

From Vision to Language: Investigating Causal Information Flow in Multimodal Decision-Making

Davide Testa^{1,2}, Hugh Mee Wong³, Alessandro Lenci⁴, Bernardo Magnini¹, Albert Gatt³

¹Fondazione Bruno Kessler (FBK) - Trento, Italy, ²Università di Roma La Sapienza - Rome, Italy,

³Utrecht University - Utrecht, The Netherlands, ⁴ University of Pisa - Pisa, Italy,

{dtesta, magnini}@fbk.eu, alessandro.lenci@unipi.it, {h.m.wong, a.gatt}@uu.nl

Abstract

Vision-Language Models are commonly evaluated through their final predictions, but understanding whether these decisions are grounded in visual evidence requires tracing how visual information contributes to language-based decisions. With this in mind, we investigate cross-modal information flow in a video-based generative multiple-choice-like setting by applying causal interventions to video-text attention pathways and examining its effects across layers. We target spatial, causal, and temporal visual reasoning. Our results show that visual information is integrated mainly when the model processes the candidate answer options, which serve as the primary textual grounding sites for the final decision. We further show that nouns play an important role as semantic anchors during multimodal enrichment, while verbs are more relevant when temporal relations are processed. Finally, we identify a distinct pattern in temporal reasoning, suggesting that VLMs struggle to reconstruct sequential information across video frames, but we remark that such fragility may also reflect linguistic biases associated with specific temporal expressions used for defining the relation between events within a scene.

Code and data available on [Github](#).

1 Introduction

Modern Large Language Models (LLMs) and Vision-Language Models (VLMs) are commonly evaluated through benchmark-based and behavioral assessments. In this setting, abilities such as language understanding (Wang et al., 2018, 2019), visual grounding (Parcalabescu et al., 2022), or event reasoning (Li et al., 2024b) are inferred from the model’s input-output behavior on diagnostic tasks (Warstadt et al., 2020; Ribeiro et al., 2020; Chang and Bergen, 2024). While this provides an essential way to compare models and quantify their performance, it does not explain which internal computa-

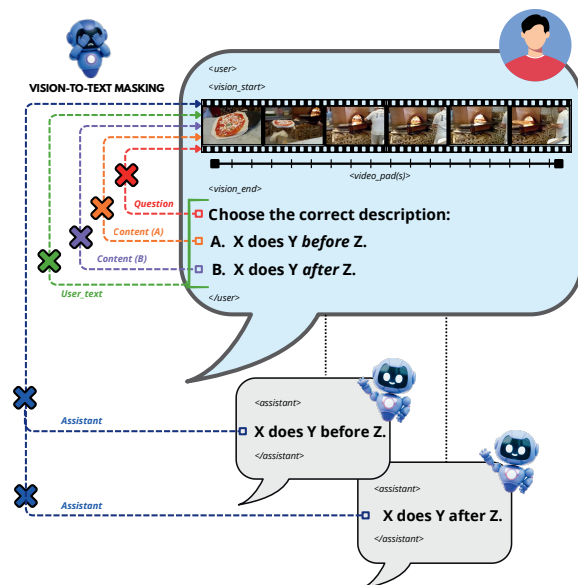


Figure 1: Overview of the Attention Knockout approach with prompt structure and region-based masking.

tions support a given prediction, nor whether models are relying on their input in the expected manner. This limitation is especially relevant for VLMs: in multimodal settings, a correct answer may suggest that the model has successfully integrated visual and linguistic information, but the final prediction alone does not reveal how this integration is implemented internally. The model may rely on visual evidence, exploit regularities in the textual prompt, or combine both in ways that remain difficult to disentangle (Wu et al., 2026). As a result, evaluating whether a VLM can solve a task is not sufficient to understand how visual information contributes to its decision (Sim et al., 2025). This problem becomes even more pronounced when videos are provided as input. Compared to static images, videos require models to process visual information across time, including objects, actions, event dynamics, and relations such as spatial configurations, causal-

ity, and temporal order (Li et al., 2024b; Liu et al., 2024; Loginova and Loguinova, 2025; Song et al., 2025). These dimensions make video-based tasks a useful testbed for studying multimodal reasoning, as they require us to trace how visual information flows through the model’s computation and supports language-based decisions.

To understand how visual evidence supports multimodal decisions, this paper investigates the internal cross-modal dynamics through which VLMs use video information in a decision-making setting. We focus on two research questions: **(RQ1)** Where and how does visual information interact with the textual representations used by the model to choose between alternative video descriptions? **(RQ2)** How does cross-modal information contribute across spatial, causal, and temporal reasoning?

We address the questions above in a controlled multimodal decision-making task in Italian, where a model must select the correct description between a caption and a minimally different foil associated with the same video. To study this process mechanistically, we apply Attention Knockout to video-text attention pathways and measure how blocking specific cross-modal interactions affects the plausibility of candidate answers.

Contributions. (i) a causal intervention framework for tracing *video-to-text* information flow in VLMs during caption–foil decision-making; (ii) an analysis of the prompt regions and token types that mediate visual grounding; (iii) an investigation of how these dynamics vary across spatial, causal and temporal relations.

The rest of this paper is structured as follows. Section 2 reviews the related work and contextualizes our approach. Section 3 describes the experimental setup, including the data (3.1), models (3.2), and methodology (3.3). Section 4 presents and discusses the results. Finally, Section 5 summarizes the main findings and concludes the paper.

2 Related Work

In vision-language research, one influential line of behavioral evaluation is represented by **foiling methods** (Shekhar et al., 2017a,b), which require the model to discriminate between alternatives that differ only in a targeted semantic element, turning visual understanding into a fine-grained contrastive problem (Madhyastha et al., 2018; Parcalabescu et al., 2022; Kesen et al., 2024; Samin et al., 2025). However, these methods do not reveal how visual

information is internally processed and used to support model decisions. Indeed, a correct answer does not necessarily imply that the model has grounded its decision in the relevant visual evidence, since it may instead exploit linguistic priors or dataset regularities, reaching the correct choice by relying only weakly, or not at all, on the visual modality. This type of failure, often referred to as *unimodal collapse*, highlights the need for analyses that go beyond final task accuracy (Parcalabescu et al., 2021; Parcalabescu and Frank, 2023; Chen et al., 2024; Sim et al., 2025).

This motivates a shift from behavioral studies to **Mechanistic Interpretability** (MI) research, which aims to identify the representations, circuits, and information flows that causally support model predictions. In unimodal language models, this line of work has produced influential accounts of transformer computation, including circuits for in-context learning and induction heads (Elhage et al., 2021; Olsson et al., 2022), factual recall and knowledge localization (Meng et al., 2022; Geva et al., 2023), and multi-hop or arithmetic reasoning (Stolfo et al., 2023; Biran et al., 2024). More recently, these methods have been extended to VLMs, where they have been used to study how visual information is localized and transformed across layers. These analyses suggest that visual representations can evolve toward more text-like forms and that visual information flow varies across reasoning tasks (Neo et al., 2025; Zhang et al., 2025a).

Attention mechanisms are an important target for mechanistic investigation as they regulate how information is exchanged across tokens and composed into internal representations. However, attention patterns alone are not sufficient for causal explanations, motivating intervention-based methods that operate directly on attention connections. Recent work has analyzed attention heads and attention-based information flow in VLMs, identifying heads involved in multimodal integration, grounding, object localization, and image-to-text-transfer (Golovanevsky et al., 2025; Kang et al., 2025; Kim et al., 2025). Among these methods, **Attention Knockout** selectively blocks attention edges between subsets of tokens, effectively performing activation-level ablations on attention pathways (Geva et al., 2023). This enables researchers to block attention from image tokens to query or generated tokens and test where visual information is used during inference (Neo et al., 2025; Kaduri et al., 2025; Zhang et al., 2025a). However, most

existing work on VLMs still focuses on image-based tasks (Ortu et al., 2024; Wang et al., 2025; Zhang et al., 2025b; Neo et al., 2025; Cohen et al., 2025), leaving open questions about how modality-specific circuits emerge and interact in more complex multimodal scenarios such as videos.

Our work connects two previous lines of research by applying attention knockout to a caption-foil discrimination problem. This design combines the structure of multiple-choice question-answering evaluation with the fine-grained control afforded by foiling methods, while also allowing us to examine how the model internally supports its decision.

3 Experimental Setting

We consider a generative caption-foil task based on video understanding. Given a multimodal prompt pairing a video with a text (Figure 1), the model is asked to select the most plausible description among two candidate options. The candidates correspond to a *caption* and a *foil* linked to the video, where the foil is a minimally altered version of the caption designed to introduce a subtle but semantically critical inconsistency (Shekhar et al., 2017b). This setup enables fine-grained evaluation of the model’s ability to ground linguistic reasoning in visual content.

3.1 Data

To construct our evaluation data, we leverage subsets of the MAIA benchmark (Testa et al., 2025a,b), which provides carefully curated Italian *caption-foil* pairs associated with videos. In particular, we focus on three subsets capturing different types of reasoning: (i) **SPATIAL**, involving spatial relations between entities; (ii) **TEMPORAL**, involving ordering and dynamics of events; (iii) **CAUSAL**, involving cause-effect relationships. Each subset comprises 1,600 caption-foil pairs associated with the 100 videos in the dataset, corresponding to 16 pairs per video within each subset and 4,800 pairs overall. Table 1 shows representative caption-foil pairs from the three MAIA subsets. The assignment of the caption and foil to option letters *A* / *B* within the decision-making task is then randomized to avoid position biases.

By controlling the semantic and lexical differences within caption-foil pairs, and by considering subsets that capture distinct types of reasoning, this setup provides a controlled testbed to probe and

disentangle the contribution of visual information to model predictions both at the instance level and across different reasoning categories.

3.2 Models

We tested **Qwen2.5-VL 7B** (Qwen et al., 2025) and **LLaVA-OneVision 7B** (Zhang et al., 2024; Li et al., 2024a), two VLMs from different model families that both pair a vision encoder with a decoder-only language model. In our experimental setting, each video is sampled into 16 frames, whose patches are processed by a vision encoder and mapped into visual token embeddings. These embeddings are concatenated with the textual tokens and jointly processed through self-attention within the language-model backbone. This shared architecture allows us to apply the same experimental framework to both models.

Despite this shared paradigm, the models differ in their visual and spatiotemporal encoding, positional representations, and training design, allowing us to assess whether these choices affect visual information flow.¹

3.3 Methodology

To analyze how visual information contributes to model predictions, we adopt the established **Attention Knockout** methodology. Building on work by Zhang et al. (2025b), our implementation intervenes directly on the pre-softmax attention-score matrices during the forward pass, rather than relying on hook-based approximations. The key idea is to selectively block specific attention connections and measure the resulting effect on model outputs. Concretely, we intervene directly on the self-attention mechanism of the VLM language backbone by applying a patch function (i.e., a targeted modification) to the original attention forward pass, without altering its underlying causal nature. Given a sequence of tokens, the classical attention operation $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$ can be interpreted as a set of interactions (i.e., the Attention matrix) between query tokens and key tokens, where each query token attends to key tokens in the sequence to gather relevant contextual information. Our framework selectively removes subsets of these interactions by modifying the attention mask during the forward pass. For each transformer layer, we define a set of masking con-

¹Both models are available on the Hugging Face Hub. For further information, see Appendix C.

Dataset	Option type	Sentence
SPATIAL	<i>Caption</i>	At the end of the scene, the little girl walks on a carpet .
	<i>Foil</i>	At the end of the scene, the little girl walks on a path .
TEMPORAL	<i>Caption</i>	Right after cutting the pizza, the girl hands it to a person in front of her.
	<i>Foil</i>	Before cutting the pizza, the girl hands it to a person in front of her.
CAUSAL	<i>Caption</i>	The girl is upset because of a cockroach .
	<i>Foil</i>	The girl is upset because of a loud noise .

Table 1: Caption–foil examples from MAIA subsets (English translations; originals in Table 5, Appendix B).

straints over attention connections in the form of tuples (q, k_{start}, k_{end}) , where q is a query token index and (k_{start}, k_{end}) defines a contiguous span of key tokens. For each of these tuples, attention from query q to all keys in the specified span is blocked by setting the corresponding attention scores to $-\infty$, effectively removing their contribution after the softmax computation. The same masking constraints are applied globally across all transformer layers in a single forward pass, with no layer-wise sliding window, enabling us to study how information flows through the network under controlled perturbations. A key aspect is the explicit separation between query tokens, which define where the intervention is applied, and key spans, which define the source of information being suppressed. This separation allows us to model different directions of information flow. In the main analysis, we focus on the *vision-to-text* direction, blocking attention from textual queries to visual tokens to directly test whether textual tokens in the prompt acquire task-relevant information from the visual stream and integrate visual evidence into language-token representations. The complementary *text-to-vision* direction blocks attention from visual queries to textual tokens, and works as an additional control analysis (see Appendix A.1). This intervention design enables fine-grained control over cross-modal interactions and provides a way to probe the contribution of specific input components to the model’s internal computations.

Region-Based Attention Knockout. We build the attention knockout framework in our multi-modal task setting by defining semantically meaningful regions over the input sequence (see Figure 1). Given the structured prompt, we identified the following regions: the *question*², the content of option A, the content of option B, the *assistant content* (i.e., all the response tokens), and a *vision* region corresponding to the tokens encoding the

video input. These regions are automatically identified through token-level position mappings.

In the main analysis, we apply region-based masking in the *vision-to-text* direction by selectively blocking attention from each of these textual regions to the vision one. This allows us to trace the flow of information from video to the prompt components and generated output.³

Token-level Attention knockout. While region-based masking enables a coarse-grained analysis of cross-modal interactions, it does not capture which specific linguistic elements within a region are responsible for the observed effects. We therefore extend our framework to a finer-grained token-level masking strategy, where attention constraints are defined over subsets of tokens. In this experimental setup, we focus on the content of options A and B within the multiple-choice prompt, as they represent the central decision points of the task. Concretely, instead of treating A- and B-content regions as homogeneous spans, we decompose them into subsets of tokens based on their linguistic properties. In particular, we leverage part-of-speech (POS) annotations⁴ to isolate semantically relevant categories, including NOUNS, VERBS, and PREPOSITIONS, as well as a subset-specific group, CATEGORY-CRITICAL TOKENS, corresponding to the minimal lexical elements that distinguish the caption from its foil and determine the reasoning category targeted by the pair. For example, consider the following pair in the *causal* subset: *The man fell into the water because he lost his balance* versus *The man fell into the water because he was pushed*. The underlined spans correspond to the category-critical tokens, as they encode the alternative causal explanations that determine which candidate is correct with respect to the video input. We applied the same (q, k_{start}, k_{end}) formulation for deleting specific attention edges and extended it to token-level granularity by operating on these

²In our setting, the “question” denotes the task instruction provided to the model (i.e., *Choose the correct description*), rather than an interrogative sentence about the video content.

³For completeness, we also evaluated the complementary *text-to-vision* direction in Appendix A.1, which requires an inverted text-video prompt under causal attention.

⁴Pos parsing done with the spaCy *it_core_news_sm* model.

Model	Dataset	Acc _{A/B}	Acc _{caption / foil}
Qwen2.5-VL	SPATIAL	0.69	0.67
	TEMPORAL	0.77	0.43
	CAUSAL	0.84	0.81
LLaVA-OneVision	SPATIAL	0.69	0.68
	TEMPORAL	0.73	0.44
	CAUSAL	0.83	0.81

Table 2: Behavioral performance on the caption–foil task for the three MAIA subsets, under two scoring schemes: A/B labels vs. full candidate sentences.

subsets of tokens. For instance, under the *vision-to-text* direction, we can block the contribution of visual tokens when attending to noun tokens within option A, thereby isolating the role of object-level grounding. Similarly, by masking verb or preposition tokens, we can probe how action and relational cues contribute to the integration of visual information.

Compared to region-based masking, this token-level approach provides a more precise characterization of the information flow within the model, enabling us to disentangle the contribution of different linguistic features to multimodal reasoning.

For both region- and token-based masking, we obtain layer-wise scores from the same globally masked forward pass using logit-lens readout (Nostalgebraist, 2020). At each layer, we apply the model’s final normalization and language-model head to the intermediate representations. We then use a *forced-target scoring* procedure: each caption or foil is inserted in the assistant-response position, as shown in Figure 1, and the probability of each token is computed autoregressively, conditioned on the multimodal input and the preceding target tokens. The sentence-level log-probability is then obtained by summing the autoregressive log-probabilities of its tokens, allowing us to compare the plausibility assigned to the two candidate sentences.

As a preliminary check (Table 2), we compare full-sentence scoring with A/B-label scoring,⁵ which reproduces the behavioral trends reported in Testa et al. (2025a). Full-sentence scoring yields comparable results on Spatial and Causal items but substantially lower accuracy on Temporal items,

⁵For A/B label scoring, we append *Answer only with A or B* to the original prompt in Figure 1, thereby replicating the original experimental setting. For full-sentence scoring, the predicted answer is the candidate with the higher forced-target sentence score; accuracy is therefore computed as $\frac{1}{N} \sum_{i=1}^N \mathbb{1} [\log P_M(C_i | x_i) > \log P_M(F_i | x_i)]$, where C_i and F_i denote the caption and foil, respectively; x_i is the corresponding multimodal input; and M is the model.

with performance dropping below chance level.⁶ Comparison with A/B-label scoring suggests that this drop may partly reflect lexical biases in autoregressive sentence scoring, particularly around temporal markers, rather than temporal reasoning alone (see Section 4.2).

4 Results

We now report the results of both the region-based and token-level attention knockout experiments.

Figure 2 shows, for each tested model and subsets, the layer-wise difference between the log-probability assigned to the caption and the foil under each tested masking condition, computed as $\Delta = \log P_C - \log P_F$. Positive values therefore indicate that the model assigns higher probability to the correct caption, whereas negative values indicate a preference for the foil. The results are reported across different prompt regions for which attention to the video tokens has been masked.⁷ The *none* condition corresponds to the standard multimodal forward pass without attention knockout. The *text-only* condition included in Figure 2 serves as a unimodal baseline, measuring the extent to which the models can solve the task by relying exclusively on linguistic priors, without access to multimodal information from the video. Its corresponding final-layer accuracy scores, together with those of the multimodal (*none*) condition, are reported in Table 3. The table further complements the layer-wise log-probability analysis with a final-layer decision-level evaluation, showing how each Attention Knockout intervention affects model predictions with reference to *none*.

Across both models and all three subsets, the *text-only* condition consistently underperforms the full-input *none* condition, indicating that linguistic priors alone are insufficient to fully solve the task.

Having established that access to the video improves final-layer accuracy, we next examine how information is processed across layers. Overall, both models exhibit a progressive pattern in the processing of Causal and Spatial information. This trend appears independently of the masking condition: it is visible both in the *none* setting and when attention to the video is blocked for specific

⁶We nevertheless adopt full-sentence scoring because it provides the token-level plausibility scores required for our mechanistic analyses.

⁷As noted above, we report here only the results for the *vision-to-text* masking direction, since the same general trends are also observed in the *text-to-vision* setting. The corresponding results are provided in Appendix A.1.

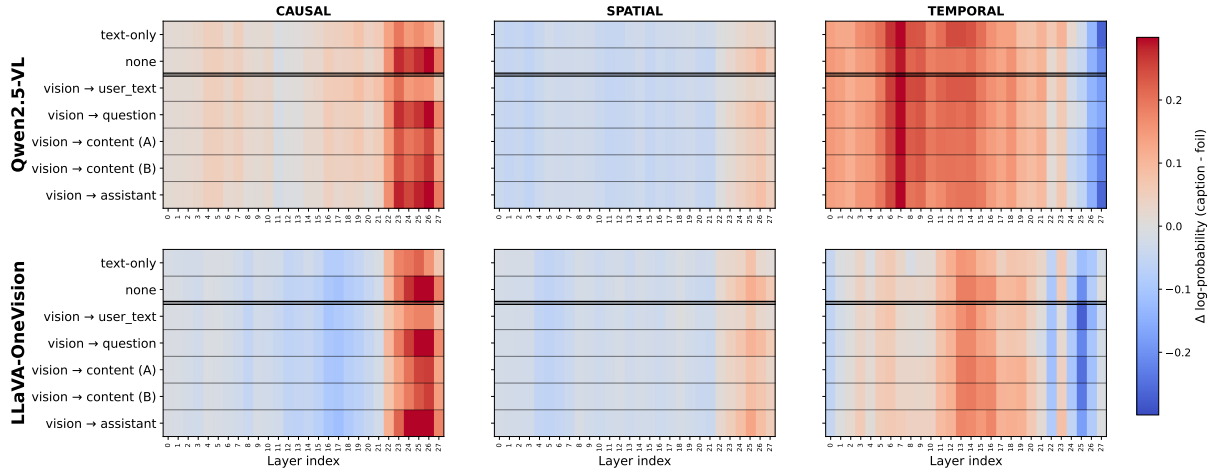


Figure 2: Region-based attention knockout results. Rows indicate the textual prompt region prevented from attending to visual tokens (i.e., *Vision* $\rightarrow$ *Region*). Colors show the layer-wise caption–foil log-probability difference. *Text-only* and *None* condition represent our two baselines.

Model	Dataset	text-only Vs. vision		vision $\rightarrow$ question		vision $\rightarrow$ content(A)		vision $\rightarrow$ content(B)		vision $\rightarrow$ user_text		vision $\rightarrow$ assistant	
		Acc	Acc	Flip rate	Δ Acc	Flip rate	Δ Acc	Flip rate	Δ Acc	Flip rate	Δ Acc	Flip rate	Δ Acc
Qwen2.5-VL	SPATIAL	0.54	0.67	8.19%	+0.06	17.62%	−5.0	15.0%	−2.62	26.38%	−11.75	8.38%	−0.37
	TEMPORAL	0.39	0.43	8.25%	−1.50	14.62%	−6.5	12.12%	−4.5	18.75%	−5.13	6.06%	−0.69
	CAUSAL	0.62	0.81	4.69%	−0.81	14.62%	−7.87	10.88%	−4.87	21.62%	−15.50	4.38%	+0.25
LLaVA-OneVision	SPATIAL	0.55	0.68	8.06%	−0.56	24.62%	−7.87	16.50%	−6.50	23.75%	−10.63	4.5%	−1.38
	TEMPORAL	0.38	0.44	9.25%	−4.87	18.88%	−0.87	10.81%	−0.44	23.81%	−10.56	5.62%	−1.00
	CAUSAL	0.66	0.81	5.44%	−0.31	19.12%	−8.75	17.50%	−8.50	20.06%	−12.56	3.19%	+0.19

Table 3: Final-layer results for the *text-only* baseline and region-based Attention Knockout. We report accuracy for *text-only* and, for each knockout condition, Flip Rate and Δ Accuracy relative to the unmasked *none* condition. Flip Rate is the proportion of changed predictions ($C \rightarrow F$ or $F \rightarrow C$), while $\Delta\text{Acc} = \text{Acc}_{\text{condition}} - \text{Acc}_{\text{none}}$ is the accuracy change in percentage points; negative values indicate that the intervention harms performance.

prompt regions. In early and middle layers, the models exhibit uncertainty with respect to the two answers, as shown by values close to zero, or even biased toward the incorrect foil, as indicated by negative values. This can be observed, for instance, around layers 4–6 for the Causal subset in LLaVA-OneVision and around layer 4 in Qwen2.5-VL. In later layers, from layer 22 onward (with Qwen2.5-VL showing this shift slightly earlier), the trend reverses: both models increasingly assign higher probability to the correct caption, confirming it as the final prediction.

A different pattern emerges for Temporal information. In Figure 2, which averages over all items, Temporal cases appear to differ from Spatial and Causal ones, with the preference for the caption decreasing in later layers and the foil often receiving higher probability. However, the split analysis ($C > F$ vs. $F > C$) in Appendix A.3 shows that this averaged pattern is mainly driven by the incorrect cases ($F > C$): both models initially favor the caption, while the foil becomes more prominent

only in late layers. Thus, the apparent opposite trajectory in the averaged plot only reflects a late-layer shift in the incorrect split, indicating that Temporal caption–foil decisions are determined later in the network and are more fragile than Spatial/Causal. The difference between Causal and Spatial subsets, on the one hand, and the Temporal subset, on the other, also emerges when considering the effect of masking attention to the video across different prompt regions. For Spatial and Causal data, both models clearly show a different behavior with reference to the *none* condition when video attention is blocked within the user-text region. This part structurally includes the question, and the two answer options A and B, as illustrated in Figure 1. In both Spatial and Causal subsets, the strongest effect of video masking is observed when the intervention targets the two answer options, especially in late layers. Conversely, blocking the connection between video tokens and the question does not lead to substantial differences in model behavior. This finding is consistent with Zhang et al. (2025b),

who showed that, in multimodal input processing, visual information is mainly used by text tokens in early and middle layers. These text tokens then propagate the visually enriched information to the assistant tokens for response generation. In other words, the visual signal survives through textual representations, which become the effective carriers of visual information, rather than the video tokens themselves. At the same time, our layer-wise analysis reveals a different temporal profile: the strongest effects of visually enriched textual tokens emerge mainly in mid-to-late layers, while direct *vision* $\rightarrow$ *assistant* masking produces only minor differences with respect to the *none* condition. This suggests that visual information is not routed to the final decision through a direct video-to-assistant pathway, but through textual tokens that remain (visually) informative and exploited until the late stages of processing. In the specific case of our decision-making task setting, these results suggest that cross-modal information is primarily constructed when the model processes the two candidate options. The question itself does not appear to be the critical area of multimodal grounding, and the same happens for the assistant region. Therefore, the video becomes most relevant only when the model evaluates the semantic content of the caption and foil. The decision-level results in Table 3 corroborate the patterns observed in the heatmaps. Masking the answer-option content generally produces the highest flip rates and the largest changes in accuracy. The negative Δ Accuracy values further indicate that suppressing vision-to-option attention disrupts more correct predictions than it recovers, supporting the role of visual information during answer evaluation. Conversely, masking the question and assistant regions produces lower flip rates and smaller changes in accuracy, consistently with their limited effects in the heatmaps.

The Temporal subset again deviates from this pattern. For both models, the effect of masking video attention within the broader user text region, compared to other regions, is much less visible than in the Spatial and Causal subsets. Surprisingly, in LLaVA-OneVision we also observe an effect of the intervention around layers 20-21 when masking the question region, i.e., the instruction *Choose the correct description*. Moreover, this effect is more apparent in middle layers than in late layers, unlike what happens for Spatial and Causal data, where the strongest differences tend to emerge toward the end of the network. Interestingly, for the

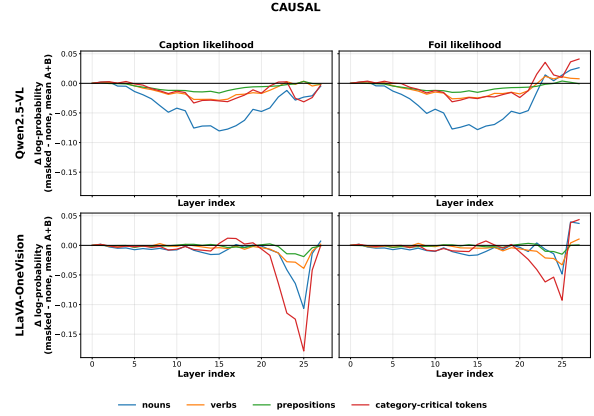


Figure 3: Layer-wise effect of token-level attention knockout on (causal) candidate log-probability.

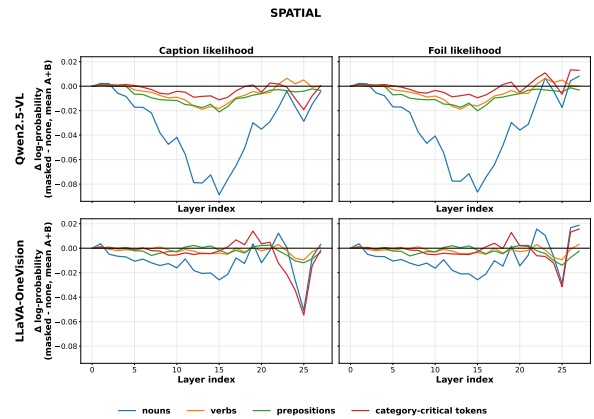


Figure 4: Layer-wise effect of token-level attention knockout on (spatial) candidate log-probability.

same model, at layer 23, the *vision* $\rightarrow$ *assistant* attention knockout has a positive effect: the model assigns higher probability to the caption, in contrast both to the other masking regions and to the *none* condition. This late-layer effect is noteworthy because it occurs at a stage where multimodal information should, in principle, have already been largely processed and propagated through visually enriched text representations. This suggests that, in the Temporal subset, the model may not fully rely on such textual carriers of visual information. Therefore, it appears to still depend on direct access to the video in this final prompt region, possibly in an attempt to recover additional visual evidence needed to solve the task.

4.1 Token-level Analysis

Starting from the results above, we further investigate the part of the user-text region that emerged as the most relevant for constructing multimodal

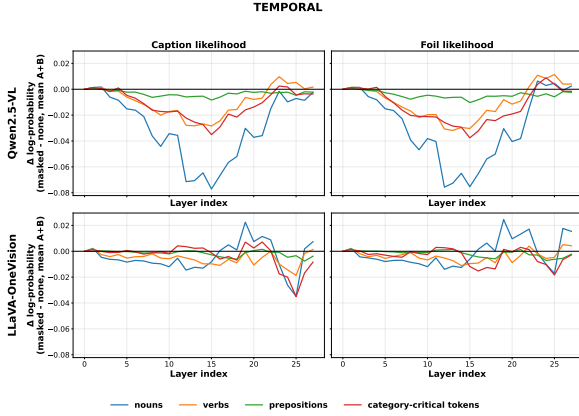


Figure 5: Layer-wise effect of token-level attention knockout on (temporal) candidate log-probability.

information in our setting: the two candidate options A and B. We therefore move to a finer-grained token-level analysis, asking which linguistic elements within the options are most involved in the exchange of information with the visual input.

Figures 3, 4, and 5 show the effect of ablating the *vision* $\rightarrow$ *token(s)* connection within the A and B options. More specifically, each line shows the change in the caption or foil log-probability when *vision-to-text* attention knockout is applied to specific token groups, compared to the *none* condition ($\Delta = \log P_{\text{masked}} - \log P_{\text{none}}$).⁸ A first clear pattern is that the strongest drop in performance across layers is observed for Qwen2.5-VL, where masking access to the video for specific linguistic elements within the options produces a more pronounced effect on both caption and foil plausibility. This suggests that Qwen2.5-VL relies more strongly on visual information while processing the candidate answers, and that grounding is actively performed in this region of the prompt. By contrast, LLaVA-OneVision is less affected by this intervention overall, with the exception of a consistent effect around layer 25 across all subsets.

A second common trend concerns the type of linguistic elements most affected by the *vision* $\rightarrow$ *token(s)* attention knockout. Across models and subsets, nouns are consistently among the most affected categories, both when evaluating the caption and the foil. This suggests that visual information is primarily routed through lexical elements denoting visually grounded entities, which act as semantic

anchors between the visual input and the linguistic representation of the candidate sentence (Shahgiri et al., 2026). At the same time, LLaVA-OneVision shows a more localized pattern, with the largest drops often occurring for CATEGORY-CRITICAL TOKENS, especially for the captions. These tokens - including nouns, verbs and/or prepositions - correspond to the lexical material that encodes the specific spatial, causal, or temporal contrast between the two candidate answers. This suggests that multimodal enrichment is especially concentrated on the task-critical spans that determine which option is visually correct, rather than being distributed more generally across broad linguistic categories. Finally, the Temporal subset again differs from the patterns observed for Spatial and Causal data. First, the use of visual information by linguistic elements within the options appears more irregular, with frequent peaks, drops, and recoveries across layers. Second, compared to the Spatial and Causal subsets, verbs show a stronger dependence on the visual input (Figure 5). This difference aligns with the nature of the Temporal task: while Spatial and Causal items can often be solved by grounding entities or salient visual states, Temporal items require the model to track actions, event dynamics, and their ordering across the video. The stronger involvement of verbs may therefore reflect the need to process visually grounded event information, rather than only visually grounded entities.

As an additional robustness check (i.e., to further assess whether these patterns are tied to the original Italian data), we replicated the analysis on English translations of 200-item samples drawn from each of the original MAIA subsets. The results, reported in Appendix A.2, broadly confirm the trends observed in Italian for both region-based and token-level attention knockout. However, although further analysis is needed, the Temporal subset shows partially different dynamics in English, appearing more stable across layers while still differing from Spatial and Causal cases and suggesting that temporal-relation processing may also be affected by the language of the input.

4.2 The Temporality Issue

These observations raise two related questions: (i) why is Temporal information processed differently, leading to a different and weaker ability to solve the task?; (ii) why does the Temporal subset not behave similarly to the Causal subset, even though

⁸For space reasons, the reported values correspond to average changes in log-probability obtained by collapsing the results across the A and B options. The same general trends are observed in the non-aggregated results.

Temporal marker	<i>N</i>	<i>Text-only baseline</i>		<i>Vision (None)</i>	
		Qwen	LLaVA	Qwen	LLaVA
After	443	0.42	0.36	0.56	0.65
Before	115	0.82	0.75	0.70	0.40
Beginning / Start	42	0.45	0.88	0.60	0.86
End	106	0.59	0.35	0.71	0.62
When	195	0.42	0.27	0.41	0.28
While	523	0.12	0.25	0.13	0.17
Others	176	0.63	0.60	0.66	0.68

Table 4: Last-layer accuracy by temporal marker.

temporal structure is also relevant for identifying cause-effect relations? A possible explanation is that Temporal and Causal items involve different forms of temporal information. In Causal examples, identifying the cause and the effect of an event may often rely on commonsense knowledge (some of which is also available to the language model, and need not rely on the sample-specific visual inputs) and on the recognition of salient visual evidence within individual frames linked to textual representation (e.g., Nouns and Verbs). By contrast, Temporal examples require the model to identify and preserve the sequential order of events, such as whether event A occurs before event B, and in some cases to ground the answer in the absolute temporal position of the video, such as the beginning/ end of the scene. The Spatial subset further supports this hypothesis. Although some Spatial examples contain temporal expressions, such as *at the beginning of the scene*, temporality is not the source of the semantic inconsistency between caption and foil. The model can therefore solve the task by retrieving the relevant visual information from a specific moment or frame of the video. In Temporal examples, instead, the temporal relation itself is the critical information that must be disentangled. We therefore hypothesize that the difficulty observed here is caused by the need to recover and maintain sequential information across the video, a process that appears to be particularly challenging for both models, and an issue that recent work has shown to be challenging also for state of the art VLMs (Song et al., 2025; Wallat et al., 2025).

Temporal markers and linguistic biases While this visual-sequential difficulty likely plays a central role, it may not fully explain the below-chance Temporal performance (see Table 2). To assess whether the linguistic formulation of the candidates also contributes, we identify the temporal markers present in the caption–foil pairs and analyze model

preferences for each marker (Table 4). The results reveal clear construction-specific asymmetries: in the *text-only* condition, both models strongly favor captions containing *Before* over those containing *After* (Qwen: 0.82 vs. 0.42; LLaVA: 0.75 vs. 0.36). Moreover, *When* and *While* remain below chance even in the *none* condition, where full visual input is available (*When*: 0.41/0.28; *While*: 0.13/0.17 for Qwen/LLaVA). With *While* being the largest group, these patterns suggest that linguistic biases may contribute substantially to the aggregate below-chance performance, compounding the difficulty of recovering temporal relations from video.

5 Conclusion

In this work, we investigated the internal cross-modal dynamics through which VLMs use visual evidence in a video-based decision-making setting. Using layer-wise Attention Knockout, we traced *video-to-text* information flow by selectively blocking attention pathways between visual tokens and different textual regions of the prompt, focusing on spatial, causal, and temporal reasoning. The analysis is useful to provide insight into where and how visual information enters the textual representations used for model decisions. Overall, our results show that: (i) visual information mainly contributes when the model processes the candidate answer options, rather than the task instruction or the assistant response, suggesting that candidate options are the primary site of cross-modal grounding and that models rely on these visually enriched representations, rather than on the video directly, when selecting the correct candidate; (ii) the token-level analysis further clarifies which linguistic elements mediate this process, with nouns acting as semantic anchors for visually grounded entities, verbs becoming more prominent in temporal examples, and, in some cases, the affected words being exactly those that distinguish the correct option from the incorrect one; and (iii) spatial and causal relations follow similar processing dynamics, whereas temporal relations exhibit more fragile behavior, suggesting that models struggle to maintain and exploit sequential information across the video, with this difficulty potentially reflecting, in part, linguistic biases associated with expressions used for describing temporal relations between events within a scene.

Limitations

One limitation of this work is that we evaluate only two relatively small VLMs. This restricts the generalizability of our findings across the rapidly evolving landscape of multimodal models. However, our goal was not to provide a leaderboard-style comparison of model performance, but to conduct a controlled diagnostic analysis of how visual information contributes to model predictions. This choice allowed us to examine whether similar information-flow patterns emerge across quite different architectures, while keeping the analysis and visualization of layer-wise interventions manageable within the space constraints of the main paper. A second limitation concerns the scoring formulation used in the mechanistic analyses. While we report both A/B-label scoring and full-sentence scoring as behavioral reference points, we do not systematically analyze the differences between them. Our mechanistic analysis required sentence-level scoring, since we needed access to the token-level structure of each candidate answer and to measure how the plausibility of the full option changes under attention interventions, rather than only whether the model assigns higher probability to the label *A* or *B*. However, the drop observed for Temporal items suggests that the choice of scoring a full sentence may affect model behavior, and should be investigated more directly in future work.

Acknowledgments

This work has been carried out while Davide Testa was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Fondazione Bruno Kessler (FBK). HMW is funded by the Dutch Research Council (NWO) through the AiNed Fellowship Grant NGF.1607.22.002. Bernardo Magnini was supported by the PNRR MUR project [PE0000013-FAIR](#) (Spoke 2). Albert Gatt was supported by the NWO through the AiNed grant NGF.1609.23.020.

References

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. [Hopping too late: Exploring the limitations of large language models on multi-hop queries](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14113–14130, Miami, Florida, USA. Association for Computational Linguistics.

Tyler A. Chang and Benjamin K. Bergen. 2024. [Language model behavior: A comprehensive survey](#). *Computational Linguistics*, 50(1):293–350.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024. [Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16449–16469, Miami, Florida, USA. Association for Computational Linguistics.

Ido Cohen, Daniela Gottesman, Mor Geva, and Raja Giryes. 2025. [Performance gap in entity knowledge extraction across modalities in vision language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29095–29108, Vienna, Austria. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2021/framework/index.html>.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. [Dissecting recall of factual associations in auto-regressive language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. 2025. [What do VLMs NOTICE? A mechanistic interpretability pipeline for Gaussian-noise-free text-image corruption and evaluation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11462–11482, Albuquerque, New Mexico. Association for Computational Linguistics.

Omri Kaduri, Shai Bagon, and Tali Dekel. 2025. What’s in the image? A deep-dive into the vision of vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 14549–14558.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9339–9350.

Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu,

- Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and Erkut Erdem. 2024. [ViLMA: A zero-shot benchmark for linguistic and temporal grounding in video-language models](#). In *The Twelfth International Conference on Learning Representations*.
- Jinyeong Kim, Seil Kang, Jiwoo Park, Junhyeok Kim, and Seong Jae Hwang. 2025. [Interpreting attention heads for image-to-text information flow in large vision-language models](#). In *Mechanistic Interpretability Workshop at NeurIPS 2025*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. [LLaVA-OneVision: Easy visual task transfer](#). Preprint, arXiv:2408.03326.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024b. [MVBench: A comprehensive multi-modal video understanding benchmark](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.
- Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Shishuo Chen, Xu Sun, and Lu Hou. 2024. [TempCompass: Do video LLMs really understand videos?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772, Bangkok, Thailand. Association for Computational Linguistics.
- Olga Loginova and Sofía Ortega Loguinova. 2025. [Deep temporal reasoning in video language models: A cross-linguistic evaluation of action duration and completion through perfect times](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20472–20502, Vienna, Austria. Association for Computational Linguistics.
- Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2018. [Defoiling foiled image captions](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 433–438, New Orleans, Louisiana. Association for Computational Linguistics.
- Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information Processing Systems*.
- Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2025. [Towards interpreting visual information processing in vision-language models](#). In *International Conference on Learning Representations*, volume 2025, pages 57172–57189.
- Nostalgebraist. 2020. [Interpreting GPT: The logit lens](#).
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.
- Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. [Competition of mechanisms: Tracing how language models handle facts and counterfactuals](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, Bangkok, Thailand. Association for Computational Linguistics.
- Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. [VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.
- Letitia Parcalabescu and Anette Frank. 2023. [MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, Toronto, Canada. Association for Computational Linguistics.
- Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. [Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks](#). In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 32–44, Groningen, Netherlands (Online). Association for Computational Linguistics.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). Preprint, arXiv:2412.15115.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- Ahnaf Mozib Samin, M Firoz Ahmed, and Md. Mush-taq Shahriyar Rafee. 2025. [ColorFoil: Investigating color blindness in large vision and language models](#).

- In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 294–300, Albuquerque, USA. Association for Computational Linguistics.
- Haz Sameen Shahgir, Xiaofu Chen, Yu Fu, Erfan Shayegani, Nael Abu-Ghazaleh, Yova Kementchedzhieva, and Yue Dong. 2026. [VLMs need words: Vision language models ignore visual detail in favor of semantic anchors](#). *Preprint*, arXiv:2604.02486.
- Ravi Shekhar, Sandro Pezzelle, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017a. [Vision and language integration: Moving beyond objects](#). In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short papers*.
- Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017b. [FOIL it! Find one mismatch between image and language caption](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.
- Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025. [Can VLMs actually see and read? A survey on modality collapse in vision-language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470, Vienna, Austria. Association for Computational Linguistics.
- Yingjin Song, Yupei Du, Denis Paperno, and Albert Gatt. 2025. [Burn after reading: Do multimodal large language models truly capture order of events in image sequences?](#) In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24316–24342, Vienna, Austria. Association for Computational Linguistics.
- Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. [A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.
- Davide Testa, Giovanni Bonetta, Raffaella Bernardi, Alessandro Bondielli, Alessandro Lenci, Alessio Miaschi, Lucia Passaro, and Bernardo Magnini. 2025a. [All-in-one: Understanding and generation in multimodal reasoning with the MAIA benchmark](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20030–20050, Suzhou, China. Association for Computational Linguistics.
- Davide Testa, Giovanni Bonetta, Raffaella Bernardi, Alessandro Bondielli, Alessandro Lenci, Alessio Miaschi, Lucia C. Passaro, and Bernardo Magnini. 2025b. [MAIA: A benchmark for multimodal AI assessment](#). In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 1121–1134, Cagliari, Italy. CEUR Workshop Proceedings.
- Jonas Wallat, Abdelrahman Abdallah, Adam Jatowt, and Avishek Anand. 2025. [A study into investigating temporal robustness of LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15685–15705, Vienna, Austria. Association for Computational Linguistics.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. [SuperGLUE: A stickier benchmark for general-purpose language understanding systems](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Sudong Wang, Yunjian Zhang, Yao Zhu, Jianing Li, Zizhe Wang, Yanwei Liu, and Xiangyang Ji. 2025. [Towards understanding how knowledge evolves in large vision-language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29858–29868.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Hongxuan Wu, Yukun Zhang, and Xueqing Zhou. 2026. [How vision becomes language: A layer-wise information-theoretic analysis of multimodal reasoning](#). *Preprint*, arXiv:2602.15580.
- Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. 2025a. [From redundancy to relevance: Information flow in LVLMs across reasoning tasks](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2289–2299, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yuanhan Zhang, Bo Li, Haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. [LLaVA-NeXT: A strong zero-shot video understanding model](#).

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

A Additional Experiments and Analysis

A.1 Text-to-vision Masking Direction

In this appendix, we report the complementary analysis conducted with the *text-to-vision* masking direction. While the main paper focuses on the *vision-to-text* setting (Section 3), where textual query tokens are prevented from attending to visual tokens, this additional experiment investigates the opposite direction of information flow. Specifically, we test whether visual-token representations are affected when their attention to textual regions is selectively blocked.

This analysis requires a different prompt configuration from the one used in the main experiments. In the standard Video-Text (VT) prompt, the video input precedes the textual prompt, as shown in Figure 1. Under the causal attention mechanism of decoder-only transformer backbones, each token can only attend to previous tokens in the sequence. As a consequence, if visual tokens appear before textual tokens, they cannot attend to the text, making a *text-to-vision* intervention ill-defined: there would be no text-to-vision attention connections to remove.

To make this intervention meaningful, we therefore adopted an inverted Text-Video (TV) prompt, where the textual prompt is placed before the video input, as shown in Figure 6. In this configuration, visual tokens occur after the textual tokens and can therefore attend to them under the standard causal mask. This allows us to selectively block attention from visual query tokens to specific textual key regions, such as the task instruction, option A, option B, or the full user text. In this way, the *text-to-vision* setup provides a complementary diagnostic view of whether textual information contributes to the construction of visual-token representations.

We report these results only for Qwen2.5-VL. LLaVA-OneVision is not included in this analysis because its chat template enforces a fixed video-text ordering: even when the input is provided in Text-Video order, the internal template reconstructs the prompt in the default Video-Text configuration. As a result, the inverted prompt required for a valid *text-to-vision* intervention cannot be reliably ap-

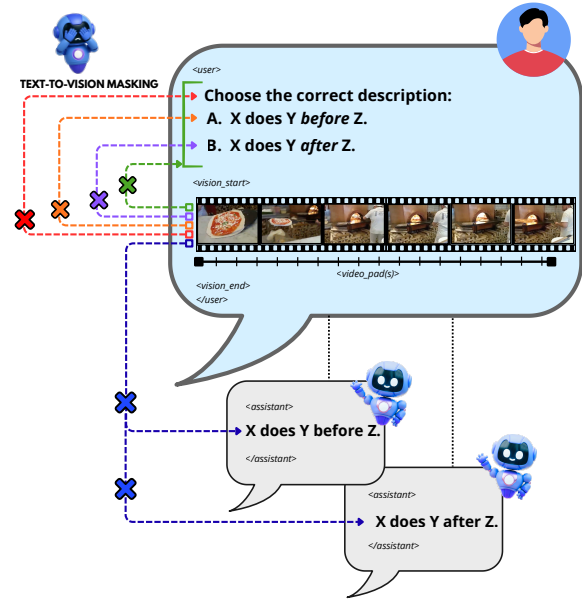


Figure 6: Overview of the Attention Knockout approach with inverted Text-Video prompt used for the *text-to-vision* masking direction.

plied to this model family. Figure 7 shows the results of the *text-to-vision* region-based knockout for Qwen2.5-VL. Despite the non-standard prompt order, the observed trends are broadly consistent with the main *vision-to-text* analysis discussed in Section 4. Blocking attention from visual tokens to textual regions affects the final caption-foil log-probability difference across layers, suggesting that textual regions also contribute to the construction of multimodal representations. However, the magnitude of these effects is generally lower than in the main *vision-to-text* setting.

The *text-to-vision* results show a weaker and less localized pattern than the main *vision-to-text* analysis. The most visible effects emerge when masking the broader *user text* region, whereas the isolated masking of option A and option B produces only marginal changes. This does not mean that the candidate descriptions are irrelevant, since both options are contained within the broader *user text* span. Instead, it suggests that in the inverted Text-Video configuration, the contribution of textual information to visual-token representations is detectable mainly at the level of the full textual context, rather than at the level of the individual answer-specific regions.

Accordingly, we treat the *text-to-vision* analysis as a complementary control rather than as a fully symmetrical counterpart to the main *vision-to-text*

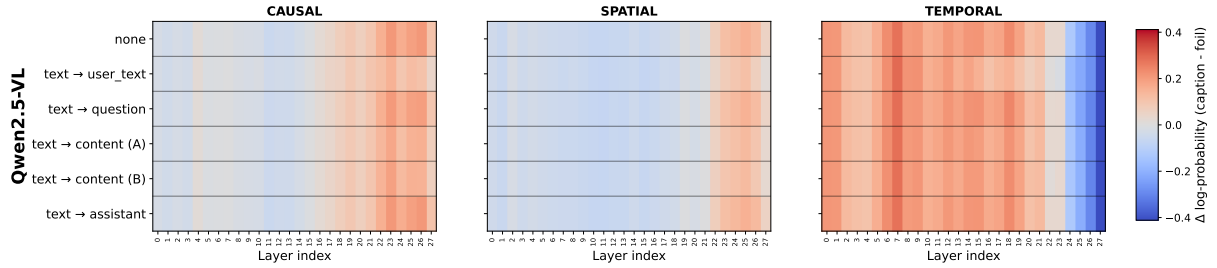


Figure 7: Region-based attention knockout results. Rows indicate the visual prompt region prevented from attending to textual tokens (i.e., *Text* $\rightarrow$ *Region*). Colors show the layer-wise caption-foil log-probability difference.

experiments. The results do not contradict the main findings, but they indicate that the reverse direction produces weaker and less clearly localized effects. This pattern may also be influenced by the non-standard nature of the Text-Video prompt order, which is used here only to make the intervention well-defined under causal attention.

Finally, the Temporal subset shows a behavior that differs from the Spatial and Causal subsets, in line with the main results. This reinforces the interpretation that temporal reasoning poses a qualitatively different challenge for the model, and that its difficulty is not simply an artifact of the standard Video-Text prompt order. Overall, the *text-to-vision* analysis supports the conclusions drawn in the main paper while remaining complementary to them: the main evidence points to a dominant *video-to-textual-representation* pathway, whereas this appendix shows that the reverse conditioning direction produces compatible, but weaker, effects.

A.2 Cross-lingual analysis

To test whether the observed information-flow patterns depend on the original Italian formulation of the MAIA data, we repeated the same attention knockout experiments on English translations of 200-item samples drawn from each of the three subsets. This analysis is particularly relevant because both tested models are likely to have been exposed more extensively to English during training, and may therefore process English prompts more robustly than Italian ones.

Figure 8, reports the region-based results, while Figures 9, 10, and 11 report the token-level results. Overall, the English setting confirms the main trends observed on the Italian data. In the region-based analysis, the strongest effects still emerge when visual access is blocked for textual regions related to the candidate options, supporting the interpretation that these regions act as the main

sites where cross-modal information is integrated. In the token-level analysis, visually grounded lexical categories show patterns comparable to those observed in the original Italian setting, with nouns remaining central for grounding. However, the English Temporal subset exhibits partially different dynamics. In the region-based analysis, the layer-wise behavior appears more stable than in the Italian setting, although the two models still show opposite tendencies in their final predictions. Moreover, token-level interventions suggest a reduced impact of verbs compared to the original Italian data. While these observations are not sufficient to draw definitive conclusions, they indicate that the internal processing of temporal relations may also be influenced by the linguistic formulation of the input, opening interesting directions for future cross-lingual analyses.

A.3 Correct Vs. Incorrect Split Analysis

In the main analysis, we report heatmaps averaged over all items, independently of whether the model assigns higher probability to the caption or to the foil in the final layer. To better understand whether the observed layer-wise patterns are driven by correct or incorrect decisions, we further split the results according to the final caption-foil preference in the *none* condition, where no attention masking is applied. Specifically, we distinguish cases in which the final score satisfies $C > F$, corresponding to a correct preference for the caption, from cases in which $F > C$, where the model assigns higher probability to the foil. Figures 12, 13, and 14 report the corresponding heatmaps for the Causal, Spatial, and Temporal subsets, respectively. Overall, the split analysis shows that the averaged patterns observed in the main paper are stable for Causal and partially for Spatial items.

For Causal items, the distinction between correct and incorrect cases is highly consistent across both

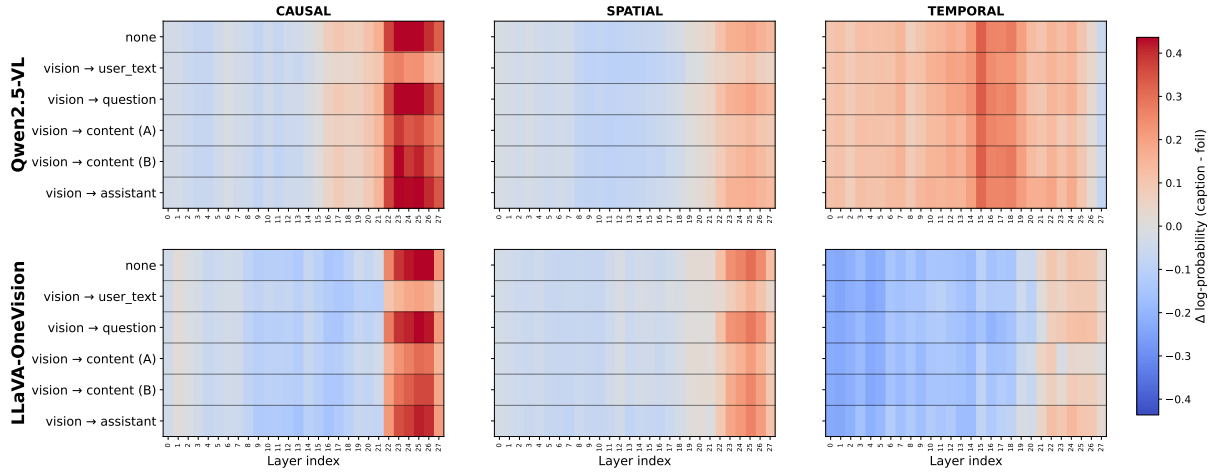


Figure 8: Region-based attention knockout results for English translations of MAIA subdatasets. Rows indicate the textual prompt region prevented from attending to visual tokens (i.e., *Vision* → *Region*). Colors show the layer-wise caption-foil log-probability difference.

Dataset	Option type	Sentence
SPATIAL	Caption	Alla fine della scena, la bambina sta camminando su un tappeto .
	Foil	Alla fine della scena, la bambina sta camminando per strada .
TEMPORAL	Caption	Subito dopo aver tagliato la pizza, la ragazza la porge alla persona che ha di fronte.
	Foil	Prima di tagliare la pizza, la ragazza la porge alla persona che ha di fronte.
CAUSAL	Caption	La ragazza è infastidita a causa di uno scarafaggio .
	Foil	La ragazza è infastidita a causa di un forte rumore .

Table 5: Original Italian examples of caption-foil pairs from the three MAIA subsets.

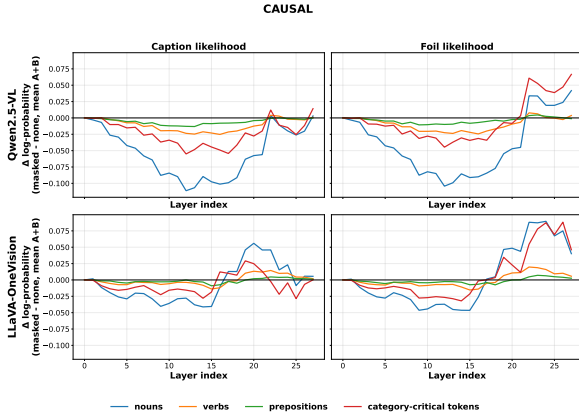


Figure 9: Layer-wise effect of token-level attention knockout on (causal) candidate log-probability. English data

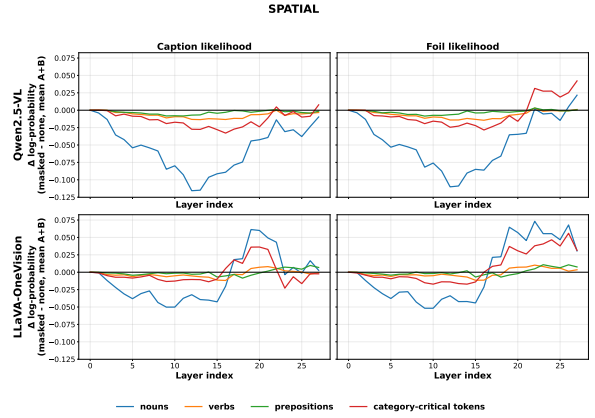


Figure 10: Layer-wise effect of token-level attention knockout on (spatial) candidate log-probability. English data.

models. When the model is correct ($C > F$), the caption becomes increasingly favored in the final layers. Conversely, when the model is incorrect ($F > C$), the foil remains preferred in the final layers, with a relatively stable negative pattern. This suggests that, for Causal items, the final outcome is largely reflected in a coherent layer-wise separation between correct and incorrect trajectories.

Spatial items show a less clear pattern. In the

$C > F$ split, both models converge toward the caption in the final layers, similarly to the Causal subset. However, in the $F > C$ split, the trajectory is less stable. For Qwen2.5-VL, the model appears to favor the caption around layers 22–25 before shifting toward the foil at the final layer. LLaVA-OneVision shows a similar trend, with a gradual movement toward the caption starting around the same layers, followed by a final shift toward the

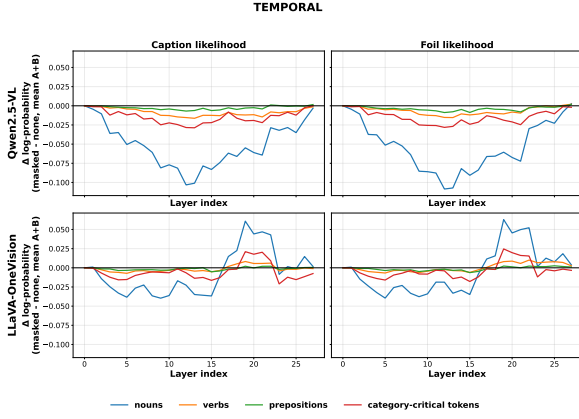


Figure 11: Layer-wise effect of token-level attention knockout on (temporal) candidate log-probability. English data.

foil. Thus, although Spatial items are more similar to Causal than to Temporal in the averaged analysis (Figure 2), the split view reveals that incorrect Spatial decisions can also be resolved very late in the network, suggesting uncertainty between the caption and the foil.

The Temporal subset shows the most distinctive behavior. In the averaged heatmaps reported in the main paper, Temporal items appear to follow an opposite trajectory compared to Spatial and Causal items, with the foil receiving higher probability in the later layers. However, the split analysis in Figure 14 reveals that this pattern is mainly driven by the $F > C$ cases, i.e., by the items for which the model ultimately assigns higher probability to the foil in the *none* condition. When the model is correct ($C > F$), the deltas are positive from the middle layers in both models, and even from earlier layers in Qwen2.5-VL. Although both models show some uncertainty in the penultimate layers, approximately between layers 22 and 25, they ultimately converge toward higher probability for the caption by layer 26. In contrast, when the model is incorrect ($F > C$), the foil begins to receive higher probability only in the late layers: around layer 22 for Qwen2.5-VL, and somewhat earlier for LLaVA-OneVision, which nevertheless shows greater variability between the two alternatives. These results indicate that the apparent Temporal inversion observed in the averaged heatmaps should not be interpreted as a uniform property of all Temporal items. Rather, it reflects a late-layer shift specific to the incorrect subset. In other words, for Temporal items, the choice between caption and foil appears to be determined late in the network: in

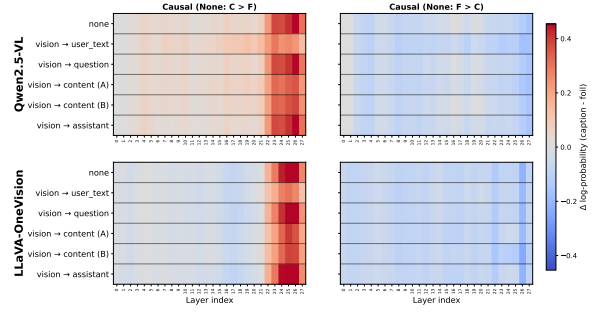


Figure 12: Δ log-probability (caption – foil) across layers and token regions on causal items, split by model outcome: correct predictions ($C > F$, left) and incorrect predictions ($F > C$, right).

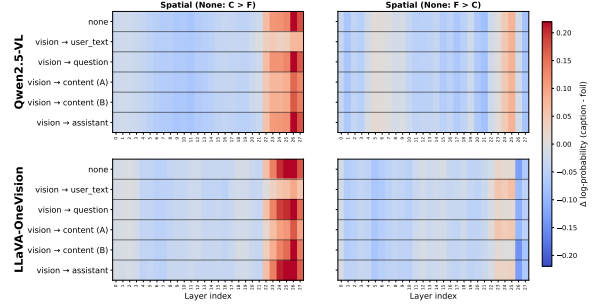


Figure 13: Δ log-probability (caption – foil) across layers and token regions on spatial items, split by model outcome: correct predictions ($C > F$, left) and incorrect predictions ($F > C$, right).

successful cases, the model eventually stabilizes on the caption, whereas in unsuccessful cases the foil becomes dominant only in the final layers. This supports the interpretation that Temporal reasoning is more fragile and depends on the model’s ability to preserve and exploit sequential information across the video until the last stages of computation.

B Additional Materials

This Appendix section presents additional materials not included in the main paper.

Table 5 shows the original Italian caption-foil pairs examples which were presented in English in Section 3.1.

C Experiments

This appendix section will contain additional details on our experimental settings, including a description of the VLMs used. Attention Knockout experiments of subsets of MAIA were conducted using A100 GPUs (40GB). Overall, the total computational budget was on the order of $\sim 2,000$ GPU

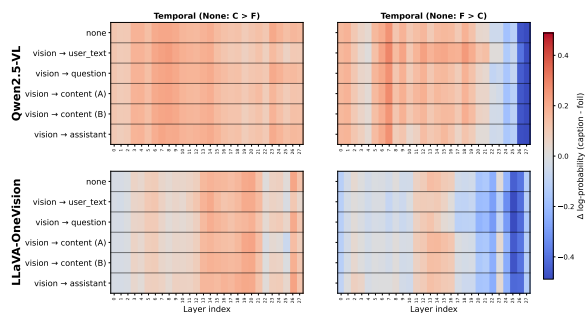


Figure 14: Δ log-probability (caption – foil) across layers and token regions on temporal items, split by model outcome: correct predictions ($C > F$, left) and incorrect predictions ($F > C$, right).

hours.

C.1 Models tested

LLaVa-OneVision. (Li et al., 2024a; Zhang et al., 2024): 7B parameter model that builds on the LLaVA framework with a Qwen2 LLM backbone to serve as a general-purpose vision-language assistant; pre-trained on extensive multimodal data to deliver robust cross-modal reasoning. *Hugging Face* model: lmms-lab/llava-onevision-qwen2-7b-ov.

Qwen2.5-VL. (Qwen et al., 2025): 7B parameter VLM of the Qwen family using the *Qwen2.5* LLM decoder; key enhancements are related to grounding, working with longer videos and capturing events. It was pre-trained on comprehensive visual and textual datasets and fine-tuned for detailed, context-aware responses. *Hugging Face* model: Qwen/Qwen2.5-VL-7B-Instruct.

C.2 Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check and help in coding. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.