Evaluating and Addressing Multimodal Hallucination

Pranav Avadhanam Aarya Deshpande Rishit Malpani Gurraj Singh
Kunal Singh
Samad Syed Vikram Varikooty Jeffrey Wang

1 Introduction

Multimodal Large Language Models (MLLMs) are models that can process text, visuals, and even audio to answer questions that require understanding across multiple modalities: for example emotion recognition and video question-answering. One major issue that continues to occur with MLLMs is that they exhibit *modality bias*: where they tend to overrely on certain modalities (often text), rather than using information from other modalities that might be essential to a given task (Chen et al., 2025; Bhosale et al., 2023). For example, an MLLM detecting the emotion of a person saying "I feel great!" can often over-rely on the lexical content (ie. the words involved) while ignoring the audial tones that suggest they are being sarcastic.

This problem is worth studying because it is at the heart of what makes multimodal models multimodal: if a model is not truly integrating all modalities, the ultimate advantage of multimodal models over normal text-based models is not being realized. Solving this problem can lead to more reliable applications of MLLMs in healthcare (working across medical images, clinical recordings, medical notes) assistive robotics, algorithmic trading, and other multimodal applications. That is, this work can help enable models that accurately process the visuals and audio that they ingest from the real world, in conjunction with a text-based backbone.

2 Hypothesis and Research Question

Overarching Research Question: To what extent are MLLMs modality-biased and how do you resolve this?

Specific Hypothesis:

H1: Modality-separated PragCoT (see related works) improves performance on the MUSTARD++ Sarcasm Benchmark because it addresses Modality Attribution Hallucination wherein a model makes up whether it has truly used a modality or not.

H2.1: On a given video Q&A, multimodal models are reliable in attributing what was the most important modality that contributed to their response (Attribution Faithfulness Score (AFS) > .75) because of sufficient latent understanding of how modalities interact. See Methodology for in-depth definition of AFS.

H2.2: MLLMs have a high Lexical Override Rate (LOR) (> .5) because previous work (Chen et al. 2025) has suggested Lexical Bias as a core defect of MLLMs.

H2.3: MLLMs have a negative Transcript Injection Bias. i.e. adding a transcript that matches audio lexical content leads to better performance on AVUT than not adding a matching transcript.

H3.1: Forcing modality separation in prompting (Modality Separated Prompting) on the AVUT dataset will lead to improved performance because it addresses Modality Attribution Hallucination by explicitly forcing the model to look at each modality separately.

H3.2: Using an empirically determined *Modality Importance Scoring* (relative importance of each modality for determining answers to the given benchmark; see Park et al. 2024) to weight how many

tokens are allocated to the text descriptions of each modality in Modality Separated Prompting will improve performance on the AVUT dataset because it more efficiently allocates a finite per-modality description budget to where it is needed the most.

H4: MLLMs hallucinate modality attribution in restricted-modality settings because they cite unavailable modality evidence in their reasoning. In the TV-only condition, the model sometimes references audio cues such as “tone” or “delivery,” suggesting that its explanations are not always faithfully grounded in the modalities provided.

H5: If audio-visual models genuinely integrate evidence across modalities, a contradictory transcript injected into the text channel should not be sufficient to override correct answers when the audio and video continue to support them. We hypothesize the opposite: that Qwen2.5-Omni-7B and Gemma-3n-E2b abandon correct AV-grounded answers under adversarial text injection at a rate well above zero, and that those abandonments are specifically directed by the content of the injection rather than scattered randomly across alternatives.

3 Related Works

Modality bias in audio- and video-language models. A growing body of work documents that nominally multimodal models often default to a single modality, typically text, even when other channels carry the relevant information. Bhosale et al. (2023) showed this on multimodal sarcasm detection, where text alone often matched or exceeded full multimodal performance. (Chen et al. 2025) gave the most direct demonstration in the audio domain, showing that audio-language models on emotion-recognition tasks rely heavily on transcribed lexical content rather than acoustic cues. Their LISTEN protocol uses counterfactual ablation, systematically removing modalities to isolate each one's contribution, as a behavioral test for modality reliance. Our pipeline adapts this counterfactual logic from audio-LLMs to omnimodal Video-LLMs and extends it from "is the model using modality X" to "does the model know whether it is using modality X."

Mitigations: prompting and post-training. Several recent methods aim to push multimodal models away from text dominance. Saha et al. (2025) introduced PragCoT, a three-stage prompting framework (Perception, Decoding, Reasoning) that asks the model to describe each modality before decoding (ie. forming concrete classifications of irony/metaphor/facial action units) and reasoning over them. They acknowledged that PragCoT's perception phase is vulnerable to "hallucinated perception," where the model invents visual evidence to match the text it has already encoded. Li et al. (2026) (SarcasmMiner) approached the same problem through RL post-training with a generative reward model, raising MUStARD F1 from 59.83% to 70.22%. Wang et al. (2026) (EmotionThinker) demonstrated a related prosody-aware RL approach for speech emotion recognition.

Our contribution (see Experiment H3) turns the diagnostic methodology of counterfactual ablation, the systematic removal of modalities, into a functional inference pipeline: Modality-Separated Prompting (ModSep). Unlike PragCoT, which uses a single "omni-modal" pass, our Modality-Separated Prompting physically isolates the model’s perception: we force the model to generate a blind visual-only description and a blind text-only description and finally a blind audio-only description, all in separate passes. Finally, the model reasons off of these modality-separated evidences. One issue with PragCoT was that it relied on the model itself to “separate” a given video input into text, audio, and visual descriptions. However, as Li et al. have described in their analysis, telling a model to separate input across all 3 modalities is liable to hallucination: a model can make up that an observation it derived from text is from visuals, hence giving a false impression that the model is truly using all modalities and not suffering from modality bias. Explicitly separating the inputs by modality, *before* they ever reach the model, is one way of addressing this insufficiency in PragCoT.

To make more concrete asymmetry in the representation of modalities in question-answering, Park et al. introduced the notion of a *Modality Importance Score* (MIS) which is a way of calculating the relative importance of each modality with respect to a given dataset (for example: "Text" have a normalized weighting of 0.7 to suggest a dataset heavily relies upon Text) (Park et al. 2024). For example, if Text has weight of 0.7, then 70% of the total token budget allocated towards modality descriptions in the ModSep pipeline will be for a textual description of the provided video. To our knowledge, there is no prior work on applying MIS to *weight* the number of tokens used for each modality separated description (we are effectively combining MIS and ModSep).

While these methods all target the model's modality use behaviorally, they evaluate success through downstream accuracy rather than directly measuring whether the intervention reaches the model's internal modality reliance. We also provide a complementary diagnostic via Experiment H2: it measures the gap that these methods aim to close, and could in principle be applied to evaluate any of them.

Self-explanation and introspection in LLMs. Beyond multimodal settings, a separate literature questions whether language models' self-reports about their own reasoning faithfully reflect their internal computations (Lanham et al. 2023; Turpin et al. 2023). The general finding is that chain-of-thought rationales and post-hoc explanations can be plausible without being causally accurate. Our work imports this concern into the multimodal setting: when a Video-LLM reports relying on the audio, is that report faithful in the sense that ablating the audio changes the answer? To our knowledge, this specific framing, attribution faithfulness as a measurable quantity for multimodal models, has not been studied directly.

Datasets. We initially used MUStARD(++) (Castro et al., 2019; Bhosale et al., 2023), the standard sarcasm benchmark, but moved to AVUTBenchmark (Yang et al. 2025) for the reasons described in the results of H1: MUStARD prompting interventions did not beat the simple text+video baseline, and inspection of clips revealed label noise and short, low-context segments that confounded interpretation. Also, AVUT is purpose-built to be audio-centric and labels questions by task type, enabling per-task analysis. We also consider Video Q&A to be a richer task type to ask questions about than 1/0 sarcasm classification.

MUStARD++ (an expansion on the original dataset) consists of a set of a 8-10 second video clips from various television sitcoms with labels on whether the clip depicts sarcasm or not (Bhosale et al.). Overall dataset consists of 1365 examples evenly split between the two classes.

AVUT consists of around 2600 videos averaging 1 minute in length from a variety of media sources (YouTube, TikTok, Television) (Yang et al.). AVUT consists of a human-annotated (with QA questions and answers) subset of 698 videos and 1,734 QA 4 multiple-choice pairs known as AV-Human. AV-Human is what we consider in this paper. Crucially, AVUT has already filtered out questions that an LLM can answer simply based on the question text, making the dataset more indicative of multimodal abilities.

4 Methodology

4.1 H1

Our methodology follows the project pipeline shown in the repository. We used MUStARD as the main sarcasm evaluation set and Qwen2.5-Omni-7B as the multimodal model. The repo structure separates the work into setup, baseline testing, diagnosis, prompting, and fine-tuning. First, the data scripts prepare the MUStARD clips and transcripts. Then the baseline scripts test whether the model can detect sarcasm from the original multimodal inputs. After that, the diagnostic and prompting scripts test whether the model is actually using the intended modalities or just relying on text shortcuts.

To test modality importance, we used the per-modality descriptions that Qwen2.5-Omni-7B had already generated for each MUStARD clip. Instead of rerunning the whole model with different raw inputs, we kept the same saved descriptions and changed only which descriptions were passed into the final Yes/No gate. This gave us text-only, video-only, audio-only, audio+text, video+text, and fused video+audio+text

conditions. Because all conditions used the same model and the same final gate, the comparison focused on the value of each modality description rather than differences in model setup.

We evaluated these conditions on all 690 balanced MUStARD clips. For each condition, the final gate predicted whether the utterance was sarcastic. We then measured accuracy, F1, sarcasm recall, and Yes-rate. This setup was designed to answer a specific concern: maybe the earlier gating variant failed only because the gate was poorly calibrated. The ablation separates that issue from the modality question by asking what happens when the gate receives different modality evidence while everything else stays fixed.

The results showed that single-modality accuracy was very similar across text, video, and audio descriptions. However, audio had much higher sarcasm recall than text or video. This suggests that audio carries a strong sarcasm signal, especially for catching sarcastic clips. At the same time, audio also predicted "Yes" more often on non-sarcastic clips, which reduced its precision and kept its overall accuracy close to the other single-modality settings. The fused condition still performed best, which supports the idea that sarcasm detection needs multiple channels rather than one isolated modality.

After the quantitative ablation, we inspected the saved descriptions to understand the failure cases. We focused especially on clips where the audio-only gate predicted "No" even though the ground truth was sarcastic. In those cases, the final gate was often not the main problem. Instead, the audio description itself was wrong. For example, in clip 1_70, Penny's line is delivered with clear mocking emphasis, but the model describes the audio as flat and lacking dramatic pitch variation. Given that description, the gate's "No" prediction is reasonable. The failure happens earlier, when the model fails to describe the actual prosody.

This led to our leakage finding. When the model is asked to describe only audio, the description does not always reflect only prosody. It can also be influenced by the spoken words. In other words, the model does not cleanly separate "what was said" from "how it was said." This matters because our method depends on the assumption that a modality-specific prompt creates a modality-specific description. The results show that this assumption is not always safe for Qwen2.5-Omni-7B.

The overall methodology therefore combines three parts: controlled ablation, standard metric evaluation, and qualitative inspection. The ablation tells us which modality descriptions help the final gate. The metrics show how each condition behaves across the full dataset. The qualitative inspection explains why some conditions fail. Together, these steps show that the problem is not only whether the model predicts sarcasm correctly, but whether its intermediate modality descriptions are faithful to the evidence they are supposed to represent.

4.2 H2 Modality attribution diagnostic on AVUT

To test whether multimodal models can faithfully report which modality drove their answer, we ran a seven-stage ablation pipeline on 1,443 AVUT-Human questions using Qwen2.5-Omni-7B, and a parallel six-stage version on a 600-sample balanced subset (100 per task) using Gemma-3n-E2B. Both runs use the same Whisper transcripts and the same mismatched-transcript pairings.

Stage	Inputs	Purpose
S1	Question + options only (Qwen2.5-1.5B-Instruct)	Text-shortcut floor
S2	Audio + question	Audio-only
S3	Video frames + question	Visual-only
S4	Audio + video + question	Reference full-AV condition
S5	S4 + matched ASR transcript	Matched transcript injection
S6	S4 + raw response, then "Which modality did you use?"	Self-reported attribution
S7	S4 + transcript from a different same-task video	Lexical override probe (Gemma only)
S8	S4 + prosody-first prompt	Prosody verbalization (Qwen only)

The same 600 questions and the same Whisper transcripts and mismatched-transcript pairings are used for both models, making every cross-model comparison directly aligned. Qwen used 1,443 questions for stages S1 to S6 and S8 but did not run S7; Gemma used the 600-sample subset for S1–S7 but did not run S8. The Qwen pipeline ran ~17 hours on a single RTX 6000 Ada (48GB); the Gemma pipeline ran ~6 hours on a single A100-40GB.

Three diagnostic metrics combine the stages:

Attribution Faithfulness Score (AFS): For each S6 case where the model's self-reported #1 modality is *falsifiable* (i.e., not every single-modality stage already gives the same answer as S4), we check whether ablating that modality changes the answer relative to S4. AFS for a task is faithful / (faithful + confabulated) over the falsifiable subset. AFS = 1.0 means perfect attribution; AFS = 0.5 is chance. Trivially-unfalsifiable cases (every single-modality stage agrees with S4) are excluded from the denominator; without this filter, AFS would be artificially deflated.

Lexical Override Rate (LOR): Of the questions the model got *right* at S4, what fraction flip to a different answer at S7 (when handed a contradictory transcript)? LOR is a direct audio-vs-text trust probe. Computed only for Gemma since S7 was not run on Qwen.

Transcript Injection Bias (TIB): Acc(S4) − Acc(S5). Negative TIB means matched transcripts help; positive means they hurt.

We additionally report per-stage Expected Calibration Error (ECE) and $Delta$Confidence (the drop in self-reported confidence when a modality is removed) as exploratory signals.

4.3 H3

Following the results from Experiment H1 in which modality separation did not lead to improved performance on MUSTARD++ dataset, we investigate whether on the AV-Human (ie. AVUT-Human) dataset there would be improved performance. The rationale is that AVUT is sampled to truly test for multimodal (ie. more than solely textual) understanding and so a method, like Modality Separation, would fare better in the AVUT setting.

We work with Gemini API and its Gemini Flash 2.5 model due to time & compute restrictions.

We formally define ModalitySeparation for AVUT as follows:

ModalitySeparation: Given a fixed MCQ prompt, perform 3 separate passes (audio-only, visual-only, and text-only) where for each pass on modality X, the model M only receives modality X and no other modality. From that pass the model is instructed to output a text description of all the information that can be determined from that modality alone (desc_X).

We then perform an immediate answer classification, feeding the MLLM desc_X for X in {visual (ie. video frames), audio, and language}: this is ModalitySeparated Prompting.

ModalitySeparation + Reasoning: Alternatively, after receiving all desc_X's, we also run provided gemini 2.5 flash (a reasoning model) with a fixed internal reasoning budget of 768 tokens. This is ModalitySeparated Prompting + Reasoning.

Modality Importance Scoring (MIS): Following from Park et al. we consider how much each modality factors into a given answer on the AVUT dataset. The general idea is that if we can take a smaller sample of the AVUT dataset (with no intersection with the test set) and calculate a relative weighting of how importance the text_desc, audio_desc, and visual_desc's are to generating a final MCQ answer, we can use this weighting to provide more tokens for certain modality descriptions over others (ex. Given weightings of .3, .4, .3 and a 1536 token description budget, we may restrict text_desc, audio_desc, and visual_desc to contain max 460, 614, and 463 tokens). The hypothesis was that a more optimal allocation of description tokens will lead to better performance. The rest of the ModalitySeparation (+ Reasoning) pipeline is identical.

MIS is calculated as follows (see Park et al. 2024) for question q_i: MIS^i_X = perf(q_i | M_X^+) - perf(q_i | M_X^-). Where X is in M = {audio, visual, text}, M_X^+ is the set of all nonempty subsets of M containing X and M_X^- is the set of all nonempty subsets of M not containing X. We then sum MIS^i_X over all questions q_i for each X. The softmax of this sum is then taken to get final relative weighting: softmax($\Sigma_i MIS^i_X$).

Vanilla: Our baseline was passing a single mp4 file to gemini API with no additional prompting (though reasoning was possible).

In terms of experiment configurations: due to compute limitations a downsampling of 2 frames per second (and 3 frames per second for alignment tasks) was used and a video constant rate factor of 20 applied. As mentioned, the reasoning budget is 768 tokens. Per-modality description tokens vary from 512 to 1024 in the first and second diagrams respectively – this was done to see if the initial 512 description budget was too severe to gather all the information relevant to a question for a given modality. Consult README for full details.

MIS used a disjoint subset of 20 samples. Test runs were performed on the same subset of 30 samples from AV-Human (ideally this quantity would be larger given more compute).

Rationale: Given our research hypothesis is the evaluation of ModalitySeparation and application MIS as an additional inquiry, our approach (the direct evaluation of these models' performance on AV-Human) makes the most sense – especially considering the higher quality of AV-Human data in comparison to AVUT overall.

Final code can be found at https://github.com/PranavAvadhanam/proj-multimodal (with assistance from the Cursor IDE and Claude Code, see attached prompts).

4.4 H4

We used Qwen2.5-Omni-7B with a LoRA adapter trained on a 1,000-sample AVUT subset, then evaluated it on MUStARD sarcasm clips. The notebook first checks the AVUT and MUStARD data, fixes video paths, trains the LoRA model with swift sft, and then runs inference with the saved checkpoint. The same model setup is then used for the MUStARD reasoning experiments so that differences come from the input condition, not from changing models.

For the main comparison, we created two MUStARD reasoning datasets. The first condition is TAV, where the model is allowed to use text, audio, and visual information. Its prompt asks the model to explain using text/context, facial expression, tone/prosody, gesture/body language, and mismatch between words and delivery. The second condition is TV, where the model is told to use only text and visual information. In this condition, the prompt explicitly says not to use audio, tone, voice, prosody, or speech sound, and the runtime disables audio with USE_AUDIO_IN_VIDEO=False. This makes the TV setting the key test case for modality attribution hallucination.

Both conditions use the same final decision format. The model must give a brief explanation and then end

with exactly Final Answer: Yes or Final Answer: No. We then parse the final answer with a regular expression and compare it to the MUStARD gold label. This allows us to measure ordinary prediction accuracy while also preserving the explanation text for qualitative analysis. The notebook runs this evaluation for the full MUStARD reasoning files and stores separate result files for the LoRA TAV and LoRA TV conditions.

To find hallucination examples, we focused on cases where both TAV and TV predicted the correct label. This produced 291 common-correct samples, and we inspected the top 100 examples. This restriction matters because it separates explanation faithfulness from prediction accuracy. We are not only asking whether the model got the answer right; we are asking whether the reasoning it gave was grounded in the modalities it actually received. In the TV condition, any explanation that relies on "tone," "delivery," "prosody," or similar audio-based evidence is treated as a modality attribution failure because audio was explicitly removed.

This setup lets us test a specific failure mode: the model may answer correctly while still giving an unfaithful explanation. If the TV-only model cites audio-like evidence, then the explanation is not a reliable account of what the model observed. Instead, it suggests the model is filling in likely sarcasm cues from text or context priors. This is why the methodology combines quantitative filtering with qualitative inspection: the numbers identify comparable correct cases, while the explanations reveal whether the model's modality attribution is faithful.

4.5 H5

The Text Override Rate (TOR) is an adversarial diagnostic for measuring whether an audio-visual model's answer remains anchored to its non-text modalities when contradictory information is injected through the text channel. The motivation is that genuine multimodal integration should be robust to single-channel attacks: if a model truly grounds its answer in audio and visual evidence, then a contradictory transcript should not be sufficient to flip its answer when the audio and video themselves still point at the correct option. TOR operationalizes this expectation by injecting an adversarially generated transcript and measuring how often the model's answer abandons the correct option as a result.

For each AVUT question, an adversarial transcript is generated using GPT-4o in a wrong-option-anchor mode. A wrong MCQ option is selected as the target, and GPT-4o is prompted to produce a short, naturalistic first-person utterance that, if trusted, would lead a listener to select that option. The transcript is then injected into the model's prompt alongside the unmodified audio and video as a purported transcription of the spoken content. The wrong-option anchoring ensures that the attack has a specific direction: the fake transcript points at exactly one of the three distractors, leaving the other two unsupported by any source.

The analysis is restricted to the four speech-content task types (AIE, AVCM, AVOM, AVTM, n = 1,156). Adversarial transcripts are not well-defined for the non-speech audio tasks (ACC, AEL), where the relevant signal is acoustic rather than lexical and there is no spoken content for a fake transcript to contradict. For each question we run two inference passes through Qwen2.5-Omni-7B. The baseline pass provides the full audiovisual input with a standard MCQ prompt. The adversarial pass provides the same audiovisual input with the fake transcript injected. Each question is classified as one of: *eligible-override* (baseline correct, adversarial pass selects the target wrong option), *eligible-resistant* (baseline correct, adversarial pass holds the correct answer), *eligible-confused*(baseline correct, adversarial pass selects a third option not supported by any source), or *ineligible* (baseline incorrect, so attack outcomes cannot be cleanly attributed). All metrics are computed only over the eligible subset.

Two complementary metrics are reported. TOR-flip is the rate at which the model abandons the correct answer under adversarial injection, defined as TOR-flip = (#override + #confused) / #eligible. This is the metric that directly answers the modality-bias question: a low value indicates the model's audio-visual grounding survives text-channel attacks, and a high value indicates that injected text is sufficient to override otherwise correct AV reasoning. The reference point is approximately zero, since under unmodified input the model is by construction holding the gold answer on every eligible question. TOR-target is the narrower rate

at which the model adopts the specific wrong option named by the fake transcript, defined as TOR-target = #override / #eligible. This metric quantifies adversarial steerability rather than modality bias: it measures not just whether the attack works, but whether it works in the direction the attacker intended. The reference point here is 1/3, the rate at which a model destabilized by injected text but not specifically steered by its content would land on the target option by chance from among the three remaining wrong options. We additionally report the steered fraction, defined as #override / (#override + #confused), which is the proportion of flips that follow the injection. A steered fraction near 1.0 indicates that flips are nearly all attacker-directed; a steered fraction near 1/3 would indicate that flips are random thrashing rather than directed override.

5 Results

5.1 H1

A natural concern from the H1 methodology is that the gating variant may have failed because the gate was poorly calibrated, not because of the modalities themselves. To separate these possibilities, we ran a follow-up ablation on all 690 MUStARD clips. We reused the per-modality descriptions already generated by Qwen2.5-Omni-7B and changed only which descriptions were given to the final Yes/No gate. All conditions used the same model.

Gate input	**Accuracy**	**F1**	**Recall (sarcasm)**	**Yes-rate**
Text-only baseline (Qwen-1.5B, no description)	57.0%	47.1%	38.3%	31.3%
Text description only	58.6%	54.5%	49.6%	41.0%
Video description only	58.3%	49.5%	40.9%	32.6%
Audio description only	59.1%	67.1%	83.5%	74.3%
Audio + text descriptions	61.9%	65.3%	—	60.0%
Video + text descriptions	61.9%	63.0%	—	53.0%
Fused video+audio+text, one pass (S3 baseline)	63.9%	69.9%	83.8%	69.9%

Table 3. MUStARD modality-importance ablation: final Yes/No predictions when only selected saved modality descriptions are fed to the contrastive gate. Results are over 690 balanced clips using Qwen2.5-Omni-7B descriptions and gate.

Two findings stand out. First, the single-modality accuracy scores are very close, ranging from 58.3% to 59.1%. However, the audio-only gate identifies sarcastic clips with much higher recall: 83.5%, or 288 out of 345 sarcastic clips. This is more than twice the recall of the text-only and video-only conditions. Audio is therefore the strongest sarcasm signal by recall. However, it also predicts "Yes" more often on genuine clips, giving it lower precision, around 56%, compared with 60–63% for text and video. This is why the accuracy gap remains small.

Second, every single-modality condition remains 4–5 percentage points below the fused baseline. This supports the conclusion that no single channel is enough on its own. The best performance still comes from combining modalities.

The third finding is the most important, and it comes from inspecting the saved descriptions. We listened to clips where the audio gate predicted "No" even though the ground truth was sarcastic. A repeated pattern appeared: the model's audio description was often wrong in a specific way. For example, in clip 1_70 from BBT, Penny says, "I don't think I'll be able to stop thinking about it." The ground truth is sarcastic, and the audio is delivered with clear emphatic, mocking inflection. However, the model describes the audio as "dry

and somewhat flat, lacking warmth or enthusiasm" with "no noticeable emphasis or dramatic pitch variation." Given that description, the downstream gate's "No" prediction makes sense. The gate is not the main problem. The problem is that the audio description failed to capture the prosody that was actually present.

This pattern suggests a failure we call **modality leakage**. When Qwen2.5-Omni-7B is asked to describe only the audio, the description does not purely reflect prosody. It also appears to be influenced by the spoken words. In other words, the model does not treat "what was said" and "how it was said" as fully separate streams. Because the text and audio information are processed through overlapping model representations, the audio description can become contaminated by lexical content.

This is not simply a prompting problem. Clearer instructions may help somewhat, but they do not guarantee clean modality separation. The issue is tied to how omnimodal foundation models are trained and represented internally. As a result, modality-separated prompting is not a clean experimental tool for this model. The descriptions cannot be assumed to isolate the modality they were prompted to analyze.

This finding motivates the framing in Experiment H4: even when researchers explicitly try to isolate one modality's contribution, the model's own outputs can leak across modalities. This weakens the assumption that prompting alone can produce faithful per-modality reasoning.

5.2 H2

5.3 Analysis of Qwen Results:

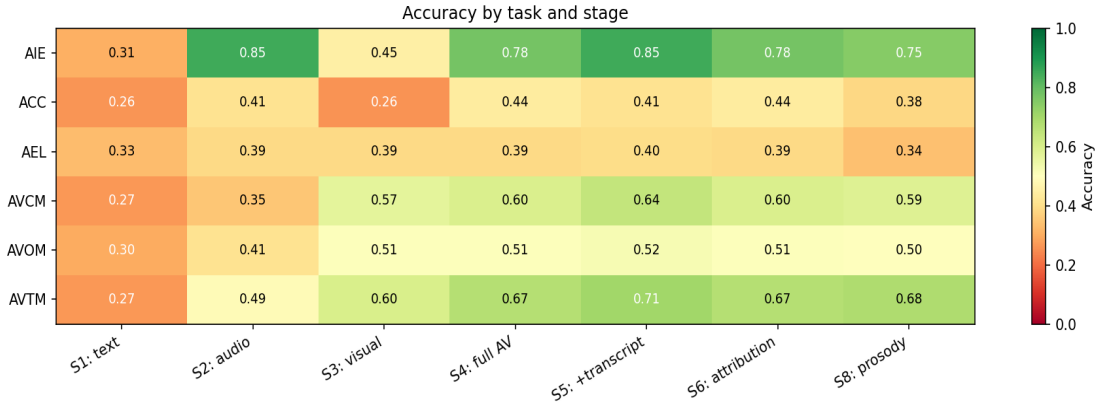

Figure H2.1: Accuracy by task (rows) and stage (columns). Color encodes accuracy. Single-modality audio tasks (top three rows: AIE, ACC, AEL) light up under S2 (audio only), while cross-modal binding tasks (bottom three rows: AVCM, AVOM, AVTM) light up under S3 (visual only). All tasks improve under S4 (full audiovisual). Stage S6 (attribution follow-up) and S8 (prosody verbalization) match S4 closely, indicating these instructional variants do not change accuracy.

The Qwen2.5-Omni-7B accuracy heatmap (Figure H2.1) shows the expected stage progression. S1 (text only, 0.29) sits near the four-option chance line (0.25), confirming AVUT's text-shortcut filtering holds for our model. Audio-only and visual-only stages tie at 0.50, full AV reaches 0.59, and matched transcripts add another 3 pp to 0.62. The 9-point gap between S4 and either single-modality stage indicates the model integrates across channels rather than collapsing onto one. The per-task pattern is consistent with task definitions: pure-audio tasks (AIE, ACC, AEL) show S2
geq S3, while cross-modal binding tasks (AVCM, AVOM, AVTM) show S3 >= S2 (the visual side carries more standalone signal for character, object, and on-screen-text recognition).

Task	n	S1 (text)	S2 (audio)	S3 (visual)	S4 (full AV)	S5 (+transcript)
AIE	289	0.31	0.85	0.45	0.78	0.85
ACC	117	0.26	0.41	0.26	0.44	0.41
AEL	170	0.33	0.39	0.39	0.39	0.40
AVCM	289	0.27	0.35	0.57	0.60	0.64
AVOM	289	0.30	0.41	0.51	0.51	0.52
AVTM	289	0.27	0.49	0.60	0.67	0.71
Overall	1443	0.29	0.50	0.50	0.59	0.62

Table H2.1: Accuracy by task and stage. AIE = Audio Information Extraction, ACC = Audio Content Counting, AEL = Audio Event Location, AVCM = Audio-Visual Character Matching, AVOM = Audio-Visual Object Matching, AVTM = Audio-Visual Text Matching. Stage S6 (attribution follow-up) and S8 (prosody verbalization) accuracies are within noise of S4 and are shown in Figure H2.1.

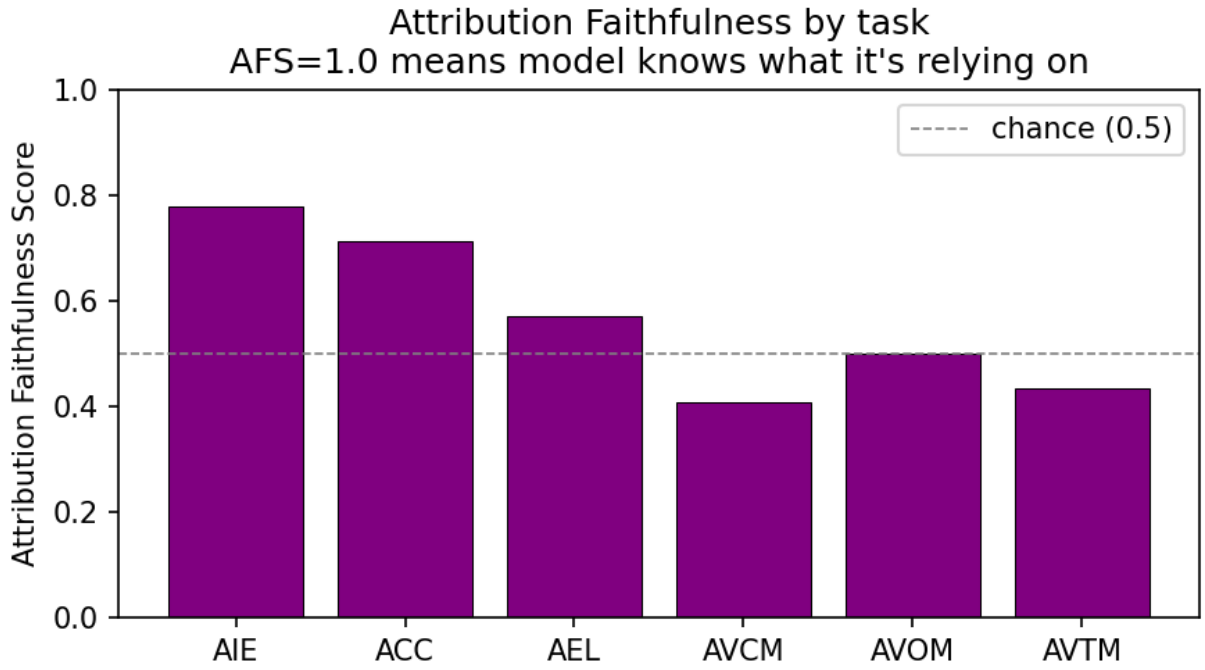

Figure H2.2: Attribution Faithfulness Score by task. AFS = 1.0 means the model's stated modality reliance always survives ablation of that modality; AFS = 0.5 is chance. The three single-modality audio tasks (AIE, ACC, AEL) score above chance; the three cross-modal binding tasks (AVCM, AVOM, AVTM) score at or below chance.

Figure H2.2 shows the headline AFS result. The six tasks split cleanly: single-modality audio tasks sit above the chance line (AIE 0.78, ACC 0.71, AEL 0.57), while cross-modal binding tasks sit at or below it (AVOM 0.50, AVTM 0.43, AVCM 0.41). The gap from highest to lowest task is 0.37. After filtering trivial questions, each task has between 63 and 167 falsifiable cases; AVCM (lowest) has 167 falsifiable cases of which 99 (59%) were confabulations, AIE (highest) has 140 of which 31 (22%) were confabulations. The pattern is not driven by small denominators on either end.

Task	Acc(S4)	Acc(S5)	TIB
AIE	0.78	0.85	−0.073
ACC	0.44	0.41	+0.029
AEL	0.39	0.40	−0.006
AVCM	0.60	0.64	−0.046
AVOM	0.51	0.52	−0.017
AVTM	0.67	0.71	−0.037

Table H2.2: Transcript Injection Bias by task. Negative TIB means the transcript helped.

This rejects H2.1's predicted high-AFS regime: rather than the model being reliable across the board (AFS > 0.75), its introspective access to modality use degrades systematically when the task requires cross-modal binding. On AIE-type questions ("what is the speaker's emotional tone"), the model reports relying on audio and removing audio does change its answer. On AVCM-type questions ("which character is speaking"), the model still reports relying on a specific modality, but ablating that modality often does not change anything, suggesting actual computation drew on different cues than claimed.

One complication warrants future work. Cross-modal binding tasks have higher S3 (visual-only) accuracy than audio-only tasks, meaning the visual channel often suffices alone. The model may be claiming audio reliance while actually answering from vision; ablating audio then leaves the answer unchanged. This is exactly the failure mode AFS is designed to detect, but distinguishing "wrong about which modality" from "used both, ablation insufficient" would require a more refined protocol than binary single-modality removal.

On TIB (Table 2), Qwen shows small uniformly negative effects: per-task TIB ranges from −0.073 (AIE) to +0.029 (ACC), with overall −0.033. H2.3's prediction of negative TIB is weakly confirmed (5 of 6 tasks non-positive), but the magnitude is small enough that transcript injection does not provide a useful lever for task-specific intervention.

5.4 Analysis of Gemma Results:

To test whether the AFS/TIB/LOR framework picks up Qwen-specific artifacts or generalizes across architectures, we ran the same pipeline on Gemma-3n-E2B-IT (a ~2B-active-parameter omnimodal model, 3.5 $times$ smaller than Qwen2.5-Omni-7B) on a 600-sample balanced subset (100 per task) of AVUT-Human. We also added a seventh stage (S7) that injects a mismatched transcript from a different same-task video, enabling computation of LOR.

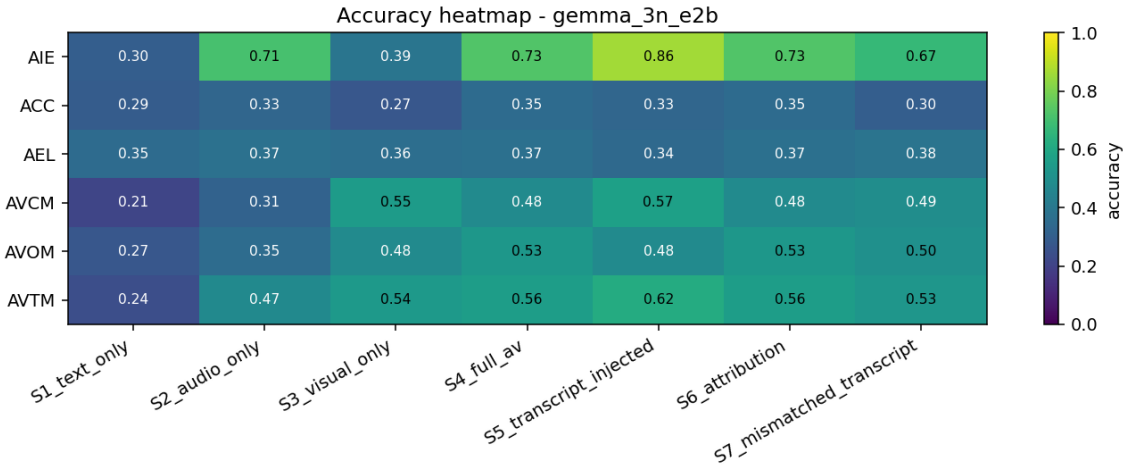

Figure H2.3: Accuracy by task and stage for Gemma-3n-E2B-IT on the 600-sample subset. The qualitative pattern matches Qwen: AIE solves cleanly, ACC and AEL stay flat near chance, cross-modal binding tasks gain from the visual stream, and S5 (matched transcript) lifts most tasks. S7 (mismatched transcript) sits near or slightly below S4, foreshadowing the moderate LOR finding.

Stage	Gemma	Qwen	detla (Qwen - Gemma)
S1 (text only)	0.277	0.289	+0.012
S2 (audio only)	0.427	0.499	+0.072
S3 (visual only)	0.431	0.495	+0.064
S4 (full AV)	0.503	0.591	+0.088
S5 (+matched transcript)	0.533	0.624	+0.091
S6 (attribution)	0.503	0.591	+0.088
S7 (+mismatched transcript)	0.478	—	—

Table H2.3: Cross-model accuracy comparison on identical inputs. Qwen-Omni outperforms Gemma uniformly by 7 - 9 pp across all stages, consistent with the parameter scaling, but the qualitative shape of the results (single-modality near-tie, full-AV gain, transcript helps, S6 $approx$ *S4) is identical.*

AFS cross-model comparison: On Gemma, overall AFS sits at 0.494 across 334 falsifiable cases (165 faithful, 169 confabulated): essentially a coin flip. The task-conditional pattern matches Qwen exactly: pure-audio tasks (AIE 0.70, ACC 0.66, AEL 0.63) sit meaningfully above chance, while cross-modal matching tasks (AVTM 0.40, AVCM 0.35, AVOM 0.31) sit well below.

Task	AFS (Gemma)	AFS (Qwen)
AIE	0.70	0.78
ACC	0.66	0.71
AEL	0.63	0.57
AVTM	0.40	0.43
AVCM	0.35	0.41
AVOM	0.31	0.50

Table H2.4: Per-task AFS, Gemma vs Qwen. The qualitative split (pure-audio above chance, cross-modal binding at or below chance) holds for both models. The largest cross-model divergence is AVOM, where Gemma confabulates substantially more.

AFS is not picking up an artifact specific to one architecture since the qualitative pattern survives a 3.5 times model-size change and suggests our findings are cross-model. The smaller model confabulates *more* on cross-modal binding (AVOM particularly), suggesting that whatever capacity supports faithful self-attribution scales with model size, but the failure mode itself is shared.

LOR rejects H2.2: Of the 300 questions Gemma answered correctly at S4, only 49 flipped to a different answer when handed a contradictory same-task transcript at S7 (251 stayed). Overall LOR = 0.163, well below the 0.5 threshold predicted by H2.2 and far from the catastrophic lexical override predicted for audio-only LLMs (Chen et al. 2025). Per-task LOR ranges from 0.110 (AIE) to 0.286 (ACC, where the small S4-correct denominator of 35 makes the rate noisy). H2.2 is rejected: Gemma trusts its audio over a contradictory transcript on roughly 84% of audio-correct cases. The visual stream may be acting as a partial guard against transcript override that audio-only LLMs lack.

ltaConf reveals a verbal/behavioral attribution gap: Although Gemma's verbalized self-attribution (AFS) is unreliable, its confidence drops under ablation are calibrated correctly. Audio removal lowers Gemma's

confidence by 1.85 to 6.79 pp across all six tasks; visual removal moves confidence by essentially zero (-1.36 to +0.15). The largest audio-removal drop is on AIE (+6.79 pp), the task with the highest AFS. The model's *behavior* tracks audio importance correctly even when its *verbal description* of that importance is wrong half the time. We read this as a verbal/behavioral attribution gap: the model's internal modality routing appears reasonable but the layer that *describes* the routing is the broken one.

TIB cross-model: Gemma TIB = −0.030, almost identical to Qwen's −0.033, with the same per-task ranking (AIE most positively affected by transcripts, AVOM the only task where transcripts hurt). H2.3 is weakly confirmed in both models in the same direction with the same magnitude.

5.5 H3

AVUT Multimodal QA — Results Summary | 512 tokens / modality describe

ModalitySeparation describe cap: 512 max output tokens per modality describe call (GEMINI_MAX_OUTPUT_TOKENS_DESCRIBE unless overridden). | Model: Gemini 2.5 Flash

`"+ Reason"` variants: Gemini reasoning / thinking budget on for final MCQ (768).

	Overall	Audio Info Extraction	Audio Counting	Audio Event Localization	AV Content Matching	AV Onset Matching	AV Temporal Matching	N	Time (s)	Describe max (tok)
Vanilla	**66.7%**	80.0%	0.0%	100.0%	100.0%	60.0%	60.0%	30	215	—
Vanilla + Reason	**76.7%**	80.0%	60.0%	80.0%	100.0%	60.0%	80.0%	30	142	—
ModalitySeparation	**63.3%**	80.0%	40.0%	100.0%	100.0%	20.0%	40.0%	30	707	512
ModalitySeparation + Reason	**70.0%**	80.0%	80.0%	100.0%	100.0%	20.0%	40.0%	30	676	512
ModalitySeparation + MIS	**70.0%**	100.0%	40.0%	100.0%	100.0%	40.0%	40.0%	30	800	512
ModalitySeparation + MIS + Reason	**46.7%**	80.0%	20.0%	60.0%	80.0%	0.0%	40.0%	30	779	512

Key:
• Heatmap colouring applies only under task columns and Overall — red = lower accuracy; green = higher.
• Accuracy is fraction correct on scored rows; blanks under a task mean that task had zero scored items in this run.
• N = prompts scored OK; Time (s) = aggregate wall time recorded in that metrics file.
• Describe max (tok) = Gemini max_output_tokens limit for ModalitySeparation describe calls (each of text/audio/visual); Vanilla shows "—."
• MIS rows may show different numbers per modality if recorded in metrics; otherwise GEMINI_MAX_OUTPUT_TOKENS_DESCRIBE fallback.

Task abbreviations:
AIE = Audio Info Extraction
ACC = Audio Counting
AEL = Audio Event Localization
AVCM = AV Content Matching
AVOM = AV Onset Matching
AVTM = AV Temporal Matching

Environment fallback — GEMINI_MAX_OUTPUT_TOKENS_DESCRIBE: 512

Table H3.1: Experiment H3 Results for 512 tokens per Modality Description
* Typo in Column Labels: should be AV Object Matching, AV Text Matching

AVUT Multimodal QA — Results Summary | 1024 tokens / modality describe

ModalitySeparation describe cap: 1024 max output tokens per modality describe call (GEMINI_MAX_OUTPUT_TOKENS_DESCRIBE unless overridden). | Model: gemini-2.5-flash

`"+ Reason"` variants: Gemini reasoning / thinking budget on for final MCQ (768).

	Overall	Audio Info Extraction	Audio Counting	Audio Event Localization	AV Content Matching	AV Onset Matching	AV Temporal Matching	N	Time (s)	Describe max (tok)
Vanilla	**66.7%**	80.0%	0.0%	100.0%	100.0%	60.0%	60.0%	30	215	—
Vanilla + Reason	**76.7%**	80.0%	60.0%	80.0%	100.0%	60.0%	80.0%	30	142	—
ModalitySeparation	**63.3%**	80.0%	40.0%	100.0%	100.0%	20.0%	40.0%	30	707	1024
ModalitySeparation + Reason	**70.0%**	80.0%	80.0%	100.0%	100.0%	20.0%	40.0%	30	676	1024
ModalitySeparation + MIS	**56.7%**	100.0%	20.0%	100.0%	80.0%	20.0%	20.0%	30	718	1024
ModalitySeparation + MIS + Reason	**66.7%**	100.0%	60.0%	80.0%	100.0%	20.0%	40.0%	30	804	1024

Key:
• Heatmap colouring applies only under task columns and Overall — red = lower accuracy; green = higher.
• Accuracy is fraction correct on scored rows; blanks under a task mean that task had zero scored items in this run.
• N = prompts scored OK; Time (s) = aggregate wall time recorded in that metrics file.
• Describe max (tok) = Gemini max_output_tokens limit for ModalitySeparation describe calls (each of text/audio/visual); Vanilla shows "—."
• MIS rows may show different numbers per modality if recorded in metrics; otherwise GEMINI_MAX_OUTPUT_TOKENS_DESCRIBE fallback.

Task abbreviations:
AIE = Audio Info Extraction
ACC = Audio Counting
AEL = Audio Event Localization
AVCM = AV Content Matching
AVOM = AV Onset Matching
AVTM = AV Temporal Matching

Environment fallback — GEMINI_MAX_OUTPUT_TOKENS_DESCRIBE: 1024

Table H3.2: Experiment H3 Results for 1024 tokens per Modality Description
* Typo in Column Labels: should be AV Object Matching, AV Text Matching

Key:
• Heatmap colouring applies only under task columns and Overall — red ≈ lower accuracy; green ≈ higher.
• Accuracy is fraction correct on scored rows; blanks under a task mean that task had zero scored items in this run.
• N = prompts scored OK; Time (s) ≈ aggregate wall time recorded in that metrics file.
• Describe max (tok) = Gemini max_output_tokens limit for ModalitySeparation describe calls (each of text/audio/visual); Vanilla shows "—."
• MIS rows may show different numbers per modality if recorded in metrics; otherwise GEMINI_MAX_OUTPUT_TOKENS_DESCRIBE fallback.

Task abbreviations:
AIE = Audio Info Extraction
ACC = Audio Counting
AEL = Audio Event Localization
AVCM = AV Character Matching
AVOM = AV Object Matching
AVTM = AV Text Matching

Environment fallback — GEMINI_MAX_OUTPUT_TOKENS_DESCRIBE: 1024

Test	Text	Audio	Visual	Tokens/desc
ModalitySeparation	0.3663	0.2853	0.3484	512
ModalitySeparation + Reason	0.3502	0.3276	0.3222	512
ModalitySeparation	0.3520	0.3132	0.3348	1024
ModalitySeparation + Reason	0.3366	0.3096	0.3538	1024

Table H3.3: Softmax Relative Weights of Each Modality per MIS

Analysis: Our hypothesis was that ModalitySeparation and its variants would show greater performance overall than the corresponding Vanilla/Vanilla+Reason results. Interestingly, the hypothesis is rejected in all prompting variations except for ModalitySeparation + MIS on 512 tokens/description.

Evaluation of ModalitySeparation: Similar to the analysis in H1, we might suspect that "modality leakage" has occurred (Chen et al.) since ModalitySeparation *underperforms* relative vanilla : where even if the model is only passed audio content, it tends to focus on the lexical (text-derived/words) content in the audio rather than nonverbal cues. In particular, ModalitySeparation does the poorest on AudioVisual Object and Text Matching. One example of an AV Object matching question is as follows:
{"qa_id": 27, "video_id": 9, "url": "https://www.youtube.com/shorts/w7CRNeCyp48?si=d5fojN7Qhff4vF3T", "video_type": "vlog", "task_type": "Audio Object Matching", [. . .] "question": "When the audio says, 'and just creativity in general was just so inspiring and fun', what is the woman doing?",

"option_a": "She is charging her phone.",

"option_b": "She is taking a sip of her drink.",

"option_c": "She is arranging her odds and ends.",

"option_d": "She is editing her video.", "answer": "D"}

Curiously, the question itself involves lexical content only, contradicting the analysis in H1. Looking at the options, however, they seem to be associated with non-verbal cues: a sip of coffee, the sound of a desk being cleaned up, and the mouse movements of video editing. Thus it does seem the case that "modality leakage" explains the poor performance of ModalitySeparation.

Evaluation of ModalitySeparation + MIS/Description Token Budget: The poor performance on ModalitySeparation + MIS may follow from the "modality leakage" hypothesis. It is worth noting that ModalitySeparation + MIS outperformed Modality Separation and Vanilla on a 512 token description budget but not the 1024 token description budget. This checks with our underlying assumption since MIS (ie. efficient allocation of tokens) should work best when tokens are most scarce (the 512 token case).

One explanation for why a larger 1024-token budget *does not* necessarily imply better performance is that too many tokens may lead a model to "overdescribe" a given modality and start hallucinating descriptions out of a pressure to fill up the budget. This can be seen in how ModalitySeparation + MIS under 1024 tokens/modality degraded on AV Matching tasks: it is possible the model started "making up" objects in the scene with its description tokens left to spare leading to skewed accuracy.

Reasoning seems to have improved model performance overall, except for the case of ModalitySeparation + MIS + Reasoning under 512 tokens/modality – here we similarly hypothesize that scarcity in description relative the 768 tokens used for internal reasoning might have led the model to start making up "evidence" in its chain of thought when it ran out of descriptive detail.

MIS Weights: Finally we note that the relative importance of each modality per MIS is roughly even, with bias towards text and visuals relative audio. This checks with the understanding that models seem to have lexical bias, perhaps a slight visual bias of MLLMs over audio must be also considered.

5.6 H4

The qualitative hallucination analysis tests whether multimodal explanations remain faithful to the modalities actually provided to the model. Prior multimodal sarcasm work argues that sarcasm depends on interactions between text, audio/prosody, and visual cues, including tone, facial expression, and delivery (Castro et al., 2019; Bhosale et al., 2023). Therefore, in the TV-only condition, where the model receives text and visual information but no audio, a faithful explanation should ground its reasoning only in textual context, facial expression, body language, gesture, and scene information. If the model instead cites audio-dependent evidence such as "tone," "delivery," "sarcastic tone," or "inflection," then the explanation is not fully modality-faithful.

To identify these failures, we inspected the top 100 common-correct examples where both the TAV model and the TV-only model predicted the correct sarcasm label. This restriction isolates cases where the final answer is correct, allowing us to study whether the reasoning process is still faithful. Recent work on modality bias and audio-centric video understanding suggests that MLLMs can over-rely on lexical or cross-modal priors rather than grounding their answers in the relevant modality evidence (Chen et al., 2025; Yang et al., 2025). In several TV-only examples, the model correctly predicts sarcasm but explains its decision using audio-like cues, suggesting that the explanation may be a post-hoc rationale rather than a faithful account of input use.

The strongest hallucination cases occur when the TV-only model explicitly relies on unavailable audio evidence rather than merely using ambiguous language. In Sample 35, the TV-only explanation states that the "tone and delivery imply a sarcastic or ironic undertone." In Sample 50, it claims that the speaker's "body language and tone suggest a lack of enthusiasm." In Sample 56, it describes the interaction as having a "playful or sarcastic tone." In Sample 8, it explains the result by saying sarcasm is often delivered with a casual tone. In Sample 7, it says the "tone seems to imply a lack of concern or interest." Because these explanations were generated in the TV-only setting, each case shows the model citing evidence from a modality that it could not actually observe.

These findings suggest that modality-separated prompting alone is not sufficient to guarantee faithful modality attribution. Although PragCoT-style methods aim to improve multimodal sarcasm reasoning by separating pragmatic evidence across modalities (Saha, 2025), our examples show that the model can still infer likely audio cues from transcript or scene context and present them as observed evidence. This supports the hypothesis that some multimodal explanations are post-hoc rationalizations rather than faithful modality attributions.

Overall, the results provide evidence for Modality Attribution Hallucination in restricted-modality reasoning. The TV-only model's use of audio-like evidence shows that correct predictions do not necessarily imply faithful explanations. This helps explain why modality-separated reasoning did not produce a large or consistent performance improvement: the model can still contaminate modality-specific reasoning with prior assumptions about what sarcasm usually sounds like. Thus, the main failure is not only prediction error, but explanation error: the model sometimes reports using a modality that was not actually present.

5.7 H5

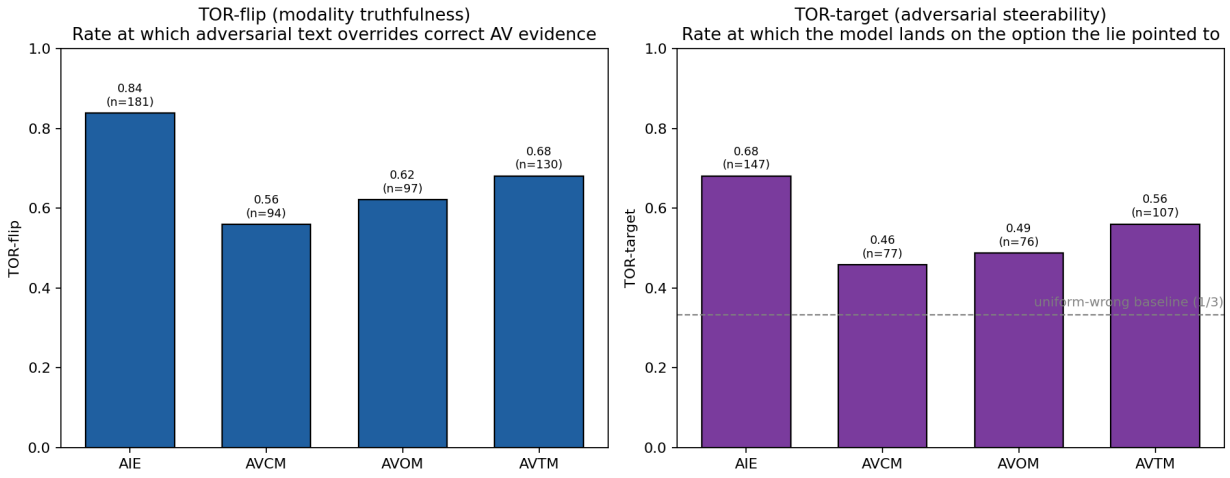

Figure H5.1 Textual Override Rate Comparison

The headline finding is that Qwen2.5-Omni-7B's audio-visual grounding does not survive text-channel attacks. Across all 731 eligible questions, TOR-flip = 0.69: the model abandons a correct answer on roughly seven out of every ten attempts when an adversarial transcript is injected, even though the audio and video continue to support the correct option throughout. This is a direct demonstration of modality bias toward text. A model that genuinely integrated all three channels in a load-bearing way should produce a TOR-flip near zero, since two out of three channels still point at the right answer; a TOR-flip of 0.69 indicates that the text channel can override the audio-visual consensus on the majority of questions.

The per-task pattern of TOR-flip is consistent with task structure. AIE is the most vulnerable (TOR-flip = 0.84): its answers are determined directly by the lexical content of speech, which is exactly what the adversarial transcript attacks, and the model has no alternative channel to fall back on. AVTM follows at 0.68, reflecting its inherent text-adjacency — matching spoken content to on-screen text is a task type where lexical interpretation is already central, and the model appears predisposed to trust injected lexical content in this setting. AVCM (0.56) and AVOM (0.62) are the most resistant, but still well above any reasonable robustness threshold. Even on these tasks, where the visual channel carries substantial standalone signal for character and object recognition, more than half of correct answers do not survive text injection. The TOR-target results confirm that flips are not random thrashing but are predominantly following the specific direction of the attack. The overall steered fraction is 0.81 — of the 502 flips, 407 land on the option the fake transcript was specifically designed to support, well above the 1/3 rate that would be expected if the injection were merely destabilizing the model. The steered fraction is uniform across tasks (0.78 to 0.82), indicating that when this model is pulled off a correct answer by injected text, it is reliably pulled toward whatever the injection points at, regardless of task type. The TOR-target panel confirms this directly: every task sits at 0.46 or higher against a 1/3 baseline, and AIE reaches 0.68.

5.8 Conclusions and Future Work

Overall, our project results suggest that while MLLMs are themselves poor diagnosers of their own modality biases, those biases can be accurately determined via ablations. The AFS diagnostic in H2 makes this concrete: across both Qwen2.5-Omni-7B and the 3.5
$times$ smaller Gemma-3n-E2B, the model's self-reported modality reliance was reliable on pure-audio tasks (AFS 0.57–0.78) but fell to chance or below on cross-modal binding tasks (0.31–0.50), with overall AFS on Gemma sitting at 0.494, essentially a coin flip. That this task-conditional confabulation pattern survives a

substantial change in model size suggests it is a property of the omnimodal architecture family rather than an artifact of any single model. H4's qualitative TV-only analysis points in the same direction: even when a modality is physically removed, the model can still describe it as if it had been present.

A more nuanced finding from H2 is the gap between verbal and behavioral attribution. Although Gemma's verbalized self-reports are unreliable (AFS ~0.49), its confidence drops appropriately when audio is ablated (1.85 to 6.79 pp) and barely moves when visual is ablated (-1.36 to +0.15). The model's internal modality routing appears reasonable while the layer that describes that routing is the broken one. This suggests behavioral probes may be a more honest evaluation surface than self-reports for any future modality-bias mitigation work.

The picture on lexical bias is more nuanced than prior work might predict. H2.2 hypothesized a high Lexical Override Rate (LOR > 0.5) following findings on audio-only LLMs (Chen et al. 2025), but Gemma's LOR sits at 0.163, or when a contradictory same-task transcript is injected alongside intact audio and video, the model abandons its correct answer only ~16% of the time. The catastrophic lexical override documented for audio-only LLMs does not appear to transfer wholesale to omnimodal models, where the visual stream may be acting as a partial guard (Chen et al. 2025). H5's adversarial probe complicates this picture: when the injected transcript is specifically engineered by GPT-4o to anchor a wrong option, Qwen's TOR-flip rises to 0.69. The two findings are consistent: same-task transcript swaps without directional anchoring produce moderate override (~16%), while adversarially crafted misinformation produces severe override (~69%). Modern omnimodal models appear robust to incidental lexical contradiction but remain vulnerable to targeted text-channel attacks.

On the prompting side, the overall picture is that modality separation, while theoretically appealing, does not work in practice, with modest positive results for Modality Importance Scoring as a way to navigate a limited description budget.Our work on interpretability/modality ablations suggests that the modality biases in modern MLLMs remain significant. Furthermore, progress in addressing cross-modal hallucinations in MLLMs will likely have to come from model architecture and training mix, as post-training/prompting does not show as much promise.

On the performance side, given more compute in the future, it is worth exploring post-training (such as DPO or Reward Model Reranking) on subsets of these multimodal benchmarks to investigate how much modals can be post-trained to correct their modality attributions and lexical biases and whether the underlying issue is endemic to the MLLM architecture itself. This would be an extension of SarcasmMiner, but perhaps including a *trained* MIS as a factor in the reward model (Li et al. 2026).

One good direction for future work is to measure the extent of cross-modal hallucination on distinct, more “real-world” datasets – we’ve been working with video datasets derived from TV, do these same biases hold (and are they exacerbated?) on real-world multimodal medical or robotics benchmarks?

Works Cited

Bhosale, S., Chaudhuri, A., Williams, A.L.R., Tiwari, D., Dutta, A., Zhu, X., Bhattacharyya, P., Kanojia, D. (2023). Sarcasm in Sight and Sound: Benchmarking and Expansion to Improve Multimodal Sarcasm Detection. arXiv preprint arXiv:2310.01430.

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., Poria, S. (2019). Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). arXiv preprint arXiv:1906.01815.

Chaubey, A., Pang, J., Siniukov, M., Soleymani, M. (2026). AVERE: Improving Audiovisual Emotion Reasoning with Preference Optimization. arXiv preprint arXiv:2602.07054.

Chen, J., Guo, Z., Chun, J., Wang, P., Perrault, A., Elsner, M. (2025). Do Audio LLMs Really LISTEN, or Just Transcribe? Measuring Lexical vs. Acoustic Emotion Cues Reliance. arXiv preprint arXiv:2510.10444.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C.E., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukovsiut.e, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T.,

Maxwell, T.D., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S., & Perez, E. (2023). Measuring Faithfulness in Chain-of-Thought Reasoning. *ArXiv, abs/2307.13702*.

Li, Z., Chen, Y., Lai, H., Gao, X., Nayak, S., Coler, M. (2026). SarcasmMiner: A Dual-Track Post-Training Framework for Robust Audio-Visual Sarcasm Reasoning. arXiv preprint arXiv:2603.05275.

Park, J., Jang, K.J., Alasaly, B., Mopidevi, S., Zolensky, A., Eaton, E., Lee, I., & Johnson, K. (2024). Assessing Modality Bias in Video Question Answering Benchmarks with Multimodal Large Language Models. *ArXiv, abs/2408.12763*.

Saha, A., Suresh, V., Hospedales, T., Demberg, V. (2025). MUStReason: A Benchmark for Diagnosing Pragmatic Reasoning in Video-LMs for Multimodal Sarcasm Detection. arXiv preprint arXiv:2510.23727.

Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *ArXiv, abs/2305.04388*.

Wang, D., Liu, S., Zhang, T., Chen, Y., Li, J., Meng, H. (2026). EmotionThinker: Prosody-Aware Reinforcement Learning for Explainable Speech Emotion Reasoning. arXiv preprint arXiv:2601.15668..