Prior Dominance in Audio-Visual LLMs: When Generative Models Memorize Over Reasoning Under Cross-modal Conflict

Adarsh Sudheer [*1] **David Li** [*1] **Omar Elbanna** [1] **Ishaan Kodarapu** [1] **Arjun Bahuguna** [†1] **Vasu Sharma** [†1]

Abstract

Audio hallucinations remain a critical failure mode in Audio-Visual LLMs, where models substitute memorized distributional priors for actual audio grounding when forced to process conflicting cross-modal inputs. Using the logit lens and head shutoff ablations on VideoLLaMA 2-7B-AV, we track the internal trajectory of cross-modal reasoning. We find that model commitment concentrates at layer 25.5 ± 1 across all configurations, regardless of alignment intervention. Crucially, 21 conflict-resolution heads cluster at layers 15-18, structurally upstream of this commitment point, indicating the model detects conflict early but overwrites it during generation. Behaviorally, all three fine-tuned configurations and off-the-shelf InternVideo2 collapse to near-chance on conflict examples, suffering a 32.3% accuracy drop and 17.3% instruction failure rate while shifting output priors. These findings identify late-layer prior commitment, not early-stage temporal alignment, as the key intervention target. Code and data to reproduce our mechanistic audit and behavioral evaluations are available at https://github.com/AdarshSudheer09/AVHBench-dmai.

1. Introduction

A central challenge for Audio-Visual LLMs is mitigating audio hallucinations, specifically determining when models reproduce memorized visual priors rather than grounding their outputs in the actual audio stream. Audio-visual LLMs are increasingly evaluated on benchmarks for audio grounding and synchronized understanding (Wang et al., 2024; Cheng et al., 2024; Sung-Bin et al., 2025; Jung et al., 2025b; Leng et al., 2024). Our central question is this: can such a model reason across modalities when the two streams conflict? If a video of a dog is paired with audio of an engine, each is plausible alone but inconsistent jointly. Solving this requires the model to compare modality-specific semantic content internally rather than fall back on surface-level co-occurrence, which makes it a useful probe for what is actually happening inside the network under cross-modal stress.

This setting is useful because agreement examples are solvable via single-modality shortcuts, whereas conflict examples reveal whether the model genuinely compares streams or defaults to a prior. The prevailing assumption that greater multi-modal alignment improves reasoning has driven work in contrastive pre-training, temporal synchronization, and token-level fusion. However, our findings challenge this assumption, implying that the failure is not just architectural, but computational: a late-layer mechanism overwrites earlier conflict-sensitive computations. Using the logit lens and head shutoff ablations, we find that the model's final prediction relies on an internally preferred response pattern rather than cross-modal reasoning. This manifests behaviorally as prior dominance: in our experiments with VideoLLaMA 2-7B-AV, increasingly strong forced alignment methods (ACTC, TATI, AMD) change answer bias, but do not improve grounded inference. Off-the-shelf InternVideo2 (Wang et al., 2024) exhibits the same vulnerability: while achieving 60.1% accuracy on the full AVHBench dataset, performance drops to 27.8% under contradiction. The contributions of this paper are as follows. First, it localizes prior dominance to a specific layer range (25.5 ± 1) using the logit lens across $1,281$ conflict samples and four model configurations. Second, it provides causal evidence via head shutoff ablations that conflict-resolution heads (layers 15-18) are structurally upstream of the commitment point, explaining why alignment-stage interventions cannot suppress prior dominance. Third, it shows the snap layer is configuration-invariant, suggesting prior dominance is a backbone property rather than a result of the alignment

[*]Equal contribution
† Senior author.

Accepted to Learning to Listen: ICML 2026 Workshop on Machine Learning for Audio (non-archival). [1]Algoverse AI Research, Palo Alto, California, USA. Correspondence to: Adarsh Sudheer <adarshsudheer09@gmail.com>, David Li <davidli07712@gmail.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

procedure.

2. Problem Setup and Mechanistic Framework

Let X_v and X_a denote video and audio streams. In a conflict example, each stream is individually coherent but the pair is incompatible. We propose a two-stage mechanistic hypothesis for processing these inputs:

Stage 1 - Conflict Detection: The model's middle layers (approximately 14-18) contain attention heads that are causally sensitive to cross-modal conflict.

Stage 2 - Prior Commitment: Despite successful conflict detection in Stage 1, late MLP layers (approximately 24-27) commit the residual stream to a high-probability output prior, overwriting the conflict signal before generation. This is consistent with MLP layers acting as distributional knowledge stores (Meng et al., 2022), which may similarly impose shortcuts under modality conflict.

We probe where the model's answer becomes committed using a logit-lens operator,

$$\hat{p}^{(\ell)} = \mathrm{softmax}\big(W_U \,\mathrm{RMSNorm}(h^{(\ell)})\big),$$

where $h^{(\ell)}$ is the residual stream at layer ℓ and W_U is the final unembedding matrix. We define the snap layer ℓ^* as the earliest layer after which the top prediction remains unchanged through the final layer. We use ℓ^* as a probe for the memorization-over-reasoning transition: an early commitment indicates the model has retrieved a learned prior rather than reasoned from cross-modal evidence.

3. Methodology

We investigate audio hallucination and late-layer commitment by evaluating both plain and fine-tuned models. We utilize InternVideo2 (Wang et al., 2024) to establish the baseline audio-visual collapse under cross-modal conflict. We then conduct an in-depth behavioral and mechanistic audit using VideoLLaMA 2-7B-AV (Cheng et al., 2024), applying three distinct alignment configurations to test whether post-training interventions improve audio grounding or merely alter the mechanistic commitment structure.

3.1. Evaluation Dataset

Audio-visual conflict examples are an effective probe for audio hallucination analysis, as agreement examples often reward visual-only shortcuts. We curated an adversarial conflict split ($N = 1,281$) from AVCD and AVHBench using automated filtering to isolate severe audio-visual contradictions (e.g., a dog visual paired with car audio). We measure the influence of various alignment methods on resolving audio hallucinations using this filtered split.

Because our automated filtering inadvertently introduced a 92% "No" label skew, which likely enabled illusory accuracy gains through rejection bias, our final evaluations only use the full balanced AVHBench dataset (N ≈ 6,408).

3.2. Three-Stage Alignment Pipeline

We investigate the effects of a three-stage fine-tuning pipeline on VideoLLaMA2-7B, designed as an ablation series to test whether post-training interventions alter the late-layer commitment mechanism identified in Section 2. All pipeline stages were implemented using LoRA on the frozen VideoLLaMA2-7B-AV base model; full training and hardware configurations are detailed in Appendix B.

3.2.1. Audio-Conditioned Token Concatenation (ACTC)

ACTC serves as our baseline integration. Audio features from BEATs are projected into a shared embedding space and concatenated sequentially with the visual token prefix. While providing access to both modalities, tokens remain temporally unaligned, leaving the model free to attend to either modality without explicit cross-modal synchronization.

3.2.2. Timestamp-Aware Token Interleaving (TATI)

TATI introduces temporal synchronization by replacing sequential concatenation with a synchronous 1:1 interleaving of visual and audio tokens, adapted from AVicuna (Tang et al., 2024). Each audio token is placed adjacent to its corresponding video frame. A shared learnable temporal embedding E_{temp} is added to both modalities at each timestep t:

$$\tilde{v}_t = v_t + E_{\text{temp}}(t), \quad \tilde{a}_t = a_t + E_{\text{temp}}(t) \tag{1}$$

This forces the model to process each visual frame alongside its temporally corresponding audio token, with the intention of reducing temporal confusion during multi-modal reasoning.

3.2.3. Asymmetric Modality Dropout (AMD)

AMD acts as a regularizer during fine-tuning. At each step, a stochastic dropout mask is applied to either the visual or audio sequence with probability P_{mask}, blinding the model to one modality. While intended to prevent single-modality reliance, this risks teaching the model that one modality suffices.

3.3. Mechanistic Audit

We audit the baseline VideoLLaMA2-7B (Cheng et al., 2024) using 100 samples from the conflict split. For each attention head in layers 14–27, we shut off that head (replacing its contribution with its mean) and re-run inference

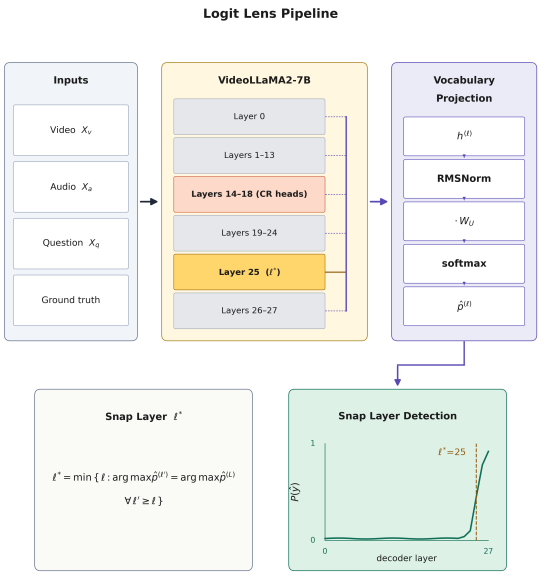

Figure 1. The Logit Lens capture pipeline and snap layer detection. Residual streams are extracted across all layers to track the trajectory of the predicted target token.

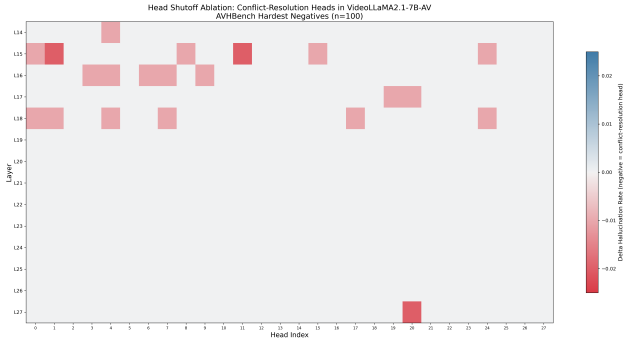

Figure 2. Distribution of hallucination and conflict resolution heads throughout layers 14–27 of VideoLLaMA2-7B. A positive Δ (Blue) implies hallucination while a negative Δ (Red) represents conflict-resolution

(Wang et al., 2023). We define Δ as the change in hallucination rate (fraction failing exact-match Yes/No) relative to the unablated baseline: $\Delta \geq 0.01$ marks a hallucination head (ablation reduces hallucinations), $\Delta \leq -0.01$ marks a conflict-resolution head (ablation increases them). For top candidates, we follow up with activation patching from clean to conflict samples to test generalization.

The audit identifies 21 conflict-resolution heads clustered in layers 15–18; no heads pass the hallucination threshold. We do not interpret this absence as evidence of non-localizability: it can also reflect insufficient audit power, distributed computation, or threshold sensitivity. The audit was not run on the LoRA-tuned configurations.

3.4. Logit Lens Capture Protocol

We apply the logit lens (nostalgebraist, 2020) across all layers of VideoLLaMA2-7B to locate the generative prior's emergence. For all 1,281 conflict samples, we project the residual stream $h^{(\ell)}$ through the final RMSNorm and W_U to yield a per-layer distribution $\hat{p}^{(\ell)}$. We track $P(\text{Yes})$, $P(\text{No})$, and the predicted token across layers, allowing us to define the snap layer ℓ^* where the top prediction stabilizes, revealing where the network commits to its prior.

3.5. Evaluation Protocol

InternVideo2 and each ablation checkpoint were evaluated on the full AVHBench dataset ($N \approx 6{,}408$) as well as the curated 1,281-sample conflict split. All evaluations used greedy decoding with a temperature of 0.01. Yes/No accuracy was computed via exact string match, while captioning outputs were analyzed qualitatively given the known failure of exact match metrics for open-ended generation.

4. Results & Discussion

4.1. Mechanistic Origins: Early Detection and Late-Layer Overwrite

We find that all four configurations of VideoLLaMA2-7B snap to their final-layer prediction at approximately layer 25. Further, no configuration demonstrated significant target-token probability mass before layer 18. This directly confirms the structural gap proposed in our two-stage hypothesis: the prior forms well beyond the model's conflict-resolution heads, which our ablations localized entirely between layers 15 and 18. Because the snap point ℓ^* consistently occurs at layer 25.5 $\pm$1 across all configurations, the contribution of upstream conflict-resolution heads is actively suppressed before generation.

Logit lens analysis confirms that the shift in priors is mechanistic: each individual capture commits to a fixed direction independently of ground truth (matching the bias documented in Section 4.3). Figure 3 tracks the probability of the committed target token, illustrating this late-stage stabilization. Alignment fine-tuning shifts which memorized prior fires, without replacing prior retrieval with grounded cross-modal reasoning.

Furthermore, this structural routing remains invariant to temporal alignment. As detailed in Appendix A (Figure 4), we track the cosine similarity of AV tokens in the deep layers. After correcting for a temporal embedding artifact by subtracting the shared E_{temp} vector, AV token cosine similarity remains identical across all configurations. The alignment pipeline alters the surface-level bias, but the underlying mechanism, Stage 1 detection followed by Stage 2

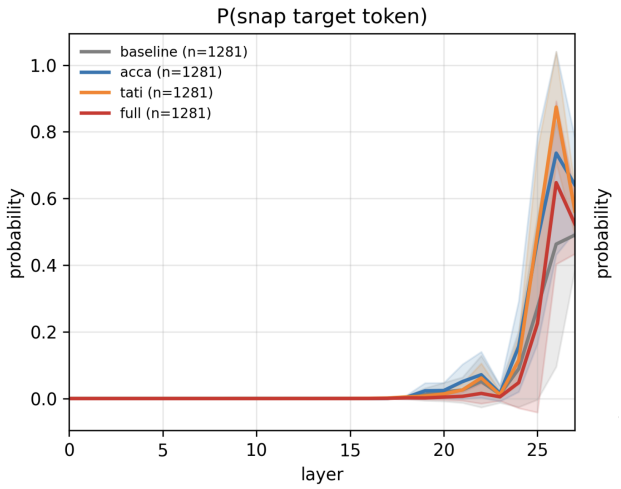

Figure 3. Logit-lens probability trajectories for the target token. Across configurations, the prediction stabilizes around layer 25.5 ± 1, indicating a consistent late-layer commitment point.

overwrite, remains structurally unchanged.

4.2. Behavioral Corroboration: Cross-Model Generalization

Evaluating off-the-shelf InternVideo2 on AVHBench establishes this vulnerability prior to alignment interventions. While achieving 60.1% accuracy on the full dataset, performance drops to 27.8% on the conflict set ($\Delta = 32.3\%$). This drop well below chance confirms the model is actively misled. Further, conflict triggers a 17.3% instruction failure rate (1,108 out of 6,408 inferences) in which the model abandons the "Yes/No" format for unconstrained filler, establishing prior dominance as a baseline vulnerability.

4.3. Behavioral Signatures of Prior Dominance

Table 1 shows two distinct patterns that should be read separately: accuracy remains near chance across all configurations, while the Yes/No output prior shifts substantially. ACTC is slightly Yes-leaning, TATI is more balanced, and the full pipeline becomes strongly No-leaning. For context, the off-the-shelf VideoLLaMA 2-7B-AV base model achieves 51.7% on the AV Matching task, showing that these alignment interventions truly lower cross-modal conflict resolution compared to pretrained baselines.

The full pipeline over-corrects toward `No`, with 3446 `No` predictions against 1853 `Yes` predictions, yet this bias shift does not yield better conflict resolution. Mechanistically, this confirms the model swaps one memorized shortcut for another, rather than fixing grounded cross-modal reasoning.

Open-ended captioning shows the same pattern. Across checkpoints, the model begins with variants of "A person is playing a musical instrument..." (occurring in 71.2% of all open-ended captioning outputs), even when neither modality supports that claim. Because the surface continuation changes while the semantic scaffold remains stable, we interpret this as behavioral evidence for prior dominance: the model preserves fluency while losing grounding. Appendix D covers this correlation and possible causal interventions.

Table 1. Exact-string Yes/No results on AVHBench. "Scored N" is the number of samples whose greedy completion is exactly `Yes` or `No`; remaining generations are omitted from this table. The main pattern is stable across rows: answer bias shifts, but accuracy stays near chance on the scored subset. All three methods failed to exceed VideoLLaMA 2-7B-AV's base 51.7% accuracy. **Note:** Accuracies are reported on the full AVHBench balanced set. Standard errors for these estimates are $< 0.8\%$ across all configurations, confirming that the observed shifts in output priors are statistically significant.

Configuration	Scored N	Overall Acc	GT=Yes Acc	GT=No Acc	Yes/No Preds
ACTC	5165	49.8%	55.5%	44.1%	2895 / 2270
ACTC + TATI	5288	49.0%	46.9%	51.1%	2529 / 2759
ACTC + TATI + AMD	5299	50.2%	35.1%	65.2%	1853 / 3446
InternVideo2	5302	52.3%	51.3%	53.3%	2600 / 2702

5. Limitations

Our mechanistic evidence is limited to the VideoLLaMA 2 architecture, though our behavioral baseline includes InternVideo2. Table 1 is restricted to exact `Yes`/`No` completions and should be read as a strict probe.

The interleaving process increases total sequence length, confounding temporal alignment effects with context window size. Mechanistically, the head shutoff audit used 100 samples from the conflict split, which carries the 92% No-label imbalance noted in Section 3.1. Head classifications may reflect this imbalance rather than genuine conflict-resolution function. Bootstrap validation across balanced subsets remains future work. Finally, the logit lens projects intermediate residual states through the final unembedding matrix, which implicitly assumes representational alignment between intermediate and final layers. This assumption is not guaranteed and should be tested with tuned lens methods in future work (Belrose et al., 2023).

6. Conclusion

Under audio-visual conflict, neither base InternVideo2 nor aligned VideoLLaMA 2-7B-AV shows strong resilience against audio hallucinations, as alignment interventions merely reshape answer bias rather than improve audio grounding. The late-layer commitment pattern suggests that improving early audio-visual synchronization is not enough to guarantee audio-grounded computation at generation time.

We find evidence consistent with prior dominance, specifi-

cally with a late-layer commitment to an internally preferred visual or text prior, as a practical driver of audio hallucinations across both plain and aligned AV-LLMs. The immediate implication is for AV-LLM training: audio prior dominance is localized, structurally stable across fine-tuning, and therefore targetable. Future work should target the late-layer commitment mechanism directly, for example, through steering vectors applied at layers 24-27 to suppress ungrounded token probabilities, or through training objectives that penalize early snap-layer stabilization.

References

Belrose, N., Ostrovsky, I., McKinney, L., Furman, Z., Smith, L., Halawi, D., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., and Bing, L. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Chowdhury, S., Gani, H., Anand, N., Nag, S., Gao, R., Elhoseiny, M., Khan, S., and Manocha, D. Aurelia: Test-time reasoning distillation in audio-visual llms. *arXiv preprint arXiv:2503.23219*, 2025.

Chung, S., Kim, S. Y., Chee, Y., and Ro, Y. M. Mad: Modality-adaptive decoding for mitigating cross-modal hallucinations in multimodal large language models. *arXiv preprint arXiv:2601.21181*, 2026.

Guo, Y., Ma, S., Ma, S., Bao, X., Xie, C.-W., Zheng, K., Weng, T., Sun, S., Zheng, Y., and Zou, W. Aligned better, listen better for audio-visual large language models. *arXiv preprint arXiv:2504.02061*, 2025.

Jung, C., Jang, Y., Choi, J., and Chung, J. S. Fork-merge decoding: Enhancing multimodal understanding in audio-visual large language models. *arXiv preprint arXiv:2505.20873*, 2025a.

Jung, C., Jang, Y., and Chung, J. S. Avcd: Mitigating hallucinations in audio-visual large language models through contrastive decoding. *arXiv preprint arXiv:2505.20862*, 2025b.

Leng, S., Xing, Y., Cheng, Z., Zhou, Y., Zhang, H., Li, X., Zhao, D., Lu, S., Miao, C., and Bing, L. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*, 2024.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. *arXiv preprint arXiv:2202.05262*, 2022.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Sung-Bin, K., Hyun-Bin, O., Lee, J., Senocak, A., Chung, J. S., and Oh, T.-H. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *International Conference on Learning Representations*, 2025.

Tang, Y., Shimada, D., Bi, J., and Xu, C. Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue. *arXiv preprint arXiv:2403.16276*, 2024.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2023.

Wang, Y., Li, K., Li, X., Yu, J., He, Y., Wang, C., Chen, G., Pei, B., Zheng, R., Xu, J., Wang, Z., et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024.

A. Appendix

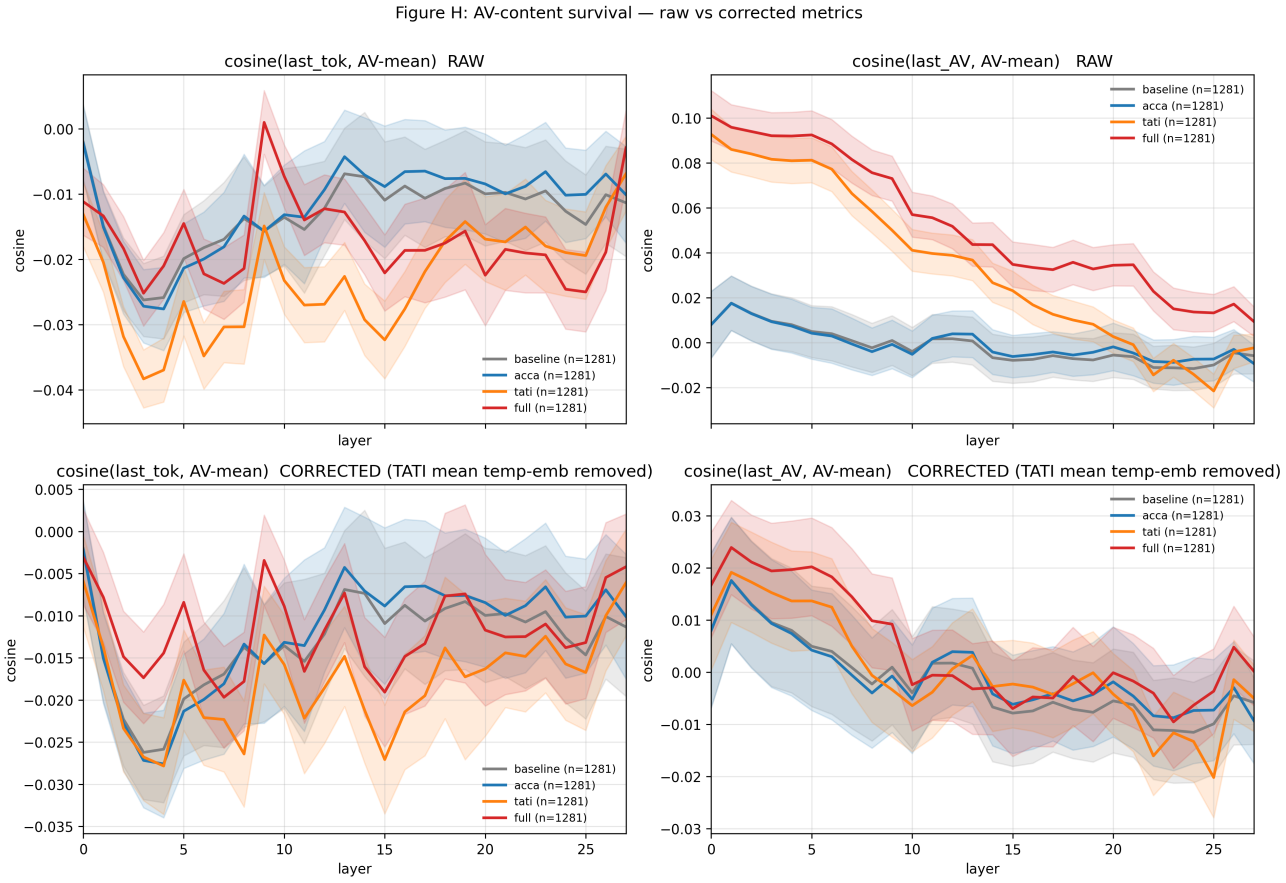

Figure 4. Exploratory cosine-similarity proxy $s_{AV}^{(\ell)}$ across layers. We report this only as a heuristic representation diagnostic, not as a direct measure of modality attribution.

Figure 4 plots an exploratory proxy, $s_{AV}^{(\ell)}$, defined as cosine similarity between late-layer audio and visual token states after removing the shared timestamp embedding used by TATI-style interleaving. We include this plot only as suggestive context. Because deep residual states have already mixed information through attention and MLP updates, this subtraction should not be interpreted as an exact decomposition of modality content. Accordingly, the main paper does not rely on this proxy for its core claim.

B. Training Setup

All three pipeline stages were implemented using Low-Rank Adaptation on top of the frozen VideoLLaMA 2-7B-AV backbone. Supervised fine-tuning was conducted on 8 A100 GPUs with a batch size of 64. Custom vision and audio projectors were trained alongside the LoRA adapters to map modality-specific features into the shared language-model embedding space.

C. Expanded Related Works

Recent benchmarks such as AVHBench (Sung-Bin et al., 2025), AVCD (Jung et al., 2025b), and broader multi-modal hallucination evaluations (Leng et al., 2024) establish the importance of testing grounded audio-visual reasoning. Temporal interleaving and synchronization are widely used design choices in AV-LLMs, including AVicuna-style interleaving (Tang et al., 2024), while recent methods also target hallucination at training time or decoding time through stronger alignment or adaptive decoding (Guo et al., 2025; Chung et al., 2026; Jung et al., 2025a; Chowdhury et al., 2025). Our paper is complementary to that line of work: instead of proposing a new alignment module, it asks whether these alignment pressures

improve grounded inference specifically under contradiction. Our mechanistic interpretability methodology stems from techniques developed for uni-modal transformers. We use the logit lens (nostalgebraist, 2020) to track per-layer predictions, and note that future works should aim to validate these results using the tuned lens (Belrose et al., 2023). Head shutoff ablations follow the methodology of (Wang et al., 2023). Late-layer MLP commitments identified in Stage 2 are consistent with work showing that late MLP layers serve as key-value stores of knowledge (Meng et al., 2022), which suggests that the same layers may be responsible for prior dominance in multi-modal LLMs.

D. Discussion & Future Mitigation Strategies

While our mechanistic audit identifies a clear temporal gap between conflict-detection (layers 15–18) and prior commitment (layer 25.5), we acknowledge that this observation is currently correlational. It remains possible that the late-layer stabilization reflects downstream information propagation rather than a dedicated "prior dominance" mechanism. To distinguish between these hypotheses, future work should employ causal interventions, specifically, steering vectors or residual stream dampening, applied to layers 20–24. By dampening the model's internal representation of the "prior" immediately after the conflict-detection heads, one could test whether the model is forced to rely on earlier, grounded audio-visual computations. Additionally, while our audit uses an operational threshold of $\Delta = \pm 0.01$, future work would benefit from statistical calibration through bootstrapping or randomized ablation baselines to rigorously quantify the impact of conflict-resolution heads. Finally, extending this analysis to structurally distinct architectures remains a priority to verify if the "snap layer" is a universal property of autoregressive multimodal systems.