TRACING ARROW-OF-TIME: DIAGNOSING AND ADDRESSING TEMPORAL INFORMATION LOSS IN VIDEO-LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

The Arrow-of-Time (AoT) task requires distinguishing forward from backward video playback by recognizing temporally irreversible events. Prior work has revealed a substantial gap between humans and Video-LLMs on AoT in end-to-end evaluations. Here, we use AoT as a diagnostic lens to localize where this gap arises: do visual backbones fail to encode temporal information, or is such information lost when transferred to the LLM? We combine vision-encoder probing with downstream video question answering (VQA) to trace temporal information across the encoder, projector, and LLM. We find that video-centric encoders with explicit temporal modeling encode substantially stronger temporal signals than the evaluated frame-centric encoders. However, encoding alone is insufficient: projector design and token compression can disrupt temporal information before it reaches the LLM. Moreover, layer-wise analysis reveals the distinction between probe-level recoverability and LLM-level usability, underscoring the complementary roles of probing and VQA. Guided by these findings, we combine a temporally aware video encoder with a time-preserving projector and achieve 98.1% accuracy on AoT$_{PPB}$, surpassing human performance. Finally, we show that AoT supervision improves temporal reasoning across both short- and long-video benchmarks. Our results position AoT as both a diagnostic lens and an effective supervision for improving broader temporal reasoning in Video-LLMs.

1 INTRODUCTION

Modern Video-LLMs typically couple a vision encoder with an LLM through a projector. Their visual backbones range from frame-centric encoders, which process videos as sequences of individual frames (Bai et al., 2025a;b), to video-centric encoders that explicitly model temporal dynamics across frames (Wang et al., 2024; Bardes et al., 2024; Assran et al., 2025). As videos introduce substantially more visual tokens than images, recent work increasingly integrates token compression into the projector (McKinzie et al., 2024; Zhang et al., 2025; Li et al., 2025a; Cha et al., 2024b), enabling longer videos or more frames within a fixed context budget. Despite promising performance on broad video understanding benchmarks (Chen et al., 2024b; Mangalam et al., 2023; Xiao et al., 2021; Fu et al., 2025), current Video-LLMs remain weak at temporal reasoning, a key prerequisite for understanding real-world dynamics (Gao et al., 2025; Chow et al., 2025; Yu et al., 2026).

A particularly striking example is the Arrow of Time (AoT): the assumption that events unfold irreversibly from past to future under constraints such as gravity, entropy, and causality (Eddington & Whittaker, 1928; Layzer, 1975). AoT operationalizes this principle as a binary task: given a short video, determine whether it is playing forward or backward (Figure 1(a)). Because forward and reversed videos share the same objects, actions, and semantic content, AoT isolates sensitivity to asymmetric moment-to-moment dynamics and provides a controlled lens for assessing motion-level, temporally asymmetric visual evidence. Beyond temporal reasoning in Video-LLMs, AoT has also attracted attention in fields such as physical AI (Azzolini et al., 2025; Schmidt & Jiang, 2024).

Existing AoT datasets (Pickup et al., 2014; Wei et al., 2018a) typically lack human baselines and offer limited motion diversity. To address this, Matta et al. (2025) introduced AoT$_{PPB}$, a psychophysics-grounded benchmark spanning diverse motion categories and enabling direct comparison with human

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

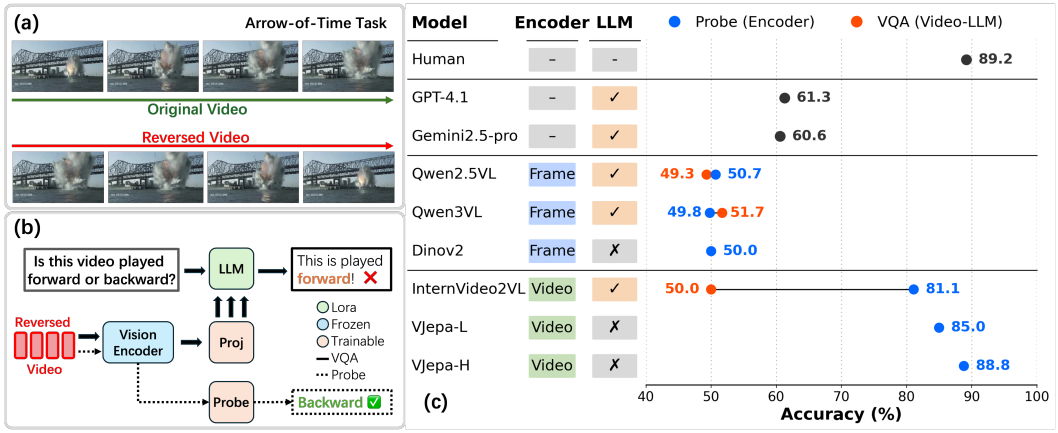


Figure 1: **(a) The AoT task**: determining whether a video clip is played forward or backward by recognizing physical irreversibility. **(b) Two AoT evaluation protocols**: probing frozen vision encoders with a lightweight classifier (probe), and fine-tuning Video-LLMs by recasting forward/backward videos as video question answering (VQA) instances. **(c) AoT performance discrepancy**: video-centric encoders perform strongly on AoT probing, whereas other models remain near chance level.

performance (Hanyu et al., 2023). Yet frontier Video-LLMs remain substantially below humans on AoT$_{PPB}$. Prior work mainly targets individual stages, such as training objectives (Xue et al., 2025; Bagad et al., 2023) or auxiliary temporal modules (Rasekh et al., 2026), leaving unclear how effectively temporal information is encoded and transferred across Video-LLM components. In this work, we use AoT$_{PPB}$ as a diagnostic benchmark and combine component-level probing with end-to-end VQA evaluation (Figure 1(b)) to localize temporal bottlenecks in Video-LLMs. We further investigate whether AoT can serve not only as a diagnostic task but also as an effective supervision for broader temporal reasoning.

We begin by addressing **RQ1: Can temporal information be encoded by the vision encoder?** To answer this question, we categorize vision encoders into two types: (1) *frame-centric encoders*, which primarily model individual frames and rely on positional mechanisms to represent temporal order (Bai et al., 2025a;b; Oquab et al., 2023); and (2) *video-centric encoders*, which explicitly model temporal dynamics across multiple frames (Tong et al., 2022; Wang et al., 2024; Bardes et al., 2024). Probing shows that explicit temporal modeling is critical for AoT: video-centric encoders consistently encode substantially stronger temporal signals than the evaluated frame encoders (Figure 1(c)).

However, strong temporal representations do not necessarily translate into strong VQA performance, suggesting that temporal information may be lost during transfer to the LLM. We therefore study **RQ2: How can temporal information be effectively transferred from video-centric encoders to LLMs?** We first compare projector architectures and observe substantial differences in temporal information transfer. To test whether these differences arise from visual information compression, we then fix the projector architecture and token budget and vary only whether compression preserves the temporal dimension. Both analyses show that preserving temporal information during vision-language projection is critical for downstream AoT performance.

In addition, we observe a performance gap between probing and VQA even when both use the same visual representations. This motivates a layer-wise analysis, which shows that temporal information is recoverable by probes from early encoder layers, whereas strong VQA performance is achieved only with later-layer representations. This contrast highlights their complementary roles: probing measures whether temporal information is encoded, while VQA measures whether it can be effectively used by the LLM. The results also suggest that CLIP-style vision-language alignment (Radford et al., 2021) may hinder temporal information transfer.

By tracing whether temporal information is encoded, transferred, and ultimately utilized, we provide a component-level explanation for why current Video-LLMs struggle with AoT. Guided by these findings, we combine a temporally aware video encoder with a time-preserving projector and an LLM, achieving 98.1% accuracy on AoT$_{PPB}$ and surpassing human performance.

Finally, we investigate **RQ3: Can AoT supervision improve Video-LLM performance on broader temporal reasoning tasks?** Our results show that AoT supervision provides an effective training signal. The resulting gains extend beyond AoT recognition to broader temporal reasoning, covering three additional short-video benchmarks and two long-video benchmarks. Overall, our findings position AoT as both a diagnostic lens for identifying temporal bottlenecks in Video-LLMs and an effective supervision signal for improving broader temporal reasoning.

2 Related Work

Video Large Language Models. Video-LLMs extend LLMs to dynamic visual inputs by coupling visual encoders with language models through projection layers, achieving strong performance across video understanding tasks (Chen et al., 2024b; Wang et al., 2023; Xiao et al., 2021; Fu et al., 2025). *Frame-centric encoders* process videos as individual frames and rely on temporal positional embeddings or token aggregation to represent temporal information (Bai et al., 2025a;b; Weng et al., 2024; Jin et al., 2024), whereas *video-centric encoders* explicitly model dynamics across frames (Tong et al., 2022; Wang et al., 2024; Chen et al., 2025). Projectors map visual features into the LLM embedding space (Ryoo et al., 2025; Zhang et al., 2023; Wang et al., 2024; Bai et al., 2025b; Cheng et al., 2024), while recent visual token compression techniques improve computational efficiency and enable longer-video processing (Zhang et al., 2025; Li et al., 2025a). Despite these advances, current Video-LLMs remain limited in temporal reasoning (Xue et al., 2025; Joseph et al., 2026; Shi et al., 2026; Cores et al., 2024). Our work studies this limitation by tracing how temporal information is encoded in visual backbones and transferred through projectors to the LLM.

Temporal Reasoning in Video-LLMs. General video benchmarks evaluate broad understanding abilities (Fu et al., 2025; Li et al., 2024a; Mangalam et al., 2023; Xiao et al., 2021), but several studies show limited sensitivity to temporal reasoning because many questions can be answered from static cues (Cores et al., 2024; Xue et al., 2025; Buch et al., 2022). This has motivated benchmarks targeting temporal localization, physical commonsense, and causality (Cores et al., 2024; Li et al., 2024b; Matta et al., 2025; Xue et al., 2025; Bordes et al., 2025). Among them, the AoT task (Hanyu et al., 2023; Xue et al., 2025; Pickup et al., 2014; Wei et al., 2018b) provides a direct diagnostic of temporal reasoning, yet frontier Video-LLMs remain substantially below human performance (Matta et al., 2025). Existing approaches mainly target individual stages: ArrowRL (Xue et al., 2025) and Test-of-Time (Bagad et al., 2023) improve training through reverse-aware reinforcement learning and contrastive training with reversed videos, while STAVEQ2 (Rasekh et al., 2026) enhances temporal feature extraction with dedicated temporal-attention modules. However, these works do not systematically trace how temporal information is encoded, preserved, and transferred across the Video-LLM pipeline, leaving architectural bottlenecks unclear and performance far below humans. We therefore use AoT as a diagnostic lens to study temporal information flow across Video-LLM components and further examine its potential as a supervision signal for broader temporal reasoning, while achieving beyond-human performance on the AoT task.

3 Preliminaries

The Arrow-of-Time (AoT) Task The AoT task assesses whether a model can infer the temporal direction of a video, i.e., whether a clip is played forward (F) or backward (B). Formally, let a forward video clip be a sequence of T frames $\mathbf{v} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, where $\mathbf{x}_n \in \mathbb{R}^{H \times W \times C}$, with H and W denoting the spatial resolution and C the number of channels. Given the time-reversal operation $\mathcal{R}(\mathbf{v}) = (\mathbf{x}_T, \ldots, \mathbf{x}_1)$, the observed clip is defined as $\tilde{\mathbf{v}} = \mathbf{v}$ if $y = \texttt{F}$ and $\tilde{\mathbf{v}} = \mathcal{R}(\mathbf{v})$ if $y = \texttt{B}$, where y denotes the temporal direction of $\tilde{\mathbf{v}}$.

Video Large Language Models (Video-LLM) A typical Video-LLM consists of three components: a vision encoder, a projector, and an LLM. Given an input video, the vision encoder extracts visual representations, the projector maps these representations into visual tokens compatible with the LLM, and the LLM generates responses conditioned on both the visual tokens and the text instructions. We distinguish between two types of **vision encoders**. *Frame-centric encoders* process videos mainly as individual frames and rely on additional mechanisms for temporal modeling. *Video-centric encoders*, in contrast, explicitly model temporal dynamics across frames. The **projector** compresses the visual representations into a smaller number of visual tokens and maps them into the **LLM** embedding

space. In this work, we empirically investigate which configuration of vision encoder and projector best encodes and transfers temporal information.

4 RQ1: CAN TEMPORAL INFORMATION BE ENCODED BY THE VISION ENCODER?

4.1 EVALUATION PROTOCOLS

Vision-Encoder Probing. To assess whether vision encoders capture temporal information, we train an attentive probe (Bardes et al., 2024) on top of frozen encoder, which consists of attention pooling followed by an MLP classifier. Given a video clip $\tilde{\mathbf{v}} \in \mathbb{R}^{T\times H\times W\times C}$, a frozen vision encoder maps the input video $\tilde{\mathbf{v}}$ to visual representations $\mathbf{H} \in \mathbb{R}^{T'\times H'\times W'\times d}$, where $T' = T/\tau$, $H' = H/p$, and $W' = W/p$. Here, τ denotes the temporal tubelet size, p denotes the spatial patch size, and d is the hidden dimension. We flatten the spatiotemporal dimensions of $\mathbf{H}$ into a token sequence $\mathbf{H} \in \mathbb{R}^{N\times d}$, where $N = T'H'W'$. Using the video representations, the probe predicts whether the video is played forward or backward, measuring temporal information left. Details are provided in Appendix C

End-to-End VQA Evaluation. While probing measures whether temporal information is encoded in the vision encoder, it does not reveal whether this information can be effectively utilized by the LLM. We therefore complement probing with end-to-end VQA evaluation, where the Video-LLM predicts whether each video is played forward or backward. Comparing probing and VQA performance helps distinguish failures of temporal encoding from failures of information transfer.

4.2 EXPERIMENTAL SETTINGS

Models. For probing, we evaluate two groups of visual encoders: *video-centric encoders*, including V-JEPA-L/H (Bardes et al., 2024) and InternVideo2-1B$_{\text{stage1/2}}$ (Wang et al., 2024); and *frame-centric encoders*, including Qwen2.5-VL$_{\text{ViT}}$ (Bai et al., 2025b), Qwen3-VL$_{\text{ViT}}$ (Bai et al., 2025a) and DINOv2-1B (Oquab et al., 2023). For VQA, we evaluate InternVideo2-VL-8B, Qwen2.5-VL-7B and Qwen3-VL-8B, both before and after fine-tuning on AoT VQA data constructed from the same videos used for probe training. During fine-tuning, we freeze the vision encoder, train the projector, and adapt the LLM with LoRA (Hu et al., 2022). Further details on the InternVideo2 variants are provided in Appendix A. We also include two proprietary models, GPT-4.1 and Gemini 2.5 Pro, which achieve the strongest performance on AoT$_{PPB}$.

Training Data. We use videos from the training split of Something-Something v2 (SSv2) (Goyal et al., 2017). To construct balanced forward/backward supervision, we randomly reverse 50% of the videos during training. Following previous works (Wang et al., 2024; Li et al., 2024a), videos are resized to 224×224 resolution, and we uniformly sample 16 frames per video clip. Other training details are given in Appendix C. For Video-LLM fine-tuning, forward/backward videos are converted into VQA-style samples using manually crafted templates, which are provided in Appendix E.2.

AoT Evaluation Data. We evaluate trained probes and Video-LLMs on psychophysics-grounded AoT_{PPB}. This benchmark consists of videos depicting physically irreversible events and enables direct comparison to human. Each video is evaluated in both forward and reverse temporal order, using 16 input frames at a 224×224 resolution. We report the overall accuracy.

4.3 RESULTS

Table 1 summarizes the overall accuracy on AoT$_{PPB}$. **In the probing setting**, video-centric encoders substantially outperform frame-centric encoders, except for the language-aligned InternVideo2$_{\textit{stage2}}$, which we further analyze in Section 5. VJepa-L, VJepa-H, and InternVideo2$_{\textit{stage1}}$ achieve strong performance, reaching 85.0%, 88.8%, and 81.1% accuracy, respectively. In contrast, frame-centric encoders remain near chance level, with Qwen2.5VL$_{\text{ViT}}$ and Qwen3VL$_{\text{ViT}}$ achieving 50.7% and 49.8%, respectively. These results indicate that AoT information can be effectively encoded by video-centric encoders with explicit cross-frame temporal modeling, whereas frame-centric encoders fail to capture such dynamics. In contrast, Video-LLMs perform poorly **under the VQA evaluation**. GPT-4.1 reaches 61.3%, still far below the human accuracy of 89.2%, while other models remain near chance, even after AoT fine-tuning. Notably, InternVideo2$_{\textit{stage1}}$ achieves 81.1% under probing,

Table 1: **Accuracy on AoT_{PPB}.** Human accuracy is **89.2%**. The AoT column indicates whether the model is trained on the AoT task. The Encoder column denotes the encoder type.

Probing

Model	AoT	Encoder	Acc.
Qwen2.5VL$_{ViT}$	✓	Frame	50.7
Qwen3VL$_{ViT}$	✓	Frame	49.8
DINOv2	✓	Frame	50.0
VJepa-L	✓	Video	85.0
VJepa-H	✓	Video	**88.8**
InternVideo2$_{stage1}$	✓	Video	81.1
InternVideo2$_{stage2}$	✓	Video	49.5

VQA

Model	AoT	Encoder	Acc.
GPT-4.1	–	–	**61.3**
Gemini2.5-pro	–	–	60.6
Qwen2.5VL-7B	✗	Frame	49.3
Qwen3VL-8B	✗	Frame	50.2
InternVideo2VL	✗	Video	50.0
Qwen2.5VL-7B	✓	Frame	49.3
Qwen3VL-8B	✗	Frame	51.7
InternVideo2VL	✓	Video	50.0

whereas InternVideo2VL reaches only 50.0% in VQA, indicating that temporal information encoded by the visual backbone is not effectively transferred to the LLM.

Overall, these results support two conclusions. First, temporal information can be effectively encoded in *video-centric encoders* with explicit temporal modeling. Second, current Video-LLMs fail to effectively transfer this information to LLMs, revealing an overlooked bottleneck in temporal information transfer. This motivates RQ2: How can temporal information be effectively transferred from video-centric encoders to LLMs?

5 RQ2: How Can Temporal Information Be Effectively Transferred from Video-Centric Encoders to LLMs?

RQ1 shows that video-centric encoders can encode strong temporal information, yet this information is not always accessible to the LLM. This suggests that the bottleneck lies not only in temporal encoding, but also in how temporal information is preserved and transferred sequentially. We therefore compare different *projector architectures* to assess how well they preserve temporal information. Second, we fix the projector architecture and token budget while varying the *visual token compression* settings for preserving the temporal dimension. Finally, because probe-recoverable temporal information may still be ineffective for the LLM, we conduct a *layer-wise analysis* to examine how temporal representations evolve across encoder depth and when they become usable by the LLM.

5.1 Projector Analysis

Experimental Setup for Projector Architecture Analysis. The projector maps video representations into a compact set of visual tokens consumed by the LLM. Since AoT recognition relies on temporal irreversibility cues, an important design consideration is whether the projector preserves the temporal information encoded by the video backbone during vision-language projection. We compare five projector architectures: Q-Former (Wang et al., 2024), BiLSTM (Fateh et al., 2026), C-Abstractor (Cha et al., 2024a), gated cross-attention (Gated-XATTN) (Alayrac et al., 2022), and an MLP projector. We do not compress the temporal dimension for the projector inputs. Additional architectural details are provided in Appendix B. All other experimental settings follow RQ1 (Section 4).

Experimental Setup for Visual Token Compression Strategy Analysis. Architectural differences alone do not isolate the role of temporal preservation. We therefore conduct a more controlled comparison using different visual token compression strategies. Since the MLP projector achieves the strongest performance among the five architectures (Figure 2(a)), we fix the projector architecture to MLP and vary only how the visual representations are compressed. Under a fixed visual-token budget, we consider two variants: (1) *Time-Preserved MLP*, which preserves the temporal dimension T' while compressing only the spatial dimensions $H' \times W'$; and (2) *Time-Compressed MLP*, which additionally compresses along the temporal dimension T'. By keeping the projector architecture and token budget fixed while changing which dimensions are compressed, this comparison directly tests whether preserving temporal information is important for downstream AoT reasoning.

Results of Projector Architecture Analysis. Figure 2(a) presents the results with InternVideo2$_{stage1}$ across five projector architectures. While four projectors achieve over 90% accuracy, Q-Former,

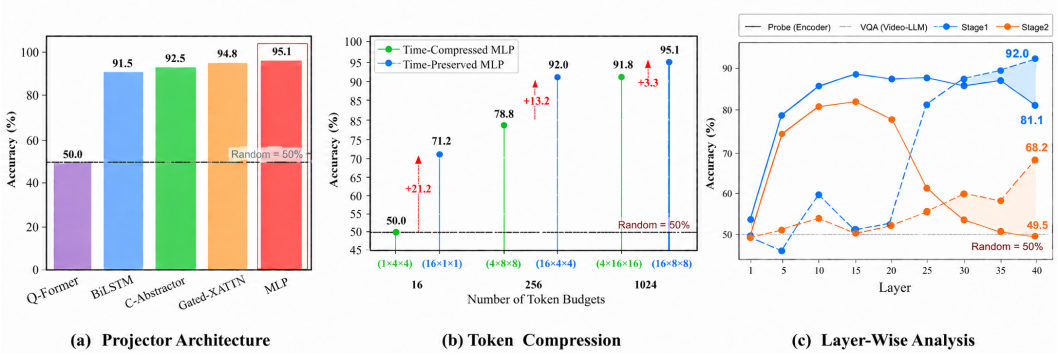


Figure 2: (a) We compare five projector architectures on AoT$_{PPB}$. (b) MLP projectors compress video representations $\mathbf{H} \in \mathbb{R}^{T' \times H' \times W' \times d}$ into a compact set of visual tokens. We compare Time-Compressed MLP and Time-Preserved MLP. (c) We analyze the discrepancy between probe recoverability and LLM usebility, InternVideo2$_{stage1}$ and InternVideo2$_{stage2}$ through layer-wise evaluation. Layer-wise analysis is conducted under $16 \times 4 \times 4$ setting for efficiency.

which is adopted by InternVideo2VL, performs close to chance level, suggesting a substantial loss of temporal information during projection. Prior works have shown that Q-Former can degrade spatial information in image understanding (Cha et al., 2024b; Lin et al., 2024), and our results suggest that a similar limitation may extend to temporal information in video representations. As our study focuses on empirical analysis, we leave a systematic investigation of this mechanism to future work. In contrast, the MLP projector achieves the highest accuracy among the five architectures, reaching 95.1%. These results suggest that preserving the temporal information encoded by the video backbone is an important consideration in projector design. We next test this hypothesis more directly by holding the projector architecture fixed and explicitly controlling whether temporal information is preserved during visual-token compression.

Results of Visual Token Compression Strategy Analysis. To directly examine the importance of preserving temporal information, we fix the projector architecture to MLP and compare Time-Preserved MLP with Time-Compressed MLP under matched visual-token budgets. Given video representations of size $T' \times H' \times W' = 16 \times 16 \times 16$, we evaluate both strategies with 16, 256, and 1024 visual tokens. As shown in Figure 2(b), the Time-Preserved MLP consistently outperforms the Time-Compressed MLP across all token budgets, with the performance gap becoming larger as compression becomes more aggressive. Because the two strategies use the same projector architecture and token budget but differ in whether temporal resolution is preserved, their performance gap provides direct evidence that retaining temporal information is important for downstream AoT reasoning. The effect becomes particularly clear when the temporal dimension is fully compressed to $T' = 1$, where AoT performance collapses to chance level.

Overall, the projector and compression analyses consistently point to the same conclusion: effective transfer of temporal information requires preserving temporal information during vision-language projection. Additional results with InternVideo2$_{stage2}$ are provided in Appendix D.2.

5.2 Layer-wise Analysis

The analyses above establish temporal preservation as critical for transferring information from the video encoder to the LLM. Yet probing accuracy does not map directly to downstream VQA performance. Notably, InternVideo2$_{stage2}$ achieves 73.1% VQA accuracy (Appendix D.2) despite near-chance probing performance (49.5%, Table 1), suggesting that probe-recoverable information and LLM-utilizable information are not necessarily equivalent. To investigate this discrepancy, we perform a layer-wise analysis by extracting representations from different encoder layers and evaluating each with both probing and downstream VQA. This allows us to trace how temporal information evolves across encoder depth.

Experimental Setup for Layer-Wise Analysis. We adopt InternVideo2$_{stage1}$ ($Stage1$) and InternVideo2$_{stage2}$ ($Stage2$) for layer-wise analysis. For layer-wise Video-LLM fine-tuning, we

Table 2: **Main Results.** The +AoT column indicates whether AoT data is included during instruction tuning. Baseline models are not directly comparable to ours due to substantially larger training data and different experimental settings.

Model	+AoT	Short Video							Long Video	
		AoT_{PPB}	VITATECS direction	TVBench Averaged	TVBench AS	TVBench AL	TVBench MD	MVBench	MLVU	Video-MME Long
Chance Rate	–	50.0	50.0	33.3	50.0	25.0	25.0	27.3	25.0	25.0
GPT-4.1	–	61.3	88.0	56.1	83.3	53.1	49.6	62.1	86.6	60.7
Gemini2.5-pro	–	60.6	89.6	67.8	87.6	71.9	94.0	67.5	81.2	64.8
InternVideo2VL	–	50.0	76.4	40.7	65.2	38.8	27.0	59.9	47.4	37.0
Qwen2VL-7B	–	50.0	84.5	42.9	64.5	31.9	16.4	60.8	62.5	52.9
Qwen2.5VL-7B	–	49.3	74.8	47.1	65.9	39.4	31.5	61.0	67.9	52.8
Mistral-7B										
+ $\text{InternVideo2}_{stage1}$	✗	48.6 (–)	78.4 (–)	40.1 (–)	64.6 (–)	34.4 (–)	25.4 (–)	56.3 (–)	47.7 (–)	34.7 (–)
+ $\text{InternVideo2}_{stage1}$	✓	94.8 (+46.2)	81.8 (+3.4)	41.4 (+1.3)	68.7 (+4.1)	38.1 (+3.7)	29.7 (+4.3)	56.1 (-0.2)	49.2 (+1.5)	35.9 (+1.2)
+ $\text{InternVideo2}_{stage2}$	✗	50.0 (–)	76.4 (–)	39.0 (–)	64.9 (–)	29.4 (–)	29.3 (–)	55.4 (–)	45.6 (–)	34.6 (–)
+ $\text{InternVideo2}_{stage2}$	✓	79.3 (+29.3)	77.9 (+1.5)	39.6 (+0.6)	67.3 (+2.4)	31.3 (+1.9)	25.4 (-3.9)	55.4 (+0.0)	46.9 (+1.3)	33.2 (-1.2)
+ VJepa-H	✗	41.7 (–)	71.3 (–)	39.0 (–)	62.7 (–)	33.8 (–)	25.9 (–)	50.6 (–)	39.7 (–)	31.7 (–)
+ VJepa-H	✓	93.6 (+51.9)	76.1 (+4.8)	39.4 (+0.4)	60.2 (-2.5)	33.8 (+0.0)	25.9 (+0.0)	51.6 (+1.0)	40.0 (+0.3)	32.2 (+0.5)
Qwen2.5-7B										
+ $\text{InternVideo2}_{stage1}$	✗	50.7 (–)	81.3 (–)	41.6 (–)	73.2 (–)	33.8 (–)	21.6 (–)	56.6 (–)	50.4 (–)	42.1 (–)
+ $\text{InternVideo2}_{stage1}$	✓	98.1 (+47.4)	82.1 (+0.8)	42.8 (+1.2)	73.9 (+0.7)	36.3 (+2.5)	28.9 (+7.3)	56.2 (-0.4)	51.5 (+1.1)	42.8 (+0.7)
+ $\text{InternVideo2}_{stage2}$	✗	41.0 (–)	80.6 (–)	40.7 (–)	72.3 (–)	33.8 (–)	22.4 (–)	56.0 (–)	48.7 (–)	40.3 (–)
+ $\text{InternVideo2}_{stage2}$	✓	82.6 (+41.6)	82.2 (+1.6)	40.8 (+0.1)	70.5 (-1.8)	35.0 (+1.2)	23.3 (+0.9)	57.9 (+1.9)	48.2 (-0.5)	41.3 (+1.0)
+ VJepa-H	✗	45.7 (–)	75.1 (–)	37.3 (–)	60.6 (–)	28.1 (–)	27.2 (–)	46.8 (–)	46.5 (–)	35.4 (–)
+ VJepa-H	✓	92.9 (+47.2)	77.4 (+2.3)	38.5 (+1.8)	63.8 (+3.2)	33.1 (+5.0)	28.4 (+1.2)	48.8 (+2.0)	46.5 (+0.0)	36.2 (+0.8)

follow InternVideo2VL using Mistral-7B-Instruct as language backbone. We adopt Time-Preserved MLP under a 256-token budget for efficiency. The projector reduces only the spatial resolution from 16×16 to 4×4 while keeping the temporal dimension unchanged at $T' = 16$. The layer-wise probing settings follow RQ1 (Section 4). Additional layer-wise results under different settings are provided in Appendix D.1.

Layer-Wise Probing Results. As shown in Figure 2(c), in the probing setting, both $Stage1$ and $Stage2$ begin near chance level in the first layer, but their performance rapidly improves from the fifth layer. For $Stage1$, probing accuracy rises to nearly human-level performance by the 15th layer. $Stage2$ exhibits a similar pattern, indicating that AoT-relevant temporal information continues to be encoded even after language alignment. Together, these results confirm that video-centric encoders can encode temporal information required for AoT task.

Layer-Wise VQA Results. The VQA results reveal a complementary trend. For $Stage1$, VQA accuracy remains close to chance in early layers but increases sharply in the 25th layer, eventually surpassing human performance. By contrast, $Stage2$ yields consistently weaker AoT performance. Specifically, probing accuracy for $Stage2$ drops markedly after the 25th layer and approaches chance level in the final layer. This indicates that the temporal information in $Stage2$ is less effectively transferred to the LLM, likely transforming temporal information into a more implicit and language-conditioned representation.

In summary, this layer-wise discrepancy reveals that the presence of recoverable temporal information does not necessarily imply that the representation is equally effective for the LLM. Probing therefore measures whether temporal information is accessible in the representation, while downstream VQA additionally reflects whether that information is represented in a form that can be effectively exploited by the LLM, highlighting the complementary roles of our two AoT settings.

6 RQ3: Can AoT supervision improve Video-LLM performance on broader temporal reasoning tasks?

Based on the findings of RQ1 and RQ2, we adopt a Video-LLM configuration designed for effective temporal information transfer, combining temporal-aware video encoder and time-preserved projector with LLMs. We then investigate whether AoT supervision can improve Video-LLM performance on broader temporal reasoning tasks.

6.1 EXPERIMENTS OF INSTRUCTION-TUNING

Model. To assess generalization across encoder and language model families, we evaluate InternVideo2$_{stage1}$, InternVideo2$_{stage2}$, and VJepa2-Huge as video encoders, combined with Mistral-7B-Instruct and Qwen2.5-7B-Instruct (Qwen et al., 2025) as language models. We adopt the best-performing projector design from RQ2 (Section 5): a Time-Preserved MLP that reduces the spatial resolution to 8×8 while maintaining the temporal resolution at $T' = 16$, yielding 1024 visual tokens. During fine-tuning, the video encoder remains frozen, the projector is trained from scratch, and the LLM is adapted using LoRA. Additional training details are provided in Appendix C.

Instruction-Tuning Data. We fine-tune Video-LLMs on approximately 1 million general-purpose visual instruction-tuning samples, following the data configuration of Maaz et al. (2024a). These data provide the models with general visual understanding capabilities and serve as our baseline training setting without AoT supervision. Additional dataset details are provided in Appendix E.

AoT VQA Data. To study AoT supervision, we augment the instruction-tuning mixture with AoT VQA samples from the training splits of Something-Something V2 (SSv2) and Moments-in-Time (MiT), motivated by our probing results in Appendix D.1. Both datasets contain rich motion and physical dynamics in short clips averaging 3.5 seconds, reducing artifacts associated with long videos and providing effective temporal supervision. The MiT samples are drawn from a split disjoint from that used by AoT$_{PPB}$, ensuring no overlap between training and evaluation data.

6.2 BENCHMARKS

We evaluate our models on six benchmarks spanning both short and long videos to assess the effectiveness of AoT supervision. Although our AoT training data consist exclusively of short clips, our evaluation includes long-video benchmarks to examine cross-duration generalization. Following Chen et al. (2024a), we sample 16 and 32 frames for short and long video evaluation, respectively.

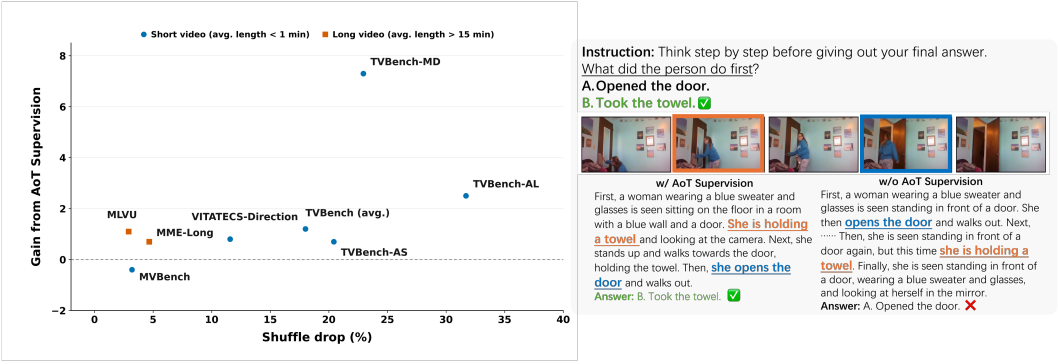
Short-Video Benchmarks: We evaluate on AoT$_{PPB}$, TVBench (Cores et al., 2024), VITATECS-Direction (Li et al., 2024b), and MVBench (Li et al., 2024a), all of which contain videos shorter than one minute. AoT$_{PPB}$, TVBench, and VITATECS-Direction primarily assess temporal understanding, covering physical irreversibility, temporal localization, and action ordering, respectively, whereas MVBench evaluates broader video understanding capabilities. **Long-Video Benchmarks:** We evaluate on MLVU (Zhou et al., 2025) and the long-video subset of Video-MME (Fu et al., 2025) to assess whether temporal reasoning learned from short-video AoT supervision generalizes to longer videos. Both benchmarks contain videos with averaged durations exceeding 15 minutes.

6.3 MAIN RESULTS

Table 2 summarizes the main results. We include two proprietary models, GPT-4.1 and Gemini 2.5 Pro, which achieve the strongest reported performance on AoT$_{PPB}$, along with three open-weight baselines, InternVideo2VL, Qwen2-VL-7B, and Qwen2.5-VL-7B, evaluated under the protocol described above. Results on broader benchmarks are provided for reference only, as differences in training data and experimental settings preclude direct comparison with our models.

AoT supervision improves both AoT recognition and broader temporal reasoning. Adding AoT supervision yields substantial gains on AoT$_{PPB}$ across all configurations. Beyond AoT$_{PPB}$, these improvements transfer to broader temporal reasoning and video understanding benchmarks, with gains of up to 1.8 points on TVBench, 3.4 points on VITATECS-Direction, and 2.0 points on MVBench. On TVBench, the improvements are concentrated in temporally sensitive subtasks, including Action Sequence (AS), Action Localization (AL), and Moving Direction (MD). Complete TVBench results are provided in Appendix F. Although our AoT VQA training data consist only of short clips, the resulting improvements extend to long-video benchmarks, with gains of up to 1.5 points on MLVU and 1.2 points on Video-MME-Long. These results show that the benefits of AoT supervision transfer across video durations despite the substantial mismatch between training and evaluation video lengths. More importantly, the gains are varying across tasks: tasks that rely more heavily on temporal information tend to benefit more from AoT supervision.

Temporal sensitivity analysis for broader temporal reasoning tasks. Although AoT supervision improves performance across several downstream tasks, the magnitude of the gains varies substantially



(a) (b)

Figure 3: **(a) Temporal Sensitivity Analysis.** We analyze why different tasks benefit differently from AoT supervision by measuring their sensitivity to frame shuffling. Results are averaged over three runs. **(b) Case Study from TVBench.** AoT supervision helps the Video-LLM correctly identify that the person took the towel before opening the door, whereas the model without AoT supervision hallucinates an inconsistent temporal order.

across benchmarks. We hypothesize that this variation partly reflects differences in how strongly each benchmark actually depends on temporal information: if a task can be solved largely from static or order-insensitive cues, improved temporal reasoning may yield little observable gain. Following Xue et al. (2025), we test this hypothesis by measuring temporal sensitivity as the relative performance drop after randomly shuffling video frames. As shown in Figure 3(a), tasks that are more sensitive to frame order generally benefit more from AoT supervision. For example, MD and AL show large performance drops under frame shuffling (22.9% and 31.7%) and correspondingly large gains from AoT supervision (+7.3 and +2.5), whereas tasks with low shuffle sensitivity exhibit limited or even negative gains. Long-video benchmarks are also comparatively insensitive to frame shuffling. These results suggest that the heterogeneous downstream gains are closely related to benchmark temporal sensitivity: AoT supervision provides the clearest benefits when successful prediction genuinely requires temporal information, while temporally insensitive benchmarks may obscure such improvements. As a sanity check, shuffling frames reduces AoT_{PPB} accuracy from 98.1% to near chance (Appendix D.3), confirming that AoT_{PPB} itself strongly depends on temporal order.

Stage-2 encoders benefit less from AoT supervision. Across both LLMs, $\text{InternVideo2}_{stage1}$ consistently outperforms $\text{InternVideo2}_{stage2}$ after AoT supervision and exhibits larger gains on broader temporal reasoning tasks. This observation is consistent with our findings in RQ2, where CLIP-style training makes temporal information less explicitly preserved and, consequently, less transferable to downstream temporal reasoning.

7 Conclusion and Limitations

We investigate temporal bottlenecks in Video-LLMs through complementary probing and downstream evaluation. Our results show that encoding temporal information in visual representations does not necessarily guarantee that it can be effectively exploited by the LLM: projector architectures, temporal compression can affect how temporal information is transferred across the model. Guided by these findings, we develop a temporally aware Video-LLM that surpasses reported human performance on AoT_{PPB} and further show that AoT supervision improves broader temporal reasoning and video understanding. Together, these results establish AoT as both a diagnostic tool for localizing temporal bottlenecks and an effective supervision signal for enhancing temporal understanding in Video-LLMs.

Our study has two limitations. First, we construct AoT VQA data without specifically optimizing for data quality. As a result, the supervision may contain temporally ambiguous clips, which could reduce the effectiveness of AoT supervision. Curating higher-quality AoT data with stronger temporal irreversibility is therefore a promising direction for future work. Second, due to limited computational resources, our experiments are restricted to 7B LLMs. Evaluating how AoT supervision scales to larger models remains an important avenue for future research.

8 AI USE STATEMENT

We used generative AI tools to improve the readability and presentation of the manuscript, assist with a portion of code, and create or modify figures. All AI-assisted text, code, and figures were reviewed and verified by the authors. We did not use AI tools to formulate research hypotheses, design the methodology or experiments, implement the core methods, analyze the results, or draw scientific conclusions.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. Test of time: Instilling video-language models with a sense of time. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2503–2516. IEEE, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

Adrien Bardes, Quentin Garrido, Jean Ponce, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv:2404.08471*, 2024.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2917–2927, 2022.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13817–13827, June 2024a.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Yejin Bang, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. *arXiv preprint arXiv:2512.10942*, 2025.

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.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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*, pp. 24185–24198, 2024b.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

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

Arthur Stanley Eddington and Edmund Taylor Whittaker. *The nature of the physical world*, volume 27. Macmillan New York, 1928.

Fawad Javed Fateh, Umer Ahmed, Hamza Khan, Zeeshan Zia, and Quoc-Huy Tran. Temporalvlm: Video llms for temporal reasoning in long videos. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1427–1447, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, Stone Tao, Mengyang Liu, Jiaxi Yang, Chao-Jung Lai, Chuanyang Jin, Jiannan Xiang, Benhao Huang, Zeming Chen, David Danks, Hao Su, Tianmin Shu, Ziqiao Ma, Lianhui Qin, and Zhiting Hu. Do vision-language models have internal world models? towards an atomic evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26170–26195, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1342. URL https://aclanthology.org/2025.findings-acl.1342/.

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 *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

Nao Hanyu, Kei Watanabe, and Shigeru Kitazawa. Ready to detect a reversal of time's arrow: a psychophysical study using short video clips in daily scenes. *Royal Society open science*, 10(4): 230036, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13700–13710, 2024.

Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. Interpreting physics in video world models. *arXiv preprint arXiv:2602.07050*, 2026.

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.

David Layzer. The arrow of time. *Scientific American*, 233(6):56–69, 1975.

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8592–8603. IEEE, 2025a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025b.

Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2024b.

Junyan Lin, Haoran Chen, Dawei Zhu, and Xiaoyu Shen. To preserve or to compress: An in-depth study of connector selection in multimodal large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5666–5680, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.325. URL https://aclanthology.org/2024.emnlp-main.325/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024b.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Shiho Matta, Lis Kanashiro Pereira, Peitao Han, Fei Cheng, and Shigeru Kitazawa. Which way does time flow? a psychophysics-grounded evaluation for vision-language models. *arXiv preprint arXiv:2510.26241*, 2025.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pp. 304–323. Springer, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T Freeman. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2035–2042, 2014.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Ali Rasekh, Erfan Bagheri Soula, Omid Daliran, Simon Gottschalk, and Mohsen Fayyaz. Enhancing temporal understanding in video-LLMs through stacked temporal attention in vision encoders. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=2EJrs3gUO6.

Michael S. Ryoo, Honglu Zhou, Shrikant Kendre, Can Qin, Le Xue, Manli Shu, Jongwoo Park, Kanchana Ranasinghe, Silvio Savarese, Ran Xu, Caiming Xiong, and Juan Carlos Niebles. xgen-mm-vid (blip-3-video): You only need 32 tokens to represent a video even in vlms, 2025. URL https://arxiv.org/abs/2410.16267.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *International Conference on Learning Representations*, volume 2024, pp. 9379–9395, 2024.

Yumeng Shi, Quanyu Long, Yin Wu, and Wenya Wang. Causality matters: How temporal information emerges in video language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9006–9014, 2026.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2023.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pp. 396–416. Springer, 2024.

Donglai Wei, Joseph J. Lim, Andrew Zisserman, and William T. Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018a.

Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8052–8060, 2018b.

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pp. 453–470. Springer, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=OYciB30Z4n.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Sihyun Yu, Nanye Ma, Pinzhi Huang, Hyunseok Lee, Shusheng Yang, June Suk Choi, Ellis Brown, Oscar Michel, Boyang Zheng, Jinwoo Shin, et al. Benchmarking visual state tracking in multimodal video understanding. *arXiv preprint arXiv:2606.03920*, 2026.

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*, 2025.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

A INTERNVIDEO2 MODEL DETAILS

InternVideo2$_{\text{stage1}}$: Vision Encoder Pre-training. InternVideo2$_{\text{stage1}}$ (Wang et al., 2024) is trained as a video-centric encoder through masked video reconstruction. Given uniformly sampled video frames, a large portion of video patches is masked, and the encoder is optimized to reconstruct the video representations distilled from video expert models (Tong et al., 2022; Chen et al., 2024b). This stage mainly aims to learn spatiotemporal representations from large-scale video data, providing the basis for subsequent multimodal alignment.

InternVideo2$_{\text{stage2}}$: Language Alignment for Vision Encoder. InternVideo2$_{\text{stage2}}$ further aligns the InternVideo2$_{\text{stage1}}$ with language by training on video–text and image–text data. The model is primarily optimized with CLIP-style training objective, enabling the visual representations to be aligned with language. Compared with $stage1$, this stage emphasizes language alignment rather than purely video-centric representation learning.

Instruction Tuning (InternVideo2VL). InternVideo2VL couples InternVideo2$_{\text{stage2}}$ with Mistral-7B via a Q-Former projector and is trained on multimodal instruction-following data. The video encoder and Q-Former are fully trainable, while the LLM is adapted with LoRA. This stage adapts the model for conversation and question answering, allowing the LLM to generate language responses conditioned on video representations.

B PROJECTOR DESIGNS

For the projector architecture comparison, all five designs use a visual-token budget of 1024. Given video representations with spatiotemporal dimensions $T' \times H' \times W' = 16 \times 16 \times 16$, Q-Former produces 1024 output tokens using learnable queries. The other four designs reduce the spatial resolution to 8×8 while preserving the temporal resolution at $T' = 16$, yielding $16 \times 8 \times 8 = 1024$ visual tokens. Architectural details are provided below.

Q-Former: Following Li et al. (2023), we initialize Q-Former with pretrained *bert-base-uncased* weights and randomly initialize its cross-attention layers. Cross-attention layers are inserted every two Transformer layers. Motivated by the finding in InternVideo2VL (Wang et al., 2024) that conditioning Q-Former on textual instructions can lead to overfitting, we also omit textual instruction conditioning.

BiLSTM: We employ a BiLSTM-based projector (Fateh et al., 2026) to model temporal dependencies. After spatial downsampling, we treat the features at each spatial location as a separate temporal sequence. We apply layer normalization and process each sequence with a single-layer bidirectional LSTM with a hidden size of 1024, aggregating temporal context in both directions while retaining an output at each time step. The resulting features undergo layer normalization, dropout, and a linear projection to the LLM hidden dimension. Finally, the outputs are flattened into a sequence of visual tokens and fed into the LLM.

C-Abstractor: Inspired by the C-Abstractor of Cha et al. (2024b), we implement a convolutional projector for video representations. Input features are first layer-normalized and projected to an intermediate channel dimension of 1024. Two spatial residual blocks refine the features before a learnable convolution with kernel size and stride $(1, 2, 2)$ reduces the spatial resolution from 16×16 to 8×8 while preserving the temporal dimension. Two additional spatial residual blocks further refine the downsampled features. Each residual block consists of group normalization, a depthwise $1 \times 3 \times 3$ convolution, a GELU activation, a pointwise convolution, and dropout with a rate of 0.1.

Gated-XATTN: Following Alayrac et al. (2022), we use gated cross-attention modules to inject visual information into selected LLM layers. The visual features are spatially downsampled while preserving the temporal resolution, flattened into a sequence of 1024 tokens, and projected to the LLM hidden dimension. Text hidden states serve as queries, and the projected visual tokens serve as keys and values. The cross-attention output is integrated through a learnable gated residual connection, followed by a gated feed-forward residual block with an expansion ratio of 4. The gates are initialized to zero, and the modules are inserted every four decoder layers.

MLP: Following Maaz et al. (2024a), we use an MLP projector to map visual features into the LLM embedding space. We apply adaptive average pooling over the spatial dimensions (Yao et al., 2024) to reduce the number of visual tokens while preserving temporal resolution. For the time-compressed variant, we additionally apply pooling over the temporal dimension. Each resulting token is independently transformed by a two-layer MLP with layer normalization and GELU activations. The output features are flattened into a sequence of visual tokens.

C ADDITIONAL EXPERIMENTS

Probing setup. Given a video clip $\tilde{\mathbf{v}} \in \mathbb{R}^{T \times H \times W \times C}$, a frozen vision encoder maps the input video $\tilde{\mathbf{v}}$ to visual representations $\mathbf{H} \in \mathbb{R}^{T' \times H' \times W' \times d}$, where $T' = T/\tau$, $H' = H/p$, and $W' = W/p$. Here, τ denotes the temporal tubelet size, p denotes the spatial patch size, and d is the hidden dimension. We flatten the spatiotemporal dimensions of $\mathbf{H}$ into a token sequence $\mathbf{H} \in \mathbb{R}^{N \times d}$, where $N = T'H'W'$. Given a learnable query token $\mathbf{q} \in \mathbb{R}^d$, attentive probing computes $\tilde{\mathbf{q}} = \mathrm{softmax}\Big((\mathbf{q}\mathbf{W}_Q)(\mathbf{H}\mathbf{W}_K)^\top/\sqrt{d}\Big)\,(\mathbf{H}\mathbf{W}_V)$, followed by $p(y \mid \tilde{\mathbf{v}}) = \mathrm{softmax}(\mathrm{MLP}(\tilde{\mathbf{q}}))$, where $y \in \{\mathtt{F}, \mathtt{B}\}$. Here, $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable projection matrices. We train the probe for 1 epoch with a batch size of 4 and a learning rate of 1×10^{-4}. The attention-based pooling layer consists of one cross-attention layer with 16 attention heads and one additional learnable query token. The probe classifier is optimized with a standard cross-entropy loss. Probing dataset statistics is summarized in Table 3.

Video-LLM fine-tuning setup. For Video-LLM fine-tuning, we keep the video encoder frozen, train the projector, and update the LLM with LoRA. We use LoRA rank 16 and $\alpha = 32$. Models are

Table 3: **Statistics of AoT Probing Datasets.**

Dataset	# Duration	# Action	# Size
Training			
$SSv2$	4.03s	172	169k
MiT_{170k}	3.00s	339	170k
$K400_{170k}$	9.57s	400	170k
Evaluation			
AoT_{PPB}	3.00s	-	424

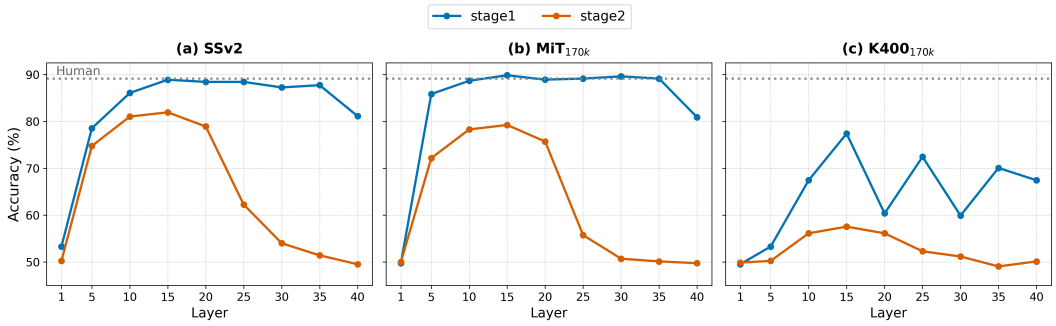


Figure 4: **Layer-wise AoT probing across probe training datasets.** Probes are trained on SSv2, MiT_{170k}, and K400_{170k}, and evaluated on AoT_{PPB}. $\text{InternVideo2}_{stage1}$ maintains strong AoT performance across deeper layers, whereas $\text{InternVideo2}_{stage2}$ peaks at early layers and then drops sharply.

trained for 1 epoch with an effective batch size of 16, a learning rate of 2×10^{-5}, and a weight decay of 0.02.

Computational resource. We conduct our experiments with 8 A100 80G GPUs.

D Additional Ablations

D.1 Layer-Wise Probing on Various Settings

Study on probe training datasets. To examine whether the layer-wise AoT dynamics are robust to the choice of probe training data, we train attentive probes on three video datasets: SSv2, Moments-in-Time (MiT), and K400 (Kay et al., 2017). For a fair comparison, we sample MiT_{170k} and K400_{170k} subsets from MiT and K400, respectively, matching the scale of SSv2 as shown in Table 3. We uniformly sample videos from each action class to construct these subsets. These datasets differ in average video duration and action distribution. For all settings, we construct balanced forward/backward supervision by randomly reversing 50% of the training videos, and evaluate the trained probes on AoT_{PPB}.

Figure 4 shows that the overall trend is consistent across datasets. For SSv2 and MiT_{170k}, temporal information is encoded in the intermediate layers of $\text{InternVideo2}_{stage1}$ and $\text{InternVideo2}_{stage2}$, while the language-aligned stage shows clear degradation in later layers. K400_{170k} yields weaker and less stable probing performance, likely because its videos are longer than those in SSv2 and MiT, causing uniform frame sampling to lose more temporal information. Nevertheless, the same qualitative pattern remains: $\text{InternVideo2}_{stage1}$ preserves stronger temporal information than language-aligned $\text{InternVideo2}_{stage2}$.

Study on video-centric encoder size. We further examine whether the layer-wise AoT performance is affected by encoder size. Figure 5(a) compares $\text{InternVideo2}_{stage2}$ with 1B and 6B parameters, using probes trained on SSv2. Both models exhibit the same overall trend: temporal information is encoded in the early layers and then degrades in later layers. The 6B encoder achieves higher peak probing accuracy and maintains strong performance over more layers, suggesting that larger video-centric encoders encode stronger AoT-relevant temporal information. However, its performance still drops

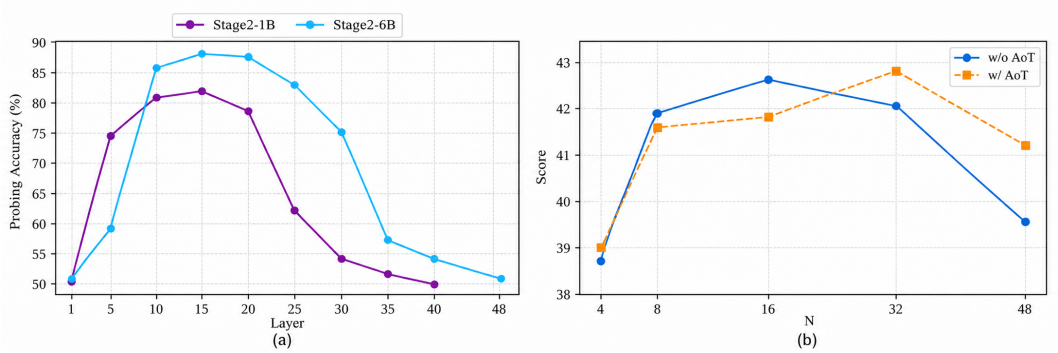


Figure 5: **(a) Layer-wise probing performance of InternVideo2$_{stage2}$ 1B and 6B.** Probes are trained on SSv2. **(b) Frame ablation on Video-MME.** We conduct frame ablation using the best-performing configuration from Table 2, which combines Qwen2.5-7B with InternVideo2$_{stage1}$ through the Time-Preserved MLP.

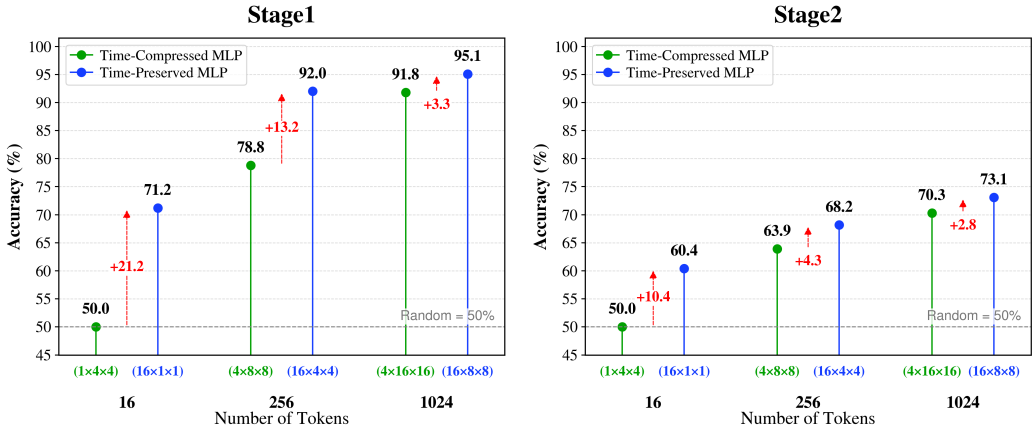


Figure 6: **Ablation of Token Compression on AoT$_{PPB}$.** We compare MLP-based projectors with different pooling strategies under three visual-token budgets. We report overall accuracy on AoT$_{PPB}$. Stage1 and Stage2 denote InternVideo2$_{stage1}$ and InternVideo2$_{stage2}$.

toward chance level in the late layers, indicating that increasing encoder size does not eliminate the degradation of explicitly accessible temporal information after language alignment.

D.2 Token Compression Strategy Analysis

Figure 6 compares two token compression strategies for MLP projectors under three visual-token budgets. The Time-Preserved MLP consistently outperforms the Time-Compressed MLP across both encoder stages and all token budgets, with the performance gap widening as compression becomes more aggressive. For InternVideo2$_{stage1}$, the gap increases from 3.3 points at 1024 tokens to 21.2 points at 16 tokens. Notably, collapsing the temporal dimension to $T' = 1$ reduces accuracy to chance level. For InternVideo2$_{stage2}$, the Time-Preserved MLP achieves accuracies of 60.4%, 68.2%, and 73.1% with 16, 256, and 1024 tokens, respectively. InternVideo2$_{stage2}$ exhibits a similar trend but consistently underperforms InternVideo2$_{stage1}$ under matched projector settings and token budgets, motivating our layer-wise analysis of how language alignment affects temporal representations. Increasing the visual-token budget also consistently improves accuracy for both compression strategies, suggesting that retaining more visual information benefits AoT recognition.

Overall, these results highlight the importance of preserving temporal resolution when transferring AoT-relevant information from video-centric encoders to LLMs.

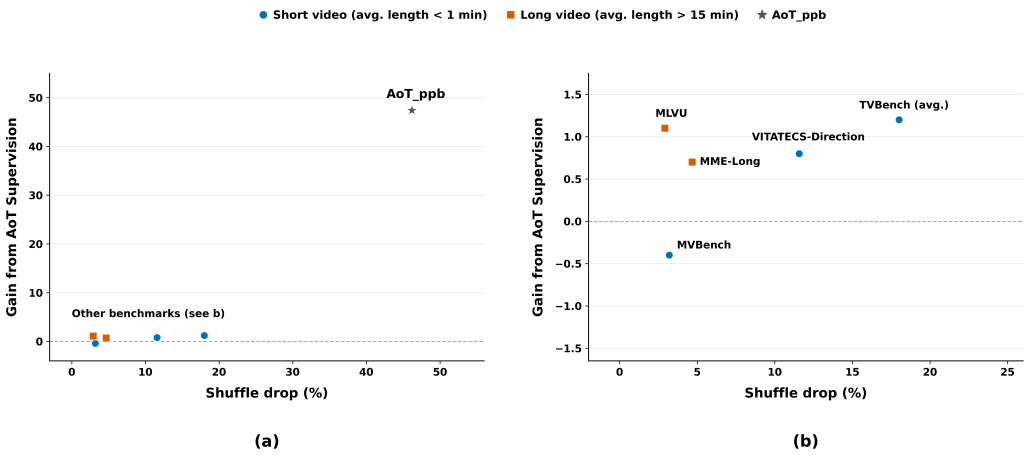


Figure 7: **Temporal Sensitivity of Evaluation Benchmarks.** Relative performance drop caused by frame shuffle versus gain from AoT supervision across (a) all benchmarks and (b) benchmarks excluding AoT$_{PPB}$. Colors distinguish short- and long-video benchmarks; the star marks AoT$_{PPB}$.

D.3 Temporal Sensitivity of Video Benchmarks

To understand when AoT supervision is most effective, we measure benchmark temporal sensitivity by the relative performance drop after randomly shuffling video frames. As shown in Figure 3(a), AoT$_{PPB}$ is highly sensitive to temporal order, with a 46.18% relative drop under frame shuffling, and also receives the largest gain from AoT supervision (+47.4 points).

A similar trend appears across other benchmarks (Figure 7(b)): temporally sensitive tasks such as TVBench show larger gains, whereas low-sensitivity benchmarks such as MVBench show limited or negative improvements. Notably, AoT supervision also benefits long-video benchmarks, suggesting that its effectiveness is more closely related to temporal dependence than video duration alone.

Overall, these results suggest that the effectiveness of AoT supervision is associated with how strongly a task depends on temporal information, rather than video duration alone. This further supports AoT as a useful supervision signal for improving temporally sensitive video understanding.

D.4 Frame Ablation on Long-Video Benchmark

We investigate the effect of the number of sampled frames on long-video understanding by varying the frame count from 4 to 48 while keeping all other settings fixed. For this ablation, we adopt the best-performing configuration from Table 2, which combines Qwen2.5-7B with InternVideo2$_{stage1}$ through the Time-Preserved MLP.

As shown in Figure 5(b), increasing the number of sampled frames generally improves performance up to a moderate frame budget, while further increasing the frame count yields diminishing or no additional gains. Without AoT supervision, performance peaks at 42.6 with 16 frames and slightly decreases to 42.1 with 32 frames. In contrast, with AoT supervision, the best performance of 42.8 is achieved with 32 frames, and the performance gains persist at 48 frames. These results suggest that the benefits of AoT supervision for long-video understanding become more evident with sufficient frame coverage. Nevertheless, our AoT training data consist only of short clips to minimize potential artifacts associated with long videos, and we sample 16 frames during training due to computational constraints. As a result, the supervision does not explicitly target long-horizon temporal dependencies. Extending AoT supervision to longer videos and larger frame budgets may further improve generalization to long-video understanding.

Table 4: **AoT VQA templates.**

Template 1 *Instruction:* Examine the motion and temporal cues in the video to determine whether it is forward or backward. *Question:* Based on the temporal cues, is the video forward or backward? *Answer:* forward / backward
Template 2 *Instruction:* Determine whether the video is played normally or backward based on physical and temporal consistency. *Question:* Is this clip more consistent with real-world forward dynamics or backward playback? *Answer:* forward / backward
Template 3 *Instruction:* Identify the arrow of time in the video by deciding whether it runs forward or backward. *Question:* What is the arrow of time of this video? *Answer:* forward / backward
Template 4 *Instruction:* Analyze the temporal progression of the video and classify it as forward-time or reversed-time. *Question:* Which of the following best describes the temporal direction of this clip? *Answer:* forward / backward

E VIDEO-LLM TUNING DATASET

E.1 INSTRUCTION TUNING DATASETS

Following Maaz et al. (2024a), we use approximately one million visual instruction-tuning samples, comprising both video–text and image–text instruction data. The dataset is constructed from a diverse mixture of multimodal instruction data, including VideoChat (Li et al., 2025b), K710 (Kay et al., 2017), SSv2 (Goyal et al., 2017), VideoChatGPT (Maaz et al., 2024b), CLEVRER (Yi et al., 2019), Next-QA (Xiao et al., 2021), VCG-112k (Maaz et al., 2024a), VCG-human (Maaz et al., 2024a), WebVid (Bain et al., 2021), ActivityNet (Caba Heilbron et al., 2015), COCO (Lin et al., 2014). These data cover diverse instruction formats, including action classification, temporal reasoning, image/video captioning, image/video question answering, and video conversation, providing broad supervision for aligning visual representations with the language model.

E.2 AOT VQA DATASETS

We combine the SSv2 training split and MiT_{170k} (Table 3). To convert forward/backward videos into VQA-style samples, we manually design four instruction templates. We use multiple templates to reduce instruction-specific bias and prevent the model from overfitting to a fixed question pattern. For each video, we uniformly select one template and assign the answer according to its playback direction, i.e., `forward` or `backward`. The templates are shown in Table 4.

F CATEGORICAL RESULTS ON TVBENCH

To better understand how AoT supervision affects different types of temporal reasoning, we report category-level results on TVBench in Table 5. Overall, adding AoT supervision yields consistent improvements in the averaged TVBench score. With Mistral-7B, AoT supervision improves $\text{InternVideo2}_{stage1}$ from 40.1 to 41.4 and $\text{InternVideo2}_{stage2}$ from 39.0 to 39.6. With Qwen2.5-7B, it improves $\text{InternVideo2}_{stage1}$ from 41.6 to 42.8 and $\text{InternVideo2}_{stage2}$ from 40.7 to 40.8.

At the category level, the gains are not uniform. AoT supervision tends to improve categories that require recognizing motion or temporal changes. For example, with Mistral-7B and $\text{InternVideo2}_{stage1}$, AoT supervision improves Action Sequence (AS) from 64.6 to 68.7, Action Antonym (AA) from 52.2 to 55.3, and Egocentric Sequence (ES) from 14.5 to 19.5. In contrast, categories such as Object Shuffle (OS), which rely more on object perception, show smaller or mixed changes. This

Table 5: **Categorical Results on TVBench.** The +AoT column indicates whether AoT data is included during instruction tuning. Colored deltas report performance changes relative to the corresponding counterpart trained without AoT. Baseline models are not directly comparable to ours due to substantially larger training data and different experimental settings.

Model	+AoT	TVBench Averaged	TVBench									
			AC	OC	AS	OS	ST	AL	AA	UA	ES	MD
GPT-4.1		56.1	27.1	27.7	83.3	35.6	77.3	53.1	85.3	59.8	62.5	49.6
Gemini2.5-pro		67.8	43.7	32.4	87.6	36.4	80.5	71.9	81.9	62.2	78.5	94.0
InternVideo2VL		40.7	28.2	46.0	65.2	33.8	63.8	38.8	57.8	29.2	17.5	26.7
Qwen2VL-7B		42.9	26.5	60.8	64.5	36.0	69.2	31.9	62.5	24.4	37.0	16.4
Qwen2.5VL-7B		47.1	30.8	53.4	65.9	37.8	81.1	39.4	64.7	44.5	22.0	31.5
Mistral-7B												
+ InternVideo2$_{stage1}$	✗	40.1 (–)	25.0	39.8	64.6	38.7	69.7	34.4	52.2	36.6	14.5	25.4
+ InternVideo2$_{stage1}$	✓	41.4 (+1.3)	25.0	37.8	68.7	35.1	68.1	38.1	55.3	36.6	19.5	29.7
+ InternVideo2$_{stage2}$	✗	39.0 (–)	26.9	41.9	64.9	34.7	69.8	29.4	49.1	32.9	11.5	29.3
+ InternVideo2$_{stage2}$	✓	39.6 (+0.6)	25.4	37.2	67.3	32.4	71.4	31.3	52.2	34.1	19.0	25.4
+ VJepa-H	✗	39.0 (–)	24.8	39.2	62.7	40.0	69.2	33.8	48.1	32.9	13.5	25.9
+ VJepa-H	✓	39.4 (+0.4)	25.0	49.3	60.2	34.7	68.1	33.8	49.1	32.9	15.0	25.9
Qwen2.5-7B												
+ InternVideo2$_{stage1}$	✗	41.6 (–)	25.1	43.2	73.2	34.2	77.3	33.8	55.6	32.9	18.5	21.6
+ InternVideo2$_{stage1}$	✓	42.8 (+1.2)	27.1	48.6	73.9	34.2	75.1	36.3	53.8	31.7	18.0	28.9
+ InternVideo2$_{stage2}$	✗	40.7 (–)	25.4	37.8	72.3	38.7	70.8	33.8	56.9	32.9	16.0	22.4
+ InternVideo2$_{stage2}$	✓	40.8 (+0.1)	25.9	40.5	70.5	35.1	64.3	35.0	57.5	32.9	22.5	23.3
+ VJepa-H	✗	37.3 (–)	25.0	35.1	60.6	33.3	62.2	28.1	50.0	35.4	16.5	27.2
+ VJepa-H	✓	38.5 (+1.2)	27.6	32.4	63.8	36.4	62.2	33.1	50.6	31.7	19.0	28.4

suggests that AoT supervision mainly strengthens temporal sensitivity. These results indicate that AoT supervision provides a useful but targeted temporal training signal.