Temporal Fidelity: Probing the Temporal Resolution Limits of Video Large Language Models

Anonymous ACL submission

Abstract

Temporal understanding—perceiving when events occur and in what order—is a prerequisite for any non-trivial video comprehension, from activity recognition to causal reasoning. Yet whether Video LLMs genuinely perceive time, or merely exploit contextual shortcuts in naturalistic benchmarks, remains untested: standard evaluations permit non-temporal cues, and raw accuracy cannot distinguish genuine perception from response bias. We introduce **Temporal Fidelity**, a psychophysics-inspired framework that isolates temporal perception using synthetic stimuli where timing is the only discriminative cue. Signal detection analysis across seven open-weight Video LLMs reveals that **most models have near-zero genuine sensitivity** ($d' < 0.1$ for three models), with raw accuracy entirely explained by response bias rather than temporal perception. Layer-wise probing uncovers a **representation–behavior gap**: temporal information is encoded near-perfectly in hidden states (probe accuracy ≥ 0.92) yet fails to surface in behavioral responses ($\Delta \geq 0.40$ for four models). These findings suggest that improving temporal reasoning in Video LLMs requires generation-side interventions—temporal instruction tuning, contrastive decoding, or temporal chain-of-thought—rather than better visual encoders alone.

1 Introduction

Video large language models (Video LLMs) report above-chance accuracy on temporal benchmarks—temporal order, duration estimation, event sequencing (Li et al., 2024a; Fu et al., 2025; Liu et al., 2024; Cai et al., 2024; Li et al., 2024b). But how much of this reflects genuine temporal perception? These benchmarks use naturalistic videos where scene context, object states, and causal structure provide non-temporal shortcuts to temporal questions (Lei et al., 2022; Buch et al., 2022; Cores et al., 2025; Varma et al., 2025); indeed, models maintain similar performance even when video frames are temporally shuffled (Feng et al., 2025).

Compounding this, standard accuracy metrics conflate genuine perception with response bias. A model that *always* answers "NO" to a temporal order question scores 55%—seemingly above chance—while perceiving nothing at all. Without controlled stimuli and bias-corrected metrics, we cannot tell whether Video LLMs reason about time or exploit correlations. While Bagad et al. (2023) showed that contrastive video-language models struggle with temporal order using synthetic stimuli, their work predates generative Video LLMs and does not disentangle sensitivity from response bias. Concurrent work by Upadhyay et al. (2026) evaluates Video LLMs on purely temporal noise-like stimuli but observes a floor effect (0% for all models), which is not diagnostic across a range of competence levels.

We address this with **Temporal Fidelity**, a psychophysics-inspired framework using synthetic stimuli where the *only* discriminative cue is timing (Figure 1). We apply signal detection theory (SDT; Green and Swets 1966) to decompose accuracy into sensitivity (d') and bias (c), and probe frozen hidden states to localize where temporal information exists but fails to surface.

Our experiments on seven open-weight Video LLMs reveal three findings. First, most models have *zero* genuine temporal sensitivity ($d' < 0.1$ for 3/7) despite above-chance raw accuracy. Second, a stark representation–behavior gap: temporal information is encoded near-perfectly in hidden states (probe accuracy ≥ 0.92) yet not used behaviorally ($\Delta \geq 0.40$ for four models). Third, only Molmo2 achieves strong behaviorally effective sensitivity ($d' = 3.92$), suggesting the bottleneck lies in language generation, not visual encoding.

2 Experimental Framework

Stimuli. Each video shows two colored shapes on a gray background moving sequentially with controlled inter-event intervals. We use 13 log-spaced intervals (33–10,000 ms) frame-aligned at 30 fps; the 33 ms minimum corresponds to one frame, and log-spacing follows Weber's law (Weber, 1834) to ensure perceptually uniform difficulty scaling. Main dataset: 1,300 videos (100 per interval, sufficient for stable d' estimation); simultaneity dataset: 600 videos across 6 offsets (0–333 ms). Order of movement is balanced.

Models. We evaluate seven 7–8B open-weight Video LLMs chosen to cover four distinct vision encoder families (CLIP, SigLIP, InternViT/InternVideo2, ViT-mRoPE) and three LLM backbones (Mistral, Qwen, InternLM): LLaVA-NeXT-Video (Zhang et al., 2024), Video-LLaMA2 (Cheng et al., 2024), Qwen2.5-VL (Bai et al., 2025), Molmo2 (Clark et al., 2026), InternVL2.5 (Chen et al., 2024), VideoLLaMA3 (Zhang et al., 2025a), and VideoChat-Flash (Li et al., 2026).

Experiments. **E1 (Order):** "Did shape A move BEFORE shape B? YES/NO" across 13 intervals. **E2 (Interval):** "How much time passed?" with 4 ordinal bins (A: <1s, B: 1–2s, C: 2–5s, D: >5s). **E3 (Simultaneity):** "Did both shapes move at the same time? YES/NO" at 6 offsets. **E4 (Frame rate):** E1 repeated at 8, 16, 30 fps. **E5 (Layer probing):** ℓ_2-regularized logistic regression ($C = 1$, `lbfgs` solver, `max_iter=1000`) on frozen encoder and LLM hidden states at 8 evenly-spaced layers (5-fold CV), following Hewitt and Manning (2019); Tenney et al. (2019); Belinkov (2022).

Metrics. We compute d' with loglinear correction (Hautus, 1995), response criterion c, balanced accuracy, Spearman's ρ for accuracy-interval correlations, quadratic-weighted κ_w (Cohen, 1960) for E2, and 95% bootstrap CIs.

3 Results

3.1 Most Models Have Zero Temporal Sensitivity

Table 1 presents our central finding. InternVL2.5 and LLaVA-NeXT-Video achieve 54.8% raw accuracy—seemingly above chance—but their balanced accuracy is exactly 0.500 and $d' = 0.06$, because both always answer "NO" ($P(\text{YES}) = 0$, $c = 3.17$); Video-LLaMA2 mirrors this from the opposite extreme, always answering "YES." Without SDT, these bias-inflated accuracies could be misread as temporal competence.

Only **Molmo2** shows strong sensitivity ($d' = 3.92$) with balanced responses ($P(\text{YES}) = 0.572$) and a strong accuracy-interval correlation ($\rho = 0.87$, $p < .001$). VideoChat-Flash ($d' = 2.33$) and Qwen2.5-VL ($d' = 1.13$) show moderate and weak sensitivity, respectively, but extreme NO biases suppress their behavioral accuracy well below what their sensitivity would predict (Figure 2).

3.2 The Representation–Behavior Gap

Table 2 exposes a stark dissociation between representation and behavior. Linear probes achieve perfect accuracy (1.00) on frozen encoder features for all 6 encoder-accessible models—yet four of them show a gap of $\Delta \geq 0.40$, with InternVL2.5 and LLaVA-NeXT-Video at BAcc $= 0.500$ (chance). For these two models, the gap is 0.50: perfect representation, zero behavior. VideoLLaMA3, despite no parseable behavioral output, still shows near-perfect encoder (1.00) and LLM (0.92) representations. Only Molmo2 closes the gap ($\Delta = 0.11$; Figure 3).

This suggests the bottleneck lies not in visual encoding but in the **language generation pathway**. Video-LLaMA2 is the sole exception: its Q-Former fails to propagate temporal information even at the representation level (LLM probe = 0.46). Shuffled-label controls (permuting labels 10×) confirm genuine encoding for the six encoder-accessible models: real probes achieve ≥ 0.92 while shuffled probes drop to $\sim$0.38 (chance: 0.25); Video-LLaMA2 is excluded as its real and shuffled probes are both 0.46, consistent with no temporal signal reaching the LLM.

3.3 Interval, Simultaneity, and Frame Rate

Interval estimation (E2). Only Molmo2 achieves above-chance BAcc (0.455 vs. 0.250) and meaningful agreement ($\kappa_w = 0.76$). All other models collapse to 1–2 response categories (e.g., LLaVA-NeXT-Video: $P(B) = 1.0$; VideoChat-Flash: $P(A) = 1.0$; Table 3, Figure 6).

Simultaneity (E3). *Every* model adopts a fixed response (always YES or always NO), yielding BAcc $= 0.500$ universally (Table 4). Even Molmo2 fails here ($P(\text{YES}) = 0$), despite strong order discrimination. This suggests simultaneity

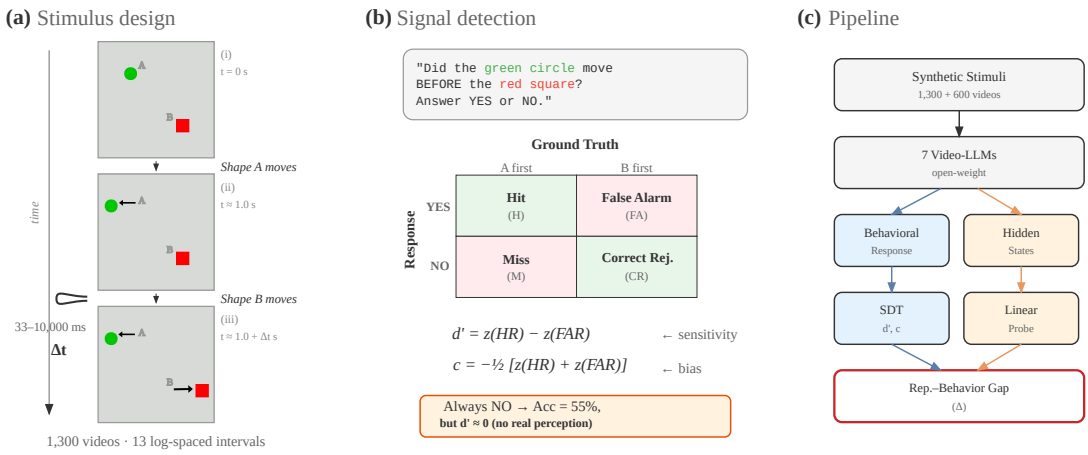

Figure 1: The Temporal Fidelity framework. (a) Two shapes move sequentially with a controlled interval Δt; timing is the only cue. (b) Signal detection theory separates genuine sensitivity (d') from response bias (c): a model that always says "NO" scores 55% yet has $d' \approx 0$. (c) Behavioral outputs are analyzed via SDT while frozen hidden states are probed with linear classifiers to localize where temporal information exists.

Model	**Encoder / LLM**	P(YES)	**BAcc**	d'	c	ρ	p
Molmo2	SigLIP2 / Qwen3-8B	0.572	**.891**	**3.92** [3.82, 4.03]	−1.18	**.87**	<.001
VideoChat-Flash	InternVideo2 / Qwen2-7B	0.088	.597	2.33 [2.21, 2.44]	2.03	−.03	.929
Qwen2.5-VL	ViT-mRoPE / Qwen2.5-7B	0.008	.509	1.13 [0.84, 1.33]	2.63	−.03	.929
InternVL2.5	InternViT / InternLM2-8B	0.000	.500	0.06 [0.03, 0.09]	3.17	−.26	.398
LLaVA-NeXT-Video	CLIP-ViT-L / Mistral-7B	0.000	.500	0.06 [0.03, 0.09]	3.17	−.26	.398
Video-LLaMA2	CLIP-ViT-L / Mistral-7B	1.000	.500	−0.06 [−.09, −.03]	−3.17	.26	.398
VideoLLaMA3	SigLIP-DiffFP / Qwen2.5-7B			*No parseable responses*			

Table 1: Temporal order results (E1). Only Molmo2 achieves high d'; most models default to a fixed response. d': sensitivity [95% bootstrap CI], c: response criterion (positive = NO bias), ρ: Spearman correlation with log-interval. VideoLLaMA3 produced no parseable output.

Model	**BAcc**	**Enc**	**LLM**	**ΔLLM**
Molmo2	.891	1.00	1.00	+.11
VideoChat-Flash	.597	1.00	1.00	+.40
Qwen2.5-VL	.509	1.00	1.00	+.49
InternVL2.5	.500	1.00	1.00	+.50
LLaVA-NeXT-Video	.500	1.00	1.00	+.50
VideoLLaMA3‡	–	1.00	.92	–
Video-LLaMA2†	.500	–	.46	−.04

Table 2: The representation–behavior gap. Linear probes on frozen features (E5) vs. behavioral balanced accuracy (E1). A positive Δ means temporal information is encoded but not used for generation. †Q-Former blocks direct encoder access. ‡No parseable behavioral output.

detection may require qualitatively different processing than order judgment.

Frame rate (E4). Molmo2 maintains sensitivity across frame rates ($d' = 3.63$ at 8 fps, 4.05 at 16 fps, 3.92 at 30 fps; Table 5). Bias-dominated models are invariant to frame rate, consistent with their lack of temporal processing.

4 Discussion and Conclusion

Raw accuracy overestimates temporal competence. A model that always answers "NO" achieves 54.8% on temporal order while having zero genuine sensitivity; on simultaneity, always-NO yields 83.3%. If existing benchmarks contain similar class imbalances, published rankings may substantially reflect response bias. We advocate for d' and balanced accuracy as standard temporal evaluation metrics.

The bottleneck appears downstream of vision. The representation–behavior gap (Table 2) shows temporal information is encoded yet not used. This points to a failure in language generation: instruction-tuned models may have learned response priors that override perceptual evidence. This finding converges with concurrent work using complementary methods: Shi et al. (2026) trace how temporal cues are synthesized through

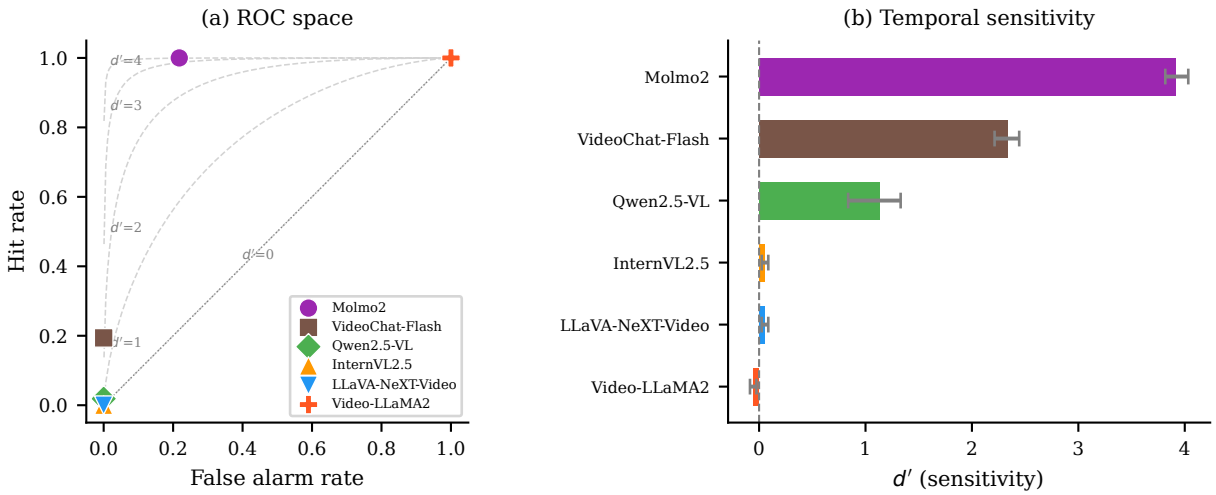

Figure 2: Signal detection analysis (E1). (a) In ROC space, Molmo2 ($d' \approx 4$) separates cleanly from the bias-dominated cluster near the diagonal. Video-LLaMA2 (always YES) sits at the top-right corner. (b) Ranked d' confirms that most models have near-zero genuine sensitivity.

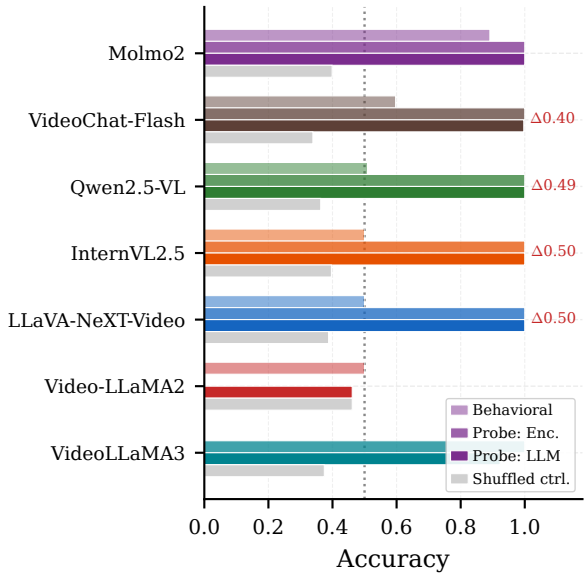

Figure 3: Temporal information is encoded but not used. Light bars: behavioral BAcc; dark bars: linear probe accuracy on frozen features. Gray bars show shuffled-label controls (∼0.38 for non-degenerate models, near the 0.25 chance floor), confirming the probe captures genuine temporal signal. Red arrows highlight gaps exceeding 0.40.

inter-frame attention and find positional encodings are not the primary bottleneck, while Zhang et al. (2025b) identify temporally specialized attention heads whose amplification improves generation—both pointing to the generation pathway rather than encoding as the locus of failure; a similar representation–action gap has been documented for non-temporal modalities (Nguyen Quang et al., 2026). That Molmo2 alone closes the gap—with SigLIP2 as encoder and Qwen3 as backbone, both newer-generation components—suggests that architectural or training advances in either the vision encoder or the LLM may be necessary to bridge this disconnect. That Molmo2 closes this gap with newer-generation encoder and LLM components points to architectural advances as the most tractable path; for existing models, generation-side interventions such as contrastive decoding or temporal chain-of-thought are lower-cost alternatives worth exploring.

Simultaneity is qualitatively harder. The universal failure on E3—including Molmo2 ($d' = 3.92$ on order)—parallels the human distinction between temporal order and simultaneity processing (Pöppel, 1997).

We release all code, stimuli, and results at `https://anonymous.4open.science/r/temporal-fidelity-C64E/` to support future work on closing the representation–behavior gap.

Limitations

Our synthetic stimuli are intentionally minimal (colored shapes on gray backgrounds), and whether models that fail on pure timing also fail on temporal judgments in naturalistic videos, where complementary cues are available, remains open. We evaluate only 7–8B models; larger or closed-source models may show different patterns, and evaluating frontier models would test whether the representation–behavior gap is a property of scale or architecture. Prompt sensitivity analysis (Table 6) shows that forced format can flip

bias polarity entirely (e.g., InternVL2.5: always-NO $\rightarrow$ $P(\text{YES}) = 0.96$) while BAcc stays at chance; prompt format modulates bias direction but not sensitivity. Linear probes provide a lower bound on encoded temporal information; nonlinear probes might recover additional signal. All inference and probing experiments were conducted on a single NVIDIA B200 GPU.

Ethics Statement

This work uses only synthetic stimuli (procedurally generated colored shapes on gray backgrounds) and evaluates publicly released open-weight models; no human subjects, personal data, or proprietary systems are involved. The framework is diagnostic and does not introduce new model capabilities or dual-use risks. All datasets, code, and model checkpoints used are publicly available under permissive licenses. We release our stimuli and evaluation code to support reproducibility and further research.

Acknowledgments

We used AI assistants for coding assistance and proofreading.

References

Piyush Bagad, Makarand Tapaswi, and Cees G. M. Snoek. 2023. Test of time: Instilling video-language models with a sense of time. In *Proceedings of CVPR*.

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, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the "video" in video-language understanding. In *Proceedings of CVPR*.

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

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

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

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, and 2 others. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. 2025. Lost in time: A new temporal benchmark for videollms. In *Proceedings of BMVC*.

Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. 2025. Breaking down video LLM benchmarks: Knowledge, spatial perception, or true temporal understanding? In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multimodal llms in video analysis. In *Proceedings of CVPR*.

David M. Green and John A. Swets. 1966. *Signal Detection Theory and Psychophysics*. Wiley, New York.

Michael J. Hautus. 1995. Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, 27(1):46–51.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*.

Ira J. Hirsh and Carl E. Sherrick. 1961. Perceived order in different sense modalities. *Journal of Experimental Psychology*, 62(5):423–432.

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2022. Revealing single frame bias for video-and-language learning. In *Proceedings of ACL*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of CVPR*.

Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. 2024b. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *Proceedings of ECCV*.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2026. Videochat-flash: Hierarchical compression for long-context video modeling. In *Proceedings of ICLR*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*.

Trung Nguyen Quang, Yiming Gao, Fanyi Pu, Kaichen Zhang, Shuo Sun, and Ziwei Liu. 2026. Senses wide shut: A representation–action gap in omni-modal LLMs. *arXiv preprint arXiv:2605.13737*.

Ernst Pöppel. 1997. A hierarchical model of temporal perception. *Trends in Cognitive Sciences*, 1(2):56–61.

Yumeng Shi, Quanyu Long, Yin Wu, and Wenya Wang. 2026. Causality matters: How temporal information emerges in video language models. In *Proceedings of AAAI*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of ACL*.

Ujjwal Upadhyay, Mukul Ranjan, Zhiqiang Shen, and Mohamed Elhoseiny. 2026. Time blindness: Why video-language models can't see what humans can? In *Proceedings of CVPR (to appear)*.

Maya Varma, Jean-Benoit Delbrouck, Sophie Ostmeier, Akshay S. Chaudhari, and Curtis Langlotz. 2025. Trove: Discovering error-inducing static feature biases in temporal vision-language models. In *Proceedings of NeurIPS*.

Ernst Heinrich Weber. 1834. *De Pulsu, Resorptione, Auditu et Tactu: Annotationes Anatomicae et Physiologicae*. Koehler, Leipzig.

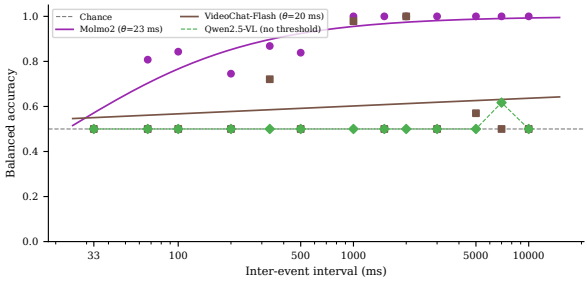

Figure 4: Psychometric functions for temporal order (E1). Sigmoid fits are shown for models exceeding $d' > 1$; the threshold θ marks the 75% correct point. Models below $d' = 1$ produce flat curves at chance, confirming they do not discriminate at any interval.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025a. Videollama3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Yiming Zhang, Zhuokai Zhao, Chengzhang Yu, Kun Wang, Zhendong Chu, Qiankun Li, Zihan Chen, Yang Liu, Zenghui Ding, Yining Sun, and Qingsong Wen. 2025b. Circuitprobe: Tracing visual temporal evidence flow in video language models. *arXiv preprint arXiv:2507.19420*.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-NeXT: A strong zero-shot video understanding model. Blog post. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

A Psychometric Functions

Figure 4 shows psychometric functions (balanced accuracy vs. log-interval) with sigmoid fits for models achieving $d' > 1$. Molmo2 exhibits a steep sigmoid with a threshold (θ) of 50 ms—within the range of human temporal order thresholds for visual stimuli (20–60 ms; Hirsh and Sherrick 1961). VideoChat-Flash shows a similar threshold ($\theta = 50$ ms) but its extreme NO bias ($c = 2.03$) prevents this sensitivity from translating into balanced behavioral accuracy. Models with $d' < 1$ remain flat at BAcc $= 0.50$ across all intervals, confirming zero temporal discrimination regardless of stimulus difficulty.

B Layer-wise Probing

Figure 5 shows probe accuracy across normalized layer depth for both encoder and LLM stages.

Model	$P(A)$	$P(B)$	$P(C)$	$P(D)$	Acc	BAcc	κ_w	H
Molmo2	.697	.000	.215	.088	.617	.455	.76	1.15
VideoChat-Flash	1.00	.000	.000	.000	.462	.250	.00	0.00
Qwen2.5-VL	.490	.510	.000	.000	.455	.402	.19	1.00
InternVL2.5	.273	.727	.000	.000	.358	.350	.18	0.85
LLaVA-NeXT-Video	.000	1.00	.000	.000	.154	.250	.00	0.00
Video-LLaMA2	.871	.000	.000	.129	.344	.186	−.27	0.56

Table 3: Interval estimation (E2). $P(A)$–$P(D)$: response proportions for each ordinal bin. κ_w: quadratic-weighted Cohen's kappa. H: response entropy in bits (max 2.0 for 4 classes). VideoLLaMA3 excluded (no parseable output).

Model	P(YES)	Acc	BAcc	d'	Strategy
Molmo2	0.000	.833	.500	0.51	Always NO
VideoChat-Flash	0.000	.833	.500	0.51	Always NO
Qwen2.5-VL	1.000	.167	.500	−0.51	Always YES
InternVL2.5	0.000	.833	.500	0.51	Always NO
LLaVA-NeXT-Video	0.000	.833	.500	0.51	Always NO
Video-LLaMA2	1.000	.167	.500	−0.51	Always YES
VideoLLaMA3‡	–	–	–	–	–

Table 4: Simultaneity detection (E3). All models adopt a fixed response, yielding BAcc $= 0.500$. The d' values are loglinear correction artifacts on degenerate response patterns. ‡No parseable output.

Model	BAcc_8	d'_8	BAcc_{16}	d'_{16}	BAcc_{30}	d'_{30}
Molmo2	.842	3.63	.904	4.05	.891	3.92
VideoChat-Flash	.514	1.29	.542	1.78	.597	2.33
Qwen2.5-VL	.502	0.53	.518	1.37	.509	1.13
InternVL2.5	.500	0.02	.500	−0.02	.500	0.06
LLaVA-NeXT-Video	.500	0.02	.500	−0.02	.500	0.06
Video-LLaMA2	.500	−0.02	.500	0.02	.500	−0.06

Table 5: Frame rate control (E4). Balanced accuracy and d' for temporal order at 8, 16, and 30 fps. Bias-dominated models (bottom three) give identical responses at all frame rates; their near-zero d' fluctuations reflect loglinear correction on varying class sizes, not behavioral change.

In the encoder (panel a), most models reach perfect or near-perfect probe accuracy by mid-depth. Notably, Qwen2.5-VL (∼0.84) and InternVL2.5 (∼0.88) start lower at layer 0 but converge to 1.00, suggesting temporal information is progressively refined; LLaVA-NeXT-Video begins at 1.00 but dips to ∼0.80 at mid-depth before recovering. In the LLM (panel b), all models except Video-LLaMA2 maintain ≥ 0.92 accuracy across all layers—temporal information, once encoded, persists through the entire language model. Video-LLaMA2 is flat at 0.46 across all LLM layers, consistent with its Q-Former discarding temporal information before it reaches the LLM.

C Interval Estimation

Table 3 reports per-model response distributions $P(A)$–$P(D)$ over the four ordinal bins, along with response entropy H (in bits; max 2.0 for 4 classes) measuring how spread responses are across categories. Molmo2 distributes responses across three bins ($H = 1.15$), achieving above-chance balanced accuracy (0.455 vs. 0.250 chance) and strong ordinal agreement ($\kappa_w = 0.76$). All other models collapse to one or two categories ($H \leq 1.00$): VideoChat-Flash and LLaVA-NeXT-Video always predict a single bin ($H = 0.00$), while Qwen2.5-VL and InternVL2.5 split responses between A and B only.

Figure 6 shows confusion matrices for interval estimation (E2). Molmo2 is the only model exhibiting a discriminative pattern: short intervals (A: <1 s) are classified near-perfectly, with orderly confusion between adjacent bins at longer durations.

D Simultaneity Detection

Table 4 provides the full simultaneity detection results summarized in §3. The d' values of ± 0.51 are artifacts of the loglinear correction applied to degenerate $0/N$ or N/N hit and false-alarm rates, not indicative of genuine sensitivity. The 83.3% raw accuracy for always-NO models reflects class imbalance (100 simultaneous vs. 500 non-simultaneous stimuli), further illustrating how raw accuracy overstates competence when response distributions are skewed.

E Frame Rate and Response Bias

Table 5 expands the frame rate results from §3. Notably, VideoChat-Flash shows monotonically increasing d' with frame rate ($1.29 \rightarrow 1.78 \rightarrow 2.33$), suggesting it benefits from finer temporal sampling—unlike Molmo2, whose sensitivity plateaus by 16 fps. Bias-dominated models (InternVL2.5, LLaVA-NeXT-Video, Video-LLaMA2) remain at BAcc $= 0.500$ regardless of frame rate; their near-zero d' fluctuations reflect loglinear correction artifacts on varying class sizes across independently generated datasets, not behavioral change.

Figure 7 shows the effect of frame rate on temporal order discrimination. Molmo2 maintains high balanced accuracy across 8, 16, and 30 fps (panel a), with the psychometric curve shape preserved at all frame rates. Panel (b) confirms that bias-dominated models remain flat at BAcc $= 0.50$ at 30 fps regardless of interval, consistent with fixed-response behavior.

Figure 8 summarizes response biases across

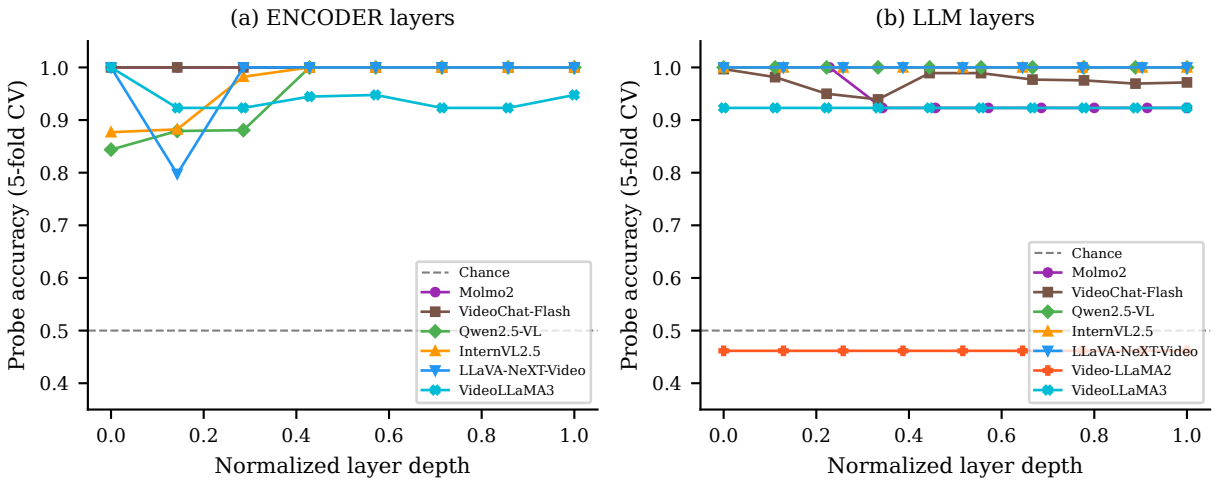

Figure 5: Probe accuracy across normalized layer depth (E5). (a) In the encoder, most models reach perfect accuracy by mid-depth; Qwen2.5-VL and InternVL2.5 start lower at layer 0 (∼0.84 and ∼0.88) and converge, while LLaVA-NeXT-Video dips to ∼0.80 at mid-depth before recovering. (b) In the LLM, temporal information persists through all layers for 6/7 models. Video-LLaMA2 remains flat at 0.46 (above the 0.25 chance floor), consistent with its Q-Former discarding temporal signal.

all three tasks. For temporal order (panel a), only Molmo2 produces balanced responses ($P(\text{YES}) \approx 0.57$); all others are at the extremes. For simultaneity (panel b), every model adopts a fixed response. For interval estimation (panel c), most models collapse to a single bin, with only Molmo2 distributing responses across categories A, C, and D.

F Prompt Sensitivity

Table 6 reports SDT metrics across three prompt variants for 4/7 models. Bias-dominated models maintain BAcc $= 0.500$ on both the original and reversed variants, confirming zero sensitivity regardless of question direction. However, the forced-choice format flips bias polarity entirely for two models: InternVL2.5 shifts from always-NO ($c = +3.17$) to near-always-YES ($c = -2.26$), and Video-LLaMA2 mirrors this in reverse—yet BAcc remains at chance in both cases. Qwen2.5-VL shows a modest BAcc improvement under forced format ($0.509 \rightarrow 0.606$), suggesting the format may unlock weak latent sensitivity. Only Molmo2 maintains genuine discriminability across all variants ($d' = 1.54$–3.92), with the forced format yielding the most balanced criterion ($c = -0.22$) and highest BAcc (0.956).

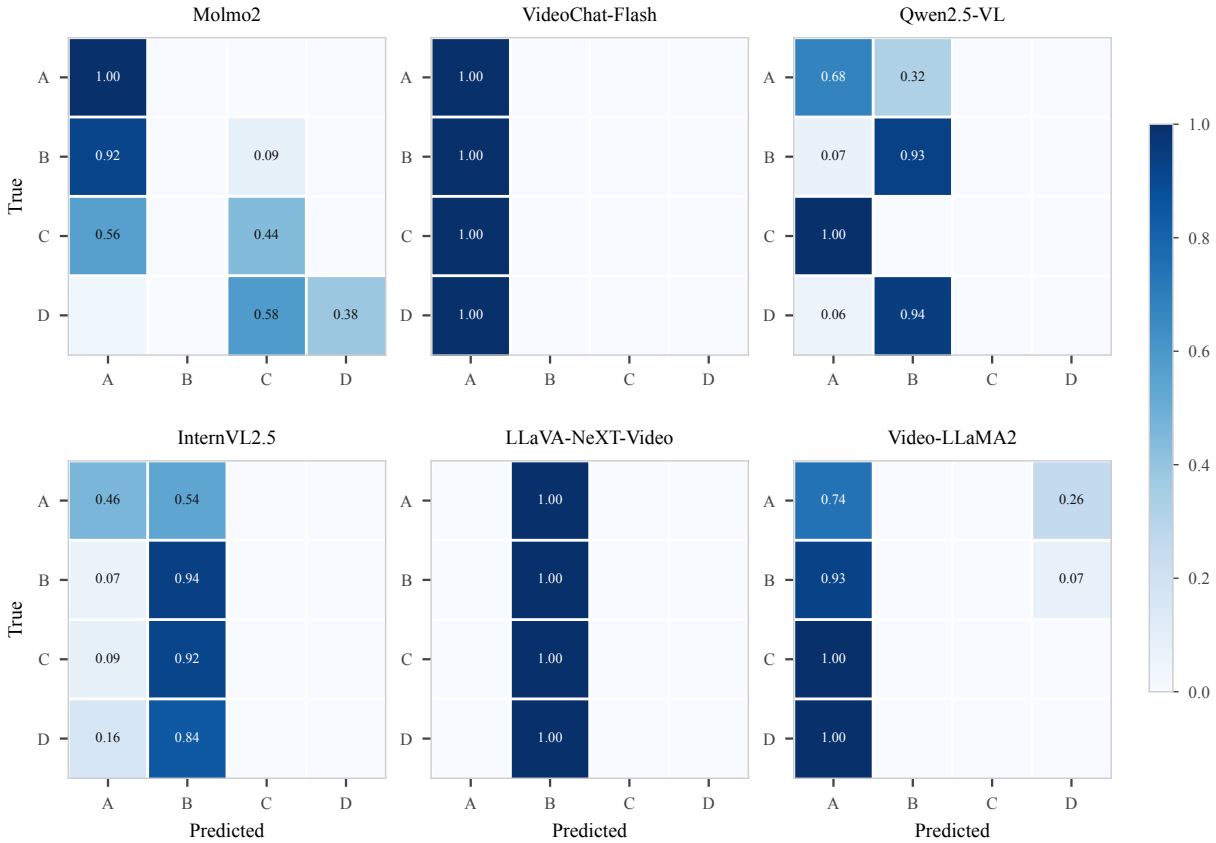

Figure 6: Interval estimation confusion matrices (E2). Molmo2 is the only model with a discriminative diagonal pattern; all others collapse to one or two bins regardless of the true interval.

Model	**Variant**	$P(\text{YES})$	**Acc**	**BAcc**	d'	c
Molmo2	Original	0.572	.881	.891	3.92	-1.18
	Reversed	0.775	.648	.676	1.54	-1.06
	Forced	0.476	.955	.956	3.49	-0.22
InternVL2.5	Original	0.000	.548	.500	0.06	$+3.17$
	Reversed	0.000	.548	.500	0.06	$+3.17$
	Forced	0.959	.412	.455	-1.86	-2.26
Qwen2.5-VL	Original	0.008	.556	.509	1.13	$+2.63$
	Reversed	0.477	.511	.509	0.04	$+0.06$
	Forced	0.370	.618	.606	0.57	$+0.32$
Video-LLaMA2	Original	1.000	.452	.500	-0.06	-3.17
	Reversed	1.000	.452	.500	-0.06	-3.17
	Forced	0.000	.548	.500	0.06	$+3.17$

Table 6: Prompt sensitivity (E1, 4/7 models). Three variants: original ("Did A move BEFORE B?"), reversed ("Did B move BEFORE A?"), and forced ("Which moved first, A or B?", requiring an A/B choice). d' and c use loglinear correction. $c > 0$: NO-bias; $c < 0$: YES-bias. Near-zero d' for InternVL2.5 and Video-LLaMA2 reflects loglinear correction on degenerate responses, not genuine sensitivity.

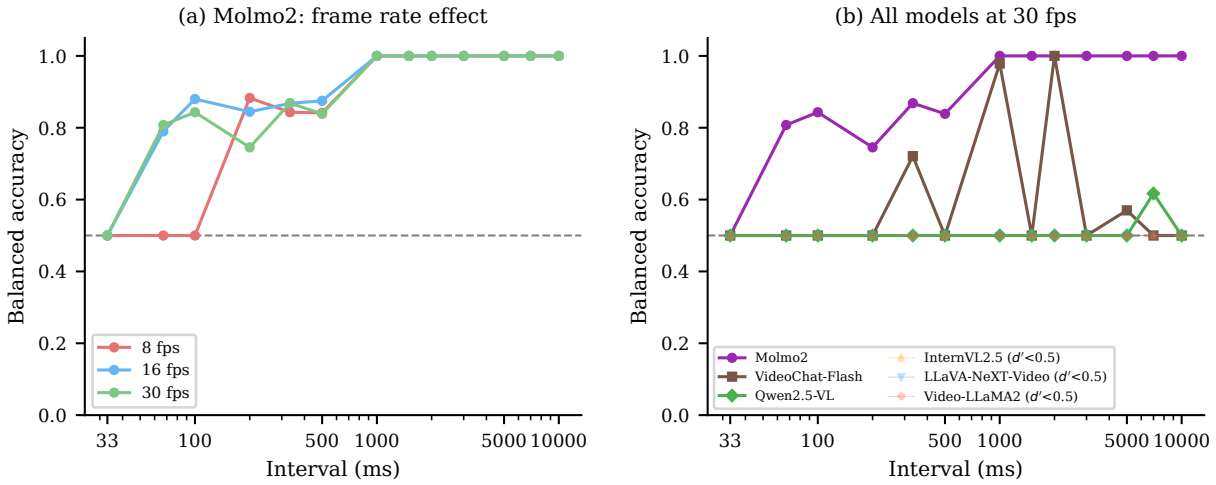

Figure 7: Frame rate robustness (E4). (a) Molmo2 retains high balanced accuracy across 8, 16, and 30 fps, with consistent psychometric curve shapes. (b) At 30 fps, only Molmo2 and VideoChat-Flash produce above-chance accuracy patterns; the rest are flat.

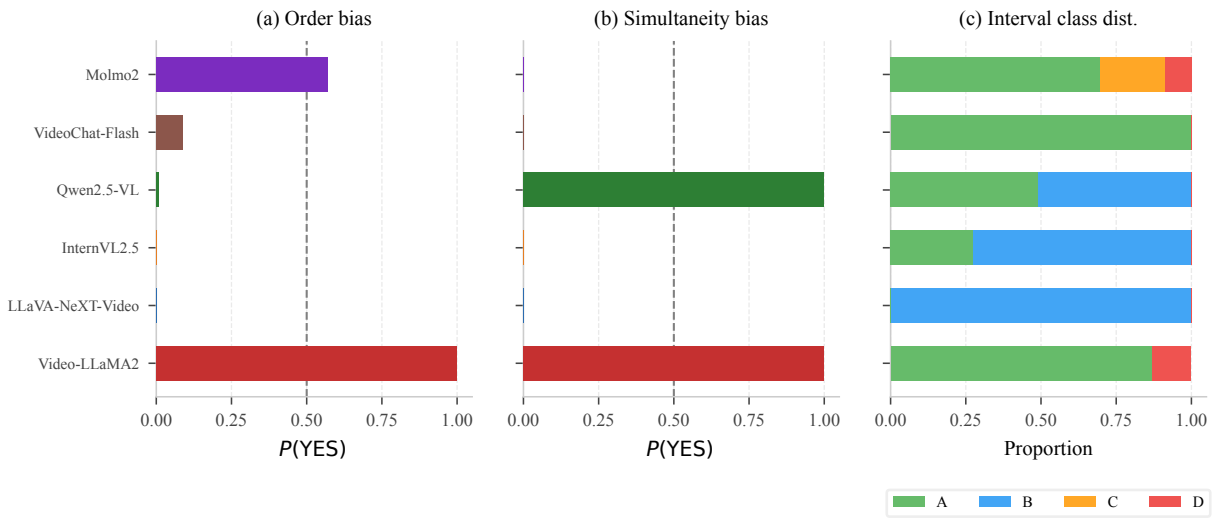

Figure 8: Response bias across tasks. (a) For temporal order, most models answer exclusively YES or NO; only Molmo2 gives balanced responses. (b) For simultaneity, every model adopts a fixed response. (c) For interval estimation, most models select a single bin regardless of the true duration.