COMPETENCE IS NOT CONTROL: CAUSAL DECOMPOSITION IN MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sentiment analysis combines video, audio, and text to infer sentiment. Prior work has reported uneven use of these modalities, with models favoring text over visual and acoustic cues. We examine whether correct video-only predictions are preserved when audio and text are added, and how joint input changes video's influence. We focus on cases with correct video-only predictions but incorrect audio-only and text-only predictions. Across three models and three datasets, joint input yields error rates of 51.4%–87.6% on these selected training examples and 83%–97% on neutral examples. Video replacement and gradient analyses show reduced visual sensitivity under joint input, although video still improves accuracy over audio and text alone. On Qwen2.5-Omni, transplanting video-only states read by attention at the answer prompt into runs with all three inputs shifts scores toward the correct label, even with attention weights fixed. Allowing attention to adjust strengthens this effect, with less attention to the answer prompt and more to video, audio, and text on average. With the receiving run fixed, changing the source video has a larger effect for states formed from video alone than under joint input. Matching update size substantially reduces but preserves this difference. For neutral examples, it persists under two update-size controls across two models and two new datasets, for both correct and incorrect joint predictions. On Qwen2.5-Omni and MiniCPM-o 4.5, states from other videos with the same annotated sentiment also shift scores toward the correct label on average. Class-averaged states produce class-specific effects even after matching update size. Together, these findings distinguish a modality's standalone correctness from its influence on joint predictions.

1 INTRODUCTION

Multimodal sentiment analysis aims to identify the sentiment expressed in human communication by drawing on information from multiple modalities (Bagher Zadeh et al., 2018; Poria et al., 2019; Yu et al., 2020). Recent evaluations show that multimodal models can achieve strong sentiment-classification performance after task-specific fine-tuning (Luo et al., 2025b). However, modality use is uneven: general evaluations find both audio and visual underused relative to text (Liu et al., 2022; Luo et al., 2025a), and conflict evaluations on CA-MER report over-reliance on audio relative to video (Han et al., 2025), while EmoMM reports reduced video contributions under conflict and missingness (Sun et al., 2026). We therefore explore how video influences sentiment predictions.

We ask whether a model retains a correct prediction from video alone (*V-only*) when audio and text are added (*Full*). We select examples where V-only matches the dataset gold label but audio-only and text-only predictions are both incorrect. Unlike the annotation-defined conflicts in CA-MER and EmoMM, this model-based selection establishes a correct prediction in the same model as a starting point for investigating internal mechanisms. We evaluate three models on three datasets. In all nine settings, Full misclassifies more than half of the selected examples. As shown in Figure 1, the most consistent finding concerns neutral examples: 83%–97% become incorrect under Full, while positive and negative results vary more.

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

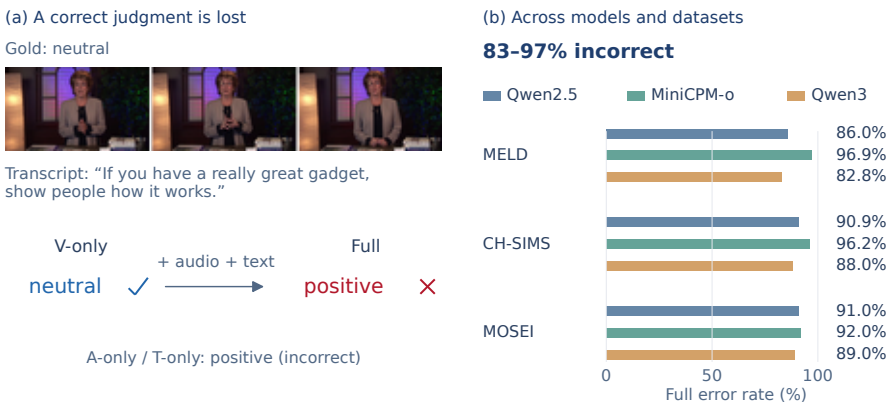


Figure 1: Adding audio and text often overturns correct video-only neutral predictions when audio-only and text-only predictions are wrong. (a) A Qwen2.5-Omni example from CMU-MOSEI. (b) Full error rates on repeated-run stable cohorts (Table 11).

We next test whether adding audio and text makes the model less sensitive to video. Replacing positive videos with negative ones changes sentiment scores less under Full than under V-only. Gradient-based sensitivity analysis (Baehrens et al., 2010; Simonyan et al., 2014) likewise shows that sentiment scores are less sensitive to visual-token states under Full than under V-only.

We next ask whether V-only states can still influence the joint prediction. Using activation patching (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024), we transplant them into Full at the classification suffix, the fixed answer-prompt tokens shared by both conditions. Scores shift toward the correct label, even with attention held fixed. Letting attention adapt increases the gain further, with less layer-averaged attention on the suffix and more on video and audio. Changing the source video matters more when states form under V-only than under Full, in the same receiving context. Magnitude matching substantially reduces but preserves this advantage in every label. For neutral clips, it also persists under both magnitude controls on a second model and new datasets, in both successful and failed predictions.

We next ask whether these intervention effects require states from the video itself. On both Qwen2.5-Omni and MiniCPM-o 4.5, V-only states from unrelated videos with the same gold label increase the gold margin on average. Class prototypes constructed using gold labels also produce label-selective effects in both models. These results identify a transferable, class-related component of the intervention effects.

Our contributions are as follows:

- We show that **adding audio and text often overturns correct video-only predictions**. Across three models and three datasets, this occurs in 83%–97% of neutral examples where video-only predictions are correct but audio-only and text-only predictions are both wrong.
- We find **reduced visual influence under joint input**: both video replacement and gradient analyses show weaker responses to changes in video content and internal visual states under Full than under V-only.
- We show that **transplanting V-only states into Full increases the gold margin even when attention weights are held fixed**. Allowing attention to adapt increases this gain further. Averaged across layers, heads, and examples, attention decreases on the classification suffix and increases on video, audio, and text.
- We find that **state formation context changes video-dependent output influence**. V-only sources produce larger effects than Full sources in matched receiving contexts; magnitude matching substantially reduces but preserves the advantage, and for neutral clips it persists under both controls on new data, in both successful and failed joint predictions.
- We identify **transferable, class-related influence across videos**: same-label donor states and gold-label prototypes produce label-directed effects.

2 PRELIMINARIES

Multimodal Sentiment Analysis. We study sentiment classification from video, audio, and transcribed text (Tsai et al., 2019; Liu et al., 2022), with labels $\mathcal{Y} = \{\text{negative}, \text{neutral}, \text{positive}\}$. V-only, A-only, and T-only denote video-only, audio-only, and text-only inputs (V, A, and T in formulas and in compounds such as V-correct); Full combines all three, and A+T combines audio and text. Each generated label is evaluated against the same dataset gold label y. We call correctness under a specified input condition *competence*.

Attention and Value States. For one Transformer attention head (Vaswani et al., 2017), input representations H yield Queries $Q = HW_Q$, Keys $K = HW_K$, and Values $U = HW_U$ through learned projections. Values are vectors within an attention head, distinct from the residual stream passed between Transformer blocks. We write U for Values to avoid confusion with video V. The causal head output at position i is

$$\alpha_{ij} = \mathrm{softmax}_{j \le i}\left(q_i k_j^\top / \sqrt{d_k}\right), \qquad o_i = \sum_{j \le i} \alpha_{ij} u_j, \tag{1}$$

where q_i, k_j, u_j are rows of Q, K, U, d_k is the Query/Key dimension, and o_i precedes the output projection; layer and head indices and positional transformations are omitted. Queries and Keys thus determine how strongly each accessible position is read, while Values provide its content (Elhage et al., 2021). Because each layer aggregates over preceding positions, identical prompt tokens can carry different Values when the preceding modality inputs differ. Replacing Values while retaining the current attention weights changes the content read at that layer.

Activation Patching and Output Influence. Activation patching replaces internal activations to measure their output influence (Vig et al., 2020; Meng et al., 2022; Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). A source run supplies recorded activations a_s to replace aligned activations a_r in a receiving run, or host. The host example is the recipient; other examples supplying videos or source states are donors. With host input X_r and parameters θ fixed,

$$z_r = f_\theta(X_r), \qquad \widetilde{z}_r = f_\theta(X_r;\, a_r \leftarrow a_s), \tag{2}$$

where f_θ returns first-answer logits (unnormalized scores) for the label tokens; downstream activations can change after replacement. Comparing patched and unmodified host outputs isolates the effect of this replacement. We measure movement toward y using the gold margin and its change:

$$g(z, y) = z_y - \log \sum_{c \in \mathcal{Y} \setminus \{y\}} \exp z_c, \qquad \Delta g = g(\widetilde{z}_r, y) - g(z_r, y). \tag{3}$$

Here z_c is the logit for label c; g is the log-odds of y under a softmax over $\mathcal{Y}$. Positive Δg favors the gold label relative to its alternatives; answer corrections are counted separately. We call this output influence under intervention *control*. Competence concerns correctness under a specified input condition; control concerns how transplanted states influence host predictions. In our transplant experiments, V-only and Full *states* are the Values recorded at the classification suffix under that input (Section 4.2); gradient analyses instead use residual-stream states at every token position.

3 SENTIMENT PREDICTIONS ACROSS INPUT CONDITIONS

We test whether sentiment predictions that are correct under video-only input remain correct when audio and text are added. We evaluate Qwen2.5-Omni-7B (Xu et al., 2025a), MiniCPM-o 4.5 (OpenBMB, 2026), and Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b) on MELD (Poria et al., 2019), CH-SIMS (Yu et al., 2020), and CMU-MOSEI (Bagher Zadeh et al., 2018). All models use frozen weights and greedy decoding. We use three sentiment labels: negative, neutral, and positive. Dataset-specific label mappings, inference settings, and prompts appear in Appendices B–D.

For each model and dataset, we screen the official training split for examples with correct V-only and incorrect A-only and T-only predictions, all scored against the same gold label. For each gold label, we stop after collecting 100 qualifying examples or exhausting the candidate pool. We then evaluate Full on the selected examples and report the fraction it misclassifies (Full error rate). Selection and aggregation details appear in Appendix B.3.

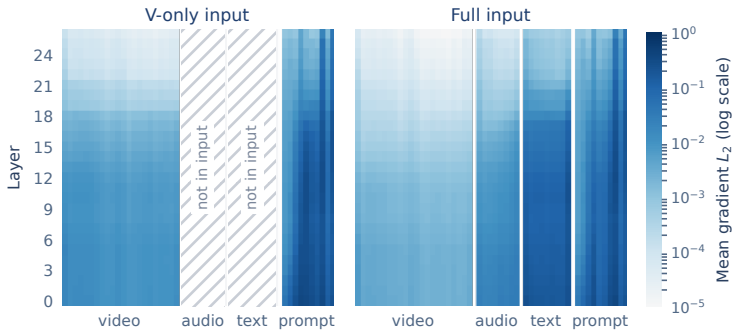


Figure 2: Visual sensitivity decreases under Full input. Mean gradient norms by layer and token type for Qwen2.5-Omni on 141 CMU-MOSEI clips with correct V-only predictions. Hatched regions indicate absent inputs. Appendix F.1 provides details of aggregation and layer selection.

V-only competence often does not carry over to Full. Across models, we select 1,830 cases from 1,159 distinct clips. Pooled Full error rates are 51.4%–87.6% overall and 83.0%–97.0% for neutral examples. Appendix E.1 reports intervals, clip counts, and rates averaged equally over labels.

Adding video input still helps. On the same selected examples, pooled accuracy is 3.7%–9.7% under A+T and 12.4%–48.6% under Full, and Full is higher in all nine settings (Appendix E.2). When A-only and T-only give the same wrong label, a majority vote is always wrong, yet Full is correct in 405 of these 1,585 cases. In 1,150 of the remaining 1,180, Full returns that shared wrong label (Appendix E.2). Video thus contributes, but when audio and text agree, Full usually follows them.

The neutral result persists across repeated runs. Among examples whose answers remain unchanged across three runs under all four input conditions, Full error rates range from 82.8% to 96.9% (Figure 1b). Changing the option order or rewording the instructions also leaves high error rates. On Qwen2.5-Omni, we evaluate the original prompt and both variants on 2,608 MELD test clips, selecting qualifying examples separately for each prompt. Across the three prompts, Full error rates range from 79.8% to 88.4% for neutral examples (163 to 168 selected clips per prompt), 89.3% to 94.7% for negative examples, and 69.0% to 80.0% for positive examples (Appendix E.3).

To estimate how often the selection condition occurs across models and datasets, we use fixed test subsets of 400 clips per dataset, chosen before inference. On these subsets, 3%–13% of clips have correct V-only and incorrect A-only and T-only predictions, depending on the model and dataset (Appendix E.3). On the same subsets, V-only predicts the neutral label for 58.0%–93.8% of all clips and for 54.4%–92.9% of clips labeled negative or positive (Appendix F.3). Given this bias, Section 4.1 tests whether correct neutral predictions depend on the video.

4 Context-Dependent Sensitivity and State Influence

We first measure how visual sensitivity changes between V-only and Full, then transplant V-only states into Full at a shared suffix interface and compare them with states formed under Full.

4.1 Full inputs reduce visual sensitivity

We ask whether adding audio and text makes predictions less sensitive to video. For 141 CMU-MOSEI clips with correct Qwen2.5-Omni V-only predictions, we compare V-only and Full using the gradient of the difference between each pair of label logits with respect to every token's residual state at each layer (Baehrens et al., 2010; Simonyan et al., 2014). Its L_2 norm measures local sensitivity to changes in that state. Figure 2 shows lower visual sensitivity under Full: the mean gradient norms at video positions fall to 29%, 23%, and 14% of their V-only levels at layers 0, 9, and 18 (Appendix F.1). The index averaged over visual tokens and layers is also lower under Full in every clip for every label pair (Appendix F.1). Under Full, mean gradient norms over layers 0

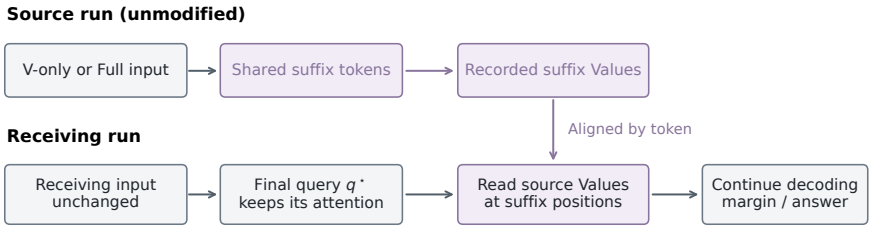


Figure 3: The receiving run's final query $q^\star$ reads source suffix Values at aligned suffix positions. At each intervened layer, its attention weights and non-suffix contributions stay unchanged; $q^\star$ itself is not transferred.

through 18 are higher for audio than video tokens in nearly every clip, and higher for transcript tokens in every clip (Appendix F.1).

Replacing the video itself shows the same contrast. For CMU-MOSEI test clips selected as in Section 3, we swap in videos from other clips, keeping the original audio and transcript under Full. A video from a clip with a different label lowers the V-only gold margin by more than three units relative to one with the same label, but moves the Full margin by less than half a unit (Appendix F.2). For neutral clips, these replacements also test whether V-only predictions depend on the video despite the model's preference for the neutral label. V-only answers neutral on 97.4% of replacements with neutral videos but only 9.8% and 11.8% with negative and positive ones; Full answers neutral on 4.6%–5.2% in all three cases. On a separate set of clips, V-only answers neutral whenever the frames are set to gray or their mean color, shuffled, or blurred (Appendix F.3). The replacement results therefore show that V-only neutral predictions depend on the video, while the tested frame disruptions consistently elicit neutral answers.

We next ask whether V-only states can still move the Full prediction when transplanted, and whether states formed under V-only and Full differ in this influence.

4.2 Where and how we intervene

To compare states formed under V-only and Full, we intervene at suffix positions whose tokens match across the two inputs. Every prompt ends with an answer-choice block that lists positive, negative, and neutral and asks for only one answer choice. This block and the following chat boundary tokens form the *classification suffix*. Its Values can differ across the two input conditions. Figure 3 shows Value transfer at these aligned positions. The Qwen2.5-Omni suffix contains 19 tokens, excluding the final query $q^\star$ that predicts the first answer token. A 12-clip development study supported the suffix interface: suffix transplants increased gold margins, whereas video-token transplants had little effect. The all-layer scope was fixed in an earlier pilot. Appendix G.1 reports the position and layer comparison and exact token boundaries.

For one head and layer, source s supplies Values u_s to receiving run r. Let $\mathcal{S}_r$ denote suffix positions, π their source alignment, and $\alpha_{r,q^\star j}$ the final query's current attention weights. We update the receiving head output $o_{r,q^\star}$ before output projection:

$$\widetilde{o}_{r,q^\star} = o_{r,q^\star} + \sum_{j \in \mathcal{S}_r} \alpha_{r,q^\star j} \big(u_{s,\pi(j)} - u_{r,j}\big). \tag{4}$$

This substitutes source Values in the final-query read without directly changing current-layer Queries, Keys, or attention weights. Other query positions remain unchanged; later final-query states and attention can respond to earlier interventions. Source Values are recorded from an unmodified run. We measure gold-margin changes as defined in Equation 3 and report generated-answer corrections separately.

4.3 Transplanting V-only states improves incorrect Full predictions

States formed under V-only can correct errors made under Full. We test this on CMU-MOSEI and MELD clips for which Qwen2.5-Omni-7B predicts correctly under V-only but incorrectly under A-

Table 1: Own V-only suffix Values transplanted into Full on the 156 failures. Before and After are mean gold margins with clips weighted equally; Corrected counts generated answers that become correct. Table 23 splits results by dataset and split.

	Gold margin ↑: higher favors the correct label			
Gold label (n)	Before	After	Δg	Corrected
Negative (35)	-3.373	-0.632	$+2.741$	6/35
Neutral (60)	-2.197	0.458	$+2.655$	47/60
Positive (61)	-2.340	-0.078	$+2.262$	35/61

only, T-only, and Full. Keeping Full inputs fixed, we transplant each clip's own V-only suffix Values across all heads and layers. Gold margins increase in every clip, and 88/156 initially incorrect answers become correct (Table 1; Appendix H).

On average, the transplant reduces the final query's attention to the suffix and increases it to video, audio, and text. To test whether the margin gain requires this redistribution, we repeat the transplant while fixing the final query's attention weights at every layer to those of the unmodified Full run. Across sentiment labels, this retains 69.8% to 77.5% of the mean margin gain obtained with adaptive attention. The gain remains positive in every clip, and allowing attention to adapt increases it in 155/156 clips. The supplied Values therefore improve the gold margin under the original attention weights, while attention adaptation amplifies this effect (Appendix H.2).

4.4 Comparing video-content effects across source contexts

We next test whether the effect of changing the source video depends on the inputs used to form its states. On the same clips as Section 4.3, we compare each clip's own video with donor videos from the same dataset with different annotated sentiment labels. We form source states under V-only or Full input and transplant them into the same Full receiving run. For Full sources, only the video changes; the recipient's audio and transcript remain fixed. Within each source condition, we measure the gold-margin difference between transplants from the own video and from donor videos, averaged over donors (Appendix G).

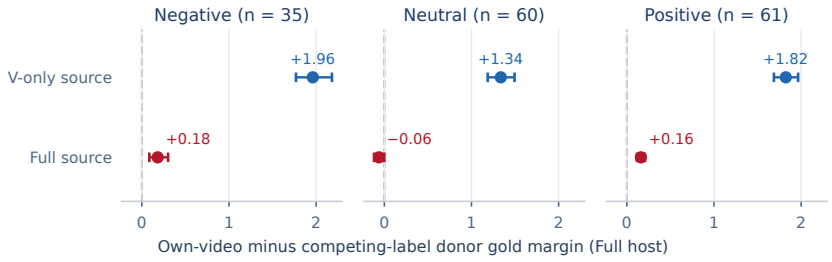


Figure 4: Own video minus mean donor gold margin in a fixed Full host, using V-only (blue) or Full (red) source states. Donor labels differ from the recipient's. Bars show descriptive 95% bootstrap intervals; Appendix G gives the donor selection and resampling details.

Changing the source video has a larger mean effect when states are formed under V-only than under Full (Figure 4). Across the three labels, the mean own-versus-donor contrast ranges from $+1.34$ to $+1.96$ for V-only sources and from -0.06 to $+0.18$ for Full sources. The V-only contrast exceeds the Full contrast in 155/156 clips. This contrast measures control: how the output responds to the source video under intervention, not how much video information the states contain. A remaining question is whether the larger contrast comes simply from larger updates. Section 5 tests this with magnitude controls and checks replication on new data.

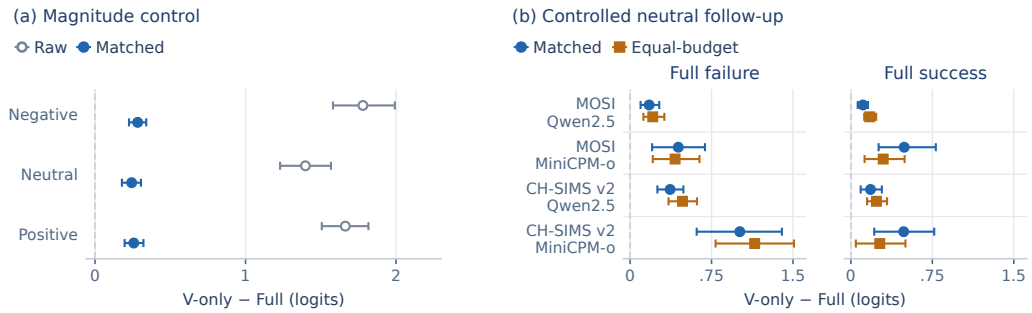


Figure 5: Source contrasts remain positive under magnitude controls. (a) Raw and matched G_F on Qwen2.5-Omni failures. (b) Neutral groups on new data, by native Full outcome: matched G_F and contrasts between directions applied from the same Full state with equal norm budgets. Positive values favor V-only sources. Bars show descriptive 95% source-cluster intervals (Appendices I and J.1).

5 Controls and Replication of Source Effects

Update magnitude. We test whether the larger source contrast under V-only is explained by update magnitude. Let $E(s, h)$ denote the own-video minus mean donor gold margin when states formed under source condition $s \in \{V, F\}$ (V-only or Full) are transplanted into host h. The source contrast in a Full host is $G_F = E(V, F) - E(F, F)$.

Using the same recipients and donors as Section 4.4, we first match the update magnitude at each layer while preserving each own-video transplant. This reduces G_F to about one sixth of its raw value, but the mean remains positive in every label (Figure 5a). A second control compares the two source directions from the same native Full state, using the same norm budget at each layer. The direction derived from V-only states still has the larger mean gold-margin effect in every label (Appendix I).

New data and Full successes. The source contrast remains positive for neutral examples on new data, including both correct and incorrect Full predictions. We evaluate Qwen2.5-Omni and MiniCPM-o on CMU-MOSI (Zadeh et al., 2016) and a CH-SIMS v2 subset excluding all original CH-SIMS sources (Liu et al., 2022). Selection requires V-only correctness, without restrictions on A-only or T-only answers, and examples are grouped by their original Full outcome. Raw G_F is positive in all eight neutral combinations of model, dataset, and Full outcome (Appendix J). A follow-up using the same recipients and donors finds positive source differences under both magnitude controls in all eight groups, with all descriptive 95% source-cluster intervals above zero (Figure 5b). Other sentiment labels have smaller groups and less consistent results (Appendix J.1).

Failure prediction. We also ask whether the source contrast helps predict which clips Full misclassifies, beyond unimodal confidence and input length. The prespecified score uses a V-only host, $G_V = E(V, V) - E(F, V)$, so that neither native Full logits nor Full correctness directly enters it. The analysis did not establish predictive support under the prespecified criteria in any of the four combinations of model and dataset. Three combinations had insufficient sample sizes (Appendix J).

6 Class-Directed Control and Cross-Video Transfer

We next ask whether improving Full predictions requires states from the recipient's own video, or whether states from other videos with the same sentiment can also help. We test V-only suffix Values on the Qwen2.5-Omni failures from Section 4.3, using individual donors and class prototypes. These are oracle diagnostics: donors and prototypes are defined using gold labels.

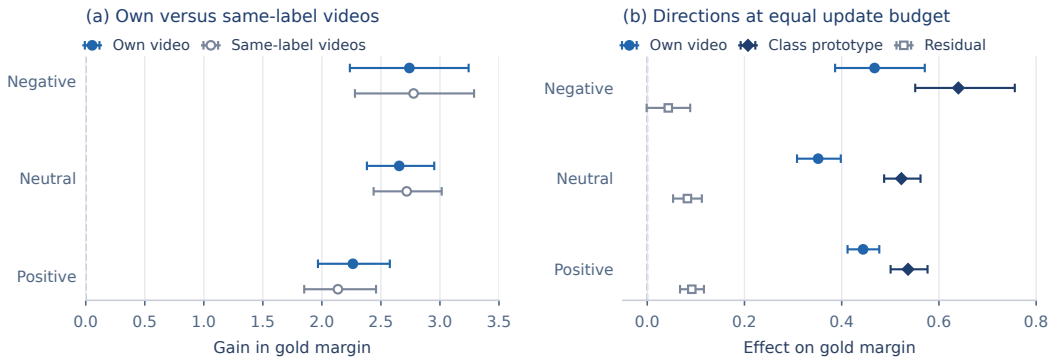


Figure 6: Cross-video transfer on the 156 Qwen2.5-Omni failures. (a) Gold-margin gains from the own V-only video and from same-label videos. (b) Effects at equal update budget from native Full: the own-video direction, the class-prototype direction, and the own-video residual after projecting out the two prototype contrasts. Bars are descriptive 95% paired intervals (Appendices I and K.1).

6.1 Transfer across videos and class selectivity

States from other videos with the recipient's gold label raise the mean gold margin in all three labels (Figure 6a) and correct some Full errors in each label (Appendix K.1). Beneficial interventions therefore do not require states from the recipient's own video. Mean gains differ from the own-video gains by at most 0.13, and all three own-minus-same-label intervals include zero.

We then test whether averaging states within a sentiment class preserves a selective influence toward that class. We form class prototypes by averaging V-only suffix Values from training videos with the same label. Selectivity measures how much more an injected prototype favors its own label than the other prototypes do, averaged over the three labels. It is positive for every recipient at all tested update magnitudes in both prototype banks, including when update norms are matched to the own source (Appendix K.2).

6.2 Class prototypes have stronger effects at equal update magnitude

To compare individual videos and class prototypes at equal update magnitude, we construct directions contrasting the own video with donors of competing labels, or the gold prototype with competing prototypes. Starting from the same native Full state and matching update norms at each layer, the prototype direction has a larger mean effect on the gold margin than the own-video direction in all three labels (Figure 6b).

We then test how much intervention effect remains after removing the components along the two prototype contrasts. We project the own-video direction off their span and renormalize the residual to the same update budget. The residual produces 9%–24% of the own-video direction's effect. Its intervals are above zero for neutral and positive failures, while the interval for negative failures includes zero; results also vary by dataset (Appendix I). These comparisons support a shared class-related component of the intervention effect. Because the residual is renormalized, these ratios do not give an additive decomposition of the original effect.

6.3 Transfer and class selectivity extend to a second model

Transfer across videos and class selectivity also appear in MiniCPM-o 4.5 on Full failures from CMU-MOSEI and MELD. Both own-video and same-label V-only states raise the mean gold margin in all three labels, and prototype selectivity remains positive in every label after matching update norms. The own-minus-same-label interval excludes zero only for negative failures, favoring same-label sources. Appendix L reports the gains, answer corrections, and matched-source comparisons.

Together, these results show that beneficial interventions can draw on states from other videos with the same sentiment, and that averaging states within a class preserves a selective influence on predictions.

7 Related Work

Modality interactions and conflict. Multimodal sentiment research models complementary cues (Tsai et al., 2019) and analyzes how dominance, complementarity, and conflict affect predictions (Wang et al., 2021). Joint training can also underuse a modality: multimodal networks can underperform their best unimodal counterparts, and modalities compete during optimization (Wang et al., 2020; Peng et al., 2022; Huang et al., 2022). Studies report text dominance in sentiment models (Liu et al., 2022; Luo et al., 2025a); under annotation-defined emotion conflicts, models over-rely on audio relative to video (Han et al., 2025) or marginalize video evidence (Sun et al., 2026), while constructed image–audio–text conflicts can elicit visual preference (Yan et al., 2026). These findings make reliance on a modality a question about the model and task, rather than a fixed hierarchy. Modality Sabotage connects this issue to correctness, showing how confident errors from modality-specific agents mislead fused predictions (Zhang et al., 2025a). Unlike training-time accounts of modality competition, we study a frozen model at inference, when video alone yields a correct answer but audio and text individually do not.

Internal signals and output influence. Gradient analyses and layerwise probes characterize modality imbalance (Kwon et al., 2025; Yan et al., 2026). Causal tracing and attention knockout have localized visual information storage and transfer in multimodal models (Basu et al., 2024; Zhang et al., 2025b). Path patching further identifies attention heads that promote or resist hallucinations under visual–textual conflict (Jiang et al., 2026). Others separate what a model represents from what influences its output: visual attributes remain encoded during visual–linguistic conflicts (Nooralahzadeh et al., 2026), perception–premise mismatches are encoded but not rejected (Quang et al., 2026), and cue aggregation can be distinguished from emotional expression (Zhang et al., 2026). We compare states formed under video-only and joint input in matched receiving contexts, separating source-context effects from attention adaptation and testing their transfer across videos.

Interventions and shared sentiment representations. Linear sentiment directions and activation additions can influence outputs across examples (Tigges et al., 2023; Turner et al., 2023), raising the question of whether an intervention supplies example-specific information or a shared class signal. Patching effects also depend on the replacement operation and output metric (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024), and subspace interventions can act through dormant pathways (Makelov et al., 2024). We compare source contexts at matched intervention magnitudes and test same-label donors and gold-label class prototypes to characterize shared class effects and dependence on source context.

8 Conclusion

In this paper, we investigate how video contributes to sentiment predictions in multimodal models, using behavioral comparisons and activation patching to examine its influence under joint input. Across three models and three datasets, adding audio and text can overturn correct video-only predictions on the selected conflict examples. Joint input reduces visual sensitivity, although video still improves accuracy over audio and text alone. On Qwen2.5-Omni, transplanting video-only states into Full at shared answer-prompt positions shifts scores toward the correct label even with attention fixed; attention adaptation strengthens the effect. In a fixed receiving run, changing the source video has a larger mean effect for states formed under video-only than joint input. Matching update magnitude substantially reduces but preserves this difference. For neutral examples, both magnitude controls preserve it across two models and new datasets, for correct and incorrect Full predictions. In both patched models, states from other videos with the same annotated sentiment raise the mean gold margin, and class prototypes retain positive selectivity after matching update magnitude. Together, these findings distinguish correctness under a single modality from influence on joint predictions.

AI USE STATEMENT

Generative AI tools were used for drafting and language editing, and for writing and debugging code. All AI-assisted content was reviewed by the authors, who take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

Appendices B–D specify data selection, models, and prompts. Subsequent appendices give intervention definitions, donor assignments, statistical estimands, complete outcomes, and numerical controls. Table 2 summarizes study designs and sample reuse. Each experiment specifies its selection rules, design timing, and applicable numerical stability and sensitivity checks.

REFERENCES

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(61):1803–1831, 2010. URL https://www.jmlr.org/papers/v11/baehrens10a.html.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.nips.cc/paper_files/paper/2024/hash/0dfe31d6e703e138d46a7d2fced38b7c-Abstract-Conference.html.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Zhiyuan Han, Beier Zhu, Yanlong Xu, Peipei Song, and Xun Yang. Benchmarking and bridging emotion conflicts for multimodal emotion reasoning. *arXiv preprint arXiv:2508.01181*, 2025. doi: 10.48550/arXiv.2508.01181. URL https://arxiv.org/abs/2508.01181.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. doi: 10.48550/arXiv.2404.15255. URL https://arxiv.org/abs/2404.15255.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259, 2022. URL https://proceedings.mlr.press/v162/huang22e.html.

Jinrui Jiang, Zhangtai Wu, Zhen Wu, and Xinyu Dai. Causal evidence for attention head imbalance in modality conflict hallucination. *arXiv preprint arXiv:2605.19250*, 2026. doi: 10.48550/arXiv.2605.19250. URL https://arxiv.org/abs/2605.19250.

Junehyoung Kwon, MiHyeon Kim, Eunju Lee, Juhwan Choi, and Youngbin Kim. See-saw modality balance: See gradient, and sew impaired vision-language balance to mitigate dominant modality bias. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the*

Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 4364–4378. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.222. URL https://aclanthology.org/2025.naacl-long.222/.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 247–258. ACM, 2022. doi: 10.1145/3536221.3556630. URL https://doi.org/10.1145/3536221.3556630.

Miaosen Luo, Yuncheng Jiang, and Sijie Mai. Towards explainable fusion and balanced learning in multimodal sentiment analysis. *arXiv preprint arXiv:2504.12151*, 2025a. doi: 10.48550/arXiv.2504.12151. URL https://arxiv.org/abs/2504.12151.

Miaosen Luo, Jiesen Long, Zequn Li, Yunying Yang, Yuncheng Jiang, and Sijie Mai. Multimodal large language models for end-to-end affective computing: Benchmarking and boosting with generative knowledge prompting. *arXiv preprint arXiv:2508.02429*, 2025b. doi: 10.48550/arXiv.2508.02429. URL https://arxiv.org/abs/2508.02429.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2202.05262.

Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts. *arXiv preprint arXiv:2604.09364*, 2026. doi: 10.48550/arXiv.2604.09364. URL https://arxiv.org/abs/2604.09364.

OpenBMB. MiniCPM-o 4.5 model card. Hugging Face model release, 2026. URL https://huggingface.co/openbmb/MiniCPM-o-4_5/tree/073dbbc8c5bc0af2d789e1ce12e7c17a6be746e1. Revision 073dbbc8c5bc; accessed 21 September 2026.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2203.15332.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

Trung Nguyen Quang, Yiming Gao, Fanyi Pu, Kaichen Zhang, Shuo Sun, and Ziwei Liu. Senses wide shut: A representation-action gap in omnimodal LLMs. *arXiv preprint arXiv:2605.13737*, 2026. doi: 10.48550/arXiv.2605.13737. URL https://arxiv.org/abs/2605.13737.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2014. URL https://arxiv.org/abs/1312.6034v2.

Yueru Sun, Yimeng Zhang, Haoyu Gu, Nuo Chen, Dong She, Xianrong Yao, Yang Gao, and Zhanpeng Jin. EmoMM: Benchmarking and steering MLLM for multimodal emotion recognition under conflict and missingness. *arXiv preprint arXiv:2605.01024*, 2026. doi: 10.48550/arXiv.2605.01024. URL https://arxiv.org/abs/2605.01024.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023. doi: 10.48550/arXiv.2310.15154. URL https://arxiv.org/abs/2310.15154.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. doi: 10.48550/arXiv.2308.10248. URL https://arxiv.org/abs/2308.10248.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. URL https://arxiv.org/abs/1706.03762.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://arxiv.org/abs/1905.12681.

Xingbo Wang, Jianben He, Zhihua Jin, Muqiao Yang, Yong Wang, and Huamin Qu. M2Lens: Visualizing and explaining multimodal models for sentiment analysis. *arXiv preprint arXiv:2107.08264*, 2021. URL https://arxiv.org/abs/2107.08264.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*, 2025a. doi: 10.48550/arXiv.2503.20215. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765*, 2025b. doi: 10.48550/arXiv.2509.17765. URL https://arxiv.org/abs/2509.17765.

Xinru Yan, Boxi Cao, Yaojie Lu, Hongyu Lin, Weixiang Zhou, Le Sun, and Xianpei Han. Beyond text-dominance: Understanding modality preference of omni-modal large language models. *arXiv preprint arXiv:2604.16902*, 2026. URL https://arxiv.org/abs/2604.16902.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.343. URL https://aclanthology.org/2020.acl-main.343/.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016. doi: 10.48550/arXiv.1606.06259. URL https://arxiv.org/abs/1606.06259.

Chengsheng Zhang, Chenghao Sun, Zhining Xie, and Xinmei Tian. Interpreting and enhancing emotional circuits in large vision-language models via cross-modal information flow. *arXiv preprint arXiv:2605.21980*, 2026. doi: 10.48550/arXiv.2605.21980. URL https://arxiv.org/abs/2605.21980. Accepted at ICML 2026.

Chenyu Zhang, Minsol Kim, Shohreh Ghorbani, Jingyao Wu, Rosalind Picard, Patricia Maes, and Paul Pu Liang. When one modality sabotages the others: A diagnostic lens on multimodal reasoning. *arXiv preprint arXiv:2511.02794*, 2025a. doi: 10.48550/arXiv.2511.02794. URL https://arxiv.org/abs/2511.02794.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Cross-modal_Information_Flow_in_Multimodal_Large_Language_Models_CVPR_2025_paper.html.

A Overview of Supplementary Evidence

Appendices B–D specify the data, inference settings, and prompts. Table 2 summarizes their purposes, designs, and shared samples. Studies that reuse recipients provide additional tests on the same sample, not independent replications. Model comparisons also differ in recipients and donor pools, so their between-model effect differences do not isolate architecture.

Table 2: Evidence roles and reuse. Counts refer to recipients or model–example pairs, not numbers of donor interventions. Design details and timing are specified with each experiment.

Study	Recipients	Design and sample reuse	Evidential role
Behavioral selection (E)	1,830 model–example pairs; 1,159 clips; 3 models, 3 datasets	Fixed-order conditional selection before Full; model-specific train cohorts	Conditional retention and paired A+T comparison
Paired gradients (F.1)	43 Qwen2.5-Omni validation clips and 98 V-correct clips from the CMU-MOSEI test subset of 400 clips	Shared with behavioral analyses; examples and contrasts fixed before gradients	Local residual sensitivity under V and Full
Video replacement (F.2)	90 Qwen2.5-Omni test examples	Separately selected conditional cohort; fixed donor comparisons	Input-level video sensitivity under V and Full
Video-content controls (F.3)	100 Qwen2.5-Omni clips: 43 CMU-MOSEI validation clips, 57 CH-SIMS test clips	Shared with behavioral checks; the 43 also belong to the transplant sample; frame controls fixed before inference	V-only answers and neutral margins under frame disruption
Interface development (G.1)	12 Qwen2.5-Omni train clips	Separate development sample excluded from transplant recipients	Choice of aligned suffix interface
Qwen2.5-Omni transplant cohort (G.2)	156 failures: 66 CMU-MOSEI, 90 MELD; 3 validation success controls	Fixed recipients and same-dataset donors shared across paired intervention comparisons	Own-state correction, attention, source–host, magnitude, transfer, and prototype controls
New-data source–host studies (J)	360 model–example pairs; Qwen2.5-Omni and MiniCPM-o on CMU-MOSI and CH-SIMS v2	Fixed recipients and donors across source contexts and magnitude controls	Raw and magnitude-controlled source effects; no model–study cohort meets all prespecified failure-prediction criteria
MiniCPM-o replication (L.1)	152 failures: 34 negative, 60 neutral, 58 positive; 9 successes	Behavioral screening and native endpoint checks; assignments fixed before intervention	Second-model own-video, same-label, prototype, and matched-source tests across CMU-MOSEI and MELD
Matched Qwen2.5-Omni sources (L.2)	131/156 failures matched	Up to 24 same-label candidates per recipient; thresholds fixed before outcomes	Logit/norm-matched sensitivity analysis

B Datasets and Cohort Construction

We evaluate sentiment classification on MELD, CH-SIMS, and CMU-MOSEI, and use CMU-MOSI and CH-SIMS v2 for the replication in Appendix J. All datasets provide video, audio, and transcripts, but their original label spaces differ. We use the same three answer labels, `negative`, `neutral`, and `positive`, with the dataset-specific gold-label rules below. For each example, V-only, A-only, T-only, Full, and A+T predictions are evaluated against one common gold label.

B.1 Label definitions

MELD. We use MELD's released three-class `Sentiment` labels (Poria et al., 2019) and evaluate each utterance independently, without neighboring dialogue context.

CH-SIMS and CH-SIMS v2. We use the multimodal sentiment score M for all input conditions, rather than the separate unimodal annotations (Yu et al., 2020). We follow the authors' three-class mapping[1]:

$$y = \begin{cases} \texttt{negative}, & M \leq -0.1, \\ \texttt{neutral}, & -0.1 < M \leq 0.1, \\ \texttt{positive}, & M > 0.1. \end{cases} \tag{5}$$

CH-SIMS v2 uses the same thresholds on its multimodal score.

CMU-MOSEI and CMU-MOSI. We map the released sentiment score s to three classes (Bagher Zadeh et al., 2018):

$$y = \begin{cases} \texttt{negative}, & s < 0, \\ \texttt{neutral}, & s = 0, \\ \texttt{positive}, & s > 0. \end{cases} \tag{6}$$

Neutral requires an exact zero; scores are not rounded. CMU-MOSI numeric sentiment is mapped by the same sign rule. This three-class evaluation is not directly comparable to Acc-2 or Acc-7.

B.2 Releases and splits

We use the released train, validation, and test splits. Table 3 reports dataset sizes before filtering, not evaluated cohort sizes.

Table 3: Sample counts before filtering. CMU-MOSEI uses the MMSA release.

Dataset	Language	Train	Validation	Test	Total
MELD	English	9,989	1,109	2,610	13,708
CH-SIMS	Chinese	1,368	456	457	2,281
CMU-MOSEI	English	16,326	1,871	4,659	22,856

We use the official MELD release and the released clip-level splits of CH-SIMS, whose source videos overlap across splits (Yu et al., 2020). For CMU-MOSEI, we use the MMSA release with source-video assignments verified against the official SDK splits. We exclude four training segments with unmatched source-video IDs, leaving 16,322 matched training segments.

Metadata and media for the replication datasets come from fixed-revision public mirrors of CMU-MOSI and CH-SIMS v2. CMU-MOSI source splits match the official SDK. The 2,281 CH-SIMS v2 rows overlapping original CH-SIMS match its labels and transcripts; new rows and CMU-MOSI mirror labels were not compared directly with the authors' downloads.

B.3 Behavioral cohorts

The behavioral analyses use two populations. The conditional cohorts of Section 3 are selected on unimodal outcomes and capped by label; the fixed test subsets preserve class proportions and are selected independently of model predictions.

For each model, dataset, and gold label, a conditional cohort contains up to 100 training-split examples with valid V-correct, A-incorrect, and T-incorrect predictions. Models share a fixed candidate order but form their own cohorts, stopping at 100 eligible examples or pool exhaustion. Selection and exclusions are independent of Full predictions. Model weights remain frozen; development examples are eligible for this behavioral analysis. In addition to the four unmatched CMU-MOSEI segments, we exclude one MELD clip with unusable video and 61 CMU-MOSEI training segments with insufficient transcript agreement: CMU-MOSEI clips require normalized word-sequence similarity of at least 0.5 and a duration discrepancy of at most 0.3 seconds relative to the matched SDK interval. We report sample sizes and Full error rates with 95% percentile intervals from 2,000 bootstrap draws, resampling dialogues for MELD and source videos for CH-SIMS and CMU-MOSEI.

[1]The SIMS three-class mapping in the versioned MMSA evaluation code.

Cohorts with fewer than 30 examples or ten source clusters are marked. Pooled error rates divide total errors by total selected examples across labels.

The fixed test subsets contain 400 test examples per dataset, drawn without replacement in a fixed within-class order that preserves class proportions (Table 4); selection and replacements are completed before inference. All three models share these examples across V-only, A-only, T-only, and Full conditions. CMU-MOSEI candidates must have usable audio and video and satisfy the transcript and duration criteria above. Six candidates fail the duration check and are replaced by the next eligible example in the same class, preserving class counts, so the CMU-MOSEI subset represents media-eligible test clips. Each condition uses one greedy prediction per example, without filtering for repeated-answer stability, and these test results are reported separately from the conditional train analysis.

Table 4: Class counts in each fixed test subset of 400 clips.

Dataset	Negative	Neutral	Positive
MELD	128	192	80
CH-SIMS	217	60	123
CMU-MOSEI	116	88	196

C Inference Settings

C.1 Checkpoints and operating modes

We use the released Qwen2.5-Omni-7B, MiniCPM-o 4.5, and Qwen3-Omni-30B-A3B-Instruct checkpoints at the linked fixed revisions, with frozen weights and speech output disabled. Qwen models use Thinker text output. MiniCPM-o uses non-thinking mode, `max_slice_nums=1`, and `use_image_id=False`.

C.2 Shared media construction

Video frames are sampled over the full clip at 1 fps and resized to fit within 448×448 pixels while preserving aspect ratio. We use the full audio track at 16 kHz in mono and the original transcripts. All conditions share the same media, with absent modalities omitted; each model uses its native processor.

C.3 Decoding and output parsing

All models run in BF16 without quantization and decode greedily (`do_sample=False`, `num_beams=1`, `max_new_tokens=16`). The parser accepts a single answer label after normalizing case, surrounding whitespace or quotation marks, and an optional final full stop; it does not extract labels from explanations or map unrecognized outputs to `neutral`. An invalid V/A/T answer excludes the example from a conditional cohort; an invalid Full or A+T answer counts as incorrect.

D Evaluation Prompts

All models receive the same English task instructions for each input condition, with transcripts kept in their original language. We use each model's native chat template without additional task-specific system instructions. The five templates share one layout, shown here for Full input:

```
Task:
Judge the overall sentiment conveyed by the clip's audio, video, and
transcript.

Transcript:
"{transcript}"

Answer choices:
positive
negative
neutral

Return only one answer choice.
```

Every template begins with the line `Task:`, a condition-specific task line (Table 5), and a blank line. For T-only, Full, and A+T, a transcript block follows: the line `Transcript:`, the line `"{transcript}"` with the placeholder replaced by the clip's transcript, and a blank line. V-only and A-only omit this block. All templates then end with the same four lines from `Answer choices:` through `neutral`, a blank line, and `Return only one answer choice.` The A+T prompt thus replaces "audio, video, and transcript" in the Full prompt with "audio and transcript."

Table 5: Task line of each prompt template. Each task line is a single line in the template; wrapping here is typographic.

Condition	Task line	Transcript block
V-only	`Judge the overall sentiment conveyed by the clip's silent video only. No transcript or audio is provided.`	no
A-only	`Judge the overall sentiment conveyed by the clip's audio only. No transcript or video frames are provided.`	no
T-only	`Judge the overall sentiment conveyed by the transcript only. No audio or video is provided.`	yes
Full	`Judge the overall sentiment conveyed by the clip's audio, video, and transcript.`	yes
A+T	`Judge the overall sentiment conveyed by the clip's audio and transcript.`	yes

E V-ONLY COMPETENCE UNDER FULL INPUT

This appendix gives complete behavioral results for the selection rule in Appendix B.3, followed by robustness checks.

E.1 FULL ERROR RATES

For model m, dataset d, and gold label ℓ, let $\mathcal{C}_{m,d,\ell}$ be the selected cohort and $n_{m,d,\ell} = |\mathcal{C}_{m,d,\ell}|$ its actual size. Conditional retention and loss are

$$R_{m,d,\ell} = \frac{1}{n_{m,d,\ell}} \sum_{i \in \mathcal{C}_{m,d,\ell}} \mathbf{1}[\hat{y}_i^{\text{Full},m} = y_i], \qquad L_{m,d,\ell} = 1 - R_{m,d,\ell}, \tag{7}$$

and L is the Full error rate. Table 6 reports it with sample sizes for all 27 model–dataset–label cohorts, with intervals as in Appendix B.3. Four model–example pairs, covering three clips, also appear in development checks.

Table 7 compares pooled error, $\sum_\ell n_\ell L_\ell / \sum_\ell n_\ell$, with equal weighting of labels, $\frac{1}{3}\sum_\ell L_\ell$. It resamples source clusters jointly across labels, retaining every clip in each sampled cluster. The largest difference is MiniCPM-o/CH-SIMS: 51.4% pooled versus 39.0% with equal label weights. Thus the pooled range does not establish a population majority in every setting. The 1,830 model–example pairs represent 1,159 unique clips, deduplicated by dataset and sample ID across models.

Table 6: Full error rates on all 27 selected V-correct, A-incorrect, T-incorrect cohorts. Qwen2.5 and Qwen3 denote Qwen2.5-Omni and Qwen3-Omni, and MOSEI denotes CMU-MOSEI; later tables use the same abbreviations. Intervals use source-cluster bootstrap. †: fewer than 30 examples or ten source clusters.

Dataset	Model	Gold label	n	Full error (%)	95% CI (%)
MELD	Qwen2.5	Neg.	100	91.0	[85.1, 95.9]
		Neu.	100	83.0	[75.2, 89.8]
		Pos.	100	80.0	[72.4, 87.2]
	MiniCPM-o	Neg.	100	51.0	[40.2, 61.1]
		Neu.	100	97.0	[93.2, 100.0]
		Pos.	100	33.0	[23.8, 42.2]
	Qwen3	Neg.	67	59.7	[48.4, 71.6]
		Neu.	100	83.0	[75.5, 90.3]
		Pos.	100	63.0	[52.3, 73.4]
CH-SIMS	Qwen2.5	Neg.	12†	66.7	[40.0, 100.0]
		Neu.	47	89.4	[76.1, 100.0]
		Pos.	25†	80.0	[57.9, 96.8]
	MiniCPM-o	Neg.	27†	11.1	[0.0, 25.0]
		Neu.	53	96.2	[91.8, 100.0]
		Pos.	31	9.7	[0.0, 21.9]
	Qwen3	Neg.	19†	63.2	[40.0, 86.7]
		Neu.	50	88.0	[80.4, 95.0]
		Pos.	38	50.0	[36.1, 63.4]
MOSEI	Qwen2.5	Neg.	18†	88.9	[71.4, 100.0]
		Neu.	100	91.0	[84.9, 96.0]
		Pos.	100	84.0	[75.7, 91.3]
	MiniCPM-o	Neg.	22†	31.8	[15.0, 50.0]
		Neu.	100	92.0	[86.5, 97.0]
		Pos.	100	42.0	[31.6, 52.1]
	Qwen3	Neg.	21†	85.7	[66.7, 100.0]
		Neu.	100	89.0	[82.4, 94.7]
		Pos.	100	66.0	[56.7, 75.7]

Table 7: Full errors under two aggregation rules on the same selected training cohorts. Pooled weights labels by their achieved sample counts; label-equal gives each of the three labels weight $1/3$. Brackets are descriptive 95% source-cluster bootstrap intervals (2,000 draws, seed 20260920), using the same draws for both rules. These conditional, label-capped cohorts do not estimate population error rates.

Dataset	Model	n	Sources	Pooled error (%)	Label-equal error (%)
MELD	Qwen2.5	300	243	84.7 [80.6, 88.4]	84.7 [80.7, 88.4]
MELD	MiniCPM-o	300	245	60.3 [54.6, 65.6]	60.3 [55.5, 64.8]
MELD	Qwen3	267	227	69.7 [63.9, 75.5]	68.6 [62.9, 74.5]
CH-SIMS	Qwen2.5	84	31	83.3 [73.6, 92.4]	78.7 [67.2, 91.2]
CH-SIMS	MiniCPM-o	111	39	51.4 [40.0, 62.9]	39.0 [33.9, 45.6]
CH-SIMS	Qwen3	107	41	70.1 [62.3, 78.2]	67.1 [58.1, 76.4]
MOSEI	Qwen2.5	218	180	87.6 [82.8, 91.9]	88.0 [81.3, 93.1]
MOSEI	MiniCPM-o	222	189	63.5 [57.1, 70.0]	55.3 [48.7, 62.3]
MOSEI	Qwen3	221	187	78.3 [72.6, 83.9]	80.2 [72.7, 86.4]

E.2 Full predictions relative to unimodal answers

Full returns an A-only or T-only answer in 1,298 of its 1,328 errors. To test whether video nonetheless helps, we compare Full with A+T on the same examples, retaining audio, transcripts, decoding, and recorded Full answers. Only the prompt's modality description changes (Appendix D). A+T contains the complete waveform and no video tensors. MiniCPM-o retains its audio padding but uses different native serialization: Full interleaves frames and audio chunks, whereas A+T supplies

Table 8: A+T and Full accuracy (%) on the fixed train cohorts. Pooled accuracy weights every example equally, as in the main text; label-equal accuracy averages the three per-label accuracies. Δ is Full minus A+T in percentage points, computed before rounding. Intervals are for Δ and use paired source-cluster bootstrap.

		Pooled				Label-equal			
Model	Dataset	A+T	Full	Δ	95% CI	A+T	Full	Δ	95% CI
Qwen2.5	MELD	9.7	15.3	+5.7	[+3.0, +8.4]	9.7	15.3	+5.7	[+3.0, +8.4]
	CH-SIMS	7.1	16.7	+9.5	[+2.7, +16.1]	7.6	21.3	+13.8	[+5.2, +22.5]
	MOSEI	4.1	12.4	+8.3	[+4.6, +12.3]	3.0	12.0	+9.0	[+4.3, +15.4]
MiniCPM-o	MELD	7.0	39.7	+32.7	[+27.4, +38.1]	7.0	39.7	+32.7	[+27.9, +37.7]
	CH-SIMS	7.2	48.6	+41.4	[+28.1, +52.6]	7.7	61.0	+53.3	[+43.0, +61.0]
	MOSEI	4.5	36.5	+32.0	[+25.8, +38.0]	4.5	44.7	+40.2	[+32.8, +47.5]
Qwen3	MELD	4.5	30.3	+25.8	[+20.3, +31.5]	4.8	31.4	+26.6	[+20.8, +32.3]
	CH-SIMS	3.7	29.9	+26.2	[+17.6, +34.2]	4.0	32.9	+29.0	[+20.3, +37.2]
	MOSEI	4.5	21.7	+17.2	[+12.1, +22.7]	3.3	19.8	+16.4	[+10.7, +23.5]

a complete audio input. With $n_{\ell,ab}$ counting the examples in label cohort ℓ with A+T correctness a and Full correctness b, where $a, b \in \{0, 1\}$, we report the label-equal paired accuracy difference

$$\Delta = \frac{1}{3} \sum_{\ell \in \mathcal{Y}} \big[\mathrm{Acc}_\ell(\mathrm{Full}) - \mathrm{Acc}_\ell(\mathrm{A{+}T}) \big] = \frac{1}{3} \sum_{\ell \in \mathcal{Y}} \frac{n_{\ell,01} - n_{\ell,10}}{n_\ell}, \tag{8}$$

without reselecting examples. Full is more accurate in all nine settings under both weighting rules (Table 8); all meet the minimum of ten examples per label for equal weighting. One MiniCPM-o A+T answer, `egative`, is invalid under the fixed parser and remains incorrect in the denominator; its Full answer is also incorrect.

We also compare Full with majority voting over unimodal answers (Table 9). A vote is wrong whenever A-only and T-only agree; three-way ties are resolved toward V-only or uniformly at random. Full has fewer errors in eight of nine settings. Qwen2.5-Omni/MELD is the exception for voting with ties resolved toward V-only. When A-only and T-only agree, Full recovers 405/1,585 answers; in 1,150/1,180 remaining errors, it returns their shared label. Video contributes, but Full usually follows agreement between audio and text.

Table 9: Full versus majority voting on the selected training examples. Three-way ties favor V-only or are broken at random. The last column counts correct Full answers when A-only and T-only agree. Abbreviations follow Table 6.

				Vote error (%)		
Dataset	Model	n	Full error (%)	ties to V	random ties	A=T: Full correct
MELD	Qwen2.5	300	84.7	81.0	93.7	33/243
	MiniCPM-o	300	60.3	82.7	94.2	88/248
	Qwen3	267	69.7	83.1	94.4	62/222
CH-SIMS	Qwen2.5	84	83.3	88.1	96.0	11/74
	MiniCPM-o	111	51.4	86.5	95.5	42/96
	Qwen3	107	70.1	92.5	97.5	29/99
MOSEI	Qwen2.5	218	87.6	91.3	97.1	25/199
	MiniCPM-o	222	63.5	90.5	96.8	70/201
	Qwen3	221	78.3	91.9	97.3	45/203
All		1830	72.6	86.6	95.5	405/1585

Table 10 compares cases with a single correct modality on the fixed test subsets. The other two modalities must agree on a wrong label. These columns contain different examples, and the pooled comparison reflects their different label distributions. After excluding neutral examples, lower Full accuracy in the V-correct group than in both other groups is supported for Qwen2.5-Omni but not consistently across models.

Table 10: Full accuracy when exactly one modality is correct and the other two agree on a wrong label. Columns use different clips from the fixed test subsets. Differences (V−A, V−T) are percentage points, with 95% percentile intervals from 4,000 bootstrap resamples of source clusters within each row (dialogues for MELD, source videos otherwise).

		Full correct			Difference (pp), 95% CI	
Gold labels	Model	V-correct	A-correct	T-correct	V−A	V−T
All	Qwen2.5	11/59	11/25	82/120	[-49.1, -2.7]	[-62.8, -35.6]
	MiniCPM-o	27/84	45/61	38/66	[-57.5, -25.5]	[-40.6, -10.0]
	Qwen3	25/102	2/27	41/70	[+2.3, +30.5]	[-49.4, -18.4]
	Pooled	63/245	58/113	161/256	[-36.9, -14.5]	[-45.3, -29.0]
Neg. + Pos.	Qwen2.5	1/18	8/21	79/115	[-56.2, -9.7]	[-75.7, -47.6]
	MiniCPM-o	24/29	45/57	38/52	[-13.4, +20.7]	[-8.3, +28.0]
	Qwen3	16/35	2/25	38/61	[+14.1, +58.5]	[-40.9, +5.9]
	Pooled	41/82	55/103	155/228	[-18.0, +10.5]	[-31.1, -5.2]

E.3 Robustness of the neutral result

Repeated runs preserve the high neutral error rate (Table 11). We repeat all four input conditions twice with identical inputs and the inference settings of Appendix C. Stable membership requires three identical, correct V-only answers and three identical, incorrect answers under each of A-only and T-only. Each group is refilled in the original candidate order until 100 stable members or pool exhaustion; Full never determines membership. Among members whose Full answers are also stable, error rates are 82.8%–96.9%. Repeats span multiple GPUs, so they measure numerical variation without isolating its hardware source.

Table 11: Neutral cohorts under three repeated runs. Stable V/A/T counts pairs whose three V-only answers are identical and correct and whose A-only and T-only answers are identical and incorrect. Full error uses pairs with identical Full answers; Full unstable counts the rest. Refilled cohorts add stable candidates in the original order (Appendix E.3). Figure 1b plots the refilled column.

		Original cohort			Refilled cohort	
Model	Dataset	n	Stable V/A/T	Full error (%)	Full error (%)	Full unstable
Qwen2.5	MELD	100	96	82/96 (85.4)	86/100 (86.0)	0
	CH-SIMS	47	45	40/44 (90.9)	40/44 (90.9)	1
	MOSEI	100	99	90/99 (90.9)	91/100 (91.0)	0
MiniCPM-o	MELD	100	97	92/95 (96.8)	95/98 (96.9)	2
	CH-SIMS	53	52	50/52 (96.2)	50/52 (96.2)	0
	MOSEI	100	97	89/97 (91.8)	92/100 (92.0)	0
Qwen3	MELD	100	100	82/99 (82.8)	82/99 (82.8)	1
	CH-SIMS	50	50	44/50 (88.0)	44/50 (88.0)	0
	MOSEI	100	100	89/100 (89.0)	89/100 (89.0)	0

Two prompt variants on Qwen2.5-Omni/MELD test sensitivity to wording. The first orders labels as neutral, positive, negative. The second keeps the order but replaces "Judge the overall sentiment conveyed by" with "Classify the overall sentiment expressed in" and "Return only one answer choice." with "Reply with exactly one of the listed labels." All three prompts are evaluated with the same media, decoding settings, and hardware. Both the prompt variants and the full-split evaluation were planned after the original-prompt results were available; the variant prompts and criteria were fixed before variant inference.

Test selection is reapplied independently under each prompt, so the groups can contain different clips. Evaluation covers the full MELD test split, including the records from the initial 400-clip check. Two of 2,610 clips exceed GPU memory before producing answers and are excluded under every prompt. Each remaining clip receives all four input conditions under all three prompts on the same GPU configuration. Both variants satisfy the primary neutral criterion: at least 20 eligible clips from ten dialogues, with the one-sided 95% error lower bound above 50% (Table 12). Intervals use

20,000 dialogue bootstrap draws from the full test pool (seed 20260925), recomputing selection in each draw.

Table 12: Prompt evaluation on 2,608 MELD test clips with Qwen2.5-Omni. Each prompt separately selects examples with correct V-only and incorrect A-only and T-only predictions. n is the selected count and k the number of dialogues. Brackets give 95% bootstrap intervals with dialogues as resampling units.

Gold	Prompt	n	k	Full errors	Full error (%)
negative	Original	28	27	25/28	89.3 [76.0, 100.0]
negative	Neutral first	38	35	36/38	94.7 [86.5, 100.0]
negative	Reworded	24	21	22/24	91.7 [79.3, 100.0]
neutral	Original	168	111	147/168	87.5 [82.1, 92.5]
neutral	Neutral first	163	105	130/163	79.8 [73.5, 85.7]
neutral	Reworded	164	106	145/164	88.4 [82.9, 93.6]
positive	Original	50	38	40/50	80.0 [65.5, 93.9]
positive	Neutral first	41	34	30/41	73.2 [56.2, 88.6]
positive	Reworded	42	32	29/42	69.0 [52.5, 85.3]

The training comparison instead keeps all 100 original neutral examples fixed (Table 13). Listing neutral first lowers Full error from 83% to 77%; rewording leaves it unchanged. The secondary non-inferiority criterion requires a one-sided 95% lower bound on the paired change above -10 percentage points. Rewording passes and placing neutral first does not.

Table 13: Prompt variants on the same 100 neutral training examples with Qwen2.5-Omni and MELD. Retained counts examples that still satisfy correct V-only and incorrect A-only and T-only predictions. Full error and its paired change use all 100 examples. Brackets give 95% dialogue-cluster bootstrap intervals.

Prompt	V-correct	Retained	Full error (%)	Change (pp)
Original	100	100	83.0 [75.5, 90.0]	–
Neutral first	99	87	77.0 [68.4, 85.3]	−6.0 [−11.4, −1.0]
Reworded	100	87	83.0 [75.3, 89.9]	0.0 [−4.0, 4.0]

Finally, the fixed test subsets of Appendix B.3 preserve class proportions. Tables 14 and 15 separate selection prevalence from Full error among selected clips. Neutral error spans 60.0%–97.1% with only 5–47 eligible examples per setting; several intervals include 50%. These estimates use one greedy run per input without stability filtering and remain separate from the training results. The sampled, media-checked subsets do not estimate prevalence in unfiltered data.

Table 14: Prevalence and conditional Full errors on the fixed test subsets. Each setting contains 400 clips. Eligible means V-only correct and A-only and T-only incorrect. Prevalence uses all 400 clips; Full error uses only eligible clips. Intervals are 95% source-cluster bootstrap intervals (2,000 draws). †: fewer than 30 eligible clips or ten source clusters.

Model	Dataset	Eligible/400	Prevalence (%)	Full errors (%)	95% CI (%)
Qwen2.5	MELD	35/400	8.75	31/35 (88.6)	[77.3, 97.4]
	CH-SIMS	19/400†	4.75	14/19 (73.7)	[50.0, 94.4]
	MOSEI	13/400†	3.25	10/13 (76.9)	[50.0, 100.0]
MiniCPM-o	MELD	45/400	11.25	35/45 (77.8)	[65.1, 89.1]
	CH-SIMS	30/400	7.50	15/30 (50.0)	[30.8, 67.7]
	MOSEI	23/400†	5.75	13/23 (56.5)	[36.4, 76.9]
Qwen3	MELD	51/400	12.75	41/51 (80.4)	[68.3, 90.9]
	CH-SIMS	42/400	10.50	30/42 (71.4)	[51.6, 88.6]
	MOSEI	22/400†	5.50	14/22 (63.6)	[42.1, 84.6]

Table 15: Full errors among test-subset eligible clips, by gold label. Entries are errors/eligible clips; these descriptive counts can have very small denominators. A dash indicates no eligible clips.

Model	Dataset	Negative	Neutral	Positive
Qwen2.5	MELD	4/4	23/27	4/4
	CH-SIMS	1/1	8/13	5/5
	MOSEI	1/1	3/5	6/7
MiniCPM-o	MELD	1/7	33/34	1/4
	CH-SIMS	1/7	13/15	1/8
	MOSEI	0/1	10/11	3/11
Qwen3	MELD	—	39/47	2/4
	CH-SIMS	2/5	17/19	11/18
	MOSEI	1/1	7/8	6/13

F Sensitivity to Video under V-only and Full Input

We compare residual gradients and video replacements under V-only and Full, then test V-only answers after disrupting video content.

F.1 Residual-gradient sensitivity to visual tokens

The analysis uses frozen Qwen2.5-Omni on two CMU-MOSEI sets. A contains the 43 validation recipients of Appendix G.2 (40 Full failures, three successes), all V-correct and A/T-incorrect. B contains all 98 V-correct clips from the fixed test subset, without A/T restrictions: 22 failures and 76 successes from 82 source videos. Technical development clips are excluded, and neither set contains a negative Full success. Groups and metrics were fixed before gradient collection. Inference uses native BF16 SDPA on an H100.

Following gradient sensitivity analysis (Baehrens et al., 2010; Simonyan et al., 2014), we differentiate label-logit differences with respect to token residual states. For clip i, condition $h \in \{V, F\}$, decoder output layer l, and physical token position p, let $x^h_{i,l,p}$ be the residual state and $z^h_{i,a}$ the first-answer logit of label token a. We measure

$$G^h_{i,ab}(l,p) = \left\| \nabla_{x^h_{i,l,p}} (z^h_{i,a} - z^h_{i,b}) \right\|_2, \qquad J^h_{i,ab} = \frac{1}{27|\mathcal{V}^h_i|} \sum_{l=0}^{26} \sum_{p \in \mathcal{V}^h_i} G^h_{i,ab}(l,p), \tag{9}$$

where $\mathcal{V}^h_i$ contains visual-token positions. The three label pairs span two independent logit differences; layers, tokens, and contrasts are not independent replicates. Layer 0 is made a gradient leaf without changing forward values, so this measures downstream residual sensitivity rather than pixel gradients. Layer 27 has structurally zero gradients at non-final tokens and is excluded from J and cross-layer summaries. Saved maps and Figure 7 retain all 28 layers; Figure 2 shows 0–26. Paired inputs have identical pixels and video grids; visual tokens align by grid order, not absolute sequence index.

Within each gold-label and Full-outcome group, clips are averaged within source video and then sources receive equal weight. Pointwise 95% intervals use 2,000 paired source bootstrap draws without multiplicity correction. Failure-minus-success comparisons resample sources jointly across outcomes. For display, groups with at most two sources show means only; groups with fewer than ten sources are marked.

$J^F < J^V$ for every clip and label pair in both sets (Table 16). Reportable failure-minus-success intervals for the absolute change include zero, so the reduction does not establish a distinction between Full failures and successes. The curves in Figure 7 show sensitivity declining through the decoder in both conditions.

Figure 2 instead shows all token types. For each clip, gradients are averaged over label pairs and binned by token order into video, audio, transcript, and prompt segments with 22, 8, 9, and 10 bins, respectively. Prompt includes task, special, and final tokens; short segments reuse tokens across bins. Maps are averaged within source and then across the 125 sources. Over layers 0–18 under

Table 16: Full/V-only visual gradient-index ratios by cohort, gold label, and Full outcome. Cells give source-weighted means [descriptive 95% intervals]. P, N, and U denote positive, negative, and neutral. An asterisk marks fewer than ten sources; groups with at most two sources show means only. Intervals are resampled separately for each stratum.

Cohort / gold	Full	Clips / sources	P–N	P–U	N–U
A / negative	failure	2 / 2*	0.285	0.320	0.205
A / neutral	failure	30 / 30	0.199 [0.175, 0.228]	0.248 [0.223, 0.274]	0.212 [0.182, 0.245]
A / neutral	success	2 / 2*	0.243	0.268	0.230
A / positive	failure	8 / 8*	0.247 [0.193, 0.309]	0.202 [0.165, 0.238]	0.328 [0.273, 0.397]
A / positive	success	1 / 1*	0.231	0.302	0.387
B / negative	failure	1 / 1*	0.196	0.190	0.148
B / neutral	failure	15 / 15	0.305 [0.250, 0.359]	0.311 [0.264, 0.360]	0.245 [0.217, 0.275]
B / neutral	success	70 / 58	0.266 [0.237, 0.298]	0.276 [0.253, 0.302]	0.256 [0.234, 0.282]
B / positive	failure	6 / 6*	0.265 [0.190, 0.350]	0.228 [0.176, 0.280]	0.414 [0.287, 0.548]
B / positive	success	6 / 6*	0.169 [0.135, 0.203]	0.195 [0.155, 0.234]	0.245 [0.213, 0.283]

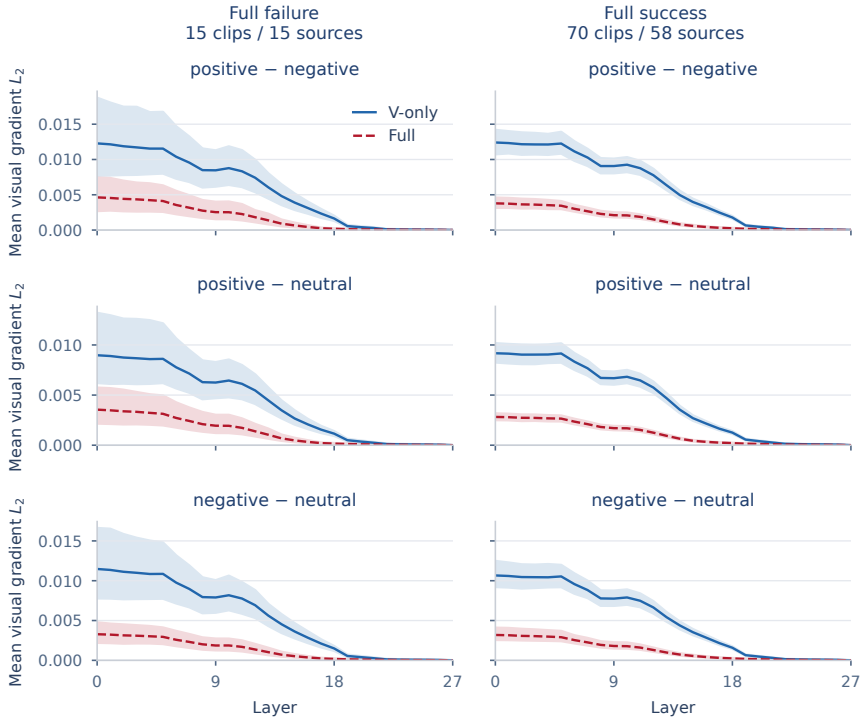


Figure 7: Layer-resolved visual sensitivity for neutral cohort-B clips, by Full outcome (columns) and label contrast (rows). Curves average visual-token gradient L_2 norms per layer with equal source-video weights; bands are pointwise 95% source-cluster bootstrap intervals (2,000 draws).

Full, mean per-token norms are 1.8×10^{-3} for video, 1.3×10^{-2} for audio, and 1.1×10^{-1} for transcripts. Audio exceeds video in 140/141 clips, and transcripts do so in all clips.

The layer ratios in Section 4.1 compare these aggregated video means. With $H^s_{v,i,\ell,b}$ the mean gradient norm over the three contrasts and tokens in video bin b for clip i from source v under input s, and N_v clips per source,

$$\mu^s_\ell = \frac{1}{125}\sum_{v=1}^{125}\frac{1}{N_v}\sum_{i\in v}\frac{1}{22}\sum_{b=1}^{22} H^s_{v,i,\ell,b}, \qquad r_\ell = \mu^F_\ell/\mu^V_\ell, \tag{10}$$

rather than averages of per-clip ratios or sums over a modality's tokens. At layers 0, 9, and 18, r_ℓ is 0.2865, 0.2271, and 0.1435, respectively, giving the reported 29%, 23%, and 14%. Native forward logits and answers are reproduced, and finite-difference checks account for BF16 rounding. Deterministic algorithms are enabled only for backward.

F.2 Sensitivity to video replacement

Whole-video replacement produces larger gold-margin changes under V-only than Full (Figure 8). Recipients meet the same V-correct, A-incorrect, T-incorrect condition, screened in fixed hash order through T, V, and A before Full evaluation. All eligible CMU-MOSEI test pools are exhausted below the target of 100 per label, yielding 90 clips (Table 17). Media criteria follow Appendix B.3; source videos referenced in prior experiments or already cached are excluded. Each recipient receives three V-correct training donors per label, selected by absolute log-duration distance with a fixed tie-breaker. Its nine donor sources are distinct and disjoint from recipients and development examples. Assignments are fixed before replacement inference.

Table 17: CMU-MOSEI test cohorts for video replacement. Each recipient satisfies V-correct, A-incorrect, T-incorrect. All candidate pools are exhausted. Full outcomes use each recipient's own video.

Gold label	Candidates	n	Sources	Full correct	Retention (%)
Neg.	445	3	1	0/3	0.0
Neu.	412	51	45	2/51	3.9
Pos.	962	36	18	12/36	33.3

Donor frames use evenly spaced rounded indices and bicubic resizing to match recipient geometry. Full keeps recipient audio and transcript; all non-video processor tensors stay fixed. Replacement changes the whole video, including person and scene. Let i identify a recipient with gold label y_i, d a donor video, and $h \in \{V, F\}$ the input condition, and let $g_i^h(d)$ be the gold margin of Equation 3, computed from the raw first-answer logits of the canonical label tokens (each label is a single token). The competing-label contrast is

$$D_{i,c}^h = \frac{1}{3} \sum_{d \in \mathcal{D}_{i,c}} g_i^h(d) - \frac{1}{3} \sum_{d \in \mathcal{D}_{i,y_i}} g_i^h(d), \qquad c \neq y_i, \tag{11}$$

where $\mathcal{D}_{i,c}$ contains three donors of label c. The common own-video baseline cancels. Contrasts are averaged over recipients for each gold and competing label. Intervals use 2,000 recipient-source bootstrap draws with donors fixed; the three negative recipients share one source and have no interval. V-only contrasts range from -5.63 to -3.12, versus -0.42 to $+0.02$ under Full.

Table 18 reports how often the output retains the recipient's original gold label, averaging donors within recipient first. This is target retention, not accuracy against a new annotation. All conditions for a clip run on the same H100, and replacing the video with itself reproduces native inputs, answers, and scores.

Table 18: Rates (%) of predicting the recipient's original gold label. Own video is the recipient's own video; the other columns average three donors with the indicated label per recipient.

Gold label	Input	Own video	Neg. donor	Neu. donor	Pos. donor
Neg.	V-only	100.0	88.9	0.0	0.0
	Full	0.0	0.0	0.0	0.0
Neu.	V-only	100.0	9.8	97.4	11.8
	Full	3.9	4.6	5.2	4.6
Pos.	V-only	100.0	0.0	0.0	88.9
	Full	33.3	20.4	29.6	28.7

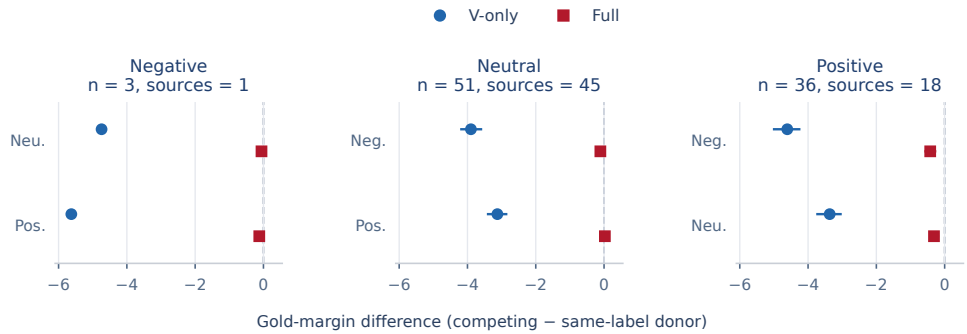


Figure 8: Video replacement effects: competing minus same-label gold margin, with competing labels in rows. Negative values favor same-label videos. Full retains recipient audio and transcript. Bars are 95% source bootstrap intervals with donors fixed; the single-source negative group has no interval.

F.3 Dependence of V-only answers on video content

On the fixed test subsets, V-only outputs neutral on 54.4%–92.9% of non-neutral examples (Table 19). Frame controls test which answers change when visual content is disrupted.

Table 19: Neutral answers under V-only and Full on the fixed test subsets of 400 clips. Negative, neutral, and positive columns give the share of examples with that gold label answered neutral; non-neutral pools negative and positive gold labels; all uses every example; precision is the share of neutral answers whose gold label is neutral. Neutral gold counts are 192, 60, and 88 of 400 for MELD, CH-SIMS, and MOSEI.

			Answered neutral (%), by gold label					
Model	Dataset	Input	Negative	Neutral	Positive	Non-neutral	All	Precision (%)
Qwen2.5	MELD	V-only	72.7	75.5	67.5	70.7	73.0	49.7
		Full	26.6	69.8	22.5	25.0	46.5	72.0
	CH-SIMS	V-only	65.4	88.3	79.7	70.6	73.2	18.1
		Full	25.8	58.3	37.4	30.0	34.2	25.5
	MOSEI	V-only	92.2	96.6	93.4	92.9	93.8	22.7
		Full	28.4	83.0	44.4	38.5	48.2	37.8
MiniCPM-o	MELD	V-only	65.6	62.5	50.0	59.6	61.0	49.2
		Full	7.0	28.6	6.2	6.7	17.2	79.7
	CH-SIMS	V-only	60.8	81.7	70.7	64.4	67.0	18.3
		Full	8.8	40.0	8.9	8.8	13.5	44.4
	MOSEI	V-only	88.8	93.2	85.7	86.9	88.2	23.2
		Full	19.0	65.9	30.1	26.0	34.8	41.7
Qwen3	MELD	V-only	80.5	75.5	63.8	74.0	74.8	48.5
		Full	19.5	51.6	13.8	17.3	33.8	73.3
	CH-SIMS	V-only	54.4	78.3	54.5	54.4	58.0	20.3
		Full	18.9	51.7	22.8	20.3	25.0	31.0
	MOSEI	V-only	80.2	89.8	77.0	78.2	80.8	24.5
		Full	19.8	69.3	36.7	30.4	39.0	39.1

We use the 43 CMU-MOSEI validation clips and all 57 remaining CH-SIMS test clips outside the fixed 400-clip subset, without further outcome selection. The latter can share sources or speakers with that subset. CMU-MOSEI uses multimodal gold labels and Full outcomes from screening; CH-SIMS uses its independent visual annotations with the thresholds of Equation 5.

Each clip receives original frames and four controls: uniform RGB 127, each frame filled with its mean color, a fixed spatial pixel permutation shared across frames, and blur by bilinear downsampling to 8×8 then upsampling. Frame count, dimensions, prompt, and processor metadata stay

fixed. Conditions run on the same device under Appendix C, with assignments fixed before new predictions.

All 400 disrupted-input predictions are neutral (Table 20). Correct non-neutral original answers switch to neutral, while neutral answers persist. Retention uses the original target because disrupted videos have no new annotation.

Table 20: All 100 clips in the video-content control study. Each control column reports retention of the original target label. Original correctness uses the multimodal gold label for CMU-MOSEI and the independent visual label for CH-SIMS. Full outcomes from screening define the CMU-MOSEI groups. Abbreviations follow Table 6.

Cohort	Label	n	Original	Gray	Mean	Shuffle	Blur
MOSEI failures	negative	2	1	0	0	0	0
MOSEI failures	neutral	30	30	30	30	30	30
MOSEI failures	positive	8	8	0	0	0	0
MOSEI successes	neutral	2	2	2	2	2	2
MOSEI successes	positive	1	1	0	0	0	0
CH-SIMS visual	negative	29	16	0	0	0	0
CH-SIMS visual	neutral	9	9	9	9	9	9
CH-SIMS visual	positive	19	3	0	0	0	0

Neutral answers persist while their margins increase on disrupted frames (Table 21). These controls show a neutral default under disruption; they do not identify whether an original neutral answer relied on recognizable visual content. The controls alter multiple visual properties together and do not isolate sentiment semantics.

Table 21: Original-minus-control neutral margin on the 30 neutral CMU-MOSEI failures. Negative values indicate higher neutral margins on control frames. Intervals use 2,000 paired source-cluster bootstrap draws and are descriptive, pointwise 95% intervals.

Control	Margin difference	95% interval
Uniform gray	-1.206	$[-1.549, -0.895]$
Frame-mean color	-1.173	$[-1.519, -0.861]$
Spatial shuffle	-0.840	$[-1.203, -0.494]$
Strong blur	-0.680	$[-1.080, -0.296]$

Paired comparisons use unmodified and disrupted inputs run on the same device. Two unmodified-input answers differ from screening predictions: a negative CMU-MOSEI answer becomes neutral, and a neutral CH-SIMS answer becomes negative. CMU-MOSEI processor tensors match the screening tensors; CH-SIMS uses different hardware and a different attention implementation from its screening run, whose complete logits are unavailable. Both cases remain in the analysis; excluding them preserves the neutral results and the switch to neutral for every remaining correct non-neutral original answer.

G State-Transplant Protocol

This appendix defines the suffix interface, recipients, donor selection, source–host contrasts, and statistical conventions for the main Qwen2.5-Omni transplant experiments. Settings specific to the controls, new-data studies, and second model appear with those analyses.

G.1 SUFFIX INTERFACE

For Qwen2.5-Omni-7B, the intervention covers the following 19 tokens after applying the native chat template:

```
Answer choices:
positive
negative
neutral

Return only one answer choice.
<|im_end|>
<|im_start|>assistant
```

The final newline, which supplies the final query $q^\star$, is excluded.

Suffix token IDs are identical across V-only and Full and are aligned by relative position. Interventions cover all heads and layers 0–27; the 28 query heads retain their native mapping to four KV heads, no further layer or position search is performed, and inference follows Appendix C.

The interface comparison uses 12 CMU-MOSEI training clips, four per label, excluded from all transplant recipients. Each is V-correct and incorrect under A-only, T-only, and Full. Table 22 compares own-video transfers at suffix and video positions, without norm matching. Video positions align by spatial and temporal coordinates; the 19-token video condition averages three fixed subsets sampled without using predictions or attention. The all-layer setting was fixed before this comparison and retained in subsequent experiments.

Table 22: Interface comparison on 12 development clips: mean gold-margin change under own-video V-only-to-Full transfer. Layer indices are zero-based.

Layers	All video tokens	19 video tokens	19 suffix tokens
Early (0–8)	$+0.029$	-0.003	-0.015
Middle (9–17)	-0.021	-0.022	$+0.102$
Late (18–27)	-0.005	-0.004	$+2.512$
All (0–27)	0.000	-0.011	$+2.403$

G.2 RECIPIENTS AND DONORS

The main analysis uses 156 V-correct, A-incorrect, T-incorrect Full failures: 40 CMU-MOSEI validation clips, 26 CMU-MOSEI test clips, and 90 MELD training clips. They comprise 35 negative, 60 neutral, and 61 positive examples; Table 23 gives the split by dataset and label. Three validation Full successes serve as controls. Recipients and donor assignments are fixed before intervention outcomes are observed.

CMU-MOSEI validation candidates are screened in fixed order through T-only, V-only, A-only, and then Full, taking at most one clip per source video. Media must contain audio and video, pass integrity checks, have transcript alignment similarity at least 0.5, and differ from the annotated duration by at most 0.3 seconds. Invalid or generation-capped answers are excluded. Validation sources are disjoint from development, the train and test inventories, and the train donor pool. Screening exhausts the eligible pool under caps of 30 failures and ten successes per label; one additional neutral failure exceeds its cap. Supplemental test screening uses the same media and source rules, targets 30 failures per non-neutral label, and exhausts the pool with three negative and 23 positive examples. Sources used in prior interventions or development are excluded; prior native predictions alone do not exclude a source. MELD supplies 30 failures per label from distinct dialogues in the behavioral training pool, with eligibility checked on the transplant hardware and development dialogues excluded.

Each recipient receives four V-correct training donors from its dataset, two per competing label. Donor sources exclude development examples and all recipients. Candidates are ranked by absolute log duration ratio and then by fixed hash order, taking distinct sources. Allocation 0 and allocation 1 each contain one donor per competing label. Assignments are shared across analyses of the validation recipients; the full donor pool contains 298 unique donors. Disjointness is defined by source

video for CMU-MOSEI and dialogue for MELD, rather than speaker identity, which is unavailable. Donor labels are dataset annotations.

Donor frames are sampled at evenly spaced rounded indices and bicubically resized to the recipient's frame count and dimensions. Adapted pixels match across V-only and Full sources; all other processor tensors remain fixed within each condition. Suffix IDs and relative order align across source and host. All technically valid assignments enter the primary analysis. A calibrated sensitivity analysis keeps donors whose adapted V-only answer matches their own gold label, requires both competing labels, and uses the same donors in all four source–host cells. Full predictions never screen donors.

G.3 Contrasts and statistical conventions

Let $z_i(s, k \to h)$ denote logits after transplanting Values formed under source condition $s \in \{V, F\}$ with video k into host $h \in \{V, F\}$. For gold label y_i, the own-versus-donor contrast is

$$E_i(s, h; d) = g\big(z_i(s, \text{own} \to h), y_i\big) - g\big(z_i(s, d \to h), y_i\big), \tag{12}$$

where own and d denote the recipient's own video and a donor video and g is the gold margin of Equation 3. E_i thus compares two transplants, not the gain over an unmodified host. Averaging within each competing label and then across labels gives $E_i(s, h)$. The primary comparison fixes the Full host,

$$G_{F,i} = E_i(V, F) - E_i(F, F), \tag{13}$$

so positive $G_{F,i}$ means that own-versus-donor margin differences are larger for V-only source Values in the fixed Full host. The supporting comparison fixes the V-only host,

$$G_{V,i} = E_i(V, V) - E_i(F, V), \tag{14}$$

which Appendix J.2 uses as the prespecified failure score. Appendix I defines the update norms and controls used to compare these effects at matched magnitude.

For the base source–host analyses, descriptive 95% intervals use 2,000 paired recipient/source-video bootstrap draws with donors held fixed; later appendices state their own draw counts and stratification. All donor-based intervals in these appendices condition on the fixed donor pool and allocation and therefore exclude uncertainty from selecting new donors. One-source groups have no interval, and failure and success groups are analyzed separately. Comparisons use common valid assignments across all four cells; a recipient without either competing label has no primary E. Invalid or capped intervention answers remain in correctness denominators, while verified finite logits remain eligible for margin analysis. Correctness always uses generated answers.

H Effects of Transplanted V-only States

This appendix reports effects by dataset and split, separates the supplied Values from attention adaptation, and checks sensitivity to donor choice.

H.1 Own-video correction by dataset and split

V-only-to-Full transfer raises the mean gold margin in every group, whereas the reverse transfer lowers it (Table 23). Generated-answer corrections are 26/40 on CMU-MOSEI validation, 14/26 on CMU-MOSEI test, and 48/90 on MELD training. Reverse transfer makes 17/40 validation answers incorrect.

All three validation success controls retain correct answers in both transfer directions. Larger success groups appear in Appendix J under a different selection rule.

Table 23: Own-video state transplantation by dataset and split. Entries are within-label mean gold margins before and after suffix Value replacement, with host inputs fixed; Δg is computed before rounding. CMU-MOSEI negative groups are small.

	Gold margin ↑: higher favors the correct label					
	V-only states → Full			Full states → V-only		
Gold label (n)	Before	After	Δg	Before	After	Δg
MOSEI validation						
Negative (2)	-2.806	-0.969	$+1.836$	0.707	-2.976	-3.683
Neutral (30)	-2.257	0.682	$+2.940$	2.275	0.737	-1.538
Positive (8)	-2.432	-0.373	$+2.059$	0.714	-1.583	-2.297
MOSEI supplemental test						
Negative (3)	-2.790	-0.420	$+2.370$	0.545	-2.356	-2.901
Positive (23)	-2.141	0.055	$+2.197$	1.262	-1.284	-2.546
MELD train replication						
Negative (30)	-3.470	-0.631	$+2.839$	0.300	-1.723	-2.023
Neutral (30)	-2.136	0.234	$+2.370$	0.801	0.016	-0.784
Positive (30)	-2.468	-0.101	$+2.367$	1.510	-1.062	-2.572

H.2 Effect under fixed native attention

Adaptive transfer moves attention from the suffix toward the input modalities (Table 24). The table averages over heads and layers and counts how widely each change occurs across clips and layers.

Table 24: Final-query attention under adaptive transfer minus native Full, in percentage points. Clip counts cover 156 failures. The last column counts layers where a majority of clips change in the direction of the mean, among layers 1–27; layer 0 is unchanged. Intervals use 2,000 recipient bootstrap draws.

Segment	Change (pp), 95% CI	Clips: up / down	Layers in mean direction
Video	$0.34\ [0.32, 0.36]$	156 / 0	24
Audio	$0.40\ [0.37, 0.42]$	156 / 0	25
Transcript	$0.35\ [0.30, 0.39]$	139 / 17	19
Suffix	$-1.48\ [-1.56, -1.40]$	0 / 156	18

To separate this redistribution from the supplied Values, the clamp fixes every final-query attention row to its native Full weights. Let $o_{F,q^\star}$ be the native pre-projection output and $\alpha_{F,q^\star j}$ its native attention weights, reconstructed in FP32. With the suffix alignment π of Equation 4, the clamped output is

$$\widetilde{o}_{F,q^\star} = \mathrm{BF16}\Big[o_{F,q^\star} + \alpha_{F,q^\star q^\star}(\widetilde{u}_{q^\star} - u_{F,q^\star}) + \sum_{j \in \mathcal{S}_r} \alpha_{F,q^\star j}(u_{V,\pi(j)} - u_{F,j})\Big], \tag{15}$$

where $\widetilde{u}_{q^\star}$ is the final query's own Value in the intervened run. This self Value and later computation may still evolve; the attention allocation itself is fixed at every layer. Identity controls reproduce their anchors, and the clamp reproduces native Full attention weights exactly.

Table 25: Gold-margin gains under fixed attention. Retained is the ratio of mean clamped to mean adaptive gain; the last column is the paired benefit of adaptation. Brackets are 95% recipient bootstrap intervals from 2,000 draws.

Gold label (n)	Clamped gain	Retained	Adaptive − clamped
Negative (35)	$2.125\ [1.707, 2.537]$	77.5%	$0.616\ [0.510, 0.724]$
Neutral (60)	$1.852\ [1.637, 2.059]$	69.8%	$0.803\ [0.705, 0.907]$
Positive (61)	$1.696\ [1.444, 1.971]$	75.0%	$0.566\ [0.491, 0.640]$

Clamped gains remain positive in every recipient. Adaptation increases the gain in 155/156 cases; the single tie is a positive CMU-MOSEI test clip with gain 0.248 in both conditions. Thus the supplied Values remain effective under native attention, and adaptation usually strengthens them.

H.3 Dependence on source context

Table 26 crosses both source contexts with both hosts on the same 156 recipients and four donors each. V-only sources produce larger own-versus-donor contrasts in every dataset, split, and label group. Figure 4 pools clips equally within label; its intervals use 2,000 recipient resamples stratified by dataset, split, and label with donors fixed. Label comparisons therefore also reflect differences in dataset composition. Own-state results are checked against Appendix H.1, and identity controls must reproduce native logits exactly.

Table 26: Own-video minus competing-label donor gold-margin contrasts on the matched 156 recipients. In $E(s,h)$, s is the source condition and h the host condition; V denotes V-only and F denotes Full.

Dataset / split	Gold label (n)	$E(V,V)$	$E(F,V)$	$E(V,F)$	$E(F,F)$
MOSEI / test	Negative (3)	$+3.017$	$+0.185$	$+2.140$	$+0.231$
MOSEI / test	Positive (23)	$+2.546$	$+0.201$	$+2.048$	$+0.212$
MOSEI / validation	Negative (2)	$+3.307$	-0.260	$+1.589$	-0.184
MOSEI / validation	Neutral (30)	$+1.348$	-0.058	$+1.589$	-0.078
MOSEI / validation	Positive (8)	$+2.078$	$+0.177$	$+1.812$	$+0.083$
MELD / train	Negative (30)	$+2.012$	$+0.163$	$+1.968$	$+0.201$
MELD / train	Neutral (30)	$+0.669$	-0.033	$+1.084$	-0.046
MELD / train	Positive (30)	$+2.154$	$+0.113$	$+1.656$	$+0.144$

Two checks on the CMU-MOSEI validation set vary donor choice. Each allocation separately gives positive label means of G_F, ranging from 1.60 to 1.94, although one neutral recipient has a negative contrast under allocation 1 (-0.151). Requiring adapted donors to remain V-correct leaves positive label means of 1.76–1.82. Five neutral failures lack a qualifying donor from one competing label, leaving 38/43 recipients for this check, including all three successes. These exclusions affect only the sensitivity analysis; all selected recipients remain in the primary analysis.

I Magnitude-Controlled Comparisons

We control update magnitude when comparing source contexts and when comparing own-video with class-prototype directions.

All controls use the 156 Qwen2.5-Omni Full failures from CMU-MOSEI and MELD, their frozen donor assignments, the 19 suffix positions, and layers 0–27, and change only the final-query attention output. Let $U^{\text{own}}_{s,l}$ and $U^{\text{don}}_{s,l}$ denote own- and donor-video suffix Values formed under source condition $s \in \{V, F\}$, and let α^F_l and $W_{O,l}$ denote native Full final-query attention on the suffix and the attention output projection. The reference budget is

$$\rho_l = \min_{s\in\{V,F\}} \left\| W_{O,l}\alpha^F_l (U^{\text{own}}_{s,l} - U^{\text{don}}_{s,l}) \right\|_2 . \tag{16}$$

The magnitude-matched control preserves each own-video transplant but rescales its donor-minus-own update to ρ_l; its V-minus-Full contrast is G_F^{matched}. The equal-budget direction control instead starts from native Full and applies symmetric half-budget updates along the V-only- or Full-derived content direction. With C_s the gold margin after the $+$ half-budget update minus that after the $-$ update, the source comparison is $C_V - C_F$. All primary H100 checks reproduce native and identity endpoints, and independent reconstruction verifies update budgets and projections.

I.1 Source advantage at matched magnitude

Both controls leave the source difference positive in every label, with intervals above zero (Table 27). Magnitude-matched G_F is about one sixth of raw G_F in each label, and the equal-budget direction

from the same native Full state also favors V-only source states. Appendix J.1 applies the same controls to new datasets and to Full successes.

Table 27: Source controls on the Qwen2.5-Omni failure sample. Positive values favor V-only source states. Matched G_F keeps each own-video transplant fixed and matches the donor update magnitude; $C_V - C_F$ compares equal-budget directions from the same native Full state. Intervals are descriptive 95% paired recipient/source-cluster bootstrap intervals with donors fixed.

Gold label (n)	Raw G_F	Matched G_F	$C_V - C_F$
Negative (35)	$1.780\ [1.581, 1.992]$	$0.283\ [0.226, 0.340]$	$0.280\ [0.211, 0.349]$
Neutral (60)	$1.398\ [1.230, 1.567]$	$0.243\ [0.179, 0.305]$	$0.426\ [0.356, 0.492]$
Positive (61)	$1.663\ [1.507, 1.817]$	$0.257\ [0.197, 0.322]$	$0.276\ [0.230, 0.323]$

I.2 Class-prototype directions and the own-video residual

The equal-budget operator compares the own-video direction with the gold-minus-competing directions of the class prototypes in Appendix K.2. To measure the effect remaining outside their span, we remove the least-squares projection onto the two prototype contrasts and renormalize the residual to ρ_l. All directions start from the same native Full state.

Table 28: Gold-margin effects at equal update budget. Negative own-minus-prototype contrasts favor the prototype. Brackets give descriptive 95% paired bootstrap intervals with donors fixed; the residual is renormalized to the original budget.

Gold label (n)	Own video	Own − prototype	Residual
Negative (35)	0.468	$-0.172\ [-0.225, -0.124]$	$0.043\ [-0.001, 0.088]$
Neutral (60)	0.352	$-0.171\ [-0.216, -0.128]$	$0.083\ [0.053, 0.112]$
Positive (61)	0.444	$-0.092\ [-0.119, -0.067]$	$0.092\ [0.067, 0.117]$

Prototype effects exceed own-video effects in every label (Table 28). The residual is smaller, with pooled intervals above zero for neutral and positive. By dataset, its interval excludes zero for CMU-MOSEI neutral and positive and MELD negative, but spans zero for CMU-MOSEI negative ($n = 5$) and MELD neutral and positive. These are effects outside the tested prototype span; renormalization prevents an additive decomposition of the original effect.

J Replication on New Data and Full Successes

We test Qwen2.5-Omni and MiniCPM-o on CMU-MOSI and CH-SIMS v2.0 (the expanded CH-SIMS dataset, abbreviated CH-SIMS v2). Recipients are V-correct and grouped by native Full outcome, without A/T correctness restrictions. These studies contain 360 model–example pairs: 65 per model on CMU-MOSI, and 116 Qwen2.5-Omni and 114 MiniCPM-o recipients on CH-SIMS v2.

Candidate clips require valid audio/video and duration 0.8–20 seconds. CMU-MOSI excludes sources used in CMU-MOSEI screening or experiments. CH-SIMS v2 excludes all 60 original CH-SIMS source IDs and selects up to 200 clips per label, at most 12 per source, in fixed hash order. Distinct source IDs can still share speakers or series. Before recipient Full inference, each model reserves up to three V-correct training donors per class from distinct sources and excludes those sources from recipients. Remaining V-correct clips enter Full failure or success groups in fixed order, capped at 20 per label and outcome on CMU-MOSI and 30 on CH-SIMS v2, with at most two clips per source per outcome. Assignments precede interventions.

The neutral confirmatory target for CH-SIMS v2 was fixed before any outputs on that dataset. This study addresses the limited sample coverage found during CMU-MOSI screening and was designed before inspecting CMU-MOSI intervention or prediction results. The datasets are analyzed separately.

J.1 RAW AND MAGNITUDE-CONTROLLED SOURCE EFFECTS

Each recipient receives one donor per competing label, selected by absolute log-duration distance with a fixed hash tie-breaker. The design crosses V-only and Full sources with both hosts using Equation 4. Donor frames match recipient geometry; Full sources retain recipient audio and transcript. Qwen2.5-Omni uses 19 suffix tokens and 28 layers; MiniCPM-o uses 23 tokens and 36 layers. Donor effects are averaged equally across the two competing labels.

All eight neutral groups have positive raw G_F with intervals above zero (Table 29). Own V-only states correct 16/20 and 9/20 neutral CMU-MOSI failures for Qwen2.5-Omni and MiniCPM-o, respectively, and 21/30 and 3/30 on CH-SIMS v2. All neutral successes remain correct.

Table 29: Raw source advantage $G_F = E(V, F) - E(F, F)$ in neutral groups. n_f/n_s counts failures/successes; MOSI denotes CMU-MOSI. Brackets are descriptive 95% source-cluster bootstrap intervals.

Dataset	Model	n_f/n_s	G_F: failures	G_F: successes
MOSI	Qwen2.5	20/20	$1.031\ [0.649, 1.463]$	$1.126\ [0.960, 1.319]$
MOSI	MiniCPM-o	20/20	$0.927\ [0.391, 1.497]$	$1.059\ [0.688, 1.485]$
CH-SIMS v2	Qwen2.5	30/30	$1.155\ [0.951, 1.353]$	$1.098\ [0.904, 1.315]$
CH-SIMS v2	MiniCPM-o	30/24	$1.522\ [0.945, 2.080]$	$0.725\ [0.410, 1.097]$

The same recipients and donors also receive both magnitude controls of Appendix I, keeping the labels and Full outcome groups from screening. The controls were designed after the raw effects were known, with their settings fixed before controlled inference. Native, raw, and controlled conditions run on the same H100. The operators retain their distinct starting states, and later attention can adapt. Within each group, intervals use 5,000 source-cluster bootstrap draws with donors fixed, retaining all clips from each sampled source. Intervals are pointwise; one-source groups have none. Identity and zero-dose runs reproduce native logits and tokens, and independent reconstruction verifies layer updates and projections.

Both controlled contrasts remain positive with intervals above zero in all eight neutral groups (Tables 30 and 31). Non-neutral groups are smaller and less consistent, including negative estimates and intervals spanning zero. The tables retain all 20 observed groups.

Table 30: Magnitude controls on Qwen2.5-Omni. n/k counts recipients/source clusters; Full refers to the outcome group from screening. Cells show means and descriptive 95% source-cluster intervals. Raw effects are measured alongside the controls. A dash denotes an unavailable interval, not zero effect.

Label	Full	n/k	Raw G_F	Matched G_F	$C_V - C_F$
CMU-MOSI					
Negative	success	6/5	$2.432\ [1.991, 2.962]$	$0.376\ [0.280, 0.504]$	$0.274\ [0.221, 0.332]$
Neutral	failure	20/19	$1.015\ [0.643, 1.435]$	$0.175\ [0.096, 0.268]$	$0.209\ [0.122, 0.315]$
Neutral	success	20/17	$1.134\ [0.973, 1.314]$	$0.111\ [0.063, 0.157]$	$0.174\ [0.122, 0.233]$
Positive	failure	5/4	$2.207\ [1.827, 2.656]$	$0.296\ [0.196, 0.354]$	$0.198\ [0.111, 0.309]$
Positive	success	14/10	$2.076\ [1.826, 2.339]$	$0.315\ [0.247, 0.387]$	$0.301\ [0.191, 0.415]$
CH-SIMS v2					
Negative	failure	1/1	$0.887\ [--]$	$0.023\ [--]$	$-0.121\ [--]$
Negative	success	25/22	$3.056\ [2.748, 3.354]$	$0.664\ [0.565, 0.754]$	$0.629\ [0.541, 0.705]$
Neutral	failure	30/24	$1.172\ [0.977, 1.362]$	$0.368\ [0.251, 0.491]$	$0.483\ [0.354, 0.616]$
Neutral	success	30/25	$1.111\ [0.928, 1.305]$	$0.180\ [0.090, 0.285]$	$0.235\ [0.149, 0.333]$
Positive	failure	4/4	$1.605\ [1.327, 1.916]$	$0.367\ [0.272, 0.543]$	$0.253\ [0.136, 0.341]$
Positive	success	26/22	$2.257\ [2.019, 2.460]$	$0.399\ [0.301, 0.486]$	$0.340\ [0.234, 0.441]$

Numerical stability checks on the H100 reproduce the initial tokens in 2,145 of 2,160 source runs, with maximum label-logit discrepancy 1.125. Six recipients change V-only correctness and three change Full correctness. The primary analysis keeps their screening groups; excluding all nine still leaves both controls' intervals above zero in every neutral group.

Table 31: The same magnitude controls on MiniCPM-o. Positive source differences favor V-only states. All observed groups are shown, including small groups and intervals spanning zero.

Label	Full	n/k	Raw G_F	Matched G_F	$C_V - C_F$
CMU-MOSI					
Negative	success	2/2	$1.914\ [1.413, 2.416]$	$0.033\ [-0.242, 0.307]$	$-0.458\ [-0.624, -0.293]$
Neutral	failure	20/18	$0.903\ [0.396, 1.398]$	$0.444\ [0.203, 0.690]$	$0.414\ [0.209, 0.639]$
Neutral	success	20/16	$1.046\ [0.666, 1.478]$	$0.491\ [0.255, 0.782]$	$0.297\ [0.126, 0.495]$
Positive	failure	3/3	$1.988\ [-0.070, 3.283]$	$0.459\ [-0.927, 1.252]$	$0.373\ [-0.230, 0.725]$
Positive	success	20/14	$1.463\ [0.902, 2.020]$	$0.380\ [0.136, 0.617]$	$0.009\ [-0.186, 0.188]$
CH-SIMS v2					
Negative	success	30/25	$1.617\ [1.144, 2.032]$	$0.393\ [0.212, 0.544]$	$0.124\ [-0.148, 0.342]$
Neutral	failure	30/24	$1.540\ [0.970, 2.085]$	$1.010\ [0.613, 1.398]$	$1.147\ [0.787, 1.509]$
Neutral	success	24/19	$0.685\ [0.349, 1.056]$	$0.486\ [0.215, 0.767]$	$0.266\ [0.045, 0.502]$
Positive	success	30/24	$1.854\ [1.298, 2.416]$	$0.794\ [0.566, 1.027]$	$0.054\ [-0.179, 0.254]$

J.2 Prespecified failure prediction

We test whether G_V predicts Full failure beyond unimodal confidence and input length. This score uses a V-only host (Equation 14), so native Full logits and correctness do not directly enter it. Association is the failure-minus-success mean G_V difference. CMU-MOSI averages it equally over labels with at least ten recipients per outcome; only neutral qualifies. CH-SIMS v2 specifies neutral alone. Prediction uses all CMU-MOSI recipients and only neutral CH-SIMS v2 recipients.

A logistic regression with L_2 regularization ($C = 1$) uses the V-predicted class, V/A/T gold margins and top-two gaps, log duration, and log transcript character count. All recipients are V-correct, so the gold label equals the V prediction. The augmented model adds G_V. Five-fold stratified cross-validation holds out source videos, with standardization fitted on training folds. Intervals use 5,000 source bootstrap draws of fixed out-of-fold predictions, excluding refitting uncertainty. AUROC retains the prespecified score direction, and outcome quotas make these comparisons conditional on outcome.

Support requires all three prespecified criteria: at least 30 recipients and 15 sources per outcome (and two eligible labels on CMU-MOSI); a positive one-sided 95% association lower bound; and log-loss improvement at least 0.01 with a positive one-sided 95% lower bound. CH-SIMS v2 applies these sample thresholds to neutral recipients. Confirmation across models requires both to pass within the same study.

No model and dataset combination meets all criteria (Tables 32 and 33). MiniCPM-o on CH-SIMS v2 improves prediction, but falls short on sample size and association. The studies have separate targets and are not pooled.

Table 32: Failure association and out-of-fold prediction. n_f/n_s and k_f/k_s count failure/success recipients and sources for prediction; CMU-MOSI association uses only its eligible neutral label. LB is the one-sided 95% association lower bound. Positive ΔAUROC and Δlog loss favor adding G_V. Intervals use source bootstrap draws of fixed predictions.

Dataset / Model	n_f/n_s	k_f/k_s	Association [95% CI]	LB	AUROC: G_V
MOSI / Qwen2.5	25/40	23/29	$-0.159\ [-0.512, 0.217]$	-0.450	0.425
MOSI / MiniCPM-o	23/42	21/31	$-0.393\ [-1.919, 1.139]$	-1.693	0.422
CH-SIMS v2 / Qwen2.5	30/30	24/25	$-0.222\ [-0.473, 0.021]$	-0.429	0.392
CH-SIMS v2 / MiniCPM-o	30/24	24/19	$0.664\ [-0.174, 1.400]$	-0.012	0.636

Dataset / Model	AUROC: base $\rightarrow +G_V$	ΔAUROC [95% CI]	Δlog loss [95% CI]
MOSI / Qwen2.5	$0.9460 \rightarrow 0.9480$	$0.002\ [-0.011, 0.015]$	$0.014\ [-0.002, 0.034]$
MOSI / MiniCPM-o	$0.9172 \rightarrow 0.9193$	$0.002\ [0.000, 0.010]$	$-0.003\ [-0.009, 0.001]$
CH-SIMS v2 / Qwen2.5	$0.9589 \rightarrow 0.9533$	$-0.006\ [-0.018, 0.004]$	$0.002\ [-0.018, 0.022]$
CH-SIMS v2 / MiniCPM-o	$0.7625 \rightarrow 0.8903$	$0.128\ [0.051, 0.211]$	$0.163\ [0.070, 0.247]$

Table 33: Prespecified failure-prediction decisions. Sample, association, and incremental log-loss criteria are defined above; all must pass. The source–host replication result is evaluated separately.

Dataset	Model	Sample	Association	Increment	Decision
MOSI	Qwen2.5	Fail	Fail	Fail	Insufficient sample
MOSI	MiniCPM-o	Fail	Fail	Fail	Insufficient sample
CH-SIMS v2	Qwen2.5	Pass	Fail	Fail	Not supported
CH-SIMS v2	MiniCPM-o	Fail	Fail	Pass	Insufficient sample

K Cross-Video Transfer and Class Prototypes

Same-label donors and class prototypes test transfer across videos. Both use gold labels and are oracle diagnostics. Equal-budget direction comparisons appear in Appendix I.2.

K.1 Same-label sources

Each of the 156 Qwen2.5-Omni failures receives two fixed same-dataset, same-label V-only donors selected without intervention outcomes. The source is read into the recipient's unchanged Full host through the same suffix operator, and donor outcomes are averaged within recipient before label means and paired intervals are computed, using 1,000,000 recipient resamples with seed 20260924. Mean same-label gains differ from own-video gains by at most 0.13, and all three own-minus-same-label intervals include zero (Table 34). In generated answers, same-label sources correct 26/70, 97/120, and 59/122 donor interventions for negative, neutral, and positive failures, against 6/35, 47/60, and 35/61 for the own video. In 29 of the 468 intervention records, the gold label ties another label at the maximum bfloat16 logit, and greedy decoding then emits the label with the lower token ID (positive, then negative, then neutral); counting these ties as uncorrected gives 21/70, 97/120, and 47/122 same-label and 6/35, 47/60, and 32/61 own-video corrections.

Table 34: Cross-video transfer on 156 Qwen2.5-Omni Full failures. Gains are changes from native Full; positive values favor the gold label. Prototype selectivity compares each injected label with the other two at the own-source update budget (bank A). Brackets give descriptive 95% paired intervals for own minus same-label sources from 1,000,000 recipient resamples stratified by dataset and split.

Gold label (n)	Own gain	Same-label gain	Own − Same-label	Prototype S
Negative (35)	2.741	2.777	-0.036 $[-0.148, 0.066]$	1.821
Neutral (60)	2.655	2.718	-0.063 $[-0.227, 0.098]$	1.914
Positive (61)	2.262	2.136	0.127 $[-0.004, 0.258]$	1.815

K.2 Class-prototype selectivity

If part of the corrective influence is shared across videos with the same label, states averaged over several such videos should favor that label. For each label, two prototype banks average V-only suffix Values from eight fixed training videos. The MELD banks are disjoint; the CMU-MOSEI banks share 5, 1, and 2 of their eight negative, neutral, and positive videos, because only 12 negative CMU-MOSEI donors are available. Every recipient receives every label prototype at raw, half-matched, and full-matched doses, where the matched doses scale the concatenated-head update before output projection to 0.5 or 1 times the own-source norm. For logits $z_i(c)$ after injecting class c, selectivity is

$$S_i = \frac{1}{3}\sum_c \left[g(z_i(c), c) - \frac{1}{2}\sum_{k \neq c} g(z_i(k), c) \right]. \tag{17}$$

Selectivity is positive for all 156 recipients in both banks and all three doses. At the full matched dose, bank-A label means are 1.821, 1.914, and 1.815, and bank-B means are 1.631, 1.669, and 1.682.

L SECOND-MODEL REPLICATION AND MATCHED SOURCES

We repeat the transfer tests with MiniCPM-o 4.5 and compare both models' own-video sources with same-label sources matched on logits and update norms.

L.1 MINICPM-O REPLICATION

MiniCPM-o 4.5 uses its native 23-token suffix and all 36 layers. The analysis includes 152 Full failures from CMU-MOSEI and MELD (34 negative, 60 neutral, and 58 positive) and three success controls per label. Native prediction checks on the H100 exclude three invalid recipients and one donor. A-only and T-only eligibility uses the initial screening predictions; V-only and Full predictions are checked on the intervention hardware. Recipient and donor source videos are disjoint. All intervention records are valid and uncapped, and reference endpoints reproduce exactly.

Own-video and same-label sources raise the gold margin in every label, and prototype selectivity is positive in every label (Table 35). The own-minus-same-label interval excludes zero only for negative failures, where it favors same-label sources. The nine success controls retain correct answers under own and same-label sources; their size supports only a descriptive check.

Table 35: MiniCPM-o replication. Gains are relative to native Full; brackets give descriptive 95% paired intervals. Corrected counts generated answers that become correct. Prototype S uses the own-source norm.

Gold label (n)	Own gain	Same-label gain	Own − Same-label	Corrected	Prototype S
Negative (34)	1.949	2.319	-0.369 $[-0.619, -0.123]$	14/34	3.031
Neutral (60)	4.391	4.363	0.028 $[-0.271, 0.304]$	14/60	3.187
Positive (58)	2.275	2.367	-0.092 $[-0.295, 0.105]$	30/58	3.294

L.2 MATCHED-SOURCE CONTROL

Own and same-label sources can differ in source logits and update norms, so own sources are compared with same-label candidates matched on both. Matching uses source-logit RMSE at most 0.35, layerwise log update-norm-ratio RMS at most $\log(1.25)$, and every norm ratio in $[0.5, 2]$, and selection precedes candidate intervention outcomes. Qwen2.5-Omni searches up to 24 same-label candidates and matches 131/156 failures: 32/35 negative, 52/60 neutral, and 47/61 positive. Own-minus-matched mean gains are 0.022 $[-0.054, 0.101]$, -0.008 $[-0.094, 0.074]$, and 0.036 $[-0.030, 0.101]$. MiniCPM-o matches 49/152 failures, with corresponding differences of -0.142 $[-0.248, -0.069]$, -0.034 $[-0.208, 0.128]$, and -0.040 $[-0.159, 0.087]$. After matching, all three Qwen2.5-Omni intervals include zero, whereas MiniCPM-o's same-label advantage for negative failures remains.