiniCPM video	6%	[3%, 10%]
Largest first-half level, MiniCPM speech (fraction of peak)	3.1%	[2.6%, 3.7%]
Section 6.2		
Largest within-modality accessibility ratio	6.9	[4.4, 12.9]
Across-modality accessibility ratio (omni-modal slices)	155	[137, 201]
Rank correlation of accessibility and interaction	0.95	[0.83, 1.00]
Variance share of modality, log accessibility	88%	[78%, 92%]
Variance share of model, log accessibility	2%	[0%, 4%]
Variance share of modality, interaction	91%	[67%, 96%]
Variance share of model, interaction	0%	[0%, 4%]
Decision-failure share, panel	40%	[39%, 42%]
Decision-failure share, LLaVA image	84%	[81%, 87%]
Decision-failure share, Qwen image	65%	[56%, 74%]
Decision-failure share, MiniCPM image	79%	[68%, 88%]
Decision-failure share, Qwen speech	42%	[37%, 47%]
Decision-failure share, MiniCPM speech	32%	[21%, 44%]
Decision-failure share, Qwen audio	29%	[27%, 32%]
Decision-failure share, MiniCPM audio	40%	[31%, 49%]
Decision-failure share, Qwen video	17%	[14%, 21%]
Decision-failure share, MiniCPM video	25%	[16%, 34%]
Median $\lvert\Gamma\rvert$, MiniCPM video repetition	9.7	[9.1, 9.9]
Median $\lvert\Gamma\rvert$, MiniCPM video rotation	0.8	[0.8, 0.8]
Section 6.3		
Pooled median $\lvert\Gamma\rvert$, speed	0.4	[0.3, 0.5]
Pooled median $\lvert\Gamma\rvert$, duration	1.2	[0.7, 1.4]
Pooled median $\lvert\Gamma\rvert$, repetition	0.8	[0.7, 0.8]
Pooled median $\lvert\Gamma\rvert$, rotation	1.1	[0.8, 1.1]
Pooled median $\lvert\Gamma\rvert$, simultaneity	0.5	[0.4, 0.5]
Median $\lvert\Gamma\rvert$, speed: Qwen2.5-Omni	0.3	[0.2, 0.4]
Median $\lvert\Gamma\rvert$, speed: Qwen3-Omni	2.9	[1.4, 3.8]
Growth factor, speed	9.3	[4.6, 13.5]
Median $\lvert\Gamma\rvert$, repetition: Qwen2.5-Omni	0.6	[0.5, 0.7]
Median $\lvert\Gamma\rvert$, repetition: Qwen3-Omni	4.8	[4.1, 7.9]
Growth factor, repetition	8.6	[6.4, 14.9]
Median $\lvert\Gamma\rvert$, simultaneity: Qwen2.5-Omni	0.4	[0.2, 0.6]
Median $\lvert\Gamma\rvert$, simultaneity: Qwen3-Omni	7.9	[7.2, 9.2]
Growth factor, simultaneity	18.1	[12.9, 34.1]

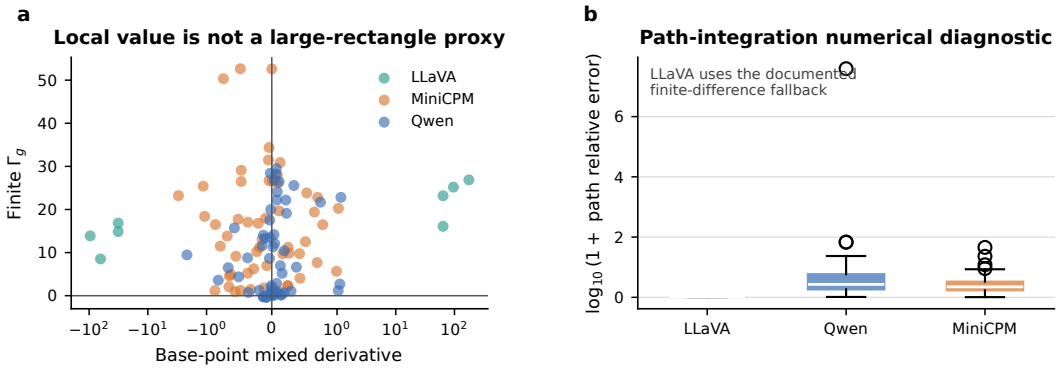

Figure 5: **Local and path-integrated numerical diagnostics.** A base-point mixed derivative can differ from a large finite contrast. The path-quadrature comparison additionally shows substantial numerical disagreement; LLaVA's fallback is not an independent confirmation. The theorem concerns the exact integral under its smoothness assumptions.

Table 8: **What the answer observable can see, and what it responds to.** Family medians. Effective rank is bounded above by two, the number of declared candidates. Normalized energy is the displacement-normalized directional functional energy along the semantic, nuisance, and random directions; the last column counts families in which the answer-changing direction has more local energy than the answer-preserving one. The LLaVA schema stores no random control. These are 8–16 families per slice: the consistency of the pattern across slices, not the precision of any single median, is what the table supports.

Model	Modality	n	Eff. rank	Semantic	Nuisance	Random	Sem. > nuis.
LLaVA	Image	8	1.514	7.6e−7	2.3e−7	—	6/8
Qwen	Image	16	1.549	1.1e−5	1.3e−5	5.0e−9	6/16
	Video	16	1.606	4.1e−7	8.7e−7	6.8e−9	5/16
	Audio	16	1.546	3.2e−7	8.7e−7	7.2e−9	6/16
	Speech	16	1.474	5.5e−7	2.7e−7	8.9e−9	10/16
MiniCPM	Image	16	1.901	2.7e−5	9.5e−5	1.9e−5	4/16
	Video	16	1.833	2.6e−5	4.6e−5	2.7e−5	7/16
	Audio	16	1.721	8.0e−4	1.0e−3	1.7e−5	6/16
	Speech	16	1.758	1.5e−3	6.0e−3	1.4e−5	3/16

I.4 Distinctness from Distributional Synergy

The closest existing account of multimodal interaction is an information-theoretic decomposition of the same data. We compare against it exactly, on the evaluated panel, and on separately scored reworded families.

Exact separation. For the matching rule of Proposition 14, changing only how often each question is asked moves MMI synergy from one bit to 0.0114 bits while the model, its four responses and Γ_s are unchanged (Fig. 6c); conversely, rescaling the score rescales Γ while every PID atom stays fixed. A conclusion drawn from a synergy estimate can therefore change when the evaluation set is resampled, and one drawn from Γ can change when the same responses are scored on a different observable. The result concerns the stated MMI construction; it does not claim that PID is invalid or that every PID definition behaves this way.

The learned PID estimator. On the panel we use the BATCH/CE-Alignment estimator of Liang et al. (2023), with an 80/20 family partition, five training seeds and 25 evaluation contexts per seed. Its family rankings are repeatable after fitting, with split-half correlation above 0.96 in every slice, but unstable across retraining: against a cross-seed Spearman threshold of 0.5, eight of nine slices fail, with correlations from 0.086 to 0.354 (Tab. 11). Only stability across fits supports a family-level ranking, so the empirical comparisons below are descriptive. The pooled association between

Table 9: **Standard representation summaries are functions of the preprocessing choice.** Family medians over the raw, centered, diagonally whitened, and norm-matched variants of the same modality tokens; each row reports the spread induced by that choice alone. Whitening uses a fixed vocabulary reference, so this is a specified family of transformations rather than an arbitrary change of basis. Twenty-two Qwen video records have undefined CKA and are excluded from that column.

Model	Modality	n	Euclid. max/min	CKA range	Cosine range
LLaVA	Image	64	8,347	0.036	0.025
Qwen	Image	64	4,019	0.100	0.037
	Video	64	3,367	0.146	0.025
	Audio	64	5,439	0.124	0.045
	Speech	64	3,118	0.228	0.049
MiniCPM	Image	64	1,210	0.047	0.053
	Video	64	839	0.045	0.081
	Audio	64	390	0.046	0.033
	Speech	64	324	0.096	0.050

estimated synergy and $|\Gamma|$ is -0.211 over 1,004 evaluations, and its sign varies by modality (Fig. 6). Negative learned synergy estimates occur and should not be read as negative exact MMI synergy.

On separately scored reworded families. We split 147 families by whether their source $|\Gamma|$ and estimated synergy lie above or below their slice's median, using per-slice thresholds fixed in an earlier disagreement analysis, and score four-cell success on independently reworded questions (Tab. 10). High-$|\Gamma|$ families succeed more often than low-$|\Gamma|$ families in all five slices, by 0.125 to 0.857, with a slice-stratified difference of 0.359 [0.257, 0.467]. The synergy contrast is -0.011 [-0.126, 0.100], and its per-slice sign is inconsistent. On the same pairs, interaction has rank correlation 0.985 with its reworded value and estimated synergy -0.049; in the one stability-passing slice, Qwen video ($n = 26$), the correlations are 0.775 and 0.029. Interaction does not, however, improve on ordinary accuracy: families that succeed at the source succeed again on 49 of 53 reworded targets and source failures on 3 of 94, and among source failures the stratified $|\Gamma|$ difference falls to 0.064 [0.000, 0.160]. The reworded targets reuse the source scenes, so the comparison concerns persistence under task-preserving rewording rather than generalization to new scenes. Adding the source measurements to ordinary diagnostics, over 34 grouped folds, changes binary AUROC by 0.00282 [-0.01204, 0.01958], so the profile does not improve per-example failure prediction either.

Table 10: **Held-out reworded success, split within each slice.** Counts are four-cell successes on separately scored paraphrased questions, among families above and below their own slice's median source $|\Gamma|$ and median PID synergy. Thresholds are reused unmodified from the earlier frozen analysis. MiniCPM paraphrase inference was not completed, so five slices enter the comparison.

Slice	High $\|\Gamma\|$	Low $\|\Gamma\|$	High PID	Low PID
LLaVA image	2/16	0/23	1/20	1/19
Qwen image	19/19	13/20	14/16	18/23
Qwen video	3/9	0/17	0/9	3/17
Qwen audio	7/15	0/14	3/13	4/16
Qwen speech	7/7	1/7	4/7	4/7
Stratified difference	$0.359\ [0.257, 0.467]$		$-0.011\ [-0.126, 0.100]$	

I.5 Stability of the Coordinate

Rewording. The paraphrase set rerenders each family's questions from frozen templates for all 22 dimensions. Automated checks find no cell or cross-cell-distinctness failures among 16,000 and 4,000 checks, and a stratified manual sample passes 50 of 50. The model-scored panel covers 200 LLaVA image families, 200 Qwen image families and 128 Qwen families in each other modality; MiniCPM is not part of it. Over 784 evaluations of 584 underlying families, signed interaction has rank correlation 0.9885 [0.9854, 0.9908] with its reworded value, and absolute interaction 0.9862 [0.9830, 0.9885]. Intervals resample underlying families jointly across models, stratified by modal-

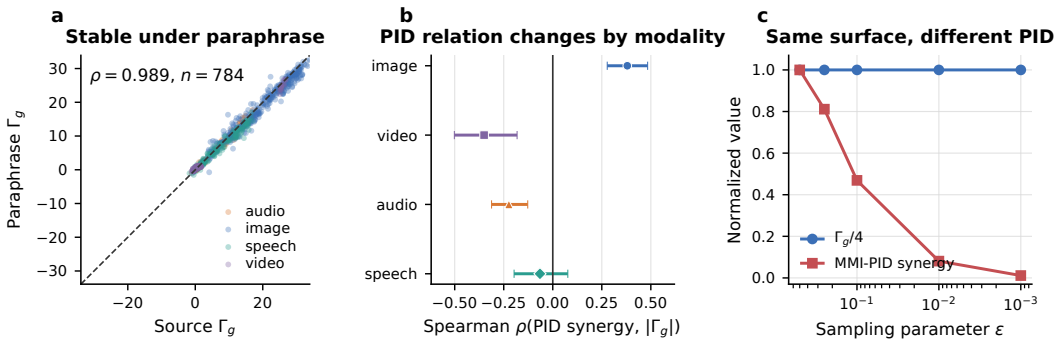

Figure 6: **Stability and the distinction from distributional synergy.** (a) Interaction is stable under the validated paraphrases. (b) Its empirical association with the evaluated PID estimate varies by modality; most PID slices have low retraining stability, so these correlations are descriptive. (c) In the exact matching-rule example, changing the input distribution changes MMI-PID synergy while leaving the four response values and interaction fixed. The figure's score notation g corresponds to s in this paper.

Table 11: **PID family-ranking stability.** The declared cross-seed Spearman threshold is 0.5. Only Qwen video passes.

Slice	Held-out families	Cross-seed ρ	Fixed-fit split-half
LLaVA image	100	0.096	0.987
Qwen image	100	0.265	0.986
Qwen video	100	0.666	0.992
Qwen audio	400	0.086	0.994
Qwen speech	200	0.354	0.987
MiniCPM image	26	0.304	0.996
MiniCPM video	26	0.093	0.961
MiniCPM audio	26	0.312	0.995
MiniCPM speech	26	0.333	0.978

ity, and recompute ranks in every draw. Within modality the correlations are 0.983 for image, 0.980 for audio, 0.979 for speech and 0.806 for video (Tab. 13). The paraphrases preserve the task semantics of the same scenes; they do not test arbitrary reformulations or new scenes.

Answer order and selected-condition wording. Swapping the order of "yes" and "no" in 64 prompts preserves every interaction sign, with rank correlation 0.9986, although the median magnitude changes by 9.4%. A separate selected-condition wording control, over 64 LLaVA and 32 families per omni-modal slice, tests answer stability rather than the four-cell contrast: several image and video slices keep every prediction, whereas MiniCPM audio agrees on only 15 of 32, with mean JS 0.26 (Tab. 12).

Execution and readout. Three identical executions of ten families per slice give zero variability, which sets the numerical floor for these runs without establishing determinism under other hardware or decoding settings. At the final layer, the layer readout's signed correlation with the output interaction exceeds 0.9996 on 1,536 comparisons. This confirms that the readout of Sec. 6.1 is implemented correctly; it is not evidence about intermediate states.

I.6 Layer Readouts, Onset, and Interventions

Readout and onset. Fig. 2 passes the pre-answer residual state at each layer through the model's final normalization and unembedding and computes Γ from the resulting yes-minus-no scores. Onset is the first depth at which the median interaction curve exceeds both the execution threshold and a quarter of its peak for three consecutive measured layers. The peak-relative term is needed because repeated executions have zero jitter: the floor alone is 10^{-6} and marks the first layer at which the

Table 12: **Wording stability depends on the measured outcome.** This 320-row control rephrases one selected canonical condition. It is separate from the 784-family full-rectangle paraphrase study and includes MiniCPM. Agreement is between predictions, not a correctness rate.

Slice	n	Agreement	Median $\lvert\Delta s\rvert$	Mean JS
LLaVA image	64	1	0.3438	$7.727e-05$
Qwen image	32	1	1.078	0.001897
Qwen video	32	1	0.3125	0.002157
Qwen audio	32	0.9688	0.2344	0.001835
Qwen speech	32	0.9062	0.625	0.01435
MiniCPM image	32	1	2.418	0.00254
MiniCPM video	32	0.9688	1.848	0.03929
MiniCPM audio	32	0.4688	3.231	0.2604
MiniCPM speech	32	0.875	2.688	0.09949

Table 13: **Paraphrase persistence with underlying-family resampling.** Shared stimuli remain grouped across models; ranks are recomputed within every bootstrap draw. Modality counts are fixed by stratified resampling.

Panel	Contrast	Evaluations	Families	Spearman [95% interval]
Pooled	signed	784	584	0.9885 [0.9854, 0.9908]
Pooled	absolute	784	584	0.9862 [0.9830, 0.9885]
Image	signed	400	200	0.9835 [0.9779, 0.9870]
Image	absolute	400	200	0.9835 [0.9777, 0.9868]
Video	signed	128	128	0.8060 [0.7142, 0.8715]
Video	absolute	128	128	0.6344 [0.4970, 0.7476]
Audio	signed	128	128	0.9801 [0.9662, 0.9870]
Audio	absolute	128	128	0.9764 [0.9586, 0.9854]
Speech	signed	128	128	0.9792 [0.9653, 0.9848]
Speech	absolute	128	128	0.9780 [0.9649, 0.9835]

median curve leaves bfloat16 rounding. Under a floor-only rule, Qwen's speech onset falls at 22% of depth, where the median interaction is below 0.2% of its peak, and the modalities appear ordered with speech first and video last; under the peak-relative rule every slice lies between 74% and 81%. Tab. 15 reports both definitions, the bootstrap interval of the onset, and the failed-to-solved peak ratio of Sec. 6.1. In the first half of the network no slice exceeds 4% of its peak. MiniCPM audio, with 3 solved families in the layer sample, and LLaVA image, with 15, have the least stable peak ratios.

Interventions. The four activation vectors at a layer decompose exactly as $h_{ue} = \mu + x_u a + x_e b + x_u x_e d$ with $d = (h_{11} - h_{10} - h_{01} + h_{00})/4$ (Proposition 15). Erasure sets d to zero and transplantation replaces it with a donor's, preserving the activation grand mean and both activation main effects, though not the output main effects after nonlinear downstream computation. Over 127 recipients, seven layers and five methods, 4,445 interventions carry factorial, main-effect and negative controls. The pre-specified intervention at layer 14, at 52% of depth and before the band in which interaction forms, changes no decisions (Tab. 14). At layer 26, inside that band, erasure changes the decision in all 76 targeted Qwen families and transplantation in 11 of 21, but off-target effects are as large as on-target ones, with ratios between 0.86 and 1.32, and main-effect controls are equally disruptive (Fig. 7). Late layers are therefore where the answer is sensitive, not where interaction is selectively encoded.

Patch recovery. The signed recovery at family i, layer ℓ and token group G is

$$\rho_{i,\ell,G} = \frac{g_{i,\mathrm{patch}(\ell,G)} - g_{i,\mathrm{cf}}}{g_{i,\mathrm{clean}} - g_{i,\mathrm{cf}} + \varepsilon}. \tag{69}$$

The profile summarizes the positive recovery curves of the modality, question and pre-answer groups by their harmonic mean, and keeps signed curves, negative recovery and over-recovery alongside it. Recovery is unbounded and a small denominator can destabilize it, as in Qwen video's estimate of

Table 14: **Primary factorial intervention at layer 14.** Effects are oriented toward erasing successes or rescuing failures.

Model	Operation	n	Mean effect [95% CI]	Transitions
Qwen	Erase successes	76	-0.0025 [$-0.0148, 0.0099$]	0
Qwen	Transplant failures	21	0.0030 [$-0.0119, 0.0179$]	0
LLaVA	Erase successes	2	0.0000 [0,0]	0
LLaVA	Transplant failures	28	0.0011 [$-0.0167, 0.0201$]	0

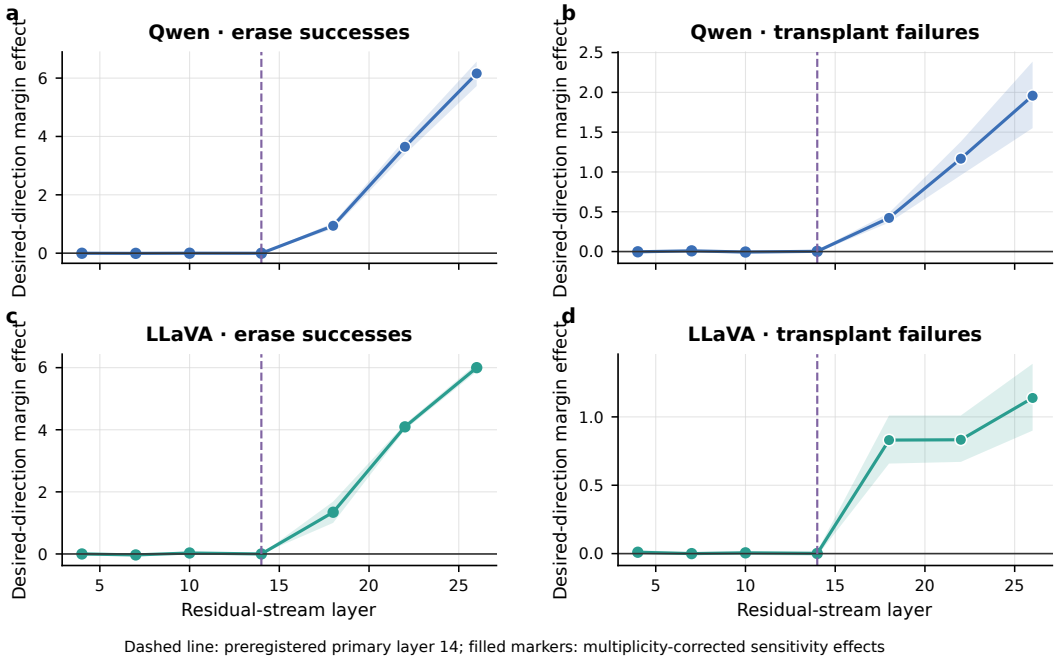

Figure 7: **Layer sensitivity of factorial erasure and transplantation.** The dashed line marks the primary layer. Later target-direction effects survive the reported multiplicity correction, but off-target and main-effect controls prevent a selective causal interpretation.

about 2,720, so it does not rank strength on its own. Recovery also depends on the corruption: in the patching panel of 352 family records, LLaVA's modality-group recovery exceeds its question-group recovery under a semantic-counterfactual corruption (0.541 versus 0.056), whereas mean-token corruption gives similar medians (0.231 and 0.237; Table 16). Natural-task patching covers 32 examples per model–dataset slice but lacks matched irrelevant-question controls and does not mediate four-cell interaction.

I.7 Modality and Model Effects on Evidence Use

Over the eight omni-modal slices, a two-way decomposition of the profile coordinates into modality and model effects attributes 88% [78%, 92%] of the variance in log accessibility to the modality and 2% [0%, 4%] to the model, and 91% [67%, 96%] and 0% [0%, 4%] of the variance in interaction (Tab. 7). The modality factor has four levels and the model factor two, so the shares are not adjusted for degrees of freedom, and each modality carries its own semantic tasks. The ratio comparison in Sec. 6.2 does not depend on either point: within a modality the two models differ in accessibility by at most a factor of 6.9, while across modalities accessibility spans a factor of 155.

I.8 Profile Cards and Normalization

The cards. Each of the nine slice cards and three aggregate cards stores point estimates, 95% intervals, the scale S_f (the median absolute clean answer margin on development data) and reasons

Table 15: **Internal and reliability coordinates for every slice.** Onset is the first depth at which the readout's median interaction exceeds a quarter of its peak for three consecutive layers, with a 2,000-draw family bootstrap interval. The floor-only onset of the original cards is shown for comparison. Peak ratio divides the peak median interaction of failed families by that of solved families. Mediation is unbounded; Qwen video's large estimate is denominator-sensitive. Execution variability is zero in the tested repetitions.

Slice	Onset [95% interval]	Floor-only onset	Peak ratio	C_{med}	V_{exec}
LLaVA image	0.81 [0.74, 0.81]	0.44	0.78	0.15	0
Qwen image	0.74 [0.74, 0.74]	0.56	0.42	0.26	0
Qwen video	0.74 [0.74, 0.74]	0.59	0.05	2,720	0
Qwen audio	0.74 [0.74, 0.74]	0.41	0.11	0.26	0
Qwen speech	0.74 [0.70, 0.74]	0.22	0.27	0.30	0
MiniCPM image	0.77 [0.77, 0.80]	0.49	0.69	0.25	0
MiniCPM video	0.77 [0.69, 0.83]	0.49	0.06	0.62	0
MiniCPM audio	0.74 [0.63, 0.77]	0.34	0.23	0.26	0
MiniCPM speech	0.77 [0.77, 0.77]	0.23	0.29	0.32	0

Table 16: **Recovery depends on the corruption baseline.** LLaVA image, 32 families. Each entry is the median across families of the mean signed recovery over measured layers. This differs from the positive-curve harmonic mediation coordinate. Negative and greater-than-one recoveries are retained in the full compendium.

Patched group	Counterfactual	Mean	Resampled	Zero
Modality	0.541	0.2309	0.1956	−0.0107
Question	0.0562	0.2367	0.2303	0.1694
Pre-answer	0.2893	0.2893	0.2346	−0.3439

for missing values. Table 17 reports all seven coordinates for five slices, and Tab. 15 gives mediation, onset and execution for all nine. Aggregate cards take medians of modality-specific coordinates and cannot substitute for matched slice comparisons; since no two full slices have matched accuracy, Appendix I.10 compares models on the same families instead. For a task-sufficient text surrogate $a(x)$, the surrogate gap of a family is

$$g_i^{\mathrm{surr}} = \mathrm{JS}(P(\cdot \mid U(x), E(q)), P(\cdot \mid E(a(x)), E(q))) \,, \tag{70}$$

reported as its median divided by S_f. If both channels factor through the same task statistic and a shared response kernel, the gap is zero; the converse need not hold, and two channels can agree while both are wrong.

The same success can hide different evidence use. MiniCPM solves 48.4% of families in both image and speech. Image has larger raw semantic JS (0.551 versus 0.443) and interaction (22.6 versus 11.7), whereas speech has a larger surrogate gap (0.0837 versus 0.0002); MiniCPM audio combines weak semantic separation (0.0144) with a large surrogate gap (0.822). The tasks differ, so this is not a ranking of modality ability, but one success rate accompanies different sensitivity and channel agreement.

Normalization. Dividing by S_f can change orderings between slices: across 36 slice pairs it reverses six accessibility, eight interaction and one surrogate-gap ordering, and raw-versus-scaled rank correlations across the nine slices are 0.833, 0.700 and 0.983 (Tab. 18). Between models within a modality, only one interaction ordering reverses. In MiniCPM, for example, speech has larger normalized accessibility and interaction than image despite smaller raw values, because its margin scale is smaller, so raw values are always reported alongside normalized ones. Scaling cannot change rankings within a slice.

Nuisance reference and selectivity. For Qwen and MiniCPM, the nuisance records pool variants of both semantic states against the U_0E_0 reference, so half of them change the reference answer and cannot measure answer-preserving sensitivity. Selectivity therefore uses only the answer-preserving

Table 17: **All seven FIP components, reported alongside success.** Point estimates for five illustrative slices; selectivity uses matched answer-preserving U_0 nuisance variants. The clean margin scale S_f supplies the indented, relative-to-margin summaries A_{sem}, C_{ctx}, and G_{surr}; the preceding rows give their raw medians. L_{onset} is a fraction of measured depth. V_{exec} also uses S_f and is zero in these runs. Intervals are in Tables 15 and 19.

Component	LLaVA image	Qwen image	MiniCPM image	Qwen speech	MiniCPM speech
Success (%)	3.8	78.6	48.4	60.9	48.4
Scale S_f	8.922	8.250	10.359	2.438	3.578
Evidence use					
Accessibility: raw JS	0.0657	0.9145	0.5507	0.4843	0.4425
Scaled A_{sem}	0.0074	0.1108	0.0532	0.1987	0.1237
Conditionality: raw $\lvert\Gamma\rvert$	13.406	24.203	22.625	7.813	11.711
Scaled C_{ctx}	1.503	2.934	2.184	3.205	3.273
Selectivity $R_{\text{sel}}^{U_0}$	7.93×10^{4}	2.21×10^{5}	4.18×10^{5}	2.29×10^{3}	44.47
Response agreement and internal behavior					
Surrogate gap: raw JS	3.97×10^{-5}	9.19×10^{-4}	2.10×10^{-4}	0.00437	0.08366
Scaled G_{surr}	4.45×10^{-6}	1.11×10^{-4}	2.03×10^{-5}	0.00179	0.02338
Mediation C_{med}	0.146	0.259	0.247	0.300	0.316
Onset L_{onset}	0.81	0.74	0.77	0.74	0.77
Measurement reliability					
Execution V_{exec}	0	0	0	0	0

$R_{\text{sel}}^{U_0}$ uses answer-preserving variants relative to U_0E_0, family-median nuisance aggregation, and floor 10^{-8} in raw JS units. It is a restricted reference measurement; large image ratios reflect very small nuisance denominators. Raw terms and intervals appear in Tab. 19. Zero execution variability is an observed result under the tested repetitions, not an assumed value.

Table 18: **Raw and normalized quantities.** Raw values are family medians; each normalized value divides its raw counterpart by the stored slice scale S_f. Cross-modality comparisons differ in task content.

Slice	S_f	Raw JS	A_{sem}	Raw $\lvert\Gamma\rvert$	C_{ctx}	Raw gap	G_{surr}
LLaVA image	8.922	0.06569	0.007363	13.41	1.503	3.97e-05	4.45e-06
Qwen image	8.25	0.9145	0.1108	24.2	2.934	0.0009187	0.0001114
Qwen video	2.375	0.003047	0.001283	0.6875	0.2895	0.02762	0.01163
Qwen audio	1.312	0.04544	0.03462	1.938	1.476	0.1184	0.09021
Qwen speech	2.438	0.4843	0.1987	7.812	3.205	0.004369	0.001793
MiniCPM image	10.36	0.5507	0.05316	22.62	2.184	0.0002103	2.03e-05
MiniCPM video	4.579	0.01804	0.003941	5.688	1.242	0.005092	0.001112
MiniCPM audio	2.859	0.01435	0.005018	3.234	1.131	0.822	0.2875
MiniCPM speech	3.578	0.4425	0.1237	11.71	3.273	0.08366	0.02338

U_0 variants, as LLaVA's records already do: $R_{\text{sel}}^{U_0} = \text{median}_i\, a_i/(\text{median}_i\, n_i^{U_0} + 10^{-8})$, where $n_i^{U_0}$ is the median JS over family i's answer-preserving variants and the numerator and denominator are resampled jointly by family (Tab. 19). The convention matters: Qwen image median nuisance JS is 0.0342 over the pooled records and 5.31×10^{-5} over the answer-preserving subset. The restriction does not supply a symmetric U_1 experiment. The semantic median exceeds the nuisance median in every slice, but LLaVA's ratio of about 7.93×10^4 coexists with low accessibility because its nuisance response is nearly zero, which is why the profile is reported as a table with raw terms rather than as a ranking figure. Null choice matters in the same way: MiniCPM video's median full-minus-null margin is 7.33 against a blank input, 2.36 against a resampled input and -6.64 against mean tokens (Tab. 5).

Table 19: **Matched U_0-side selectivity.** First take each family's median over answer-preserving nuisance variants, then the median across families. $R_{\text{sel}}^{U_0}$ is the semantic median divided by this nuisance median plus 10^{-8}. Intervals jointly resample families 2,000 times with the floor fixed; the analysis does not reconstruct matched U_1 comparisons.

Slice	Families	Preserved variants	Semantic JS	Nuisance JS	$R_{\text{sel}}^{U_0}$	95% interval
LLaVA image	500	4500	0.06569	8.181e-07	7.932e+04	[4.454e+04, 1.792e+05]
Qwen image	500	4500	0.9145	4.125e-06	2.212e+05	[1.407e+05, 3.144e+05]
Qwen video	500	3000	0.003047	0.001767	1.724	[1.331, 2.067]
Qwen audio	2000	16000	0.04544	0.0004202	108.2	[95.35, 127.9]
Qwen speech	1000	8501	0.4843	0.0002112	2293	[1843, 2857]
MiniCPM image	128	1152	0.5507	1.306e-06	4.185e+05	[2.319e+05, 8.585e+05]
MiniCPM video	128	768	0.01804	0.0006112	29.52	[12.44, 109.7]
MiniCPM audio	128	1024	0.01435	0.00264	5.434	[3.176, 8.408]
MiniCPM speech	128	1088	0.4425	0.00995	44.47	[24.53, 88.86]

I.9 Calibration and the Two Kinds of Failure

A shared offset recovers decisions without changing interaction. One offset per slice is fitted on 100 calibration families and subtracted from every cell of the remaining families. LLaVA solves 14 of 400 remaining image families; after the offset its scores support all four answers on 266, rescuing 252 failures and losing none, with raw interaction unchanged (Tab. 20). Fig. 8 shows one such family, in which every score is positive but the two cells requiring "yes" score highest. Without refitting, the same offset raises LLaVA's success on previously scored paraphrases from 8 of 158 to 113 of 158. The operation also improves Qwen image and audio, leaves Qwen video unchanged, and loses slightly more families than it rescues on Qwen speech; Tab. 23 reports every slice, including MiniCPM, where only 28 calibration-disjoint families per modality remain. Across all 5,012 evaluations the offset changes 840 decisions, counting rescues and losses, and changes Γ by at most 8.9×10^{-16}; that audit requires strictly positive gold margins, whereas the tables use the canonical tie convention of Appendix H.

Two response signatures behind the same zero. A four-cell failure can mean that the question-dependent contrast is absent, so that no decision rule can recover it, or that it is present but every cell falls on one side of the threshold. The cleanest instance lies within one slice: MiniCPM solves none of its video repetition or video rotation families, whose median $|\Gamma|$ are 9.66 and 0.77 while their semantic response distances are 0.0132 and 0.0013. Across the panel (Tab. 22), LLaVA's failing image dimensions keep median $|\Gamma|$ between 8.66 and 13.00, whereas Qwen's failing video and audio dimensions and MiniCPM's video rotation stay below 1.10. Because each group is dominated by one model, the within-slice MiniCPM contrast is what shows that the signatures belong to the task rather than only to the system. Tables 6 and 21 place these failures beside successful dimensions.

Two fifths of controlled failures are decision failures. Applying Proposition 3 to every failed family, 1,250 of 3,094 failures (40.4%) admit a constant offset that makes all four cells correct. The share ranges from 17.0% for Qwen video and 24.7% for MiniCPM video, through 29.2% for Qwen audio, 31.8% for MiniCPM speech, 40.0% for MiniCPM audio and 41.9% for Qwen speech, to 65.4% for Qwen image, 78.8% for MiniCPM image and 84.2% for LLaVA image. Feasibility uses a separate optimal threshold for each labelled family, so it bounds what a shared offset can recover: 322 of LLaVA's 386 remaining failures are feasible and its shared offset rescues 252, while 56 of Qwen video's 334 are feasible and its offset rescues none. Neither accuracy, which is zero for both groups by construction, nor semantic response distance, which is below 0.08 in every cell of Tab. 22, separates the two signatures; the signed four-score contrast does.

Low-divergence failures can still be recovered. Among LLaVA's 386 remaining image failures, 179 have semantic JS below 0.05, yet 123 of these have separable score groups and 71 are rescued by the shared offset. At cutoffs of 0.025 and 0.01 the counts are 71 feasible and 32 rescued of 123, and 30 and 6 of 69 (Fig. 9). A saturated candidate distribution makes semantic edits look small even when the scores separate the answers, so low JS alone does not indicate missing information. The effect has limits: all 334 remaining Qwen video failures have JS below 0.05, only 56 are feasible, and none is rescued.

Table 20: **Accuracy can change while response interaction stays fixed.** Each correction subtracts one offset from every cell in its slice. Evaluation excludes that slice's 100 calibration families. Counts use the canonical tie convention. MiniCPM's small remaining partitions are shown in Appendix I.9.

Slice	n	Raw success	Corrected	Rescued	Lost
LLaVA image	400	14	266	252	0
Qwen image	400	317	332	32	17
Qwen video	400	66	66	0	0
Qwen audio	1,900	604	726	202	80
Qwen speech	900	545	540	26	31

A measured LLaVA rectangle

	E_0	E_1
U_0	12.28	1.00
U_1	1.94	14.58

All scores are positive: four-cell failure.
Subtract 5.44: both off-diagonals become negative.
$\Gamma \approx 23.92$ in both cases.

Calibration-disjoint LLaVA families

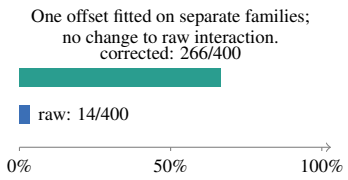

Figure 8: **A failure need not imply absent interaction.** Left: saved family `clevr_real_color_0010`, with rounded yes-minus-no scores. The diagonal has the larger scores even though every decision is "yes." The slice's frozen offset separates the two score groups. Right: the same correction evaluated across calibration-disjoint image families. This supports distinguishing decision bias from question-dependent response structure; it does not establish that every failure is calibratable.

I.10 Cross-Model Agreement and Matched Families

Agreement by accuracy and by threshold feasibility. We score each of the 22 semantic dimensions twice, by four-cell success and by threshold feasibility, which asks whether the four scores separate the answer groups regardless of where the decision boundary sits. A dimension is *contested* when one slice solves at least 50% of its families and another fewer than 5%. Six dimensions are contested by success and three by feasibility. Five resolve: material, size, left/right and front/behind through LLaVA's image failures, and event order through MiniCPM's audio failures. Two become contested, repetition and rotation, where feasibility separates models that success leaves tied at zero. LLaVA's image slice illustrates the effect: from success to feasibility, color rises from 0.12 to 1.00, material from 0.00 to 0.97, shape from 0.14 to 0.99, left/right from 0.00 to 0.85, size from 0.00 to 0.79, front/behind from 0.00 to 0.75 and count from 0.00 to 0.59. The dimensions contested under feasibility, direction, repetition and rotation, differ in what the models represent rather than in how they decide.

Identical answers can conceal different responses. We join model outputs on shared families for all six same-modality pairs, three image pairs and Qwen–MiniCPM for video, audio and speech, and keep cohorts with identical correctness. Cohorts are defined by outcome, not by any profile value, so they describe observed behavior rather than causal effects of architecture. In every pair with at least 19 shared successes, some response coordinate differs with a bootstrap interval excluding zero although every answer agrees, and the coordinate differs between pairs: semantic response distance for LLaVA–Qwen image (0.641 [0.604, 0.681]) and Qwen–MiniCPM image (-0.149 [-0.206, -0.094]), surrogate disagreement for Qwen–MiniCPM speech (0.035 [0.017, 0.052]), and both for Qwen–MiniCPM video (Tab. 24). On the 57 speech families both models solve, semantic JS does not separate them (0.7153 against 0.7578; difference 0.0425 [-0.0212, 0.1029]), but MiniCPM disagrees more with the text surrogate (0.0409 against 0.0063; difference 0.0346 [0.0171, 0.0523]) and moves more under nuisance edits (0.01637 against 0.00023; difference 0.01614 [0.00989, 0.02318]). On the 85 video families both models fail, semantic JS stays small (0.0028 and 0.0271) while the surrogate gap differs by 0.2112 [0.2048, 0.2185]; on the 42 shared speech failures it differs by 0.4658 [0.3686, 0.5568]. Larger nuisance JS measures response movement rather than an accuracy

Table 21: **Response measurements distinguish semantic tasks beyond their success rates.** Success counts require all four cells. Absolute interaction and semantic JS are raw family medians. Selected dimensions illustrate successful discrimination, weak response separation, and failures with retained score interaction; complete success counts are in Appendix I.12. Raw interaction magnitudes should be compared within the declared scoring interface.

Slice	Semantic dimension	Successes	$\lvert\Gamma\rvert$	Sem. JS
Qwen image	Color	71/72	29.17	0.9825
	Count	15/71	11.22	0.2192
Qwen video	Event order	83/83	26.03	0.9761
	Repetition	0/83	0.47	0.0007
Qwen audio	Event presence	267/400	11.87	0.5841
	Repetition	0/400	0.75	0.0136
Qwen speech	Keyword identity	161/167	13.72	0.7607
	Affirmation / negation	2/166	2.34	0.1433
MiniCPM video	Repetition	0/21	9.66	0.0132
	Rotation	0/21	0.77	0.0013
MiniCPM speech	Keyword identity	21/22	19.04	0.8981
	Affirmation / negation	1/22	3.23	0.1281

Table 22: **Two response signatures behind the same zero.** All slice–dimension cells with no four-cell successes, split by median absolute interaction. Both groups have near-zero semantic response distance, so probability-space separation does not distinguish them. The threshold at $\lvert\Gamma\rvert = 5$ is a presentational split of a continuum, not a fitted classifier.

Slice	Dimension	n	$\lvert\Gamma\rvert$	Sem. JS
Structure retained, decisions fail				
LLaVA image	Size	71	13.000	0.0619
LLaVA image	Material	72	12.750	0.0683
LLaVA image	Left / right	71	11.969	0.0773
MiniCPM video	Repetition	21	9.656	0.0132
LLaVA image	Front / behind	71	8.656	0.0186
Structure absent				
Qwen video	Rotation	83	1.063	0.0075
MiniCPM video	Rotation	21	0.773	0.0013
Qwen video	Direction	84	0.750	0.0017
Qwen audio	Repetition	400	0.750	0.0136
Qwen video	Duration	83	0.687	0.0034
Qwen video	Repetition	83	0.469	0.0007
Qwen audio	Simultaneous / sequential	400	0.438	0.0083
Qwen video	Speed	84	0.313	0.0006

loss, and larger surrogate JS measures channel disagreement rather than proof that the text channel is correct.

Estimands and intervals. Semantic, surrogate and nuisance distances are raw per-family JS; nuisance sensitivity uses the answer-preserving U_0 records. Intervals use 2,000 paired family bootstrap draws stratified by semantic dimension and correctness cohort; they are pointwise and not adjusted for multiplicity.

I.11 Natural Benchmarks

Unchanged accuracy, changed answers. Under the paper's candidate-scoring and null interface, 36,000 model–example evaluations span GQA (Hudson & Manning, 2019), POPE (Li et al., 2023), NExT-QA (Xiao et al., 2021), Clotho-AQA (Lipping et al., 2022), MMAU (Sakshi et al., 2024) and Spoken-SQuAD (Li et al., 2018) (Tab. 25); these are not official leaderboard scores. On Clotho-AQA, Qwen reaches 31.30% with the audio and 31.45% with a mean-token replacement, yet restoring the audio changes 926 of 2,000 predictions, 248 to correct and 251 to incorrect. On Spoken-

Table 23: **Complete frozen-offset comparison.** The last two columns exclude calibration families. Counts use canonical success.

Slice	Offset	All n	All raw	All corrected	Held-out raw	Corrected
LLaVA image	5.435	500	19	333	14/400	266/400
Qwen image	2.016	500	393	414	317/400	332/400
Qwen video	1.435	500	83	83	66/400	66/400
Qwen audio	0.882	2,000	639	760	604/1,900	726/1,900
Qwen speech	0.365	1,000	609	602	545/900	540/900
MiniCPM image	4.679	128	62	97	13/28	21/28
MiniCPM video	−0.954	128	43	43	9/28	9/28
MiniCPM audio	−3.362	128	8	22	3/28	6/28
MiniCPM speech	−1.292	128	62	56	12/28	9/28

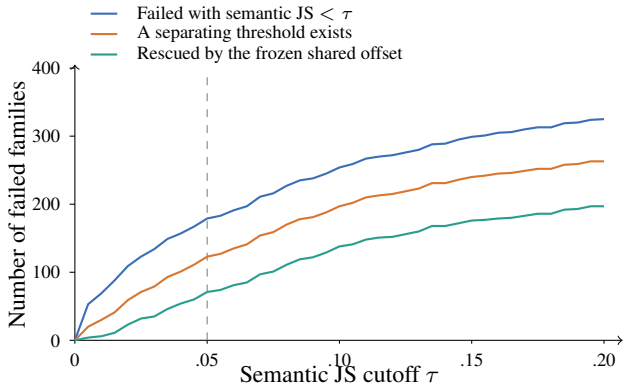

Figure 9: **Weak probability separation can conceal usable score structure.** Cumulative counts among 386 failed LLaVA image families excluded from calibration. At the dashed cutoff $\tau = .05$, 179 failures have low semantic JS; 123 admit a separating threshold and 71 are rescued by the fixed shared offset. The plot samples thresholds every .005; it is not a fitted rescue rule. All slices and the full threshold grid are retained in the analysis artifacts.

SQuAD, full input and a resampled null both give 284 of 1,000 correct answers, with 19 helpful and 19 harmful changes among 56. Helpful and harmful changes can cancel in aggregate, so a small full–null accuracy difference establishes neither that a model ignores the modality nor that the benchmark does not require it.

How much of a benchmark score survives modality removal. For each of 14 model–dataset combinations we report the accuracy gap between the full input and the strongest of three nulls (Tab. 25). Only POPE exceeds ten points; ten combinations have a gap below seven points, and two are negative, with Qwen scoring 0.3145 on Clotho-AQA with mean-token audio against 0.3130 with the recording, and 0.3100 on Spoken-SQuAD against 0.2840. The null also changes the answer: on Qwen GQA the gain from the image is 8.0 points over mean tokens but 0.4 points over a blank image, and MiniCPM video's median full-minus-null margin is 7.33, 2.36 or −6.64 depending on the null. A benchmark whose aggregate is insensitive to its modality can still measure something, but it cannot certify that a model uses that modality; that question needs controlled families. Text surrogates are used only where their information can be justified, and the Clotho study's 200-record stratified human audit does not certify every record of the 2,000-example evaluation.

I.12 The Capability Cliff in Full

A cliff, not a gradient. Pooled across the nine slices (Tab. 26), speed, repetition, rotation, duration and simultaneity are solved in 0 of 1,268 evaluations (1,133 distinct families), and repetition is empty in all four slices where it is tested (0 of 529). Affirmation/negation (3 of 188) and counting (19 of 161) are nearly empty and relative loudness follows (77 of 426), whereas keyword, named-entity

Table 24: **Different responses on the same families with the same correctness outcome.** Raw JS means for Qwen and MiniCPM; Δ is MiniCPM minus Qwen. Intervals are pointwise 95% paired family bootstraps stratified by semantic dimension, conditional on the observed outcome cohort. These are descriptive comparisons, not effects of model architecture.

Task	Cohort	n	Response	Qwen	MiniCPM	Δ [95% interval]
Video	Both correct	21	Semantic JS	0.9746	0.9960	0.0214 [0.0192, 0.0238]
Video	Both correct	21	Surrogate JS	2.50×10^{-5}	2.44×10^{-4}	2.19×10^{-4} [1.43×10^{-4}, 3.00×10^{-4}]
Video	Both correct	21	Nuisance JS	5.13×10^{-5}	1.16×10^{-4}	6.42×10^{-5} [2.83×10^{-5}, 1.03×10^{-4}]
Video	Both fail	85	Semantic JS	0.0028	0.0271	0.0243 [0.0131, 0.0378]
Video	Both fail	85	Surrogate JS	0.0272	0.2384	0.2112 [0.2048, 0.2185]
Video	Both fail	85	Nuisance JS	0.0027	0.0079	0.0051 [0.0033, 0.0072]
Speech	Both correct	57	Semantic JS	0.7153	0.7578	0.0425 [-0.0212, 0.1029]
Speech	Both correct	57	Surrogate JS	0.0063	0.0409	0.0346 [0.0171, 0.0523]
Speech	Both correct	57	Nuisance JS	2.26×10^{-4}	0.0164	0.0161 [0.0099, 0.0232]
Speech	Both fail	42	Semantic JS	0.1522	0.2182	0.0660 [-0.0175, 0.1596]
Speech	Both fail	42	Surrogate JS	0.0401	0.5059	0.4658 [0.3686, 0.5568]
Speech	Both fail	42	Nuisance JS	0.0011	0.0315	0.0305 [0.0196, 0.0431]

Table 25: **Modality accessibility of natural benchmarks under the declared interface.** Accuracy with the full input and with the strongest of three nulls (mean projected tokens, blank input, family-resampled input), over 36,000 model–example evaluations. The gap is full minus strongest null. These are repository candidate-scoring subsets under a fixed prompt and candidate set, not official leaderboard protocols; a small gap is a property of this evaluation of this benchmark, not proof that the task is solvable without the modality.

Model	Dataset	n	Full	Best null	Gap
Qwen	POPE	6,000	0.902	0.544	0.358
MiniCPM	POPE	6,000	0.885	0.551	0.334
LLaVA	POPE	6,000	0.807	0.523	0.284
MiniCPM	NExT-QA	2,000	0.265	0.198	0.067
LLaVA	GQA	2,000	0.329	0.272	0.057
MiniCPM	MMAU	1,000	0.445	0.408	0.037
Qwen	NExT-QA	2,000	0.240	0.204	0.036
MiniCPM	GQA	2,000	0.288	0.254	0.034
Qwen	MMAU	1,000	0.469	0.437	0.032
MiniCPM	Clotho-AQA	2,000	0.284	0.267	0.017
MiniCPM	Spoken-SQuAD	1,000	0.265	0.258	0.007
Qwen	GQA	2,000	0.336	0.332	0.004
Qwen	Clotho-AQA	2,000	0.313	0.315	−0.002
Qwen	Spoken-SQuAD	1,000	0.284	0.310	−0.026

and number identity are solved in 91–96% of families and event order in 75%. Identity and ordering are accessible; properties defined by counting or by comparing magnitudes over time are not. The failures recur across both time-based modalities and in models with different encoders, which a deficiency of one encoder does not explain. Slice by slice, 20 of 55 slice–dimension cells are at exactly zero, 15 exceed 75%, and only 10 lie between 25% and 75% (Tab. 28). Qwen video is the extreme case: event order is solved in 83 of 83 families and the other five dimensions in 0 of 417, so its aggregate accuracy of 16.6% is the mixing ratio of one solved dimension and five empty ones rather than a measure of moderate competence. A zero count means that no family passed all four cells, not that every cell was answered incorrectly.

An interaction floor. Of the 5,012 evaluations, 1,132 have median $|\Gamma| < 1$ and exactly one of them is solved; below $|\Gamma| < 2$ the success rate is 0.013 over 1,687 families, and the smallest $|\Gamma|$ among all 1,918 successes is 0.938 (Tab. 27). This is the empirical form of Theorem 1. Because $|\Gamma|$ uses four candidate scores and no gold labels, it can flag a new checkerboard task as unsolved before any annotation. Empty dimensions are not all alike, however: their median $|\Gamma|$ can be near 0.4 or near 12 (Appendix I.9).

Table 26: **A capability cliff, not a gradient.** Every controlled semantic dimension, pooled over the model–modality slices in which it is evaluated, sorted by four-cell success. Interaction and semantic response distance are family medians. The measurements are constructed on the declared suite and scoring interface; they characterize these dimensions as operationalized here, not every possible test of the underlying concept. n counts model–family evaluations.

Semantic dimension	Slices	n	Success	$\lvert\Gamma\rvert$	Sem. JS
Speed	2	105	0/105	0.375	0.0010
Repetition	4	529	0/529	0.750	0.0082
Rotation	2	104	0/104	1.063	0.0043
Duration	2	105	0/105	1.188	0.0034
Simultaneous / sequential	2	425	0/425	0.500	0.0080
Affirmation / negation	2	188	3/188	2.375	0.1396
Count	3	161	19/161	8.594	0.0434
Relative loudness	2	426	77/426	1.719	0.0227
Direction	2	106	22/106	0.812	0.0035
Spoken ordering	2	187	60/187	3.062	0.0603
Front / behind	3	160	53/160	11.812	0.0745
Proposition truth	2	188	77/188	5.828	0.3964
Size	3	160	67/160	18.605	0.2579
Left / right	3	160	68/160	13.969	0.1860
Material	3	162	75/162	19.754	0.2207
Color	3	163	95/163	26.531	0.7493
Shape	3	162	97/162	27.156	0.6172
Event presence	2	426	274/426	11.641	0.5566
Event order	4	530	400/530	11.984	0.5157
Number identity	2	188	171/188	11.234	0.6432
Named-entity identity	2	188	178/188	12.531	0.7172
Keyword identity	2	189	182/189	13.969	0.7712

Table 27: **An interaction floor on solvability.** Cumulative counts over all 5,012 controlled evaluations. Interaction is a contrast of the four yes-minus-no scores and is computed without gold labels, so the floor is usable as a screen on an unannotated task. It is a necessary-condition screen: a large interaction does not imply success, as the zero-success dimensions with retained structure show.

$\lvert\Gamma\rvert$ below	Families	Solved	Rate
0.5	559	0	0.0000
1	1132	1	0.0009
2	1687	22	0.0130
3	1994	53	0.0266
4	2168	83	0.0383
5	2307	106	0.0459
6	2433	155	0.0637
8	2673	249	0.0932

The same dimension can be solved by one model and empty in another. Video direction is solved in 22 of 22 MiniCPM families and 0 of 84 Qwen families on the same construction, with median $\lvert\Gamma\rvert$ of 32.57 against 0.75 and median semantic JS of 0.8664 against 0.0017, while both models solve video event order perfectly. The difference is therefore specific to the property rather than to video competence.

I.13 Held-Out Generalization and Solvability

Design and pre-specified criteria. Two objections follow from Sec. 6.3: the empty dimensions might be a property of three similarly sized systems, or they might not be answerable at all. One experiment tests both. It evaluates a model outside the panel on the controlled suite and, in the same pass, rescores every family with the modality replaced by a text description of the same scene, holding the questions, labels and scoring fixed. The prediction is fixed before the model is loaded: it records the frozen panel's per-dimension outcomes, the dimensions expected to be dead and live,

Table 28: **Complete controlled success counts.** Entries are successful four-cell families / evaluated families. A dash indicates an unevaluated model–modality combination.

Modality	Semantic dimension	Qwen	MiniCPM	LLaVA
Image	Color	71/72	15/19	9/72
	Material	71/72	4/18	0/72
	Shape	71/72	16/18	10/72
	Size	63/71	4/18	0/71
	Left / right	58/71	10/18	0/71
	Front / behind	44/71	9/18	0/71
	Count	15/71	4/19	0/71
Video	Direction	0/84	22/22	–
	Speed	0/84	0/21	–
	Event order	83/83	21/21	–
	Repetition	0/83	0/21	–
	Rotation	0/83	0/21	–
	Duration	0/83	0/22	–
Audio	Event order	295/400	1/26	–
	Event presence	267/400	7/26	–
	Repetition	0/400	0/25	–
	Relative loudness	77/400	0/26	–
	Simultaneous / sequential	0/400	0/25	–
Speech	Keyword identity	161/167	21/22	–
	Named-entity identity	159/167	19/21	–
	Number identity	156/167	15/21	–
	Proposition truth	71/167	6/21	–
	Spoken order	60/166	0/21	–
	Affirmation / negation	2/166	1/22	–

and the pass criteria, and it is hashed together with the panel outputs it reads. The procedure checks that each dead dimension has zero panel successes and never overwrites an existing prediction. The criteria are success of at most 0.05 and median $|\Gamma| < 2$ on dead dimensions, success of at least 0.50 on live dimensions, and success of at least 0.50 from text on dead dimensions. No verdict is returned unless a live dimension is scored in the same run, because a model that always answers "no" satisfies every dead criterion.

Surrogates and nulls. Video descriptions come from the dataset generator's description function applied to the stored generator arguments of U_0 and U_1, so both states are described by the code that rendered them. Audio descriptions use the stored template and its declared inversion, filled with the generator's class labels so that the two descriptions do not differ in surface form. Surrogates are defined for video and audio; speech families serve only as live controls. Mean-token and resampling nulls are fitted on the first sixteen families of each dataset, stratified by dimension, as in the main experiments. The Phi-4 runs use a scoring-only adapter without null statistics, which the four-cell outcomes do not require.

The dimensions are answerable, and the cliff persists at scale. Qwen2.5-Omni was scored on 288 families over seven dimensions, Qwen3-Omni-30B-A3B on 384 over ten, and Phi-4-multimodal-instruct on 224 over seven (Tables 29 and 30). All five dimensions are solved from text in every family, 192 of 192 for each Qwen model, against 0–4.7% from the modality, so the questions, answer format, label structure and scoring do not produce the zeros. The control does not show that the media are unambiguous to a human observer. On Qwen3-Omni, with 4.3 times the parameters and a mixture-of-experts language model, the five dimensions stay at or below 4.7%, three at exactly zero, while the live controls reach 0.719 to 1.000.

The interaction prediction misses. Median $|\Gamma|$ on Qwen3-Omni is 2.906 for speed, 4.812 for repetition and 7.906 for simultaneity, above the pre-specified bound of 2 and nine to eighteen times the Qwen2.5-Omni values of 0.313, 0.562 and 0.438. In the terms of Sec. 6.2, these dimensions move from absent structure towards structure that the decision rule does not yet use. Accuracy

Table 29: **Held-out generalization and solvability.** Four-cell success from the modality and from a text description of the same scene, on the five dimensions with no successes anywhere in the frozen panel, with live controls below. Qwen3-Omni-30B-A3B is outside the panel: it is $4.3\times$ the parameter count and a later generation. The criteria were fixed before scoring: success $\leq .05$ and median $|\Gamma| < 2$ on dead dimensions, success $\geq .50$ on live dimensions and on the text control.

		Qwen2.5-Omni 7B		Qwen3-Omni 30B-A3B		
Dimension	n	Modality	Text	Modality	$\lvert\Gamma\rvert$	Text
Zero-success dimensions of the frozen panel						
Speed	32	0.000	1.000	0.000	2.906	1.000
Duration	32	0.000	1.000	0.000	1.500	1.000
Rotation	32	0.000	1.000	0.000	0.500	1.000
Repetition	64	0.000	1.000	0.047	4.812	1.000
Simultaneous / sequential	32	0.000	1.000	0.031	7.906	1.000
Live controls						
Event presence	32	0.719	1.000	0.719	18.625	1.000
Event order	64	0.891	1.000	0.953	30.234	1.000
Number identity	32	—	—	0.875	25.078	—
Named-entity identity	32	—	—	0.969	30.547	—
Keyword identity	32	—	—	1.000	28.375	—

reports both models at zero, so only the response measurement registers the change. We report the miss rather than adjusting the bound; two models on three dimensions do not establish how capabilities emerge in general or which training change produced the shift.

Backbone diversity is not established. All three panel models and Qwen3-Omni use a Qwen-family language model. Phi-4-multimodal-instruct, which does not, solves no audio family, including both live controls, where Qwen2.5-Omni reaches 0.891 and 0.719. Its audio does reach the response, with the yes-minus-no margin moving from $+1.50$ to -1.75 when the recording is removed in a verified example, but median $|\Gamma|$ stays at 0.250 across its 128 audio families. Because its live controls fail, its zeros on the dead dimensions carry no information about the cliff, and other language-model families remain untested. One held-out model is one model, and a dead dimension is a statement about this suite's operationalization of the property.

Table 30: **Phi-4-multimodal, complete counts.** Every audio dimension is at zero, including both live controls, where Qwen2.5-Omni reaches 0.891 and 0.719 on the same families. The speech dimensions are partially solved. Because the live controls fail, the audio zeros are uninformative about the capability cliff and are reported here rather than as supporting evidence.

Modality	Dimension	Role	n	Success	$\lvert\Gamma\rvert$
Audio	Event order	live	32	0.000	0.250
	Event presence	live	32	0.000	0.750
	Repetition	dead	32	0.000	0.250
	Simultaneous / sequential	dead	32	0.000	0.250
Speech	Keyword identity	live	32	0.094	5.250
	Number identity	live	32	0.500	7.000
	Named-entity identity	live	32	0.594	5.750

I.14 Further Results

Cost and coverage. The finite contrast needs four candidate scores and no probe fitting or differentiation. On a shared audit of 400 model–family evaluations on one H100, in which both pipelines score the same 1,600 canonical cells, it takes 787 seconds end to end against 2,459 seconds for the PID workflow, which estimates a different quantity and includes representation extraction and fitting (Appendix I.4). Internal and nuisance probes add cost. The reporting conditions are in Appendix G.5.

Table 31: **Modality benefit does not establish task-aligned interaction.** Benefit means a strictly positive gold-answer margin change at U_0E_0 relative to each of the blank, mean-token, and resampled nulls. The last three columns partition those benefit cases by signed interaction. Zero uses tolerance 10^{-8}; all such cases here are exactly zero in the stored scores. These are descriptive counts on the full controlled panel.

Slice	Evaluated	Benefit under all nulls	$\Gamma > 0$	$\Gamma = 0$	$\Gamma < 0$
LLaVA/image	500	418	417	0	1
Qwen/image	500	450	450	0	0
Qwen/video	500	370	299	4	67
Qwen/audio	2000	1392	1334	18	40
Qwen/speech	1000	857	850	0	7
MiniCPM/image	128	95	95	0	0
MiniCPM/video	128	36	36	0	0
MiniCPM/audio	128	85	85	0	0
MiniCPM/speech	128	76	75	0	1

Null-ablation benefit and interaction. A null comparison asks whether the modality improves the selected answer relative to a replacement; the mixed contrast asks whether its effect changes across questions in the direction the task requires. The two can disagree. Of Qwen's 500 video families, 370 have a strictly positive gold-margin change at U_0E_0 against each of the blank, mean-token and resampled nulls, yet 67 of these have negative and four zero interaction; in Qwen audio, 40 and 18 of 1,392 such families do (Tab. 31). Misalignment is 18.1% of benefit cases in Qwen video and 4.2% in Qwen audio, but 0.8% in Qwen speech, 0.2% in LLaVA image and zero in the other five slices. It is not explained by low competence alone, since MiniCPM audio, the weakest slice at 6% success, has none. Where the two measurements agree, the cheaper null comparison suffices; where they disagree, the signed contrast identifies families in which a favorable ablation result coexists with a wrongly oriented joint response. Zero interaction on one rectangle does not establish global additivity.

I.15 Limitations

Every result is relative to the declared answer observable and to this suite's operationalization of each semantic property, and every evaluated language model except Phi-4 belongs to the Qwen family. The measurement is not a universal prompt invariance: a separate selected-cell wording test agrees on only 47% of MiniCPM audio answers (Appendix I.5). It does not localize interaction to a selectively manipulable direction (Appendix I.6). It complements accuracy rather than replacing it: it does not improve per-example failure prediction beyond source behavior.