The Fourteenth International Conference on Learning Representations ICLR 2026

Map the Flow: Revealing Hidden Pathways of Information in VideoLLMs

Minji Kim [1★] Taekyung Kim [2★] Bohyung Han [1]

[1] Seoul National University
[2] NAVER AI Lab
★ Equal Contribution

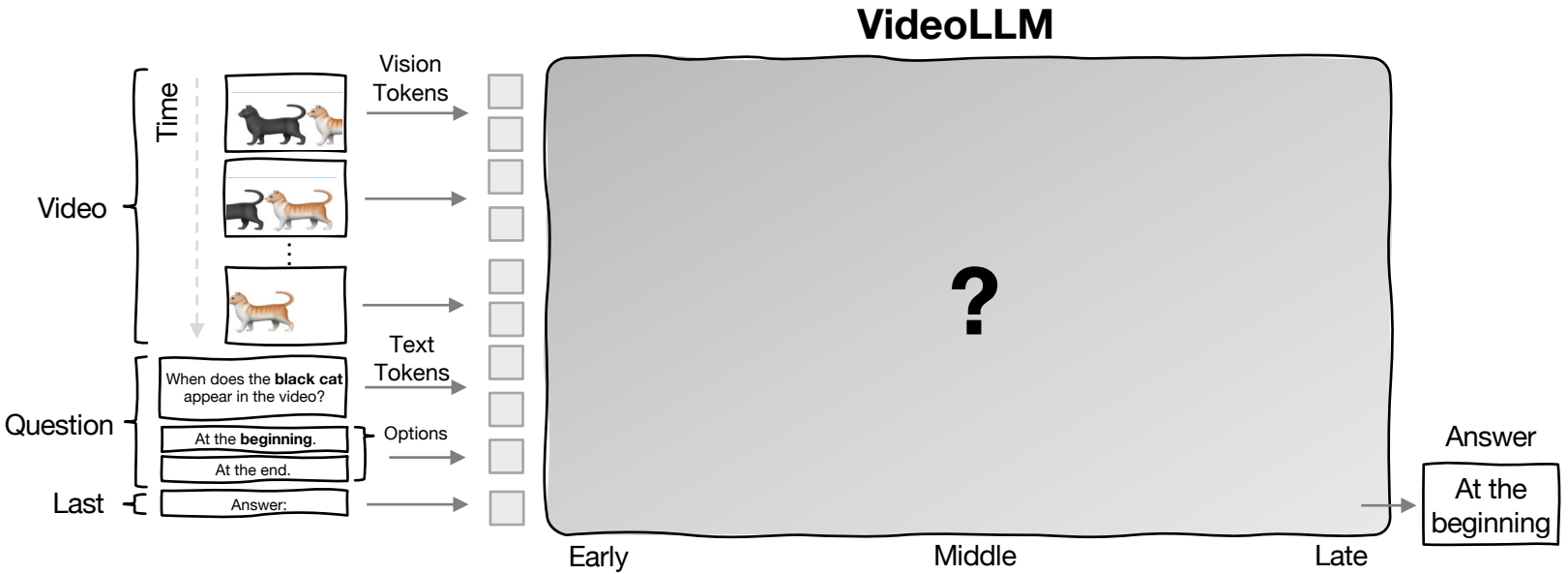
ComputerVisionLab
Seoul National University

NAVER AI LAB

ICLR

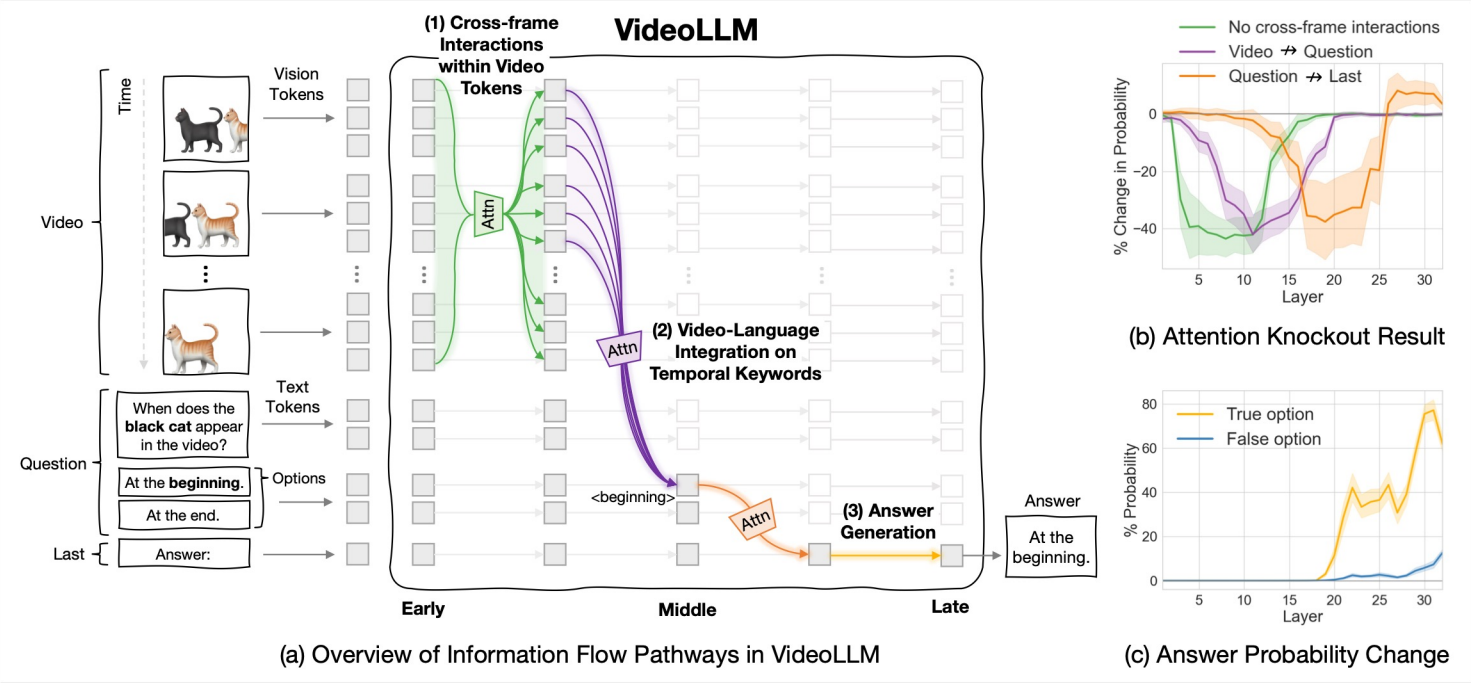
Motivation

- Despite recent progress in VideoLLMs for video question answering (VideoQA),
- We lack understanding of their **internal mechanisms**: **where** and **how** do they extract and propagate temporal information from videos to generate answer?



Overview

- This paper presents a **mechanistic analysis** on **where** and **how** information flows in VideoLLMs across layers and modalities
- Our finding: structured **3-stage information flow** for temporal reasoning



(a) Overview of Information Flow Pathways in VideoLLM

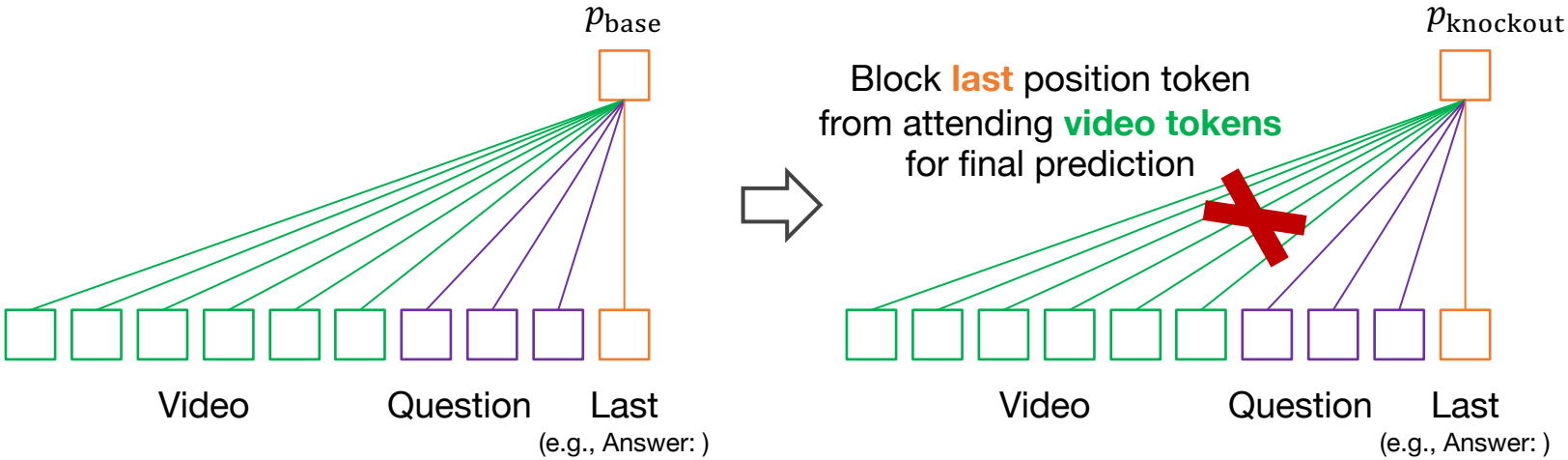
(b) Attention Knockout Result

(c) Answer Probability Change

Experimental Setup

- **Attention Knockout on** TVBench (temporal reasoning tasks)
  - Disconnects attention pairs and tracks the **drop in probability of the final answer** to quantify their impact
  - Layer & token group sweeping → Larger drop, larger impact

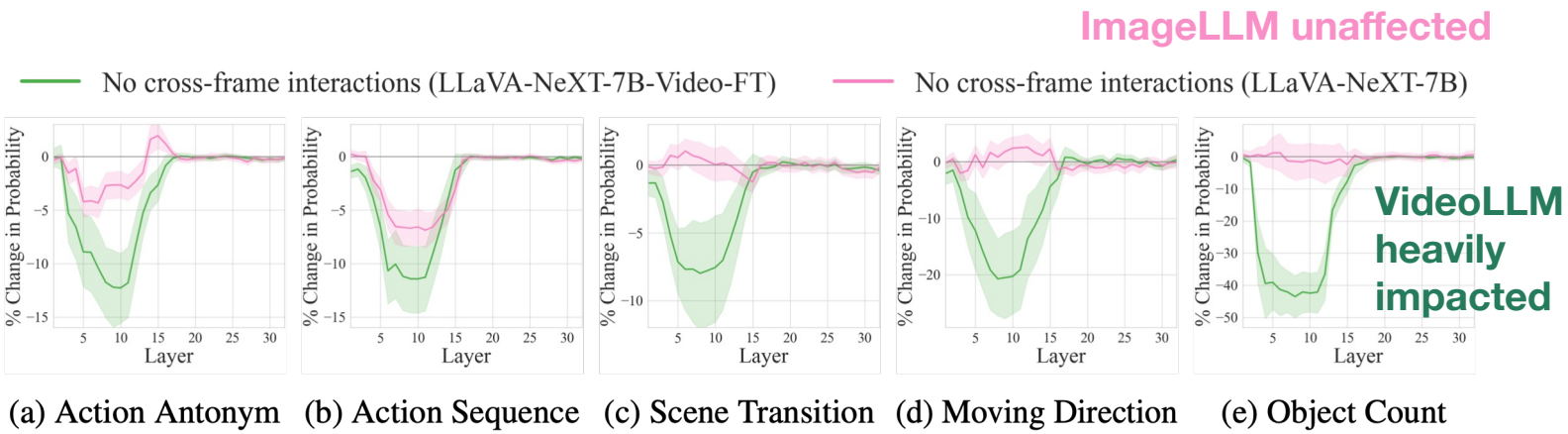
$$\text{Relative Change} = ((p_{\text{knockout}} - p_{\text{base}})/p_{\text{base}}) \times 100$$



Phase 1 · Spatiotemporal Encoding in Early-to-Middle Layers

How do VideoLLMs encode spatiotemporal information from flattened sequence of video tokens?

- ImageLLM vs. VideoLLM:
  VideoQA fine-tuning boosts cross-frame interactions in early-to-middle layers.



Phase 1 · Spatiotemporal Encoding in Early-to-Middle Layers

Impact of early-stage temporal interactions on answer generation

- Blocking cross-frame attention in first half layers causes incorrect or opposite answers.

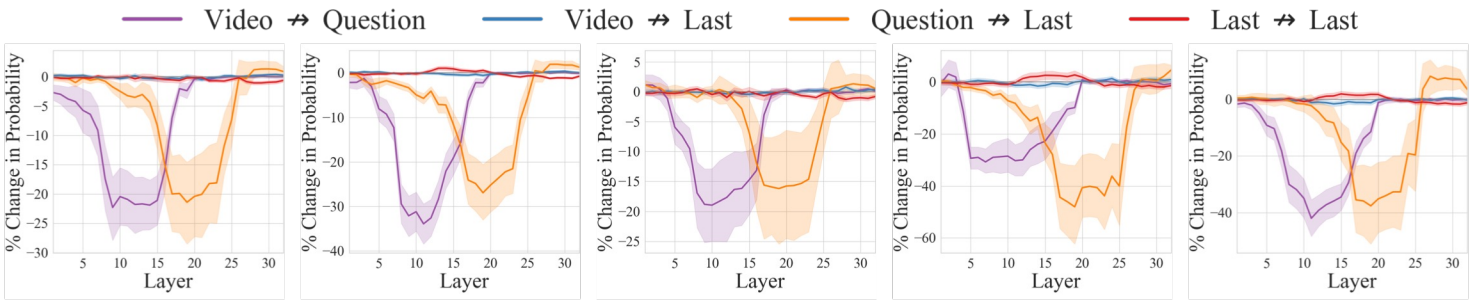
Task	Acc Drop	Answer Example
Action Antonym	−24.1%	*Baseline*: The action being performed in the video is to stand up . *Knockout*: The action being performed in the video is to sit on a chair .
Action Sequence	−20.2%	*Baseline*: The action the person is doing first is to open the plastic bag . *Knockout*: 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.

→ VideoLLMs rely heavily on cross-frame interactions in the early stage to reason about temporal events

Phase 2 · Video-to-Language Integration in Middle Layers

How are the temporal concepts in the question extracted from video tokens and propagated to text tokens?

- Overall cross-modal information flow follows Video → Question → Last, rather than direct transfer from Video to Last.

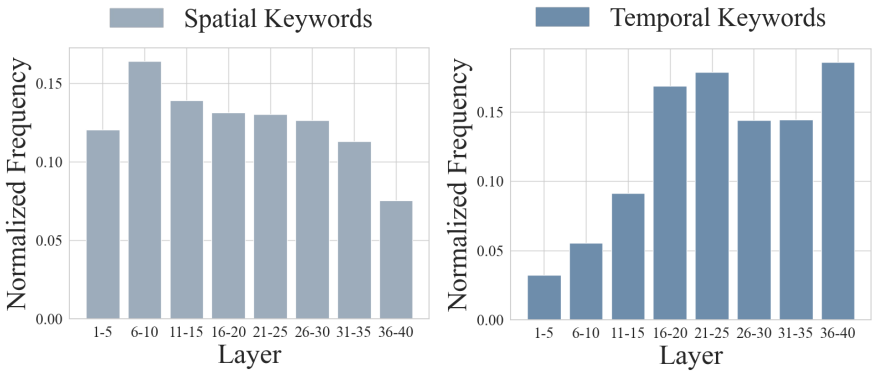


(a) Action Antonym (b) Action Sequence (c) Scene Transition (d) Moving Direction (e) Object Count

Phase 2 · Video-to-Language Integration in Middle Layers

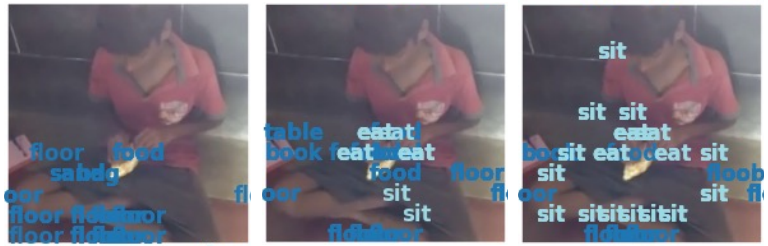
Emergence of spatial and temporal concepts in video tokens

- Analyzing semantic concepts via **Logit Lens** (video token probing via language model head) → **temporal concepts emerge later than spatial concepts**



Spatial grounding first

Temporal dynamics encoded afterward

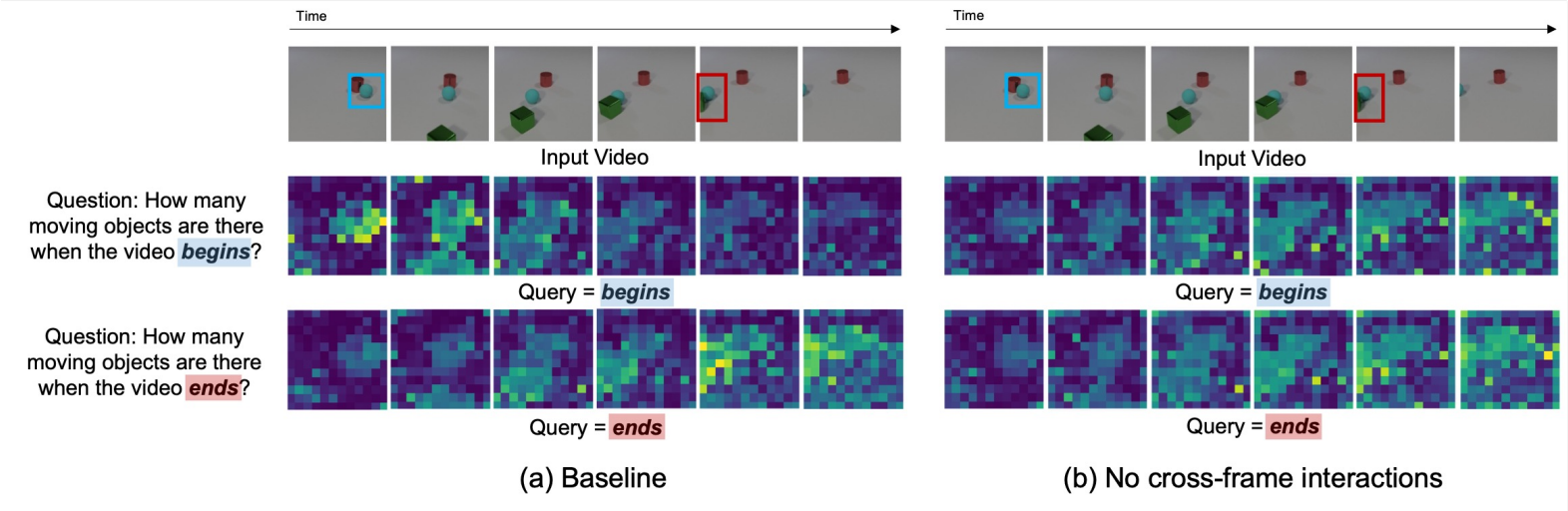


Early Layer | Middle Layer | Late Layer

Phase 2 · Video-to-Language Integration in Middle Layers

Video-language alignment enables selective spatiotemporal propagation

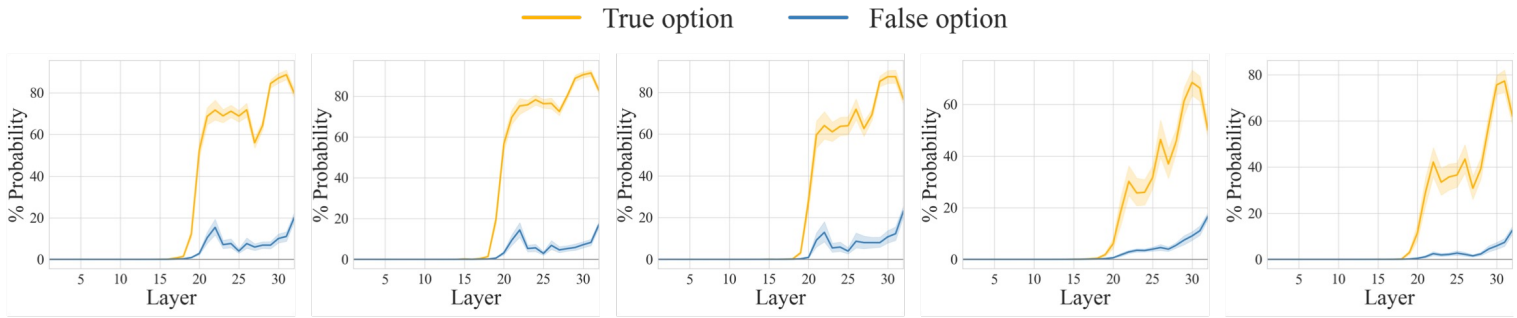
- How are the emergent concepts in videos propagated through text tokens?
- (1) Temporal visual information is aligned with temporal concept vocabularies
  (2) Such alignment emerges specifically through cross-frame interactions



Phase 3 · Answer Generation in Middle-to-Late Layers

When does the model become ready to generate an answer?

- Layer-wise answer probability rises sharply after video-language integration, indicating that the model is **ready to predict correct answers** after middle layers.



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

Practical Implications · Attention Pruning

Effective information flow pathways for solving VideoQA tasks

- Disabling all but the critical pathways retains performance comparable to full-attention baselines, e.g., **58% are prunable** in LLaVA-NeXT-7B-Video-FT.

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

Take-away

- VideoLLMs follow structured **3-stage information flow** for temporal reasoning
- Attention pruning with these **sparse** information pathways suffice for VideoQA

- Future implications of our findings
  - **Pathway regularization** during training to prevent shortcut learning
  - Establishing **temporal concepts in earlier layers** before V-L integration to reduce hallucinations
  - **Early-exit** strategies to reduce inference overhead


Thank
you