THE LOW-FREQUENCY TRAP: WHY SCALING DOESN'T SOLVE SIMPLE TEMPORAL COUNTING

Sarvesh Baskar * **Muhammad R. Islam** * **Zikui Cai** †
Ankit Nakhawa **Anirudh Satheesh** **Tom Goldstein** **Furong Huang**
University of Maryland, College Park
zikui@umd.edu

ABSTRACT

Large multimodal models demonstrate strong performance on complex video understanding benchmarks, leading to the expectation that they should trivialy handle simple temporal reasoning tasks. In this work, we show that this assumption is fundamentally flawed. Using parametric profiling – systematically varying event frequency, event count, and temporal span – we uncover a striking failure mode: state-of-the-art video–language models fail catastrophically on conceptually simple tasks. While performance generally degrades as event frequency increases (as expected), we observe a counter-intuitive collapse in the easy regime: even at low frequencies (0.5 – 1 Hz) with visually distinct events, performance plummets once the event count exceeds a trivial threshold (e.g., $N > 4$). Moreover, scaling model size from 8B to 235B does not resolve this limitation; large and small models exhibit nearly identical capability boundaries. Our analysis suggests that errors arise not from high-level reasoning or counting per se, but from systematic temporal misinterpretation, including event merging, hallucinated intermediate states, and color-based temporal interpolation. These results reveal a blind spot in current models' temporal abstraction that is masked by aggregate benchmark scores and largely invariant to scale. Our findings highlight the need for diagnostic evaluation beyond average accuracy and suggest that scaling alone is insufficient to resolve fundamental limitations in temporal event reasoning.

1 INTRODUCTION

While recent multimodal foundation models claim near-human performance on complex video benchmarks (Google, 2025; OpenAI, 2025; Bai et al., 2025; Wang et al., 2025), they harbor a paradoxical blindness to the simplest of tasks. In this work, we focus on simple, visually unambiguous temporal tasks—such as counting discrete events in videos with low event frequency and small event counts—that, under standard assumptions, should be strictly easier than the complex scenarios where modern models already excel. To probe these assumptions, we adopt parametric profiling, a diagnostic evaluation methodology in which task difficulty is varied along interpretable axes (e.g., event frequency, number of events, and temporal span) while holding other factors constant. This approach allows us to examine not just whether models succeed on average, but where and how their performance degrades.

While we observe the expected trend that higher-frequency events (which are visually faster and harder to segment) lead to lower performance, we uncover a critical anomaly at the other end of the spectrum. Models struggle to maintain accurate counts even when events are slow, distinct, and widely spaced. specifically, once the number of events exceeds a small number (e.g., 4), accuracy drops drastically, even in low-frequency settings (< 1 Hz) where temporal crowding is non-existent.

Even more strikingly, we find that scaling model size does not reliably resolve these failures. A 235B-parameter model does not meaningfully outperform an 8B-parameter model under parametric profiling, exhibiting nearly identical capability boundaries and error patterns. While larger models often achieve higher average accuracy, they do not expand the range of temporal configurations they

* Equal contribution. † Corresponding author.

can robustly handle. This suggests that the observed limitations are not due to insufficient capacity or optimization, but rather reflect deeper representational or inductive biases shared across scales.

Our results challenge the assumption that temporal reasoning abilities naturally improve with scale and highlight a blind spot in current evaluation practices (Tong et al., 2024). Standard benchmarks, which emphasize average-case performance on fixed datasets, may fail to reveal such boundary behaviors, allowing fundamental limitations to go unnoticed (Agarwal et al., 2025). By contrast, parametric profiling exposes non-monotonicity, scale invariance, and structured failure modes that are critical for understanding what current models can and cannot do.

In summary, this paper makes three contributions:

- **Temporal capability profiling**: we introduce a parametric evaluation framework that systematically varies event frequency and event count to probe temporal reasoning boundaries;
- **Low-Frequency Trap**: we identify a counter-intuitive failure regime where VLMs fail to count sparse events even when visual perception is trivial;
- **Scale-invariant limitation**: we show that this capability boundary persists across model scales and multiple VLM architectures.

Together, these findings underscore the need for diagnostic evaluation beyond leaderboard metrics and suggest that scaling alone is insufficient to address fundamental limitations in temporal event understanding.

2 Parametric Profiling

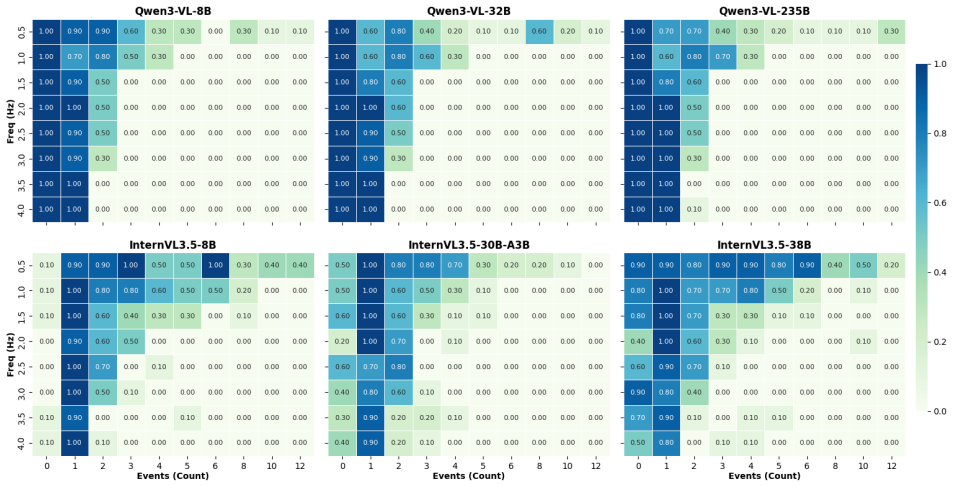

Figure 1: Visualizing the Low-Frequency Trap. Heatmaps show success rates of Qwen3-VL and InternVL3.5 families for counting state transitions across event frequencies and counts. While high-frequency degradation is expected, models also suffer a catastrophic collapse at the bottom of the spectrum: failing to count events once $N > 4$, even at slow speeds (0.5 Hz).

We evaluate temporal event reasoning by profiling models on a controlled sweep of event frequency and event count for state-transition video tasks (Figure 1). To generate our evaluation data, we build upon the MORSE framework (Cai et al., 2025), which is a programmatically controllable video benchmark that provides robust deterministic generation. Leveraging its foundational generation capabilities, we designed minimalistic, conceptually simple video tasks—specifically State Transition, Blinking, and Bouncing Ball—that are highly optimized for parametric profiling. For visual examples of the generated frames for each task, please refer to Figure 4. A comprehensive discussion of our parametric difficulty control can be found in Appendix D.

Scenes are intentionally simple (single salient object, minimal clutter, no occlusion, deterministic dynamics) so failures cannot be attributed to visual complexity or ambiguous semantics. These tasks

are diagnostic probes, not realistic benchmarks, designed to isolate whether models can segment continuous input into discrete events and reason over temporal structure.

Crucially, lower values along these axes correspond to tasks that should be strictly easier under standard assumptions about temporal reasoning and counting. This parametric formulation allows us to examine not only average performance, but also how accuracy evolves as a function of each parameter, revealing non-linearities and boundary effects that are invisible to aggregate metrics.

3 Negative Results

3.1 The "Easy" Tasks Are Not Solved

Figure 1 presents model performance across the parameter space defined by event count and frequency. Standard intuition suggests that performance should be robust at low frequencies (where events are slow and clearly separated) and degrade only as frequency increases (where events become rapid and blurred). Even when events are slow, distinct, and visually trivial (0.5 Hz), models fail to count beyond a small threshold (e.g., $N = 4$). This reveals a 'Simplicity Ceiling': reducing perceptual difficulty does not unlock better reasoning capabilities. The model's inability to count to 5—despite having ample time and clear frames—proves that the bottleneck is not visual perception, but a fundamental deficit in temporal state tracking.

3.2 Scaling Does Not Improve Boundary Behavior

We compare the performance of the 8B to 235B models across the same parameter grid in Figure 1. While the larger model often achieves higher average accuracy, it does not expand the region of the parameter space where the task can be solved reliably. The boundary between success and failure remains largely unchanged, and the two models exhibit strikingly similar performance contours.

In particular, the low-frequency, low-event-count regime that induces failures in the smaller model also induces failures in the larger model. The large model does not recover performance in these regions, nor does it exhibit smoother degradation. Instead, both models fail abruptly and in qualitatively similar ways.

This result suggests that increased parameter count improves performance within the model's existing capability envelope but does not fundamentally alter how temporal structure is represented. The persistence of identical failure modes across scales indicates that these limitations are unlikely to be resolved through scaling alone.

4 Analysis: Why Is It Not Better?

The negative results above raise a natural question: why do models fail most severely on conceptually simple temporal configurations, and why does scale not alleviate this failure? In this section, we analyze model predictions and error patterns to identify consistent mechanisms underlying these behaviors.

Temporal Interpolation Hypothesis We hypothesize that current multimodal models often represent temporal dynamics as continuous appearance interpolation rather than as sequences of discrete events. Under this view, models encode changes in visual attributes (e.g., position, color, or state) as smooth transitions over time, without explicitly segmenting the video into event boundaries.

This hypothesis explains several aspects of the observed behavior. When events occur frequently, continuous interpolation provides sufficient signal to approximate event counts or transitions. However, when events are sparse and separated by long temporal gaps, interpolation-based representations collapse: discrete events are merged, skipped, or hallucinated, leading to systematic counting and ordering errors.

To test this hypothesis, we examine model predictions on videos where the same discrete event occurs under different temporal spacings while all frames and object appearances remain unchanged. We find that increasing the temporal gap between events—without altering the visual evidence—leads to a higher incidence of event merging and miscounting. This sensitivity to temporal spacing, rather

than visual content, supports the interpretation that models rely on implicit temporal smoothing rather than explicit event tracking.

Color and Event Misinterpretation Evidence A particularly revealing failure mode involves color-based event reasoning. In tasks where object color changes discretely at event boundaries (e.g., color toggles upon contact), models frequently predict intermediate or blended colors that never appear in the video. These predictions persist even when the queried color corresponds to a single, clearly visible frame.

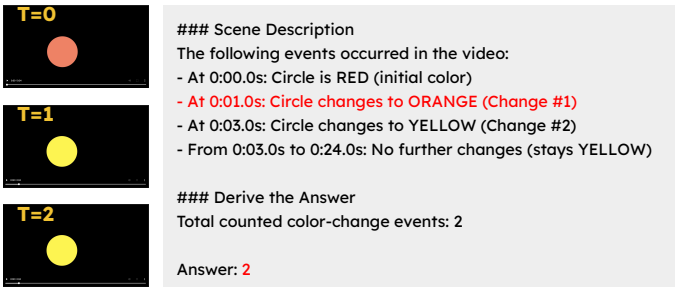

Figure 2: Example of the structured reasoning-trace format for a state-transition counting task. Sampled frames (T=0–2) show the circle changing color (red → orange → yellow), while the event log summarizes the detected transitions and yields a total of 2 color-change events.

Figure 2 illustrates representative examples in which models hallucinate temporally interpolated colors or attribute a color change to the wrong event. Importantly, these errors occur regardless of model scale and are more prevalent in low-frequency settings. This suggests that models do not reliably bind discrete visual attributes to specific temporal events, but instead infer attributes from smoothed trajectories over time.

Beyond color, we observe analogous patterns in event identity and order: models conflate successive events, attribute properties from one event to another, or infer additional events that are not present. Taken together, these behaviors indicate that failures arise from temporal misrepresentation, not from an inability to recognize objects, colors, or language instructions in isolation.

Boundary Conditions and Failure Regimes The observed failures exhibit sharp and reproducible boundary conditions. Performance degrades abruptly once event frequency falls below a task-dependent threshold, even when event count remains small. Conversely, increasing event frequency often restores performance, despite increasing nominal task difficulty.

These boundaries are consistent across models, prompts, and random seeds, suggesting that they reflect inherent representational limits rather than optimization instability. Notably, the boundary locations do not shift meaningfully with model scale, reinforcing the conclusion that additional capacity improves performance only within an existing representational regime.

Crucially, we do not observe gradual degradation near the boundary. Instead, performance transitions rapidly from near-perfect to near-random, indicating that models either successfully approximate the task using interpolation-based cues or fail entirely when those cues break down. This all-or-nothing behavior further supports the hypothesis that current models lack a robust mechanism for explicit temporal event abstraction.

5 Conclusion

We present a counter-intuitive negative result: state-of-the-art VLMs, regardless of scale, fail to count simple events once they are slow and widely spaced. This "Low-Frequency Trap" reveals that current architectures rely on continuous temporal interpolation rather than discrete state tracking. Our findings suggest that leaderboard saturation on standard benchmarks masks fundamental deficits in temporal reasoning, and that scaling alone is insufficient to bridge this gap without new inductive biases for object persistence and discrete causality.

REFERENCES

Amit Agarwal, Hitesh Laxmichand Patel, Srikant Panda, Hansa Meghwani, Jyotika Singh, Karan Dua, Paul Li, Tao Sheng, Sujith Ravi, and Dan Roth. Rci: A score for evaluating global and local reasoning in multimodal benchmarks, 2025. URL https://arxiv.org/abs/2509.23673.

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, 2025.

Zikui Cai, Andrew Wang, Anirudh Satheesh, Ankit Nakhawa, Hyunwoo Jae, Keenan Powell, Minghui Liu, Neel Jay, Sungbin Oh, Xiyao Wang, et al. Morse-500: A programmatically controllable video benchmark to stress-test multimodal reasoning. 2025. URL https://arxiv.org/abs/2506.05523.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. URL https://arxiv.org/abs/2403.20330.

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, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding, 2026. URL https://arxiv.org/abs/2601.10611.

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, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. URL https://arxiv.org/abs/2405.21075.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024. URL https://arxiv.org/abs/2404.12390.

Google. A new era of intelligence with gemini 3, 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-3/#note-from-ceo. Accessed: 2026-02-01.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. In Proceedings of the 42nd International Conference on Machine Learning (ICML 2025), 2025. URL https://arxiv.org/abs/2501.05444.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models, 2024. URL https://arxiv.org/abs/2402.07865.

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. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024. URL https://arxiv.org/abs/2311.17005.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos?, 2024. URL https://arxiv.org/abs/2403.00476.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL https://arxiv.org/abs/2310.02255.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023. URL https://arxiv.org/abs/2308.09126.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2025. URL https://arxiv.org/abs/2410.05229.

OpenAI. Introducing gpt-5, 2025. URL https://openai.com/index/introducing-gpt-5/. Accessed: 2026-02-01.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025. URL https://arxiv.org/abs/2506.06941.

Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. Video-mmlu: A massive multi-discipline lecture understanding benchmark, 2025. URL https://arxiv.org/abs/2504.14693.

Yiyou Sun, Shawn Hu, Georgia Zhou, Ken Zheng, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization, 2025. URL https://arxiv.org/abs/2506.18880.

Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, Chaoran Tao, Zhiyuan Guo, Jizhou Yu, Tianhao Cheng, Zhiheng Xi, Changhao Jiang, Zhangyue Yin, Yining Zheng, Weifeng Ge, Guanhua Chen, Tao Gui, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. Game-rl: Synthesizing multimodal verifiable game data to boost vlms' general reasoning, 2025. URL https://arxiv.org/abs/2505.13886.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265, 2025.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

A Cross-Architecture Temporal Capability Profiling

To examine whether the observed failure patterns are specific to a single model family, we repeat the temporal capability profiling experiments on additional state-of-the-art vision–language model architectures. In addition to the Qwen3-VL (Bai et al. (2025)) models discussed in the main text, we evaluate two independent VLM families: InternVL3.5 (Wang et al. (2025)) and Molmo2 (Clark et al. (2026)).

The evaluation protocol remains identical across all models. For each configuration defined by event frequency and number of events, we generate synthetic videos using the same programmatic generators and evaluate models using the same zero-shot instruction-following prompt. Performance is measured as exact-match accuracy on the final answer, averaged across multiple randomized scene variants. This controlled setup allows direct comparison of capability boundaries across architectures without confounding differences in task distribution or visual complexity.

Figure 4 illustrates representative frames for the three temporal tasks used in our evaluation: *State Transition*, *Blinking*, and *Bouncing Ball*. These tasks are intentionally minimalistic, containing a

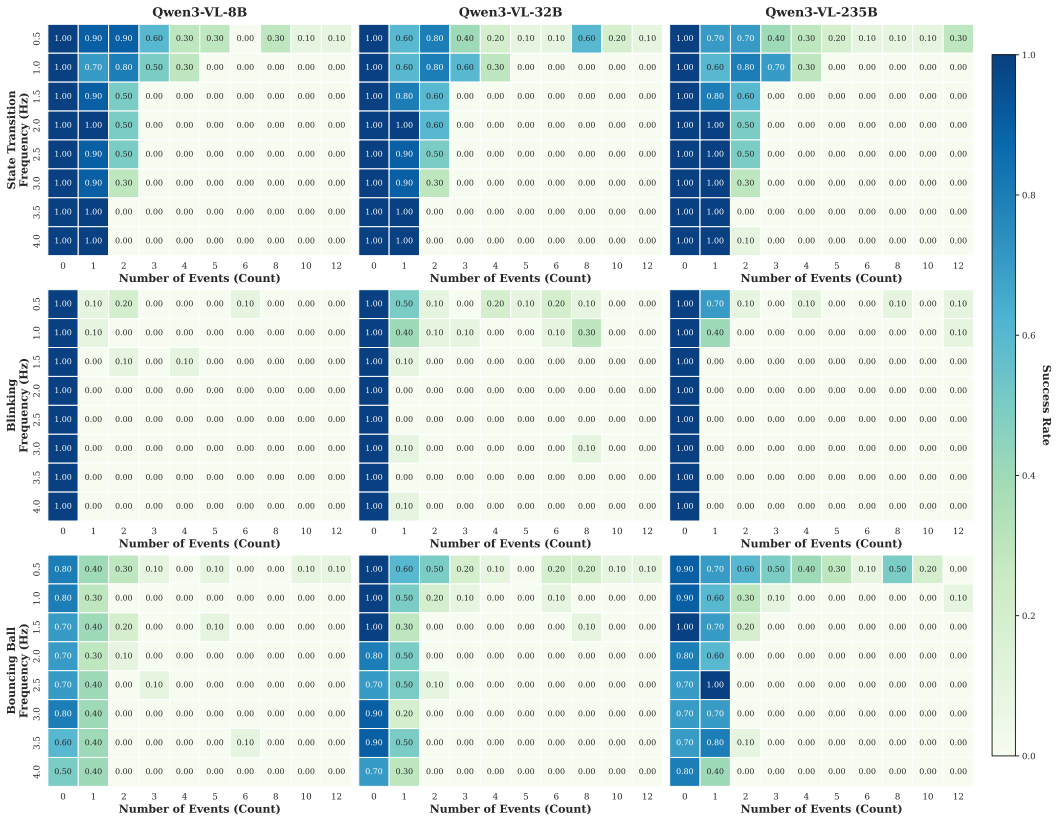

Figure 3: Visualizing the Low-Frequency Trap. Heatmaps show success rates of Qwen3-VL across three scales (8B, 32B, 235B) for counting state transitions across event frequencies and counts. While high-frequency degradation is expected, models also suffer a catastrophic collapse at the bottom of the spectrum: failing to count events once $N > 4$, even at slow speeds (0.5 Hz).

single salient object and discrete event transitions, ensuring that failures cannot be attributed to visual clutter or semantic ambiguity.

Figures 3, 5, and 6 present the resulting temporal capability heatmaps across model scales within each architecture. Consistent with the results reported for Qwen3-VL, we observe similar capability boundaries across architectures. In particular, performance drops sharply once the number of discrete

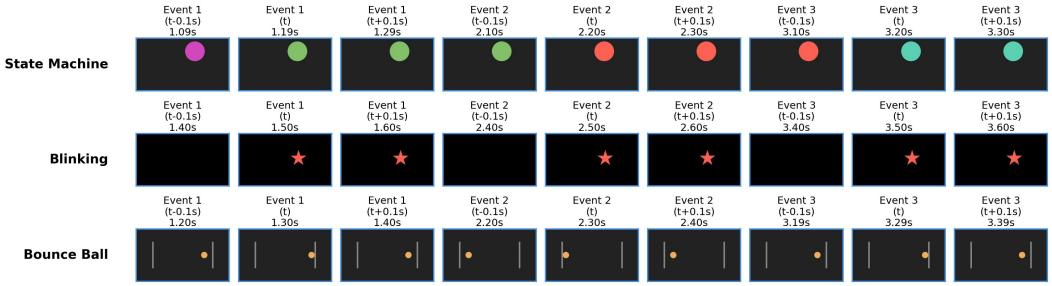

Figure 4: **Representative frames from the three simple temporal tasks.** Each row shows a different task: *State Transition*, where an object changes among discrete color states; *Blinking*, where an object toggles on and off; and *Bouncing Ball*, where a ball undergoes repeated bounce events between boundaries. Columns visualize frames immediately before, at, and after each event time, illustrating that the events are visually simple, temporally localized, and unambiguous. These examples highlight the controlled nature of the tasks used for temporal capability profiling.

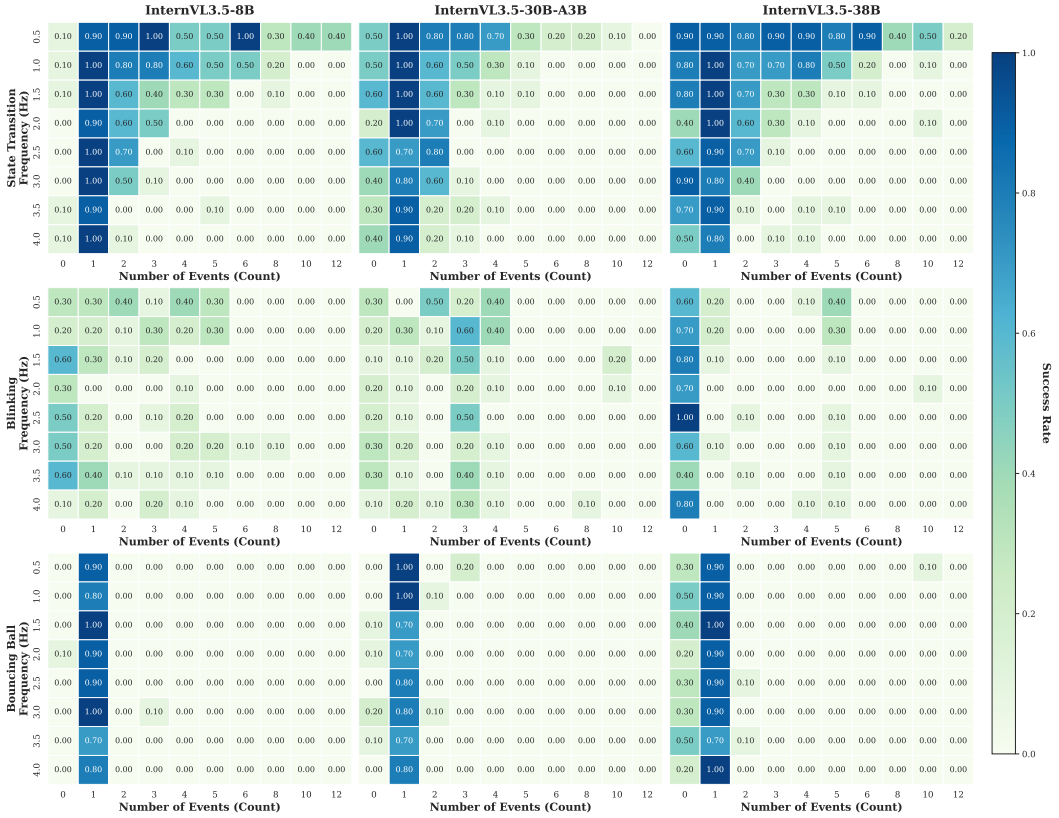

Figure 5: **Temporal capability heatmaps across models and tasks (Molmo2 & InternVL).** Exact-match success rate as a function of event frequency (Hz; y-axis) and number of events (x-axis) for three temporal tasks (State Transition, Blinking, Bouncing Ball). Columns compare InternVL3.5-8B, InternVL3.5-30B-A3B and InternVL3.5-38B

events exceeds a small threshold, even when events occur at low frequencies where perceptual segmentation should be straightforward.

While absolute accuracy varies across architectures and model sizes, the qualitative failure pattern remains consistent. Across all tested models, increasing parameter count generally improves performance within the region where the task is already solvable but does not substantially expand the boundary of configurations that can be handled reliably. In other words, larger models tend to achieve higher accuracy within the existing capability envelope rather than extending the range of temporal configurations that can be solved.

These cross-architecture results suggest that the low-frequency failure mode identified in Section 3 is not unique to a single VLM design. Instead, similar capability boundaries emerge across multiple independently developed architectures, indicating that the observed limitation may reflect broader representational biases in current multimodal models. At the same time, we emphasize that our study does not exhaustively evaluate all VLM families; rather, these results demonstrate that the phenomenon persists across several diverse model architectures trained with different design choices and scaling strategies.

B One possible failure analysis

C Related Work

Multimodal Benchmarks: Saturation and Shortcuts. Standard benchmarks face a dual challenge of saturation and ambiguity. Static suites like MathVista (Lu et al., 2024), MMMU (Yue et al., 2024),

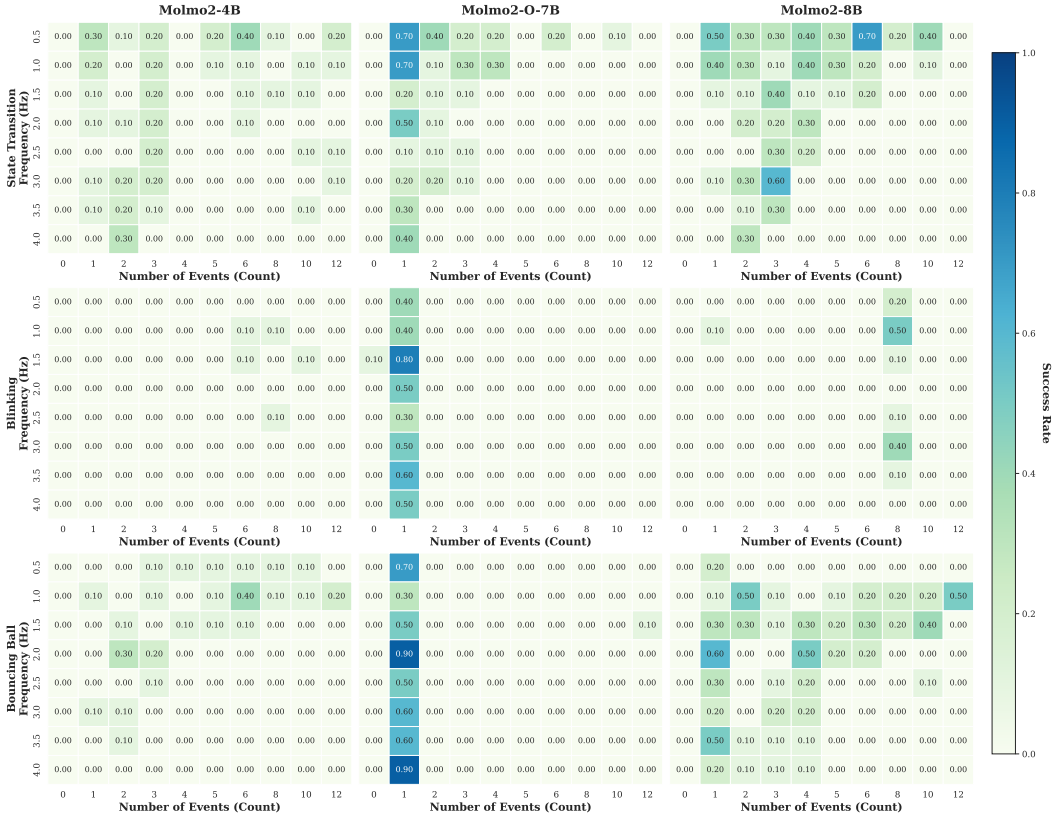

Figure 6: **Temporal capability heatmaps across models and tasks.** Exact-match success rate as a function of event frequency (Hz; y-axis) and number of events (x-axis) for three temporal tasks (State Transition, Blinking, Bouncing Ball). Columns compare Molmo2-4B, Molmo2-O-7B and Molmo2-8B

and MMStar (Chen et al., 2024) provide strong reasoning signals but ignore temporal dynamics. Conversely, video-centric benchmarks such as Video-MME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), and EgoSchema (Mangalam et al., 2023) introduce temporal context but often prioritize retrieval or holistic description. Crucially, recent work suggests many video" tasks can be solved via single-frame shortcuts, blurring the line between perception and reasoning (Song et al., 2025). While benchmarks like MVBench (Li et al., 2024) and TempCompass (Liu et al., 2024) attempt to isolate specific temporal skills, they remain black-box" evaluations: they assess *if* a model failed, but provide limited insight into *why*—whether due to perception (missing an object), memory (forgetting state), or reasoning (applying the wrong rule).

Controllable Complexity Evaluation. To address this opacity, evaluation is shifting toward *controllable complexity manipulation*. In the language domain, GSM-Symbolic (Mirzadeh et al., 2025) demonstrates that LLMs exhibit significant variance when mathematical templates are varied, exposing brittleness in seemingly robust capabilities. Similarly, Shojaee et al. (Shojaee et al., 2025) use controllable puzzle environments to reveal "accuracy collapse" in Large Reasoning Models beyond specific complexity thresholds. Frameworks like OMEGA (Sun et al., 2025) and Game-RL (Tong et al., 2025) further formalize this by evaluating generalization across specific axes (exploratory, compositional) or via synthesized game tasks. These approaches establish the foundation for our programmatic difficulty control: rather than relying on fixed crowd-sourced videos (Hao et al., 2025), reliable evaluation requires procedural control over difficulty parameters to prevent saturation.

Disentangling Perception from Reasoning. A critical limitation of end-to-end evaluation is the confounding of visual recognition and logical inference. BLINK (Fu et al., 2024) and Prismatic (Karamcheti et al., 2024) demonstrate that many SOTA failures stem from fundamental perception

deficits (e.g., assessing depth or spatial overlap) rather than reasoning failures. However, these diagnostics are largely static. In the temporal domain, disentanglement is harder; a model might fail to count an action because it missed the motion (perception) or because it lost the count (state tracking). RCI (Agarwal et al., 2025) attempts to separate these factors via intervention, but lacks a unified generation framework. MORSE extends this disentanglement to video by using **executable traces** to explicitly map the boundaries between what a model *sees* and how it *reasons* over time.

D PARAMETRIC PROFILING DETAILS

We formalize multimodal evaluation as capability profiling: measuring model performance as a function of interpretable capability demands, rather than as accuracy on a fixed test set. Standard benchmarks evaluate models at a small number of discrete points in the space of possible task configurations. In contrast, real-world deployment requires understanding how performance degrades as perceptual, temporal, and spatial demands increase. Capability profiling makes this dependency explicit and measurable.

Concretely, we view evaluation as estimating a function:

$$\mathcal{P}_\theta(\mathbf{x}) = \mathbb{E}_{\tau \sim \mathcal{D}(\mathbf{x})} \left[\mathbb{I}(\mathcal{M}(\tau) = y_\tau) \right] \tag{1}$$

where $\mathcal{P}_\theta(\mathbf{x})$ denotes the performance of model θ under the difficulty configuration $\mathbf{x} \in \mathbb{R}^d$ (composed of parameters such as frequency, object count, and spatial scale). Here, $\mathcal{D}(\mathbf{x})$ represents the task distribution that is programmatically generated and conditioned on $\mathbf{x}$, τ is a specific task instance sampled from this distribution, and $\mathbb{I}(\cdot)$ is the indicator function evaluating whether the model's prediction $\mathcal{M}(\tau)$ matches the ground truth y_τ. This reframing enables identification of capability boundaries where performance collapses, which we instantiate parametric difficulty control D.1.

D.1 PARAMETRIC DIFFICULTY CONTROL AND CAPABILITY AXES

While reasoning traces reveal where a model fails, they do not reveal how capability limits are structured. To expose these limits, we introduce parametric difficulty control, where tasks are generated by explicitly varying interpretable capability demands along independent axes. This transforms evaluation from a binary outcome into a structured mapping from capability demand to performance. We conceptualize multimodal reasoning as requiring a combination of interacting capabilities, including but not limited to:

- Perceptual resolution: the ability to perceive small objects, fine spatial details, or rapid visual changes.
- Temporal resolution: the ability to detect events that occur at increasing frequency or short duration.
- Temporal range (working memory): the ability to maintain and update state across long sequences of events.
- Object persistence: tracking entities through motion, occlusion, and transformation.
- Cardinality and counting: enumerating objects or events under varying visual and temporal conditions.

In natural benchmarks, these demands are typically entangled: as a question becomes harder, multiple capabilities degrade simultaneously, making it unclear which limitation is responsible for failure. Prior work has shown that performance collapses as complexity increases (e.g., Cai et al. (2025)), but such results cannot disentangle whether the bottleneck lies in perception, memory, or inference. Parametric difficulty control addresses this limitation by varying one capability at a time while holding others constant. This enables precise measurement of how performance changes as a function of specific demands and reveals capability thresholds, performance cliffs, and non-graceful degradation patterns that are invisible to aggregate accuracy.

D.2 DETAILED EXPERIMENTAL SETTING IN THIS PAPER

We programmatically render short clips from a deterministic generator while varying (i) the rate at which discrete state changes occur (e.g., 0.5–4.0 Hz) and (ii) the total number of state-change

events (e.g., 0–12). To reduce overfitting to surface appearance, we generate 10 seeded variants per cell that randomize nuisance visuals (e.g., color palette, object shape/size, and initial position or distances) while keeping the underlying dynamics and event schedule fixed. We run the same zero-shot, instruction-following prompt across Qwen3-VL (Bai et al. (2025)) and InternVL3.5 (Wang et al. (2025)) families, and ask a fixed event-level question (e.g., count the number of transitions). Ground-truth answers are computed directly from the executable trace. We score each trial by exact-match on the normalized final answer (number/label), and report success rate as the mean accuracy across seeds per configuration.

D.3 Why This Is Not a Benchmark Artifact

A natural concern is that the observed failures may be artifacts of synthetic data, prompt sensitivity, or mismatches between training and evaluation distributions. We emphasize that this setup is intentionally simpler than standard video understanding benchmarks along multiple dimensions: fewer objects, fewer events, clearer visual cues, and simpler language. Moreover, our conclusions do not rely on absolute accuracy values, but on relative trends across parameter settings and model scales. The key findings—non-monotonic performance and scale-invariant failure boundaries—persist across prompt variants, random seeds, and task instantiations. Because all task configurations differ only along explicitly controlled parameters, there is no confounding change in visual appearance, semantics, or linguistic complexity. Rather than replacing benchmarks, parametric profiling complements them by exposing capability boundaries that aggregate benchmark scores can mask. As such, the observed failures should be interpreted not as dataset flaws, but as evidence of systematic limitations in how current models represent and reason about temporal structure.