MECHANISTIC SIGNS OF MODALITY COLLAPSE IN MULTIMODAL EVENT UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modality collapse occurs when a multimodal model relies on one dominant modality and ignores others. So far, this phenomenon has been established behaviorally, by modifying the input prompt and analyzing the model's predictions. Such analyses cannot show *whether* a model's computational flow actually leveraged information in a modality. In this work, we focus on the task of identifying important moments in football games and mechanistically verify modality collapse through causal intervention. We separately ablate a moment's video and its commentary transcription from the model's internal representations rather than the prompt, and quantify the impact each modality has on the final prediction. Using the signs of these impacts, we place each moment in one of four diagnostic quadrants: synergy, video collapse, commentary collapse, or conflict. Collapse takes opposite directions across the two classes: where it occurs, important moments are decided by the video and non-important ones by the commentary. Even when a model classifies a moment correctly, collapse costs it confidence. Synergy occurs predominantly for important moments and conflict for non-important ones. Within the important class, synergy monotonically decays from prototypically important moments (goals) to contextually important ones (shots-on-target). To complement these intervention-based results, we also observe how each modality's attention share and the model's decision evolve across layers. Together, our findings indicate that collapse is as much a characteristic of the moment as of the model. Future work can build on these insights to causally localize *where* collapse occurs for enabling inference-time mitigation.

1 INTRODUCTION

Surikuchi et al. (2026) analyzed the influence of modalities on multimodal large language models (MLLMs) by measuring changes in their confidence (via logit differences) when classifying whether a moment in a football game is important or not, under various combinations of modalities in the input prompt: video and commentary (LV), commentary only (L), or video only (V). Across two representative MLLMs, they observed that the visual modality is more crucial for models to correctly identify important moments, whereas accompanying commentaries are more crucial for identifying non-important moments. Furthermore, they showed that for contextually important moments (such as corners or shots-on-target), multimodal input does not lead to any increase in confidence of models. They attributed this evidence to a phenomenon called *modality collapse* in multimodal models (Zhang et al., 2024; Chaudhuri et al., 2025), wherein a model fails to synergistically integrate information from all the input modalities as a result of unimodal dominance, and noticed that Qwen2.5-Omni-7B (Xu et al., 2025) exhibits this more severely relative to Qwen2.5-VL-7B-Instruct (Bai et al., 2025).

All of these observations and the modality collapse claim are based on behavioral evidence—changes in model confidence as a function of variations (including/excluding a modality) in the input prompt. However, changes to the input prompt also result in an input distribution shift altering the latent representations of models and thereby altering their output probability distributions (Hooker et al., 2019; Hase et al., 2021). This confound can limit the explanatory power when isolating the true influence of modalities, because variations in the input prompt can shift the final predictions.

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.

Consequently, to clearly determine how modalities interact and influence the model, we move beyond behavioral analyses and look into the internals of models to discover the representational pathways causally responsible for their predictions. To this end, we focus on mechanistically verifying the modality collapse phenomenon by structurally intervening at the level of underlying architectural mechanisms of models. Specifically, we knock out information from each modality through causal interventions while keeping the input prompts unchanged, thereby isolating the true, unconfounded influence of each modality on model outputs. We consider this as an essential progression from black-box behavioral observations towards a mechanistic understanding of multimodal integration. Our contributions are:

- We adapt attention knockout to test whether each modality is used, ablating the video and commentary spans at all layers simultaneously. We also introduce a four-quadrant diagnostic (synergy, commentary collapse, video collapse, conflict) based on the signs of their causal impacts.
- We find that collapse takes opposite directions across the two classes: commentary collapse for important moments and video collapse for non-important ones. Collapse costs models confidence even when they classify a moment correctly, and synergy decays monotonically from prototypically important moments (e.g., goals) to contextually important ones (e.g., shots-on-target).
- Using attention share and logit lens, we show that attention allocation hardly differs across classes while causal impacts invert. This suggests that collapse is as much a characteristic of the moment as of the model.

2 Related Work

Bereska & Gavves (2024) divide mechanistic interpretability methods into two categories: interventional and observational. Interventional approaches (such as activation patching) involve manipulating the internal mechanisms of models to measure consequences (searching for causality), whereas observational approaches are useful to monitor internal states of models without modifying them.

Recent work surveying the state of mechanistic interpretability for MLLMs highlights a substantial gap in the availability of methods compared to large language models (LLMs) and recommends efforts towards adapting existing LLM-specific approaches to interpret MLLMs (Lin et al., 2025; Liu et al., 2025b).

Within the interventional category, *causal ablation* methods (Bereska & Gavves, 2024; Lepori et al., 2025) can be most readily adapted. They are based on the principle that removing a model's component should alter its prediction. Recent work applies this idea to MLLMs: Zhang et al. (2025) separately block the attention edges of both vision and language modality positions across a sliding window of layers, to identify where the two modalities are integrated. Kaduri et al. (2025) block image tokens as a group from reaching text and generated tokens, to localize where visual information is consumed. Gou et al. (2026) adopt attention knockouts in Video-LLMs to study how different aspects of video information (spatial and temporal) reach the generated answer in video question answering.

A commonality across all these previous adaptations is their objective to locate *where* in the model a modality is processed. For instance, Zhang et al. (2025) achieve this by moving the attention knockout systematically across a window of model layers. We adopt their core intervention technique of ablating modality-specific attention edges, but for a different purpose: instead of testing *where* the two modalities are integrated in the model, we adapt it to study *whether* both modalities are actually used, or if there is a collapse. Specifically, we perform independent knockouts on the video and commentary spans of each moment within the annotated MOMENTS subset released by Surikuchi et al. (2026). Crucially, we apply this ablation by intervening at all layers simultaneously, rather than one window at a time. We elaborate our approach in the next section.

3 MODALITY COLLAPSE: VERIFICATION

3.1 INTERVENTION THROUGH ATTENTION KNOCKOUT

In this section, we describe our causal ablation approach to verify modality collapse in MLLMs. First, we formalize the attention mechanism in MLLMs and then use that notation to define our intervention through knockout, including quantification of results.

We start by formalizing the binary classification task defined in Surikuchi et al. (2026) and focus only on the LV setting, given that both MLLMs support this specific combination of multimodal inputs. Specifically, each football game moment is provided as a single multimodal prompt input (comprising the moment's video frames, its associated commentary transcription, and the task instruction) to the model, which is mapped into an ordered sequence of tokens by the model's tokenizer: $x = (x_1, \ldots, x_T)$ of width T, with position indices denoted by the set $\mathcal{T} = \{1, \ldots, T\}$. To differentiate where the vision and language components are positioned relative to the prompt instruction, we partition $\mathcal{T}$ as:

$$\mathcal{T} = \mathcal{I} \cup V \cup L \cup \{t^*\} \tag{1}$$

where V and L are sets of indices corresponding to the video and commentary tokens, respectively, with $\mathcal{I}$ representing the remaining prompt tokens. We exclude boundary markers such as <V_START> from the video tokens set because they sink attention, which tends to skew/corrupt the analysis. This is a fundamental architectural phenomenon of multimodal decoder models wherein the model dumps its attention on positional markers when the actual context is not semantically required (Kang et al., 2025). Considering that the classification output of autoregressive models is at the final token position in a sequence, we denote it as t^*, which for a given input x would transform either into the ground-truth class y or the other class $\bar{y}$. Therefore, to have an influence on the output of the model, information in a modality (V or L) must flow to this position t^*.

Attention weights and attention share. In a decoder-based transformer (the architecture of the two MLLMs used in this work), cross-positional information routing is handled only through attention mechanisms. At any given decoder layer and attention head, the final position t^* integrates information from its preceding tokens by computing a weighted average over the sequence:

$$\alpha_t = \text{softmax}_t\left(\frac{q_{t^*} \cdot k_t}{\sqrt{d_h}}\right), \quad \sum_{t=1}^{T} \alpha_t = 1, \quad o = \sum_{t=1}^{T} \alpha_t v_t \tag{2}$$

where q, k, and v are the query, key, and value vectors, d_h is the head dimension, and o is the head's output at t^*. Here, we want to emphasize the order of these operations. First, the weights (α) are determined by the queries and keys, and only then applied to the values to produce outputs. From this formulation, we define the *attention share* $\mu(S)$ (Lepori et al., 2025) allocated to a set/span of modality positions $S \in \{V, L\}$ as:[1]

$$\mu(S) = \sum_{t \in S} \alpha_t \tag{3}$$

Attention knockout. The order of operations in equation 2 enables two distinct stages during which a modality span's contribution can be ablated: the attention weights (α) or the values (v) they read. We intervene on the attention weights of a modality span by setting their pre-softmax scores to $-\infty$ on the incoming edge set $\{(t, s) : s \in S, s \neq t\}$ (at every decoder layer and query head).[2] We choose $-\infty$ for the ablated scores because softmax transforms them into an attention weight of 0, thereby blocking any information flow from the targeted span S. This causally ablates the targeted modality span S from the model's computational flow, and forces the attention weights of all other

[1] This holds at every layer and attention head, and the intervention is applied to all of them.

[2] As we explain in Appendix A, intervening on values would be an inappropriate choice when verifying modality collapse.

positions (outside S) to renormalize:[3]

$$\tilde{\alpha}_t = \begin{cases} 0, & t \in S \\ \frac{\alpha_t}{1-\mu(S)}, & t \notin S \end{cases} \quad \text{with} \quad \sum_t \tilde{\alpha}_t = 1 \tag{4}$$

Unlike approaches that modify the input prompt, this method keeps all the tokens in their original intended positions. This avoids shifts in the relative positional embeddings, thereby eliminating the confound of shifts in the underlying semantic distributions of models.

Measuring modality-span impact. We use the same notation from Surikuchi et al. (2026) to represent the difference between the logits (ΔZ) predicted for the ground-truth class y and the other class $\bar{y}$ for a given moment:

$$\Delta Z = Z_y - Z_{\bar{y}} \tag{5}$$

We then quantify the causal impact of ablating a modality span S as:

$$\text{impact}_S = \Delta Z - \Delta \tilde{Z}_S \tag{6}$$

where $\Delta \tilde{Z}_S$ is the same quantity recomputed with the knockout applied to S. Because ΔZ is oriented towards the ground truth (rather than towards the model output), $\text{impact}_S > 0$ would indicate that the information provided by modality S to the model drove it towards the correct class, whereas an $\text{impact}_S \leq 0$ means the modality either distracted the model or was ignored by it. By mapping the resulting impact values for each moment as a point $(\text{impact}_V, \text{impact}_L)$ on a 2D coordinate system, the signs of the causal impacts place each moment into one of the four diagnostic quadrants:

- *Synergy* ($\text{impact}_V > 0$, $\text{impact}_L > 0$): Both modalities contribute positively towards predicting the ground-truth.
- *Commentary Collapse* ($\text{impact}_V > 0$, $\text{impact}_L \leq 0$): The model relies primarily on the video and the commentary is unused or irrelevant.
- *Video Collapse* ($\text{impact}_V \leq 0$, $\text{impact}_L > 0$): The model relies primarily on the commentary and the video is unused or irrelevant.
- *Conflict* ($\text{impact}_V \leq 0$, $\text{impact}_L \leq 0$): Neither modality is useful for the model to predict the ground-truth class.

We use this taxonomy for validating and interpreting our experiments and results outlined below.

3.2 Experiments

We evaluate the intervention on two MLLMs—Qwen2.5-VL-7B-Instruct (**VL**) (Bai et al., 2025) and Qwen2.5-Omni-7B (**Omni**) (Xu et al., 2025)—using the subset of football game moments manually annotated with their event category from the MOMENTS dataset.[4]

Setup. Our test data subset comprises 146 moments (103 important and 43 non-important), spread across three categories: GOAL (48), CORNER/THROW-IN (49), and SHOT-ON-TARGET (49). Each prompt has a median width of ∼11650 tokens, of which the video span V occupies ∼11520 tokens and the commentary span L occupies ∼46 tokens. The two modality spans, therefore, differ substantially in width. To ensure this difference does not confound our measurements, we conduct a control ablation by knocking out a random contiguous span of video tokens matched in width to the commentary (see Appendix B). In this experiment, we observe a negligible impact of the controlled ablation compared to full video ablation, reflecting that the causal impact of the visual modality is driven by its overall semantics rather than its sequence width.

[3] Given that attention weights over the sequence must sum to 1 for preserving numerical stability in subsequent layer normalizations and feed-forward networks of the transformer architecture, zeroing of weights on S requires the attention weights at non-ablated positions to scale proportionally: by a factor of $(1 - \mu(S))^{-1}$.

[4] Subset of MOMENTS with category annotations: `https://github.com/akskuchi/MOMENTS/blob/main/analyses/category_annotations.json`

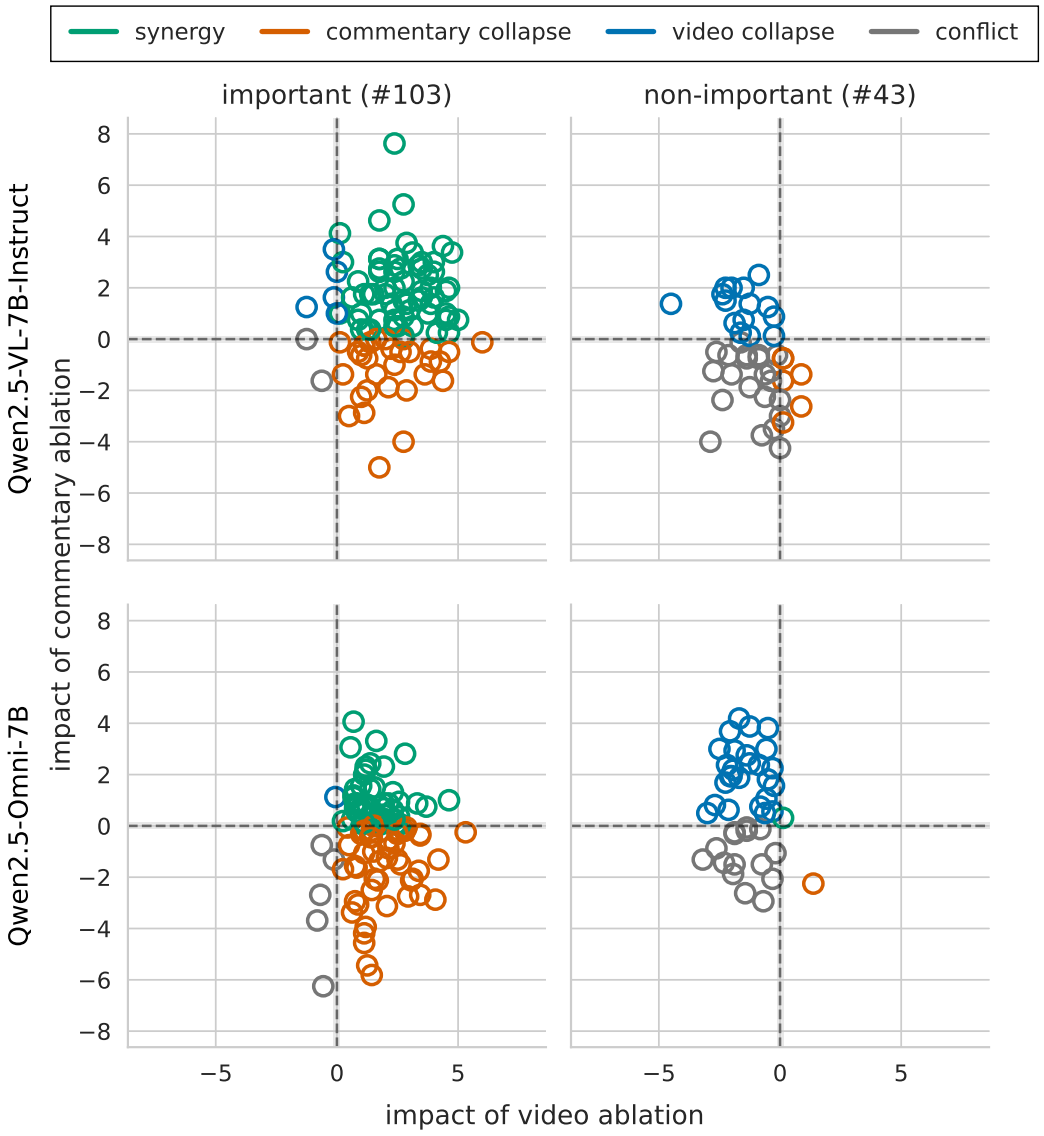

Figure 1: Causal ablation impacts—models on rows and classes on columns. Each moment is a point at $(\text{impact}_V, \text{impact}_L)$, and the quadrant it falls in is its label. Figure 4 (Appendix C) breaks down important moments by category.

Results. Figure 1 shows that collapse occurs in opposite directions across the two classes. For important moments, *commentary collapse* is prevalent: 30% using VL; 47% using Omni, against 5% and 1% of *video collapse*. For non-important moments, *video collapse* is prevalent: 35% and 60%, against 14% and 2% of *commentary collapse*. These mechanistic intervention results causally verify the behavioral findings established in Surikuchi et al. (2026). By ablating each modality from the computational flow of models rather than from the prompt, we eliminate the behavioral confounds. The other two quadrants divide in a similar manner: *synergy* occurs almost only for important moments (63% using VL; 48% using Omni) and *conflict* mostly for non-important ones (51% using VL; 35% using Omni). For moments in the *conflict* quadrant, ablating either modality moves the prediction towards the ground-truth class (y) suggesting a potential 'conflict' between information present in different modalities.

		Synergy	Commentary Collapse	Video Collapse	Conflict
VL	correct (n =114)	65 (57%, +5.49)	28 (25%, +2.75)	16 (14%, +1.75)	5 (4%, +1.70)
	incorrect (n =32)	0 (0%, –)	9 (28%, −0.60)	4 (13%, −1.00)	19 (59%, −1.70)
Omni	correct (n =117)	49 (42%, +5.78)	45 (38%, +3.37)	17 (15%, +1.27)	6 (5%, +1.46)
	incorrect (n =29)	1 (3%, −1.75)	4 (14%, −1.67)	10 (34%, −1.27)	14 (48%, −2.63)

Table 1: Distribution of classification results across quadrants. Each cell reports the number of moments (their share of the row, and their mean model confidence ΔZ). ΔZ is positive for correctly classified moments and negative for incorrect ones.

In Table 1, we group moments based on classification correctness of MLLMs, and report for each diagnostic quadrant the number of moments that fall in it, along with mean model confidence (ΔZ) for the group. The *synergy* quadrant almost exclusively accounts for correctly classified moments (0% of VL misclassifications; 3% of Omni). With the *conflict* quadrant, we observe the opposite trend: it predominantly contains misclassified moments (59% of VL misclassifications; 48% of Omni). While both the collapse quadrants contain correctly classified moments, the model confidence is substantially lower compared to correct classifications in the *synergy* quadrant. This shows that the diagnostic quadrants are directly indicative of both correctness and confidence of MLLMs in classifying moments.

With this insight, we next examine what determines the quadrant a moment falls in. Figure 4 (Appendix C) shows that *synergy* monotonically decays from prototypically important moments (e.g., goals) to moments whose importance is defined contextually (e.g., corners, shots-on-target). Considering the direction in GOAL → CORNER/THROW-IN → SHOT-ON-TARGET among important moments, we observe a 79 → 59 → 39% drop using VL and 73 → 33 → 18% using Omni. This drop in *synergy* tends to lead specifically towards *commentary collapse*, while *video collapse* is almost negligible throughout. This suggests that the degree of modality collapse depends on the characteristics of moments: the same model integrates both modalities on GOAL moments and shows collapse predominantly for CORNER/THROW-IN and SHOT-ON-TARGET moments.

Based on the overall results across models, we also find that Omni shows less *synergy* on important moments compared to VL (48% vs. 63%), more *commentary collapse* (47% vs. 30%), more *video collapse* on non-important moments (60% vs. 35%), and a stronger decay on *synergy* as importance becomes more contextual (73 → 33 → 18% vs. 79 → 59 → 39%). Considering that both the MLLMs are assessed on the same data and with the same intervention, we conclude that VL integrates multiple modalities better than Omni, and this mechanistically verifies the same observation made in Surikuchi et al. (2026).

4 Observational Analyses

By applying an interventional approach (at every layer simultaneously), we established *whether* a modality is used by models. In this section, we analyze the forward pass of models without intervening, to uncover layer-level insights about modality usage. To this end, we leverage two methods: attention share (Lepori et al., 2025) and logit lens (nostalgebraist, 2020).

To trace how models allocate attention across the input prompt, we define the layer-level **attention share** $\mu(S)$ (see equation 3) as the proportion of the final token t^*'s attention budget allocated to a target modality span $S \in \{V, L\}$. For each decoder layer, this share is calculated by summing the attention weights over all positions $t \in S$, and averaging the result across all attention heads.

Logit lens applies the model's final norm and the unembedding matrix to the hidden state at t^* after each layer, enabling us to obtain layer-level logits for y and $\bar{y}$. We then compute the corresponding ΔZ^l for every layer (l) to track how the YES/NO decision forms across different layers of MLLMs.

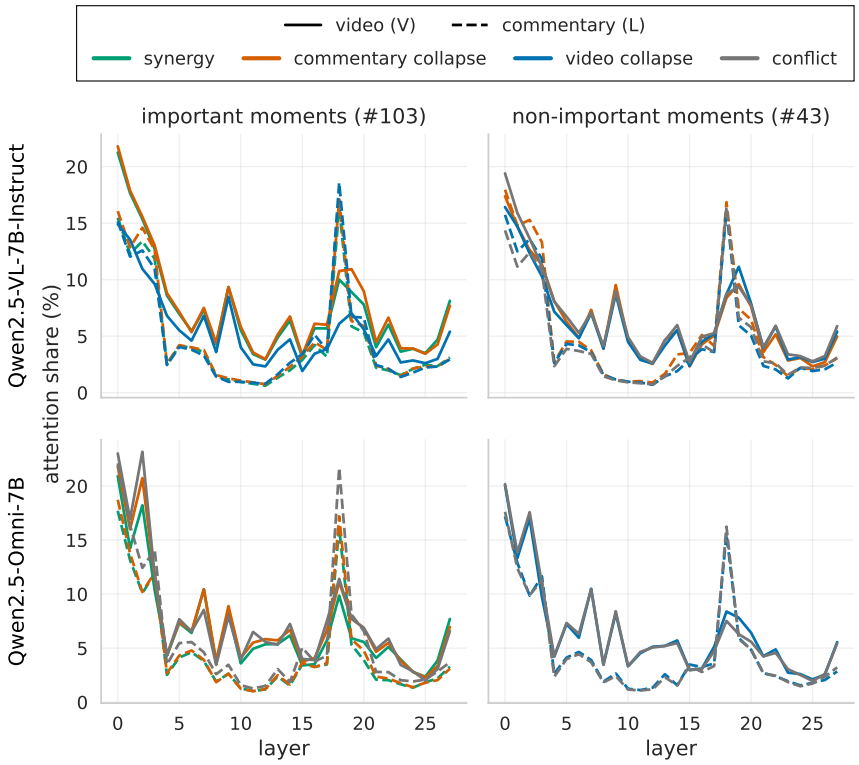

Figure 2: Attention share of each modality span ($S \in \{V, L\}$) across layers, with models on rows and classes on columns.

Results. Figure 2 visualizes layer-level modality span attention shares and Figure 3 shows logit difference evolution across layers. Beyond the first few layers, we find that ∼90% of the final token t^*'s attention falls on neither modality and goes instead to the system prompt and the task instruction. Of the remaining attention, the video receives a larger share than the commentary at nearly every layer, in both MLLMs for both classes of moments. However, this only shows how models distribute their attention across the prompt, and not when they decide on an answer. For this purpose, we use the logit lens and find that moments in the commentary collapse quadrant settle by layer 2, whereas most moments in the *video collapse* quadrant settle only at later layers (layer 23 in VL; 25 in Omni). This difference also corresponds to class predictions: when MLLMs classify a moment as important, they settle on that answer by layer 2 and classifying a moment as non-important requires a late-layer decision crossover (see Figure 3). Consequently, both models classify the majority of important moments correctly (95% using VL; 97% using Omni) compared to non-important ones (37% using VL; 40% using Omni). This difference in accuracy reflects that for most non-important moments, the late-layer crossover fails to happen. For non-important moments that models do correctly classify, causal ablation places most of them in the *video collapse* quadrant (11 of 16 using VL; 16 of 17 using Omni), confirming that commentaries drive the decision crossover toward the non-important class. This further supports the same claim made in Surikuchi et al. (2026).

Here we want to underline that attention share in itself should not be interpreted as a causal contribution (Bibal et al., 2022). However, we find that when aggregated, the attention share results

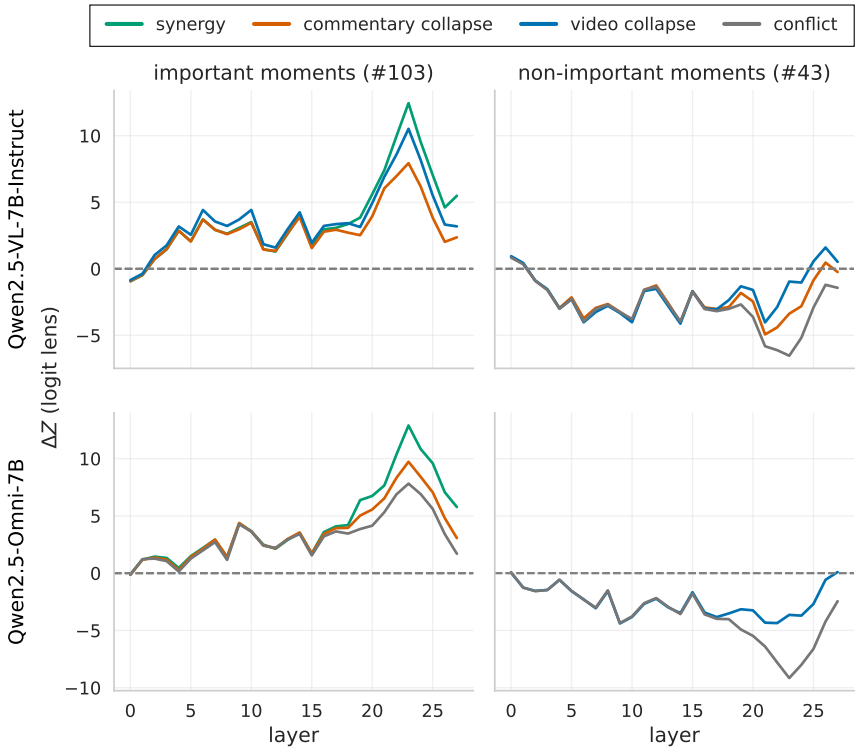

Figure 3: Layer-level ΔZ from the logit lens, with models on rows and classes on columns.

agree with the causal ablation results reported in Section 3: MLLMs attend to the video more than the commentary, and the ablation results show that video carries a larger aggregate causal effect on the prediction.[5] But that agreement is only at the aggregate level: the attention allocation hardly differs between important and non-important moments—using VL, V share moves from 7.4% for important to 6.6% for non-important, but L share stays at ∼4.7%—while causal ablation results invert between both classes. These layer-level observations are only intended to inform where in the model a particular modality is active rather than whether it is contributing to the decision.

5 Discussion and Conclusion

In this work, we transitioned from behavioral output-centric analyses to mechanistically interpreting modality integration in MLLMs. Through causal ablation, we verified that modality collapse is indeed a phenomenon within the internal representations and computational flow of models, and not merely a behavioral consequence of input prompt-level modifications. Both MLLMs exhibit a severe commentary collapse for contextually important moments (e.g., shots-on-target) and video collapse for non-important moments.

We also observed a systematic decay in synergy for contextually important moments (e.g., corners) compared to prototypically important ones (goals), which suggests that under contextual complexity, MLLMs resort to unimodal priors (commentary collapse) for resolution, instead of synergistically integrating signals present in different modalities. Consequently, even if MLLMs classify moments correctly by relying on a single modality, collapse costs models their confidence. The logit lens complements these findings: moments in the commentary collapse quadrant settle early (often by layer 2), whereas those in the video collapse quadrant settle only in the late layers. This reflects a general pattern in both MLLMs, where a decision is formed in the earliest layers and revised (if at all) only in the late layers, in line with existing work (Zhang et al., 2025; Lin et al., 2025).

[5] By aggregate, we mean at the level of the whole modality span S without separating classes.

This is also what makes commentary collapse the dominant failure mode for contextually important moments: the signal from the commentary arrives late (after the model settles on an answer).

Together, our interventional and observational results also mechanistically confirm the attention-contribution paradox (Guo et al., 2024): attention weights function only as routing gates and cannot be independently considered for measuring causal influence. Both MLLMs allocate similar attention share to the video of a moment regardless of its importance, and yet ablating the video moves the prediction in opposite directions across the two classes. Given that attention share stays approximately the same across classes, whereas ablation results invert between classes and decay across categories, we consider modality collapse to also be a characteristic of the moment and not just the model. We believe this likely stems from systematic biases in multimodal training data, wherein prototypically important events such as goals are overrepresented and paired with well-aligned commentaries, whereas contextually important events such as corners or shots-on-target tend to be underrepresented (and have misaligned, complex commentaries).

Our study also has limitations. We analyze two MLLMs, the same ones evaluated by Surikuchi et al. (2026); while both are widely adopted in the research community, extending the analysis to other model families would establish how far our findings generalize. Similarly, our analysis relies on a relatively small subset of 146 moments. At the same time, these moments are manually annotated with their event category, providing a focused and reliable testbed for fine-grained, category-level analysis, and our approach is not specific to this task and can readily be extended to other multimodal datasets. Methodologically, the video and commentary spans differ in width by $\sim$250$\times$, so ablating the video removes almost the entire sequence. Although our width-matched control shows that impact depends on a span's content rather than its width, ablating different proportions of the video (e.g., 10%, 25%, 50%) can reveal more about this relationship. We also ablate one modality span at a time, which limits our method from separating redundancy from synergy; we group both under the same quadrant. Crucially, our approach establishes *whether* a modality is used but does not localize *where* it is leveraged. Causally localizing collapse to specific components, e.g., through causal tracing (Li et al., 2026), can identify the responsible shortcut pathways and enable inference-time mitigation, such as model editing and modality boosting techniques that re-scale attention share to neglected commentary tokens without fine-tuning (Meng et al., 2022; Hase et al., 2023; Liu et al., 2025a). Our findings can also inform model architectures and learning procedures that dynamically adapt to differences in input data samples, instead of relying on static fusion.

AI USE STATEMENT

We used AI assistance for literature review and validation, and for adjusting and refining figures and tables. We reviewed and verified all AI-assisted outputs and take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

Both MLLMs we analyze, Qwen2.5-VL-7B-Instruct and Qwen2.5-Omni-7B, are publicly available, and the category-annotated subset of MOMENTS used in our experiments is publicly released.[6] Section 3 fully specifies our intervention: the partition of prompt positions, the attention knockout applied at every decoder layer and attention head, and the impact measure used to assign diagnostic quadrants. Appendices A and B detail the alternative intervention and the width-matched control experiment. Our code will be released upon acceptance.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025. URL https://arxiv.org/abs/2502.13923.

[6] https://github.com/akskuchi/MOMENTS/blob/main/analyses/category_annotations.json

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ePUVetPKu6. Survey Certification, Expert Certification.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3889–3900, 2022. URL https://aclanthology.org/2022.acl-long.269/.

Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. A closer look at multimodal representation collapse. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Vf9f7eNX6T.

Chenhui Gou, Ziyu Ma, Zicheng Duan, Haoyu He, Feng Chen, Akide Liu, Bohan Zhuang, Jianfei Cai, and Hamid Rezatofighi. An empirical study on how video-llms answer video questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18587–18597, 06 2026.

Zhiyu Guo, Hidetaka Kamigaito, and Taro Watanabe. Attention score is not all you need for token importance indicator in KV cache reduction: Value also matters. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21158–21166, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1178. URL https://aclanthology.org/2024.emnlp-main.1178/.

Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3650–3666. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/1def1713ebf17722cbe300cfc1c88558-Paper.pdf.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=EldbUlZtbd.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558, 06 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7uDI7w5RQA.

Michael A. Lepori, Michael Curtis Mozer, and Asma Ghandeharioun. Racing thoughts: Explaining contextualization errors in large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3020–3036, Albuquerque, New Mexico, 4 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.155. URL https://aclanthology.org/2025.naacl-long.155/.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability

and hallucination mitigation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40: 31645–31653, 03 2026. doi: 10.1609/aaai.v40i37.40431. URL https://ojs.aaai.org/index.php/AAAI/article/view/40431.

Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A. Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, Ying Shen, Barry Menglong Yao, Zhiyang Xu, Qin Liu, Yuxiang Zhang, Yan Sun, Shilong Liu, Li Shen, Hongxuan Li, Soheil Feizi, and Lifu Huang. A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models, 2025. URL https://arxiv.org/abs/2502.17516.

Qing Liu, Jianhao Zhang, Ou Wu, Michael Ng, and Yi Du. Alphaedit+: Model editing in the presence of conflicting and inconsistent knowledge, 2025a. URL https://openreview.net/forum?id=zhI2L3mDoM.

Yiming Liu, Yuhui Zhang, and Serena Yeung-Levy. Mechanistic interpretability meets vision language models: Insights and limitations. In *The Fourth Blogpost Track at ICLR 2025*, 2025b. URL https://openreview.net/forum?id=pZqvfsUpeh.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. Quantifying context mixing in transformers. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3378–3400, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.245. URL https://aclanthology.org/2023.eacl-main.245/.

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

Aditya K Surikuchi, Raquel Fernández, and Sandro Pezzelle. Where is the multimodal goal post? on the ability of foundation models to recognize contextually important moments, 2026. URL https://arxiv.org/abs/2601.16333.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, et al. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Yedi Zhang, Peter E. Latham, and Andrew M Saxe. Understanding unimodal bias in multimodal deep linear networks. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=CTEMHDSwIj.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791, 06 2025.

A Intervention through Value Zeroing

		Off-task (after ablating V)	Off-task (after ablating L)	impact$_V$ (median)	impact$_L$ (median)
VL	Value zeroing	**32%**	0%	4.44	1.75
	Attention knockout	0%	0%	1.94	1.38
Omni	Value zeroing	**64%**	0%	5.87	1.50
	Attention knockout	0%	0%	1.44	1.16

Table 2: Comparison of two interventional choices: attention knockout vs. value zeroing.

An alternative intervention choice is *Value Zeroing* (Mohebbi et al., 2023), which leaves the attention weights α as they are and replaces value representations of the target modality span S with zero vectors ($v_t = 0$ for $t \in S$). Intervening on a modality span's attention weights and renormalizing weights of the remaining positions to sum to 1—equation 4—ensures that the model output is an average over the surviving tokens. However, value zeroing yeilds the same average multiplied by (1 - $\mu(S)$). Therefore, besides removing the modality span's content, zeroing shrinks the influence of the surviving tokens (in proportion to $\mu(S)$). That proportion is not the same for the two modalities: video receives a larger share of attention than commentary at nearly every layer (see Section 4).

Consequently, this throws the model off-task (models output neither YES nor NO) when intervening on the video span—for 32% of moments using VL and 64% using Omni—compared to 0% when intervening on the commentary span. The same asymmetry is reflected in the magnitudes of measured interventional impacts: the median of impact$_V$ is 2.3× (VL) and 4.1× (Omni) larger under value zeroing relative to attention knockout, while the median of impact$_L$ is only 1.3× (see Table 2). Intervening on values therefore inflates the measured impact of the modality that receives more attention (video), and would under-report its collapse. To this end, we intervene on attention weights that removes each modality span's content through the operation without any span-dependent scaling.

B Controlled Intervention: Modality Span Width

		impact$_S$ (median)	impact$_S$ (maximum)
VL	commentary (L) ablation	1.38	7.62
	video (V; full width) ablation	1.94	6.00
	video (random —L— width of V) ablation	**0.00**	0.25
Omni	commentary (L) ablation	1.16	6.25
	video (V; full width) ablation	1.44	5.31
	video (random —L— width of V) ablation	**0.06**	0.31

Table 3: We conduct a control experiment in which we abalte a random contiguous span of the video that matches the width of the corresponding commentary and compare the impact$_S$ ($S \in \{V, L\}$) with ablating the entire video.

As shown in Table 3, knocking out a random contiguous span of the video modality moves model confidence by 0.00 for VL and 0.06 for Omni and by a maximum of 0.31 for both MLLMs. The causal impact of a modality is therefore driven by its semantic content rather than its sequence width.

C Category-level Ablation Impacts

Figure 4 breaks down important moments by category.

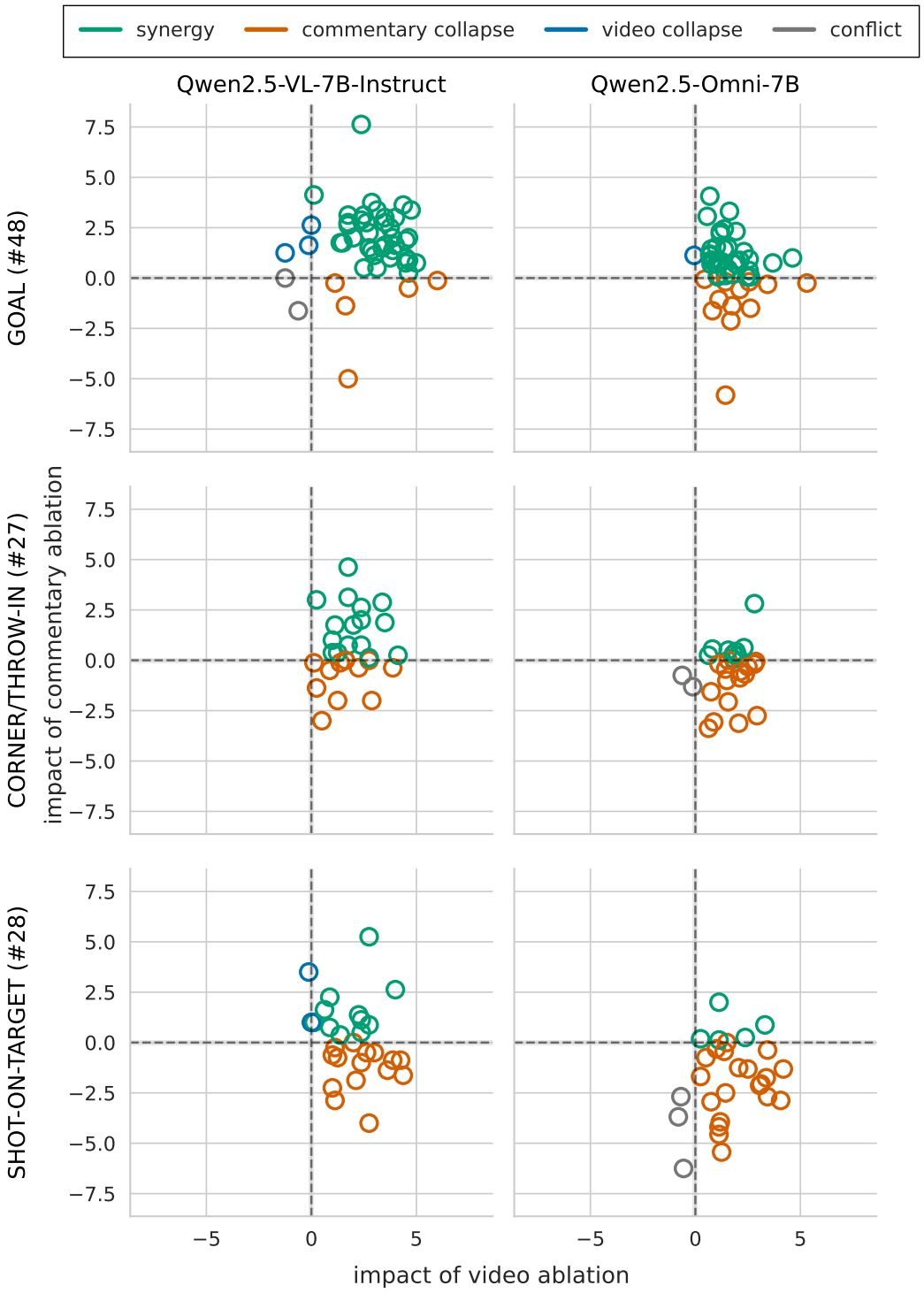

Figure 4: Causal ablation impacts for important moments with categories on rows and models on columns.