MAP THE FLOW: REVEALING HIDDEN PATHWAYS OF INFORMATION IN VIDEOLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

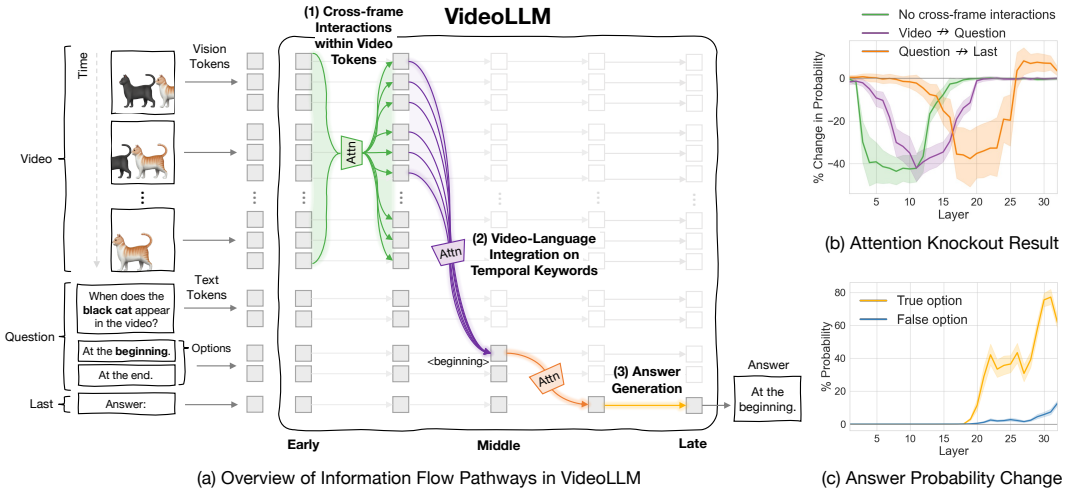
Video Large Language Models (VideoLLMs) extend the capabilities of vision-language models to spatiotemporal inputs, enabling tasks such as video question answering (VideoQA). Despite recent advances in VideoLLMs, their internal mechanisms on *where* and *how* they extract and propagate video and textual information remain less explored. In this study, we investigate the internal information flow of VideoLLMs using mechanistic interpretability techniques. Our analysis reveals consistent patterns across diverse VideoQA tasks: (1) temporal reasoning in VideoLLMs initiates with active cross-frame interactions in early-to-middle layers, (2) followed by progressive video-language integration in middle layers. This is facilitated by alignment between video representations and linguistic embeddings containing temporal concepts. (3) Upon completion of this integration, the model is ready to generate correct answers in middle-to-late layers. (4) Based on our analysis, we show that VideoLLMs can retain their VideoQA performance by selecting these effective information pathways while suppressing substantial amount of attention edges, e.g., 58% in LLaVA-NeXT-7B-Video-FT. These findings provide a blueprint on how VideoLLMs perform temporal reasoning and offer practical insights for improving model interpretability and downstream generalization.

1 INTRODUCTION

Multimodal large language models (MLLMs) (Bai et al., 2023; Chen et al., 2024b;c;a; Liu et al., 2023; 2024a; Wang et al., 2024a) have achieved remarkable success in vision-language tasks by combining powerful auto-regressive language models with vision encoders. Building upon the success of MLLMs, recent efforts have extended these architectures to videos, giving rise to video large language models (VideoLLMs) (Maaz et al., 2024b; Lin et al., 2024; Xu et al., 2024; Wang et al., 2024c) that process spatiotemporal information alongside text. These models have shown promising results on video question answering (VideoQA) tasks, which demand temporal reasoning over multiple frames.

Most prior studies on VideoLLMs have focused on *external* designs of the models, such as scaling video instruction tuning datasets (Li et al., 2024a; Maaz et al., 2024b;a; Li et al., 2024b), key frame selection (Tan et al., 2024; Korbar et al., 2024; Wang et al., 2024b), and compression of input video tokens (Li et al., 2024c; Du et al., 2025; Xu et al., 2024; Zhang et al., 2025b; Jin et al., 2024; Weng et al., 2024; Shen et al., 2024). However, little is known about the *internal* mechanisms of *where* and *how* these models extract relevant temporal information from given videos and propagate it through text tokens to generate final answers. Although recent studies on image-based MLLMs (Neo et al., 2025; Zhang et al., 2024b) have identified their structured behaviors for image-text inputs, it remains unclear whether these findings remain preserved in VideoLLMs and what novel capabilities are acquired through video-text alignment beyond image-text pretraining.

In this study, we aim to provide a *complete blueprint* that reveals the systematic behaviors of VideoLLMs on temporal reasoning tasks, with a focus on the information flow across different layers and modalities. To understand how VideoLLMs generate an *answer* from a given (*video*, *question*) pair, we decompose the temporal reasoning process into several stages and investigate the following key questions: (1) How do VideoLLMs encode spatiotemporal information from the given flattened sequence of video tokens? (2) How are the queried temporal concepts in the question extracted from video tokens and propagated to text tokens? (3) At what stage does the model become ready



(a) Overview of Information Flow Pathways in VideoLLM

(b) Attention Knockout Result

(c) Answer Probability Change

Figure 1: **Summary of our findings on VideoLLMs' information flow. (a)** Temporal reasoning begins with cross-frame interactions within video tokens at early-middle layers [green], followed by video-language integration into temporal keywords in the question [purple]. This information is conveyed to the last token at middle-late layers [orange], where answer generation occurs [yellow]. **(b)** These effective pathways are identified via Attention Knockout, which disconnects attention pairs and tracks the drop in probability of the final answer to quantify their impact. **(c)** Layer-wise answer probability rises immediately after video-language integration, indicating that the model is ready to predict correct answers after the middle layers.

to generate an answer? (4) Can we identify effective information flow pathways sufficient to solve VideoQA tasks?

To answer these questions, we take a mechanistic interpretability perspective (Rai et al., 2024; Nanda et al., 2023; Geva et al., 2023) and reverse-engineer the internal computations of VideoLLMs. Our analysis reveals consistent patterns in how VideoLLMs process video-language information across various VideoQA tasks. Our key findings are summarized as follows:

- **Active temporal interaction within video tokens in early-to-middle layers (§ 3.2):** Temporal reasoning begins by building spatiotemporal representations from video tokens through focused cross-frame attention in early-to-middle layers. Our analysis using Attention Knockout (Geva et al., 2023), which selectively disconnects attention edges to quantify their impact, shows this capability is uniquely acquired through VideoQA instruction tuning from base ImageLLMs.
- **Video-language integration on temporal keywords in middle layers (§ 3.3):** Analyzing semantic concepts in video tokens though Logit Lens (nostalgebraist, 2020) show that temporal concepts are emergent among video tokens in the vocabulary space. Alignment between these representations and temporal keyword embeddings facilitates selective video-language integration over relevant question tokens in early-to-middle layers, which is followed by information converging to the last position token in middle-to-late layers.
- **Answer generation at middle-to-late layers (§ 3.4):** Tracing layer-wise answer probability at the last token reveals that the model is prepared to generate a correct answer immediately once the video-language integration concludes after middle layers.
- **Effective information flow pathways are sufficient for solving VideoQA tasks (§ 3.5):** To validate above findings, we disable all information pathways except those identified as critical. Evaluation on VideoQA benchmarks shows that the models retain performance comparable to baselines, demonstrating that these effective pathways suffice for accurate answer generation.

Our findings provide a first step in understanding the internal mechanisms of VideoLLMs for temporal reasoning. Code and data will be made publicly available.

2 PRELIMINARY

2.1 VIDEO LARGE LANGUAGE MODELS (VIDEOLLMS)

Video and Instruction Tokenization Given an input video $V \in \mathbb{R}^{T\times H\times W\times 3}$, where T and $H \times W$ denote the number of frames and the spatial resolution, we patchify each frame into non-overlapping patches of size $p \times p$, resulting in a total of $N_v = T \times \frac{H}{p} \times \frac{W}{p}$ patches. These patches are processed by a vision encoder $f(\cdot)$ to produce a sequence of video token representations $\{\mathbf{v}_i\}_{i=1}^{N_v}$ where $\mathbf{v}_i \in \mathbb{R}^d$. On the other hand, the instruction texts $\mathbf{t}$ of length N_T is processed using a tokenizer of the language model component in the VideoLLM, which acts as a lookup table of word embeddings, resulting in a sequence of text tokens $\{\mathbf{t}_i\}_{i=1}^{N_T}$. The video and text tokens are then combined as $\{\mathbf{v}_1, ..., \mathbf{v}_{N_v}, \mathbf{t}_1, ..., \mathbf{t}_{N_T}\} \in \mathbb{R}^{(N_v+N_T)\times d}$ and fed into the VideoLLM for multimodal processing.

Multi-head Attention Layers with Causal Modeling Each transformer layer consists of linear projection matrices $\mathbf{W}_q^l, \mathbf{W}_k^l, \mathbf{W}_v^l, \mathbf{W}_o^l \in \mathbb{R}^{d\times d_H}$ with the projection dimension d_H, which are used to derive the query, key, value, and output representations, respectively. Given the input $\mathbf{x}^{l-1}$ from the previous layer, the model computes the query, key, and value by $\mathbf{q}^l = \mathbf{x}^{l-1}\mathbf{W}_q^l$, $\mathbf{k}^l = \mathbf{x}^{l-1}\mathbf{W}_k^l$, $\mathbf{v}^l = \mathbf{x}^{l-1}\mathbf{W}_v^l$. These projections are computed independently for each attention head, evenly splitting the query, key, and value into $\{\mathbf{q}^{l,i}\}_{i=1}^H$, $\{\mathbf{k}^{l,i}\}_{i=1}^H$, and $\{\mathbf{v}^{l,i}\}_{i=1}^H$ for $\mathbf{H}$ heads. Since VideoLLMs adopt causal attention to preserve the autoregressive nature of generation, the attention output for each head is computed using scaled dot-product attention:

$$\text{Attention}(\mathbf{q}^{l,i}, \mathbf{k}^{l,i}, \mathbf{v}^{l,i}) = \text{softmax}\left(\frac{\mathbf{q}^{l,i}(\mathbf{k}^{l,i})^\top}{\sqrt{d_H}} + \mathbf{M}^{l,i}\right)\mathbf{v}^{l,i}, \tag{1}$$

where d denotes the dimensionality of the key vectors and $\mathbf{M}^{l,i}$ is a causal mask. The outputs of all heads are concatenated and projected through $\mathbf{W}_o^l$ to form the final output of the multi-head attention module at layer l.

2.2 ATTENTION KNOCKOUT

Attention Knockout (Geva et al., 2023) selectively disables specific attention connections between tokens during inference. This technique allows us to causally trace the contributions of different modalities or frames. By ablating particular attention paths and measuring the impact on predictions, we can uncover the mechanisms by which information propagates through the model, revealing knowledge localization and the functional roles of individual components.

In practice, to prevent information flow from source tokens (e.g., video inputs or earlier frames) to target tokens (e.g., later frames, question, or answer tokens), we set the value of the attention mask $\mathbf{M}^{l,i}$ at position (s, t) to $-\infty$ in Eqn. 1, where s and t denote the positions of the source and target tokens, respectively. This replacement ensures that the token representations at position t cannot attend to the representations at position s during further attention computations, effectively blocking targeted token interactions in the multi-head attention layers.

In VideoQA, a model generates an answer a from a given video-question pair (v, q), where the question may contain n number of options $o = [o_1; o_2; ...; o_n]$. The model initially predicts the answer a with the highest probability p_{base} at the last token position of the input sequence. We trace the relative change in probability $\%p_{\text{change}} = ((p_{\text{knockout}} - p_{\text{base}})/p_{\text{base}}) \times 100$, where p_{knockout} is the updated probability for the same answer a derived after intervention.

3 INFORMATION FLOW DYNAMICS IN VIDEOLLMS

In this section, we investigate the behaviors of VideoLLMs in VideoQA tasks. Our analyses reveal effective information flow pathways of VideoLLMs and organize into four key findings: (1) temporal interactions occur effectively among video tokens in the early-to-middle layers (§3.2); (2) video information is selectively propagated to their relevant temporal reasoning vocabulary tokens and integrated with textual information (§3.3); and (3) answer generation emerges near the completion of the video-language integration and progresses through the mid-to-late layers (§3.4). We further

Table 1: **Overview of tasks and data for our analyses.** We adopt five tasks from TVBench (Cores et al., 2024), a multiple-choice VideoQA benchmark covering diverse temporal reasoning types.

Task	Reasoning Type	Question Example	Option Example
Action Antonym	Action recognition Sequential ordering	What is the action being performed in the video?	2 temporally opposite actions e.g., Wear jacket.; Take off jacket.
Action Sequence	Action recognition Temporal localization	What did the person do first?	2 actions that actually happened on the video e.g., Put down the blanket.; Took the towel.
Scene Transition	Scene recognition Sequential ordering	What's the right option for how the scenes in the video change?	From {scene1} to {scene2}. From {scene2} to {scene1}.
Moving Direction	Moving object properties	Which direction does the gray cube move in the video?	Down and to the right.; Down and to the left. Up and to the right.; Up and to the left.
Object Count	Moving object properties Temporal localization	How many metal objects are moving when the video begins?	0; 1; 2; 3

discuss the impact of discarding ineffective information flow pathways of VideoLLMs on the VideoQA performance (§3.5). We present the analysis setup (§3.1) and provide detailed examinations in the following sections. Our analyses focus on multiple-choice VideoQA samples for structured evaluation, with open-ended question extensions in the Appendix (§C).

3.1 Experimental Setup

Task and Data We construct our data by selecting five tasks from TVBench (Cores et al., 2024), a multiple-choice VideoQA benchmark designed to evaluate temporal understanding strictly without static bias. As shown in Table 1, our data is constructed with tasks reasoning about diverse attributes under temporally challenging situations. Furthermore, we restrict our analysis to examples where the model outputs the correct answer to ensure meaningful causal tracing. This filtering step focuses our study on samples where the model successfully reasons about the visual and temporal content, eliminating noisy instances due to random guesses or misunderstandings.

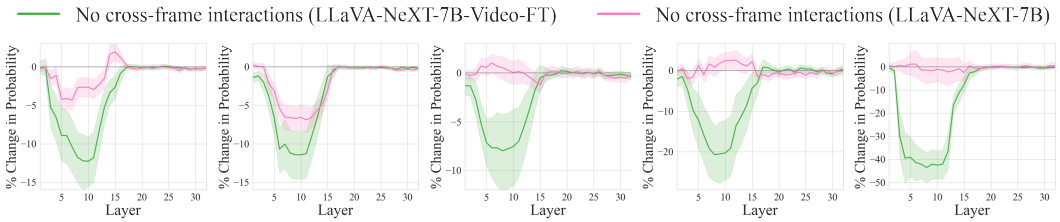
Models We study the behavior of MLLMs that are fine-tuned with video instruction tuning. Specifically, we focus on models originally trained on static image-text data and later fine-tuned on video datasets to analyze unique properties learned during the video instruction tuning procedure. To this end, we fine-tune LLaVA-NeXT-7B (Liu et al., 2024b) with VideoChat2-IT (Li et al., 2024b) for 3 epochs and use this model for the analysis in our main paper. For convenience, we refer to this model as LLaVA-NeXT-7B-Video-FT. For both training and inference, we use 8-frame sampling with 144 tokens per frame. We further extend our analyses on other VideoLLMs with diverse architectures, including LLaVA-NeXT-13B-Video-FT, Mini-InternVL-4B-Video-FT (Gao et al., 2024), and VideoLLaMA3-7B (Zhang et al., 2025a) in the Appendix.

3.2 Temporal Interaction within Video Tokens

To solve VideoQA tasks, VideoLLMs must extract temporally distributed information from videos and generate a correct answer in the last token position. In this subsection, we focus on how VideoLLMs internally encode the spatiotemporal information from the given flattened sequence of video tokens.

Training with VideoQA data boosts cross-frame interactions in the early-to-middle layers. In ImageQA tasks like object identification, answers are often derived by simply pinpointing specific regions at a token level. However, VideoQA tasks present a unique challenge where visual data is spread across a sequence of frames, requiring models to interleave information across frames to capture requisite temporal concepts. To assess how such difference between ImageQA and VideoQA influences the internal mechanisms of trained models, we compare the Attention Knockout results of MLLMs trained solely on image data (i.e., LLaVA-NeXT-7B) and those fine-tuned on video data (i.e., LLaVA-NeXT-7B-Video-FT). Specifically, for each layer l in the MLLM, we block the vision tokens from attending to the tokens in previous frames within a window[1] of $k = 9$ layers around the l^{th} layer, and plot the relative change of prediction probability of answers. Figure 2 shows that blocking the cross-frame interactions in the early-to-middle layers consistently impacts the performance of

[1]We observed that a narrow blocking lets information bypass the knockout while wide windows robustly induce pronounced drops. Thus, we adopt $k = 9$ from (Geva et al., 2023). See §F.6 for details.



(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure 2: **Change in prediction probability when disconnecting cross-frame attention edges.** Blocking cross-frame interactions in early-to-middle layers significantly harms LLaVA-NeXT-7B-Video-FT's prediction, while LLaVA-NeXT-7B remains mostly unaffected.

Table 2: **Impact of cross-frame attention on answer generation.** We block cross-frame attention in the first half of the total layers and measure the resulting accuracy drop. Answers in the third column are taken from open-ended responses from each case. Without cross-frame attention, the model generates incorrect or even opposite answers to the given videos.

Task	Acc Drop	Answer Example
Action Antonym	−24.1%	*Baseline*: The action being performed in the video is to stand up . *Knockout*: The action being performed in the video is to sit on a chair .
Action Sequence	−20.2%	*Baseline*: The action the person is doing first is to open the plastic bag . *Knockout*: The action the person is doing first is to put a bag in the microwave .
Scene Transition	−18.0%	*Baseline*: The scene in the video changes from the bedroom to the street . *Knockout*: The scene in the video changes from the street to a different location .
Moving Direction	−44.8%	*Baseline*: The purple sphere moves to the right in the video. *Knockout*: The purple sphere moves to the left in the video.
Object Count	−60.8%	*Baseline*: The number of moving objects is zero when the video begins. *Knockout*: The number of moving objects is three when the video begins.

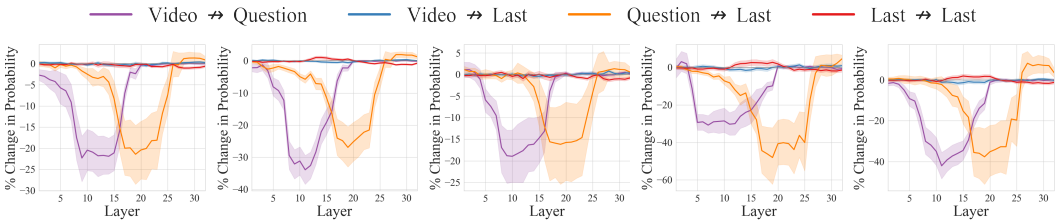
the VideoLLM across all tasks. In contrast, ImageLLM does not exhibit similar correlations in most tasks. This suggests that stronger cross-frame interaction is built during VideoQA training.

How much do temporal interactions affect answer generation? To investigate the extent to which temporal interactions in the early-to-middle layers impact final answer generation, we block cross-frame attention in the first half of the model's layers (i.e., layers 1 to 16), and examine how this intervention influences the baseline's performance. In Table 2, this intervention leads to accuracy drops of at least 18% among samples originally answered correctly with full causal attention. In the third column, where we provide examples of the model's open-ended responses, we observe that the model generates incorrect or even opposite answers to the given videos and instructions across all tasks. These findings suggest that VideoLLMs rely heavily on cross-frame interactions in the early stage to reason about temporal events.

3.3 Video-language Integration on Temporal Reasoning Keywords

Having shown that cross-frame interactions build spatiotemporal representations in early layers, we now examine how this video information integrates with text tokens. As a first step, we trace the overall video-to-language information flow in Figure 3, showing that VideoLLMs follow a structured video → question → last-position token pathway. Building on this understanding, we investigate how VideoLLMs selectively propagate spatiotemporal information through temporal reasoning keywords.

Emergence of temporal concepts in video tokens. Which semantic concepts are extracted from video tokens, and how do they emerge across layers? To answer this, we employ Logit Lens (nostalgebraist, 2020) to trace vocabulary evolution across layers. Specifically, we project hidden states of video tokens at all layers through the language model head to obtain logits, then count the occurrence of spatial and temporal keywords to examine their distribution across layers. We use LLaVA-NeXT-



(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure 3: **Overall cross-modal information flow in VideoLLMs.** We analyze changes in the prediction probability when intervening on attention edges between video, question, and last token (i.e., the starting position for answer generation), following the protocol of (Zhang et al., 2024b). Information from the video tokens is conveyed to the question tokens in the early-to-middle layers, followed by the transfer of information from the question tokens to the last token in the middle-to-late layers. *Source* ↛ *Target* indicates blocking attention edges from source tokens to the target tokens.

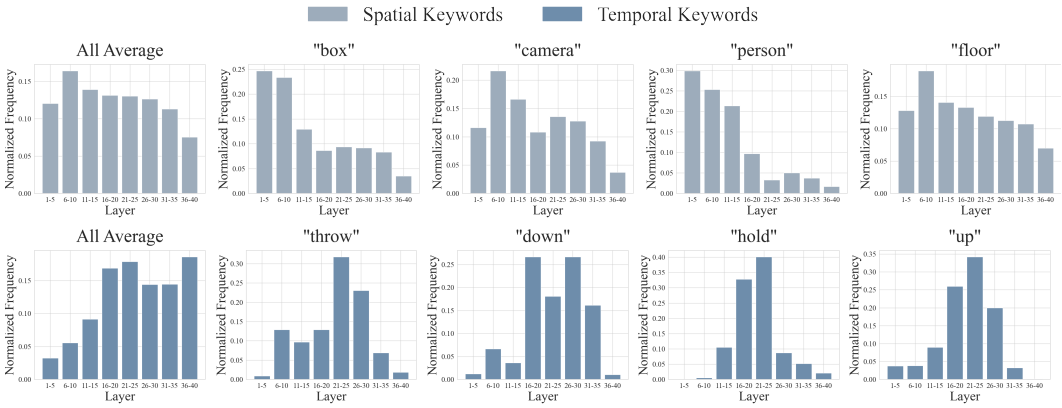


Figure 4: **Normalized frequency of spatial and temporal keywords extracted from video tokens via Logit Lens.** Spatial concepts start to appear in the very early layers, whereas temporal concepts develop later in the middle layers. Full list of keywords is shown in Table D.

13B-Video-FT across Action Sequence videos, with spatial and temporal keywords parsed from the question prompts. Figure 4 shows that both spatial and temporal concepts are captured in video tokens, but with distinct emergence patterns: spatial concepts start to appear in very early layers, while temporal concepts develop in middle layers.

Video-language alignment enables selective spatiotemporal propagation. How are the emergent concepts in videos propagated through text tokens? we analyze the propagation of spatiotemporal information to question tokens and compare it against the propagation of static vision information. We qualitatively show two aspects: (1) temporal visual information is aligned with temporal concept vocabularies, and (2) such alignment emerges specifically through cross-frame interactions.

To this end, we compare video-to-question attention while varying temporal concept words in the question (e.g., "begins", "ends"). As illustrated in Fig. 5 (a), when the temporal interactions are enabled, attention maps highlight the semantically relevant temporal segment of the video corresponding to the temporal meanings of the words "begins" and "ends". This demonstrates that spatiotemporal interactions enable the selective exploitation of semantically crucial information within space and time, allowing question tokens to focus on the most relevant spatiotemporal regions across the entire video. In contrast, when temporal interactions are blocked, the VideoLLMs fail to associate temporal concept vocabulary with relevant video content and instead exhibits positional bias based on positional proximity toward question tokens, as shown in Fig. 5 (b). These findings denote that VideoLLMs implicitly learn to align spatiotemporal representations with linguistic embeddings corresponding to temporal concepts through the video instruction tuning.

Core checkpoint: where and how video-text information is integrated. Interleaving the two previous findings raises a natural question of how video information is propagated to the last position

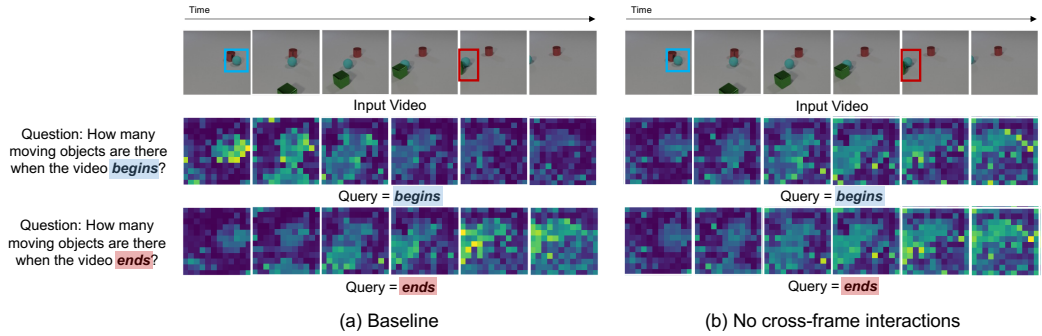


Figure 5: **Visualization of video-to-question attention maps.** Queries are "begins" and "ends" question tokens; keys are video tokens. (a) With spatiotemporal interactions, each question token attends to semantically relevant regions: "begins" focuses on blue sphere at start, "ends" on blue sphere and green square at end. (b) When temporal interactions among video tokens are blocked, video-text alignment fails and text tokens instead attend to positionally proximate regions rather than semantically relevant ones.

via temporal reasoning keywords. However, explicitly linking the pathways among these keywords is challenging since their presence and significance vary across questions. Given that multiple choice options consistently act as temporal keywords, we analyze the information flow through the options. We segment the full question prompt into fine-grained components: the non-option question (e.g., "Question: What is the action being performed in the video?"), the true option (e.g., "(A) Wear jacket"), and the false option (e.g., "(B) Take off jacket"). We then examine where the last token primarily derives information. Figure 6 reveals that information from non-option question tokens does not effectively flow to the last token, whereas the information on true option is propagated to the last token in the middle-to-late layers. This indicates that video-language integration completes at the options tokens.

However, pathways toward options tokens may vary across VideoQA tasks. To validate this hypothesis, we split the pathways toward the options into two different routes (i.e., *video* $\rightarrow$ *true option* and *video* $\rightarrow$ *non-option question* $\rightarrow$ *true option*) and trace how flow patterns differ across questions. Figure 7 shows various task-specific behaviors: in Action Antonym, Action Sequence, and Scene Transition, video information is primarily transferred directly to the true option tokens (see purple line), with relatively minor contribution from non-option question tokens. Conversely, in Moving Direction, video information related to the queried target object is first transferred to non-option question tokens (see red line), after which it flows through the true option to select the correct moving direction (see red dotted line). In Object Count, both direct and indirect flows are observed.

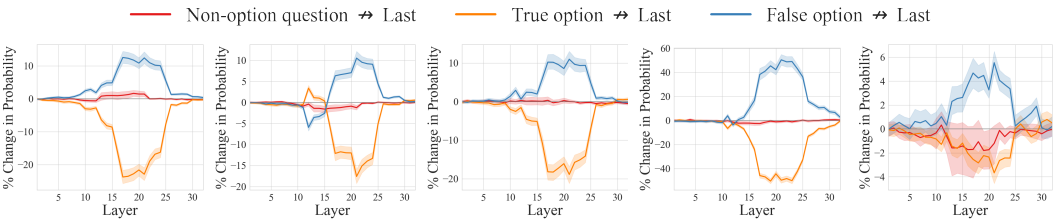
Together, these findings indicate that option tokens serve as the decisive integration point, with the precise pathways varying across task types.

3.4 Inherited Answer Generation Behavior at Middle-to-Late Layers

We further examine the role of layers beyond the information propagation stage. Regarding the prior study (Zhang et al., 2024b) that the last layers of MLLMs primarily focus on linguistic completion, we trace the progression of the answer generation. Specifically, we probe the layer-wise hidden representations at the last token position to follow their probabilities toward the true and false options. Figure 8 shows that the prediction probability for the true option rises sharply and immediately from around the 20th layer, corresponding to the point where video-to-question flow is completed. Furthermore, the probability for the true option increases distinctly, rather than gradually considering false candidates before selection. This suggests that the decision point for a correct answer is heavily dependent on the success of the video-to-language propagation in the middle layers.

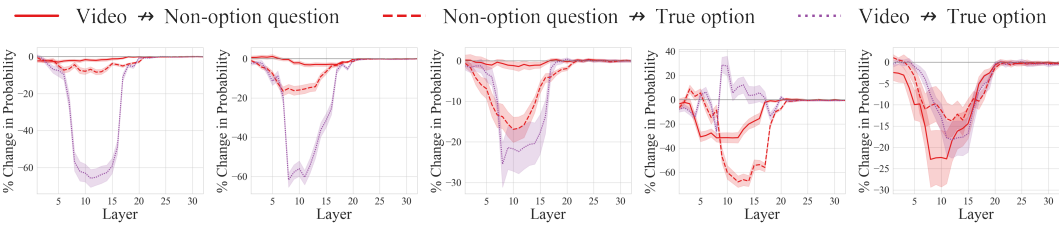
3.5 Dominant Contribution of Effective Information Flow to VideoQA

We have discovered the effective information flow pathways of the temporal reasoning process within VideoLLMs. This raises a natural question regarding the contribution of these pathways to overall VideoQA performance. To assess the impact of effective information pathways, we conduct



(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure 6: **Change in the prediction probability when intervening on attention edges from different parts of the question tokens to the last token.** *Source* ↛ *Target* indicates blocking attention edges from source positions to the target positions. Most of the information flowing to the last token in the middle-to-late layers derives from the true option tokens, rather than broader context in non-option question. Note that the observed probability rise in *false option* ↛ *last* is likely because removing the false option makes the task easier to solve.



(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

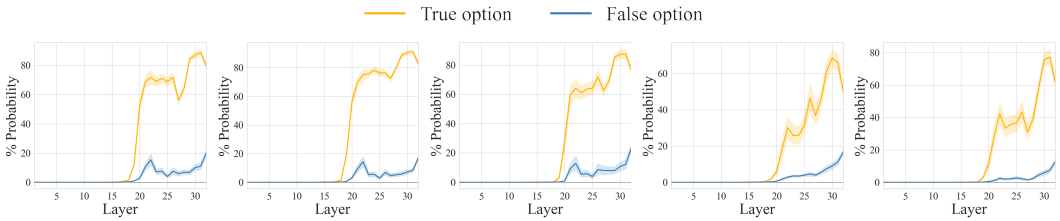
Figure 7: **Change in the prediction probability when intervening on attention edges to the true option position.** *Source* ↛ *Target* indicates blocking attention edges from source positions to the target positions. Information from video tokens consistently converges to the true option tokens in early-to-middle layers, while routing to non-option question tokens varies depending on the task.

a quantitative analysis by evaluating VideoLLMs on VideoQA benchmarks after retaining only the identified effective token interactions, while disabling all others [2]. Table 3 summarizes the performance on TVBench (Cores et al., 2024) and TOMATO (Shangguan et al., 2024) benchmarks. While attention restricted to effective pathways suppresses a substantial portion of attention edges (e.g., using only 42% in LLaVA-NeXT-7B-Video-FT), it results in only marginal accuracy decreases across both benchmarks. However, randomly blocking the same proportion of attention edges causes a significant performance drop. These results underscore the validity of our analysis on the effective information flow pathways.

4 Related Work

Video Large Language Models (VideoLLMs) Research on video understanding has increasingly focused on leveraging image-level pre-trained MLLMs by fine-tuning them for video-language tasks such as VideoQA, video captioning, and video conversation. To improve the temporal reasoning ability of VideoLLMs, much of the existing studies have concentrated on the *external* aspect of the VideoLLM backbone itself, such as scaling video instruction tuning datasets (Li et al., 2024a; Maaz et al., 2024b;a; Li et al., 2024b), selecting key frames (Tan et al., 2024; Korbar et al., 2024; Wang et al., 2024b), retaining memory banks (Song et al., 2024; He et al., 2024), and compressing the input video tokens (Li et al., 2024c; Du et al., 2025; Xu et al., 2024; Zhang et al., 2025b; Jin et al., 2024; Weng et al., 2024; Shen et al., 2024). In contrast, our study focuses on the *inner* mechanism of VideoLLMs and investigates how temporal reasoning occurs.

[2]The layer ranges for effective pathways are determined from empirical patterns in Sections 3.2–3.4, selecting those with significant probability drops when blocked (See Table E for details). We enable *cross-frame interactions* in early-to-middle layers (e.g., L6-15), *video* → *question* in early-to-middle layers (e.g., L6-20), and *question* → *last* in middle-to-late layers (e.g., L16-25). *Video* → *last* and *last* → *last* connections are disabled across all layers. Flows to *video* and *question* tokens are blocked at late layers, as these are no longer used.



(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure 8: **Layerwise prediction probability for true and false options in the last token position.** The probability for the true option starts to rise immediately after the middle layers.

Table 3: **Impact of effective flow pathways on performance in TVBench and TOMATO.** The number of attention edges is the count of valid (query, key) pairs over all attention layers. When we enable attention only for effective pathways, VideoQA performance is retained across diverse tasks and models even though suppressing a substantial portion of attention edges.

Model	# Video Tokens	Attention Type	# Attention Edges	TVBench	TOMATO
LLaVA-NeXT-7B-Video-FT	8×12×12	Full causal attention	25.7M (100%)	51.5	30.2
		Effective pathways	10.8M (42%)	51.2	29.2
		Random blocking	10.8M (42%)	40.1	23.1
LLaVA-NeXT-13B-Video-FT	8×12×12	Full causal attention	32.2M (100%)	55.1	27.2
		Effective pathways	14.3M (37%)	54.6	27.4
		Random blocking	14.3M (37%)	41.5	23.8
Mini-InternVL-4B-Video-FT	8×16×16	Full causal attention	74.6M (100%)	56.0	32.2
		Effective pathways	29.6M (40%)	56.0	31.2
		Random blocking	29.6M (40%)	41.0	25.9
VideoLLaMA3-7B	8×12×12	Full causal attention	19.9M (100%)	55.2	28.0
		Effective pathways	11.4M (58%)	57.2	28.7
		Random blocking	11.4M (58%)	22.2	13.9

Mechanistic Interpretability of Multimodal Models Mechanistic interpretability (Rai et al., 2024; Nanda et al., 2023; Geva et al., 2023) is an emerging area that seeks to understand neural networks by reverse-engineering their internal computations. Recently, several studies (Palit et al., 2023; Yu & Ananiadou, 2024; Cohen et al., 2024; Basu et al., 2024; Neo et al., 2025; Zhang et al., 2024b) have applied mechanistic interpretability techniques to MLLMs to explain their inner mechanisms. (Basu et al., 2024) focused on how information is stored and retrieved from the model parameters of MLLMs. On the other hand, (Neo et al., 2025) examined the object identification task and explored how the object-level information flows and emerges. Most recently, (Zhang et al., 2024b) systematically studied cross-modal information flow in MLLMs, revealing that many models exhibit a single-stream information transfer pattern from vision to language. Inspired by these methodologies, we extend attention-based knockout and layerwise logit probing techniques to VideoLLMs to understand temporal reasoning mechanisms.

VideoQA Benchmarks VideoQA benchmarks have advanced to validate temporal understanding for VideoLLMs. VCGBench (Maaz et al., 2024b) and Video-MME (Fu et al., 2024) provide wide coverage of aspects that appear within videos, establishing strong general-purpose baselines. TVBench (Cores et al., 2024) is designed to require true temporal reasoning and to discourage single frame shortcuts. TempCompass (Liu et al., 2024c) evaluates multiple temporal aspects through varied VideoQA formats to measure time aware perception. Vinoground (Zhang et al., 2024a) uses counterfactual short video pairs to assess fine grained temporal distinctions beyond static cues. TemporalBench (Cai et al., 2024) focuses on detailed temporal dynamics such as event order frequency and change patterns. TOMATO (Shangguan et al., 2024) introduces expert written tasks that force reasoning about event evolution across frames. MotionBench (Hong et al., 2025) targets fine grained motion comprehension using motion centric questions LongVideoBench (Wu et al., 2024) evaluates long form interleaved video language understanding where models must retrieve

relevant moments and maintain evidence over extended contexts. Among the benchmarks, we mainly analyze on TVBench, TOMATO, LongVideoBench, and Video-MME.

5 Conclusion

We have conducted a comprehensive mechanistic analysis to understand where and how VideLLMs extract and propagate video and text information for VideoQA tasks. Our study reveals that temporal reasoning initiates from the encoding of spatiotemporal representations among video tokens during the early-to-middle layers through active cross-frame interactions. This processed information is then transferred to semantically aligned temporal concept tokens in the question. Furthermore, we have observed that critical information needed to determine the correct answers is conveyed to the last position tokens in the middle-to-late layers, eventually contributing to answer generation. These effective pathways alone prove sufficient for solving VideoQA tasks. Our findings provide practical insights into the internal working mechanisms of VideoLLMs and open new research directions for their interpretability and generalization.

Reproducibility Statement. To ensure reproducibility, we provide comprehensive implementation details in Appendix H, including model architectures, training and inference strategies, and experimental configurations. Our Attention Knockout and Logit Lens analyses are implemented based on the public codebase from (Geva et al., 2023). All experiments use publicly available datasets (TVBench (Cores et al., 2024) and TOMATO (Shangguan et al., 2024)) and models (LLaVA-NeXT (Liu et al., 2024b), Mini-InternVL (Gao et al., 2024), and VideoLLaMA3 (Zhang et al., 2025a)). We will release our complete code and models upon publication.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 28

Sangmin Bae, Jongwoo Ko, Hwanjun Song, and Se-Young Yun. Fast and robust early-exiting framework for autoregressive language models with synchronized parallel decoding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5910–5924, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.362. URL https://aclanthology.org/2023.emnlp-main.362/. 26

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 27

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *NeurIPS*, 2024. 9

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Towards fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024. 9

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a. 1

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 2024b. 1

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024c. 1

Ido Cohen, Daniela Gottesman, Mor Geva, and Raja Giryes. Performance gap in entity knowledge extraction across modalities in vision language models. *arXiv preprint arXiv:2412.14133*, 2024. 9

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in time: A new temporal benchmark for videollms. *arXiv preprint arXiv:2410.07752*, 2024. 4, 8, 9, 10

Yifan Du, Yuqi Huo, Kun Zhou, Zijia Zhao, Haoyu Lu, Han Huang, Xin Zhao, Bingning Wang, weipeng chen, and Ji-Rong Wen. Exploring the design space of visual context representation in video MLLMs. In *ICLR*, 2025. 1, 8

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJg7KhVKPH. 26

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 9, 24

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2024. 4, 10, 15, 28

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *EMNLP*, 2023. 2, 3, 4, 9, 10, 26, 28

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 27

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 27

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *CVPR*, 2024. 8

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8450–8460, June 2025. 9

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*, 2024. 1, 8

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 27

Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. Text-conditioned resampler for long form video understanding. In *ECCV*, 2024. 1, 8

Dongxu Li, Xiaohan Wang, et al. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2403.08173*, 2024a. 1, 8, 27

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024b. 1, 4, 8, 27

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, 2024c. 1, 8

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *CVPR*, 2016. 27

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. 2024. 1

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *ICLR*, 2023. 1

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024a. 1

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024b. 4, 10, 15, 27, 28

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.517. URL https://aclanthology.org/2024.findings-acl.517/. 9

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024a. 1, 8

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2024b. 1, 8, 9, 20, 27

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023. 2, 9

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *ICLR*, 2025. 1, 9

nostalgebraist. Interpreting GPT: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, August 2020. Accessed: 2025-02-22. 2, 5, 28

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In *ICCV*, 2023. 9

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, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115. 28

Alec Radford, Jong Wook Kim, Christopher Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pam Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 27, 28

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024. 2, 9

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. NeurIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088. 26

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. *arXiv preprint arXiv:2410.23266*, 2024. 8, 9, 10

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 1, 8

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*, 2024. 8

Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer, Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. In *CVPR*, 2024. 1, 8

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a. 1

Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Lou, Ming C Lin, and Shan Yang. Vila: Efficient video-language alignment for video question answering. In *ECCV*, 2024b. 1, 8

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV)*, 2024c. 1

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *ECCV*, 2024. 1, 8

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL https://arxiv.org/abs/2407.15754. 9, 23, 24

Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Hong Zhou, Mike Zheng Shou, and Xiang Bai. A large cross-modal video retrieval dataset with reading comprehension. *Pattern Recognition*, 2025. 27

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 27

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 1, 8

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *ICLR*, 2019. 27

Zeping Yu and Sophia Ananiadou. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering. *arXiv preprint arXiv:2411.10950*, 2024. 9

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023. 28

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a. 4, 10, 28

Jianrui Zhang, Cai Mu, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv*, 2024a. URL https://arxiv.org/abs/2410.02763. 9

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*, 2025b. 1, 8

Tianxing Zhang, Haoran Shi, Xiang Lisa Li, Zhou Yu, Chelsea Finn, and Tatsunori Hashimoto. Cross-modal information flow in multimodal large language models. *arXiv preprint arXiv:2411.18620*, 2024b. 1, 6, 7, 9

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 2023. 27, 28

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 27

APPENDIX

A SCALABILITY OF OUR FINDINGS

To verify the scalability of our findings, we investigate a larger-scale VideoLLM. Specifically, we fine-tune LLaVA-Next-13B (Liu et al., 2024b) on video instruction tuning datasets, resulting in **LLaVA-NeXT-13B-Video-FT**. In this section, we analyze the information flow in the 13B model and show that the effective information flow pathways identified in the smaller model remain consistent at scale.

Active temporal interaction within video tokens. As shown in Fig. A, blocking cross-frame interactions in the early-to-middle layers consistently degrades the performance of the VideoLLM across all tasks (green), different from its ImageLLM baseline (pink). Notably, LLaVA-NeXT-13B-Video-FT exhibits similar trends to LLaVA-NeXT-7B-Video-FT, highlighting that our findings on active temporal interactions among video tokens generalize across model scales.

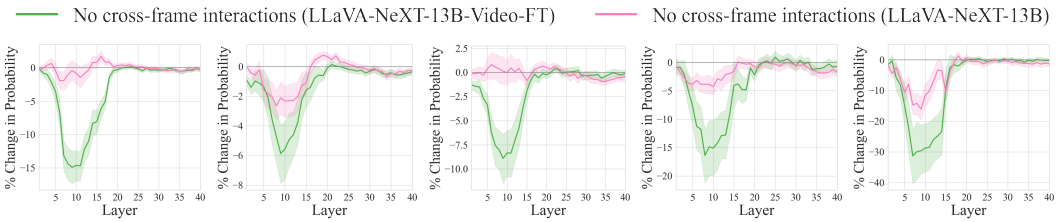
Video-language integration on temporal keywords. We further analyze the integration mechanism of video and text information in the larger-scale VideoLLM. Fig. B illustrates the impact of blocking attentions between video, question, and last position tokens. Consistent with the trends observed in the smaller-scale model, video information is not directly transmitted to the last token but is instead routed through the question tokens. To further trace how video and language information reaches the option tokens, we analyze the attention flow toward the answer options. As in Fig. D, information from video tokens flows to the true option tokens either directly or indirectly via non-option question tokens. The balance between these pathways varies across tasks, suggesting task-specific patterns in cross-modal integration. These consistent results showcase that our findings on video-language integration hold across scales.

Answer generation. Fig. E illustrates that generation occurs only after video and language information has been fused, primarily through the option tokens. The main difference from the smaller-scale VideoLLM lies in the specific layer range where this transition from integration to generation occurs, while the overall pattern remains consistent.

B GENERALIZATION OF OUR FINDINGS ACROSS VIDEOLLMS

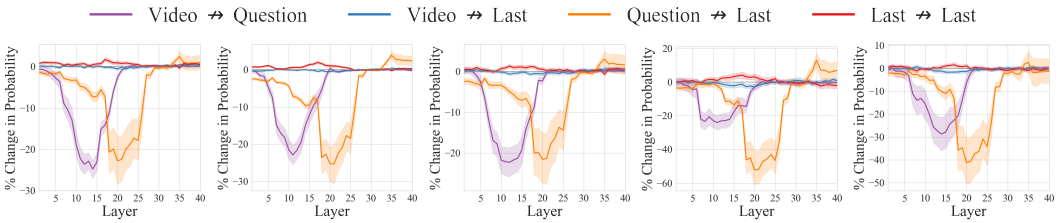
In this section, we validate the generalization of our findings to other VideoLLMs. Specifically, we employ **VideoLLaMA3-7B** and **Mini-InternVL-4B-Video-FT**, a VideoLLM obtained by fine-tuning Mini-InternVL-4B (Gao et al., 2024) on video instruction tuning datasets. We verify VideoLLaMA3-7B and Mini-InternVL-4B-Video-FT to validate whether our findings on effective information flow generalize across different VideoLLMs.

Active temporal interaction within video tokens. Fig. F and Fig. K show that blocking the attention between video tokens leads to a greater performance drop across all tasks, indicating that Mini-InternVL-4B-Video-FT and VideoLLaMA3-7B also learn stronger temporal interaction through VideoQA training than its ImageLLM counterpart. Moreover, this behavior appears in the early-to-middle layers, similar to the behaviors of the LLaVA-NeXT series.



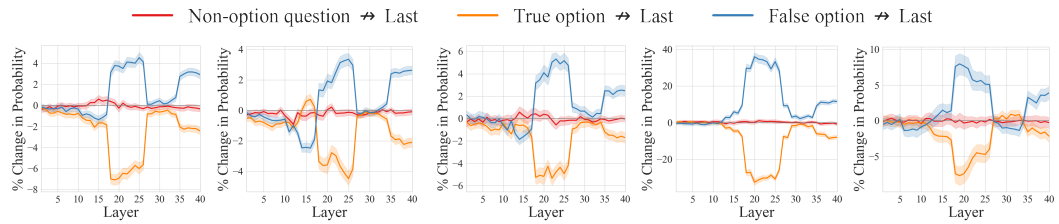
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure A: **Change in prediction probability when disconnecting cross-frame attention edges in LLaVA-NeXT-13B-Video-FT and LLaVA-NeXT-13B.**



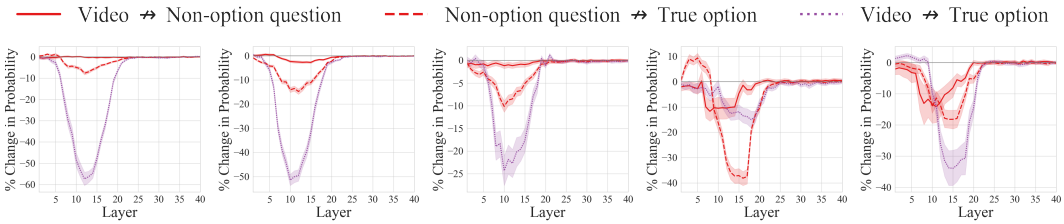
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure B: **Change in the prediction probability of LLaVA-NeXT-13B-Video-FT when intervening on attention edges between video, question, and last token.**



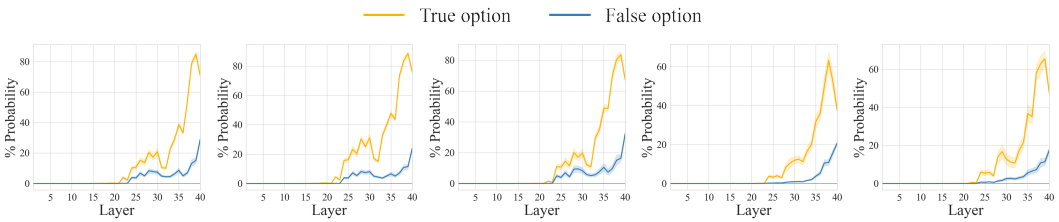
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure C: **Change in the prediction probability of LLaVA-NeXT-13B-Video-FT when intervening on attention edges from different parts of the question tokens to the last token.**



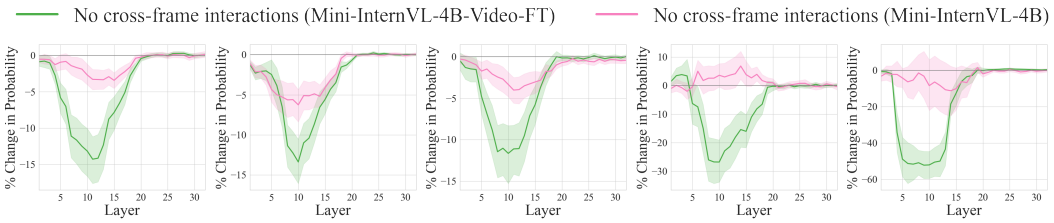
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure D: **Change in the prediction probability of LLaVA-NeXT-13B-Video-FT when intervening on attention edges to the true option position.**



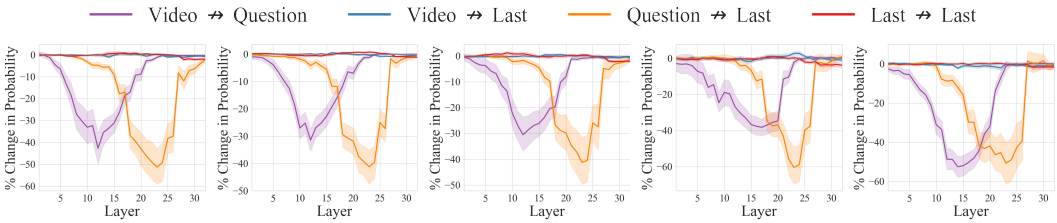
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure E: **Layerwise prediction probability of LLaVA-NeXT-13B-Video-FT for true and false options in the last token position.**



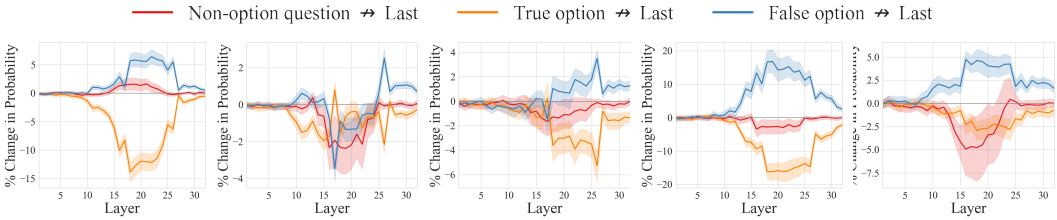
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure F: **Change in prediction probability when disconnecting cross-frame attention edges in Mini-InternVL-4B-Video-FT and Mini-InternVL-4B.**



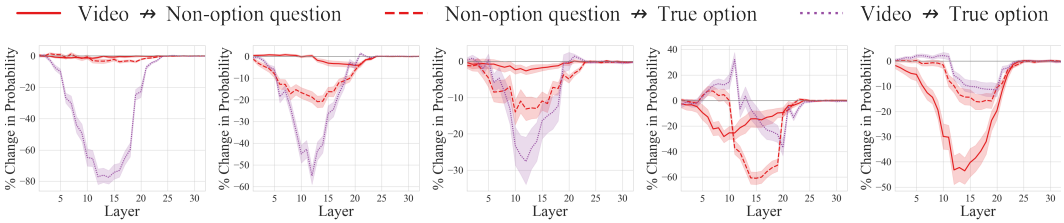
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure G: **Change in the prediction probability of Mini-InternVL-4B-Video-FT when intervening on attention edges between video, question, and last token.**



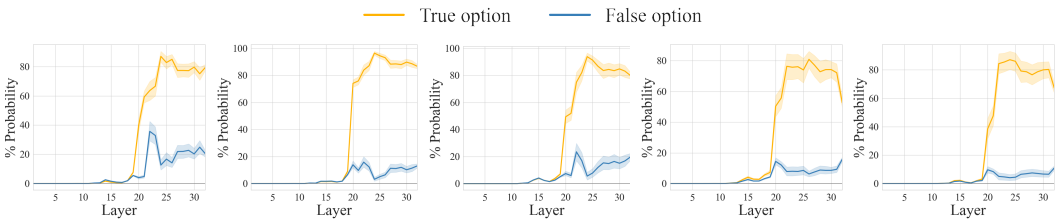
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure H: **Change in the prediction probability of Mini-InternVL-4B-Video-FT when intervening on attention edges from different parts of the question tokens to the last token.**



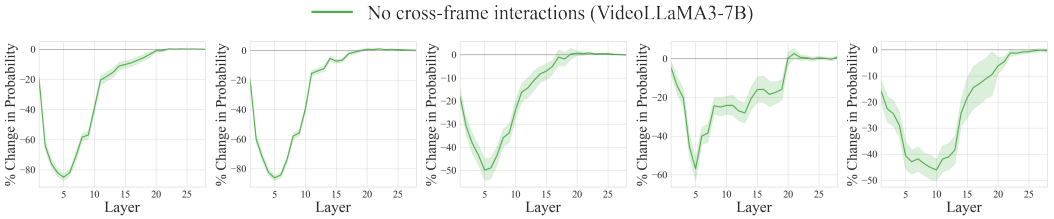
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure I: **Change in the prediction probability of Mini-InternVL-4B-Video-FT when intervening on attention edges to the true option position.**



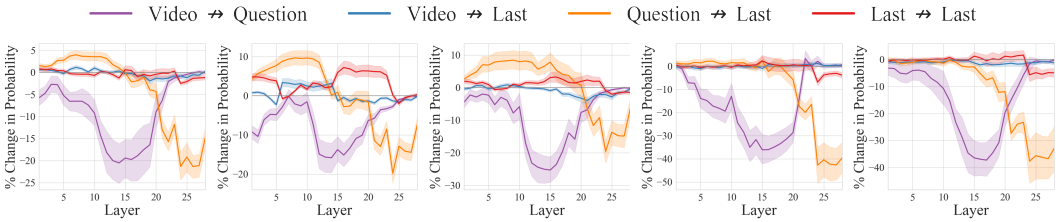
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure J: **Layerwise prediction probability of Mini-InternVL-4B-Video-FT for true and false options in the last token position.**



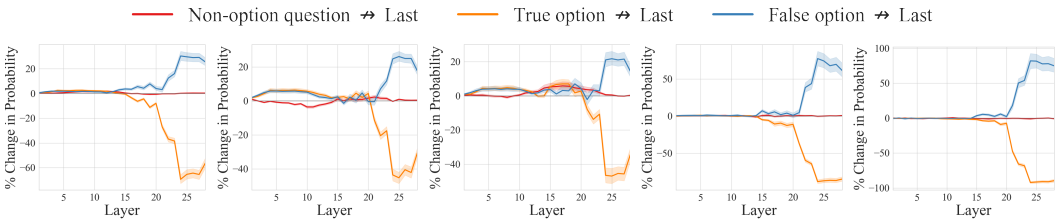
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure K: **Change in prediction probability when disconnecting cross-frame attention edges in VideoLLaMA3-7B.**



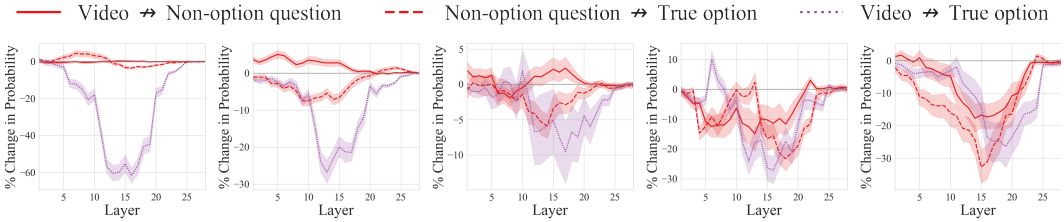
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure L: **Change in the prediction probability of VideoLLaMA3-7B when intervening on attention edges between video, question, and last token.**



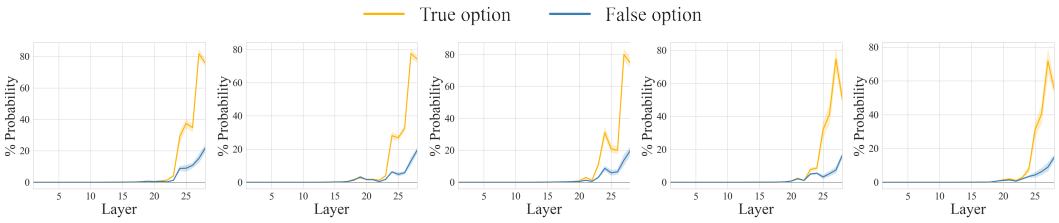
(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure M: **Change in the prediction probability of VideoLLaMA3-7B when intervening on attention edges from different parts of the question tokens to the last token.**



(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure N: **Change in the prediction probability of VideoLLaMA3-7B when intervening on attention edges to the true option position.**



(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure O: **Layerwise prediction probability of VideoLLaMA3-7B for true and false options in the last token position.**

Video-language integration on temporal keywords. We analyze how Mini-InternVL-4B-Video-FT and VideoLLaMA3-7B integrate video and language information in response to temporally grounded questions. As shown in Fig. G and Fig. L, video information is transmitted to question tokens in the early-to-middle layers, and only later transferred to the last position for answer generation. Fig. I and Fig. N further show that the video-language information is also gathered in the true option tokens, and the pathways toward the option tokens vary across the VideoQA tasks. These results demonstrate that our findings on video-language integration generally hold across various VideoLLMs.

Answer generation. Fig. J and Fig. O show that although Mini-InternVL-4B-Video-FT tends to exhibit a sharp rise in generation probability across various VideoQA tasks, its overall behavior remains consistent with that of LLaVA-based VideoLLMs, where the probability begins to increase near the end of the video-language integration process.

C Analysis on Open-ended VideoQA

In open-ended video question-answer tasks, the input prompt does not include keywords related to the ground truth answers. Thus, the information flow to the final token may differ because the model generates its answer using new vocabulary rather than selecting from given multiple-choice options. In this section, we investigate whether the difference in prompt format affects the information flow.

C.1 Single Token Generation

Open-ended analysis setup. The input prompt formats in TVBench are modified by removing the options and adopting a sentence completion style. For example: *"USER: <video> USER: Question: Which direction does the gray cube move in the video? ASSISTANT: The gray cube moves to the ___."* To avoid ambiguity, we select tasks where the first tokenized sub-word of the model's possible answer is relatively constrained, such as Action Antonym, Moving Direction, and Object Count. We adopt LLaVA-NeXT-7B-Video-FT and LLaVA-NeXT-7B for this analysis.

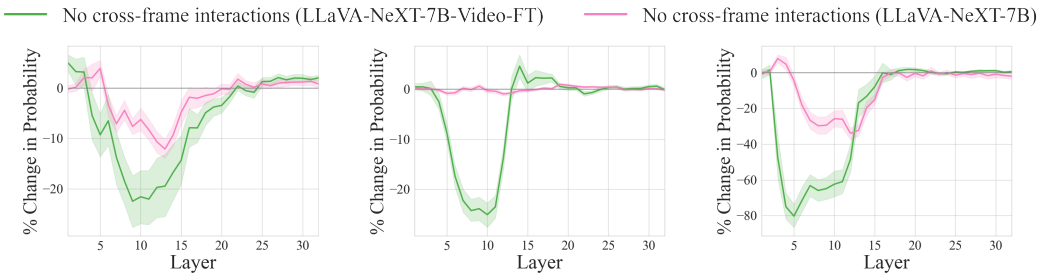
Active temporal interaction in open-ended VideoQA. We examine the impact of temporal interaction within video tokens in open-ended VideoQA. Using the same attention-blocking setup as in previous evaluations, we observed that disabling cross-frame interactions in early-to-middle layers leads to a significant decrease in answer probability even in the open-ended questions answering problems, as shown in Fig. P. These results suggest that active temporal interaction is a general mechanism leveraged by VideoLLMs, regardless of the format of the question answering problems.

Video-language integration. Unlike multiple-choice question answering, open-ended question answering does not explicitly provide candidate answers. Consequently, the input text lacks explicit temporal reasoning keywords that directly reference temporal information in the video. We hypothesize that, in the absence of true option tokens, the last token itself becomes the core checkpoint for video-language integration. To this end, we examine the information flow from the video and question tokens to the last token. Fig. Q shows two different routes: *video* $\rightarrow$ *last* (purple lines) and *video* $\rightarrow$ *non-option question* $\rightarrow$ *last* (red lines). While video information may first pass through question tokens, the final integration converges at the last token. This behavior aligns with what we have observed in multiple-choice tasks, where video and language information merge at a core checkpoint, although this checkpoint shifts to the last token position in the open-ended case.

Answer generation. We observe that answer generation occurs at middle-to-late layers also for open-ended problems. As shown in Fig. R, the prediction probability rises from around layer 20, similar to the trend observed in multi-choice question answering problems.

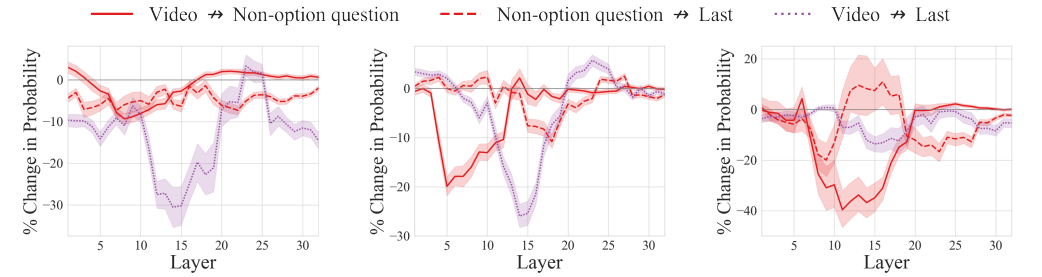
C.2 Multiple Token Generation

Building on our previous finding that the last token serves as the key checkpoint for video-language integration in open-ended generation, we further examine how checkpoints emerge within responses and how information propagates through these checkpoints as the model continues to generate multiple tokens.



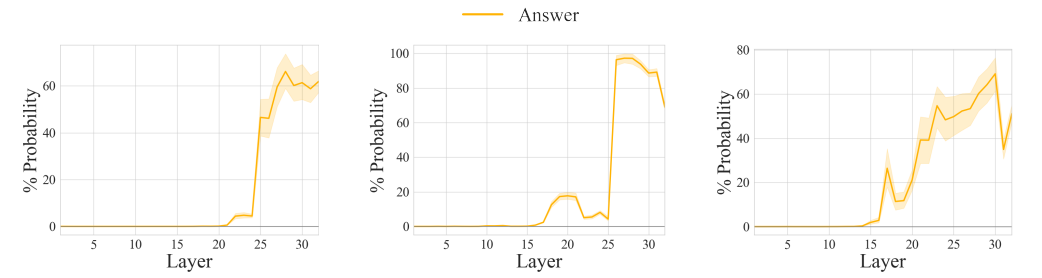
(a) Action Antonym (b) Moving Direction (c) Object Count

Figure P: **Change in prediction probability in open-ended QA format when disconnecting cross-frame attention edges.** LLaVA-NeXT-7B-Video-FT shows a stronger correlation with cross-frame interactions and the final answer probability compared to LLaVA-NeXT-7B.



(a) Action Antonym (b) Moving Direction (c) Object Count

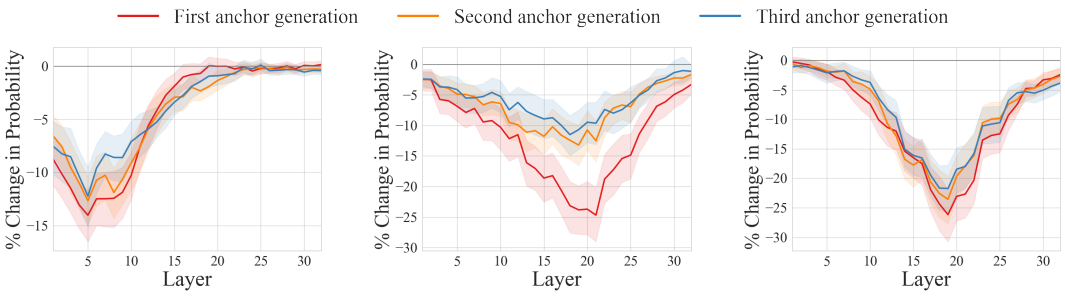
Figure Q: **Change in the prediction probability in open-ended QA format when intervening on attention edges between video, non-option question, and last token.** *Source* ↛ *Target* indicates blocking attention edges from source tokens to the target tokens. In the absence of explicit temporal keywords in the open-ended format, the last position itself serves as a checkpoint for video-text integration in the middle layers.



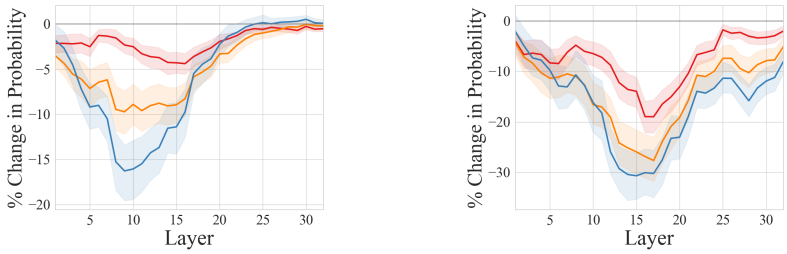
(a) Action Antonym (b) Moving Direction (c) Object Count

Figure R: **Layerwise prediction probability for ground truth answers in open-ended QA format in the last token position.** The probability for the ground truth starts to rise immediately after the middle layers.

Experimental setup. We adopt the Temporal QA subset of VCGBench (Maaz et al., 2024b), a video conversation benchmark that includes diverse reasoning-based QA examples. Specifically, we analyze how information flows among the video, question, generated response, and the last position, as the model continues to generate multiple temporal vocabularies. While automatically identifying temporal vocabulary in generated responses is challenging, verbs serve as strong candidates, since they often contain action and time-related semantics crucial for solving VideoQA tasks. To extract temporal vocabulary from model responses, we employed spaCy's en-core-web-lg model to detect verbs and used their token positions as semantic anchors for our analysis.



(a) Video ↛ Question (b) Question ↛ Last (c) Video ↛ Last



(d) Video ↛ Non-last response (e) Non-last Response ↛ Last

Figure S: **Open-ended generation analysis on VCGBench.** We set verbs in the generated response as potential semantic anchors containing temporal vocabulary and measure a probability drop at generating each anchor when intervening attention edges from *Source* ↛ *Target* tokens.

Then, we trace the probability change after Attention Knockout at each stage of anchor generation. For instance, given the question *"What is happening in this video?"* with baseline response *"A boy swings a bat and runs to the bases..."*, the detected anchors are [swings, runs, ...]. We then analyze information flow at different generation stages: generating the first anchor with no prior context (e.g., prompt: *"USER: <video> What is happening in this video? ASSISTANT: A boy"*, target: *"swings"*), generating the second anchor after one (e.g., prompt: *"USER: <video> What is happening in this video? ASSISTANT: A boy swings a bat and"*, target: *"runs"*), and so on.

Results. Figure S depicts the information flow of different routes: *video* → *question* → *last* (a-b), *video* → *last* (c), and *video* → *non-last response* → *last* (d-e). Our results show that, as generation continues, newly produced temporal verbs in the response increasingly function as additional core checkpoints. This is evidenced by a clear shift in the dominant sources feeding the last position. As the number of temporal anchor increases, the contribution of response to last consistently grows, while the relative importance of video to last and question to last decreases (Fig. S(a-c)). We also observe a structural change in the route through which video evidence reaches the final prediction. At the initial anchor generation with no given anchor, the model relies more on video to question to last (Fig. S(a-b)), whereas with larger number of anchor it increasingly depends on video to response to last (Fig. S(d-e)).

We observe consistent monotonic trends as the number of generated anchors from 1 to 3, and therefore expect the similar patterns to hold for more anchors. Overall, open-ended generation exhibits the same checkpoint-driven information flow pattern observed in multi-choice QA. The model dynamically forms new checkpoints around temporal concepts, and effective information flow reorganizes accordingly, confirming that our core claims generalize to open-ended VideoQA.

D Visualization of Logit Lens Analysis

To further investigate the emergence of temporal concepts in videos (Section 3.3), we visualize the token positions corresponding to spatial and temporal concepts extracted by Logit Lens. As shown in Fig T, static concepts emerge first, and temporal concepts appear later as layers go deeper.

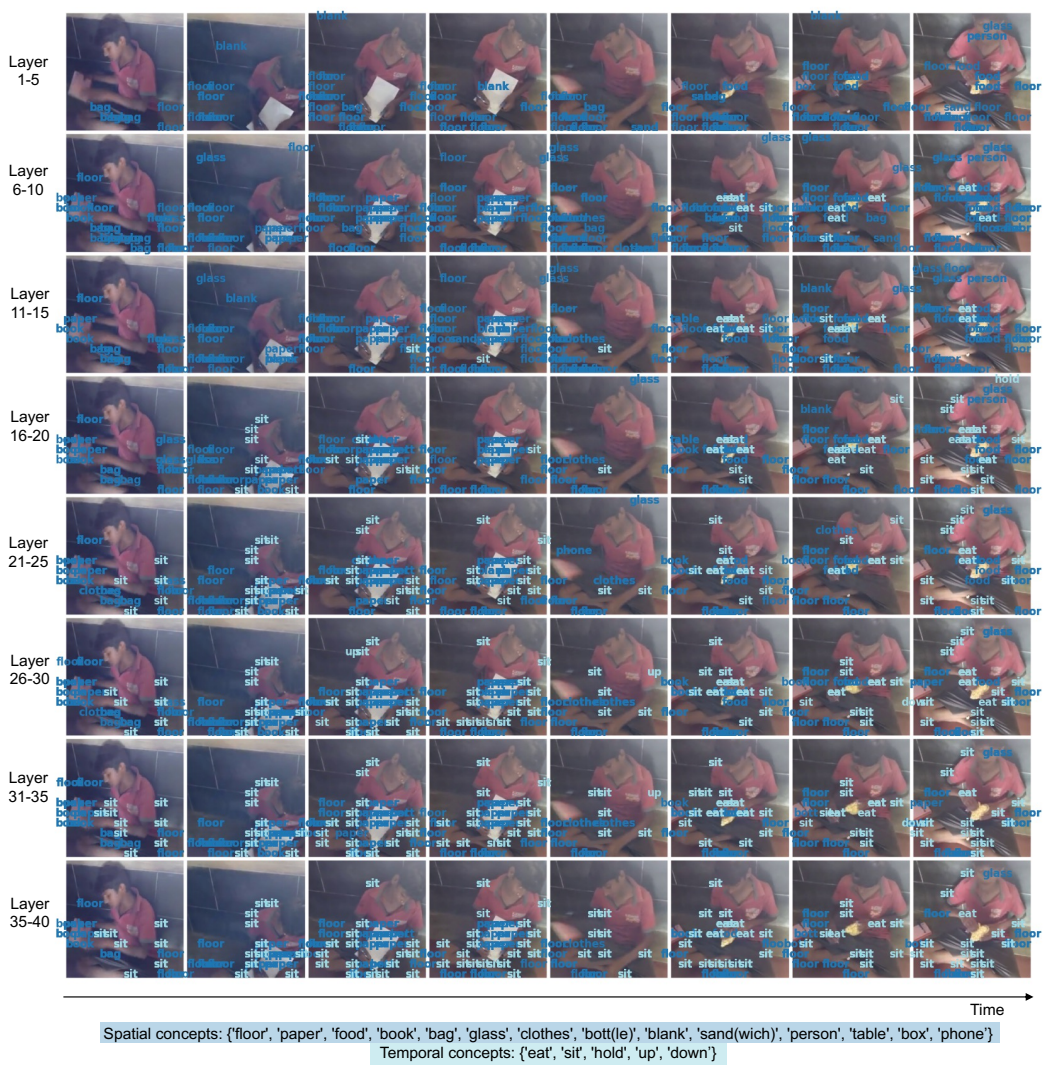


Figure T: **Visualization of spatial and temporal concepts extracted by Logit Lens.** Spatial concepts tend to settle on salient regions early, and temporal concepts then emerge mainly on the remaining tokens rather than replacing already stabilized spatial tokens.

Moreover, the emergent positions of temporal concepts spatially aligns with their relevant foreground regions. For example, the concept "sit" aligns with the region around a seated person. Beyond this overall trend, we additionally observe that spatial concepts tend to settle on salient regions early, and temporal concepts then emerge mainly on the remaining tokens rather than replacing already stabilized spatial tokens. We interpret this behavior as a consequence of priority in spatial localization. Specifically, foreground regions are first mapped with spatial concepts that describe salient entities or attributes, and temporal concepts tend to emerge afterward. We hypothesize that this order of emergence enforces temporal concepts to occupy the remaining token positions not already taken by spatial concepts, which explains the positioning mechanisms of temporal concepts.

E Additional Visualization of Video-to-Question Attention Maps

We extend our analysis of emergent video concepts propagated through text tokens to various VideoQA tasks (Fig. 5) to verify the generalizability of our findings. To this end, we further conduct a qualitative analysis of video-to-question attention maps on the action antonym, action sequence, scene

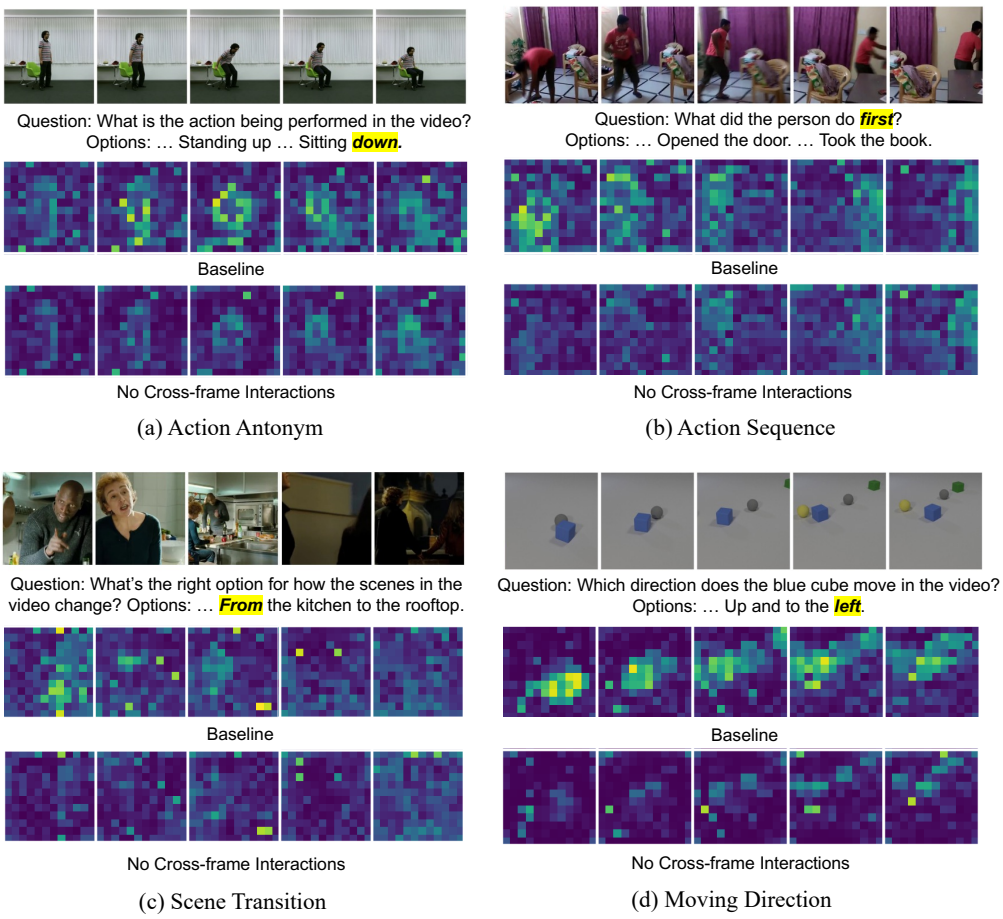


Figure U: **More visualizations of video-to-question attention maps.** Queries are highlighted in yellow, and keys correspond to video tokens. The baseline model attends to visually relevant tokens that align with the semantics of each query token, whereas disabling cross-frame interactions makes the model's attention less adaptive and limits its ability to infer the correct temporal context.

transition, and moving direction tasks. As depicted in Fig. U, the baseline model consistently trigger the attentions on video tokens that are semantically aligned with the highlighted query words such as *down*, *first*, *from*, and *left*. For example, in the action antonym (Fig. U(a)) and action sequence tasks (Fig. U(b)), the query tokens focus on frames around the critical action change, while in the scene transition (Fig. U(c)) and moving direction (Fig. U(d)) tasks they concentrate on frames that capture the transition of the scene or the motion of the object. However, when cross-frame interactions are disabled, the attention of the video tokens is not triggered by their relevant temporal vocabulary, showing undistinguishable attention maps across frames. This demonstrates the lack of capability to capture temporal relationships and scene changes when the cross-frame interaction is blocked. These consistent patterns across diverse tasks support that our attention-based analysis and the associated findings are broadly applicable to VideoQA models beyond the specific task.

F Further Analysis

F.1 Quantitative validation of effective information flow pathways on long-form videos

We extend the effective pathway analysis to long video question-answering problems. To this end, we disabled the ineffective pathways of LLaVA-NeXT-7B-Video-FT as configured in Section 3.5 and evaluate the performance on LongVideoBench (Wu et al., 2024). Table A showcases that

Table A: **Impact of effective information flow pathways on LongVideoQA performance.** The total number of attention edges is calculated by counting valid (query, key) pairs over all attention layers.

Case	Total Number of Attention Edges	Object-referred Event	Object before/ after Object	Scene-referred Object Tracking	All
Full causal attention	25.7M (100%)	52.9	40.9	44.4	46.1
Attention in effective pathways	10.8M (42%)	54.0	39.4	43.2	45.5

LLaVA-NeXT-7B-Video-FT also retains competitive performance on long-form videos understanding using only 42% of the original attention edges, with only a marginal accuracy drop of 0.6%p. This validates that our findings on the internal mechanisms of VideoLLMs generalize across various video question-answering tasks.

F.2 Generalizability of our analysis to various benchmarks

We further investigate how the information flow changes with different forms of input videos and question types, including long video understanding in LongVideoBench (Wu et al., 2024) and spatial understanding in Video-MME (Fu et al., 2024).

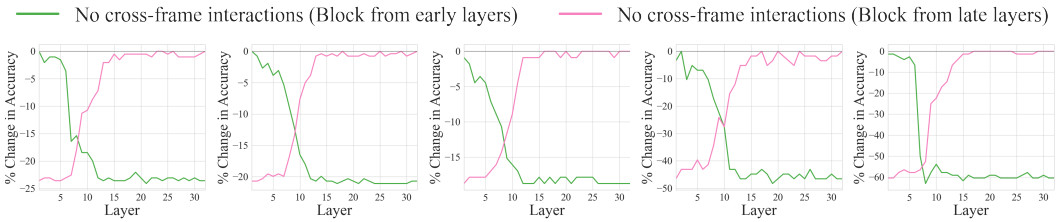
Long video understanding. We validate VideoLLaMA3-7B on LongVideoBench (Wu et al., 2024) using the same setup in the main paper. Fig. X, Y, Z, and AA, highlights the results with 8 frames 12x12 tokens per each frame. Fig. AB, AC, AD, AE also shows results with 24 frame inputs. Overall, the effective layer ranges for long-form VideoQA maintain similar patterns to short video benchmarks. A notable difference is that the probability drop from cross-frame attention and video-to-question components is relatively smaller compared to short video tasks (See Fig. K and Fig. L). We conjecture this to two factors: (1) long video benchmarks typically do not require every frame to be equally informative, reducing the need for comprehensive visual processing, and (2) questions in long video benchmarks contain more descriptive information, causing the model to rely more heavily on textual cues from the question rather than visual content.

Spatial understanding. To compare the patterns driven from spatial and temporal reasoning tasks, we adopt action recognition and spatial perception tasks from Video-MME (Fu et al., 2024). Results with LLaVA-NeXT-7B-Video-FT are shown in Fig. AF, AG, AH, and AI. When we block the cross-frame interaction, the action recognition task exhibits a clear and consistent performance drop, indicating that temporal aggregation is crucial for this setting. In contrast, the spatial perception task shows a much smaller average degradation and a much larger variance across samples. We conjecture that this pattern arises because some spatial perception questions incidentally benefits from temporal information, whereas many others can be answered from static scenes. Therefore, the impact of blocking cross-frame interaction ranges from almost no change to a significant drop.

F.3 Failure case analysis

To offer a mechanistic understanding of why the model makes wrong answers, we extend our analysis to failure cases. We focus on two aspects: (1) samples where the VideoLLM is highly confident in false options, and (2) samples when the VideoLLM highly rely on static scene information rather than temporal reasoning.

To this end, we analyze how the probability of a wrong answer changes in the failure VideoQA samples. As can be seen in Fig. AK and Fig. AL, the cross-modal flow patterns routing false options are the same as those with successful samples routing true options, indicating that a root cause could be in an earlier stage of building video representations. In contrast, when we intervene on the cross-frame interaction as in Fig. AJ, failed cases split into two patterns. In some examples, the probability of the incorrect option decreases (green), while in others it increases (pink), and thus no single consistent behavior emerges.



(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Figure V: **Change in accuracy when gradually disconnecting cross-frame attention edges in each layer.** The green line shows the accuracy change when gradually blocking cross-frame attention from the first layer up to the l^{th} layer, whereas the pink line shows the accuracy change when blocking from the l^{th} layer to the last layer.

Table B: **Impact of cross-frame attention in the second half layers of VideoLLMs on answer generation.** We block cross-frame attention in the first and second half of the total layers and measure the resulting accuracy drop (%). While disabling cross-frame attention in the first half layers significantly degrades accuracy, disabling it in the second half layers barely impact performance.

Case	Action Antonym	Action Sequence	Scene Transition	Moving Direction	Object Count
First half layers	24.1	20.2	18.0	44.8	60.8
Second half layers	0.5	0.7	0.8	1.7	1.2

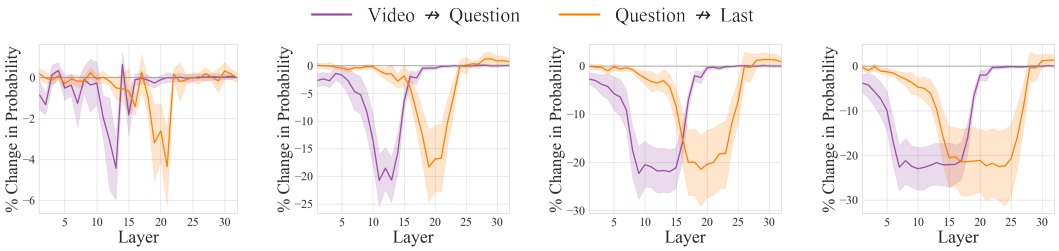
We conjecture that, in the first type, the VideoLLM is already overconfident in false options, where the erroneous signal may come from cross-frame interaction or misaligned between specific video and language. In the second type, a plausible interpretation is that the model primarily relied on per-frame static scene information. Thus, when cross-frame interaction is interrupted, the model even more emphasizes the rationale from static scenes, which in turn reinforces confidence in false options relevant to those static scenes.

F.4 Signal leakage check

Our attention knockout setup blocks cross frame interaction within a local layer window, which may not capture potential residual signal leakage through bypassing pathways. To address this concern, we extend the intervention in two complementary ways. Instead of blocking cross-frame interaction only in the first half of the layers, we progressively expand the knockout from the first layer up to the N-th layer for $N = 1, \ldots, 32$. We also conduct the reverse intervention by blocking from the last layer backward. This gradual design explicitly check whether auxiliary signals from video tokens continue to propagate beyond early-to-middle layers, which would contradict the effective information flow range identified in our analysis. As shown in Fig. V, progressively expanding the knockout does not lead to any additional significant performance drop unless the intervention overlaps with early-to-middle layers, the effective cross frame interaction range that we identified. These results indicate that the cross-frame interaction active occurs in the early-to-middle layer, supporting the validity of our findings beyond a specific knockout configurations.

F.5 Impact of blocking cross-frame attention in the second half layers of VideoLLMs

We further examine the impact of blocking cross-frame attention in the second half layers. In Table B, the accuracy drop is marginal when temporal interactions are blocked in the latter layers, compared to the earlier layers. This supports our claim that active temporal interaction within video tokens occurs in early-to-middle layers.



(a) $k = 1$ (b) $k = 5$ (c) $k = 9$ (d) $k = 13$

Figure W: **Impact of window size k on Attention Knockout.** Following Geva et al. (2023), we take $k = 9$ as our default choice.

Table C: **Coverage of models.** We adopt MLLMs with diverse sizes, base vision encoders, and base LLMs to ensure generalizability.

Model	Size	Base Vision Encoder	Base LLM
Mini-InternVL-4B	4B	InternViT-300M-448px	Phi-3-mini-128k-instruct
LLaVA-NeXT-7B	7B	CLIP-ViT-L-336px	Vicuna-7B-v1.5
LLaVA-NeXT-13B	13B	CLIP-ViT-L-336px	Vicuna-13B-v1.5
VideoLLaMA3-7B	7B	siglip-so400m-patch14-384	Qwen2.5-7B

F.6 Robustness of our analyses on choice of window size k

We observed the robustness of our analyses when using window sizes with sufficient width, and therefore chose to follow the k value of 9 as used in (Geva et al., 2023). Specifically, to examine the robustness of our analyses to the window sizes, we conducted an extended analyses by varying the window sizes in 1, 5, 9, 13. As shown in Table W, when the window size is extremely small (e.g., k=1), the narrow attention block is easily bypassed and VideoLLMs can still transmit information through remaining effective information pathways. This leads to only marginal probability drops across the layers. In contrast, with wider windows (k=5,9,13), we observe significant probability drops, which validates the robustness of our analyses across various choice of window sizes.

G Discussion

Discussion on the probability increases in the last layers. We observe an increase in the true option probability when attention from the question to the last position is knocked out in later layers in some cases. Interleaving the attention knockout analysis from question to last (Fig. 3 and B) and the layerwise prediction probability analysis(Fig. 8 and E) suggests that propagating information from question tokens to the last token boosts the true option, as this flow consolidates evidence in the middle layers. Therefore, the belief of VideoLLMs is already stabilized, so keeping this pathway open mainly acts as a broad amplifier, increasing probabilities for both true and false options. Thus, blocking the pathway only at the final layers preserves earlier propagated evidence for the true option in the hidden states while preventing further amplification of false options. There, the true option probability can rises relative to the false ones and can even increase in absolute terms, matching the behavior in Fig. 1(b).

Future applications of our findings. We further discuss how our findings can be leveraged in practice. For training, our analysis suggests that current VideoLLMs rely on a relatively narrow set of dominant information pathways, so intentionally blocking these pathways during training could regularize the model to explore alternative pathways, thereby better utilizing the representational capacity of VideoLLMs. For testing, our identification of effective information flow ranges implies that tokens beyond these ranges contribute marginally to the final decision, which opens a path to apply early-exiting strategies (Elbayad et al., 2020; Schuster et al., 2022; Bae et al., 2023) that adaptively stop computation for such tokens to reduce inference cost while preserving accuracy.

Table D: **List of vocabularies used for semantic concept extraction.** Keywords are parsed from Action Sequence question prompts and converted to lowercase and present tense to avoid interference from linguistic completion in later layers.

Spatial Keywords	bag, bed, blank, book, box, cabinet, camera, clothes, cup, door, floor, food, glass, laptop, paper, person, phone, sandwich, table
Temporal Keywords	close, down, drink, eat, hold, on, open, put, sit, take, throw, tidy, up

Table E: **Effective pathway layer ranges for different VideoLLMs.** (a) Layer ranges for effective pathways across different models, determined by selecting 5-layer intervals with significant probability drops from Attention Knockout analysis. (b) Detailed knockout results showing probability drops across layer intervals in Action Antonym task. Significant drops ($< -5\%$) are highlighted in gray; N/A indicates unavailable layers.

(a) Effective pathway layer ranges

Model	Cross-frame Interactions	Video-to-Question	Question-to-Last
LLaVA-NeXT-7B-Video-FT	L6-15	L6-20	L16-25
LLaVA-NeXT-13B-Video-FT	L6-15	L6-20	L16-30
Mini-InternVL-4B-Video-FT	L6-15	L6-20	L11-30
VideoLLaMA3-7B	L1-15	L6-20	L21-28

(b) Attention Knockout results in Action Antonym

Model	Interaction	L1-5	L6-10	L11-15	L16-20	L21-25	L26-30	L31-35	L36-40
LLaVA-NeXT-7B-Video-FT	Cross-frame	-4.2	-11.1	-6.3	-0.2	0	-0.2	-0.2	N/A
	Video-to-Question	-3.9	-15.1	-21.5	-5.6	-0.2	0	0	N/A
	Question-to-Last	-0.3	-1.2	-4.5	-19.3	-15.1	0.7	1.1	N/A
LLaVA-NeXT-13B-Video-FT	Cross-frame	-0.7	-11.2	-11.1	-2.1	-0.2	-0.2	-0.2	-0.3
	Video-to-Question	-1.2	-16.7	-29.1	-9.2	-0.3	-0.2	-0.2	-0.2
	Question-to-Last	-1.8	-2	-4.6	-21.7	-28.4	-5.7	-0.1	-1.9
Mini-InternVL-4B-Video-FT	Cross-frame	-2.3	-11.1	-11.5	-3.3	0	0.2	0	N/A
	Video-to-Question	-2.4	-24.4	-35.0	-15.9	-1.3	-0.3	-0.2	N/A
	Question-to-Last	0	-1.3	-5.9	-30.8	-46.8	-14.1	-3.2	N/A
VideoLLaMA3-7B	Cross-frame	-65.1	-61.4	-14.9	-5.0	0	0.2	N/A	N/A
	Video-to-Question	-4.3	-7.2	-18.5	-16.3	-2.2	0	N/A	N/A
	Question-to-Last	2	3.7	1.9	-2.7	-16.2	-19.1	N/A	N/A

H Implementation Details

We describe the implementation details for the VideoLLMs and their training setup. Table C shows the details of the VideoLLMs.

Training setup. Our video instruction tuning data is derived from VideoChat2-IT (Li et al., 2024b), comprising 874k samples covering tasks such as VideoQA, captioning, reasoning, classification, and conversation. These samples are from diverse video understanding benchmarks, including VideoChatGPT-100k (Maaz et al., 2024b), VideoChat-11k (Li et al., 2024a), Webvid (Bain et al., 2021), YouCook2 (Zhou et al., 2018), TextVR (Wu et al., 2025), NExT-QA (Xiao et al., 2021), CLEVRER (Yi et al., 2019), TGIF (Li et al., 2016), Ego4D (Grauman et al., 2022), Kinetics-710 (Kay et al., 2017), and Something Something V2 (Goyal et al., 2017). We freeze the vision encoder while fully fine-tuning the MLP projector and LLM backbone. Our experiments are conducted with NVIDIA A6000 GPUs.

- **LLaVA-NeXT-7B-Video-FT.** During training, we initialize the model with LLaVA-NeXT-7B (Liu et al., 2024b), which employs CLIP-ViT-L-336px (Radford et al., 2021) as the vision encoder and Vicuna-7B-v1.5 (Zheng et al., 2023) as the language model. We utilize a batch size of 128 and train for 3 epochs. The base learning rate is initially set to 2e-5 and is decayed to 5e-6 using a cosine scheduler, with a warmup ratio of 0.2. For both training

and inference, we uniformly sample 8 frames as input and resize each frame into 336×336 pixels. These frames are then processed through a vision encoder to extract 8×24×24 patch embeddings. Next, we use an MLP projector to project these embeddings, followed by average spatial pooling to generate 8×12×12 video tokens.

- **LLaVA-NeXT-13B-Video-FT.** Similarly, we initialize the model with LLaVA-NeXT-13B (Liu et al., 2024b), which utilizes CLIP-ViT-L-336px (Radford et al., 2021) as the vision encoder and Vicuna-13B-v1.5 (Zheng et al., 2023) for the language model component. The model is trained for 1 epoch using the same training recipe and video token sampling strategy as LLaVA-NeXT-7B-Video-FT.
- **Mini-InternVL-4B-Video-FT.** We start with Mini-InternVL-4B (Gao et al., 2024), which adopts InternViT-300M-448px as the vision encoder and Phi-3-mini (Abdin et al., 2024) as the LLM backbone. We use a batch size of 128 with a learning rate of 4e-5, which decays to zero following a cosine schedule with a warmup ratio of 0.03 for a total of 3 epoch. For both training and inference, we uniformly sample 8 frames as input and resize each frame into 448×448 pixels. These frames are passed through the vision encoder, producing 8×32×32 patch embeddings. After applying the MLP projection, we put 8×16×16 video tokens as the input of the language model.
- **VideoLLaMA3-7B.** VideoLLaMA3-7B (Zhang et al., 2025a) uses SigLIP (Zhai et al., 2023) as a vision encoder and Qwen2.5-7B (Qwen et al., 2025) as a LLM backbone. We directly use VideoLLaMA3-7B without fine-tuning and put 8×12×12 video tokens as the input.

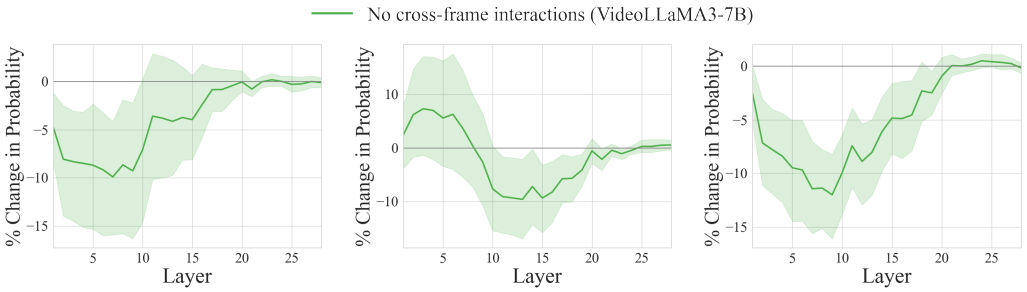
Implementation details for Attention Knockout. In VideoQA, a model generates an answer a from a given video-question pair (v, q), where the question may contain n number of options $o = [o_1; o_2; ...; o_n]$. We employ Attention Knockout (Geva et al., 2023) to measure the information flow between different input parts. Specifically, the model initially predicts the answer a with the highest probability p_{base} at the last token position of the input sequence. After applying Attention Knockout as explained in § 2.2, we trace the relative change in probability $\%p_{\text{change}} = ((p_{\text{knockout}} - p_{\text{base}})/p_{\text{base}}) \times 100$, where p_{knockout} is the updated probability for the same answer a derived after intervention. Unless otherwise stated, we apply Attention Knockout within a window size of $k = 9$ layers around the l^{th} layer of MLLMs, and trace the probability change for the first tokenized subword of the complete answer.

Implementation details for Logit Lens. To quantify emergence of spatial and temporal semantic concepts in video tokens, we employ Logit Lens (nostalgebraist, 2020). We trace top-1 logits by projecting intermediate representations of all video tokens across layers using the language model head. We use Action Sequence videos with LLaVA-NeXT-13B-Video-FT. For the vocabulary pool, we parse spatial and temporal keywords from Action Sequence question prompts (Table D). To trace initial concept emergence, we convert parsed words to lowercase present tense, as linguistic completion occurs in later layers and can impact the analysis.

Implementation details for effective pathway analysis. To identify effective pathways, we use Attention Knockout results from Action Antonym tasks (Table E). We divide layers into 5-layer intervals, calculate average probability drops, and select intervals with significant drops ($< -5\%$) as effective layers. We then enable cross-frame interactions, *video* → *question*, and *question* → *last* flows only within these effective layers while disabling *video* → *last* and *last* → *last* connections across all layers. Additionally, flows to *video* and *question* tokens are blocked in late layers as these tokens are no longer needed (e.g., after layers 20 and 25 respectively in LLaVA-NeXT-7B-Video-FT).

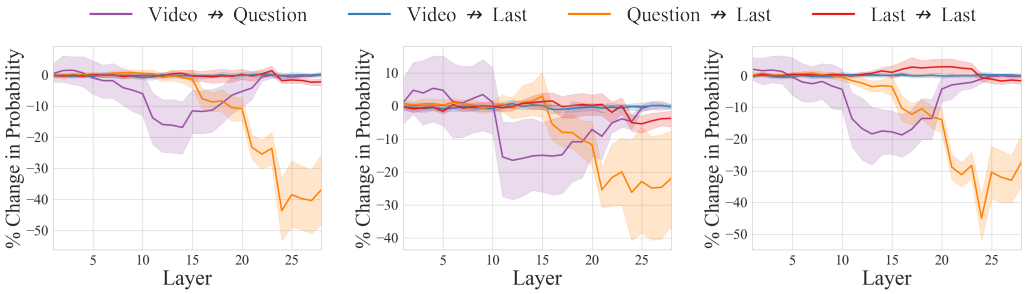
I The Usage of Large Language Models.

In this work, LLMs were used only to polish manuscript clarity, fix grammatical errors, and enhance readability. Specifically, all initial writing was done by the authors, with LLMs used afterwards for sentence-level polishing in part of the manuscript. LLMs were not involved in research ideation and experimental design. All core contributions, methodologies, and findings are the result of the authors' original work.



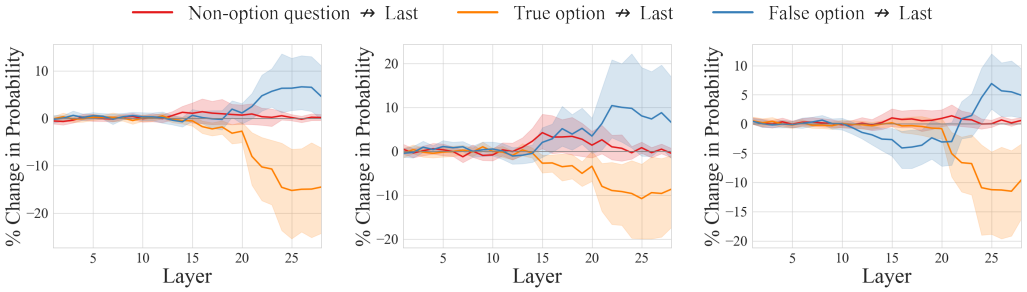
(a) O2E (b) O3O (c) SOS

Figure X: **LongVideoBench with 8 frame inputs: Prediction probability change after disconnecting cross-frame attention edges in VideoLLaMA3-7B.** Object-referred Event (O2E), Object before/after Object (O3O), Scene-referred Object Tracking (SOS) subsets are used.



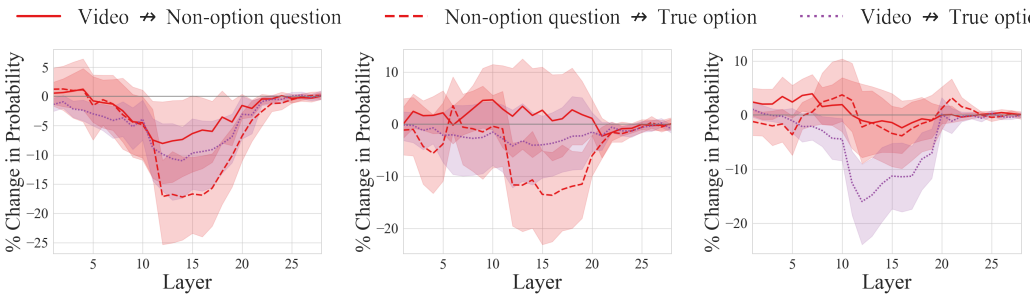
(a) O2E (b) O3O (c) SOS

Figure Y: **LongVideoBench with 8 frame inputs: Prediction probability change when intervening on attention edges between video, question, and last token VideoLLaMA3-7B.**



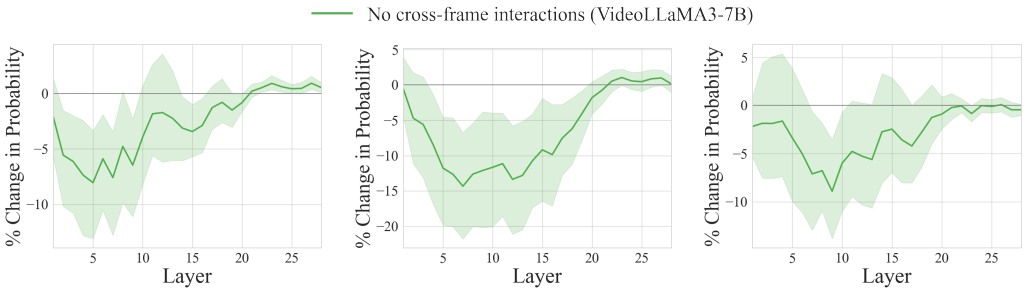
(a) O2E (b) O3O (c) SOS

Figure Z: **LongVideoBench with 8 frame inputs: Prediction probability change when intervening on attention edges from different parts of the question tokens to the last token in VideoLLaMA3-7B.**



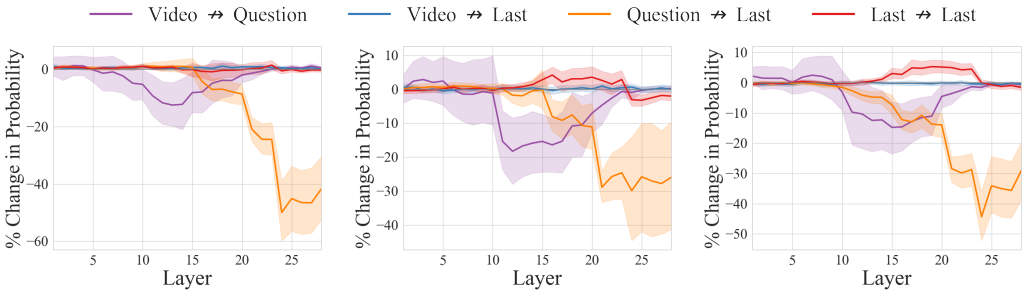
(a) O2E (b) O3O (c) SOS

Figure AA: **LongVideoBench with 8 frame inputs: Prediction probability change when intervening on attention edges to the true option position in VideoLLaMA3-7B.**



(a) O2E (b) O3O (c) SOS

Figure AB: **LongVideoBench with 24 frame inputs: Prediction probability change after disconnecting cross-frame attention edges in VideoLLaMA3-7B.** Object-referred Event (O2E), Object before/after Object (O3O), Scene-referred Object Tracking (SOS) subsets are used.

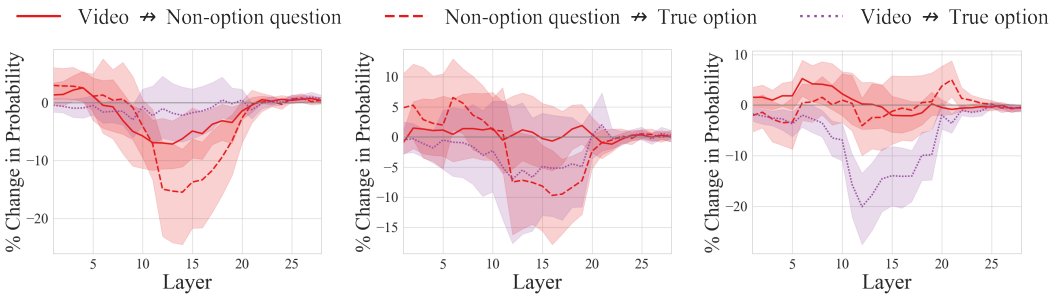


(a) O2E (b) O3O (c) SOS

Figure AC: **LongVideoBench with 24 frame inputs: Prediction probability change when intervening on attention edges between video, question, and last token VideoLLaMA3-7B.**

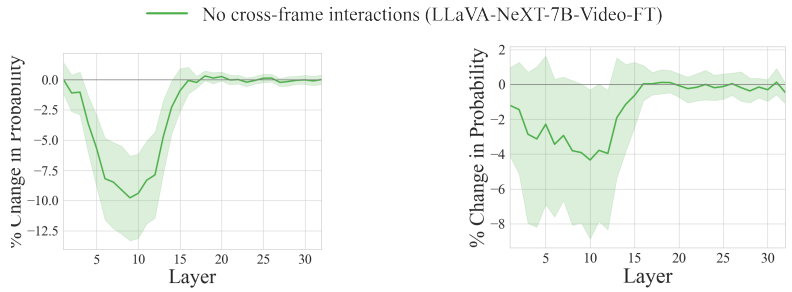
(a) O2E (b) O3O (c) SOS

Figure AD: **LongVideoBench with 24 frame inputs: Prediction probability change when intervening on attention edges from different parts of the question tokens to the last token in VideoLLaMA3-7B.**



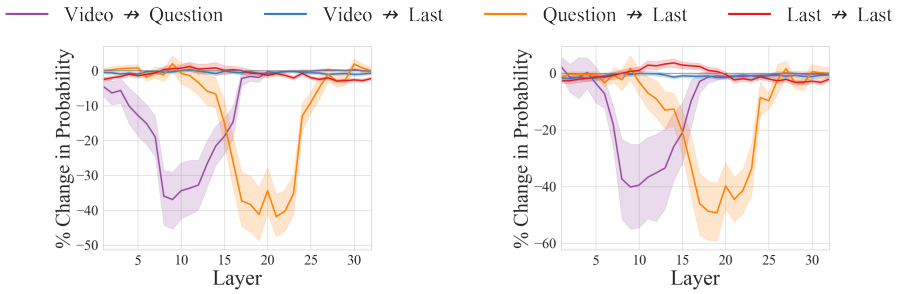
(a) O2E (b) O3O (c) SOS

Figure AE: **LongVideoBench with 24 frame inputs: Prediction probability change when intervening on attention edges to the true option position in VideoLLaMA3-7B.**



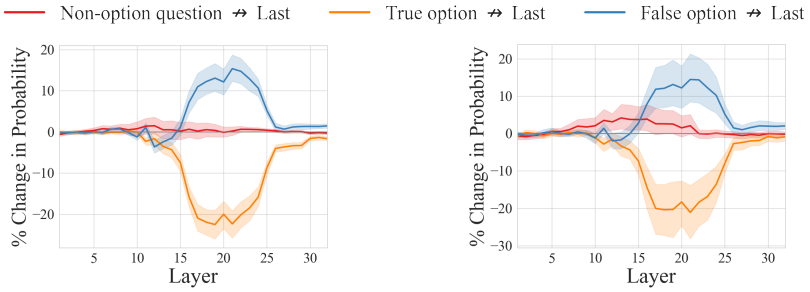
(a) Action Recognition (b) Spatial Perception

Figure AF: **Video-MME: Change in the prediction probability when disconnecting cross-frame attention edges.** The spatial perception task shows a much smaller drop, as it contains questions that can be answered with static scenes.



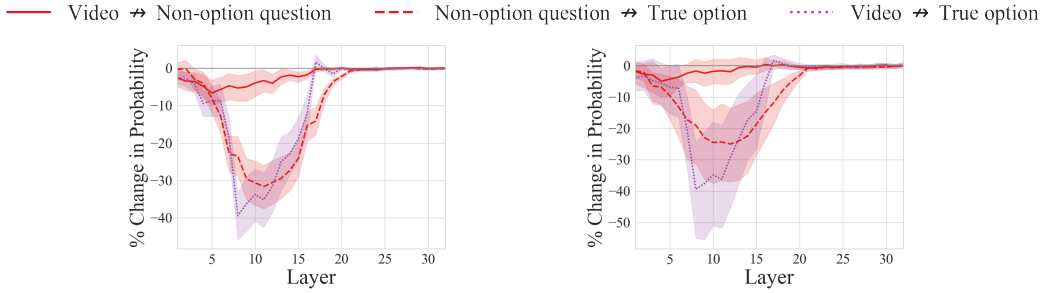
(a) Action Recognition (b) Spatial Perception

Figure AG: **Video-MME: Change in the prediction probability when intervening on attention edges between video, question, and last token.**



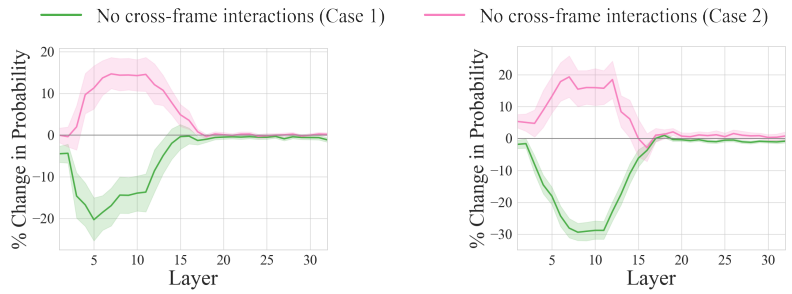
(a) Action Recognition (b) Spatial Perception

Figure AH: **Video-MME: Change in prediction probability when intervening on attention edges from different parts of the question tokens to the last token.**



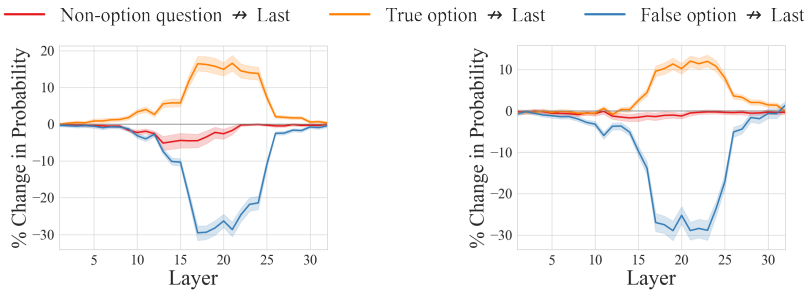
(a) Action Recognition (b) Spatial Perception

Figure AI: **Video-MME: Change in the prediction probability when intervening on attention edges to the true option position.**



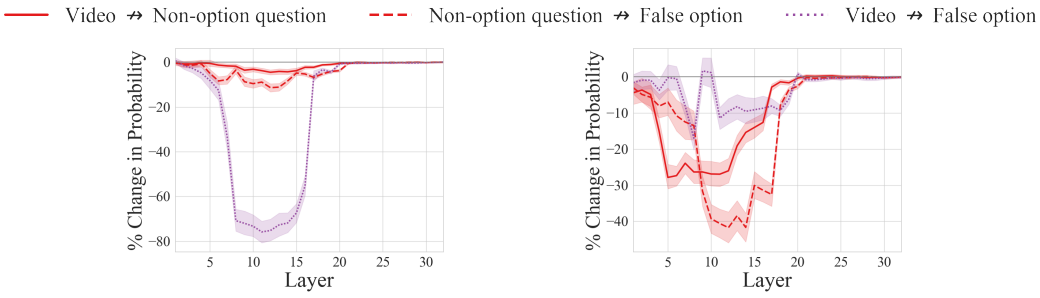
(a) Action Antonym (b) Moving Direction

Figure AJ: **Failure case analysis: Change in the prediction probability when disconnecting cross-frame attention edges.** We observe two distinct patterns in the failure samples. In Case 1 (green), the incorrect option probability decreases after disabling cross-frame attention edges, suggesting that the erroneous signal carried by these edges was a primary cause of the model's misprediction. In Case 2 (pink), the incorrect option probability instead increases, indicating a form of static bias where the model even more emphasizes unhelpful static information.



(a) Action Antonym (b) Moving Direction

Figure AK: **Failure case analysis: Change in the prediction probability when intervening on attention edges from different parts of the question tokens to the last token.** *Source ↛ Target* indicates blocking attention edges from source tokens to the target tokens.



(a) Action Antonym (b) Moving Direction

Figure AL: **Failure case analysis: Change in the prediction probability when intervening on attention edges to the false option position.** *Source ↛ Target* indicates blocking attention edges from source positions to the target positions.