DISAGREEMENT IS NOT DEGRADATION: EVALUATING RELIABILITY IN MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal reliability evaluations can conflate annotation disagreement, input degradation, task utility, and output-format compliance. We audit these distinctions with three omni-modal models: a 21,714-request, seven-view evaluation on 1,034 CH-SIMS v2 clips, followed by paired raw-media transformations and fixed MELD and CMU-MOSEI cohorts. Adding a transcript to audio–visual input increases Qwen2.5-Omni's error on the annotation-disagreeing CH-SIMS subset, while the corresponding MiniCPM-o estimate remains uncertain. A training-median control reveals different target distributions: none of the three model-minus-prior intervals on the disagreeing subset excludes zero. Qwen3-Omni's format failures further separate schema compliance from task performance. Narrow syntax rules developed post hoc on the original run recover its malformed numerical responses; fixed before the follow-up, complete-fence normalization also restores high coverage on the external tasks. Additional modalities do not yield uniform gains: on the locked MELD cohort, MiniCPM-o's text accuracy exceeds its text–audio–visual accuracy by 9.96 percentage points despite complete strict coverage. Paired physical transformations have nonuniform predictive effects, and an earlier matched-supervision feature study does not support a factorized reliability gate. The evidence supports joint reporting of absolute error, paired input effects, train-only prior controls, output coverage, and distinct failure causes. It does not establish a human-validated perceptual-quality effect or a superior fusion architecture.

1 INTRODUCTION

A multimodal sentiment system can receive clearly expressed signals that do not agree. It can also receive damaged signals that happen to agree. A single notion of reliability risks confusing four questions: do modality-specific annotations agree, has the observed input been degraded, is an input useful for an overall target, and does the model satisfy an output contract? An unimodal label is not a physical-quality label, and a malformed response is not sufficient evidence that its sentiment estimate is wrong.

Omni-modal models allow comparisons of different input combinations on the same clip. An audio–visual prediction and one that additionally receives a transcript share their target, but their difference does not identify an internal conflict-resolution mechanism. Speech already conveys words, frames can contain subtitles, and model-specific processors allocate different token budgets. Natural annotation groups can also have different sentiment-intensity distributions. Comparing group errors without a target-only reference can give a misleading account of task difficulty.

Prior research already studies emotion conflict and missingness, perceptual versus propositional evidence, modality noise, and reliability-aware fusion (Sun et al., 2026; Piao et al., 2026; Mao et al., 2022; Meng et al., 2026). Our contribution is a bounded empirical audit of how these measurements interact in a continuous-sentiment setting and two external tasks. We retain an unfavorable architecture result, distinguish format adherence from recovered numerical predictions, and test a train-only prior before interpreting group differences. Neither multiple models nor additional corruption conditions alone constitute a novel benchmark claim.

The completed evidence has three components: a frozen three-model, seven-view evaluation; a transparent post-hoc syntax and target-distribution audit; and an earlier 39-fit feature study that fails

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.

to support a proposed factorization benefit under matched supervision. A paired raw-media extension tests physical transformations, and locked MELD/MOSEI cohorts examine whether format and input-utility distinctions persist in additional tasks. We distinguish complete execution from valid responses, and both from evidence of improved task performance.

2 TARGETS AND ESTIMANDS

For clip i, let $x_i = (x_i^T, x_i^A, x_i^V)$ denote transcript, audio and sampled visual frames. Human unimodal labels are y_i^T, y_i^A, y_i^V; the overall sentiment target $y_i^M \in [-1, 1]$ is shared across all views. A frozen model produces a string whose strict parsing yields a prediction $p_{i,v}$ or an invalid flag $z_{i,v} = 0$.

Natural annotation strata. At $\tau = 0.2$, group D contains clips with at least one unimodal label $\geq \tau$ and another $\leq -\tau$. Group A has all three labels $\geq \tau$ or all three $\leq -\tau$; N contains the remainder. These human-label groups are never supplied to a model. They describe annotation disagreement, not verified degradation, sarcasm, deception, or a speaker's concealed emotion. N need not have a neutral overall target.

Coverage and paired input effects. We report valid-output MAE and strict all-intended risk separately:

$$R_v^{\text{valid}}(G) = \frac{\sum_{i \in G} z_{i,v} |p_{i,v} - y_i^M|}{\sum_{i \in G} z_{i,v}}, \tag{1}$$

$$R_v^{\text{strict}}(G) = \frac{1}{|G|} \sum_{i \in G} \left[z_{i,v} |p_{i,v} - y_i^M| + (1 - z_{i,v})\, 2 \right]. \tag{2}$$

The invalid penalty is the maximum possible absolute error on $[-1, 1]$; it defines an operational loss, not a standard MAE. A view contrast $\Delta_{v \to w} = R_w - R_v$ pairs the same clips under a named parsing policy. A small or negative change does not establish low absolute risk. Input deletion changes supplied evidence; it is not a validated corruption of retained evidence.

A target-only reference. The constant $b = \text{median}_{i \in \text{train}}\, y_i^M$ is fitted without test labels. We report its risk R_b and paired model excess risk $R_v - R_b$. This can reveal target-strength differences across groups, but does not adjust away all confounding. Here $b = 0$. We chose this control after observing the Omni test results; the analysis is exploratory.

3 FROZEN OMNI EVALUATION

Data version and dependence. CH-SIMS and CH-SIMS v2 provide modality-specific human sentiment annotations (Yu et al., 2020; Liu et al., 2022). Our acquired raw-media manifest contains 4,402 clips: 2,721 train, 647 validation and 1,034 test. This is a derived manifest: one training clip was removed from the acquired 4,403-clip release after a test-input duplicate audit. Scoring uses verified native feature-pickle M/T/A/V labels, not an interchangeable CSV release. The test set contains 459/149/426 clips in $A/D/N$, from 142 source videos overall. Distributed label files are not interchangeable; targets and groups are bound to a fixed scoring manifest. There are 129, 120 and 142 distinct train, validation and test sources, respectively. Train–validation, train–test and validation–test intersections contain 119, 127 and 117 sources. These are not source-disjoint partitions; source clustering does not repair cross-split dependence or certify unseen-speaker generalization. In the benchmark annotation protocol, acoustic labels were elicited with words obscured and visual labels without audio or subtitles. Their disagreement is defined by those restricted annotation conditions, whereas our original audio retains words and frames may retain subtitles (Liu et al., 2022).

Models and media. We evaluate official checkpoints of Qwen2.5-Omni-7B (Q25), MiniCPM-o-4.5 (Mini), and Qwen3-Omni-30B-A3B-Instruct (Q3). Each receives the same clip identities under T, A, V, TA, TV, AV, TAV and predicts the overall M target. The Chinese instruction is fixed, decoding is greedy, and the generation budget is 64 tokens. There is no task fine-tuning or repeated

Table 1: TAV scoring policies. Strict-valid MAE conditions on success; strict risk penalizes invalid outputs by 2. Post-hoc recovered MAE uses all 1,034 clips. These columns measure different estimands.

Model	Strict valid	Strict-valid MAE	Strict risk	Recovered MAE
Q25	1034/1034	.3342	.3342	.3342
Mini	1034/1034	.2674	.2674	.2674
Q3	750/1034	.3637	.8132	.3539

decoding seed. Audio is the complete 16-kHz mono waveform. Visual input uses up to 16 evenly spaced frames, approximately one per second, with long edge at most 448 pixels. Audio retains words and frames retain embedded subtitles. Native processors, time encodings and token counts differ; these are frozen pipeline comparisons, not equalized-compute or isolated-semantic-channel tests.

All $1,034 \times 3 \times 7 = 21,714$ requests were recorded, with no missing, OOM, timeout or unattempted cells. Q25 and Mini satisfy the strict output contract on all 7,238 requests each. Q3 has 6,252 strict-valid responses and 986 invalid responses. Invalid records are retained without retries.

Statistics and timing. Original comparisons use 2,000 source-cluster bootstrap draws, seed 0, shared across contrasts. Intervals condition on fixed weights, processors, clips, prompts and one decode; they do not include retraining or prompting variability. There is no family-wise correction. The post-hoc syntax analysis uses the same draw plan for its new estimand. New prior and target-bin analyses use 2,000 draws with seed 20260922 and remain exploratory. Group contrasts do not identify a causal effect of disagreement.

4 Completed results

4.1 Format compliance and numerical prediction differ

Q3's 986 strict-invalid outputs comprise 800 complete Markdown-fenced objects and 186 objects with one extra quote after an unquoted numeric value. They end normally after 10–15 generated tokens, below the 64-token limit. Restricted normalization removes a complete outer fence or that single unmatched quote, then reapplies the original schema, numeric and range checks. It changes no numeric value and reads no label. All three models are processed with the same rules.

Every response in this matrix becomes parseable, but this is not improved model instruction following. Recovery was designed after inspecting failures and remains a sensitivity analysis beside strict results. Arbitrary number extraction, truncation completion, range clipping and error-based selection are excluded.

Recovered TAV MAE intervals are .3342 [.3134,.3533] for Q25, .2674 [.2498,.2860] for Mini, and .3539 [.3288,.3813] for Q3. The paired Q25-minus-Q3 difference is $-.0197$ $[-.0460, .0056]$, which does not establish a difference under this analysis. Mini-minus-Q3 is $-.0865$ $[-.1114, -.0649]$. These fixed-setting comparisons are not a universal model-family ranking. Q3's strict-valid subset must not silently replace the intended cohort.

4.2 Adding evidence has model-dependent effects

Figure 1 shows recovered MAE across seven views. Q25's AV estimate (.3089) is lower than TAV (.3342); Mini's AV and TAV estimates are .2695 and .2674. On D, Q25's AV and TAV errors are .3792 and .4456. The prespecified $AV - TAV$ contrast is $-.0664$ $[-.1090, -.0317]$: adding the transcript worsens error here. Mini's corresponding contrast is $+.0215$ $[-.0133, .0609]$, leaving its direction uncertain. This does not show that text is intrinsically unreliable or that a particular attention mechanism causes Q25's change.

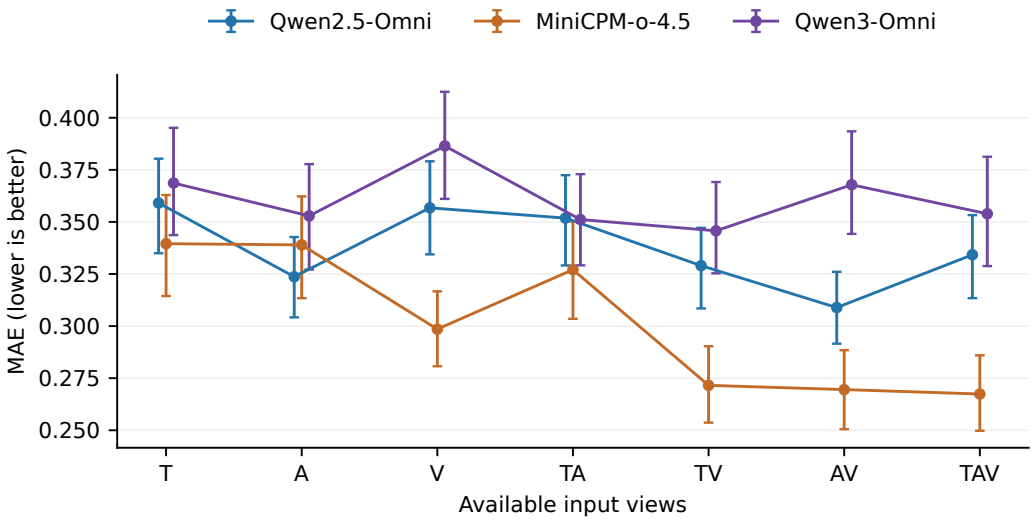

Figure 1: Seven views of the same 1,034 clips. Recovered MAE and conditional source-cluster 95% intervals; the Q3 estimand is post hoc. More supplied modalities do not guarantee lower target error.

Table 2: Exploratory prior control. TAV recovered MAE; the constant uses training labels only. N denotes residual annotation patterns, not neutral M.

Group	Clips	Train median	Q25	Mini	Q3
A	459	.6776	.3218	.2115	.3190
D	149	.3960	.4456	.3872	.4591
N	426	.3601	.3087	.2857	.3548

4.3 A PRIOR CONTROL CHANGES THE INTERPRETATION OF GROUPS

Mini's TAV MAE is .2115 in A and .3872 in D. This associates disagreement with higher error, but the target-only reference gives a necessary qualification. The train-median prior has MAE .6776 in A and .3960 in D (Table 2), showing different target-strength distributions. The prior is easier to beat in A.

On D, model-minus-prior differences are $+.0497$ $[-.0055, .1119]$ for Q25, $-.0087$ $[-.0833, .0710]$ for Mini, and $+.0631$ $[-.0127, .1507]$ for Q3. None excludes zero. This is insufficient evidence that any of the three improves on the train-only prior in this subgroup; it is not an equivalence test and does not prove an absence of emotional reasoning. All three improve on the overall prior MAE (.5062), with paired intervals excluding zero in this exploratory analysis.

We retain every group-by-target cell using the training nonzero $|M|$ median .6 as cutoff, with exact zero separate. The A zero cell has one clip; the strong-negative D cell has six. These sparse strata do not justify a stable adjusted disagreement effect. No favorable bins are selected or promoted to independent confirmation.

4.4 MATCHED FEATURE CONTROLS DO NOT SUPPORT FACTORIZATION

An earlier supervised study tests whether multiplying a utility gate by a separate intervention detector helps prediction. Thirteen variants and three seeds (39 fits) control auxiliary labels, augmentation, gradient paths and approximate parameter budget. Mean/std pooling and small MLPs operate on released features, without reproducing full prior temporal encoders. The main factorized and utility variants share detector BCE, unimodal supervision and optimization; only factorization inserts the detector into fusion. Seen-condition MAEs are .2742 and .2718, respectively. The

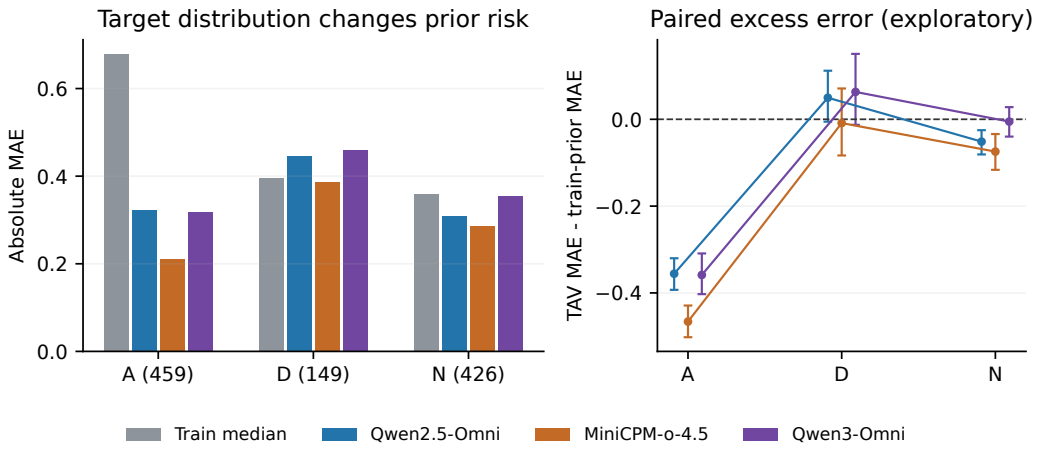

Figure 2: Absolute risk and excess risk relative to the training-median constant. Target distributions differ across natural groups. Exploratory intervals overlapping zero do not establish equivalence.

factorized-minus-utility contrast in paired error increments is $+.000768$ $[.000075, .001482]$, with seed directions differing. This does not support the proposed factorization benefit.

On D, factorized original MAE is .4631 and all-missing MAE is .3980. Information removal can lower error when a target distribution lies close to the model's prior; a favorable increment alone obscures original error. These supervised feature results cannot be fairly ranked against zero-shot raw-media models. The complete feature protocol and all predefined ablations appear in Appendices I and L; this negative result supplies supporting evidence, not a new architecture claim.

5 Related Work and Remaining Distinction

EmoMM also uses CH-SIMS v2 unimodal annotations, conflict/missingness settings and sentiment error metrics, with CHASE steering (Sun et al., 2026). Tri-PvP crosses perceptual and propositional evidence in vision and audio, constructs three-way conflicts, and judges free-form modality preferences (Piao et al., 2026). Natural annotation strata and continuous overall sentiment measure a different object; our study does not isolate evidence form as Tri-PvP does. We claim neither the first conflict benchmark nor the first observation of input preference.

Robust-MSA studies modality noise and NIAT studies noise-based training (Mao et al., 2022; Yuan et al., 2023). CONFER uses conflict-aware evidence negotiation; MRCF addresses incomplete observations (Hou et al., 2026; Meng et al., 2026). Our small matched comparison tests one factorization hypothesis, rather than refuting those systems. Any contribution must rest on reproducible interactions among absolute risk, priors, paired effects and output coverage. Applying blur or noise alone is insufficient.

6 Paired Raw-Media Follow-Up

The follow-up crosses the fixed annotation strata with four additional transformations of the same TAV inputs: audio noise at 20/10 dB and frame blur at $\sigma = 1/2$ pixels. Clean references are rerun within this campaign; labels and untouched modalities stay fixed. Appendix E specifies physical checks, cohorts, failure accounting and frozen statistics. The design was informed by the earlier test results and is not an independent preregistered study. No completed human-quality study is available; a 28-task author pilot is disclosed in Appendix E.

Coverage boundary. The final sealed follow-up contains 27,588 raw responses, 12 predeclared data restrictions and six interrupted requests out of 27,606 intended cells; no unattempted cell remains. A response is not necessarily a valid prediction. All three models retain the two physically

Table 3: CH-SIMS test follow-up. Each condition intends 1,034 clips. Entries are valid-only MAE / all-intended risk. Q25 and Mini have identical statistics across all parsing tiers; Q3 here uses fence-plus-quote normalization. The two rejected 10-dB clips remain in each denominator. Q3 additionally retains one clean and two stronger-blur interruptions. All strict and fence-only results and valid counts are in Appendix F.

Condition	Q25	Mini	Q3 (normalized)
Clean	.3342 / .3342	.2674 / .2674	.3542 / .3558
Audio 20 dB	.3394 / .3394	.2662 / .2662	.3519 / .3519
Audio 10 dB	.3390 / .3422	.2668 / .2701	.3530 / .3561
Blur $\sigma = 1$	.3408 / .3408	.2721 / .2721	.3527 / .3527
Blur $\sigma = 2$	.3480 / .3480	.2920 / .2920	.3560 / .3591

rejected 10-dB audio clips; the other six data restrictions concern MOSEI audiovisual correspondence. Operational penalties for data restrictions and interruptions describe execution coverage, not observed semantic errors on those requests.

All conditions and parsing sensitivity. Table 3 reports every assigned condition. At $\sigma = 2$, paired all-intended risk increments are .0137 [.0069,.0214] for Q25 and .0246 [.0123,.0371] for Mini, relative to each model's new clean reference. Several other condition intervals overlap zero. Thus, the measured physical interventions do not uniformly worsen target error. On D, Mini's stronger blur increment is .0564 [.0254,.0856], while Q25's is .0174 [$-.0036$,.0389]. Separate group intervals do not establish a group-by-condition interaction. All twelve condition contrasts, including uncertain and negative estimates, are retained at every parsing tier in the appendix. These are fixed-pipeline predictive stress results, not proof of a conflict-resolution mechanism or a human-validated quality scale.

For Q3, stronger blur has a strict operational-risk change of $-.1040$ [$-.1638, -.0450$], while complete-fence plus numerical-quote normalization gives a risk change of .0034 [$-.0092$,.0152]. The latter valid-pair MAE change is .0018 [$-.0099$,.0123] on 1,031 pairs. Strict valid counts rise from 750 for clean to 818 for stronger blur; after normalization they are 1,033 and 1,032. An apparent improvement under the strict penalty therefore does not establish improved sentiment inference. New and historical clean predictions are compared separately; the new clean reference is never selected according to which makes a transformation look favorable.

6.1 External cohorts separate format and task utility

The locked MELD and MOSEI cohorts each contain 512 clips, evaluated under T, AV and TAV. MELD has 4,608 responses; MOSEI has 4,602 responses and six predeclared alignment restrictions. Neither external matrix contains unattempted or interrupted requests. They measure utterance-level seven-class emotion recognition and continuous sentiment, respectively, rather than replicating the CH-SIMS annotation strata or physical interventions.

Q3's strict-valid counts for $T/AV/TAV$ are 0/88/28 on MELD and 3/100/69 on MOSEI (512 intended per view). The complete-fence normalization fixed before follow-up inference raises these to 512/510/511 and 512/480/505. The additional numerical quote rule changes no external result. This extends the format-versus-prediction distinction to fixed external tasks; it does not convert recovered strings into native schema compliance.

Additional modalities do not uniformly improve on text. On MELD, Mini achieves 62.89% accuracy with T and 52.93% with TAV, with complete strict coverage in both. The paired accuracy difference is 9.96 percentage points [6.25,13.58]; the macro-F1 difference is .0456 [.0033,.0826]. Q25's strict $T - TAV$ all-intended accuracy difference is -1.56 percentage points [-4.97,1.62]. Conversely, adding text to AV improves all-intended accuracy for all three models under the fixed fence normalization. The results do not identify conflict as the mechanism, or imply that every additional input is harmful.

On MOSEI's native $[-3, 3]$ scale, Mini's strict $T - TAV$ operational-risk difference is $-.1628$ [$-.2637, -.0609$], while its valid-pair MAE difference is $+.0352$ [$-.0176$,.0868] across 493

pairs. The former includes a failure penalty of 6; the latter conditions on both responses being valid. They answer different questions. All models, views, parsing tiers, predefined contrasts and fixed duplicate/alignment sensitivity cohorts are retained in Appendix F and the anonymous `results/followup-aggregate.json` attachment. These subset evaluations are not full-test leaderboard comparisons or evidence of human-validated quality, emotional masking, or unseen-speaker generalization.

7 LIMITATIONS

Scope. The Omni evaluations use fixed Chinese and English benchmark cohorts, one prompt per task, one media budget and one decode per request. Pretraining contamination is unknown. Source overlap and label-version differences limit generalization. Resampling only addresses dependence within the evaluated cohort. Post-hoc recovery and prior analyses cannot become preregistered evidence; many unadjusted comparisons and sparse cells limit inference. Speech semantics, subtitles and nonidentical processors prevent clean causal modality attribution. No model internals were intervened upon. Emotional masking and private mental states are outside the measured construct. The external datasets lack corresponding unimodal conflict labels and use restricted utterance-level cohorts; human quality checks remain incomplete.

8 CONCLUSION

The audit shows why sentiment reliability cannot be summarized by an error increment or strict-format score. Model-dependent input effects coexist with different target-only difficulty, and syntax recovery changes the numerical interpretation of Q3 failures without changing its format compliance. Matched feature controls fail to support the proposed factorization advantage. Absolute risk, paired changes, a train-only prior and output coverage should be reported together. The completed physical-stress blocks show why assigned transformation strength should not replace measured paired target loss. External utterance-level cohorts show related format and utility distinctions, but do not establish perceptual-quality or unrestricted cross-domain claims.

REPRODUCIBILITY STATEMENT

Raw responses, failures, manifests, model revisions and scoring receipts are retained with hashes. Tables and figures derive from sealed predictions; exploratory analyses retain their timing. Appendices give the feature-study protocol, full predefined ablations, media and model contracts, and follow-up scoring rules. The anonymous supplement includes complete aggregate statistics, manuscript sources and three exact frozen scoring modules for algorithm inspection. It omits private execution bindings, raw responses, row-level labels, inference adapters and the earlier training implementation, so it is not a self-contained end-to-end reproduction package. Acquired media are not redistributed without permission.

ETHICS STATEMENT

The study evaluates dataset annotations and model predictions, not an individual's concealed feelings, intentions or clinical state. Natural annotation disagreement is not a diagnosis of emotional masking. Existing data and model resources retain their provider terms; no raw audiovisual media are included in the submission. Human-quality judgments are not claimed without completed and documented ratings.

AI USE STATEMENT

Generative AI assistants supported literature discovery and summarization, conceptual framing and hypothesis refinement, experimental design, data format and integrity checks, implementation, debugging, automated experiment execution, statistical interpretation, figures, manuscript drafting and editing. They also helped implement synthetic audio engineering probes and specified noise/blur

transformations; no synthetic target labels or human ratings were substituted for observations. AI-assisted software was checked using automated tests and separate assistant audits, with input/output hashes and failures retained. These checks do not constitute independent human replication. The authors are responsible for the final claims, citations and artifacts.

REFERENCES

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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Bojing Hou, Ruohao Li, Yitong Zhu, Luwen Yu, and Yuyang Wang. CONFER: Conflict-aware evidence negotiation for regime-calibrated weak supervision in multimodal emotion recognition. *arXiv preprint arXiv:2608.07867*, 2026. URL https://arxiv.org/abs/2608.07867.

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 *International Conference on Multimodal Interaction*, 2022. doi: 10.1145/3536221.3556630. URL https://arxiv.org/abs/2209.02604.

Huisheng Mao, Baozheng Zhang, Hua Xu, Ziqi Yuan, and Yihe Liu. Robust-MSA: Understanding the impact of modality noise on multimodal sentiment analysis. *arXiv preprint arXiv:2211.13484*, 2022. URL https://arxiv.org/abs/2211.13484.

Chunlei Meng, Jacqueline J. Pang, Pengbin Feng, Zhenyu Yu, Chun Ouyang, and Zhongxue Gan. Rethinking modality reliability in multimodal sentiment analysis with incomplete observations. *arXiv preprint arXiv:2608.03611*, 2026. URL https://arxiv.org/abs/2608.03611. Version 2.

Yen-Ting Piao, Shu-Yun Chen, Chin-Hui Chu, Chun-Wei Chen, Shih-Yun Shan Kuan, Hung-yi Lee, and Yun-Nung Chen. Tri-PvP: Exposing modality bias in omni-modal large language models through perceptual-propositional evidence conflicts. *arXiv preprint arXiv:2609.06011*, 2026. URL https://arxiv.org/abs/2609.06011v1. Version 1.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

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. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20351–20371. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1018. URL https://aclanthology.org/2026.findings-acl.1018/.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.343. URL https://aclanthology.org/2020.acl-main.343/.

Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 2023. doi: 10.1109/TMM.2023.3267882. URL https://doi.org/10.1109/TMM.2023.3267882. Early-access author version.

A ALL VIEWS AND TARGET-BIN CELLS

Table 4: Recovered MAE for every view and group. Overall/A/D/N denominators are 1034/459/149/426 for every row; recovery is post hoc.

Model	View	Overall	A	D	N
Q25	T	0.3591	0.3222	0.5295	0.3392
Q25	A	0.3237	0.2878	0.4631	0.3136
Q25	V	0.3568	0.4015	0.3919	0.2962
Q25	TA	0.3518	0.3277	0.4846	0.3315
Q25	TV	0.3290	0.3166	0.4356	0.3052
Q25	AV	0.3089	0.2978	0.3792	0.2962
Q25	TAV	0.3342	0.3218	0.4456	0.3087
Mini	T	0.3396	0.2505	0.5275	0.3697
Mini	A	0.3390	0.2808	0.4839	0.3509
Mini	V	0.2985	0.2821	0.4087	0.2777
Mini	TA	0.3271	0.2745	0.4732	0.3326
Mini	TV	0.2716	0.2242	0.4235	0.2695
Mini	AV	0.2695	0.2076	0.4087	0.2876
Mini	TAV	0.2674	0.2115	0.3872	0.2857
Q3	T	0.3687	0.3168	0.5342	0.3667
Q3	A	0.3529	0.2999	0.5346	0.3465
Q3	V	0.3865	0.4499	0.3966	0.3146
Q3	TA	0.3512	0.3169	0.4624	0.3492
Q3	TV	0.3457	0.3074	0.4537	0.3493
Q3	AV	0.3678	0.3887	0.4362	0.3215
Q3	TAV	0.3539	0.3190	0.4591	0.3548

Table 5: Exploratory group/target cells, with all sparse cells retained. Strong means $|M| > .6$, weak means $0 < |M| \leq .6$, based on the training nonzero absolute-target median. Point estimates are descriptive; the A/zero cell has one observation.

Target bin	Group	n	Prior	Q25	Mini	Q3
Negative, strong	A	144	0.8750	0.3597	0.2444	0.4229
Negative, strong	D	6	0.8333	0.8333	0.6667	0.7333
Negative, strong	N	14	0.8571	0.6643	0.3429	0.7000
Negative, weak	A	131	0.4779	0.3443	0.1908	0.3779
Negative, weak	D	56	0.3857	0.4036	0.3429	0.5357
Negative, weak	N	152	0.3842	0.3355	0.2796	0.4135
Exact zero	A	1	0.0000	0.0000	0.5000	0.0000
Exact zero	D	16	0.0000	0.1437	0.3125	0.2687
Exact zero	N	77	0.0000	0.1260	0.3143	0.2312
Positive, weak	A	99	0.5253	0.2707	0.2465	0.2869
Positive, weak	D	60	0.3933	0.5300	0.4450	0.4767
Positive, weak	N	162	0.4037	0.3265	0.2963	0.3593
Positive, strong	A	84	0.8381	0.2857	0.1429	0.0905
Positive, strong	D	11	0.8000	0.4273	0.2545	0.1000
Positive, strong	N	21	0.8381	0.4095	0.1048	0.1190

B PARSING SENSITIVITY

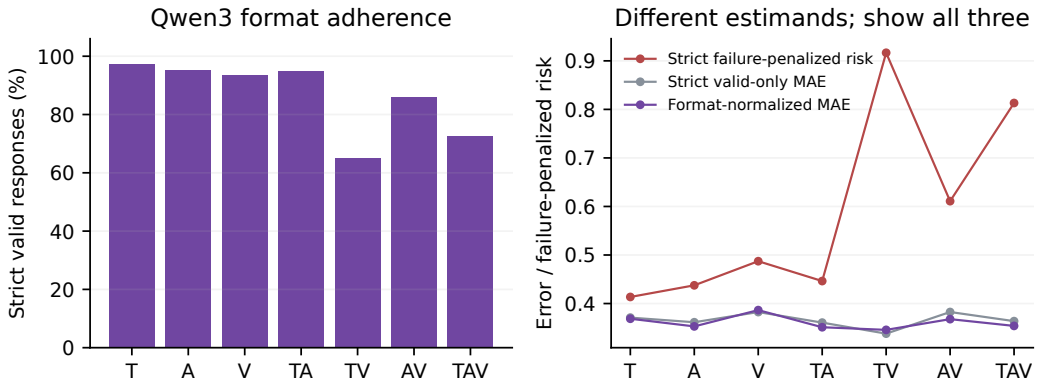

Figure 3: Q3 strict-valid MAE, all-intended risk and recovered numeric MAE are different estimands. Strict-invalid counts for T/A/V/TA/TV/AV/TAV are 27/48/67/54/360/146/284.

The strict parser requires the whole response to match the original JSON schema with one finite numeric sentiment value in $[-1, 1]$. Recovery recognizes a complete outer code fence or the documented unmatched quote, then reapplies the original validation. Ground-truth labels are never read by recovery. Outputs and parsing decisions are retained. These rules account for the present failures, not all possible generation failures. Future runs must retain unrecoverable cases in intended denominators.

C MODEL IDENTITIES AND MEDIA CONTRACT

Official checkpoints are Qwen/Qwen2.5-Omni-7B, openbmb/MiniCPM-o-4_5, and Qwen/Qwen3-Omni-30B-A3B-Instruct. Frozen revision prefixes are `ae9e1690543f`, `503e754207c9`, and `26291f793822`, respectively. Full hashes and runtime package versions are retained in the manifests. No quantization or training is added.

For duration d, visual sampling uses $n = \min(16, \lceil d \rceil)$ frames at nominal coordinates $(j+1/2)d/n$, $j = 0, \ldots, n-1$. Nominal and actual decoded timestamps are retained. Resizing preserves aspect ratio with long edge at most 448 pixels. Full audio is 16-kHz mono PCM16. The native model processors further transform these inputs; common canonical media do not imply identical tokens.

D PHYSICAL AND HUMAN CHECKS

Noise uses full-waveform RMS and a per-clip PCG64 direction shared across levels. Near silence fails closed. PCM16 uses explicit saturation and nearest rounding without gain normalization. Required checks include realized SNR error at most .25 dB and saturated fraction at most .001. Blur operates on canonical RGB frames with reflect boundaries, truncate 3 and nearest uint8 rounding. Dimensions, timestamps and unchanged-modality hashes are verified. Physical fidelity alone does not establish perceived quality or preserved sentiment semantics.

The human candidate set has 96 different sources. Each clip yields three audio-only and three silent-visual tasks, giving 576 tasks per rater from 480 bundles; clean audio and visual inputs are rated separately. Six cyclically randomized blocks each contain a clip once. Targets, groups, condition names and model outputs are hidden. Repeated content and perceptible transformations prevent full blinding. The 1–5 scales assess clarity and affect-cue readability, with explicit non-applicability and playback-failure options. Operational thresholds and missingness rules are frozen before collecting ratings. No A/D difference is required to pass. The incomplete single-author pilot and its limitations are disclosed in Appendix E; it does not meet the frozen human-evaluation requirements.

E RAW-MEDIA FOLLOW-UP PROTOCOL

The follow-up was designed after the earlier seven-view results. Its inputs, parsing rules, labels and analysis code were frozen before its own model predictions. This timing does not make it an independently preregistered confirmation of the earlier observations.

Paired transformations. Each CH-SIMS clip has a new clean TAV reference and four paired conditions: additive audio noise at 20 and 10 dB and Gaussian frame blur at $\sigma = 1$ and 2 pixels. Noise direction is fixed per clip using PCG64 and shared between strengths. Scaling uses complete-waveform RMS at 16 kHz mono; PCM16 encoding applies explicit rounding and saturation, without gain normalization. Physical acceptance requires SNR error at most 0.25 dB and saturated fraction at most .001. Two test clips fail the latter check at 10 dB, leaving six model cells data-unverified without an inference call. Thresholds and clips are not changed after this check.

Blur operates on canonical RGB frames with reflect boundaries and truncation at three standard deviations. Dimensions, frame order and timestamps remain fixed. Untouched inputs are checked against clean content hashes. These are specified physical transformations, not established perceptual-quality levels. The original sentiment annotation is retained as a predictive-stress reference; its validity after transformation is not reannotated.

Cohorts and external-task protocols. The CH-SIMS follow-up includes all 1,034 test clips and a locked 192-clip validation cohort with 97/34/61 clips in $A/D/N$. The validation clips were available in previous supervised checkpoint selection and share sources with other splits. They are not untouched external validation.

The external cohorts contain 512 locked test clips each from CMU-MOSEI (Bagher Zadeh et al., 2018) and MELD (Poria et al., 2019), with T, AV and TAV supplied to each model. MOSEI predicts continuous sentiment in $[-3, 3]$ using the acquired MMSA label/split release; equivalence to other CMU SDK folds is not assumed. MELD predicts one of seven emotion labels for the target utterance without preceding dialogue context. Audio retains language and frames retain any visible text. Neither dataset supplies the corresponding CH-SIMS unimodal human labels; we do not construct an equivalent gold disagreement group. Subset and context choices preclude direct comparison with full-test or dialogue-context leaderboards.

A 103.2-second MOSEI clip has unresolved correspondence between its long media and short transcript. Its text input is retained, while six AV/TAV model cells are data-unverified. Another selected MOSEI test clip has a byte-identical training counterpart, excluded from the 24-clip training pilot; no task training is performed. Two MELD test IDs have identical media, different official transcripts and the same gold label. The original IDs remain in the intended cohort. Fixed sensitivity cohorts exclude the MOSEI duplicate, long clip, or both (511, 511 and 510 clips), and the two flagged MELD IDs (510 clips). They supplement rather than replace the original cohorts.

Execution and provenance. The intended matrix contains 27,606 cells: 15,510 CH-SIMS test, 2,880 CH-SIMS validation, and 4,608 for each external dataset. Twelve cells have predeclared data restrictions. Complete responses, including malformed strings, are retained without requery. An interrupted attempt is also retained and never sent again during recovery. Unattempted cells remain explicit. Recovery binds each original row to its actual execution and merges only disjoint request identities; it does not replace old hardware or run identities with a new allocation.

The models retain their frozen revisions, native processors, separate environments, bfloat16 settings, greedy decoding, batch size one and 64-token generation cap. Parallel recovery workers process disjoint whole dataset/split/model branches. Assignment uses training-only latency and preserves order within each branch. Each assigned model repeats its original 182 training and engineering checks before formal inference. Label accuracy and strict format success are not preflight acceptance criteria. A physical resource interruption does not become evidence of model incapacity.

One identical-environment transport restart is permitted for the next unattempted request within a branch; the failed request is not retried. A second transport error or implementation/OOM error stops that branch. Native input-context limits are recorded separately. Scoring begins only after all relevant writers exit and prediction identities, hashes and provenance are sealed. Labels and anno-

tation groups are not supplied to model adapters, but shared-filesystem access remains technically possible; this is application-level separation, not operating-system isolation.

Scoring. Strict JSON parsing and the two fixed syntax-normalization tiers are reported separately. A complete outer Markdown fence may be removed; only numerical sentiment permits the previously specified unmatched-quote repair. There is no numeric clipping, arbitrary extraction, label-based repair, MELD fuzzy matching or output selection. These rules predate follow-up predictions, although their original development from earlier failures was post hoc.

Valid-only MAE is separated from all-intended operational risk, with failure penalties 2 for CH-SIMS and 6 for MOSEI. Data restrictions, execution errors, unattempted cells and invalid formats are counted separately. Incomplete execution can inflate operational risk without establishing poor predictive ability. MELD reports coverage, valid-only and all-intended accuracy, seven-gold-class macro/weighted F1, and an invalid prediction column in its seven-row confusion matrix. Invalid outputs add false negatives to their true classes; they do not create an eighth gold class.

Follow-up intervals use 2,000 shared source-cluster bootstrap draws with seed 20260922: source video for CH-SIMS/MOSEI and episode for MELD. They condition on fixed models, processors, prompts and one noise realization; multiple comparisons are not corrected. All natural strata, target-strength bins, train-median controls, condition-minus-new-clean contrasts and fixed external sensitivity sets are retained. Historical clean outputs serve only as a separate reproducibility check, not a selectable reference.

Human evidence. A separate protocol plans 576 audio-only or silent-visual tasks per rater from 96 source-distinct training clips, with two independent raters scoring clarity and affect-cue readability. We received an incomplete author pilot: 28 tasks (56 scores), comprising 19 audio and 9 visual tasks from 28 distinct clips, from one hypothesis-aware author. No second-rater export or within-clip clean/transformed rating pairs were available. Clarity scores were 3/4/5 on 4/11/13 tasks; affect-cue readability scores were 3/4/5 on 1/12/15 tasks. These are descriptive counts for the received subset, not representative estimates or a completed human evaluation. The subset cannot establish paired perceptual effects or inter-rater agreement; the planned completeness and paired-coverage requirements remain unmet. Missing ratings are not imputed. Ratings neither gate inference nor choose conditions, parsing or reported subsets. The absence of a completed study limits perceptual interpretation and does not support claims about preserved emotional semantics or emotional masking. The original design made human checks an execution gate; a later amendment allowed inference to proceed while ratings were pending. Engineering input checks and the evidential requirements for perceptual claims were unchanged.

F Complete follow-up aggregate results

All entries are transcribed from the frozen scoring output. Intervals are conditional 95% source-cluster bootstrap intervals (2,000 shared draws), without multiplicity correction. All intended cells remain in operational-risk denominators. MAE conditions on a valid response; paired MAE differences condition on both outputs being valid. NA denotes an undefined statistic, not zero. † retains the frozen exploratory flag for sparse cohorts (fewer than 50 valid clips or 10 source clusters). Full natural strata, target bins, prior controls, sensitivity cohorts and confusion matrices are retained in the anonymous `results/followup-aggregate.json` attachment.

F.1 CH-SIMS test

Cohort: 1034 clips, 142 source clusters, 15510 intended requests. Terminal counts: output = 15501; data_unverified = 6; runtime_error = 3.

The sentiment range is $[-1, 1]$ with operational failure penalty 2.

Strict. F/D/R/U denote format errors, data restrictions, runtime errors and unattempted requests, respectively.

Model/input	Valid/intended	Valid MAE	Operational risk [CI]	F/D/R/U
Q25/Audio 10 dB	1032/1034	0.3390	0.3422 [0.3225, 0.3615]	0/2/0/0
Q25/Audio 20 dB	1034/1034	0.3394	0.3394 [0.3200, 0.3591]	0/0/0/0
Q25/Clean	1034/1034	0.3342	0.3342 [0.3145, 0.3545]	0/0/0/0
Q25/Blur 1	1034/1034	0.3408	0.3408 [0.3207, 0.3621]	0/0/0/0
Q25/Blur 2	1034/1034	0.3480	0.3480 [0.3275, 0.3690]	0/0/0/0
Mini/Audio 10 dB	1032/1034	0.2668	0.2701 [0.2514, 0.2884]	0/2/0/0
Mini/Audio 20 dB	1034/1034	0.2662	0.2662 [0.2478, 0.2847]	0/0/0/0
Mini/Clean	1034/1034	0.2674	0.2674 [0.2490, 0.2863]	0/0/0/0
Mini/Blur 1	1034/1034	0.2721	0.2721 [0.2519, 0.2923]	0/0/0/0
Mini/Blur 2	1034/1034	0.2920	0.2920 [0.2700, 0.3141]	0/0/0/0
Q3/Audio 10 dB	781/1034	0.3584	0.7601 [0.7043, 0.8208]	251/2/0/0
Q3/Audio 20 dB	767/1034	0.3583	0.7822 [0.7111, 0.8609]	267/0/0/0
Q3/Clean	750/1034	0.3637	0.8132 [0.7522, 0.8779]	283/0/1/0
Q3/Blur 1	777/1034	0.3560	0.7646 [0.7003, 0.8277]	257/0/0/0
Q3/Blur 2	818/1034	0.3683	0.7092 [0.6525, 0.7745]	214/0/2/0

Every predefined overall comparison is retained below. Differences are left minus right; lower error is better. MELD differences concern classification error, not accuracy. Valid-pair columns can have different denominators; they are not differences between two separate valid-only means.

Comparison	Valid pairs	Intended error difference [CI]	Valid-pair error difference [CI]
Q25/A20 - clean	1034	0.0051 [-0.0022, 0.0133]	0.0051 [-0.0022, 0.0133]
Q25/A10 - clean	1032	0.0079 [-0.0004, 0.0169]	0.0043 [-0.0035, 0.0118]
Q25/V1 - clean	1034	0.0066 [-0.0005, 0.0144]	0.0066 [-0.0005, 0.0144]
Q25/V2 - clean	1034	0.0137 [0.0069, 0.0214]	0.0137 [0.0069, 0.0214]
Mini/A20 - clean	1034	-0.0013 [-0.0062, 0.0040]	-0.0013 [-0.0062, 0.0040]
Mini/A10 - clean	1032	0.0027 [-0.0058, 0.0115]	-0.0010 [-0.0077, 0.0057]
Mini/V1 - clean	1034	0.0047 [-0.0023, 0.0122]	0.0047 [-0.0023, 0.0122]
Mini/V2 - clean	1034	0.0246 [0.0123, 0.0371]	0.0246 [0.0123, 0.0371]
Q3/A20 - clean	637	-0.0309 [-0.0769, 0.0193]	-0.0009 [-0.0111, 0.0073]
Q3/A10 - clean	643	-0.0531 [-0.0999, -0.0048]	-0.0022 [-0.0137, 0.0089]
Q3/V1 - clean	632	-0.0485 [-0.1051, 0.0065]	-0.0071 [-0.0200, 0.0040]
Q3/V2 - clean	654	-0.1040 [-0.1638, -0.0450]	0.0009 [-0.0140, 0.0143]
Q25 - Mini/TAV/clean	1034	0.0668 [0.0434, 0.0916]	0.0668 [0.0434, 0.0916]
Q25 - Q3/TAV/clean	750	-0.4789 [-0.5481, -0.4206]	-0.0300 [-0.0622, -0.0008]
Mini - Q3/TAV/clean	750	-0.5457 [-0.6112, -0.4842]	-0.0969 [-0.1327, -0.0667]

Complete-fence normalization. F/D/R/U denote format errors, data restrictions, runtime errors and unattempted requests, respectively.

Model/input	Valid/intended	Valid MAE	Operational risk [CI]	F/D/R/U
Q25/Audio 10 dB	1032/1034	0.3390	0.3422 [0.3225, 0.3615]	0/2/0/0
Q25/Audio 20 dB	1034/1034	0.3394	0.3394 [0.3200, 0.3591]	0/0/0/0
Q25/Clean	1034/1034	0.3342	0.3342 [0.3145, 0.3545]	0/0/0/0
Q25/Blur 1	1034/1034	0.3408	0.3408 [0.3207, 0.3621]	0/0/0/0
Q25/Blur 2	1034/1034	0.3480	0.3480 [0.3275, 0.3690]	0/0/0/0
Mini/Audio 10 dB	1032/1034	0.2668	0.2701 [0.2514, 0.2884]	0/2/0/0
Mini/Audio 20 dB	1034/1034	0.2662	0.2662 [0.2478, 0.2847]	0/0/0/0
Mini/Clean	1034/1034	0.2674	0.2674 [0.2490, 0.2863]	0/0/0/0
Mini/Blur 1	1034/1034	0.2721	0.2721 [0.2519, 0.2923]	0/0/0/0
Mini/Blur 2	1034/1034	0.2920	0.2920 [0.2700, 0.3141]	0/0/0/0
Q3/Audio 10 dB	999/1034	0.3597	0.4152 [0.3833, 0.4523]	33/2/0/0
Q3/Audio 20 dB	1004/1034	0.3577	0.4054 [0.3742, 0.4410]	30/0/0/0
Q3/Clean	1008/1034	0.3598	0.4011 [0.3719, 0.4362]	25/0/1/0
Q3/Blur 1	1006/1034	0.3576	0.4020 [0.3715, 0.4364]	28/0/0/0
Q3/Blur 2	1000/1034	0.3630	0.4169 [0.3847, 0.4526]	32/0/2/0

Every predefined overall comparison is retained below. Differences are left minus right; lower error is better. MELD differences concern classification error, not accuracy. Valid-pair columns can have different denominators; they are not differences between two separate valid-only means.

Comparison	Valid pairs	Intended error difference [CI]	Valid-pair error difference [CI]
Q25/A20 - clean	1034	0.0051 [-0.0022, 0.0133]	0.0051 [-0.0022, 0.0133]
Q25/A10 - clean	1032	0.0079 [-0.0004, 0.0169]	0.0043 [-0.0035, 0.0118]
Q25/V1 - clean	1034	0.0066 [-0.0005, 0.0144]	0.0066 [-0.0005, 0.0144]
Q25/V2 - clean	1034	0.0137 [0.0069, 0.0214]	0.0137 [0.0069, 0.0214]
Mini/A20 - clean	1034	-0.0013 [-0.0062, 0.0040]	-0.0013 [-0.0062, 0.0040]
Mini/A10 - clean	1032	0.0027 [-0.0058, 0.0115]	-0.0010 [-0.0077, 0.0057]
Mini/V1 - clean	1034	0.0047 [-0.0023, 0.0122]	0.0047 [-0.0023, 0.0122]
Mini/V2 - clean	1034	0.0246 [0.0123, 0.0371]	0.0246 [0.0123, 0.0371]
Q3/A20 - clean	998	0.0043 [-0.0123, 0.0206]	-0.0021 [-0.0120, 0.0068]
Q3/A10 - clean	993	0.0141 [-0.0016, 0.0307]	-0.0012 [-0.0101, 0.0077]
Q3/V1 - clean	996	0.0010 [-0.0168, 0.0170]	-0.0048 [-0.0145, 0.0037]
Q3/V2 - clean	994	0.0158 [-0.0022, 0.0338]	0.0018 [-0.0093, 0.0116]
Q25 - Mini/TAV/clean	1034	0.0668 [0.0434, 0.0916]	0.0668 [0.0434, 0.0916]
Q25 - Q3/TAV/clean	1008	-0.0668 [-0.1036, -0.0350]	-0.0201 [-0.0484, 0.0045]
Mini - Q3/TAV/clean	1008	-0.1337 [-0.1674, -0.1037]	-0.0882 [-0.1153, -0.0644]

Fence plus numerical-quote normalization. F/D/R/U denote format errors, data restrictions, runtime errors and unattempted requests, respectively.

Model/input	Valid/intended	Valid MAE	Operational risk [CI]	F/D/R/U
Q25/Audio 10 dB	1032/1034	0.3390	0.3422 [0.3225, 0.3615]	0/2/0/0
Q25/Audio 20 dB	1034/1034	0.3394	0.3394 [0.3200, 0.3591]	0/0/0/0
Q25/Clean	1034/1034	0.3342	0.3342 [0.3145, 0.3545]	0/0/0/0
Q25/Blur 1	1034/1034	0.3408	0.3408 [0.3207, 0.3621]	0/0/0/0
Q25/Blur 2	1034/1034	0.3480	0.3480 [0.3275, 0.3690]	0/0/0/0
Mini/Audio 10 dB	1032/1034	0.2668	0.2701 [0.2514, 0.2884]	0/2/0/0
Mini/Audio 20 dB	1034/1034	0.2662	0.2662 [0.2478, 0.2847]	0/0/0/0
Mini/Clean	1034/1034	0.2674	0.2674 [0.2490, 0.2863]	0/0/0/0
Mini/Blur 1	1034/1034	0.2721	0.2721 [0.2519, 0.2923]	0/0/0/0
Mini/Blur 2	1034/1034	0.2920	0.2920 [0.2700, 0.3141]	0/0/0/0
Q3/Audio 10 dB	1032/1034	0.3530	0.3561 [0.3329, 0.3821]	0/2/0/0
Q3/Audio 20 dB	1034/1034	0.3519	0.3519 [0.3281, 0.3783]	0/0/0/0
Q3/Clean	1033/1034	0.3542	0.3558 [0.3312, 0.3820]	0/0/1/0
Q3/Blur 1	1034/1034	0.3527	0.3527 [0.3279, 0.3784]	0/0/0/0
Q3/Blur 2	1032/1034	0.3560	0.3591 [0.3339, 0.3863]	0/0/2/0

Every predefined overall comparison is retained below. Differences are left minus right; lower error is better. MELD differences concern classification error, not accuracy. Valid-pair columns can have different denominators; they are not differences between two separate valid-only means.

Comparison	Valid pairs	Intended error difference [CI]	Valid-pair error difference [CI]
Q25/A20 - clean	1034	0.0051 [-0.0022, 0.0133]	0.0051 [-0.0022, 0.0133]
Q25/A10 - clean	1032	0.0079 [-0.0004, 0.0169]	0.0043 [-0.0035, 0.0118]
Q25/V1 - clean	1034	0.0066 [-0.0005, 0.0144]	0.0066 [-0.0005, 0.0144]
Q25/V2 - clean	1034	0.0137 [0.0069, 0.0214]	0.0137 [0.0069, 0.0214]
Mini/A20 - clean	1034	-0.0013 [-0.0062, 0.0040]	-0.0013 [-0.0062, 0.0040]
Mini/A10 - clean	1032	0.0027 [-0.0058, 0.0115]	-0.0010 [-0.0077, 0.0057]
Mini/V1 - clean	1034	0.0047 [-0.0023, 0.0122]	0.0047 [-0.0023, 0.0122]
Mini/V2 - clean	1034	0.0246 [0.0123, 0.0371]	0.0246 [0.0123, 0.0371]
Q3/A20 - clean	1033	-0.0039 [-0.0131, 0.0054]	-0.0020 [-0.0115, 0.0065]
Q3/A10 - clean	1031	0.0004 [-0.0108, 0.0120]	-0.0014 [-0.0116, 0.0086]
Q3/V1 - clean	1033	-0.0031 [-0.0127, 0.0061]	-0.0013 [-0.0107, 0.0074]
Q3/V2 - clean	1031	0.0034 [-0.0092, 0.0152]	0.0018 [-0.0099, 0.0123]
Q25 - Mini/TAV/clean	1034	0.0668 [0.0434, 0.0916]	0.0668 [0.0434, 0.0916]
Q25 - Q3/TAV/clean	1033	-0.0215 [-0.0495, 0.0029]	-0.0196 [-0.0471, 0.0044]
Mini - Q3/TAV/clean	1033	-0.0883 [-0.1138, -0.0668]	-0.0865 [-0.1127, -0.0634]

New versus historical clean. This diagnostic retains the new clean response as the paired reference for every condition comparison. The following differences are new minus historical; identical refers to numerical predictions on valid pairs.

Model/tier	Pairs	Identical	Intended risk difference [CI]	Valid-pair MAE difference [CI]
Q25/S	1034	1034	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Mini/S	1034	1034	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Q3/S	750	750	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Q25/F	1034	1034	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Mini/F	1034	1034	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Q3/F	1008	1008	0.0018 [0.0000, 0.0060]	0.0000 [0.0000, 0.0000]
Q25/FQ	1034	1034	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Mini/FQ	1034	1034	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Q3/FQ	1033	1033	0.0018 [0.0000, 0.0060]	0.0000 [0.0000, 0.0000]

F.2 CH-SIMS VALIDATION

Cohort: 192 clips, 82 source clusters, 2880 intended requests. Terminal counts: output = 2877; runtime_error = 3.

The sentiment range is $[-1, 1]$ with operational failure penalty 2.

Strict. F/D/R/U denote format errors, data restrictions, runtime errors and unattempted requests, respectively.

Model/input	Valid/intended	Valid MAE	Operational risk [CI]	F/D/R/U
Q25/Audio 10 dB	192/192	0.3734	0.3734 [0.3223, 0.4296]	0/0/0/0
Q25/Audio 20 dB	191/192	0.3613	0.3698 [0.3219, 0.4241]	0/0/1/0
Q25/Clean	192/192	0.3589	0.3589 [0.3134, 0.4090]	0/0/0/0
Q25/Blur 1	192/192	0.3625	0.3625 [0.3190, 0.4079]	0/0/0/0
Q25/Blur 2	192/192	0.3714	0.3714 [0.3225, 0.4253]	0/0/0/0
Mini/Audio 10 dB	192/192	0.2630	0.2630 [0.2284, 0.3010]	0/0/0/0
Mini/Audio 20 dB	192/192	0.2552	0.2552 [0.2218, 0.2927]	0/0/0/0
Mini/Clean	191/192	0.2487	0.2578 [0.2200, 0.3030]	0/0/1/0
Mini/Blur 1	192/192	0.2688	0.2688 [0.2337, 0.3061]	0/0/0/0
Mini/Blur 2	192/192	0.2755	0.2755 [0.2376, 0.3164]	0/0/0/0
Q3/Audio 10 dB	125/192	0.3336	0.9151 [0.7729, 1.0582]	67/0/0/0
Q3/Audio 20 dB	137/192	0.3577	0.8281 [0.6840, 0.9768]	54/0/1/0
Q3/Clean	133/192	0.3850	0.8812 [0.7560, 1.0153]	59/0/0/0
Q3/Blur 1	138/192	0.3812	0.8365 [0.7183, 0.9562]	54/0/0/0
Q3/Blur 2	146/192	0.3397	0.7375 [0.6141, 0.8561]	46/0/0/0

Every predefined overall comparison is retained below. Differences are left minus right; lower error is better. MELD differences concern classification error, not accuracy. Valid-pair columns can have different denominators; they are not differences between two separate valid-only means.

Comparison	Valid pairs	Intended error difference [CI]	Valid-pair error difference [CI]
Q25/A20 - clean	191	0.0109 [-0.0088, 0.0350]	0.0026 [-0.0115, 0.0185]
Q25/A10 - clean	192	0.0146 [-0.0045, 0.0349]	0.0146 [-0.0045, 0.0349]
Q25/V1 - clean	192	0.0036 [-0.0176, 0.0270]	0.0036 [-0.0176, 0.0270]
Q25/V2 - clean	192	0.0125 [-0.0080, 0.0335]	0.0125 [-0.0080, 0.0335]
Mini/A20 - clean	191	-0.0026 [-0.0302, 0.0199]	0.0068 [-0.0069, 0.0231]
Mini/A10 - clean	191	0.0052 [-0.0246, 0.0332]	0.0147 [-0.0052, 0.0368]
Mini/V1 - clean	191	0.0109 [-0.0230, 0.0441]	0.0204 [-0.0081, 0.0497]
Mini/V2 - clean	191	0.0177 [-0.0209, 0.0554]	0.0272 [-0.0047, 0.0616]
Q3/A20 - clean	116	-0.0531 [-0.1520, 0.0431]	-0.0250 [-0.0513, -0.0043]
Q3/A10 - clean	107	0.0339 [-0.0818, 0.1497]	-0.0589 [-0.1069, -0.0165]
Q3/V1 - clean	116	-0.0448 [-0.1473, 0.0529]	-0.0190 [-0.0453, 0.0094]
Q3/V2 - clean	119	-0.1438 [-0.2438, -0.0462]	-0.0286 [-0.0600, 0.0024]
Q25 - Mini/TAV/clean	191	0.1010 [0.0438, 0.1557]	0.1099 [0.0575, 0.1621]
Q25 - Q3/TAV/clean	133	-0.5224 [-0.6497, -0.4025]	-0.0218 [-0.0855, 0.0355]
Mini - Q3/TAV/clean	132	-0.6234 [-0.7641, -0.4922]	-0.1432 [-0.2135, -0.0801]

Complete-fence normalization. F/D/R/U denote format errors, data restrictions, runtime errors and unattempted requests, respectively.

Model/input	Valid/intended	Valid MAE	Operational risk [CI]	F/D/R/U
Q25/Audio 10 dB	192/192	0.3734	0.3734 [0.3223, 0.4296]	0/0/0/0
Q25/Audio 20 dB	191/192	0.3613	0.3698 [0.3219, 0.4241]	0/0/1/0
Q25/Clean	192/192	0.3589	0.3589 [0.3134, 0.4090]	0/0/0/0
Q25/Blur 1	192/192	0.3625	0.3625 [0.3190, 0.4079]	0/0/0/0
Q25/Blur 2	192/192	0.3714	0.3714 [0.3225, 0.4253]	0/0/0/0
Mini/Audio 10 dB	192/192	0.2630	0.2630 [0.2284, 0.3010]	0/0/0/0
Mini/Audio 20 dB	192/192	0.2552	0.2552 [0.2218, 0.2927]	0/0/0/0
Mini/Clean	191/192	0.2487	0.2578 [0.2200, 0.3030]	0/0/1/0
Mini/Blur 1	192/192	0.2688	0.2688 [0.2337, 0.3061]	0/0/0/0
Mini/Blur 2	192/192	0.2755	0.2755 [0.2376, 0.3164]	0/0/0/0
Q3/Audio 10 dB	184/192	0.3598	0.4281 [0.3600, 0.5094]	8/0/0/0
Q3/Audio 20 dB	184/192	0.3668	0.4349 [0.3685, 0.5137]	7/0/1/0
Q3/Clean	183/192	0.3732	0.4495 [0.3830, 0.5224]	9/0/0/0
Q3/Blur 1	185/192	0.3643	0.4240 [0.3556, 0.5062]	7/0/0/0
Q3/Blur 2	184/192	0.3582	0.4266 [0.3587, 0.5025]	8/0/0/0

Every predefined overall comparison is retained below. Differences are left minus right; lower error is better. MELD differences concern classification error, not accuracy. Valid-pair columns can have different denominators; they are not differences between two separate valid-only means.

Comparison	Valid pairs	Intended error difference [CI]	Valid-pair error difference [CI]
Q25/A20 - clean	191	0.0109 [-0.0088, 0.0350]	0.0026 [-0.0115, 0.0185]
Q25/A10 - clean	192	0.0146 [-0.0045, 0.0349]	0.0146 [-0.0045, 0.0349]
Q25/V1 - clean	192	0.0036 [-0.0176, 0.0270]	0.0036 [-0.0176, 0.0270]
Q25/V2 - clean	192	0.0125 [-0.0080, 0.0335]	0.0125 [-0.0080, 0.0335]
Mini/A20 - clean	191	-0.0026 [-0.0302, 0.0199]	0.0068 [-0.0069, 0.0231]
Mini/A10 - clean	191	0.0052 [-0.0246, 0.0332]	0.0147 [-0.0052, 0.0368]
Mini/V1 - clean	191	0.0109 [-0.0230, 0.0441]	0.0204 [-0.0081, 0.0497]
Mini/V2 - clean	191	0.0177 [-0.0209, 0.0554]	0.0272 [-0.0047, 0.0616]
Q3/A20 - clean	179	-0.0146 [-0.0775, 0.0475]	-0.0028 [-0.0276, 0.0218]
Q3/A10 - clean	179	-0.0214 [-0.0822, 0.0468]	-0.0117 [-0.0511, 0.0252]
Q3/V1 - clean	182	-0.0255 [-0.0712, 0.0202]	-0.0063 [-0.0297, 0.0174]
Q3/V2 - clean	182	-0.0229 [-0.0650, 0.0179]	-0.0148 [-0.0408, 0.0112]
Q25 - Mini/TAV/clean	191	0.1010 [0.0438, 0.1557]	0.1099 [0.0575, 0.1621]
Q25 - Q3/TAV/clean	183	-0.0906 [-0.1683, -0.0190]	-0.0011 [-0.0557, 0.0500]
Mini - Q3/TAV/clean	182	-0.1917 [-0.2677, -0.1198]	-0.1198 [-0.1765, -0.0717]

Fence plus numerical-quote normalization. F/D/R/U denote format errors, data restrictions, runtime errors and unattempted requests, respectively.

Model/input	Valid/intended	Valid MAE	Operational risk [CI]	F/D/R/U
Q25/Audio 10 dB	192/192	0.3734	0.3734 [0.3223, 0.4296]	0/0/0/0
Q25/Audio 20 dB	191/192	0.3613	0.3698 [0.3219, 0.4241]	0/0/1/0
Q25/Clean	192/192	0.3589	0.3589 [0.3134, 0.4090]	0/0/0/0
Q25/Blur 1	192/192	0.3625	0.3625 [0.3190, 0.4079]	0/0/0/0
Q25/Blur 2	192/192	0.3714	0.3714 [0.3225, 0.4253]	0/0/0/0
Mini/Audio 10 dB	192/192	0.2630	0.2630 [0.2284, 0.3010]	0/0/0/0
Mini/Audio 20 dB	192/192	0.2552	0.2552 [0.2218, 0.2927]	0/0/0/0
Mini/Clean	191/192	0.2487	0.2578 [0.2200, 0.3030]	0/0/1/0
Mini/Blur 1	192/192	0.2688	0.2688 [0.2337, 0.3061]	0/0/0/0
Mini/Blur 2	192/192	0.2755	0.2755 [0.2376, 0.3164]	0/0/0/0
Q3/Audio 10 dB	192/192	0.3500	0.3500 [0.2971, 0.4087]	0/0/0/0
Q3/Audio 20 dB	191/192	0.3589	0.3674 [0.3144, 0.4280]	0/0/1/0
Q3/Clean	192/192	0.3607	0.3607 [0.3162, 0.4152]	0/0/0/0
Q3/Blur 1	192/192	0.3555	0.3555 [0.3052, 0.4116]	0/0/0/0
Q3/Blur 2	192/192	0.3479	0.3479 [0.3009, 0.4030]	0/0/0/0

Every predefined overall comparison is retained below. Differences are left minus right; lower error is better. MELD differences concern classification error, not accuracy. Valid-pair columns can have different denominators; they are not differences between two separate valid-only means.

Comparison	Valid pairs	Intended error difference [CI]	Valid-pair error difference [CI]
Q25/A20 - clean	191	0.0109 [-0.0088, 0.0350]	0.0026 [-0.0115, 0.0185]
Q25/A10 - clean	192	0.0146 [-0.0045, 0.0349]	0.0146 [-0.0045, 0.0349]
Q25/V1 - clean	192	0.0036 [-0.0176, 0.0270]	0.0036 [-0.0176, 0.0270]
Q25/V2 - clean	192	0.0125 [-0.0080, 0.0335]	0.0125 [-0.0080, 0.0335]
Mini/A20 - clean	191	-0.0026 [-0.0302, 0.0199]	0.0068 [-0.0069, 0.0231]
Mini/A10 - clean	191	0.0052 [-0.0246, 0.0332]	0.0147 [-0.0052, 0.0368]
Mini/V1 - clean	191	0.0109 [-0.0230, 0.0441]	0.0204 [-0.0081, 0.0497]
Mini/V2 - clean	191	0.0177 [-0.0209, 0.0554]	0.0272 [-0.0047, 0.0616]
Q3/A20 - clean	191	0.0068 [-0.0219, 0.0391]	-0.0005 [-0.0238, 0.0226]
Q3/A10 - clean	192	-0.0107 [-0.0474, 0.0246]	-0.0107 [-0.0474, 0.0246]
Q3/V1 - clean	192	-0.0052 [-0.0270, 0.0177]	-0.0052 [-0.0270, 0.0177]
Q3/V2 - clean	192	-0.0128 [-0.0370, 0.0117]	-0.0128 [-0.0370, 0.0117]
Q25 - Mini/TAV/clean	191	0.1010 [0.0438, 0.1557]	0.1099 [0.0575, 0.1621]
Q25 - Q3/TAV/clean	192	-0.0018 [-0.0537, 0.0469]	-0.0018 [-0.0537, 0.0469]
Mini - Q3/TAV/clean	191	-0.1029 [-0.1588, -0.0526]	-0.1128 [-0.1669, -0.0665]

F.3 MELD TEST

Cohort: 512 clips, 126 source clusters, 4608 intended requests. Terminal counts: output = 4608.

MELD uses seven-class utterance-level recognition. Accuracy and F1 below use all intended clips; invalid predictions are errors. Values are proportions, not percentages.

Strict. F/D/R/U denote format errors, data restrictions, runtime errors and unattempted requests, respectively.

Model/input	Valid/intended	Accuracy [CI]	Macro F1	Weighted F1	F/D/R/U
Q25/AV	484/512	0.5117 [0.4702, 0.5497]	0.4034	0.5218	28/0/0/0
Q25/T	490/512	0.5801 [0.5395, 0.6206]	0.4329	0.5986	22/0/0/0
Q25/TAV	500/512	0.5957 [0.5535, 0.6377]	0.4328	0.5968	12/0/0/0
Mini/AV	512/512	0.4082 [0.3688, 0.4463]	0.3438	0.4088	0/0/0/0
Mini/T	512/512	0.6289 [0.5946, 0.6624]	0.4899	0.6322	0/0/0/0
Mini/TAV	512/512	0.5293 [0.4928, 0.5636]	0.4443	0.5348	0/0/0/0
Q3/AV	88/512	0.0820 [0.0569, 0.1092]	0.1671	0.1375	424/0/0/0
Q3/T	0/512	0.0000 [0.0000, 0.0000]	0.0000	0.0000	512/0/0/0
Q3/TAV	28/512	0.0371 [0.0229, 0.0524]	0.1053	0.0685	484/0/0/0

Every predefined overall comparison is retained below. Differences are left minus right; lower error is better. MELD differences concern classification error, not accuracy. Valid-pair columns can have different denominators; they are not differences between two separate valid-only means.

Comparison	Valid pairs	Intended error difference [CI]	Valid-pair error difference [CI]
Q25/T - TAV	482	0.0156 [-0.0162, 0.0497]	-0.0124 [-0.0423, 0.0203]
Q25/AV - TAV	477	0.0840 [0.0453, 0.1257]	0.0692 [0.0293, 0.1119]
Mini/T - TAV	512	-0.0996 [-0.1358, -0.0625]	-0.0996 [-0.1358, -0.0625]
Mini/AV - TAV	512	0.1211 [0.0805, 0.1592]	0.1211 [0.0805, 0.1592]
Q3/T - TAV	0	0.0371 [0.0229, 0.0524]	NA [NA]†
Q3/AV - TAV	17	-0.0449 [-0.0690, -0.0206]	0.1765 [0.0000, 0.4286]†
Q25 - Mini/TAV/clean	500	-0.0664 [-0.1151, -0.0201]	-0.0780 [-0.1279, -0.0310]
Q25 - Q3/TAV/clean	28	-0.5586 [-0.6025, -0.5143]	0.1429 [-0.0969, 0.3750]†
Mini - Q3/TAV/clean	28	-0.4922 [-0.5279, -0.4540]	0.0714 [-0.1250, 0.2814]†

Complete-fence normalization. F/D/R/U denote format errors, data restrictions, runtime errors and unattempted requests, respectively.

Model/input	Valid/intended	Accuracy [CI]	Macro F1	Weighted F1	F/D/R/U
Q25/AV	484/512	0.5117 [0.4702, 0.5497]	0.4034	0.5218	28/0/0/0
Q25/T	490/512	0.5801 [0.5395, 0.6206]	0.4329	0.5986	22/0/0/0
Q25/TAV	500/512	0.5957 [0.5535, 0.6377]	0.4328	0.5968	12/0/0/0
Mini/AV	512/512	0.4082 [0.3688, 0.4463]	0.3438	0.4088	0/0/0/0
Mini/T	512/512	0.6289 [0.5946, 0.6624]	0.4899	0.6322	0/0/0/0
Mini/TAV	512/512	0.5293 [0.4928, 0.5636]	0.4443	0.5348	0/0/0/0
Q3/AV	510/512	0.5293 [0.4899, 0.5680]	0.4222	0.5229	2/0/0/0
Q3/T	512/512	0.6289 [0.5887, 0.6653]	0.4813	0.6278	0/0/0/0
Q3/TAV	511/512	0.5781 [0.5359, 0.6209]	0.4861	0.5843	1/0/0/0

Every predefined overall comparison is retained below. Differences are left minus right; lower error is better. MELD differences concern classification error, not accuracy. Valid-pair columns can have different denominators; they are not differences between two separate valid-only means.

Comparison	Valid pairs	Intended error difference [CI]	Valid-pair error difference [CI]
Q25/T - TAV	482	0.0156 [-0.0162, 0.0497]	-0.0124 [-0.0423, 0.0203]
Q25/AV - TAV	477	0.0840 [0.0453, 0.1257]	0.0692 [0.0293, 0.1119]
Mini/T - TAV	512	-0.0996 [-0.1358, -0.0625]	-0.0996 [-0.1358, -0.0625]
Mini/AV - TAV	512	0.1211 [0.0805, 0.1592]	0.1211 [0.0805, 0.1592]
Q3/T - TAV	511	-0.0508 [-0.0825, -0.0175]	-0.0509 [-0.0827, -0.0175]
Q3/AV - TAV	509	0.0488 [0.0109, 0.0877]	0.0491 [0.0110, 0.0882]
Q25 - Mini/TAV/clean	500	-0.0664 [-0.1151, -0.0201]	-0.0780 [-0.1279, -0.0310]
Q25 - Q3/TAV/clean	499	-0.0176 [-0.0711, 0.0314]	-0.0281 [-0.0832, 0.0243]
Mini - Q3/TAV/clean	511	0.0488 [0.0057, 0.0936]	0.0509 [0.0079, 0.0958]

Fence plus numerical-quote normalization. All statistics equal the preceding complete-fence normalization tier; no additional response is recovered.

F.4 CMU-MOSEI TEST

Cohort: 512 clips, 320 source clusters, 4608 intended requests. Terminal counts: output = 4602; data_unverified = 6.

The native sentiment range is $[-3, 3]$ with operational failure penalty 6.

Strict. F/D/R/U denote format errors, data restrictions, runtime errors and unattempted requests, respectively.

Model/input	Valid/intended	Valid MAE	Operational risk [CI]	F/D/R/U
Q25/AV	511/512	0.9489	0.9587 [0.8840, 1.0319]	0/1/0/0
Q25/T	512/512	0.8077	0.8077 [0.7427, 0.8745]	0/0/0/0
Q25/TAV	511/512	0.8174	0.8275 [0.7582, 0.8966]	0/1/0/0
Mini/AV	486/512	0.7761	1.0414 [0.9287, 1.1603]	25/1/0/0
Mini/T	512/512	0.7430	0.7430 [0.6843, 0.8001]	0/0/0/0
Mini/TAV	493/512	0.7094	0.9057 [0.8088, 1.0067]	18/1/0/0
Q3/AV	100/512	0.6750	4.9600 [4.7645, 5.1526]	411/1/0/0
Q3/T	3/512	1.2778	5.9723 [5.9373, 6.0000]	509/0/0/0
Q3/TAV	69/512	0.6860	5.2839 [5.1129, 5.4354]	442/1/0/0

Every predefined overall comparison is retained below. Differences are left minus right; lower error is better. MELD differences concern classification error, not accuracy. Valid-pair columns can have different denominators; they are not differences between two separate valid-only means.

Comparison	Valid pairs	Intended error difference [CI]	Valid-pair error difference [CI]
Q25/T - TAV	511	-0.0198 [-0.0655, 0.0270]	-0.0149 [-0.0582, 0.0300]
Q25/AV - TAV	511	0.1312 [0.0581, 0.2038]	0.1315 [0.0582, 0.2042]
Mini/T - TAV	493	-0.1628 [-0.2637, -0.0609]	0.0352 [-0.0176, 0.0868]
Mini/AV - TAV	478	0.1357 [0.0384, 0.2499]	0.0630 [0.0134, 0.1155]
Q3/T - TAV	0	0.6885 [0.5345, 0.8638]	NA [NA]†
Q3/AV - TAV	29	-0.3239 [-0.5494, -0.1063]	-0.1494 [-0.3200, 0.0128]†
Q25 - Mini/TAV/clean	493	-0.0783 [-0.1753, 0.0185]	0.1108 [0.0514, 0.1691]
Q25 - Q3/TAV/clean	69	-4.4564 [-4.6293, -4.2657]	0.1208 [-0.0219, 0.2608]
Mini - Q3/TAV/clean	67	-4.3781 [-4.5682, -4.1729]	0.0060 [-0.1360, 0.1576]

Complete-fence normalization. F/D/R/U denote format errors, data restrictions, runtime errors and unattempted requests, respectively.

Model/input	Valid/intended	Valid MAE	Operational risk [CI]	F/D/R/U
Q25/AV	511/512	0.9489	0.9587 [0.8840, 1.0319]	0/1/0/0
Q25/T	512/512	0.8077	0.8077 [0.7427, 0.8745]	0/0/0/0
Q25/TAV	511/512	0.8174	0.8275 [0.7582, 0.8966]	0/1/0/0
Mini/AV	486/512	0.7761	1.0414 [0.9287, 1.1603]	25/1/0/0
Mini/T	512/512	0.7430	0.7430 [0.6843, 0.8001]	0/0/0/0
Mini/TAV	493/512	0.7094	0.9057 [0.8088, 1.0067]	18/1/0/0
Q3/AV	480/512	0.6340	0.9694 [0.8495, 1.0869]	31/1/0/0
Q3/T	512/512	0.7079	0.7079 [0.6467, 0.7652]	0/0/0/0
Q3/TAV	505/512	0.6729	0.7458 [0.6670, 0.8275]	6/1/0/0

Every predefined overall comparison is retained below. Differences are left minus right; lower error is better. MELD differences concern classification error, not accuracy. Valid-pair columns can have different denominators; they are not differences between two separate valid-only means.

Comparison	Valid pairs	Intended error difference [CI]	Valid-pair error difference [CI]
Q25/T - TAV	511	-0.0198 [-0.0655, 0.0270]	-0.0149 [-0.0582, 0.0300]
Q25/AV - TAV	511	0.1312 [0.0581, 0.2038]	0.1315 [0.0582, 0.2042]
Mini/T - TAV	493	-0.1628 [-0.2637, -0.0609]	0.0352 [-0.0176, 0.0868]
Mini/AV - TAV	478	0.1357 [0.0384, 0.2499]	0.0630 [0.0134, 0.1155]
Q3/T - TAV	505	-0.0378 [-0.1053, 0.0211]	0.0290 [-0.0121, 0.0671]
Q3/AV - TAV	476	0.2236 [0.1083, 0.3399]	-0.0417 [-0.0959, 0.0149]
Q25 - Mini/TAV/clean	493	-0.0783 [-0.1753, 0.0185]	0.1108 [0.0514, 0.1691]
Q25 - Q3/TAV/clean	505	0.0817 [0.0115, 0.1527]	0.1320 [0.0752, 0.1919]
Mini - Q3/TAV/clean	487	0.1600 [0.0519, 0.2720]	0.0218 [-0.0270, 0.0770]

Fence plus numerical-quote normalization. All statistics equal the preceding complete-fence normalization tier; no additional response is recovered.

G DATA PROVENANCE AND DEPENDENCE

We acquired the authors' supervised CH-SIMS v2 feature package with 2,722/647/1,034 training/validation/test examples. Its continuous text, acoustic, and visual dimensions are 768/25/177; their padded temporal extents are 50/925/232. Acoustic and visual valid lengths come from the package. Text valid lengths come from its original prefix attention masks. Token IDs, masks, and segment IDs are not continuous text features. We do not re-extract or retokenize the stored representations. A token audit found shorter retained prefixes and some case-normalization differences relative to a current tokenizer; it does not reconstruct the historical encoder revision.

The metadata CSV and feature package agree on all 4,403 identities, transcripts, and split assignments, but 2,289 rows differ in at least one sentiment label. We consistently use the package's native regression labels for all supervised losses, validation selection, and evaluation. The CSV provides verified source identities, not replacement targets. This decision follows the authors' clarification that their experiments used the package regression labels and that a subsequent correction concerned categorical annotations (official issue 5). It was fixed before model fitting and was not chosen from predictive results.

An exact-feature audit found one cross-split equivalence class: training `video_0049$_$0008` and test `video_0049$_$0007`. Their three complete feature arrays, valid lengths (18/209/53), transcript and packed tokens agree. Their overall labels agree, but two unimodal labels differ. Equality of features, not equality of labels, defines the exclusion. We remove only the training member in a derived release and preserve all validation/test examples, remaining targets and order. All methods use the resulting 2,721/647/1,034 split and a freshly fitted training scaler. This is test-input-aware deduplication before any fitted-model output. It does not remove near duplicates or establish raw-media identity.

The authors separately report that a corrected release has 2,721 training examples (official issue 4). The matching count does not establish that our removed example is their correction. We identify our acquired payload and derived data explicitly and do not claim exact reproduction of a final corrected official release.

The exact acquired and derived artifacts have the following SHA-256 identities. These distinguish a data version; they do not replace the provider's access terms or certify upstream extraction.

- Acquired feature pickle: `c5327d86ff1521395935d7b2929a3d3a56a9770a32edeb1feb16eac89177f8c5`.
- Acquired metadata CSV: `be7e482640e7ee24d2070033e1e2aef61ce183199286a8a18b0d6247556bf150`.
- Derived preparation manifest: `e83e155fdda50581aedb7997f814a8e4228a6dfa5da795e265e7b7018eff4efd`.

No nonfinite feature values were found. However, 51 training examples have all-zero visual features despite positive supplied lengths. We retain their original availability metadata. This observation does not distinguish missing raw video, extraction failure and other preprocessing conditions. In addition, the fitting scope of upstream feature extraction or normalization is not certified by the package. Our train-only checks apply to the additional standardization performed here.

Table 27: Accepted split composition at the fixed label threshold 0.2. A: same strong polarity in all modalities; D: opposing strong polarities; N: all other configurations.

Split	Clips	Sources	A	D	N
Training	2721	129	1435	347	939
Validation	647	120	327	115	205
Test	1034	142	459	149	426

There are 119 shared source videos between training and validation, 127 between training and test, and 117 between validation and test. The disagreeing test stratum spans 70 sources. Source identities do not certify speaker identity. Source-cluster intervals account for within-test source dependence; they do not undo train–test source overlap or establish performance on unseen speakers.

H COMPLETE INTERVENTION PROTOCOL

Each continuous feature dimension is standardized over valid retained training time steps only, using float64 population moments and a standard-deviation floor of 10^{-6}. Transformation uses float64 before returning to the array dtype; padding is unchanged and does not enter moment estimation or pooling. Gaussian interventions independently add zero-mean noise of standard deviation 0.1 or 0.3 in standardized feature units at valid positions. Temporal dropout uses probability 0.1 or 0.3 and zeroes selected complete time-step vectors, without inverted-dropout rescaling. Neither operation changes the supplied length. The model does not receive intervention masks, rates, names, or labels. If temporal dropout selects no valid time step, its applied-intervention target is zero, rather than the requested operator's nominal positive label.

For each original training example and epoch, a stable hash of example ID, training seed and epoch selects one of the $3 \times 2 \times 2 = 12$ modality/operator/severity combinations. Augmented methods receive the same original and paired view for the same seed and epoch. Loss components average the two views before applying their weights. The batch size counts 64 original examples, not 128 independent observations. The original-only variants use a single view. Shuffling uses the fixed offset 10^6 plus the training seed.

Evaluation realizations depend on example identity, modality, operator, severity and replicate; they do not depend on the fitted seed or batching. Each fitted model has one original inference sweep, 36 seen-intervention views (12 conditions times three replicas), 18 held-out block views (six conditions times three replicas), three individual missing-modality views and one all-missing view. Thus each fitted model has 59 views. Missingness and held-out operators are reported separately from the primary seen-intervention endpoint.

For held-out block removal, a sequence of valid length n receives a block of width $\min(n, \max(1, \lfloor pn + 0.5 \rfloor))$ for $p \in \{0.1, 0.3\}$, with a uniform valid start. The operation replaces the block with zeros but does not change length or availability. A short sequence may be completely erased; the audit retains and counts this outcome. Explicit missingness instead masks availability. All-missing inference remains a model output and is compared with the constant median of the training overall labels (zero), rather than being silently replaced by that prior.

I ARCHITECTURE, OPTIMIZATION, AND ABLATIONS

This section describes the supervised feature study, separately from the zero-shot raw-media experiments. Let u_m be a learned utility logit, h_m its modality representation, and d_m the intervention-detector logit. The factorized gate uses

$$q_m = \mathrm{stopgrad}(1 - \sigma(d_m)), \qquad w_m = \mathrm{softmax}_{\mathrm{available}}\big(u_m + \log(q_m + 10^{-6})\big).$$

The utility control uses u_m alone but retains the detector and the same supervision. Thus, their contrast tests inserting the detector in the gate, not whether an extra supervised objective helps. The detector receives $\mathrm{stopgrad}(h_m)$ by default. It detects assigned feature interventions, not independently measured perceptual quality.

Each encoder applies linear($2d_m$,128), ReLU, dropout(0.2), linear(128,128), ReLU to concatenated masked mean and population standard deviation. Variance has a floor of 10^{-8} before square root; an empty modality has an exactly zero pooled vector and is masked after encoding. The fusion head is linear(384,128), ReLU, dropout(0.2), linear(128,1). Unimodal heads are linear(128,1). The utility head is linear(384,128), ReLU, linear(128,3); each intervention head is linear(128,128), ReLU, linear(128,1). These dimensions describe width 128; the wider multitask control replaces 128 with 155 and 384 with 465 throughout the entire head inventory, including allocated heads unused by its losses. Predictions are real-valued, without post-hoc test calibration.

The direct and multitask variants concatenate the masked embeddings. Gated variants concatenate $3w_m h_m$. Uniform weights reproduce the direct scale when all three modalities are present, but not under partial availability. All variants instantiate the same head inventory, including heads unused by a particular loss or prediction path. Equal allocated parameter counts do not mean identical effective capacity. No unavailable modality receives a positive gate, and all gates are zero when all inputs are absent.

All formal runs use AdamW (learning rate 10^{-3}, weight decay 10^{-4}), gradient clipping at norm 5, up to 80 epochs, and patience 12. The selected checkpoint is the earliest strict minimum original validation MAE. Training seeds are 11, 29 and 47. An epoch contains 43 updates for 2,721 examples at batch size 64. Common maximum budgets and stopping rules do not imply equal realized epochs, updates, time, or FLOPs.

Table 28: Loss weights and views. Every variant includes overall L1 at weight 1. Aux: unimodal L1; Det: applied-intervention BCE; Target: target-agreement KL. The table specifies diagnostic controls, not replications of published systems.

Variant	Views/example	Aux	Det	Target
Direct, original	1	0	0	0
Direct, augmented	2	0	0	0
Multitask, original	1	1	0	0
Multitask, augmented	2	1	0	0
Agreement, augmented	2	1	0	0
Target-agreement, augmented	2	1	0	0.2
Utility, augmented	2	1	0.2	0
Detector-only gate, augmented	2	1	0.2	0
Factorized, augmented	2	1	0.2	0
Factorized joint, augmented	2	1	0.2	0
Factorized no auxiliary, augmented	2	0	0.2	0
Factorized, original	1	1	0.2	0
Wider multitask, augmented	2	1	0	0

Unimodal L1 averages over available sample–modality entries. Detector BCE uses originally available entries, including the original view's negative targets. Agreement logits use the negative mean absolute difference from other available unimodal predictions, divided by 0.5. With one present modality this logit is zero. The target-agreement control trains the utility distribution toward the available-modality soft target proportional to $\exp(-|y_m - y_M|/0.5)$, using training labels only.

Default detectors receive detached representations, and task prediction cannot backpropagate through their probabilities. The joint variant removes both boundaries simultaneously; it cannot isolate which boundary matters. Removing unimodal supervision leaves its unimodal heads untrained, so their recovery metrics are undefined. Original-only factorized training retains BCE but has zero positive intervention training examples. It removes both augmentation and positive detector supervision; it is not a pure single-factor augmentation ablation.

The wider multitask control uses width 155, selected by a parameter-count formula before fitting. For summed input dimension $d = 970$, $P_{\mathrm{MTL}}(h) = 6h^2 + (2d + 11)h + 4$, while $P_F(128) = 12(128)^2 + (2d + 21)128 + 10$. We minimize their absolute difference over integer widths 128–256, breaking ties toward the smaller width. This is a formula evaluation, not 129 trained candidates. The resulting training-active counts are 446,559 and 447,626 (a 1,067-parameter, 0.238% difference). This controls approximate training parameter budget, not FLOPs or the exact capacity of the prediction path.

J Uncertainty and supplementary analyses

For each fitted seed, we compute metrics from individual predictions before averaging results across the three seeds. We never average predictions first. Seen losses average three replicas within each condition and then weight the 12 conditions equally. Source-cluster bootstrap uses 2,000 draws with seed 0, shared across methods, fitted seeds, conditions and groups. A sampled source contributes all its examples with its sampling multiplicity. We then restrict to the relevant group and compute its metric. Intervals condition on the three fitted models; sample standard deviation across seeds is separately reported. For paired increments, between-method absolute-risk contrasts and explicitly labeled supplementary risks, intervals are the 2.5th and 97.5th bootstrap percentiles. The main single-method absolute-risk tables report seed SD only. Empty-group draws retain their counts; insufficient valid draws yield undefined intervals. Accuracy, F1 and correlation are reported with

fitted-seed dispersion, not bootstrap intervals. Uncorrected intervals across many groups are not family-wise hypothesis tests.

Regression metrics use all applicable targets. Primary binary metrics exclude true-zero targets and classify predictions ≥ 0 as positive. An auxiliary convention includes true-zero targets as positive. Both macro and weighted F1 are retained. Macro-F1 uses both fixed classes with zero for an absent class; an empty classification set is undefined. Correlation is undefined for a constant sequence or fewer than two examples. Sparse groups (fewer than 50 examples or 10 verified sources) remain visible and are exploratory for worst-group interpretation.

The predefined primary paired-increment contrasts are factorized minus utility and factorized minus augmented multitask. Supplementary ablation contrasts are no-auxiliary minus factorized, original-only factorized minus augmented factorized, wider minus standard augmented multitask, and factorized minus wider multitask. The supplementary contrasts report original MAE, seen absolute MAE and seen paired increment for overall/A/D/N, retaining all comparisons.

Threshold sensitivity uses 0.1 and 0.3 in addition to the fixed 0.2. A further analysis fixes the median nonzero absolute overall sentiment in training at $q_M = 0.6$ (2,489 nonzero training targets). It creates five bins: zero and the two polarities crossed with $0 < |y_M| \leq q_M$ versus $|y_M| > q_M$. Crossing these bins with $A/D/N$ gives at most 15 cells. Cut points remain fixed in bootstrap draws. Empty/sparse cells are retained; bins are not merged after seeing test risks. This is descriptive target stratification, not confounding removal or causal adjustment.

For factorized, utility and multitask augmentation, the fixed deletion diagnostic reports, for each originally available modality, $I_m = |f(x_{-m}) - y_M| - |f(x) - y_M|$ and $J_m = |f(x_{-m}) - f(x)|$. Pre-existing missing inputs form a separate identity check with absolute tolerance 10^{-6}, never a filtering rule. Deletion changes availability, gate normalization and input distribution; I_m/J_m cannot establish that a model used opposing emotional semantics. Gate weights are competitive fusion weights; no gate-to-intervention ranking score and direction were predefined, so gate AUROC/AP are not reported. Ranking metrics do not themselves require calibrated probabilities. Detector AUROC/AP instead evaluate the explicit applied-intervention target within modality. Paired diagnostics concatenate original and corresponding intervention views using original availability. Repeated clips are not independent samples and these diagnostics have no bootstrap intervals. Untrained detectors and single-class targets yield undefined AUROC/AP. Explicit missingness exposes availability and gives the missing detector a zero embedding; its output may reflect only biases, not perceptual quality recognition.

K Execution integrity and resources

All 39 formal fits completed with normal early stopping. Each saves training history, the selected checkpoint, validation predictions and identities, and content hashes. The main 30 fits completed 753 epochs and 32,379 updates; the additional nine completed 212 epochs and 9,116 updates. Their summed training loop time was 2,290.876 seconds (38.18 minutes); the interval from the first start to the final completion was 40.94 minutes. The separate five-fit feasibility stage took 135.073 training-loop seconds. Summed fit timers include in-loop validation and checkpoint/history I/O, but exclude data preparation and final test inference/analysis. The first-start-to-last-end interval also includes between-fit gaps. Neither measures the full GPU allocation or queue duration.

Model execution used one NVIDIA H20, Python 3.10.12, PyTorch 2.7.1+cu128, NumPy 2.2.6, CUDA 12.8 and cuDNN 90701. The maximum recorded PyTorch allocated tensor peak was 195,417,600 bytes (186.365 MiB); this is not total device/process memory or reserved memory. No FLOP or energy measurement was made. Independent checks covered the dataset export, fixed execution configurations, training completions and the definitions of analysis. Software fixtures were kept separate from empirical results. Predictions, intervention audits, completion records and statistics have separate integrity checks. Such checks reduce pipeline errors; they are not an independent scientific replication.

L SUPPLEMENTARY NUMERICAL RESULTS

All entries below are transcribed from the sealed, predefined statistics. They do not introduce new subgroups or resampling. The sentiment bins use the training cut point $q_M = 0.6$; source intervals condition on that cut point and the fitted seeds. The natural grouping remains observational. The complete machine-readable output retains full precision, per-seed estimates, undefined cases and source pointers.

L.1 MAIN FEATURE CONTROLS AND HELD-OUT CONDITIONS

These are supervised feature-model results, not zero-shot Omni comparisons.

Table 29: Main feature-study controls. R_0 and R_C are original and seen-intervention MAE; parentheses contain sample SD over three fitted seeds, not confidence intervals. D is the annotation-disagreeing group.

Method	Overall R_0	Overall R_C	D: R_0	D: R_C
Direct (orig.)	0.2858 (0.0026)	0.2895 (0.0014)	0.4470 (0.0141)	0.4528 (0.0147)
Direct (aug.)	0.2818 (0.0082)	0.2836 (0.0072)	0.4326 (0.0066)	0.4332 (0.0088)
Multitask (orig.)	0.2773 (0.0052)	0.2806 (0.0035)	0.4191 (0.0088)	0.4244 (0.0057)
Multitask (aug.)	0.2763 (0.0049)	0.2777 (0.0048)	0.4487 (0.0223)	0.4482 (0.0236)
Agreement	0.2799 (0.0022)	0.2817 (0.0024)	0.4339 (0.0046)	0.4334 (0.0045)
Target agreement	0.2724 (0.0048)	0.2730 (0.0038)	0.4391 (0.0264)	0.4369 (0.0233)
Utility	0.2705 (0.0037)	0.2718 (0.0031)	0.4480 (0.0196)	0.4457 (0.0173)
Detector-only gate	0.2744 (0.0021)	0.2766 (0.0026)	0.4292 (0.0048)	0.4303 (0.0054)
Factorized	0.2722 (0.0043)	0.2742 (0.0034)	0.4631 (0.0095)	0.4629 (0.0094)
Factorized joint	0.2765 (0.0052)	0.2780 (0.0058)	0.4532 (0.0170)	0.4497 (0.0180)

Table 30: Held-out feature conditions, averaged over three seeds (sample SD). Block removal and explicit missingness were not included in the main seen-intervention training family. These columns are absolute MAE.

Method	Held-out block	One missing	All missing
Direct (orig.)	0.2921 (0.0009)	0.3579 (0.0060)	0.5077 (0.0014)
Direct (aug.)	0.2856 (0.0062)	0.3570 (0.0031)	0.5081 (0.0017)
Multitask (orig.)	0.2840 (0.0027)	0.3535 (0.0025)	0.5080 (0.0017)
Multitask (aug.)	0.2788 (0.0044)	0.3465 (0.0028)	0.5080 (0.0018)
Agreement	0.2836 (0.0022)	0.3697 (0.0066)	0.5082 (0.0025)
Target agreement	0.2743 (0.0023)	0.3704 (0.0204)	0.5080 (0.0020)
Utility	0.2733 (0.0027)	0.4605 (0.0106)	0.5079 (0.0021)
Detector-only gate	0.2790 (0.0032)	0.3447 (0.0019)	0.5082 (0.0021)
Factorized	0.2759 (0.0036)	0.4584 (0.0303)	0.5083 (0.0013)
Factorized joint	0.2793 (0.0064)	0.4221 (0.0439)	0.5091 (0.0020)

Table 31: Additional feature ablations and their comparators. Entries are original risk, seen risk and paired increment, averaged over three seeds (sample SD). Original-only factorized training removes augmentation and positive intervention supervision together.

Method	R_0	R_C	Δ_C
Factorized no aux.	0.2791 (0.0032)	0.2818 (0.0045)	0.0027 (0.0013)
Factorized (orig.)	0.2671 (0.0048)	0.2723 (0.0059)	0.0051 (0.0017)
Wider multitask	0.2750 (0.0043)	0.2759 (0.0045)	0.0009 (0.0007)
Factorized	0.2722 (0.0043)	0.2742 (0.0034)	0.0020 (0.0010)
Multitask (aug.)	0.2763 (0.0049)	0.2777 (0.0048)	0.0014 (0.0001)

L.2 TRAINING-FITTED TARGET BINS

Table 32: All five target bins crossed with all three annotation strata. F denotes factorized augmentation; n/s denotes clips/source clusters. Sparse means fewer than 50 test examples or 10 sources. The zero-target aligned cell has one source and no valid confidence interval; it remains visible. Values are native-label MAE or paired increments.

Bin	Group	Train n/s	Test n/s	Sparse	R_0	R_C	Δ_C
negative low	A	414/113	131/66	no	0.1983	0.1977	-0.00066
negative low	D	120/60	56/31	no	0.4387	0.4371	-0.00158
negative low	N	319/103	152/70	no	0.2453	0.2451	-0.00025
negative high	A	421/119	144/81	no	0.2450	0.2481	+0.00307
negative high	D	28/22	6/6	yes	0.8965	0.8943	-0.00221
negative high	N	51/39	14/13	yes	0.5112	0.5145	+0.00324
zero	A	10/9	1/1	yes	0.5268	0.5244	-0.00235
zero	D	34/26	16/16	yes	0.3754	0.3753	-0.00007
zero	N	188/69	77/50	no	0.2849	0.2864	+0.00144
positive low	A	312/104	99/60	no	0.1718	0.1729	+0.00116
positive low	D	127/72	60/40	no	0.4713	0.4718	+0.00055
positive low	N	348/101	162/83	no	0.2518	0.2544	+0.00254
positive high	A	278/90	84/55	no	0.2281	0.2397	+0.01166
positive high	D	38/28	11/9	yes	0.4337	0.4373	+0.00367
positive high	N	33/26	21/17	yes	0.3467	0.3533	+0.00657

L.3 INTERVENTION DETECTION AND UNIMODAL RECOVERY

Table 33: Paired original/intervention detector AUROC for every seen and held-out block condition, for the overall test group. F/U/Joint are augmented factorized, utility and joint factorized models. G/D/B denote Gaussian noise, temporal dropout and held-out block masking. Subscripts 0/1/2 denote replicas; each entry is averaged over three fitted seeds. The paired pool contains two scores per base example, not twice as many independent examples. Pools containing only original-input views, and other single-class evaluation pools, have undefined AUROC/AP. Higher means more detectable assigned interventions, not better sentiment prediction or physical input-quality estimation.

Condition	F_0	F_1	F_2	U_0	U_1	U_2	$Joint_0$	$Joint_1$	$Joint_2$
G/T/0.1	0.4990	0.4994	0.4987	0.4988	0.4990	0.4988	0.4990	0.4992	0.4988
G/T/0.3	0.4910	0.4924	0.4914	0.4896	0.4918	0.4893	0.4926	0.4937	0.4923
G/A/0.1	0.5007	0.5005	0.5006	0.5004	0.5001	0.5007	0.5000	0.4996	0.4996
G/A/0.3	0.5034	0.5035	0.5029	0.5006	0.5002	0.5000	0.4941	0.4940	0.4937
G/V/0.1	0.6334	0.6330	0.6332	0.6234	0.6229	0.6237	0.9432	0.9432	0.9432
G/V/0.3	0.8924	0.8917	0.8915	0.8706	0.8703	0.8702	0.9932	0.9932	0.9932
D/T/0.1	0.6118	0.5870	0.6029	0.6220	0.6010	0.6109	0.6195	0.5903	0.6072
D/T/0.3	0.7240	0.7297	0.7276	0.7305	0.7368	0.7315	0.7315	0.7338	0.7325
D/A/0.1	0.5762	0.5762	0.5759	0.5808	0.5797	0.5806	0.6100	0.6094	0.6083
D/A/0.3	0.7412	0.7403	0.7392	0.7551	0.7530	0.7532	0.8239	0.8227	0.8208
D/V/0.1	0.8462	0.8453	0.8476	0.8268	0.8260	0.8304	0.9816	0.9827	0.9826
D/V/0.3	0.9534	0.9527	0.9527	0.9399	0.9385	0.9392	0.9920	0.9919	0.9920
B/T/0.1	0.5727	0.5745	0.5721	0.5740	0.5760	0.5741	0.5746	0.5748	0.5752
B/T/0.3	0.7240	0.7204	0.7206	0.7244	0.7234	0.7204	0.7293	0.7294	0.7256
B/A/0.1	0.5733	0.5721	0.5714	0.5771	0.5763	0.5754	0.6055	0.6052	0.6036
B/A/0.3	0.7203	0.7247	0.7239	0.7290	0.7320	0.7311	0.8009	0.8048	0.8063
B/V/0.1	0.8563	0.8564	0.8568	0.8368	0.8364	0.8368	0.9844	0.9843	0.9844
B/V/0.3	0.9549	0.9549	0.9547	0.9425	0.9420	0.9417	0.9916	0.9915	0.9918

Table 34: Original unimodal-target MAE for all listed methods and annotation groups. All models use augmentation; MTL is multitask fusion. Each entry is a fitted-seed mean (sample SD). Recovery is not a test of fusion's use of that target. Lower is better.

Method	Group	Text MAE (SD)	Audio MAE (SD)	Vision MAE (SD)
F	overall	0.2395 (0.0065)	0.4234 (0.0055)	0.2983 (0.0032)
F	A	0.2290 (0.0052)	0.5219 (0.0086)	0.3702 (0.0055)
F	D	0.2750 (0.0025)	0.4551 (0.0119)	0.2968 (0.0058)
F	N	0.2384 (0.0102)	0.3061 (0.0026)	0.2214 (0.0053)
U	overall	0.2395 (0.0044)	0.4220 (0.0025)	0.3080 (0.0061)
U	A	0.2447 (0.0053)	0.5169 (0.0083)	0.3770 (0.0044)
U	D	0.2542 (0.0099)	0.4544 (0.0186)	0.3084 (0.0178)
U	N	0.2287 (0.0123)	0.3085 (0.0089)	0.2337 (0.0057)
MTL	overall	0.2342 (0.0082)	0.4280 (0.0065)	0.3073 (0.0032)
MTL	A	0.2282 (0.0045)	0.5341 (0.0229)	0.3796 (0.0050)
MTL	D	0.2624 (0.0109)	0.4436 (0.0208)	0.3023 (0.0096)
MTL	N	0.2307 (0.0167)	0.3082 (0.0050)	0.2311 (0.0093)
Joint	overall	0.2422 (0.0064)	0.4211 (0.0091)	0.3056 (0.0081)
Joint	A	0.2391 (0.0099)	0.5147 (0.0250)	0.3684 (0.0103)
Joint	D	0.2627 (0.0023)	0.4544 (0.0149)	0.3073 (0.0163)
Joint	N	0.2382 (0.0052)	0.3086 (0.0065)	0.2373 (0.0063)

L.4 ALL PREDEFINED SUPPLEMENTARY CONTRASTS

Positive entries mean larger error or increment for the left-hand method. Intervals are 95% source-cluster percentiles conditional on fitted seeds, without multiplicity adjustment. The wider control approximately matches training-active parameters, not FLOPs. Original-only factorized training also removes positive detector supervision. No-auxiliary recovery heads are untrained. All four comparisons, groups and risks are retained below.

Table 35: Factorized no auxiliary supervision minus Factorized. Three-seed mean differences and 95% source-cluster intervals.

Group	Difference in R_0	Difference in R_C	Difference in Δ_C
overall	+0.00693 [-0.00043, +0.01419]	+0.00762 [+0.00052, +0.01480]	+0.00069 [-0.00002, +0.00142]
A	+0.01163 [+0.00250, +0.02106]	+0.01258 [+0.00383, +0.02178]	+0.00095 [+0.00015, +0.00172]
D	-0.01428 [-0.03699, +0.00733]	-0.01235 [-0.03473, +0.00904]	+0.00193 [-0.00024, +0.00406]
N	+0.00927 [-0.00070, +0.01963]	+0.00926 [-0.00023, +0.01944]	-0.00001 [-0.00128, +0.00124]

Table 36: Factorized original-only training minus Factorized augmented training. Three-seed mean differences and 95% source-cluster intervals.

Group	Difference in R_0	Difference in R_C	Difference in Δ_C
overall	-0.00505 [-0.00938, -0.00101]	-0.00192 [-0.00638, +0.00214]	+0.00313 [+0.00141, +0.00485]
A	-0.00541 [-0.01121, +0.00037]	-0.00548 [-0.01127, -0.00008]	-0.00007 [-0.00207, +0.00192]
D	-0.01110 [-0.02679, +0.00075]	-0.00521 [-0.02044, +0.00774]	+0.00589 [+0.00055, +0.01107]
N	-0.00255 [-0.00872, +0.00374]	+0.00306 [-0.00357, +0.00958]	+0.00561 [+0.00301, +0.00815]

Table 37: Wider multitask minus multitask. Three-seed mean differences and 95% source-cluster intervals.

Group	Difference in R_0	Difference in R_C	Difference in Δ_C
overall	-0.00123 [-0.00604, +0.00349]	-0.00180 [-0.00634, +0.00279]	-0.00057 [-0.00112, -0.00006]
A	+0.01353 [+0.00615, +0.02079]	+0.01287 [+0.00592, +0.01987]	-0.00066 [-0.00144, +0.00012]
D	-0.01911 [-0.03153, -0.00753]	-0.01924 [-0.03142, -0.00794]	-0.00013 [-0.00132, +0.00104]
N	-0.01088 [-0.01851, -0.00361]	-0.01149 [-0.01902, -0.00430]	-0.00062 [-0.00139, +0.00009]

Table 38: Factorized minus wider multitask. Three-seed mean differences and 95% source-cluster intervals.

Group	Difference in R_0	Difference in R_C	Difference in Δ_C
overall	-0.00286 [-0.00945, +0.00345]	-0.00171 [-0.00794, +0.00429]	+0.00115 [+0.00040, +0.00193]
A	-0.02142 [-0.03038, -0.01286]	-0.02004 [-0.02867, -0.01180]	+0.00138 [+0.00035, +0.00245]
D	+0.03346 [+0.01080, +0.05781]	+0.03394 [+0.01333, +0.05637]	+0.00048 [-0.00189, +0.00281]
N	+0.00442 [-0.00454, +0.01357]	+0.00557 [-0.00315, +0.01439]	+0.00115 [-0.00002, +0.00234]