Causality Matters: How Temporal Information Emerges in Video Language Models

Yumeng Shi, Quanyu Long, Yin Wu, Wenya Wang

Nanyang Technological University
yumeng001@e.ntu.edu.sg, quanyu001@e.ntu.edu.sg, wuyi0023@e.ntu.edu.sg, wangwy@ntu.edu.sg

Abstract

Video language models (VideoLMs) have made significant progress in multimodal understanding. However, temporal understanding, which involves identifying event order, duration, and relationships across time, still remains a core challenge. Prior works emphasize positional encodings (PEs) as a key mechanism for encoding temporal structure. Surprisingly, we find that removing or modifying PEs in video inputs yields minimal degradation in the performance of temporal understanding. In contrast, reversing the frame sequence while preserving the original PEs causes a substantial drop. To explain this behavior, we conduct substantial analysis experiments to trace how temporal information is integrated within the model. We uncover a causal information pathway: temporal cues are progressively synthesized through inter-frame attention, aggregated in the final frame, and subsequently integrated into the query tokens. This emergent mechanism shows that temporal reasoning emerges from inter-visual token interactions under the constraints of causal attention, which implicitly encodes temporal structure. Based on these insights, we propose two efficiency-oriented strategies: staged cross-modal attention and a temporal exit mechanism for early token truncation. Experiments on two benchmarks validate the effectiveness of both approaches.

Code — https://github.com/ANDgate99/Causality-Matters
Extended version — https://arxiv.org/pdf/2508.11576

1 Introduction

Video language models (VideoLMs) (Zhang et al. 2024; Lin et al. 2023), built upon large language models (LLMs) (OpenAI 2024; Yang et al. 2025), have advanced a wide range of video understanding tasks, including captioning, question answering, and temporal reasoning. Among these challenges, temporal understanding (Cai et al. 2024; Shangguan et al. 2025), defined as the ability to recognize and interpret the order, duration, and relationships of events, stands out as both fundamental and difficult. From causal reasoning in instructional videos to maintaining narrative coherence in long-form content, temporal understanding plays a vital role in enabling intelligent video-language interactions.

In the pursuit of enhancing the temporal understanding capabilities of VideoLMs (Li et al. 2025; Nguyen et al. 2025), many studies (Liu et al. 2025) have focused on what is likely the model's most direct method for modeling temporal structures: the positional encodings (PEs). It is commonly assumed that using a more sophisticated positional encoding function leads directly to improved temporal awareness. This belief has spurred a variety of modifications, from extending PEs to higher dimensions (Bai et al. 2025; Wei et al. 2025) to designing variable-rate formats (Ge et al. 2024) to better capture temporal information. However, this intense focus on an explicit signal raises a basic but often overlooked question: ① ***To what extent do PEs support temporal understanding in modern VideoLMs?***

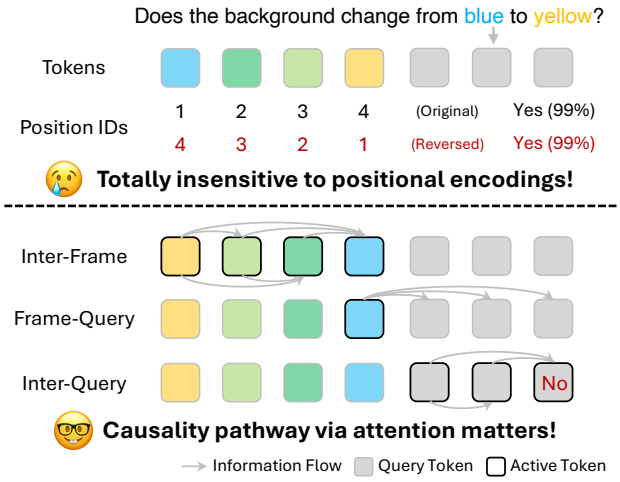

Figure 1: Temporal information emerges through a causal attention pathway instead of PEs. Reversing position IDs has little effect, but reversing the frame order while keeping position IDs aligned with the original order changes the output.

To better understand the role of PEs, our analysis begins by iteratively removing them layer by layer and selectively modifying them in the video and query inputs separately. The outcome is revealing: PEs have only a marginal impact on temporal understanding performance beyond the first layer, and while query PEs play a more significant role, video input PEs contribute little. These observations raised a more fundamental question: ② ***If the model does not pri-***

marily rely on explicit PEs for temporal understanding, what is the primary source of this capability?

This question motivates a follow-up experiment, where we evaluate the models on deliberately reversed video input sequences while preserving PEs. In contrast to the PE ablation, this reversal severely degrades the performance. These initial findings suggest that VideoLMs are highly sensitive to the order of video frames, yet do not primarily rely on explicit PEs to detect it. Building on this, we propose a new hypothesis: Currently, temporal understanding is not a fixed attribute extracted from PEs, but an emergent phenomenon derived from the order-aware processing imposed by causal masking in attention. More specifically, it emerges from: ③ ***How the causal attention mechanism permits temporal information to be generated and to flow from interaction between tokens across layers.***

We examine the hypothesis by tracing the model's temporal understanding process backward from its final output. This reveals that the model generates its answer primarily from the query, rather than directly referencing the video evidence at the final stage. Tracing this dependency further upstream, we observe a multi-stage process in the model's early and middle layers: raw information is first aggregated across the video frames, and the resulting representation is then used to enrich the query's context. In addition, we validate that this entire information pathway is not merely correlational but genuinely causal, providing direct evidence of how temporal understanding emerges within the model.

Based on experiments with the advanced Qwen2.5-VL (Bai et al. 2025) and LLaVA-OneVision (Li et al. 2024), which represent state-of-the-art VideoLMs, we identify several key takeaways about how modern VideoLMs internally process temporal information, as shown in Figure 1:

① PEs are not the primary source of temporal information currently, suggesting emphasis may shift from PE design to how models leverage them during training.

② Temporal information emerges from the causal attention mechanism's order-aware structure, highlighting the potential of sequential architectures for temporal modeling.

③ Temporal information is progressively constructed via a multi-stage causal pathway: it emerges from long-range inter-video interactions, flows into the query through the final video frame where it is aggregated, and is then processed independently by the query without further contribution from the video input.

Building on these insights, we discuss two potential application scenarios: (1) *Staged Modality Interaction for Sparse Attention*, which reduces cross-modal interaction to improve computational efficiency; (2) *Temporal Exit for KV Cache Compression*, which alleviates GPU memory pressure by discarding tokens that no longer contribute to temporal information propagation. We demonstrate the feasibility of both approaches through experiments on two datasets.

2 Related Work

2.1 Video Language Model

With the rise of LLMs (OpenAI 2024; Bi et al. 2024), numerous studies (Liu et al. 2024a; Zhang et al. 2024; Yao et al. 2024; Zhang et al. 2025a) have explored their integration into VideoLMs, which increasingly outperform traditional vision-only methods. One major direction focuses on enhancing the video encoder (Zhao et al. 2024b; Choudhury et al. 2024; Chung et al. 2025). Another line of work (Chen et al. 2024; Wei et al. 2025; Ge et al. 2024; Bai et al. 2025) adapts LLMs to better accommodate multimodal inputs. In addition, other studies aim to improve the efficiency of processing long video inputs (Fu et al. 2024; He et al. 2024; Shi, Long, and Wang 2025). Collectively, these advancements have driven progress in video-language understanding.

Format	Description	Count
Yes or No	Validate statement correctness.	2453
Multiple Choice	Select from multiple options.	1580
Caption Matching	Select the aligned caption.	1503
Captioning	Generate a caption.	2004

Table 1: Introduction for task types in TempCompass.

2.2 Temporal Understanding

VideoLMs face unique challenges in video temporal understanding, which requires detecting similarities and differences across frames. Benchmarks (Cai et al. 2024; Shangguan et al. 2025; Plizzari et al. 2025; Liu et al. 2024b) have been proposed to evaluate this capability. Besides, various approaches (Nie et al. 2024; Hu et al. 2024; Zhao et al. 2025; Fateh et al. 2024) have been introduced to improve this capability. Among them, T3 (Li et al. 2025) transfers temporal skills from synthetic text tasks to VideoLMs. Despite these advances, the internal mechanisms by which VideoLMs capture temporal information remain underexplored. Understanding how temporal information is extracted is essential for building more robust and generalizable models.

2.3 Model Mechanistic Interpretability

Many studies analyze how LLMs make predictions (Zhao et al. 2024a; Goldshmidt and Horovicz 2024; Biran et al. 2024; Wang et al. 2023). This line of research has also been extended to multimodal large language models (Yu and Ananiadou 2024; Golovanevsky et al. 2025; Zhang et al. 2025b). For instance, Basu et al. (2024) uses causal tracing to reveal how early-layer modules capture and transmit visual information. However, most existing work focuses on single-image tasks, and the interpretability of VideoLMs, particularly for temporal reasoning, remains underexplored. This gap motivates our work, which seeks to explain how temporal information is captured within VideoLMs.

3 Experimental Setting

3.1 Dataset

We conduct analysis on TempCompass (Liu et al. 2024b), which encompasses a diverse range of temporal phenomena, including action, speed, direction, attribute change, and event order. It provides a fine-grained assessment of a model's ability to reason over temporal information. The

benchmark includes four task formats, summarized in Table 1. Since our evaluation focuses on the model's first predicted token, we convert the caption generation task into a multiple-choice format, where the model is required to select the most temporally accurate option.

To validate our proposed efficiency-oriented strategies, we run downstream application experiments on the multiple-choice subset of the NExT-QA dataset (Xiao et al. 2021), which emphasizes explanatory reasoning over causal and temporal relationships. The test set contains 8,564 questions. To further assess generalizability, we additionally report results on the open-ended ActivityNet-QA dataset (Caba Heilbron et al. 2015), included in the appendix.

3.2 Model

We evaluate two representative VideoLMs: Qwen2.5-VL and LLaVA-OneVision. Qwen2.5-VL (Bai et al. 2025) uses multi-dimensional positional encodings tailored for video, making it well-suited to study PE impact on spatiotemporal modeling. In contrast, LLaVA-OneVision (Li et al. 2024) is another state-of-the-art multimodal model that does not adapt PEs for video, allowing us to derive insights that are agnostic to special PE designs. Unless specialized otherwise, we report results using Qwen2.5-VL-7B in the main text. Results for Qwen2.5-VL-3B and LLaVA-OneVision are provided in the appendix.

To facilitate a deeper understanding of VideoLM behavior in our experiments, we briefly outline the common architectural framework of modern VideoLMs.

Input Construction. VideoLMs process inputs by integrating an instruction, visual features, and a query into a unified token sequence. The instruction, providing task-specific guidance, is embedded as $\boldsymbol{T}_i \in \mathbb{R}^{L_i \times D}$. The visual input comprises sampled video frames, encoded and projected into the language embedding space as $\boldsymbol{V} \in \mathbb{R}^{T \times H \times W \times D}$, where T is the temporal resolution and H, W are the spatial dimensions. The tensor is then flattened into $\boldsymbol{T}_v \in \mathbb{R}^{THW \times D}$. The user query is embedded as $\boldsymbol{T}_q \in \mathbb{R}^{L_q \times D}$. The final model input is the concatenated sequence $\boldsymbol{T} = [\boldsymbol{T}_i; \boldsymbol{T}_v; \boldsymbol{T}_q]$, enabling joint multimodal reasoning.

Feature Interaction. The concatenated token sequence $\boldsymbol{T}$ is fed into the VideoLM. At each layer, the model computes hidden states $\boldsymbol{H} = [\boldsymbol{H}_i; \boldsymbol{H}_v; \boldsymbol{H}_q]$, corresponding to instruction, visual, and query tokens, respectively. Multimodal token interactions are governed by the attention mechanism:

$$\boldsymbol{Q} = \boldsymbol{W}_Q \boldsymbol{H}, \quad \boldsymbol{K} = \boldsymbol{W}_K \boldsymbol{H}, \quad \boldsymbol{V} = \boldsymbol{W}_V \boldsymbol{H}, \tag{1}$$

$$\boldsymbol{A} = \text{Softmax}\left(\frac{\text{PE}(\boldsymbol{Q}) \cdot \text{PE}(\boldsymbol{K})^{\text{T}}}{\sqrt{d_k}} + \boldsymbol{M}_C\right)\boldsymbol{V}, \tag{2}$$

where $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$ are learnable projection matrices, d_k is the dimensionality of each attention head, and $\boldsymbol{M}_C$ is a causal mask used for autoregressive decoding. The operator $\text{PE}(\cdot)$ denotes the position encoding function. The attention output is passed through a feed-forward network with residual connections and layer normalization to produce updated hidden states. At the final layer, a vocabulary distribution is computed to predict the next token.

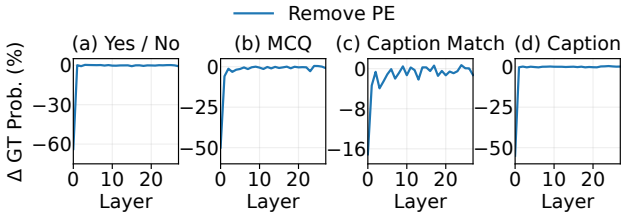

Figure 2: Effect of layer-wise PE ablation. Each point shows the change in ground-truth answer probability when the PE is removed from a layer. An obvious drop appears only at the first layer, while later layers show minimal impact.

3.3 Evaluation

For analysis experiments, we assess the model's temporal sensitivity by tracking changes in the predicted probability of the ground-truth token after incorporating certain perturbations:

$$P_C = \tilde{\boldsymbol{P}}_{\text{next}}(\boldsymbol{t}_{\text{gt}}) - \boldsymbol{P}_{\text{next}}(\boldsymbol{t}_{\text{gt}}), \tag{3}$$

where $\tilde{\boldsymbol{P}}_{\text{next}}$ is the output probability distribution under a perturbed setting, and $\boldsymbol{P}_{\text{next}}$ is the distribution from the base setting. A larger P_C reflects a stronger influence of the perturbation on temporal reasoning. For downstream application experiments, we report the standard accuracy.

3.4 Implementation Detail

Our experiments are implemented using PyTorch with the Transformers library. For video input processing, we uniformly sample 8 frames from each video. For Qwen2.5-VL, its internal frame merging mechanism results in 4 effective frames per video. To ensure a controlled evaluation setting, we constrain the model to generate only a single output token, and we analyze its probability distribution to measure temporal effects. Additional experimental configurations are detailed in relevant sections and the appendix.

4 Is Positional Encoding the Key to Temporal Understanding?

Positional encodings (PEs) are widely considered essential for modeling temporal relationships in VideoLMs, especially in recent architectures like Qwen2.5-VL that adopt advanced 3D encodings. The belief that PEs are a primary driver of temporal understanding has spurred research into more sophisticated PE designs. However, it remains unclear whether these encodings are truly the main source of temporal information. In this section, we conduct experiments that isolate the role of PEs and test the hypothesis that temporal information may instead emerge from the causal attention mechanism's inherent sensitivity to token order.

4.1 Layer-wise Temporal Sensitivity to PEs

Experiment. To probe the importance of PEs in temporal information extraction, we first quantify their contribution at different layers of the model. Specifically, we remove the PE terms from Equation 2 at one layer at a time, leaving all other components unchanged. This allows us to isolate the effect

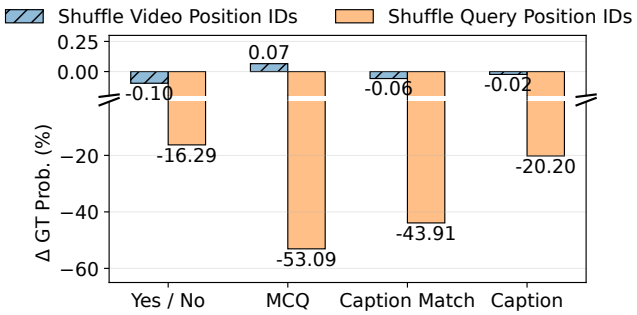

Figure 3: Effect of modality-specific position ID shuffling in the first layer. Shuffling query position IDs leads to significant performance drops, while shuffling video position IDs has minimal effect, highlighting the dominant role of PEs in the textual rather than the intended video modality.

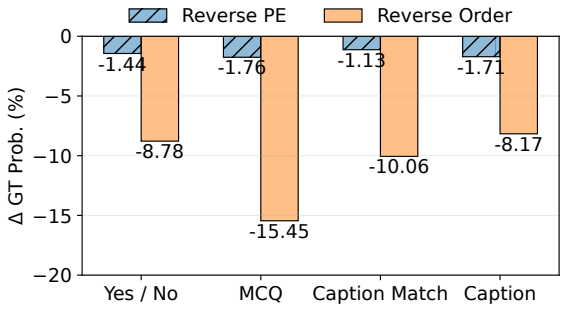

Figure 4: Effect of reversing position IDs versus frame order. Reversing the frame order results in significantly larger performance drops across all task types, suggesting that temporal understanding is primarily driven by the inherent order of frames rather than by positional encodings.

of PEs on temporal modeling at each layer, and to quantify where and how positional signals are most influential.

Results. As shown in Figure 2, the model exhibits minimal sensitivity to the removal of PEs across most layers. Across all task types, ablating PEs from intermediate and deeper layers results in minimal performance change, typically within ±2%. This suggests that these layers are relatively insensitive to the presence of explicit positional signals. A notable exception occurs in the first layer, where removing PEs causes a substantial drop in predicted probability of the correct token, up to 60%. **This stark contrast reveals that the impact of PEs for temporal understanding is restricted to the first layer of the model, as later layers exhibit minimal sensitivity to their presence or absence.**

4.2 Modality-Specific Temporal Roles of PEs

Experiment. Building on the observed drop in ground-truth probability when PEs are removed from the first layer, we further examine their role in temporal understanding through a targeted, modality-specific experiment. Specifically, we investigate whether the contribution of PEs for temporal understanding differs across input modalities. To this end, we independently shuffle the position IDs of either the video tokens or the query tokens in the first layer, while keeping the other modality unchanged. This setup enables us to directly assess the different impact of PEs on video and textual modalities during temporal modeling.

Result. As shown in Figure 3, the results reveal a clear and consistent disparity across modalities. Shuffling the position IDs of query tokens in the first layer significantly degrades performance across all task types. The most pronounced drops occur in the "Multiple Choice" task, with ground-truth probability decreasing by 53.09%. In contrast, shuffling the position IDs of video tokens leads to negligible effects. For instance, the performance decreases by only 0.02% in "Captioning" and even increases slightly by 0.07% in "Multiple Choice", suggesting that the model is largely insensitive to positional distortions in the video modality at this stage. **Although special PEs are designed to capture temporal structure in video inputs, their impact appears to be concentrated in the textual modality.**

4.3 PE vs. Order in Temporal Modeling

Experiment. Previous results indicate that PEs have limited influence on capturing temporal information for the video modality, prompting a key question: if not positional encodings, what enables VideoLMs to perform effective temporal understanding? We hypothesize that temporal understanding is primarily driven by the order of input video frames, rather than from explicit positional signals. To test this, we compare two interventions applied across all layers during inference: (1) Reverse PE: we reverse the temporal axis of the position IDs while keeping the input frame order unchanged; and (2) Reverse Order: we reverse the input frame order while preserving the original position IDs. For example, given input frames 1 to N with temporal position IDs 1 to N, "Reverse PE" uses frames 1 to N with position IDs N to 1, while "Reverse Order" uses frames N to 1 with position IDs N to 1.

Result As shown in Figure 4, reversing the position IDs results in only a modest drop in the predicted ground-truth probability, with changes ranging from -1.13% to -1.76% across task types. In contrast, reversing the input frame order leads to a significant performance decline. The largest drop is observed in the "Multiple Choice" task, with a decrease of 15.45%, and consistent trends are observed across the remaining task types. **These results demonstrate that the order of video frames plays a more crucial role than positional encodings in temporal understanding.**

5 Temporal Information Emerges from Attention-Based Causal Interactions

In Section 4, we show that positional encodings have limited influence on temporal understanding, particularly in the video modality. In contrast, altering the order of input frames has a more pronounced effect. These findings suggest that temporal information is not primarily captured through explicit positional signals, but instead emerges from the model's processing of sequential input.

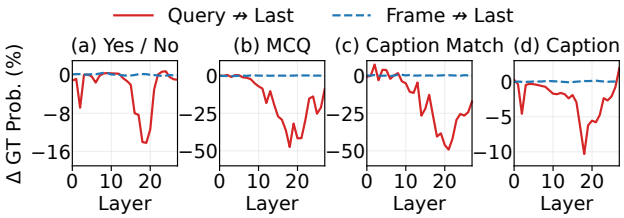

Figure 5: Effect of blocking attention from the final token to the input sources. Disabling query access causes major performance drops, while blocking video tokens has little effect. This highlights the dominant role of query tokens in final-stage temporal understanding.

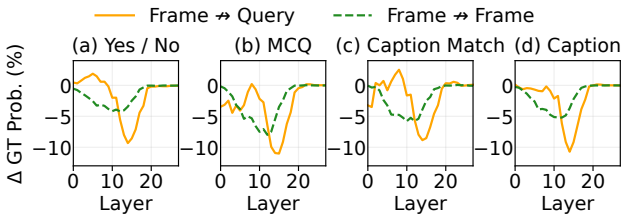

Figure 6: Effect of blocking inter-frame and frame-to-query attention. Probability drops from inter-frame blocking in early layers indicate where temporal information begins to form within the video stream. In contrast, drops resulting from frame-to-query blocking in later layers reveal when this temporal information flows to the query.

Given the causal architecture of current VideoLMs, we hypothesize that this order sensitivity arises from the attention mechanism, specifically due to the use of causal masking. Within this structure, temporal representations may be constructed progressively across layers as each token aggregates information from earlier positions.

To investigate this hypothesis, we examine how causal attention mechanisms facilitate the step-by-step construction and propagation of temporal information throughout the model. Drawing inspiration from Zhang et al. (2025b), we adopt a backward tracing approach, analyzing the model from its output back to earlier interactions. We identify where temporal information originates, how it flows between video and text modalities, and validate its causal nature through targeted analysis.

We adapt the attention knockout method (Geva et al. 2023) to the VideoLM setting in order to analyze this process. Specifically, we modify the attention computation in Equation 2 by incorporating an additional mask $\boldsymbol{M}$ that prevents target tokens in the set $\mathcal{T}$ from attending to source tokens in the set $\mathcal{S}$:

$$\tilde{\boldsymbol{A}} = \text{Softmax}\left(\frac{\text{PE}(\boldsymbol{Q}) \cdot \text{PE}(\boldsymbol{K})^{\text{T}}}{\sqrt{d_k}} + \boldsymbol{M}_C + \boldsymbol{M}\right)\boldsymbol{V}, \quad (4)$$

$$m_{i,j} = \begin{cases} -\infty, & \text{if } (i,j) \in \mathcal{T} \times \mathcal{S}, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

where $m_{i,j}$ denotes the (i,j)-th element of the mask matrix $\boldsymbol{M}$, controlling whether token i attends to token j in the attention computation. To localize where temporal information emerges and flows, we apply this knockout iteratively across layers, using a sliding window of $k = 5$ layers to reduce noise and improve stability.

5.1 Query-Driven Temporal Prediction

Experiment. To better understand the full process, we begin by identifying which input tokens directly contribute to the model's temporal understanding at the output. We perform an attention knockout to the final answer token, blocking its ability to attend to either the query tokens or the video frame tokens. Formally, we set $\mathcal{T} = \{i_{\text{last}}\}$, where i_{last} is the index of the final output token, and set $\mathcal{S}$ to the indices of either the query tokens or the video tokens, depending on the modality being ablated. This allows us to isolate which modality serves as the immediate source of temporal information during model prediction.

Result. Figure 5 reveals a clear asymmetry in the effects of blocking attention from the final output token to different input modalities. Disabling access to frame tokens has minimal impact, with ground-truth probabilities remaining near zero across all layers, suggesting that visual features are not directly referenced during final decoding. In contrast, blocking access to query tokens causes substantial and consistent performance drops, particularly in deeper layers, exceeding 40% in "Multiple Choice" and "Caption Matching" tasks. **These results indicate that temporal understanding at the output stage is predominantly carried by query tokens, which likely synthesize temporal cues accumulated earlier in the model, as further validated in the following sections. In contrast, video tokens play only a minimal direct role in the final prediction.**

5.2 Stage-wise Temporal Integration

Experiments. Although final predictions primarily rely on the query, temporal understanding necessarily requires visual information to recognize patterns of change, motion, and progression over time. To pinpoint when and how video frames contribute to this process, we design two complementary attention knockout setups: (1) Frame-to-query: we prevent query tokens from attending to video frame tokens, (2) Inter-frame: we prevent each video frame from attending to earlier frames. These interventions allow us to isolate two distinct stages in the temporal information pathway: the early construction of temporal relationships through inter-frame attention, and the later integration of this information into the query representation.

Result. As shown in Figure 6, the two interventions yield distinct layer-wise effects. Blocking frame-to-query attention leads to a sharp decline in ground-truth probability around layer 15, with the most pronounced drop over 10% in the "Captioning" task. This indicates that the middle layers play a key role in integrating temporal information into the query representation. In contrast, blocking inter-frame attention causes earlier degradation, peaking near layer 10, with probability drops ranging from 4% to 6%, suggesting that early layers are critical for forming temporal rela-

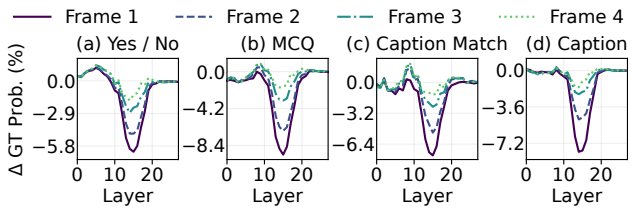

Figure 7: Effect of restricting query attention to single frames. Performance drops sharply when the query can only attend to early frames, but shows minimal drop when attending to later frames. This supports that temporal information flows forward and accumulates in the last frame.

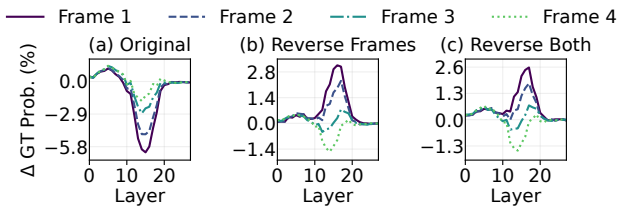

Figure 8: Effect of reversing video order in the "Yes or No" task. Attribution patterns consistently flip regardless of position IDs, suggesting that the model captures temporal information primarily through causal inference, instead of simple visual content collection.

tionships across frames. **Together, these observations indicate a two-stage temporal processing process: early layers construct inter-frame temporal relationships, which are then integrated into the query by middle layers to support temporal reasoning in later decoding.**

5.3 Last-Frame-Centered Temporal Propagation

Experiment. Building on the two-stage process where temporal cues are first constructed via inter-frame interaction and then integrated into the query, we now investigate whether certain frames play a more prominent role in transmitting temporal information to the query. To do this, we constrain the model such that the query tokens are only allowed to attend to the visual tokens from a single frame. In our setting, each video contains 4 frames. Formally for each frame $t \in \{1, 2, 3, 4\}$, we enforce this by defining the set $\mathcal{T}$ as all query token indices, and the set $\mathcal{S} = \{i_v \mid i_v \in \mathcal{I}_v \setminus \mathcal{I}_{v_t}\}$, where $\mathcal{I}_v$ is the set of all visual token indices and $\mathcal{I}v_t$ denotes those corresponding to frame t.

Result. Figure 7 illustrates the effect of restricting the query's attention to a single frame. Across all tasks, the largest drop in ground-truth probability occurs when the query attends only to early frames, particularly Frame 1. For instance, in the "Multiple Choice" task, this results in a decline of over 8%, indicating that early frames lack sufficient temporal cues for accurate prediction. In contrast, attending only to later frames, especially Frame 4, yields minimal performance loss, typically below 2%. This suggests that temporal information accumulates across frames, with the final frame serving as an aggregation point. As a result, the query performs well when attending solely to the last frame, but suffers when limited to early ones. **Overall, rather than evenly extracting and comparing information for temporal understanding, the query primarily accesses aggregated temporal cues by attending to the final, information-rich frame.**

5.4 Beyond Correlation to Causality

Experiment. The previous experiment suggests that the last frame serves as an aggregation point for temporal cues. However, it remains unclear whether this aggregation results from causal processing or merely from content accumulation regardless of frame order. To probe this, we leverage the "Yes or No" task, where reversing the video order flips the correct answer. We repeat the single-frame attribution analysis on reversed sequences to examine whether the model exhibits causal aggregation. If the model simply aggregates visual information, the query should still favor the last frame, as it continues to contain the most information. To further rule out PEs as a confounding factor, we also reverse both the frame order and the position IDs, ensuring that positional signals remain aligned with the original frame sequence.

Result. As shown in Figure 8, reversing the video frame order leads to a complete inversion of the attribution pattern observed in the previous experiment. Only attending to the initial frame, which previously had the most negative impact, now yields the most positive contribution, even increasing the ground-truth probability. Conversely, the last frame, which had minimal influence before, now causes the greatest drop, despite containing the most visual information. This inversion persists even when position IDs are also reversed, restoring the original positional alignment. **These observations confirm that the model performs order-sensitive inference through the causal attention mechanism, rather than merely accumulating information.**

5.5 Spatiotemporal Assembly across Frames

Experiments. In the last experiment, we move beyond analyzing when and how temporal information flows to the query, and instead investigate how it is causally constructed across frames. Specifically, we examine how later visual content interacts with earlier frames to capture temporal dynamics. Using the prior setup where the query attends only to Frame 4 as a baseline, we isolate how temporal information is encoded within the final frame. We compare three attention configurations to assess the inter-frame construction of temporal information from a spatiotemporal perspective: (1) Corresponding Area: for each frame, attention is restricted to spatially aligned regions in all preceding frames, meaning that each token only attends to tokens located at the same or neighboring spatial positions. As a result, the final frame attends to approximately 300 tokens, while earlier frames attend to fewer; (2) Previous Frame: each frame attends to the immediately preceding frame with about 300 tokens; and (3) Corresponding Area in Previous Frame: each frame attends to the spatially aligned region in the immediately preceding frame with 100 tokens.

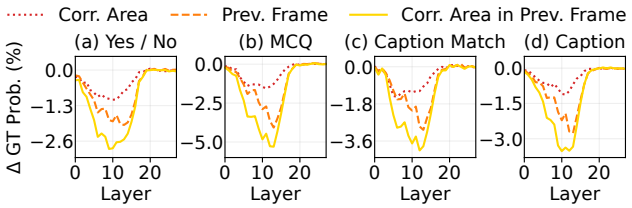

Figure 9: Effect of restricting each frame's attention to examine inter-frame spatiotemporal contributions to temporal understanding. Sparse long-range attention performs better than dense short-range attention. Early layers prefer global spatial context, and later layers focus on local detail.

Method	TempCompass				NExT-QA
	Yes/No	MCQ	Caption	Match	
Baseline	70.8	66.5	59.0	57.4	75.1
(1)	67.9	64.1	57.6	55.0	71.9
(2)	70.5	66.5	59.3	57.4	75.2
(3)	70.7	66.6	59.0	57.2	75.1

Table 2: Accuracy on TempCompass and NExT-QA under different efficiency strategies. Strategies (2) and (3) preserve performance while reducing computational or memory cost. Strategy (1) introduces a slight but acceptable drop, with potential for further optimization.

Result. Figure 9 shows that "Corresponding Area" yields the most stable performance, with ground-truth probability drops consistently under 1% across all layers. Surprisingly, "Previous Frame" leads to a larger drop, despite involving more token interactions altogether. Moreover, when comparing "Previous Frame" and "Corresponding Area in Previous Frame" with equivalent temporal coverage, early layers exhibit a preference for global spatial context. However, this advantage fades in deeper layers, indicating a transition toward local detail processing. **Overall, sparse long-range attention process inter-frame temporal integration better than dense short-range interaction, while spatial processing shifts from global to local focus over layers.**

6 Temporal Pathway-Guided Inference

Based on our findings that temporal information in VideoLMs arises from a causal pathway, starting from long-range inter-frame interactions and flowing into the query via specific multimodal interaction, we propose two inference strategies that improve efficiency. Both reduce computational and memory costs by pruning token interactions that contribute little to the final temporal representation:

- **Staged Modality Interaction for Sparse Attention**: Temporal information in VideoLMs is progressively constructed, with different layers contributing in distinct ways to the final prediction. Motivated by this observation, we design sparse attention strategies that selectively disable token interactions with limited impact on temporal modeling. For example, (1) in middle layers, query tokens can be restricted to attend only to the last video frame, and (2) in deeper layers, inter-frame interactions can be reduced. This staged sparsification lowers FLOPs while maintaining performance, providing a scalable and effective solution for efficient inference.
- **Temporal Exit for KV Cache Compression**: During inference, the key-value (KV) cache grows with sequence length, becoming a major memory bottleneck. However, our analysis shows that not all tokens remain relevant throughout all layers. Leveraging this insight, we can design pruning strategies that remove tokens from the KV cache once their contribution to the final prediction diminishes. For instance, (3) in deeper layers, the large number of frame tokens can be safely discarded. This reduces memory usage without sacrificing performance, providing an effective and compatible solution for improving efficiency in autoregressive generation.

To evaluate the effectiveness of our proposed strategies, we conduct validation experiments on the TempCompass and NExT-QA datasets using Qwen-VL-Chat-2.5-7B. The implementation details for the three strategies are as follows: (1) each query token is restricted to attend only to the last frame token in layers 10–20, (2) attention between frame tokens is disabled in layers 20–28, and (3) all frame tokens are removed from the KV cache in layers 20–28.

As shown in Table 2, strategies (2) and (3) achieve performance comparable to the baseline across all task types, with accuracy differences typically within 0.2%. In some cases, they even slightly outperform the baseline. This indicates that both computation and memory can be reduced without sacrificing performance. In contrast, strategy (1), which restricts query attention to only the last frame, leads to a slightly larger accuracy decline. Nonetheless, it remains promising and may benefit from further refinement based on our insightful findings.

7 Conclusion

This work investigates how modern VideoLMs achieve temporal understanding, a fundamental capability that remains insufficiently explored. While PEs are commonly assumed to be crucial for encoding temporal representations, our analysis shows they contribute only marginally to capturing temporal information in the video input. In contrast, reversing the video sequence while preserving positional signals aligned with the original order results in a significant drop in performance, suggesting that temporal information arises elsewhere. Our findings reveal that temporal understanding emerges from the causal attention mechanism's order-sensitive structure. Specifically, temporal information is constructed through a causal pathway: it is causally aggregated across frames, integrated into the final frame, and subsequently refined within the query. Building on this insight, we propose two efficiency-focused strategies and validate their effectiveness on benchmark datasets. We hope this work lays the foundation for better temporal modeling in VideoLMs, clarifies key challenges, and encourages further exploration of their internal mechanisms.

Acknowledgments

This research is supported by the NTU Start-Up Grant (#023284-00001), Singapore, and the MOE AcRF Tier 1 Seed Funding Grant (#025041-00001, RS37/24).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Basu, S.; Grayson, M.; Morrison, C.; Nushi, B.; Feizi, S.; and Massiceti, D. 2024. Understanding information storage and transfer in multi-modal large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 7400–7426.

Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Biran, E.; Gottesman, D.; Yang, S.; Geva, M.; and Globerson, A. 2024. Hopping Too Late: Exploring the Limitations of Large Language Models on Multi-Hop Queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 14113–14130.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 961–970.

Cai, M.; Tan, R.; Zhang, J.; Zou, B.; Zhang, K.; Feng, Y.; Zhu, F.; Gu, J.; Zhong, Y.; Shang, Y.; et al. 2024. TemporalBench: Benchmarking Fine-grained Temporal Understanding for Multimodal Video Models. In *Proceedings of the Workshop on Video-Language Models at NeurIPS 2024*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Choudhury, R.; Zhu, G.; Liu, S.; Niinuma, K.; Kitani, K. M.; and Jeni, L. 2024. Don't Look Twice: Faster Video Transformers with Run-Length Tokenization. *arXiv preprint arXiv:2411.05222*.

Chung, J.; Zhu, T.; Saez-Diez, M. G.; Niebles, J. C.; Zhou, H.; and Russakovsky, O. 2025. Unifying Specialized Visual Encoders for Video Language Models. In *Forty-second International Conference on Machine Learning*.

Fateh, F. J.; Ahmed, U.; Khan, H.; Zia, M. Z.; and Tran, Q.-H. 2024. Video LLMs for Temporal Reasoning in Long Videos. *arXiv preprint arXiv:2412.02930*.

Fu, T.; Liu, T.; Han, Q.; Dai, G.; Yan, S.; Yang, H.; Ning, X.; and Wang, Y. 2024. FrameFusion: Combining Similarity and Importance for Video Token Reduction on Large Visual Language Models. *arXiv preprint arXiv:2501.01986*.

Ge, J.; Chen, Z.; Lin, J.; Zhu, J.; Liu, X.; Dai, J.; and Zhu, X. 2024. V2pe: Improving multimodal long-context capability of vision-language models with variable visual position encoding. *arXiv preprint arXiv:2412.09616*.

Geva, M.; Bastings, J.; Filippova, K.; and Globerson, A. 2023. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12216–12235.

Goldshmidt, R.; and Horovicz, M. 2024. Tokenshap: Interpreting large language models with monte carlo shapley value estimation. *arXiv preprint arXiv:2407.10114*.

Golovanevsky, M.; Rudman, W.; Palit, V.; Eickhoff, C.; and Singh, R. 2025. What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline for Gaussian-Noise-free Text-Image Corruption and Evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 11462–11482.

He, B.; Li, H.; Jang, Y. K.; Jia, M.; Cao, X.; Shah, A.; Shrivastava, A.; and Lim, S.-N. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13504–13514.

Hu, Z.-Y.; Zhong, Y.; Huang, S.; Lyu, M.; and Wang, L. 2024. Enhancing Temporal Modeling of Video LLMs via Time Gating. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2845–2856.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, L.; Liu, Y.; Yao, L.; Zhang, P.; An, C.; Wang, L.; Sun, X.; Kong, L.; and Liu, Q. 2025. Temporal Reasoning Transfer from Text to Video. In *The Thirteenth International Conference on Learning Representations*.

Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122*.

Liu, J.; Wang, Y.; Ma, H.; Wu, X.; Ma, X.; Wei, x.; Jiao, J.; Wu, E.; and Hu, J. 2024a. Kangaroo: A Powerful Video-Language Model Supporting Long-context Video Input. *arXiv preprint arXiv:2408.15542*.

Liu, Y.; Li, S.; Liu, Y.; Wang, Y.; Ren, S.; Li, L.; Chen, S.; Sun, X.; and Hou, L. 2024b. TempCompass: Do Video LLMs Really Understand Videos? In *Findings of the Association for Computational Linguistics ACL 2024*, 8731–8772.

Liu, Z.; Guo, L.; Tang, Y.; Yue, T.; Cai, J.; Ma, K.; Liu, Q.; Chen, X.; and Liu, J. 2025. Vrope: Rotary position embedding for video large language models. *arXiv preprint arXiv:2502.11664*.

Nguyen, H.; Tran, D.; Hoang, H.; Nguyen, P.; and Narayanan, S. 2025. MOOSE: Pay Attention to Temporal Dynamics for Video Understanding via Optical Flows. *arXiv preprint arXiv:2506.01119*.

Nie, M.; Ding, D.; Wang, C.; Guo, Y.; Han, J.; Xu, H.; and Zhang, L. 2024. Slowfocus: Enhancing fine-grained temporal understanding in video llm. *Advances in Neural Information Processing Systems*, 37: 81808–81835.

OpenAI. 2024. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/.

Plizzari, C.; Tonioni, A.; Xian, Y.; Kulshrestha, A.; and Tombari, F. 2025. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24129–24138.

Shangguan, Z.; Li, C.; Ding, Y.; Zheng, Y.; Zhao, Y.; Fitzgerald, T.; and Cohan, A. 2025. TOMATO: Assessing Visual Temporal Reasoning Capabilities in Multimodal Foundation Models. In *The Thirteenth International Conference on Learning Representations*.

Shi, Y.; Long, Q.; and Wang, W. 2025. Static or Dynamic: Towards Query-Adaptive Token Selection for Video Question Answering. *arXiv preprint arXiv:2504.21403*.

Wang, L.; Li, L.; Dai, D.; Chen, D.; Zhou, H.; Meng, F.; Zhou, J.; and Sun, X. 2023. Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9840–9855.

Wei, X.; Liu, X.; Zang, Y.; Dong, X.; Zhang, P.; Cao, Y.; Tong, J.; Duan, H.; Guo, Q.; Wang, J.; et al. 2025. VideoRoPE: What Makes for Good Video Rotary Position Embedding? In *Forty-second International Conference on Machine Learning*.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*.

Yu, Z.; and Ananiadou, S. 2024. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering. *arXiv preprint arXiv:2411.10950*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025a. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, P.; Zhang, K.; Li, B.; Zeng, G.; Yang, J.; Zhang, Y.; Wang, Z.; Tan, H.; Li, C.; and Liu, Z. 2024. Long Context Transfer from Language to Vision. *arXiv preprint arXiv:2406.16852*.

Zhang, Z.; Yadav, S.; Han, F.; and Shutova, E. 2025b. Cross-modal information flow in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19781–19791.

Zhao, H.; Ji, G.-P.; Yan, R.; Xiong, H.; and Li, Z. 2025. VideoExpert: Augmented LLM for Temporal-Sensitive Video Understanding. *arXiv preprint arXiv:2504.07519*.

Zhao, H.; Yang, F.; Shen, B.; Lakkaraju, H.; and Du, M. 2024a. Towards uncovering how large language model works: An explainability perspective. *arXiv preprint arXiv:2402.10688*.

Zhao, L.; Gundavarapu, N. B.; Yuan, L.; Zhou, H.; Yan, S.; Sun, J. J.; Friedman, L.; Qian, R.; Weyand, T.; Zhao, Y.; et al. 2024b. VideoPrism: A Foundational Visual Encoder for Video Understanding. In *International Conference on Machine Learning*, 60785–60811. PMLR.