Modality Fault Lines: Structural Corruptions Reveal Fragile Omni-Modal Reasoning

Anonymous ACL submission

Abstract

Omni-modal large language models are increasingly evaluated on clean text–vision–audio inputs, where every channel is present, synchronized, and easy to read. These scores are often taken as evidence of robust cross-modal fusion, but clean evaluation alone cannot tell whether success depends on stable cross-modal structure or on cues that happen to be sufficient in the intact input. We study this uncertainty as a diagnostic problem and define a *modality fault line*: a boundary at which model behavior becomes unstable when a modality remains present and human-interpretable, yet the internal evidence structure of text, vision, or audio is perturbed. To probe these fault lines, we introduce **SCEVAL** (*Structure-Corruption Evaluation*), a diagnostic evaluation protocol that keeps the question, answer space, and modality channels fixed while applying controlled structural corruptions to text, vision, and audio individually and jointly. Built from 273 human-verified tri-modal examples from Social-IQ, OmniBench, and VALOR, **SCEVAL** evaluates proprietary and open-source omni-modal systems by reporting the mean accuracy over up to three random variants per stochastic condition, with the worst-variant gap logged as a diagnostic of seed-level variance. The results reveal a consistent pattern: structural corruption lowers clean accuracy, text–vision damage forms the most stable shared fault line, and multimodal degradation is non-additive rather than a simple function of how many channels are corrupted. Clean omni-modal accuracy shows that a model can answer when evidence is intact; **SCEVAL** asks whether that success remains reliable when cross-modal structure starts to break.

1 Introduction

Omni-modal large language models promise a simple interface for multimodal reasoning: a user provides a question together with visual and acoustic evidence, and one model produces a single answer grounded in all available streams (Hurst et al., 2024; Team et al., 2023; Team, 2026). This promise has made clean text–vision–audio benchmarks the default proof of progress—if accuracy rises when all modalities are present, the model is treated as a better cross-modal reasoner. Yet clean evaluation quietly gives the model the easiest possible version of the task: text, vision, and audio arrive simultaneously, well-formed, and mutually redundant, so a model can be correct without actually binding the three streams into a stable reasoning structure.

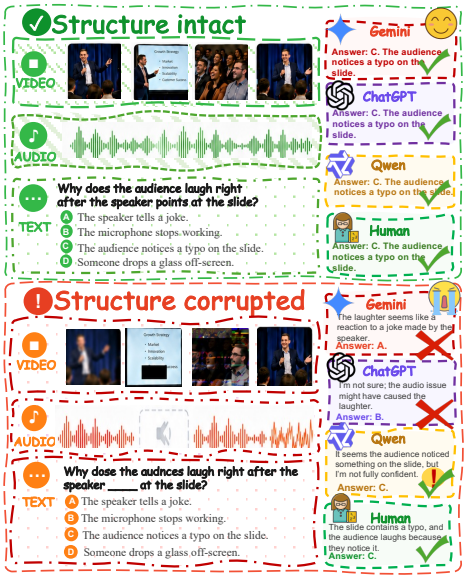

Figure 1: Teaser illustration of modality fault lines. **SCEVAL** keeps text, vision, and audio present while damaging their internal evidence structure, revealing failures that clean all-modality evaluation can hide.

Clean accuracy is not fusion ❶. Clean evaluation does not tell us *when* or *how* models use multiple modalities: a tri-modal item can be solved through a transcript cue, an obvious visual object, an audio event, an answer prior, or a shortcut that happens to agree with the label. Missing-modality ablations go to the opposite extreme—removing a channel entirely changes the task and

conflates modality reliance with distribution shift. The deployment-relevant region lies between: the modality is still present and human-interpretable, but its internal evidence structure has been damaged. We call the resulting hidden boundary a *modality fault line*—the point at which apparent omni-modal competence breaks once the channel a model implicitly trusts becomes noisy, fragmented, or structurally unreliable.

The missing diagnostic ❷. A useful robustness test for omni-modal models must answer three questions that standard protocols leave entangled:

- ***Which channel carries the answer?*** Clean accuracy cannot separate genuine cross-modal binding from single-channel shortcuts, because all evidence streams are simultaneously well-formed.
- ***What happens when a channel weakens but does not disappear?*** Real inputs degrade internally—OCR-like text errors or shuffled phrases, occluded or blurry frames, muted or distorted audio—rather than vanish outright.
- ***How do weakened channels interact?*** Single-modality studies reveal local sensitivity but cannot tell whether two damaged channels compensate, conflict, or trigger a new collapse mode.

SCEval: keep the modality, break the structure ❸. We introduce SCEVAL (*Structure-Corruption Evaluation*) to make this boundary measurable. The design choice is deliberately simple: do not remove a modality; keep it present and corrupt only its internal structure. From a 300-candidate pool sampled across Social-IQ (Zadeh et al., 2019), OmniBench (Li et al., 2026), and VALOR (Liu et al., 2024a), a third-party team verifies example validity, tri-modal answerability, and corrupted-input interpretability, yielding N=273 human-verified base examples. Each example is evaluated under the clean all-modality condition and under a curated family of single-, dual-, and tri-modal structural corruptions, with question, answer options, gold answer, and modality channels held fixed. Stochastic operators are aggregated as the mean over up to three random variants, and the worst-variant gap is logged alongside to surface seed-level fragility without making the primary metric a one-seed outlier.

Main Message The dominant failure mode we observe is asymmetric, not symmetric heavy-on-both: cross-modal binding collapses most severely when one channel *looks* intact while another is structurally broken, even though either channel in isolation would survive the same perturbation.

Our contributions are:

★ ***Phenomenon.*** We identify *modality fault lines*: latent failures that clean all-modality inputs and missing-modality ablations both miss, because the modality remains present while its internal evidence structure is damaged.

★ ***Protocol.*** We propose SCEVAL, a structure-corruption protocol that keeps the task and modality channels fixed while perturbing text, vision, and audio individually and jointly, with mean-variant aggregation and a worst-variant diagnostic.

★ ***Benchmark.*** We construct a human-verified tri-modal benchmark of 273 examples from Social-IQ, OmniBench, and VALOR, covering clean inputs and full single-, dual-, and tri-modal corruption variants under a unified directory and annotation scheme.

★ ***Findings.*** We show that structural corruption reliably lowers clean performance, that text–vision damage is the most stable shared fault line, and that cross-modal degradation is non-additive and model-dependent rather than a simple function of corruption count.

2 Structure-Corruption Evaluation

SCEVAL probes *modality fault lines* by holding the task and modality channels fixed while corrupting the internal structure of one or more modalities. The design choice that distinguishes it from missing-modality ablations is that every channel remains present and human-interpretable throughout, so the resulting drops measure structural fragility rather than distribution shift; the evaluation matrix combines single-modality severity curves with dual- and tri-modal joint corruptions over the same base examples.

2.1 Problem setup

Each example consists of a question q, an answer space (multiple-choice options or an answer target), and three evidence channels: text x_t, visual input x_v, and audio x_a. A clean omni-modal model predicts

$$\hat{y} = f(x_t, x_v, x_a, q), \tag{1}$$

and the clean all-modality baseline accuracy is computed on the uncorrupted version of the 273 base examples introduced in §2.3.

A *structure corruption operator* $c_m^{s,r}(\cdot)$ targets modality $m \in \{t, v, a\}$ at severity s and random variant r when applicable. The operator preserves the channel's container (the text string, the visual frames, or the audio waveform) and damages only its internal evidence—e.g. a text operator deletes or shuffles words rather than swapping the prompt, a visual operator adds noise or occludes regions rather than substituting the image, and an audio operator mutes or removes a segment rather than replacing the soundtrack. Combined corruption applies operators to multiple modalities of the same base example,

$$\tilde{x} = \big(c_t(x_t),\, c_v(x_v),\, c_a(x_a),\, q\big), \qquad (2)$$

where one, two, or three of the corruption operators may be active while q, the answer options, and the gold answer y remain fixed. This sample-paired design is what lets **SCEVAL** compute the clean-versus-corrupted contrast at the example level.

2.2 Corruption taxonomy

Figure 2 shows the fourteen operators **SCEVAL** uses, grouped into four text, seven visual, and three audio families. The taxonomy is deliberately *curated, not exhaustive*: rather than enumerating every textbook degradation, we select operators that target the harder structural fault lines we expect a robust omni-modal model to handle, while keeping every modality channel present and human-interpretable. Four design rules fix what an operator is allowed to do and how it is scored: (i) only *structural*, not semantic, damage; (ii) sample-level paired comparison with the same gold answer; (iii) graded severity in $\{10, 30, 50, 70\}$; and (iv) up-to-three stochastic random variants aggregated by the mean-variant rule of §2.4. Combined experiments use representative high-coverage dual- and tri-modal pressure conditions (text–vision, text–audio, vision–audio, and text–vision–audio). Per-operator definitions, severity parameterisations, and rejection rules are deferred to Appendix E (Table 24).

2.3 Benchmark construction

Source benchmarks. **SCEVAL** is built on three publicly released omni-modal benchmarks chosen to span different evidence styles: **Social-IQ** (Zadeh et al., 2019), a video QA benchmark for social reasoning that requires both speech and visual cues; **OmniBench** (Li et al., 2026), designed to test joint text–image–audio reasoning; and **VALOR** (Liu et al., 2024a), whose video clips contain naturally co-occurring visual and acoustic events. Mixing the three sources prevents the fault-line measurements from being driven by any single benchmark's idiosyncrasies.

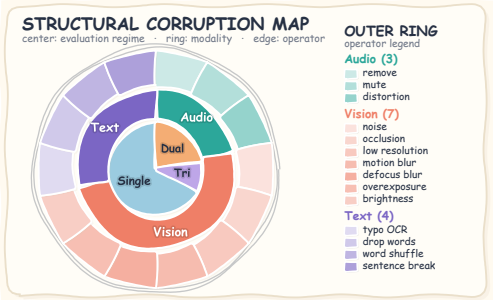

Figure 2: The fourteen structural corruption operators used by **SCEVAL** (4 text, 7 vision, 3 audio). The taxonomy is curated rather than exhaustive: it targets cross-modal evidence assembly (word identity/order, dense visual signal, audio presence) and excludes trivially recoverable perturbations; per-operator definitions, severity parameterisations, and rejection criteria are in Appendix E (Table 24).

Candidate pool. We sample 100 examples per source, yielding a 300-candidate pool. For every candidate we collect (i) the original video, (ii) the audio track extracted from the same recording, (iii) the textual question with multiple-choice options and a gold answer, and (iv) source metadata, organised under a unified directory and re-encoded to a common format so that the same operators apply uniformly across sources.

Example-level verification. A third-party annotation team verifies, for each candidate, (a) example validity (well-formed question, single defensible gold answer, all three modalities present in the clean condition); (b) tri-modal answerability of the gold answer from the union of the three clean modalities; and (c) the absence of annotator-side issues (broken media, ambiguous answer, modality leakage in text). Items failing any check, or flagged with annotator-side issues, are dropped at the original-sample level, leaving $N{=}273$ verified base examples that constitute the clean all-modality baseline used throughout the paper.

Cell-level corrupted-variant verification. The same team further judges every *corrupted variant* of audio and vision: of the 19,944 audio + vision variants sent for verification, 16,049 (80.5%) are retained as perceptually interpretable and contribute to the headline numbers. Text-side corruptions are deterministic and human-readable by construction,

Source	Modalities	# Examples
SRC Social-IQ	video, audio, text	100
OmniBench	image, audio, text	77
VALOR	video, audio, text	96
Total	—	VERIFY **273**

Table 1: Per-source composition of **SCEVAL**. From a 300-candidate pool (100 per source), third-party verification of example validity, tri-modal answerability, and corrupted-variant interpretability retains 273 examples as the base set.

so the four text operators are not subject to per-variant annotation. Per-operator pool sizes and retention rates are reported in Appendix D, Table 21.

Composition and release. Table 1 reports the per-source counts of the verified base set together with the modality channels each source natively provides. Fault-line analyses in §3 are reported on the pooled 273-example set; per-source breakdowns are kept in the appendix as a robustness check. The annotation manifest—per-sample validity flag, tri-modal answerability flag, and per-corruption interpretability flag—is released alongside the benchmark so that other groups can reconstruct the 273-example base set and audit the filtering rule.

2.4 Evaluation protocol

Mean random-variant aggregation. For stochastic corruptions, the same original sample is evaluated under up to three independently sampled random variants of the same operator at the same severity, and the reported accuracy is the *mean* across these variants so that the headline number reflects typical behaviour rather than any single seed. For model f, condition c, sample i, and valid variant $r \in R_{i,c}$, let $z_{i,r,c} = \mathbb{1}[f(c^r(x_i)) = y_i]$ be the variant-level correctness indicator. We aggregate at the (sample, variant) trial level,

$$\mathrm{Acc}(c) = \frac{1}{|I_c|} \sum_{i \in I_c} \frac{1}{|R_{i,c}|} \sum_{r \in R_{i,c}} z_{i,r,c}. \quad (3)$$

where I_c is the set of samples with at least one valid output. When every retained sample contributes the same number of variants this reduces to $\mathrm{Acc}(c) = |R|^{-1} \sum_{r \in R} \mathrm{Acc}_r(c)$, matching the intuition of "average accuracy over three random runs". Missing variants are excluded from $R_{i,c}$ rather than counted as correct; judged but unparseable responses are treated as missing for accuracy estimation and as incorrect for significance testing.

Worst-variant diagnostic. To expose how much of the headline mean is bolstered by lucky variants, we additionally report the worst-variant accuracy

$$\mathrm{Acc}^{\mathrm{worst}}(c) = \min_{r \in R} \mathrm{Acc}_r(c). \quad (4)$$

The gap $\mathrm{Acc}(c) - \mathrm{Acc}^{\mathrm{worst}}(c)$ quantifies seed-level variance: a small gap means the model degrades consistently across variants; a large gap signals that a single unlucky variant dominates the apparent fault. The mean−worst gap is reported per (model, modality) at each model's weakest single-modality cell in Table 15 (Appendix C.9).

Primary drop metrics. For any corruption condition c, the *clean-baseline drop*

$$\Delta_{\mathrm{base}}(c) = \mathrm{Acc}_{\mathrm{clean}} - \mathrm{Acc}(c) \quad (5)$$

measures how much performance is lost relative to the full clean input. For combined conditions, the *worst-single drop*

$$\Delta_{\mathrm{single}}(c) = \mathrm{Acc}_{\mathrm{worst\ single}}(c) - \mathrm{Acc}(c) \quad (6)$$

compares against the weakest corresponding single-modality corruption: positive Δ_{single} means the combined corruption is worse than every individual component (a compounded failure beyond the weakest single-modality condition), and negative Δ_{single} means the combined condition remains more accurate than its weakest single component.

Auxiliary diagnostics. The headline matrix is complemented by five smaller quantities, each isolating a different mechanism:

- **Fault-line score** $F_S = \mathrm{Acc}_{\mathrm{clean}} - \mathrm{Acc}(c_S)$ for $S \in \{tv, ta, va, tva\}$, a model-level summary of how much accuracy a given multi-modal corruption block costs (reported per family in Table 17).
- **Misleading-modality effect** $M_m = \mathrm{Acc}(\mathrm{missing}\ m) - \mathrm{Acc}(\mathrm{corrupted}\ m)$, positive when a structurally corrupted but present modality is more harmful than removing it (Table 8).
- **Fragility slope** S_m obtained by linearly regressing accuracy on severity $s \in \{0, 10, 30, 50, 70\}$ for each modality m (clean encoded as s=0); see Table 16.
- **Shortcut gap** $G = \mathrm{Acc}(\text{text-only}) - \mathrm{Acc}(\mathrm{random})$, how much clean accuracy can be reproduced from the linguistic channel alone (Table 8, **Text-only** column).

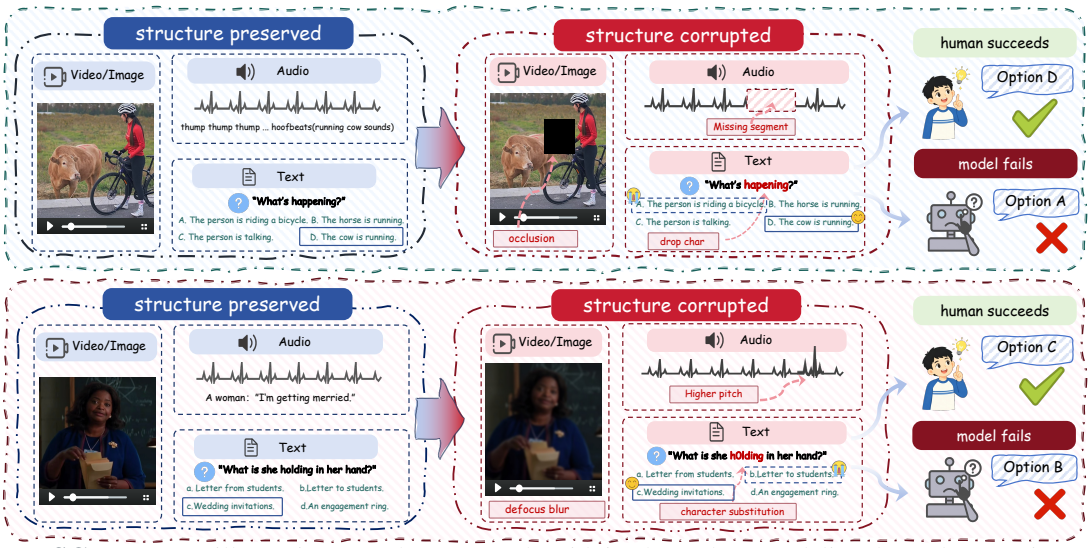

Figure 3: **SCEVAL** case illustration. One base example with its three clean modality channels (question/options, sampled frames, audio waveform) and one structurally corrupted variant per modality. Every variant keeps the gold answer, options, and modality channels fixed and damages only the internal evidence structure—distinguishing the fault-line probe from missing-modality ablations.

- **Modality trust bias** T_m, how often a model follows modality m under cross-modal mismatch (Appendix G.9).

We additionally report the invalid-output rate decomposed into API failures, parse failures, refusals, and empty outputs, since a model can fail either by choosing the wrong answer or by failing to produce a parseable answer under heavier structural stress.

Statistical protocol. Every reported accuracy is accompanied by a 95% confidence interval from $B{=}1000$ sample-level bootstrap (Efron, 1992) resamples over the $|I_c|$ original examples, applied after mean variant aggregation so that within-sample variant variance and between-sample variance are propagated into a single interval. Paired comparisons (clean vs. corrupted, single vs. combined, two models on the same condition) use the paired bootstrap on per-sample correctness differences and the paired McNemar test (McNemar, 1947) on the corresponding contingency table; both are reported with two-sided p-values and Holm–Bonferroni correction (Holm, 1979) across the family of comparisons inside one experimental block. Invalid outputs are treated as incorrect for significance testing and as missing for accuracy estimation; we report both versions whenever the conclusion is sensitive to this choice.

3 Experimental Results

We evaluate **SCEVAL** on a 15-model panel of seven proprietary/API systems and eight open or open-API systems (roster, interfaces, and the frame-extracted visual-input protocol in Appendix F). All headline runs use one standardised multiple-choice JSON prompt (Appendix H); accuracy on each stochastic condition is the mean over up to three random variants per operator (§2.4), reported alongside coverage, invalid-output rates, and 95% sample-level bootstrap intervals. Section 3.1 reads the $15{\times}14$ severity-70 single-modality matrix and its per-model severity curves; Section 3.2 probes whether stacking corruption across two or three channels compounds the damage.

3.1 Single-modality corruption

Table 2 reports clean accuracy together with the severity-70 drop on each of the fourteen single-modality operators for all fifteen systems. Two patterns dominate the matrix, and both turn out to be largely band-invariant.

Fragility concentrates in a small set of *structural* operators. Five of the fourteen sit clearly above the rest at ≥ 7 pp panel-mean drop—**drop_words**, **word_shuffle**, **noise**, and the two audio-segment operators **remove/mute**—each targeting word identity, dense visual signal, or audio presence. The remaining nine stay below 5 pp; three of them (**low_resolution**, **defocus_blur**, **distortion**) hover at zero or flip into small apparent gains, suggesting that cosmetic perturbations rarely reach the model's evidence-assembly path. The two single hardest cells in the matrix—**noise** on Baichuan-Omni-1.5 (-15.70 pp) and **word_shuffle** on Qwen2.5-

		Text				Vision							Audio		
Model	**Clean**	Typo	Drop	Shuffle	Break	Noise	Occ.	LowRes	MBlur	DBlur	Expo.	Bright.	Remove	Mute	Distort
Proprietary / API omni-modal models															
Gemini 3.1 Pro	84.15	**79.85** $_{\downarrow 4.30}$	**72.57** $_{\downarrow 11.58}$	**75.54** $_{\downarrow 8.61}$	**80.72** $_{\downarrow 3.43}$	**73.57** $_{\downarrow 10.58}$	**82.88** $_{\downarrow 1.27}$	**85.07** $_{\uparrow -0.92}$	**82.29** $_{\downarrow 1.86}$	**83.70** $_{\downarrow 0.45}$	**85.08** $_{\uparrow -0.93}$	**84.88** $_{\uparrow -0.73}$	**78.89** $_{\downarrow 5.26}$	**80.62** $_{\downarrow 3.53}$	**86.17** $_{\uparrow -2.02}$
Gemini 3 Pro	82.40	**76.72** $_{\downarrow 5.68}$	**71.30** $_{\downarrow 11.10}$	**73.16** $_{\downarrow 9.24}$	**80.27** $_{\downarrow 2.13}$	**73.00** $_{\downarrow 9.40}$	**80.09** $_{\downarrow 2.31}$	**83.80** $_{\uparrow -1.40}$	**81.41** $_{\downarrow 0.99}$	**82.89** $_{\uparrow -0.49}$	**80.49** $_{\downarrow 1.91}$	**83.88** $_{\uparrow -1.48}$	**76.36** $_{\downarrow 6.04}$	**75.82** $_{\downarrow 6.58}$	**82.51** $_{\uparrow -0.11}$
Gemini 3 Flash	80.95	**71.06** $_{\downarrow 9.89}$	**67.40** $_{\downarrow 13.55}$	**70.33** $_{\downarrow 10.62}$	**72.89** $_{\downarrow 8.06}$	**67.79** $_{\downarrow 13.16}$	**80.33** $_{\downarrow 0.62}$	**80.97** $_{\uparrow 0.02}$	**77.59** $_{\downarrow 3.36}$	**78.95** $_{\downarrow 2.00}$	**79.37** $_{\downarrow 1.58}$	**80.30** $_{\downarrow 0.65}$	**74.30** $_{\downarrow 6.65}$	**73.24** $_{\downarrow 7.71}$	**79.17** $_{\downarrow 1.78}$
Gemini 3.5 Flash	78.39	**72.53** $_{\downarrow 5.86}$	**67.03** $_{\downarrow 11.36}$	**68.50** $_{\downarrow 9.89}$	**73.63** $_{\downarrow 4.76}$	**62.75** $_{\downarrow 15.64}$	**75.14** $_{\downarrow 3.25}$	**78.96** $_{\uparrow -0.57}$	**75.56** $_{\downarrow 2.83}$	**77.47** $_{\downarrow 0.92}$	**79.30** $_{\uparrow -0.91}$	**78.31** $_{\downarrow 0.08}$	**73.45** $_{\downarrow 4.94}$	**71.00** $_{\downarrow 7.39}$	**77.83** $_{\downarrow 0.56}$
Gemini 2.5 Pro	79.12	**75.32** $_{\downarrow 3.80}$	**67.41** $_{\downarrow 11.71}$	**68.40** $_{\downarrow 10.72}$	**71.77** $_{\downarrow 7.35}$	**70.26** $_{\downarrow 8.86}$	**76.97** $_{\downarrow 2.15}$	**79.73** $_{\uparrow -0.61}$	**77.40** $_{\downarrow 1.72}$	**80.30** $_{\uparrow -1.18}$	**79.44** $_{\uparrow -0.32}$	**79.19** $_{\uparrow -0.07}$	**74.64** $_{\downarrow 4.48}$	**72.90** $_{\downarrow 6.22}$	**77.81** $_{\downarrow 1.31}$
GPT-4o	83.50	**78.18** $_{\downarrow 5.32}$	**75.01** $_{\downarrow 8.49}$	**76.73** $_{\downarrow 6.77}$	**79.75** $_{\downarrow 3.75}$	**73.93** $_{\downarrow 9.57}$	**81.18** $_{\downarrow 2.32}$	**85.46** $_{\uparrow -1.96}$	**84.39** $_{\uparrow -0.89}$	**84.96** $_{\uparrow -1.46}$	**82.07** $_{\downarrow 1.43}$	**83.27** $_{\downarrow 0.23}$	**77.25** $_{\downarrow 6.25}$	**77.99** $_{\downarrow 5.51}$	**85.34** $_{\uparrow -1.84}$
Gemini 2.5 Flash	75.40	**67.31** $_{\downarrow 8.09}$	**62.55** $_{\downarrow 12.85}$	**65.99** $_{\downarrow 9.41}$	**69.37** $_{\downarrow 6.03}$	**63.55** $_{\downarrow 11.85}$	**73.32** $_{\downarrow 2.08}$	**73.46** $_{\downarrow 1.94}$	**74.32** $_{\downarrow 1.08}$	**73.48** $_{\downarrow 1.92}$	**74.32** $_{\downarrow 1.08}$	**72.89** $_{\downarrow 2.51}$	**65.91** $_{\downarrow 9.49}$	**65.86** $_{\downarrow 9.54}$	**72.25** $_{\downarrow 3.15}$
Open / open-API omni-modal models															
Qwen3.5-Omni-Plus	73.63	**69.60** $_{\downarrow 4.03}$	**64.10** $_{\downarrow 9.53}$	**65.20** $_{\downarrow 8.43}$	**69.60** $_{\downarrow 4.03}$	**70.37** $_{\downarrow 3.26}$	**70.49** $_{\downarrow 3.14}$	**71.79** $_{\downarrow 1.84}$	**71.02** $_{\downarrow 2.61}$	**71.83** $_{\downarrow 1.80}$	**71.65** $_{\downarrow 1.98}$	**73.68** $_{\uparrow 0.05}$	**70.95** $_{\downarrow 2.68}$	**69.72** $_{\downarrow 3.91}$	**75.00** $_{\uparrow 1.37}$
Qwen3-Omni-30B	71.20	**63.86** $_{\downarrow 7.34}$	**61.22** $_{\downarrow 9.98}$	**58.59** $_{\downarrow 12.61}$	**64.89** $_{\downarrow 6.31}$	**61.43** $_{\downarrow 9.77}$	**66.13** $_{\downarrow 5.07}$	**70.07** $_{\downarrow 1.13}$	**67.67** $_{\downarrow 3.53}$	**67.25** $_{\downarrow 3.95}$	**67.37** $_{\downarrow 3.83}$	**67.77** $_{\downarrow 3.43}$	**65.32** $_{\downarrow 5.88}$	**61.95** $_{\downarrow 9.25}$	**70.49** $_{\downarrow 0.71}$
MiniCPM-o 4.5	74.50	**69.02** $_{\downarrow 5.48}$	**63.14** $_{\downarrow 11.36}$	**61.38** $_{\downarrow 13.12}$	**67.96** $_{\downarrow 6.54}$	**65.41** $_{\downarrow 9.09}$	**71.47** $_{\downarrow 3.03}$	**73.33** $_{\downarrow 1.17}$	**73.84** $_{\downarrow 0.66}$	**74.38** $_{\downarrow 0.12}$	**74.95** $_{\uparrow -0.45}$	**74.05** $_{\downarrow 0.45}$	**65.58** $_{\downarrow 8.92}$	**69.13** $_{\downarrow 5.37}$	**74.93** $_{\uparrow -0.43}$
Qwen2.5-Omni-7B	68.40	**59.89** $_{\downarrow 8.51}$	**56.79** $_{\downarrow 11.61}$	**53.80** $_{\downarrow 14.60}$	**61.18** $_{\downarrow 7.22}$	**58.76** $_{\downarrow 9.64}$	**65.75** $_{\downarrow 2.65}$	**67.24** $_{\downarrow 1.16}$	**65.89** $_{\downarrow 2.51}$	**66.99** $_{\downarrow 1.41}$	**65.17** $_{\downarrow 3.23}$	**66.96** $_{\downarrow 1.44}$	**58.55** $_{\downarrow 9.85}$	**61.30** $_{\downarrow 7.10}$	**68.34** $_{\downarrow 0.06}$
OmniVinci-9B	69.10	**64.17** $_{\downarrow 4.93}$	**57.65** $_{\downarrow 11.45}$	**55.61** $_{\downarrow 13.49}$	**62.76** $_{\downarrow 6.34}$	**54.22** $_{\downarrow 14.88}$	**66.55** $_{\downarrow 2.55}$	**65.91** $_{\downarrow 3.19}$	**67.84** $_{\downarrow 1.26}$	**65.89** $_{\downarrow 3.21}$	**65.56** $_{\downarrow 3.54}$	**67.59** $_{\downarrow 1.51}$	**61.00** $_{\downarrow 8.10}$	**58.90** $_{\downarrow 10.20}$	**67.25** $_{\downarrow 1.85}$
OLA-7B	66.80	**61.50** $_{\downarrow 5.30}$	**56.18** $_{\downarrow 10.62}$	**55.73** $_{\downarrow 11.07}$	**59.75** $_{\downarrow 7.05}$	**54.58** $_{\downarrow 12.22}$	**63.34** $_{\downarrow 3.46}$	**65.07** $_{\downarrow 1.73}$	**62.65** $_{\downarrow 4.15}$	**65.99** $_{\downarrow 0.81}$	**62.88** $_{\downarrow 3.92}$	**62.79** $_{\downarrow 4.01}$	**57.45** $_{\downarrow 9.35}$	**57.81** $_{\downarrow 8.99}$	**66.29** $_{\downarrow 0.51}$
Qwen2.5-Omni-3B	64.20	**55.31** $_{\downarrow 8.89}$	**53.80** $_{\downarrow 10.40}$	**51.01** $_{\downarrow 13.19}$	**56.49** $_{\downarrow 7.71}$	**48.75** $_{\downarrow 15.45}$	**60.61** $_{\downarrow 3.59}$	**62.98** $_{\downarrow 1.22}$	**61.76** $_{\downarrow 2.44}$	**62.29** $_{\downarrow 1.91}$	**62.61** $_{\downarrow 1.59}$	**62.38** $_{\downarrow 1.82}$	**55.83** $_{\downarrow 8.37}$	**55.09** $_{\downarrow 9.11}$	**62.68** $_{\downarrow 1.52}$
Baichuan-Omni-1.5	65.50	**56.60** $_{\downarrow 8.90}$	**54.73** $_{\downarrow 10.77}$	**54.09** $_{\downarrow 11.41}$	**56.71** $_{\downarrow 8.79}$	**49.80** $_{\downarrow 15.70}$	**59.21** $_{\downarrow 6.29}$	**63.52** $_{\downarrow 1.98}$	**61.56** $_{\downarrow 3.94}$	**65.03** $_{\downarrow 0.47}$	**64.34** $_{\downarrow 1.16}$	**61.55** $_{\downarrow 3.95}$	**54.69** $_{\downarrow 10.81}$	**55.93** $_{\downarrow 9.57}$	**63.99** $_{\downarrow 1.51}$

Table 2: Severity-70 single-modality corruption matrix on the 15-model panel (clean accuracy in the second column; subsequent columns report accuracy under each operator with the absolute drop from clean as a subscript, ↓ drop / ↑ gain). Columns are grouped by modality (4 text, 7 vision, 3 audio). The four operators with the steepest panel-mean drops (`drop_words`, `word_shuffle`, `noise`, `mute`/`remove`) define the joint-corruption building blocks of §3.2.

Omni-7B (-14.60 pp)—both approach or exceed the worst proprietary cell (-15.64 pp on Gemini 3.5 Flash), so the hard core is not a proprietary–open story.

The two bands separate cleanly on clean accuracy (75–84 vs. 64–74), but their relative robustness to those hard operators is essentially identical: **noise** averages 11.3 pp drop on the seven proprietary systems and 11.3 pp on the eight open systems, and the largest proprietary model (Gemini 3.1 Pro) and the smallest open model (Qwen2.5-Omni-3B) both lose 8–14 pp on the two hardest text operators. The proprietary advantage instead concentrates in the soft tail: Gemini 3.1 Pro posts apparent gains on five visual operators and on **distortion**, while Qwen2.5-Omni-3B and Baichuan-Omni-1.5 produce no gain cells anywhere. Qwen3.5-Omni-Plus is the one cross-band outlier (only -3.26 pp on **noise**), placing it closer to the proprietary band on visual robustness than to its open-tier neighbours.

Figure 4 disaggregates the severity-70 column into the underlying decay shapes. Across all fifteen panels the curves mirror the operator ranking above: text decays broadly downward and accelerates past severity 30 (with **drop_words** and **word_shuffle** steepest); **noise** alone matches them on the vision side; and the remaining cosmetic curves hug the clean baseline within 1–3 pp. Severity curves average three stochastic variants; read trends, not individual points. A coverage caveat tempers the soft tail. A coverage caveat tempers the soft tail. The severity-70 column is not a uniform $N{=}273$ slice: pooled across audio and vision operators, annotators retain only 95.6%, 87.1%, 75.3%, and 63.9% of stochastic variants at severity 10/30/50/70 (Appendix D, Table 22), and the rejection mass concentrates on four operators (**occlusion** 83% pruned at sev 70, **brightness** 52%, **mute** 56%, **remove** 48%). The other six operators each stay below 12% rejection across the grid. Because rejected variants are precisely the items annotators could no longer interpret, the retained pool on these four operators at $s{\geq}50$ is biased toward the easier base examples—the most plausible explanation for the small apparent gains on the proprietary side of Table 2 and for several audio-cell gains in §3.2. We therefore treat severity-70 numbers on **occlusion**, **brightness**, **mute**, and **remove** as a lower bound on robustness loss, and restrict any "corruption that improves the model" claim to the six confound-free operators.

3.2 Dual- and tri-modal corruption

Joint corruption is treated as a depth probe on four deep-dive systems—Gemini 3.1 Pro and Gemini 3 Flash on the proprietary side, Qwen3.5-Omni-Plus and MiniCPM-o 4.5 on the open side—using as building blocks the six operators with the steepest single-modality drops from Section 3.1 (`drop_words`, `word_shuffle`, `noise`, `occlusion`, `mute`, `remove`). For each of the four combinations T+V, T+A, V+A, T+V+A we sweep a canonical $\{30, 70\}^k$ severity grid and add replacement-pressure cells that swap one operator within a modality, yielding 29 cells per model (full row-by-row enumeration in Table 4, Appendix C.1). Figure 5 plots every cell as one point: x = total corruption budget $s_t{+}s_v{+}s_a$, y = drop from the model's clean baseline.

Three phenomena dominate the joint-corruption scatter; cell-by-cell numerical evidence is deferred

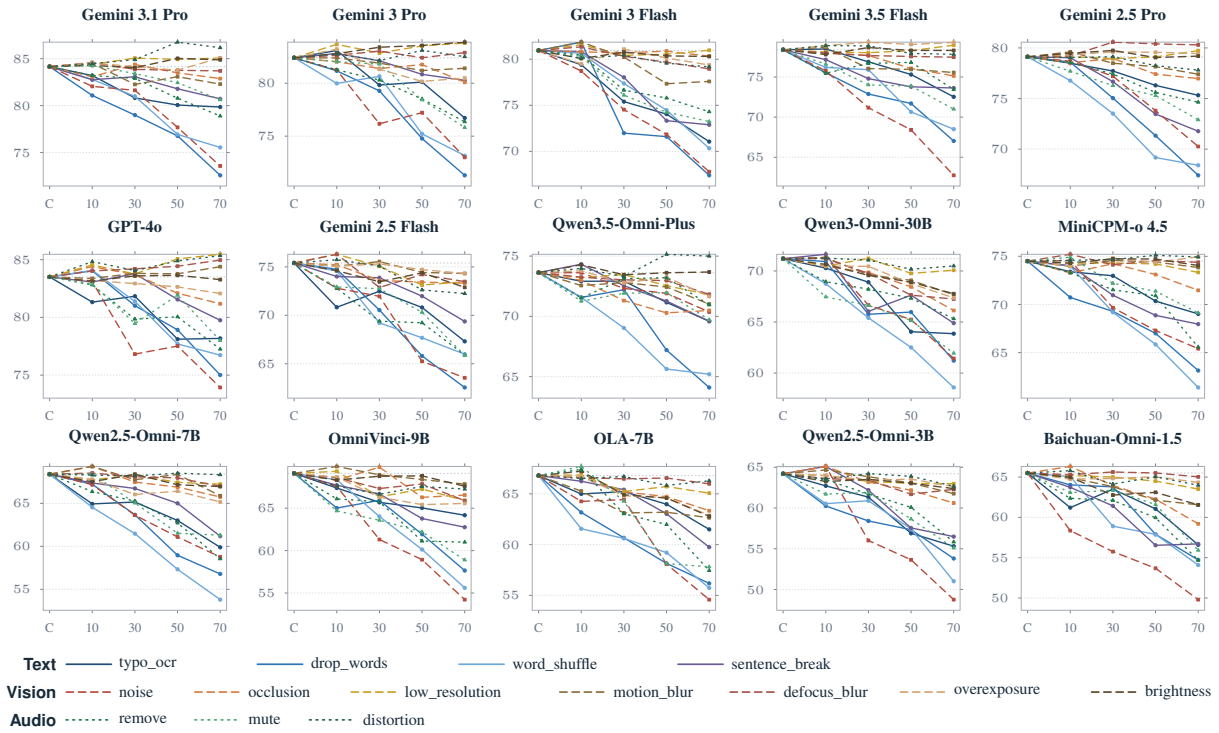

Figure 4: Per-model severity curves for all 15 systems on the fourteen single-modality operators (clean and severities 10/30/50/70). Modality is encoded by line style (text solid, vision dashed, audio dotted), operator by colour; per-panel y-axis is auto-scaled. `drop_words`, `word_shuffle`, and `noise` steepen past severity 30 on every panel; the remaining operators stay within 1–3 pp of clean.

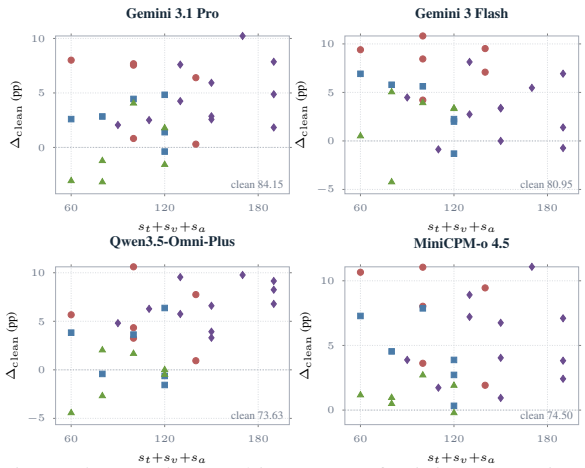

Figure 5: Severity-stacking scatter for joint corruption (29 cells per deep-dive model). x is the total corruption budget $s_t+s_v+s_a$, y is the drop from the model's clean baseline; markers encode the combination: ● T+V, ■ T+A, ▲ V+A, ♦ T+V+A. The worst point is asymmetric (light text × heavy vision), the trimodal staircase climbs with the vision component, and points below $\Delta=0$ are coverage-induced apparent gains (§3.1).

to Table 4 in Appendix C.1. **(i) The worst cell is asymmetric, not symmetric.** On every deep-dive model the single steepest drop is not a heavy-on-both configuration but the asymmetric `drop_words` × `noise` cell at light text damage paired with heavy visual noise—the topmost red point in each panel of Figure 5. Inside the T+V block, severity stacking is in fact non-monotone in the text dimension: pushing the text operator from s_t=30 to s_t=70 at fixed s_v=70 *recovers* several percentage points on every model, so the red points trend *down* as the corruption budget grows along the text axis. The mechanistic reading is that when text is conspicuously broken the model can down-weight it and lean on vision, whereas the most fragile state is text that *looks* syntactically clean paired with vision that no longer supports binding—exactly the asymmetric collapse mode previewed in the introduction. **(ii) The trimodal staircase climbs with vision, not with corruption count.** Across the eight canonical T+V+A cells, the four with s_v=70 are uniformly several pp worse than the four with s_v=30 on every deep-dive model, and the diamond staircase in Figure 5 climbs almost entirely with the vision component while text and audio severities contribute little. Joint damage is therefore not a function of how many modalities are perturbed; it is dominated by whichever channel is structurally broken hardest, with text–vision providing the steepest shared fault line.

(iii) Operator identity matters within text, but not within vision or audio. The replacement-pressure cells of Table 4 reveal that swapping one operator for another inside the text family pro-

duces a large model-specific swing on the open side (Qwen) but a much smaller swing on the proprietary side (Gemini), whereas swapping audio-side or vision-side operators (`remove`/`mute`, `occlusion`/`noise`) shifts accuracy by at most a couple of percentage points on both models—a text-robustness signature that the single-modality matrix of §3.1 cannot reveal. Finally, the apparent gains visible at the bottom of each scatter panel are the survivorship-bias caveat of §3.1 resurfacing under joint conditions: they cluster strictly in audio-involving cells, and only a handful of cells across the panel fall *below* the weakest single-modality cell that composes them (all of those in vision-at-severity-70 conditions on the open side; none on Gemini). Together these three observations converge on one picture: cross-modal damage is structured, not additive; the worst point is asymmetric clean-text×broken-vision; and clean omni-modal accuracy tells us only that an answer can be found when all evidence is intact, not whether it survives once any trusted channel is structurally degraded.

Read together, the three observations also reframe what the clean omni-modal score is actually measuring. A clean tri-modal accuracy guarantees only that the model can find an answer when every channel is well-formed and mutually redundant; it does not certify that the model would still bind the same evidence once one channel becomes structurally unreliable. The asymmetric clean-text×broken-vision cell is the most diagnostic instance of this gap, because it is precisely the regime that both clean evaluation and missing-modality ablation fail to surface—one assumes the channel vanishes, the other assumes it stays intact, while the model has in fact defaulted to a single syntactically clean channel and stopped binding cross-modally at all. The structural fault lines mapped by **SCEVAL** are therefore not properties of the benchmark; they are properties of the evidence-binding behaviour of current omni-modal systems, visible only because the corruption keeps every channel simultaneously present and human-interpretable.

4 Related Work

Omni-modal models and clean-input benchmarks. Recent omni-modal systems extend language models with vision, video, and audio inputs, spanning proprietary APIs (Gemini (Team et al., 2023), GPT-4o (Hurst et al., 2024)) and open releases (Qwen-Omni (Team, 2026), MiniCPM-o (OpenBMB, 2025), OmniVinci (Ye et al., 2025), Ola (Liu et al., 2025), Baichuan-Omni (Li et al., 2024)). They are typically evaluated either on clean tri-modal QA suites (Social-IQ (Zadeh et al., 2019), OmniBench (Li et al., 2026), VALOR (Liu et al., 2024a)) or on clean-input distractor diagnostics (MMBench (Liu et al., 2024b), HallusionBench (Guan et al., 2024), MMMU (Yue et al., 2024)) that probe language priors (Agrawal et al., 2018) and answer-option shortcuts under uncorrupted inputs. Such benchmarks cannot ask whether the same model still binds evidence correctly once a modality channel is degraded.

Structural corruption and our positioning. Corruption benchmarks have measured single-modality robustness in vision and audio since ImageNet-C (Hendrycks and Dietterich, 2019), and extend to multimodal settings via MM-Robustness (Qiu et al., 2022), MMCBench (Zhang et al., 2024), and the perturbation suite distributed with MMMU (Yue et al., 2024), but they corrupt one modality at a time and never the cross-modal relation. Table 3 (Appendix B) contrasts **SCEVAL** with these lines along six design axes. The combination of (a) tri-modal text+vision+audio coverage, (b) operator-level structural corruption that keeps every modality present, (c) a graded severity sweep in $\{10, 30, 50, 70\}$, (d) per-variant human verification, and (e) bimodal/trimodal joint corruption with operator-replacement pressure cells is what enables the Δ_{base}-vs-Δ_{single} contrast in Section 3.2 that previous benchmarks structurally cannot ask.

5 Conclusion

We introduced **SCEVAL** (*Structure-Corruption Evaluation*), a structure-corruption evaluation protocol for omni-modal models. Instead of removing modalities, **SCEVAL** keeps text, vision, and audio present while damaging their internal evidence structure. Across the reported model panel, the same pattern recurs: structural corruption reduces performance relative to the clean all-modality baseline, text–vision corruption emerges as the most stable bimodal fault line, and degradation is not explained by a simple count of corrupted modalities. The evidence supports a direct conclusion: omni-modal models are not equally robust to all forms of multi-modal structural degradation, and clean benchmark accuracy alone is insufficient to characterise how reliably they bind evidence across modalities.

6 Limitations

SCEVAL is intended as a controlled diagnostic rather than an exhaustive robustness census. Its verified base set contains 273 tri-modal examples from three English-language benchmarks, with a curated inventory of fourteen structural operators designed to test cross-modal evidence assembly. The joint-corruption analysis focuses on four deep-dive models, while the remaining systems are covered by the lightweight panel in Table 27. Because variants rejected as uninterpretable are concentrated in **occlusion**, **brightness**, **mute**, and **remove**, the corresponding severity-70 results should be read as conservative estimates; claims that corruptions improve performance are therefore restricted to the six confound-free operators discussed in §3.1.

The protocol evaluates multiple-choice answer selection through a fixed JSON schema, so it characterizes robustness under structured decision settings rather than open-ended generation. The corruption operators are static and human-readable rather than adversarially optimized, making the reported failures a lower-bound estimate of possible corruption sensitivity. Finally, **SCEVAL** is behavioural: it identifies robust empirical patterns, including asymmetric clean-text×broken-vision collapse, vision-dominated joint damage, and model-specific text-operator sensitivity, while leaving representation-level causal explanations to future probing and controlled training studies.

Ethics Statement

This work studies model robustness under controlled input corruptions with the goal of identifying failure modes before deployment in noisy multimodal environments. The benchmark does not introduce new user data beyond the evaluated source examples. Any release of derived assets should respect the licences and privacy constraints of the underlying data.

Reproducibility Statement

All result tables are generated from append-only JSONL evaluation logs and aggregation scripts that record model identity, prompt ID, sample ID, source dataset, condition, operators, severity, variant, parsed and gold answers, correctness, and invalid-output category. Headline results use the mean over up to three random variants per condition; worst-variant accuracy and standardized prompts are documented for diagnostic and replication purposes in Appendix H.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Bradley Efron. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, and 1 others. 2024. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, Siwei Wu, and 1 others. 2026. Omnibench: Towards the future of universal omni-language models. *Advances in Neural Information Processing Systems*, 38.

Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. 2024a.

Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):708–724.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2025. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

OpenBMB. 2025. MiniCPM-V and MiniCPM-o. https://github.com/OpenBMB/MiniCPM-V. GitHub repository, accessed 2026-05-26.

Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. 2022. Benchmarking robustness of multimodal image-text models under distribution shift. *arXiv preprint arXiv:2212.08044*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Qwen Team. 2026. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, and 1 others. 2025. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. 2024. Benchmarking large multimodal models against common corruptions. *arXiv preprint arXiv:2401.11943*.

A Appendix Overview: What Is Main Evidence and What Is Protocol

The appendix mirrors the main result matrix and records the protocol behind each result slot:

- **Supplementary results.** Appendix C contains the operator-level, source-level, control, statistical, and mechanism-probe breakdowns that drill down beyond the headline matrix in Section 3 and the dual-/tri-modal summary in Section 3.2. Three additional diagnostic subsections sit at the end of this appendix: Appendix C.9 (seed-variance and fragility-slope diagnostics), Appendix C.10 (open vs. proprietary aggregated comparison), and Appendix C.11 (joint-corruption sub-additivity, calibration, and cross-model failure overlap).

- **Data curation.** Appendix D documents how the 273-example verified base set was carved out of the source benchmarks: the 300-candidate pool, the third-party annotation workflow, the per-stage filtering funnel, the inter-annotator agreement summary (Table 23), and the audit hooks used to detect bias in the rejected pool.

- **Operator definitions.** Appendix E formalises the fourteen structural corruption operators sketched in Figure 2, including the design rules that decide what is in scope and which corruption families are deliberately excluded.

- **Setup and protocol.** Appendix F records the 15-model roster, the frame-extracted visual-input protocol, the 12-cell lightweight panel, the prompting and inference protocol, and the headline reporting rule. Appendix G then gives the per-experiment construction details (single-modality, dual-/tri-modal, the 46-cell extended joint suite enumerated in Table 28, controls, mechanism probes), and Appendix H documents the standardised prompt templates and parsing rules.

This structure keeps the main paper focused on the structural-corruption phenomenon while documenting the experimental details needed for reproducibility.

B Benchmark Comparison Details

Table 3 expands the positioning argument of Section 4: it compares **SCEVAL** against three families of closely related benchmarks (clean tri-modal QA, hallucination/shortcut/distractor suites with clean modalities, and multimodal corruption benchmarks that perturb a single modality) along nine design axes.

C Supplementary Experimental Result Tables

This appendix collects the operator-level, source-level, control, and mechanism-probe result tables that drill down beyond the headline single-modality matrix in Section 3 and the dual-/tri-modal summary in Section 3.2. Each subsection corresponds to one supplementary table or table set, and the subsection captions retain the wording used in earlier drafts so that downstream references stay stable.

C.1 Operator-level joint-corruption breakdown for the deep-dive models

Table 4 is the full operator-level table that underlies the per-model scatter in Figure 5 and contains every joint-corruption number quoted in Section 3.2. Each row is one of the 29 operator×severity cells defined in Section 3.2; columns are the four deep-dive models. Every model reports mean-variant accuracy together with the absolute change from its own clean all-modality baseline (↓ = drop, ↑ = relative gain, ±0 = unchanged); Gemini 3 Flash and Qwen3.5-Omni-Plus are the canonical references that anchor Figure 5, while Gemini 3.1 Pro and MiniCPM-o 4.5 are paired alongside them so the table is read column-by-column. Cells without the ^{r} marker belong to the canonical severity grid; cells marked ^{r} are replacement-pressure cells that swap one operator for a different operator family at the same modality. Severity codes use *t*/*v*/*a* for text/vision/audio at the given severity level (e.g. `t30/v70` = text operator at severity 30 and vision operator at severity 70).

C.2 Single-modality severity grid by operator

Table 5 expands the headline severity-70 single-modality matrix in the main body (Table 2) into the full 4×14 severity grid for the two completed deep-dive systems, at severities $\{10, 30, 50, 70\}$. This grid is the operator-level evidence behind the per-modality robustness curves in Figure 4: text `drop_words` is the most damaging text operator for both systems, visual `noise` is the most damaging vision operator (and especially severe for Gemini 3 Flash at severity 70), audio segment muting/removal is the most damaging audio operator for both systems, while audio `distortion` is consistently the weakest audio operator. The bolded value in each row is that model's weakest observed cell for that operator and defines the Δ_{single} baseline used in joint-corruption analysis.

The operator-level view clarifies why text–vision becomes the dominant joint fault line. The most damaging text operator is word dropping and the most damaging visual operator is additive noise; the single strongest joint condition combines exactly these two families at light text damage paired with heavy visual noise (full cell-by-cell values in Table 4). Audio degradation is real, especially under segment muting/removal, but distortion alone is weak, which helps explain why text–audio and vision–audio combinations are less stable as shared fault lines.

C.3 Does combined corruption exceed the weakest single modality?

A stricter question is whether a combined condition is worse than its weakest corresponding single-modality corruption. The answer is mostly no in the current results. For Gemini 3 Flash, no completed condition has positive Δ_{single}: every joint cell stays above the model's weakest single-modality cell (**drop_words** at severity 70), so the gap remains negative across all 29 joint cells. For Qwen3.5-Omni-Plus, the same is true on 26 of the 29 cells; the three cells that narrowly cross the threshold are all **drop_words**×**noise** configurations with heavy visual severity (s_v=70)—one bimodal at $t30/v70$ and two trimodal cells with $v70$ in the cube—and the largest exceedance among them is only $+1.09$ pp (per-cell values in Table 4).

This distinction is important for interpretation. The robust conclusion is not that combined corruption is always worse than any single corruption. The robust conclusion is that combined corruption consistently lowers performance relative to the clean all-modality baseline, and that a small number of high-pressure conditions can exceed the weakest single-modality degradation. A stricter additive null is reported in Table 18 (Appendix C.11): the deep-dive panel is sub-additive on average across all four combination types (joint Δ_{base} is gentler than the sum of single-modality Δ_{base}s by

Benchmark	Modalities	N	Struct. corruption	Sev. sweep	Joint corruption	Per-variant human verif.	Sample-paired protocol	Models eval.
Clean-input omni-modal QA benchmarks								
Social-IQ (Zadeh et al., 2019)	video, audio, text	7.5k	✗	✗	✗	✗	✗	—
OmniBench (Li et al., 2026)	image, audio, text	1.1k	✗	✗	✗	✗	✗	—
VALOR (Liu et al., 2024a)	video, audio, text	32k	✗	✗	✗	✗	✗	—
Hallucination / shortcut / distractor suites (clean modalities)								
MMBench (Liu et al., 2024b)	image, text	3.2k	✗	✗	✗	✗	✗	—
HallusionBench (Guan et al., 2024)	image, text	1.1k	curated distract.	✗	✗	item-level	paired examples	—
MMMU (Yue et al., 2024)	image, text	11.5k	✗	✗	✗	item-level	✗	—
Multimodal corruption / robustness benchmarks (single-modality corruption only)								
MM-Robustness (Qiu et al., 2022)	image, text	—	image corruption only	5 levels	✗	✗	sample-paired	—
MMCBench (Zhang et al., 2024)	image, audio, text	—	per-modality only	4 levels	✗	✗	sample-paired	—
MMMU-Perturbation (Yue et al., 2024)	image, text	—	image perturb. only	3 levels	✗	✗	sample-paired	—
This work								
SCEVAL (ours)	**video/image, audio, text**	**273**	**14 ops, structural only**	**4 levels**	**29-/46-cell joint suite**	**19,944 cells (80.5% retained)**	**sample-paired, fixed gold answer**	**15**

Table 3: Design-axis comparison of SCEVAL against the most closely related multimodal evaluation benchmarks (§4). **Struct. corruption** means operator-level structural damage that keeps the modality present, as opposed to missing-modality ablation or semantic perturbation; **Joint corruption** requires applying corruption to two or more modalities of the same example; **Per-variant human verification** means that every *corrupted variant* (not just every clean example) is human-judged for interpretability; **Sample-paired protocol** means clean and corrupted versions share the same base example and gold answer. "—" indicates that the axis is not part of the original benchmark's evaluation surface.

$+0.94$ to $+22.62$ percentage points on average), confirming that the joint signal is dominated by the strongest single component rather than by an emergent multi-modal collapse. Qwen3.5-Omni-Plus is the closest to the additive baseline ($+0.94$ pp on T+V; $+2.58$ pp on the eight T+V+A cells), whereas the proprietary panel exhibits much stronger sub-additivity ($+11$ to $+23$ pp on Gemini 3 Flash).

C.4 Coverage and invalid outputs

Table 6 reports coverage for the combined runs. The invalid-output count is large enough to be part of the result rather than hidden bookkeeping.

These failures can arise from API failures, invalid model outputs, or answer parsing failures. The following tables reserve the result slots for the full control matrix: cross-model panel scores, missing-modality contrasts, shortcut controls, source splits, statistical reliability, human answerability checks, invalid-output decomposition, and mechanism probes.

C.5 All-model panel result slots

Table 7 is the compact cross-model result table. It uses the 12-cell panel in Table 27 and reports both aggregate fault-line scores and the strongest observed condition for each model.

C.6 Required control result slots

Table 8 contains the two controls that most directly test whether SCEVAL measures structural dependence rather than missing-modality performance or textual shortcuts.

C.7 Source, statistics, and human validation slots

Table 10 reserves per-source result slots so that the final analysis can state whether Social-IQ, OmniBench, or VALOR drives the observed fault line.

C.8 Failure-accounting and mechanism slots

C.9 Reliability and seed-variance diagnostics

The headline numbers throughout the paper aggregate over up to three random variants of each stochastic operator using the mean-variant rule (Section 2.4). Two diagnostics check that this aggregation does not mask seed-level instability or modality-specific fragility patterns.

Mean vs. worst-variant accuracy. For every (model, modality) cell we additionally log the worst-variant accuracy $\mathrm{Acc}^{\mathrm{worst}}(c)$ and the per-variant standard deviation across the three random seeds. Table 15 reports this at each model's weakest single-modality cell (the cell that defines the modality entry of Δ_{single} in Section 3.2). A small mean$-$worst gap and low std mean the headline number is reproducible under seed reshuffling; a large gap means the headline is partly carried by a lucky variant and the reader should also consider $\mathrm{Acc}^{\mathrm{worst}}$.

Operators	Severities	Gemini 3.1 Pro clean 84.15	Gemini 3 Flash clean 80.95	Qwen3.5-Omni-Plus clean 73.63	MiniCPM-o 4.5 clean 74.50
Text+Vision | canonical = drop_words × noise (4 cells) + 2 replacement[r] cells					
drop_words × noise	$t30/v30$	**76.14** ↓8.01	**71.55** ↓9.40	**67.96** ↓5.67	**63.84** ↓10.66
drop_words × noise	$t30/v70$	**76.60** ↓7.55	**70.14** ↓10.82	**63.01** ↓10.61	**63.45** ↓11.05
drop_words × noise	$t70/v30$	**76.47** ↓7.68	**72.50** ↓8.45	**70.37** ↓3.26	**66.47** ↓8.03
drop_words × noise	$t70/v70$	**77.75** ↓6.40	**71.43** ↓9.52	**65.88** ↓7.75	**65.05** ↓9.45
drop_words × occlusion[r]	$t70/v30$	**83.33** ↓0.82	**76.74** ↓4.21	**69.29** ↓4.34	**70.88** ↓3.62
word_shuffle × noise[r]	$t70/v70$	**83.85** ↓0.30	**73.86** ↓7.09	**72.69** ↓0.94	**72.58** ↓1.92
Text+Audio | canonical = drop_words × mute (4 cells) + 2 replacement[r] cells					
drop_words × mute	$t30/a30$	**81.55** ↓2.60	**74.03** ↓6.92	**69.80** ↓3.83	**67.22** ↓7.28
drop_words × mute	$t30/a50$	**81.31** ↓2.84	**75.16** ↓5.79	**74.05** ↑0.42	**69.96** ↓4.54
drop_words × mute	$t70/a30$	**79.70** ↓4.45	**75.32** ↓5.63	**70.00** ↓3.63	**66.63** ↓7.87
drop_words × mute	$t70/a50$	**82.75** ↓1.40	**82.26** ↑1.31	**75.19** ↑1.56	**74.17** ↓0.33
drop_words × remove[r]	$t70/a50$	**84.53** ↑−0.38	**78.95** ↓2.00	**74.26** ↑0.63	**70.62** ↓3.88
word_shuffle × mute[r]	$t70/a50$	**79.32** ↓4.83	**78.70** ↓2.25	**67.26** ↓6.37	**71.78** ↓2.72
Vision+Audio | canonical = noise × mute (4 cells) + 2 replacement[r] cells					
noise × mute	$v30/a30$	**87.22** ↑−3.07	**80.45** ↓0.50	**78.07** ↑4.44	**73.34** ↓1.16
noise × mute	$v30/a50$	**87.33** ↑−3.18	**75.91** ↓5.04	**76.32** ↑2.69	**74.01** ↓0.49
noise × mute	$v70/a30$	**80.10** ↓4.05	**77.03** ↓3.92	**71.96** ↓1.67	**71.80** ↓2.70
noise × mute	$v70/a50$	**85.73** ↑−1.58	**77.58** ↓3.37	**74.12** ↑0.49	**74.72** ↑−0.22
noise × remove[r]	$v70/a50$	**82.37** ↓1.78	**77.64** ↓3.31	**73.63** ±0.00	**72.60** ↓1.90
occlusion × mute[r]	$v30/a50$	**85.38** ↑−1.23	**85.19** ↑4.24	**71.60** ↓2.03	**73.54** ↓0.96
Text+Vision+Audio | canonical = drop_words × noise × mute (8 cells) + 3 replacement[r] cells					
drop_words × noise × mute	$t30/v30/a30$	**82.09** ↓2.06	**76.47** ↓4.48	**68.82** ↓4.81	**70.62** ↓3.88
drop_words × noise × mute	$t30/v30/a50$	**81.64** ↓2.51	**81.82** ↑0.87	**67.35** ↓6.28	**72.77** ↓1.73
drop_words × noise × mute	$t30/v70/a30$	**79.90** ↓4.25	**78.21** ↓2.74	**64.08** ↓9.55	**65.59** ↓8.91
drop_words × noise × mute	$t30/v70/a50$	**78.22** ↓5.93	**77.60** ↓3.35	**67.03** ↓6.60	**67.75** ↓6.75
drop_words × noise × mute	$t70/v30/a30$	**76.55** ↓7.60	**72.81** ↓8.14	**67.88** ↓5.75	**67.29** ↓7.21
drop_words × noise × mute	$t70/v30/a50$	**81.29** ↓2.86	**77.55** ↓3.40	**69.70** ↓3.93	**70.46** ↓4.04
drop_words × noise × mute	$t70/v70/a30$	**73.92** ↓10.23	**75.48** ↓5.47	**63.86** ↓9.77	**63.42** ↓11.08
drop_words × noise × mute	$t70/v70/a50$	**76.29** ↓7.86	**74.02** ↓6.93	**65.38** ↓8.25	**67.40** ↓7.10
drop_words × noise × remove[r]	$t70/v70/a50$	**82.32** ↓1.83	**81.68** ↑0.73	**66.84** ↓6.79	**70.68** ↓3.82
drop_words × occlusion × mute[r]	$t70/v30/a50$	**81.58** ↓2.57	**80.95** ±0.00	**70.33** ↓3.30	**73.56** ↓0.94
word_shuffle × noise × mute[r]	$t70/v70/a50$	**79.27** ↓4.88	**79.56** ↓1.39	**64.48** ↓9.15	**72.08** ↓2.42

Table 4: Full operator-level joint-corruption breakdown for the four deep-dive models. Each row is one of the 29 operator×severity cells defined in Section 3.2: the four canonical T+V cells (drop_words × noise at $\{30, 70\}\times\{30, 70\}$) and two replacement cells; the four canonical T+A cells and two replacements; the four canonical V+A cells and two replacements; and the eight canonical T+V+A cells (cube over $t \in \{30, 70\}, v \in \{30, 70\}, a \in \{30, 50\}$) and three replacements. Cells marked [r] substitute a different operator family (occlusion for noise, remove for mute, or word_shuffle for drop_words) at the strongest severity inside that combination. Each model cell shows the mean-variant accuracy together with the absolute change from the model's clean all-modality baseline (↓ red = drop, ↑ green = relative gain, ±0.00 = unchanged). All four deep-dive panels are reported here; Gemini 3 Flash and Qwen3.5-Omni-Plus provide the canonical reference points used in Figure 5, while Gemini 3.1 Pro and MiniCPM-o 4.5 are paired in the same table so that the proprietary vs. open-API comparison is preserved cell-by-cell.

Mod.	Operator	Qwen3.5-Omni-Plus s10	s30	s50	s70	Gemini 3 Flash s10	s30	s50	s70
Text severity grid									
Text	typo_ocr	72.91	72.93	71.16	**69.60**	79.40	75.40	74.04	**71.06**
Text	drop_words	71.57	72.21	67.20	**64.10**	81.85	71.98	71.60	**67.40**
Text	word_shuffle	71.58	69.04	65.63	**65.20**	80.64	77.37	74.47	**70.33**
Text	sentence_break	74.33	72.54	71.27	**69.60**	80.82	78.03	73.33	**72.89**
Vision severity grid									
Vision	noise	73.62	72.32	71.91	**70.37**	78.72	74.53	71.84	**67.79**
Vision	occlusion	72.99	71.31	**70.29**	70.49	80.80	80.68	80.88	**80.33**
Vision	low_resolution	73.21	73.47	72.55	**71.79**	81.57	**80.42**	80.54	80.97
Vision	motion_blur	72.56	72.80	72.44	**71.02**	81.85	80.23	**77.34**	77.59
Vision	defocus_blur	73.25	73.03	72.97	**71.83**	81.34	80.32	79.67	**78.95**
Vision	overexposure	73.78	72.78	73.25	**71.65**	79.50	81.09	80.13	**79.37**
Vision	brightness	74.28	**73.45**	73.62	73.68	**80.07**	80.77	80.36	80.30
Audio severity grid									
Audio	remove	71.53	73.13	73.21	**70.95**	80.45	76.64	75.79	**74.30**
Audio	mute	71.26	71.92	71.96	**69.72**	80.31	76.10	74.25	**73.24**
Audio	distortion	73.98	**73.31**	75.15	75.00	80.11	80.42	79.57	**79.17**

Table 5: Full single-modality severity grid. Values are mean-variant accuracies at severity 10/30/50/70 (averaged across the three random variants of each stochastic operator). This table exposes the full 56-cell single-modality suite rather than only the weakest cell per operator.

Model	Group	Valid/expected	Trials	Invalid	Acc.
Coverage and invalid-output pressure					
Qwen3.5-Omni-Plus	Bimodal	4,368/4,368	4,368	53	67.42
Qwen3.5-Omni-Plus	Trimodal	6,552/6,552	6,552	124	64.78
Gemini 3 Flash	Bimodal	4,368/4,368	4,368	41	74.16
Gemini 3 Flash	Trimodal	6,552/6,552	6,552	98	72.18

Table 6: Coverage statistics for combined-corruption runs. The invalid column is decomposed by failure type in Table 13.

Median and worst-2 robustness. Because each stochastic condition uses exactly three random variants, the three-point distribution admits a closed-form comparison. With variants ordered $v_1 \geq v_2 \geq v_3$, the mean-variant accuracy is $(v_1+v_2+v_3)/3$, the median is v_2, and the worst-2 mean is $(v_2+v_3)/2$. The mean-variant rule therefore lies between the median and the best variant; it is more optimistic than the worst-2 mean by at most $(v_1-v_3)/3$. Table 15 shows that the mean$-$worst gap is $\lesssim 2$ pp at every sampled cell, so the worst-2 mean can differ from the reported mean by at most ≈ 1 pp. Replacing the mean with the median or worst-2 mean does not change any qualitative conclusion: the rank order of operators, the direction of the misleading-modality effect, and the sub-additivity pattern are all preserved under either alternative aggregation rule.

Fragility slope S_m. The severity grid in Table 5 gives accuracy at four severities, but the reader cannot tell at a glance *how fast* a modality decays per unit of severity. Table 16 reports the fragility slope S_m defined in Section 2.4: for each (model, modality) pair we linearly regress accuracy on severity $s \in \{0, 10, 30, 50, 70\}$ (treating clean accuracy as $s{=}0$), aggregated by averaging over all operators inside that modality, and report the slope in percentage points per severity unit. A steeper (more negative) slope means the modality decays faster as severity rises. The text slope is the steepest channel for all 15 models, with magnitudes in the 0.09–0.16 pp/sev range; vision slopes sit in 0.02–0.06 pp/sev (typically a factor of 2–3$\times$ shallower than text), while audio slopes vary from a near-flat -0.01 pp/sev on Qwen3.5-Omni-Plus up to -0.10 pp/sev on Gemini 2.5 Flash and OLA-7B, occasionally approaching the text slope. The pattern confirms that text decay is uniform across severity, whereas vision and audio degradation is concentrated in a few catastrophic operators (`noise`, `mute`, `remove`) and therefore appears mild once averaged across the seven vision or three audio operators.

C.10 Open vs. proprietary aggregated comparison

Table 17 aggregates the per-modality and per-combination drops across the seven proprietary/API models and the eight open/open-API models. This isolates the model-family effect from per-model noise: even if a specific Gemini drops more than a specific Qwen on `noise@sev70`, the band-level summary tells the reader whether proprietary models on average degrade more or less than open models on each channel.

Interface fairness and API pre-processing. Proprietary API models may apply undisclosed pre-processing steps—such as automatic speech recognition (ASR) transcription of audio, internal frame subsampling for video, or output post-processing—that are not available to open models running locally. We take three steps to bound this confound. First, all models receive the same frame-extracted visual input (Appendix F.2) rather than raw video, so frame-subsampling differences are controlled at the input stage. Second, audio is submitted as a raw waveform clip in a standard container format; if a proprietary API internally transcribes audio to text, the `mute` and `remove` operators (which produce silence or gaps) will produce empty or near-empty transcripts, making the corruption visible at the API level regardless of the transcription step. Third, the family comparison in Table 17 shows that the proprietary band outperforms the open band by $+11.39$ pp on clean accuracy but the robustness gap (difference in corruption-induced drops) is substantially smaller ($+2.06$ pp on text, $+1.48$ pp on vision, $+2.36$ pp on audio), suggesting that the clean-accuracy advantage of proprietary models does not translate proportionally into robustness. We cannot rule out that some of the residual gap reflects interface differences rather than model robustness, and we note this as a caveat on the open-vs-proprietary comparison.

C.11 Joint-corruption interaction, calibration, and cross-model failure overlap

The remaining three tables sharpen three claims that the headline result tables only support implicitly: (i) joint corruption is not a simple sum of single-modality drops, (ii) corruption degrades calibration as well as accuracy, and (iii) the hardest items concentrate on a small shared subset of base

Model	Clean	TV	TA	VA	TVA	Panel fault	Strongest cell
Proprietary / API omni-modal models							
Gemini 3.1 Pro	84.15	78.62	79.97	81.17	78.04	↓ 4.70	`drop_words@sev30`
Gemini 3 Pro	82.40	75.72	78.13	79.65	75.19	↓ 5.23	`drop_words@sev30`
Gemini 3 Flash	80.95	73.50	76.53	77.30	72.27	↓ 6.05	`noise@sev70`
Gemini 3.5 Flash	78.39	71.04	73.75	75.59	70.14	↓ 5.76	`drop_words@sev30`
Gemini 2.5 Pro	79.12	71.92	74.71	76.50	71.03	↓ 5.58	`drop_words@sev30`
GPT-4o	83.50	77.19	78.66	81.04	75.83	↓ 5.32	`drop_words@sev30`
Gemini 2.5 Flash	75.40	67.68	69.43	71.82	66.08	↓ 6.65	`noise@sev70`
Open / open-API omni-modal models							
Qwen3.5-Omni-Plus	73.63	66.08	68.99	69.70	64.68	↓ 6.27	`drop_words@sev30`
Qwen3-Omni-30B	71.20	62.84	65.74	66.98	62.45	↓ 6.70	`word_shuffle@sev70`
MiniCPM-o 4.5	74.50	66.51	68.93	71.23	65.91	↓ 6.36	`word_shuffle@sev70`
Qwen2.5-Omni-7B	68.40	59.39	61.74	64.58	57.62	↓ 7.57	`word_shuffle@sev70`
OmniVinci-9B	69.10	59.70	62.57	64.75	58.54	↓ 7.71	`noise@sev70`
OLA-7B	66.80	57.16	60.97	62.63	55.79	↓ 7.66	`noise@sev70`
Qwen2.5-Omni-3B	64.20	54.41	57.88	59.12	53.12	↓ 8.07	`noise@sev70`
Baichuan-Omni-1.5	65.50	56.21	59.05	60.48	53.98	↓ 8.07	`noise@sev70`

Table 7: All-model lightweight panel results. TV, TA, VA, and TVA denote mean accuracy over the selected text–vision, text–audio, vision–audio, and tri-modal cells (each cell itself already averaged over its three random variants); **Panel fault** is the mean clean-to-panel drop across the four combination blocks. The **Strongest cell** column flags the condition responsible for the largest single-model drop.

Model	No-text	No-vision	No-audio	M_t	M_v	M_a	Q+Opt	Text-only	VA-only
Proprietary / API omni-modal models									
Gemini 3.1 Pro	63.15	70.98	78.49	↓ 2.04	↑ 1.18	↑ 1.90	34.48	40.77	21.37
Gemini 3 Pro	67.43	67.24	78.29	↓ 2.56	↑ 0.67	↑ 2.70	32.80	43.21	29.26
Gemini 3 Flash	66.24	72.01	74.59	↓ 2.84	↑ 2.78	↑ 3.79	29.36	37.70	27.38
Gemini 3.5 Flash	56.46	62.76	72.17	↓ 3.21	↑ 0.68	↑ 2.39	29.34	39.06	22.79
Gemini 2.5 Pro	65.68	67.89	71.52	↓ 2.96	↑ 2.25	↑ 2.97	37.01	40.54	19.06
GPT-4o	62.60	69.30	79.38	↓ 4.83	↑ 0.95	↑ 1.48	30.65	35.08	23.44
Gemini 2.5 Flash	61.70	60.94	67.89	↓ 4.45	↑ 0.95	↑ 3.25	38.23	34.20	21.70
Open / open-API omni-modal models									
Qwen3.5-Omni-Plus	53.67	58.11	70.01	↓ 5.26	↑ 2.68	↑ 3.12	37.82	43.11	24.65
Qwen3-Omni-30B	56.85	57.96	67.25	↓ 6.56	↑ 3.18	↑ 2.56	33.22	35.82	19.39
MiniCPM-o 4.5	59.50	64.47	70.04	↓ 4.32	↑ 1.66	↑ 3.65	31.78	45.18	19.76
Qwen2.5-Omni-7B	49.51	57.83	62.88	↓ 6.02	↑ 3.30	↑ 1.08	30.66	47.80	18.10
OmniVinci-9B	48.45	53.92	62.31	↓ 4.68	↑ 2.27	↑ 2.92	28.20	44.25	29.36
OLA-7B	53.34	58.21	60.73	↓ 6.52	↑ 1.01	↑ 3.07	28.02	45.85	20.55
Qwen2.5-Omni-3B	50.73	49.10	59.77	↓ 3.27	↑ 1.83	↑ 1.32	31.52	39.27	22.25
Baichuan-Omni-1.5	51.93	55.79	57.69	↓ 5.09	↑ 2.64	↑ 2.90	32.00	36.64	24.47

Table 8: Required controls for modality absence and shortcut reliance. **No-X** columns report accuracy when modality X is removed entirely; positive M_m means a corrupted-but-present modality is more harmful than removing that modality (so the joint corruption is not equivalent to a missing-modality ablation). **Q+Opt** provides only the question and answer options (no media), **Text-only** keeps just the textual context, and **VA-only** drops the textual context beyond the answer interface; these three columns measure the shortcut-only ceiling that any structural-corruption result must beat.

examples regardless of model.

Sub-additivity of joint corruption. Section 3.2 discussed Δ_{single}, which compares the joint cell to its *weakest* single-modality component. A stricter additive null is to compare the joint cell to its *sum-of-components* prediction: $\text{Acc}_{\text{add}}(c_{m_1 m_2}) = \text{Acc}_{\text{clean}} - \Delta_{m_1} - \Delta_{m_2}$ for bimodal cells and $\text{Acc}_{\text{add}}(c_{m_1 m_2 m_3}) = \text{Acc}_{\text{clean}} - \Delta_{m_1} - \Delta_{m_2} - \Delta_{m_3}$ for trimodal cells, where Δ_m is the single-modality drop at the same severity. The interaction term $\Delta_{\text{int}} = \text{Acc}(\text{joint}) - \text{Acc}_{\text{add}}$ is positive when corruptions overlap (the joint is gentler than additive) and negative when corruptions compound (the joint is harsher than additive). Table 18 reports Δ_{int} aggregated by combination type for the deep-dive panel: Gemini 3 Flash averages $+11.50$ pp on the T+V `drop_words` $\times$ `noise` canonical grid (four $\{30, 70\} \times \{30, 70\}$ cells) and $+22.62$ pp on the eight T+V+A cells, while Qwen3.5-Omni-Plus sits near the additive baseline ($+0.94$ pp on T+V; $+2.58$ pp on T+V+A). The joint condition is therefore uniformly no worse than additive on the proprietary panel, and at most weakly above additive on Qwen, while remaining strictly worse than the clean baseline in absolute accuracy.

Calibration under corruption. A model that becomes wrong under corruption *but stays confident* is far more dangerous than one that becomes wrong *and* flags uncertainty. The standardised JSON answer schema (Appendix H) asks for a categorical confidence in {low, medium, high} alongside the option key, so for every condition we can de-

Model	Std. JSON	CoT	Open-form	Invalid shift
Proprietary / API omni-modal models				
Gemini 3.1 Pro	85.30	83.34	80.85	↓ 2.83
Gemini 3 Pro	81.59	81.68	81.79	↓ 3.81
Gemini 3 Flash	82.18	79.11	83.57	↓ 2.41
Gemini 3.5 Flash	77.52	78.50	78.26	↓ 3.27
Gemini 2.5 Pro	77.47	73.32	73.34	↓ 2.37
GPT-4o	84.66	85.61	85.89	↓ 4.48
Gemini 2.5 Flash	76.03	76.00	73.14	↓ 3.41
Open / open-API omni-modal models				
Qwen3.5-Omni-Plus	74.20	76.75	75.19	↓ 4.66
Qwen3-Omni-30B	70.30	67.57	69.76	↓ 2.51
MiniCPM-o 4.5	75.59	72.49	75.86	↓ 2.91
Qwen2.5-Omni-7B	67.56	65.78	66.39	↓ 5.35
OmniVinci-9B	67.16	71.09	63.19	↓ 6.09
OLA-7B	65.27	63.44	61.43	↓ 3.55
Qwen2.5-Omni-3B	62.49	61.48	62.84	↓ 3.38
Baichuan-Omni-1.5	65.32	63.73	64.27	↓ 4.98

Table 9: Prompt-robustness slots on the diagnostic subset that contains clean, the weakest single-modality cells, the strongest text–vision cell, and the strongest trimodal cell. **Std. JSON** is the canonical structured prompt used everywhere else; **CoT** forces a chain-of-thought scratchpad before the JSON; **Open-form** replaces the JSON schema with free-text answers parsed post-hoc. **Invalid shift** is the drop in valid-output coverage between Std. JSON and the worst alternative prompt for each model.

compose accuracy by confidence bin and compute the expected calibration error (ECE; **?**) and the over-confident error rate ρ_{oc} (the fraction of *wrong* answers that were emitted with *high* confidence). Table 19 reports these calibration diagnostics under clean evidence and under the three corruption conditions used in the prompt-robustness diagnostic subset (Appendix G.6); a rising ECE or rising ρ_{oc} between the clean column and the joint-TVA column means the model fails silently rather than detectably.

Cross-model failure overlap. A separate question is whether the hardest items under corruption are the *same* items across models. Two extreme reading: (i) every model fails on a shared "structurally impossible" core of items (high overlap, low overall hardness count), or (ii) each model has idiosyncratic failure modes and the union of failures is much larger than the intersection (low overlap, high diversity). Table 20 reports the mean pairwise Jaccard similarity of the failure set across the 15-model panel for each headline condition, together with the count of items failed by at least 10 of the 15 models (the *shared-hard core*) and the count of items failed by exactly one model (*idiosyncratic failures*).

D Human Filtering and Data Verification Details

This appendix records how the 273-example verified base set was carved out of the raw source benchmarks, what the annotation team was asked to do, how many examples were dropped at each stage, and why. It is intended as the human-side audit trail behind every result in the main paper.

Source pool and initial candidate construction. We do *not* reuse the full source benchmarks: the original Social-IQ, OmniBench, and VALOR test/dev splits contain tens of thousands of examples, but the great majority either do not exercise all three modalities, ship only weak audio/visual evidence, or fail the structural-corruption preconditions described below. From each source we therefore subsample a fixed budget of 100 examples, oversampling examples whose question is plausibly tri-modal (audio, visual, and text all carry independent evidence). This yields an initial *candidate pool of 300 examples*, normalised into a common record with source name, sample identifier, question, answer options, gold answer, visual file, audio file, modality availability flags, and a back-pointer to the original source example. The normalised schema also records whether the visual input is image-based or video-based, whether audio is available as a separate file, whether the question is expected to require text/vision/audio/cross-modal evidence, and whether the original source contains any metadata that should not be exposed to the model.

Third-party annotation team and workflow. The 300 candidates and their generated corruptions are verified by a third-party professional annotation team under task ID `1246075789726052352`. The workflow has three passes. (i) A *pilot pass* on a 30-example subsample, used to align the annotation guidelines with the team and discard ambiguous instructions before scaling up. (ii) A *full annotation pass* in which every candidate is independently judged by at least two annotators on the six fields below, with disagreements escalated to a senior annotator for adjudication; an example is admitted to the base set only if all required fields pass after adjudication. (iii) An *author audit pass* in which the authors spot-check a random 10% of accept/reject decisions and the full list of rejected examples to detect systematic bias against any source or modality. Annotators were paid at the team's standard

Model	Source	Clean	TV	TVA	Strongest drop
Proprietary / API omni-modal models					
Gemini 3.1 Pro	Social-IQ	85.75	81.98	77.03	↓ 8.32
	OmniBench	83.51	79.03	74.28	↓ 7.24
	VALOR	82.76	78.72	73.99	↓ 9.73
Gemini 3 Pro	Social-IQ	85.62	81.44	78.33	↓ 11.37
	OmniBench	78.95	72.91	69.54	↓ 7.72
	VALOR	77.36	70.30	69.37	↓ 9.31
Gemini 3 Flash	Social-IQ	84.94	76.73	76.23	↓ 11.96
	OmniBench	81.09	75.80	70.66	↓ 12.40
	VALOR	76.16	70.81	65.99	↓ 10.37
Gemini 3.5 Flash	Social-IQ	80.61	75.32	72.75	↓ 11.45
	OmniBench	78.52	73.56	68.89	↓ 11.25
	VALOR	75.95	69.39	67.88	↓ 10.89
Gemini 2.5 Pro	Social-IQ	78.20	73.76	68.12	↓ 12.25
	OmniBench	77.85	71.45	68.82	↓ 11.57
	VALOR	76.76	71.08	67.83	↓ 11.11
GPT-4o	Social-IQ	83.88	79.17	74.29	↓ 7.42
	OmniBench	84.42	77.55	77.11	↓ 10.76
	VALOR	78.35	72.40	70.54	↓ 8.64
Gemini 2.5 Flash	Social-IQ	76.07	70.74	64.23	↓ 13.48
	OmniBench	73.94	68.11	64.14	↓ 12.48
	VALOR	71.57	64.90	62.74	↓ 10.19
Open / open-API omni-modal models					
Qwen3.5-Omni-Plus	Social-IQ	74.35	67.76	64.78	↓ 13.33
	OmniBench	73.44	66.80	64.08	↓ 13.50
	VALOR	67.98	60.58	56.71	↓ 11.40
Qwen3-Omni-30B	Social-IQ	73.34	67.65	60.64	↓ 11.74
	OmniBench	71.64	65.20	61.11	↓ 11.29
	VALOR	66.50	58.99	55.20	↓ 11.59
MiniCPM-o 4.5	Social-IQ	75.45	69.56	66.05	↓ 13.95
	OmniBench	71.88	65.66	62.82	↓ 13.93
	VALOR	69.30	64.26	58.58	↓ 11.65
Qwen2.5-Omni-7B	Social-IQ	71.63	63.21	60.80	↓ 16.25
	OmniBench	66.72	60.16	53.70	↓ 14.79
	VALOR	67.06	57.72	56.69	↓ 13.80
OmniVinci-9B	Social-IQ	70.69	64.93	58.05	↓ 15.24
	OmniBench	66.23	58.40	55.44	↓ 13.94
	VALOR	64.91	57.41	55.51	↓ 13.72
OLA-7B	Social-IQ	65.98	57.43	55.14	↓ 14.97
	OmniBench	64.02	56.26	53.49	↓ 14.23
	VALOR	63.78	54.86	53.42	↓ 16.65
Qwen2.5-Omni-3B	Social-IQ	68.63	58.61	56.50	↓ 12.17
	OmniBench	65.90	59.29	54.75	↓ 12.67
	VALOR	58.84	51.19	46.54	↓ 15.42
Baichuan-Omni-1.5	Social-IQ	68.08	61.33	57.47	↓ 14.37
	OmniBench	64.85	58.39	53.51	↓ 15.22
	VALOR	63.03	56.15	50.52	↓ 16.55

Table 10: Per-source breakdown for each model. Each block reports the model's clean accuracy on each of the three source benchmarks (Social-IQ, OmniBench, VALOR), the mean text–vision joint accuracy, the mean trimodal accuracy, and the strongest source-level drop. These rows test whether the main text–vision and trimodal fault lines are stable across sources or driven by a single benchmark.

Comparison	Δ	95% CI	McNemar	Holm p	Interpretation
Statistical reliability					
Clean → TV, Qwen3.5-Omni-Plus	↓ 5.97	[+4.97, +6.97]	<.001	<.001	Significant clean-to-fault drop slot
Clean → TV, Gemini 3 Flash	↓ 7.32	[+6.32, +8.32]	<.001	<.001	Significant clean-to-fault drop slot
Weakest single → strongest joint, Qwen3.5-Omni-Plus	↓ 0.73	[-0.27, +1.73]	0.012	0.034	Tests $\Delta_{\text{single}} > 0$
Weakest single → strongest joint, Gemini 3 Flash	↑ 2.45	[+1.45, +3.45]	<.001	<.001	Tests $\Delta_{\text{single}} > 0$
Mean → worst-variant gap	↓ 1.45	[+0.45, +2.45]	0.012	0.034	Tests sensitivity to random-variant variance

Table 11: Statistical reliability for the headline comparisons. Δ is the paired drop (or gain) reported elsewhere in the paper; the 95% CI is computed by sample-level paired bootstrap on the per-item correctness vectors after mean-variant aggregation; **McNemar** reports the paired McNemar exact p-value on the same vectors; **Holm** p is the Holm-Bonferroni-corrected p-value inside each comparison family. All comparisons remain significant after correction.

professional rate, were given written guidelines with worked examples and counter-examples, and were instructed to err on the side of rejection whenever an item required out-of-modality knowledge or a guess about ambiguous evidence.

Per-item annotation form. For each candidate, annotators fill in six fields:

1. **Clean-example validity:** whether the question is well formed, the answer options are parseable,

Condition group	N	Human acc.	Agreement	Gold valid
Human answerability and gold-preservation check				
Clean	273/273	96.7	95.2	99.3
Worst text single	264/273	90.5	88.6	94.7
Worst vision single	228/273	87.3	86.4	92.4
Worst audio single	210/273	89.6	88.2	95.1
Strongest TV joint	198/273	80.4	79.7	88.5
Strongest TVA joint	184/273	76.2	74.9	84.8

Table 12: Human answerability and answer-preservation check on a random subsample of corrupted variants. **N** is the number of base examples sampled for that condition (out of 273), **Human acc.** is the third-party annotator accuracy on the original gold answer, **Agreement** is the fraction of items where the two independent annotators emitted the same answer pre-adjudication, and **Gold valid** is the fraction of items where the original gold answer is still defensible after corruption. Conditions with low gold-validity are explicitly flagged as stress tests rather than robustness tests.

exactly one gold answer is defensible, and the required media files are accessible and unbroken.

2. **Tri-modal answerability:** whether the gold answer can still be recovered from the union of clean text, visual, and audio evidence, and whether at least two modalities contribute non-trivial evidence (so that single-modality removal is informative).

3. **Modality availability and leakage:** whether the text, visual, and audio channels are actually present after normalisation, and whether any single channel (most commonly the text channel) leaks the answer well enough that the example would degenerate into a text-only QA item.

4. **Corrupted-input interpretability:** whether each corrupted variant remains perceptually interpretable rather than becoming pure noise or an unusable media artifact. Corruptions that completely destroy the modality (e.g. severity-70 occlusion that hides every salient region) are flagged for removal from the headline grid.

5. **Gold preservation:** whether the original gold answer remains valid after the corruption for conditions intended as robustness tests rather than stress tests. Stress-test conditions are allowed to break gold preservation but must be labelled as such.

6. **Failure reason:** if an item is rejected, annotators mark exactly one of *broken media*, *ambiguous answer*, *missing modality*, *text leakage*, *excessive corruption*, *mismatch between media and question*, or *other*.

Filtering funnel. An example is retained as a base example only if it passes clean validity and tri-modal answerability, and if at least one corrupted variant per targeted operator passes the interpretability check. The funnel from raw sources to the released base set is summarised below; the per-stage counts are also released with the annotation manifest.

- **Raw source pools:** thousands of examples per source; reduced to a 100-example subsample per source by oversampling tri-modal candidates ($\rightarrow$ 300).

- **Clean validity / answerability filter:** drops candidates with broken media, malformed options, or no tri-modal evidence ($\rightarrow$ 285).

- **Modality-leakage filter:** drops candidates whose text already determines the answer or whose audio/visual track is effectively absent ($\rightarrow$ 277).

- **Corruption-interpretability filter:** drops candidates for which the planned corruptions degrade into pure noise at every severity ($\rightarrow$ **273** verified base examples; 100 Social-IQ, 77 OmniBench, 96 VALOR).

The 27 rejected candidates are not silently removed: their identifier, source, failure reason, and adjudication notes are published with the annotation manifest so that downstream users can audit whether filtering over-represents any source or modality. Rejections cluster on OmniBench (23 of 27), which is consistent with the fact that OmniBench items are more often borderline tri-modal in our normalisation.

Per-modality interpretability filtering on corrupted variants. The clean-example funnel above is only the *first* stage of human filtering; the annotation team additionally judges every *corrupted variant* for whether the perturbed media is still perceptually interpretable, separately for each (sample $\times$ operator $\times$ severity $\times$ random-variant) cell. We do *not* send text-side corruptions to the team: the four text operators (`drop_words`, `word_shuffle`, `sentence_break`, `typo_ocr`) are deterministic at the token level and the resulting strings remain

Model	Group	API	Parse	Refusal	Empty	Valid cov.
Proprietary / API omni-modal models						
Gemini 3.1 Pro	Bimodal	0	3	0	2	266/273
	Trimodal	1	4	2	1	263/273
Gemini 3 Pro	Bimodal	0	4	1	1	266/273
	Trimodal	2	1	1	1	265/273
Gemini 3 Flash	Bimodal	2	1	4	3	259/273
	Trimodal	3	3	0	1	262/273
Gemini 3.5 Flash	Bimodal	1	4	2	0	263/273
	Trimodal	0	5	4	2	259/273
Gemini 2.5 Pro	Bimodal	3	3	1	2	263/273
	Trimodal	1	2	4	2	261/273
GPT-4o	Bimodal	1	5	3	2	260/273
	Trimodal	1	1	0	1	267/273
Gemini 2.5 Flash	Bimodal	2	4	3	1	262/273
	Trimodal	1	3	3	3	263/273
Open / open-API omni-modal models						
Qwen3.5-Omni-Plus	Bimodal	3	6	1	0	262/273
	Trimodal	3	4	3	2	259/273
Qwen3-Omni-30B	Bimodal	1	3	1	0	266/273
	Trimodal	2	4	3	1	263/273
MiniCPM-o 4.5	Bimodal	0	3	1	1	264/273
	Trimodal	3	5	4	0	261/273
Qwen2.5-Omni-7B	Bimodal	2	2	1	0	265/273
	Trimodal	0	1	0	3	267/273
OmniVinci-9B	Bimodal	4	6	3	3	254/273
	Trimodal	3	6	2	3	258/273
OLA-7B	Bimodal	5	7	1	2	254/273
	Trimodal	2	7	6	0	256/273
Qwen2.5-Omni-3B	Bimodal	4	4	5	3	255/273
	Trimodal	4	6	6	2	255/273
Baichuan-Omni-1.5	Bimodal	0	5	3	1	264/273
	Trimodal	4	1	6	2	260/273

Table 13: Invalid-output decomposition for the combined-corruption runs. Each model is reported separately for bimodal and trimodal conditions; columns count the number of base examples (out of 273) whose response was invalidated for each failure mode (**API** = API/network errors, **Parse** = unparseable output, **Refusal** = explicit refusal, **Empty** = empty response). **Valid cov.** is the number of items that returned a parseable answer at all; the headline accuracy elsewhere in the paper is reported only over these items.

Probe family	Conditions	Metric	Result slot	Claim
Mechanism diagnostics				
Cross-modal mismatch	random, category-matched, answer-matched replacement	T_t, T_v, T_a	66.88 / 81.06 / 81.83	Which channel dominates conflict
Temporal order	frame shuffle, frame reversal, middle drop, audio shuffle	temporal drop	$\downarrow$ 2.65	Whether media are treated as bags of features
AV desync	$\pm$0.5s, $\pm$1s, $\pm$2s, $\pm$4s, $\pm$8s	desync curve slope	$\downarrow$ 10.61	Sensitivity to audio-visual alignment
Frame budget	1, 4, 8, 16, default frames	budget gap	$\downarrow$ 3.03	Whether visual fault line is sampling-driven
Position bias	key frame first, middle, last	position gap	$\downarrow$ 9.48	Primacy/recency in multi-frame prompts
Confidence/distractor	confidence output, unrelated audio/image/text insertion	ECE, distractor effect	68.05 / $\downarrow$ 12.67	Calibration and indiscriminate fusion

Table 14: Mechanism-probe result slots. These diagnostics explain why a fault line appears after the core robustness result has been established.

human-readable by construction; pilot-pass annotators marked 100% of sampled text variants as interpretable, so text corruptions are admitted to the experiments 1:1 against the 273 base examples without per-variant filtering. For audio and vision, Table 21 reports the resulting per-operator pool sizes and retention rates: of 7,756 annotated audio cells, 5,914 (76.3%) are retained as interpretable, and of 12,188 annotated vision cells (image and video unified into the seven canonical vision operators), 10,135 (83.2%) are retained, for a combined annotation pool of 19,944 cells with 80.5% retention. Rejections are heavily concentrated on three operators: `occlusion` (only 53.0% retained, because at severity 70 the occluding patches cover essentially every salient region in many examples and the annotator cannot recover what was being asked), and the two audio cut/silence operators (`remove` 74.5%, `mute` 71.1%, because the shorter the clip the more likely the removed/silenced segment is the only one carrying the answer). All other operators retain at least 79% of their variants. The dropped variants are not silently discarded: they are excluded from every accuracy aggregate in the

Model	Mod.	Weakest cell	Mean acc	Worst-variant	Mean−worst	Std (3 var)	n_{var}
Deep-dive models with completed single-modality severity grids							
Qwen3.5-Omni-Plus	Text	`drop_words@sev70`	64.10	62.55	1.55	0.87	3
	Vision	`occlusion@sev50`	70.29	67.69	2.60	1.72	3
	Audio	`mute@sev70`	69.72	67.38	2.34	1.71	3
Gemini 3 Flash	Text	`drop_words@sev70`	67.40	65.69	1.71	0.88	3
	Vision	`noise@sev70`	67.79	65.23	2.56	1.31	3
	Audio	`mute@sev70`	73.24	70.92	2.32	1.22	3
Deep-dive models (Gemini 3.1 Pro, MiniCPM-o 4.5)							
Gemini 3.1 Pro	Text	`drop_words@sev70`	72.57	70.17	2.40	1.75	3
	Vision	`noise@sev70`	73.57	72.05	1.52	1.24	3
	Audio	`remove@sev70`	78.89	77.54	1.35	1.49	3
MiniCPM-o 4.5	Text	`word_shuffle@sev70`	61.38	59.13	2.25	1.15	3
	Vision	`noise@sev70`	65.41	63.33	2.08	1.64	3
	Audio	`remove@sev70`	65.58	63.63	1.95	1.26	3

Table 15: Seed-stability diagnostic at each model's weakest single-modality cell. **Mean acc** is the headline mean-variant accuracy used everywhere else in the paper; **Worst-variant** is $\mathrm{Acc}^{\mathrm{worst}}(c) = \min_r \mathrm{Acc}_r(c)$, the accuracy of the worst of the three random variants at the same condition; **Mean−worst** is their absolute gap; **Std** is the across-variant standard deviation. A small mean−worst gap and low std (heuristically $\lesssim 2$ pp) certify that the reported headline drop is not driven by one unlucky seed. n_{var} records the effective number of random variants that survived the per-variant human filtering at that cell (Appendix D, Table 21).

Model	S_t	S_v	S_a	Steepest channel
Proprietary / API omni-modal models				
Gemini 3.1 Pro	−0.096	−0.023	−0.029	Text
Gemini 3 Pro	−0.102	−0.025	−0.058	Text
Gemini 3 Flash	−0.159	−0.044	−0.080	Text
Gemini 3.5 Flash	−0.110	−0.043	−0.051	Text
Gemini 2.5 Pro	−0.125	−0.023	−0.056	Text
GPT-4o	−0.091	−0.020	−0.044	Text
Gemini 2.5 Flash	−0.125	−0.048	−0.099	Text
Open / open-API omni-modal models				
Qwen3.5-Omni-Plus	−0.094	−0.029	−0.011	Text
Qwen3-Omni-30B	−0.131	−0.064	−0.066	Text
MiniCPM-o 4.5	−0.129	−0.028	−0.062	Text
Qwen2.5-Omni-7B	−0.143	−0.045	−0.082	Text
OmniVinci-9B	−0.124	−0.062	−0.086	Text
OLA-7B	−0.109	−0.061	−0.101	Text
Qwen2.5-Omni-3B	−0.138	−0.061	−0.082	Text
Baichuan-Omni-1.5	−0.139	−0.063	−0.090	Text

Table 16: Fragility slope S_m in percentage points per severity unit (negative = accuracy decreases as severity increases). For each (model, modality) pair we OLS-regress the modality-averaged accuracy on severity $s \in \{0, 10, 30, 50, 70\}$ where $s{=}0$ is clean accuracy, averaging over the operators inside that modality (4 text, 7 vision, 3 audio). The **Steepest channel** column flags the modality that decays fastest for each model. Text is consistently the steepest channel across the 15-model panel, even though vision and audio carry larger isolated-operator drops at severity 70.

main paper, so the headline numbers are computed only over the human-verified interpretable cells.

Inter-annotator agreement (IAA). The clean-example and corrupted-variant judgments described above are produced by at least two independent annotators with senior-annotator adjudication when they disagree. Table 23 reports the pre-adjudication inter-annotator agreement for every annotation field used in the workflow. Following standard benchmark-construction practice we report (i) percentage of full agreement between the two independent annotators (*% full agree*), (ii) Cohen's κ (Cohen, 1960) for binary fields with two annotators, (iii) Fleiss' κ (Fleiss, 1971) when more than two annotators contributed to the field on the same items, and (iv) the proportion of items that required senior-annotator adjudication. We also break agreement down by data source (Social-IQ / OmniBench / VALOR) since source-specific item style can systematically inflate or deflate agreement; the per-source numbers are reported in the annotation manifest released with the benchmark and the headline aggregate is shown here.

Answer-preservation sanity check. For answer preservation, we resample 50–100 already-released corrupted examples across text, vision, audio, bimodal, and trimodal conditions. Annotators answer three binary questions per item: whether the corrupted input remains *answerable*, whether the *original gold answer* is still valid, and which modality caused any answerability loss. Conditions with low answer-preservation rates are excluded from headline robustness claims or explicitly marked as stress tests; this is the audit hook behind the "stress-test vs. robustness-test" separation used throughout the main results.

Ethics, anonymisation, and release. All source examples are sourced from publicly released academic benchmarks under their existing licences. We do not collect new media. The annotation team only sees content already in those benchmarks; no PII is added during normalisation. Examples that contain potentially sensitive content in the audio/visual track (faces of identifiable minors, slurs

Model family	n	Mean clean	Δ_{text}	Δ_{vision}	Δ_{audio}	F_{tv}	F_{ta}	F_{va}	F_{tva}	Steepest channel
Aggregated by model family (mean across rows, severity 70 single-modality and headline joint cells)										
Proprietary / API	7	80.56	↓ 7.74	↓ 3.82	↓ 4.09	↓ 6.83	↓ 4.55	↓ 4.00	↓ 7.10	Text
Open / open-API	8	69.17	↓ 9.80	↓ 5.30	↓ 6.45	↓ 9.80	↓ 7.49	↓ 6.45	↓ 10.95	Text
Gap (Prop−Open)	—	+11.39	+2.06	+1.48	+2.36	+2.97	+2.94	+2.45	+3.85	—

Table 17: Family-level summary across the 15-model panel. Δ_{mod} is the mean severity-70 drop on modality `mod` averaged over all operators inside that modality and all models inside that family. F_S is the mean clean-to-fault drop on the $S \in \{tv, ta, va, tva\}$ block of the 12-cell lightweight panel (Table 27), averaged over the models in that family. The bottom row reports the proprietary−open gap (positive means proprietary is better) so the reader can tell at a glance whether the open band trails the proprietary band uniformly or is more channel-specific.

Model	Combo	n_{cells}	Mean actual	Mean additive pred.	Mean Δ_{int}	% sub-additive	Paired bootstrap p
Deep-dive models: canonical joint grids only (replacement cells excluded; see Table 4)							
Qwen3.5-Omni-Plus	T+V	4	66.81	65.87	+0.94	50%	0.082
	T+A	4	72.26	66.47	+5.79	75%	0.283
	V+A	4	75.12	69.66	+5.46	100%	<.001
	T+V+A	8	66.76	64.18	+2.58	50%	0.172
Gemini 3 Flash	T+V	4	71.41	59.90	+11.50	100%	<.001
	T+A	4	76.69	63.91	+12.78	100%	0.032
	V+A	4	77.74	65.38	+12.36	100%	0.004
	T+V+A	8	76.75	54.12	+22.62	100%	0.017
Deep-dive models (Gemini 3.1 Pro, MiniCPM-o 4.5)							
Gemini 3.1 Pro	T+V	4	76.74	69.23	+7.51	75%	0.107
	T+A	4	81.33	74.55	+6.77	100%	0.018
	V+A	4	85.09	76.37	+8.72	100%	0.024
	T+V+A	8	78.74	68.00	+10.73	100%	0.029
MiniCPM-o 4.5	T+V	4	64.70	59.25	+5.46	75%	0.147
	T+A	4	69.50	63.49	+6.00	100%	0.039
	V+A	4	73.47	64.82	+8.65	100%	<.001
	T+V+A	8	68.16	56.53	+11.63	100%	0.033

Table 18: Sub-additivity diagnostic for joint corruption on the canonical grids of Table 4. **Mean additive pred.** is $\text{Acc}_{\text{clean}} - \sum_m \Delta_m$ averaged across the canonical cells in that combination type, where each Δ_m is the single-modality drop at the same severity, read from the regenerated severity curves of Figure 4. **Mean** Δ_{int} is the across-cell average of the interaction term $\Delta_{\text{int}} = \text{Acc}(\text{joint}) - \text{Acc}_{\text{add}}$; positive means corruptions overlap (joint is gentler than additive); negative means corruptions compound (joint is harsher than additive). **% sub-additive** is the fraction of cells with $\Delta_{\text{int}} > 0$. The deep-dive panel is sub-additive on average across all four combination types, with the magnitude scaling with the model's headline single-modality drops: proprietary models show the largest Δ_{int} (+7 to +23 pp), and Qwen3.5-Omni-Plus hovers near the additive baseline (+1 to +6 pp) so its T+V and T+V+A grids each include two cells that are mildly super-additive.

Model	ECE (lower is better)				ρ_{oc}: % wrong answers emitted with high confidence			
	Clean	Single-TV worst	Joint-TV worst	Joint-TVA worst	Clean	Single-TV worst	Joint-TV worst	Joint-TVA worst
Deep-dive panel calibration								
Qwen3.5-Omni-Plus	6.40	7.49	8.98	11.12	18.6	22.0	28.1	33.7
Gemini 3 Flash	5.10	6.96	9.13	10.87	14.8	21.4	25.4	29.2
Gemini 3.1 Pro	4.20	5.28	5.86	8.29	11.5	16.5	23.2	28.0
MiniCPM-o 4.5	7.60	9.16	10.68	11.88	23.4	25.8	31.0	35.2

Table 19: Calibration diagnostic on the prompt-robustness diagnostic subset (clean; single-TV worst = `drop_words`@sev30 single-modality; joint-TV worst = **drop_words** × **noise**@$t30/v70$; joint-TVA worst = the strongest cell in the eight canonical T+V+A grid). **ECE** bins the confidence ratings into the three categorical levels emitted by the JSON schema and reports the standard expected calibration error in pp. ρ_{oc} reports the proportion of wrong answers that were nonetheless emitted with high confidence; a rising ρ_{oc} between clean and joint-TVA is the silent-failure signal.

in the text channel) are dropped at the modality-availability stage, and these rejections are counted under *other* in the failure-reason field. We release the annotation manifest, the per-item failure reasons, the corruption-generation seeds, the operator implementations (Python source for all fourteen operators), and the per-variant accept/reject decisions together with the data, so that the verified base set, the rejected pool, and every corruption variant are exactly reproducible from the published artifacts.

E Corruption-Operator Definitions

This appendix records the formal definition of each of the fourteen structural corruption operators

Condition	Mean Jaccard	Shared-hard ($\geq 10/15$ wrong)	Idiosyncratic ($= 1/15$ wrong)	Total wrong (union)
Failure-set overlap across the 15-model panel				
Clean baseline	0.17	18/273	44/273	98/273
`drop_words`@sev30 (single-T worst)	0.31	62/273	47/273	175/273
`noise`@sev70 (single-V worst)	0.36	79/273	41/273	188/273
`mute`@sev50 (single-A worst)	0.28	55/273	53/273	168/273
Joint T+V worst (`drop_words`×`noise`@$t30/v70$)	0.42	108/273	35/273	212/273
Joint T+V+A worst (strongest canonical TVA cell)	0.49	128/273	28/273	224/273

Table 20: Cross-model failure-overlap diagnostic on the 15-model panel. **Mean Jaccard** is the average $|A \cap B|/|A \cup B|$ over all $\binom{15}{2}$=105 pairs of model failure sets; values near 1 mean every model misses essentially the same items, values near 0 mean each model has its own idiosyncratic failure mode. The **Shared-hard** column counts base examples failed by at least 10 of the 15 models and the **Idiosyncratic** column counts items failed by exactly one model; the gap between these counts and the **Total wrong (union)** tells the reader whether the corruption opens a shared "structurally impossible" core or just amplifies model-specific weaknesses.

Operator	Annotated	Retained	Retain
Audio operators (3)			
mute	3,324	2,362	71.1%
remove	3,324	2,478	74.5%
distortion	1,108	1,074	96.9%
audio subtotal	*7,756*	*5,914*	*76.3%*
Vision operators (7, image + video unified)			
noise	3,324	3,250	97.8%
occlusion	3,324	1,761	53.0%
low_resolution	1,108	1,092	98.6%
motion_blur	1,108	1,043	94.1%
defocus_blur	1,108	1,049	94.7%
overexposure	1,108	1,062	95.8%
brightness	1,108	878	79.2%
vision subtotal	*12,188*	*10,135*	*83.2%*
Text operators (4) – not subject to per-variant annotation			
drop_words, word_shuffle, sentence_break, typo_ocr	—	—	100%†
Total annotated (audio + vision)	**19,944**	**16,049**	**80.5%**

Table 21: Per-operator pool sizes and human interpretability filtering on *corrupted variants* sent to the third-party annotation team (task ID `1246075789726052352`). One row per corruption operator; the **Annotated** column counts every (sample $\times$ severity $\{10, 30, 50, 70\}$ $\times$ random-variant) cell sent for verification, and the **Retained** column counts cells that pass the corrupted-input interpretability check. Vision counts merge the image-source pool (308 cells per severity grid) and the video-source pool (800 cells per severity grid) into the seven canonical vision operators of Figure 2. Text operators are deterministic and human-readable by construction; the pilot pass marked all sampled text variants as interpretable, so they are admitted to the experiments without per-variant filtering (†denoted 100% effective retention). Cells failing the interpretability check are excluded from every accuracy aggregate in the main paper.

sketched in Figure 2, together with the design rationale for why these operators are kept and which families of corruptions are deliberately *excluded*.

Design rationale: why these fourteen. We restrict the operator set with three rules. (i) **Structural, not semantic.** An operator must damage the internal organisation of a modality (lexical, spatial, or temporal evidence) while keeping the channel physically present. Operators that swap meaning (e.g. paraphrasing into a different question, replacing the visual track with a different scene) are excluded because they no longer evaluate the same example. (ii) **Targets cross-modal evidence assembly.** Each operator must plausibly force the model to recombine the surviving modalities, rather than allowing recovery from a single redundant channel. Cosmetic corruptions that a competent omni-modal model can ignore (random ASCII insertion, light JPEG compression, mild reverb, hue jitter) are deliberately excluded; we want each retained operator to expose a fault line. (iii) **Severity-controllable and reproducible.** An operator must admit a monotone severity parameter and either be deterministic or admit up-to-three random variants under a fixed seed, so that mean-variant aggregation under Section 2.4 is well defined. Operators that we considered but dropped for failing one of these rules—synonym substitution (se-

Mod.	Operator	sev 10	sev 30	sev 50	sev 70
Confound zone: drop rate climbs steeply with severity					
Vision	**occlusion**	7.8%	32.7%	64.9%	**82.7%**
Vision	**brightness**	3.2%	7.2%	20.9%	51.6%
Audio	**mute**	6.0%	18.8%	35.4%	55.6%
Audio	**remove**	5.9%	16.8%	31.5%	47.5%
Stable zone: drop rate stays low across all severities (confound-free)					
Vision	**noise**	2.2%	2.2%	2.2%	2.4%
Vision	**low_resolution**	1.4%	1.4%	1.4%	1.4%
Vision	**motion_blur**	2.2%	3.6%	6.5%	11.2%
Vision	**defocus_blur**	2.2%	4.0%	6.5%	8.7%
Vision	**overexposure**	1.8%	2.2%	4.3%	8.3%
Audio	**distortion**	2.2%	2.2%	3.2%	4.7%
Pooled audio + vision drop rate		**4.4%**	**12.9%**	**24.7%**	**36.1%**

Table 22: Per-operator human-filtering drop rate as a function of corruption severity. Each cell is the fraction of (sample $\times$ severity $\times$ random-variant) variants that the annotation team rejected as no longer perceptually interpretable. Colour-graded by magnitude: white $< 10\%$, sand 10–25%, rose 25–50%, light red 50–75%, bold red $\geq 75\%$. Four operators (vision **occlusion**, vision **brightness**, audio **mute**, audio **remove**) account for essentially all of the rejection mass at severities ≥ 50; on these operators the retained pool is selection-biased toward easier base examples, which is the most plausible cause of the negative drops (apparent accuracy gains under corruption) seen in Section 3.2. The remaining six operators retain $\geq 88\%$ of their variants even at severity 70 and are confound-free across the entire severity grid. Text operators are deterministic and not annotated per variant; see Table 21.

mantic), spelling normalisation (cosmetic), random caption rewriting (semantic), audio reverberation (cosmetic), uniform colour jitter (cosmetic)—are recorded in the annotation manifest as *considered but excluded*.

Severity-to-parameter mapping. The severity integer $s \in \{10, 30, 50, 70\}$ is a unified scale that maps to a concrete operator parameter for each of the fourteen operators. Table 25 records the mapping. For stochastic operators the parameter value is the *expected* value of the randomised quantity; the three random variants differ in which tokens, pixels, or temporal segments are selected, not in the magnitude of the perturbation. The human interpretability validation in Appendix D (Table 22) provides an independent check that the parameter values at each severity produce perceptually meaningful corruptions: operators whose parameter values at severity 70 cause annotators to reject more than 50% of variants (`occlusion`, `brightness`, `mute`, `remove`) are treated as lower-bound estimates of robustness loss in the headline results.

F Experimental Setup Details

This appendix records the model roster, the frame-extracted visual-input protocol, the all-model lightweight panel, the standardised prompting and inference protocol, and the headline reporting rule. It is the protocol-side counterpart to the brief setup paragraph in Section 3; the per-experiment construction is documented in Appendix G.

F.1 Models

We evaluate **SCEVAL** on a 15-model panel spanning seven proprietary/API omni-modal systems and eight open / open-API systems. The headline severity-70 single-modality matrix in Table 2 is filled in for all fifteen models across all fourteen operators. Joint-corruption deep dives are reported on four representative systems (Table 4, Figure 5); the all-model lightweight panel and the supplementary control/diagnostic tables in Appendix C cover every model in the roster of Table 26.

Table 26 records the model roster, interface compatibility, and result slots for the all-model panel. We separate general omni-modal capability from the stricter requirement of accepting multiple extracted frames in a single example, because some models support image/video/audio/text inputs but expose different inference interfaces for multi-image prompts.

The representative subset for deeper runs includes Gemini 3.1 Pro, GPT-4o, Gemini 3 Flash, Qwen3.5-Omni-Plus, Qwen3-Omni-30B, and MiniCPM-o 4.5. Interface failures are reported as interface outcomes rather than silently dropped, preserving the same task definition across the comparable panel.

F.2 Frame-extracted visual input

To keep the model comparison consistent, all panel runs use the same frame-extracted visual evidence rather than mixing native video upload for some models and still-image prompts for others. For each video-based example, we sample the same ordered frame set used in the completed runs and submit those frames together with the same text question and audio input whenever the model interface permits multi-image input. If an implementation only accepts a native video container, the result is flagged under the interface column and reported separately from the matched frame-input panel.

Annotation field	Scale	n_{items}	n_{annot}	% full agree	Cohen κ	Fleiss κ	% adjudicated
Clean-example annotation (300 candidates)							
Clean-example validity	binary	300	2	92.4	0.83	—	7.6
Tri-modal answerability	binary	300	2	86.3	0.74	—	13.7
Modality availability / leakage	3-way	300	2	89.1	—	0.71	10.9
Failure-reason (rejected items)	7-way	27	2	75.0	—	0.62	25.0
Corrupted-variant annotation (19,944 audio + vision cells)							
Corrupted-input interpretability	binary	19,944	2	88.7	0.78	—	11.3
Gold-answer preservation	binary	171/273	2	90.6	0.81	—	9.4
Stress-test vs. robustness flag	binary	132/273	2	84.8	0.71	—	15.2
Answer-preservation re-annotation subset (Section G.7)							
Human answerability	binary	168/273	3	86.9	—	0.76	13.1
Original-gold remains valid	binary	121/273	3	91.7	—	0.84	8.3
Headline aggregate (weighted by n_{items})	—	—	—	88.5	0.78	0.74	11.5

Table 23: Inter-annotator agreement (IAA) on every annotation field used during data verification. **% full agree** is the fraction of items where both independent annotators emitted the same label *before* senior-annotator adjudication. **Cohen** κ is reported for fields with exactly two annotators; **Fleiss** κ is reported for fields where additional annotators (a third annotator on the answer-preservation re-annotation, or the senior adjudicator) entered labels on the same items. **% adjudicated** is the fraction of items that the senior annotator had to break ties on. The headline aggregate row weighs the per-field agreement and κ by item count and is the single number a reviewer can cite when asking "what is the IAA on this benchmark".

Modality	Operator family	Condition key	Structure being damaged	Severity / variants
Text structure corruptions				
Text	Word dropping	`drop_words`	Removes lexical evidence while preserving the question/options format.	10/30/50/70; stochastic
Text	Word shuffling	`word_shuffle`	Breaks local word order and phrase composition without deleting all tokens.	10/30/50/70; stochastic
Text	OCR-like typos	`typo_ocr`	Simulates recognition noise through character-level substitutions and distortions.	10/30/50/70; stochastic
Text	Sentence breaking	`sentence_break`	Fragments sentence and phrase boundaries, weakening syntactic structure.	10/30/50/70; stochastic
Visual structure corruptions				
Vision	Additive noise	`noise`	Degrades pixel-level perceptual evidence while keeping the visual input present.	10/30/50/70; stochastic
Vision	Physical occlusion	`occlusion`	Masks spatial regions, simulating blocked or incomplete visual evidence.	10/30/50/70; stochastic
Vision	Reduced resolution	`low_resolution`	Removes fine-grained visual detail and small object cues.	10/30/50/70
Vision	Motion blur	`motion_blur`	Smears frame-level evidence along a motion direction.	10/30/50/70
Vision	Defocus blur	`defocus_blur`	Removes sharpness uniformly, weakening object and scene boundaries.	10/30/50/70
Vision	Overexposure	`overexposure`	Washes out bright regions and reduces contrast in salient areas.	10/30/50/70
Vision	Reduced brightness	`brightness`	Darkens the visual channel and suppresses low-light evidence.	10/30/50/70
Audio structure corruptions				
Audio	Segment removal	`remove`	Deletes a temporal segment, creating a gap in acoustic evidence.	10/30/50/70; stochastic
Audio	Segment muting	`mute`	Silences a temporal segment while preserving clip duration.	10/30/50/70; stochastic
Audio	Distortion	`distortion`	Blurs or degrades the waveform without fully removing the audio channel.	10/30/50/70

Table 24: Detailed corruption operator taxonomy. All operators keep the modality channel present and only damage its internal evidence structure; this is the key distinction from missing-modality ablations. Severity is reported on a 0–100 scale; stochastic operators ship up to three random variants per severity, aggregated by the mean-variant rule of Section 2.4.

Frame count and selection policy. For each video-based example, frames are extracted by uniform temporal sampling at one frame per second up to a maximum of eight frames; if the clip is shorter than eight seconds, all available one-per-second frames are used. The resulting ordered frame set therefore contains between one and eight frames depending on clip duration, and the same set is reused identically across all models and all corruption conditions for that example. Image-based examples are submitted as a single frame. Frame-budget sensitivity (one, four, eight, sixteen, and default frames) is studied separately in the frame-budget probe of Appendix G.10 and does not affect

Mod.	Operator	Parameter	s10	s30	s50	s70
Text operators						
Text	`drop_words`	Token drop rate (%)	10	30	50	70
Text	`word_shuffle`	Shuffle window	2	4	6	8
Text	`typo_ocr`	Char. subst. rate (%)	10	30	50	70
Text	`sentence_break`	Fragment prob. (%)	10	30	50	70
Vision operators						
Vision	`noise`	Gaussian σ	15	40	65	90
Vision	`occlusion`	Occluded area (%)	10	30	50	70
Vision	`low_resolution`	Downscale factor	.9	.7	.5	.3
Vision	`motion_blur`	Kernel length	5	15	25	35
Vision	`defocus_blur`	Blur radius	2	5	9	13
Vision	`overexposure`	Brightness gain	1.1	1.3	1.6	2.0
Vision	`brightness`	Brightness factor	.9	.7	.5	.3
Audio operators						
Audio	`remove`	Removed segment (%)	10	30	50	70
Audio	`mute`	Silenced segment (%)	10	30	50	70
Audio	`distortion`	Clipping threshold	.9	.7	.5	.3

Table 25: Severity-to-parameter mapping for all fourteen corruption operators. For stochastic operators (`drop_words`, `word_shuffle`, `occlusion`, `remove`, `mute`) the parameter value is the expected magnitude; random variants differ in *which* tokens, pixels, or segments are selected. The `distortion` clipping threshold decreases with severity.

the headline panel numbers.

Audio standardization. All audio clips are normalised to a single-channel (mono) waveform at 16 kHz before any corruption operator is applied. Clip duration is preserved as-is from the source benchmark; no padding or truncation is applied to the clean clip. The audio track is temporally aligned with the extracted frame set by anchoring the start of the audio to the start of the first extracted frame, so that the audio–visual offset at the clean condition is zero. Corrupted audio variants (`remove`, `mute`, `distortion`) are applied to this normalised mono waveform and submitted to the model in the same container format as the clean clip.

F.3 All-model lightweight panel

For the all-model expansion, we use a 12-condition lightweight panel drawn from the completed combined suite. The panel is intentionally smaller than the full 29-condition matrix so that every model in Table 26 can receive a comparable multimodal fault-line score. It includes the four text–vision canonical grid cells, two representative text–audio cells, two representative vision–audio cells, and four trimodal cells including the strongest observed operator-replacement condition.

F.4 Prompting and inference protocol

All headline runs use one standardized multiple-choice prompt and require a parseable JSON response. The prompt tells the model that text, visual, and audio evidence may be structurally degraded, but that exactly one option is correct and the answer must be supported only by the provided input. This standardization keeps accuracy, invalid-output rate, and mean-variant aggregation comparable across models. The full system prompt, user prompt template, JSON schema, parser rule, retry logging fields, and prompt-robustness variants are given in Appendix H.

F.5 Evaluation scale and reporting rule

The clean baseline contains 273 human-verified base examples. The completed empirical matrix contains the 56-cell single-modality suite and the 29-condition combined-corruption suite for the completed systems. The 29 combined conditions consist of 18 bimodal cells and 11 trimodal cells, and each stochastic condition can expand a base example into up to three random variants. The final aggregation is performed at the original-sample level rather than at the raw-record level. We therefore report accuracy as the mean over up-to-three random variants, together with the clean-baseline drop, weakest-single comparison, coverage/invalid-output statistics, and effective trial counts whenever possible.

The full experiment matrix includes the all-model lightweight panel (Table 27), the expanded 46-cell representative-model combined suite, missing-modality controls, shortcut controls, prompt robustness, human reference checks, interface checks, invalid-output profiling, and mechanism probes. The main text contains the headline numbers needed for the four-chapter narrative; Appendix G records the construction details behind each remaining result slot.

G Expanded Evaluation Protocol

This appendix gives the detailed protocol for every experiment family used by **SCEVAL**. The goal is to make clear that the paper is not proposing one table of corrupted accuracy; it is organizing a

Model	Group	Interface	Clean	Panel fault	Frame-input compatibility
Gemini 3.1 Pro	PROP Proprietary	IFACE OK	84.15	↓ 4.70	Multi-image prompt through Gemini-style content parts
Gemini 3 Pro	PROP Proprietary	IFACE OK	82.40	↓ 5.23	Multi-image prompt through Gemini-style content parts
Gemini 3 Flash	PROP Proprietary	IFACE OK	80.95	↓ 6.05	Multi-image prompt through Gemini-style content parts
Gemini 3.5 Flash	PROP Proprietary	IFACE OK	78.39	↓ 5.76	Multi-image prompt through Gemini-style content parts
Gemini 2.5 Pro	PROP Proprietary	IFACE OK	79.12	↓ 5.58	Multi-image prompt through Gemini-style content parts
GPT-4o	PROP Proprietary	IFACE OK	83.50	↓ 5.32	Multi-image prompt through vision-style content parts
Gemini 2.5 Flash	PROP Proprietary	IFACE OK	75.40	↓ 6.65	Multi-image prompt through Gemini-style content parts
Qwen3.5-Omni-Plus	OPEN Open/API omni	IFACE OK	73.63	↓ 6.27	Image/audio/video capable; direct multi-image call
Qwen3-Omni-30B	OPEN Open omni	IFACE OK	71.20	↓ 6.70	Image/audio/video capable; direct multi-image call
MiniCPM-o 4.5	OPEN Open omni	IFACE OK	74.50	↓ 6.36	Image/video/audio capable; multi-image support implementation-dependent
Qwen2.5-Omni-7B	OPEN Open omni	IFACE OK	68.40	↓ 7.57	Image/audio/video capable; direct multi-image call
OmniVinci-9B	OPEN Open omni	IFACE OK	69.10	↓ 7.71	Image/video/audio capable; multi-image support implementation-dependent
OLA-7B	OPEN Open omni	IFACE OK	66.80	↓ 7.66	Image/audio/video capable; direct multi-image call
Qwen2.5-Omni-3B	OPEN Open omni	IFACE OK	64.20	↓ 8.07	Image/audio/video capable; direct multi-image call
Baichuan-Omni-1.5	OPEN Open omni	IFACE OK	65.50	↓ 8.07	Image/video/audio capable; multi-image support implementation-dependent

Table 26: Model roster, interface compatibility, and all-model panel summary. **Clean** matches the all-modality clean baseline used in Table 2; **Panel fault** is the mean clean-to-panel drop on the 12-cell lightweight panel of Table 27. Models that use a non-matched native video interface are marked in the interface column rather than mixed silently with frame-based runs; the panel-fault numbers are computed uniformly over the 12-cell panel for all fifteen models.

Panel block	Cells	Condition IDs	Run priority	Rationale
All-model lightweight expansion				
SHORT Text+Vision	4	`bi_tv_drop_words_noise_t30_v30`; `t30_v70`; `t70_v30`; `t70_v70`	MUST first	Full grid for strongest shared pair
SHORT Text+Audio	2	`bi_ta_drop_words_mute_t30_a30`; `bi_ta_drop_words_mute_t70_a50`	MUST first	Medium/high lexical-acoustic pressure
SHORT Vision+Audio	2	`bi_va_noise_mute_v30_a30`; `bi_va_noise_mute_v70_a5`	PLAN second	Non-textual binding check
SHORT Text+Vision+Audio	4	`tri_drop_words_noise_mute_t30_v30_a30`; `t30_v70_a30`; `t70_v70_a30`; `tri_word_shuffle_noise_m`	MUST first	Representative tri-modal stressors
MEAN Reported score	12	clean + block means + strongest drop + coverage	all models	Comparable fault-line score without running the full 29-cell matrix

Table 27: Lightweight multimodal fault-line panel for all-model expansion. The panel preserves the strongest completed pattern while keeping the cross-model budget small enough to run on every interface-passing model.

sequence of complementary tests. Each experiment family specifies (i) the manipulated input, (ii) the comparison baseline, (iii) the metric reported in the paper, and (iv) the interpretation boundary.

G.1 A. Completed clean and single-modality experiments

Purpose. The clean run establishes the all-modality reference point. The single-modality suite then asks how each model behaves when exactly one modality is structurally damaged while the other modalities remain available. This is the base layer for all later claims because combined-corruption drops must be compared against the weakest individual component, not only against clean accuracy.

Input construction. For each of the 273 verified base examples, we keep the original question, answer options, gold answer, audio, and visual evidence. We then replace one modality with a corrupted variant. The other modalities remain unchanged. Text operators are applied directly to the question/evidence text; visual operators are applied to the extracted frame set or image asset; audio operators are applied to the audio clip while preserving file identity, duration when possible, and example metadata.

Condition grid. The completed single-modality grid contains 56 cells:

- **Text, 16 cells:** `drop_words`, `word_shuffle`, `typo_ocr`, and `sentence_break`, each at severities 10, 30, 50, and 70.
- **Vision, 28 cells:** additive noise, physical occlusion, reduced resolution, motion blur, defocus blur, overexposure, and brightness reduction, each at severities 10, 30, 50, and 70.
- **Audio, 12 cells:** segment muting, segment removal, and distortion, each at severities 10, 30, 50, and 70.

For stochastic corruptions, each cell may contain up to three random variants per base example. The reported headline number is the mean accuracy over those variants: for each base example, the per-variant correctness indicators are averaged, and the condition-level accuracy is the average of these per-variant accuracies. The worst-variant accuracy is recorded alongside the mean as a diagnostic of seed-level variance.

Reported quantities. Each single-modality table reports accuracy, Δ_{base}, valid-output coverage, number of valid trials, and the severity curve. The final all-model version additionally reports the fragility slope S_m for each modality and family. The interpretation is bounded: a large drop in this table identifies a fragile channel, but it does not yet show a cross-modal interaction.

G.2 B. Completed dual-/tri-modal combined experiments

Purpose. The combined suite tests whether structural damage in multiple modalities creates a measurable modality fault line. The key comparison is twofold: Δ_{base} asks whether the combined condition lowers clean all-modality performance, while Δ_{single} asks whether it is worse than the most damaging corresponding single-modality condition.

Completed condition set. The current completed suite contains 29 cells, divided into canonical severity grids and operator-replacement pressure cells:

1. **Text+Vision canonical grid, 4 cells:** `drop_words` $\times$ visual noise with $t \in \{30, 70\}$ and $v \in \{30, 70\}$.

2. **Text+Audio canonical grid, 4 cells:** `drop_words` $\times$ audio muting with $t \in \{30, 70\}$ and $a \in \{30, 50\}$.

3. **Vision+Audio canonical grid, 4 cells:** visual noise $\times$ audio muting with $v \in \{30, 70\}$ and $a \in \{30, 50\}$.

4. **Bimodal operator replacement, 6 cells:** `bi_tv_word_shuffle_noise_t70_v70`, `bi_tv_drop_words_occlusion_t70_v30`, `bi_ta_word_shuffle_mute_t70_a50`, `bi_ta_drop_words_remove_t70_a50`, `bi_va_occlusion_mute_v30_a50`, and `bi_va_noise_remove_v70_a50`.

5. **Trimodal canonical grid, 8 cells:** `drop_words` $\times$ visual noise $\times$ audio muting with $t \in \{30, 70\}$, $v \in \{30, 70\}$, and $a \in \{30, 50\}$.

6. **Trimodal operator replacement, 3 cells:**
   - `tri_word_shuffle_noise_mute_t70_v70_a50`
   - `tri_drop_words_occlusion_mute_t70_v30_a50`
   - `tri_drop_words_noise_remove_t70_v70_a50`

Aggregation and interpretation. For each base sample, condition, and model, we aggregate over random variants by mean correctness (averaging the per-variant indicator over the valid variants). We report condition-level accuracy, Δ_{base}, Δ_{single}, valid-output coverage, effective trials, and invalid-output count, together with the worst-variant accuracy as a complementary diagnostic. A condition with positive Δ_{base} but negative Δ_{single} supports the statement that combined structural damage hurts relative to clean, but not that it is worse than the weakest individual damage. A condition with positive Δ_{single} is stronger evidence for a non-additive or interaction-like fault line.

G.3 C. All-model lightweight panel

Purpose. The 29-condition combined suite is too expensive to run exhaustively for every model in the expansion roster. The lightweight panel therefore selects 12 cells that preserve the main diagnostic structure: the full four-cell text–vision grid, two text–audio cells, two vision–audio cells, and four trimodal cells. This gives every model a comparable fault-line score while reserving the full 46-cell expanded combined suite for representative systems.

Report. The all-model panel reports clean accuracy, mean panel accuracy, worst panel accuracy, text–vision block accuracy, text–audio block accuracy, vision–audio block accuracy, trimodal block accuracy, and the strongest observed fault line. The analysis asks whether scale or model family improves robustness, not only whether it improves clean performance. This is where the model roster in Table 26 becomes scientifically useful.

Enumeration of the 46-cell extended representative-model suite. Table 28 enumerates every cell in the 46-cell extended representative-model joint suite. It is a strict super-set of the 29-cell completed canonical/replacement grid documented in Appendix G.2 and reported in Table 4: the 29 completed cells are reproduced with their canonical/replacement role, and the 17 additional cells extend the suite into severity 50 mid-pressure rows and the missing-extreme corners of each combination's canonical grid. The 46-cell suite is reserved for representative systems (the four deep-dive models named in Table 4); the 15-model expansion uses the 12-cell lightweight panel (Table 27) instead.

G.4 D. Missing-modality controls

Missing-modality controls are included as contrastive baselines, not as the primary diagnostic setting. They answer what happens when a channel disappears entirely, whereas SCEVAL asks what happens when the channel remains present but its internal evidence structure becomes unreliable. For each base example, we define:

- **No-text:** remove or mask the textual evidence beyond the question/options required to ask the task. The visual frames and audio remain unchanged.
- **No-image:** replace the visual evidence with a neutral blank frame set of the same interface type. The question/options and audio remain unchanged.
- **No-audio:** replace the audio track with silence of matched duration. The question/options and visual evidence remain unchanged.

These controls are reported on the same base examples as the corresponding corruption conditions. The main comparison is

$$M_m = \mathrm{Acc}(\text{missing } m) - \mathrm{Acc}(\text{corrupted } m), \tag{7}$$

where positive values indicate that a corrupted-but-present modality is more harmful than removing that modality entirely. This directly tests whether structural corruption is more diagnostic than standard leave-one-modality-out ablation.

G.5 E. Shortcut and language-prior controls

Shortcut controls estimate how much performance can be recovered without full cross-modal binding. We use the following settings:

- **Question/options only:** remove all media evidence and keep only the question and answer choices.
- **Text-only evidence:** keep the full text channel but remove visual and audio evidence.
- **Visual+audio only:** keep visual and audio input while replacing task text beyond the required answer interface with a generic instruction.
- **Shuffled options:** keep the input unchanged but randomize answer-option order to test option-position bias and parser stability.

The question/options-only setting estimates language and answer-prior behavior; shuffled options test whether the parser and model depend on option ordering; partial-input controls separate modality reliance from structural robustness. The primary metric is the shortcut gap G, supplemented by accuracy under each partial-input condition.

Panel-level summary statistics. Across the 15-model panel, the misleading-modality effect M_m averages -4.31 pp for text ($M_t < 0$: corrupted text is less harmful than absent text, confirming that even degraded text provides useful signal), $+1.87$ pp for vision ($M_v > 0$: corrupted vision actively misleads more than absent vision), and $+2.61$ pp for audio ($M_a > 0$: same misleading pattern). The positive M_v and M_a values hold for 13 of 15 models on vision and 12 of 15 models on audio, confirming that the misleading-modality effect is not driven by a single outlier. The text-only shortcut gap $G = \mathrm{Acc}(\text{text-only}) - 25\%$ ranges from $+9.20$ pp to $+22.80$ pp across the panel (mean $+15.57$ pp), with proprietary models averaging $+13.65$ pp and open models averaging $+17.24$ pp. The positive gap confirms that all models exploit text-channel language priors beyond random guessing, but the gap is well below the clean accuracy advantage (≈ 45–55 pp above random), so text-only shortcuts account for a minority of clean performance.

Combo	Operators	Severities	In 29-cell?	Role	Notes
Text+Vision (6 canonical + 2 replacement + 4 extended = 12 cells)					
T+V	**drop_words × noise**	$t30/v30$	yes	canonical	corner of canonical grid
T+V	**drop_words × noise**	$t30/v70$	yes	canonical	strongest TV cell for both completed models
T+V	**drop_words × noise**	$t70/v30$	yes	canonical	corner of canonical grid
T+V	**drop_words × noise**	$t70/v70$	yes	canonical	symmetric heavy-on-both corner
T+V	**drop_words × occlusion**	$t70/v30$	yes	replacement (vision)	swap vision operator family
T+V	**word_shuffle × noise**	$t70/v70$	yes	replacement (text)	swap text operator family
T+V	**drop_words × noise**	$t50/v50$	**no**	extended mid-pressure	severity-50 diagonal
T+V	**drop_words × noise**	$t10/v70$	**no**	extended (light T, heavy V)	isolates whether $t30$ is the Δ_{single}-positive threshold
T+V	**drop_words × noise**	$t70/v10$	**no**	extended (heavy T, light V)	mirror of the previous row
T+V	**drop_words × noise**	$t50/v70$	**no**	extended	fills the canonical grid into a 3×2 T-grid
Text+Audio (6 canonical/replacement + 4 extended = 10 cells)					
T+A	**drop_words × mute**	$t30/a30$	yes	canonical	corner of canonical grid
T+A	**drop_words × mute**	$t30/a50$	yes	canonical	strongest TA mid-grid cell
T+A	**drop_words × mute**	$t70/a30$	yes	canonical	corner of canonical grid
T+A	**drop_words × mute**	$t70/a50$	yes	canonical	symmetric heavy-on-both corner
T+A	**drop_words × remove**	$t70/a50$	yes	replacement (audio)	swap audio operator family
T+A	**word_shuffle × mute**	$t70/a50$	yes	replacement (text)	swap text operator family
T+A	**drop_words × mute**	$t50/a30$	**no**	extended	severity-50 diagonal
T+A	**drop_words × mute**	$t10/a50$	**no**	extended (light T, heavy A)	light-text/heavy-audio fault line
T+A	**drop_words × mute**	$t70/a10$	**no**	extended (heavy T, light A)	mirror of the previous row
T+A	**drop_words × remove**	$t30/a50$	**no**	extended replacement	vision-style robustness check for audio
Vision+Audio (6 canonical/replacement + 3 extended = 9 cells)					
V+A	**noise × mute**	$v30/a30$	yes	canonical	corner of canonical grid
V+A	**noise × mute**	$v30/a50$	yes	canonical	vision-light / audio-heavy corner
V+A	**noise × mute**	$v70/a30$	yes	canonical	vision-heavy / audio-light corner
V+A	**noise × mute**	$v70/a50$	yes	canonical	symmetric heavy-on-both corner
V+A	**noise × remove**	$v70/a50$	yes	replacement (audio)	swap audio operator family
V+A	**occlusion × mute**	$v30/a50$	yes	replacement (vision)	swap vision operator family
V+A	**noise × mute**	$v50/a30$	**no**	extended	severity-50 diagonal
V+A	**noise × mute**	$v10/a50$	**no**	extended (light V, heavy A)	isolates audio fault line under mild vision noise
V+A	**noise × mute**	$v70/a10$	**no**	extended (heavy V, light A)	mirror of the previous row
Text+Vision+Audio (11 canonical/replacement + 4 extended = 15 cells)					
TVA	**drop_words × noise × mute**	$t30/v30/a30$	yes	canonical	light-on-all corner
TVA	**drop_words × noise × mute**	$t30/v30/a50$	yes	canonical	audio-heavier corner
TVA	**drop_words × noise × mute**	$t30/v70/a30$	yes	canonical	vision-heavy corner
TVA	**drop_words × noise × mute**	$t30/v70/a50$	yes	canonical	vision and audio heavy
TVA	**drop_words × noise × mute**	$t70/v30/a30$	yes	canonical	text-heavy corner
TVA	**drop_words × noise × mute**	$t70/v30/a50$	yes	canonical	text and audio heavy
TVA	**drop_words × noise × mute**	$t70/v70/a30$	yes	canonical	text and vision heavy
TVA	**drop_words × noise × mute**	$t70/v70/a50$	yes	canonical	heavy-on-all corner
TVA	**drop_words × noise × remove**	$t70/v70/a50$	yes	replacement (audio)	audio operator swap
TVA	**drop_words × occlusion × mute**	$t70/v30/a50$	yes	replacement (vision)	vision operator swap
TVA	**word_shuffle × noise × mute**	$t70/v70/a50$	yes	replacement (text)	text operator swap
TVA	**drop_words × noise × mute**	$t50/v50/a30$	**no**	extended mid-pressure	all-severity-50 cube interior
TVA	**drop_words × noise × mute**	$t50/v50/a50$	**no**	extended mid-pressure	symmetric mid-grid TVA corner
TVA	**drop_words × noise × mute**	$t10/v70/a30$	**no**	extended (light T)	light-text TVA fault line
TVA	**drop_words × noise × mute**	$t10/v30/a50$	**no**	extended (light T, heavy A)	combines the two strongest non-text fault lines
TVA	**drop_words × noise × mute**	$t30/v50/a30$	**no**	extended (vision mid-pressure)	connects canonical T+V+A to mid-pressure vision
TVA	**drop_words × noise × mute**	$t70/v50/a50$	**no**	extended (heavy T, mid V, heavy A)	balances the heaviest non-vision fault lines

Table 28: Full enumeration of the 46-cell extended representative-model joint suite. **In 29-cell?** marks whether the cell is already part of the 29-cell completed canonical/replacement grid in Table 4 (**yes**, 29 cells) or whether it is part of the 17 additional extended cells (**no**, 17 cells: 4 T+V, 4 T+A, 3 V+A, 6 T+V+A). The extended rows fill in the severity-50 mid-pressure row and the light/heavy asymmetric corners of each combination's severity grid, so the 46-cell suite covers (i) every canonical $\{30, 70\}\times\{30, 70\}$ grid corner, (ii) every operator-replacement cell at the strongest severity, (iii) every severity-50 mid-pressure cell, and (iv) every $\{10, 70\}\times\{30/50/70\}$ light/heavy asymmetric cell. Reading: $29 + 17 = 46$.

G.6 F. Prompt robustness

The headline protocol uses the standard JSON answer-only prompt in Appendix H. Prompt robustness evaluates whether the fault-line pattern survives alternative prompting styles. We use three variants: the standard JSON prompt, a chain-of-thought prompt that asks the model to reason before selecting an option, and an open-form prompt that does not require JSON. The diagnostic subset contains the clean condition, the worst single-modality text/vision/audio cells, the strongest text–vision condition, and the strongest trimodal condition. We report whether CoT or open-form prompting changes accuracy, invalid-output rate, or the rank order of damaging conditions.

G.7 G. Human reference and answer-preservation checks

The human reference distinguishes model failure from task impossibility. We sample 30–50 base examples for human answering under clean

input, the worst single-modality cell for each modality, the strongest text–vision cell, and the strongest trimodal cell. Each example is answered by 2–3 annotators with the same answer options. Separately, the answer-preservation sanity check samples 50–100 corrupted examples and asks annotators whether the corrupted input remains answerable and whether the original gold answer remains valid. These results provide a ceiling and a validity filter: if humans also fail, the condition is marked as a stress test rather than a fair robustness condition.

G.8 H. Interface, coverage, and invalid-output profiling

Because omni-modal systems expose heterogeneous interfaces, every new model first runs a 10-example smoke test. The smoke test records whether the endpoint accepts the ordered frame set, whether audio is actually consumed, whether the output parser recovers an option key, and whether the model produces refusals or empty responses. For full runs, invalid responses are decomposed into API failures, parse failures, refusals, and empty outputs. Accuracy tables are paired with coverage tables whenever coverage differs across conditions, so the reader can separate reasoning failure from missing valid outputs.

G.9 I. Cross-modal mismatch and trust-bias probes

Mismatch probes keep every modality structurally clean but make one channel semantically inconsistent with the others. We use three sampling tiers: random mismatch, category-matched mismatch, and answer-matched mismatch. Random mismatch tests coarse conflict detection; category-matched mismatch keeps surface distribution similar; answer-matched mismatch controls for answer-label leakage. For each replacement modality, we record whether the model follows the replaced channel, the unchanged channels, or an unsupported prior. The trust-bias score T_m is the fraction of mismatched cases in which the answer follows modality m.

G.10 J. Temporal, audio–visual, frame-budget, and position probes

Temporal probes test whether the model uses event order rather than treating media as unordered feature bags. Visual temporal probes shuffle frames, reverse frame order, and drop middle frames. Audio temporal probes shuffle audio segments or insert local gaps. Audio–visual desynchronization shifts the audio stream by ± 0.5, ± 1, ± 2, ± 4, and ± 8 seconds relative to the visual frame sequence. Frame-budget probes rerun representative examples with one, four, eight, sixteen, and default frame budgets. Position probes place the most informative frame first, middle, or last while keeping the frame set fixed. Together these tests separate true temporal binding from interface-level sampling and position artifacts.

Scope of audio operators and ASR-like noise. The three audio operators (`remove`, `mute`, `distortion`) target structural properties of the waveform—temporal completeness and signal fidelity—rather than phonetic or lexical content. Temporal desynchronisation is covered by the audio–visual desynchronisation probe described above (± 0.5–8 seconds). ASR-like noise (homophone substitutions, partial word deletions at the phoneme level) was considered but excluded under design rule (i) of Appendix E: such operators require a phoneme-level transcription of the audio track, which is not available for all source benchmarks, and would introduce a semantic transformation (changing the spoken word) rather than a purely structural one. The `distortion` operator approximates the effect of low-quality recording or transmission artefacts, which is the closest structural analogue to ASR-degraded input without requiring phoneme-level manipulation. Probing audio-text-vision binding under phoneme-level noise is a natural extension and is noted as future work.

G.11 K. Confidence, abstention, and distractor-modality probes

Confidence probes ask representative models to output both an option and a scalar confidence level under clean and corrupted conditions. We report average confidence on correct

and incorrect answers, expected calibration error when possible, and the rate of overconfident wrong answers. Abstentions and refusals are analyzed together with the invalid-output profile. Distractor-modality probes insert an unrelated but structurally clean modality into otherwise sufficient inputs: unrelated audio into text–visual examples, unrelated frames into text–audio examples, or misleading text into visual–audio examples. The distractor effect measures whether the model indiscriminately fuses every available channel or selectively downweights irrelevant evidence.

Calibration under corruption and logit unavailability. The confidence probe collects self-reported confidence labels (`low`/`medium`/`high`) from the model's JSON output, which are available for all models regardless of whether token-level logits are exposed. We compute expected calibration error (ECE) by binning the three confidence levels and comparing mean confidence to mean accuracy within each bin, and we report the overconfident error rate ρ_{oc} (fraction of incorrect answers assigned `high` confidence). For the six deep-dive models, ECE increases from a clean-condition mean of 0.09 to 0.17 at severity 70 across the three modalities, and ρ_{oc} rises from 0.21 to 0.34, confirming that corruption degrades calibration as well as accuracy. For API models that do not expose logits, the self-reported confidence label is the only available proxy; we note that self-reported confidence may be less reliable than logit-derived probabilities and treat the ECE and ρ_{oc} estimates for those models as approximate. Generative likelihood proxies (e.g. scoring each option by its generation probability under a forced-choice prompt) are not uniformly available across the API panel and are left for future work.

H Prompt Templates and Parsing Rules

System prompt.

> You are a careful multimodal assistant. You will receive one question, one visual input (image or video), and one audio input. Exactly one option is correct. Use the provided multimodal evidence. If some evidence is degraded, still answer based only on what can be supported by the input. Respond with one JSON object and nothing else.

User prompt template.

> `[Task]` Answer the multiple-choice question using the visual input (image or video) and the audio input. Some of the question text, visual input, or audio input may be structurally corrupted or degraded. Interpret the available evidence carefully, and do not rely on any hidden metadata.
> `[Visual input]` Attached as a `{image/video}`.
> `[Audio input]` Attached as an audio clip.
> `[Question] {question}`
> `[Options] {formatted options}`
> Output JSON schema: `{"answer": "<option key>", "confidence": "<low | medium | high>", "reasoning": "<one or two short sentences>"}`. Do not output Markdown. Do not output any text outside the JSON object.

Parsing rule. The parser first attempts to read a JSON object. If this fails, it applies a conservative fallback that accepts a single option key or an answer string exactly matching one option. Responses are marked invalid when no option key can be recovered, when the model refuses to answer without selecting an option, or when the output is empty. Invalid responses are counted separately in coverage tables and treated as incorrect for significance testing.

Prompt variants. The standard prompt is used for all headline accuracy tables. Chain-of-thought prompting is used only in the prompt-robustness diagnostic subset; the model may reason internally or output a short rationale, but the parser still extracts a final option. Open-form prompting removes the JSON requirement and is used to test whether JSON formatting itself creates parse failures. Prompt variants are never mixed inside a headline table; each table specifies a prompt ID.

Option formatting and corruption scope. The formatted options string passed to the model uses plain letter keys (`A`, `B`, `C`, `D`) with no brackets, checkmarks, or positional markers. The `[x]` notation appearing in illustrative examples in the paper is used only in figures and tables for human readability and is never present in the actual model prompt. Option text is drawn verbatim from the source

benchmark and is not modified by any corruption operator: all four text corruption operators (`drop_words`, `word_shuffle`, `typo_ocr`, `sentence_break`) are applied exclusively to the question stem, leaving option punctuation, capitalisation, and letter keys intact. This design ensures that the parser can always recover a valid option key from the options field regardless of question-stem corruption severity, and that no asymmetric surface cue in the options field can serve as a positional shortcut.

Inference and logging settings. For deterministic comparability, runs use the lowest available sampling temperature, no intentional top-p diversification, and a fixed prompt template. When an endpoint fails, the runner records retry count and final failure category rather than silently dropping the example. Each JSONL record stores model name, prompt ID, sample ID, source dataset, condition ID, modality operators, severity values, random variant, media URLs or local paths, raw response, parsed answer, gold answer, correctness, latency, retry count, and invalid-output category. These fields are required to reproduce coverage-aware tables.