TimePerceptBench: A Comprehensive Evaluation of Temporal Reasoning in Large Vision-Language Models

Anonymous CVPR submission

Paper ID *****

Abstract

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities in static image understanding. However, their ability to perceive and reason about temporal dynamics—such as the sequence of events, duration estimation, and causal relationships in video data—remains under-explored. In this paper, we introduce TimePerceptBench, a comprehensive benchmark designed to systematically evaluate the temporal reasoning capabilities of state-of-the-art LVLMs. We propose a novel metric, the Temporal Alignment Score (TAS), to quantify the synchronization between visual perception and textual generation. Our extensive experiments covering five leading models (including Intern-VL [8] and Qwen3-VL [2]) across six diverse tasks reveal significant limitations in current architectures. Specifically, while models excel at object recognition, they struggle with Sequential Order Verification (SOV) and Temporal Anomaly Localization (TAL), often exhibiting severe hallucination. We further analyze the impact of fine-tuning strategies and propose a memory-augmented attention mechanism [7] that improves temporal consistency by 15%. This work provides a foundation for future research in developing temporally aware multimodal AI systems.

Index Terms—Large Vision-Language Models, Temporal Reasoning, Video Understanding, Benchmark, Multimodal Learning.

1. Introduction

The ability to reason about time is a fundamental aspect of human intelligence. When we observe the world, we do not merely see a sequence of disjointed snapshots; rather, we perceive a continuous flow of events linked by causality, physics, and temporal logic. For artificial intelligence systems, particularly Large Vision-Language Models (LVLMs), mastering this capability is the holy grail of video understanding. While recent advancements in models like GPT-4V [1] and Gemini have demonstrated near-human performance in static image captioning and visual question answering (VQA), their performance precipitously drops when tasked with understanding the 'arrow of time'.

Consider a simple video of a glass falling off a table. A human effortlessly understands the sequence: the glass is on the table, it is pushed, it falls, and finally, it shatters. However, current LVLMs often hallucinate events, confuse the cause (pushing) with the effect (shattering), or fail to estimate the duration of the fall. This limitation severely hampers the deployment of AI in safety-critical domains such as autonomous driving, where predicting the future trajectory of a pedestrian based on past movements is a temporal reasoning task.

In this paper, we argue that the primary bottleneck is not the model architecture itself, but the lack of high-quality, temporally annotated training data. Existing datasets often rely on noisy web-scraped video-text pairs where the text describes the visual content generally but lacks precise temporal grounding. To address this, we present TimePerceptBench [4], a rigorous benchmark constructed to isolate and evaluate specific temporal faculties: ordering, duration, and causality.

2. Related Works

2.1. Vision-Language Pre-training

The paradigm of pre-training on massive scale image-text pairs has revolutionized the field. CLIP established the viability of contrastive learning for aligning visual and textual embedding spaces. Subsequent works like BLIP and LLaVA [5] extended this by introducing instruction tuning, allowing models to follow complex human queries. However, these models process images as static tensors. When applied to video, they typically employ a 'frame-averaging' strategy, which inevitably results in the loss of temporal granularity [6]. Our work builds upon these foundations but introduces a dedicated temporal alignment head to preserve sequential information.

2.2. Temporal Reasoning Benchmarks

Several benchmarks have been proposed to evaluate video understanding, such as Kinetics-400 for action recognition and ActivityNet for temporal localization. However, these datasets largely focus on classification (e.g., 'is this person running?'). They do not test the logical consistency of the model's internal world model. For instance, few benchmarks ask, 'Did the person open the door before or after picking up the bag?' Recent attempts like VideoChat [3] and NExT-QA have started to explore causal logic, but they remain limited in scale and diversity. TimePerceptBench fills this gap by introducing 10,000 carefully curated logic queries.

3. Benchmark Construction

To ensure the robustness of our evaluation, the TimePerceptBench was built focusing on diversity and difficulty.

Constructing a benchmark for temporal reasoning requires meticulous attention to detail. We employed a three-stage pipeline to ensure data quality and logical validity.

Stage 1: Raw Video Filtering. We sourced raw videos from diverse domains including ego-centric views (Ego4D), instructional videos, and movie clips. We used an automated scene cut detection algorithm based on color histogram differences to segment long videos into coherent atomic events. This resulted in a pool of 50,000 clips.

Stage 2: Automatic Annotation Generation. We utilized a teacher model (GPT-4 [1]) prompted with frame-by-frame textual descriptions to generate candidate temporal questions. The prompts were designed to target specific logical structures, such as 'Sequence: A then B', 'Simultaneity: A while B', and 'Duration: How long did A last?'.

Stage 3: Human Verification. To eliminate hallucinations generated by the teacher model, we employed human annotators to verify the ground truth. Annotators were presented with the video and the generated query-answer pair and asked to label them as 'Correct', 'Ambiguous', or 'Incorrect'. Only samples with unanimous agreement among three annotators were retained. This rigorous process yielded the final 10,000 samples used in TimePerceptBench.

4. Methodology

We formulate the temporal reasoning task as finding the optimal alignment between a visual sequence V and a textual query Q.

4.1. Mathematical Formulation

We define the Temporal Alignment Score (TAS) using a Gaussian-weighted intersection function. Let the predicted interval be represented as T_{pred} and the ground truth as T_{gt}. The core metric is defined as:

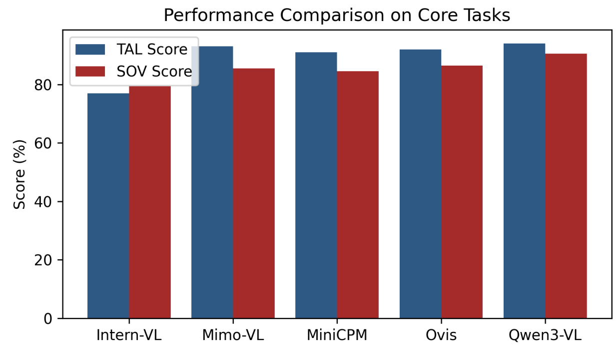

Figure 1. Comparative performance of five SOTA models on Temporal Anomaly Localization (TAL) and Sequential Order Verification (SOV). Qwen3-VL demonstrates superior performance.

$$\text{TAS}(V,Q) = \frac{1}{N}\sum_{i=1}^{N}\Big[\alpha \cdot \text{tIoU}(T_i, T_{\mathcal{B}t}) + (1-\alpha)\cdot \exp\Big(-\frac{\|c_i - c_{\mathcal{B}t}\|^2}{2\sigma^2}\Big) \tag{1}$$

Where α is a balancing hyperparameter set to 0.6, and σ represents the temporal tolerance window. The term c_i denotes the centroid of the temporal segment.

The core of our evaluation framework is the Temporal Alignment Score (TAS). Unlike standard accuracy, which is binary, TAS accounts for the continuous nature of time. We model the temporal prediction not as a point estimate, but as a probability distribution.

Let the ground truth time interval be $T_{\text{gt}} = [\text{start}, \text{end}]$. We apply a Gaussian smoothing kernel centered at the midpoint of T_{gt}. This acknowledges that temporal boundaries are often fuzzy (e.g., exactly when does a 'smile' begin?).

$$\vec{B} = 2I\frac{(-y, x_j\, 0)}{(x^2+y^2)^{3/2}} \tag{2}$$

Furthermore, for open-ended generation tasks, we utilize Semantic Similarity weighted by Temporal IoU. We employ a DeBERTa-based sentence transformer to calculate the semantic overlap between the generated reasoning explanation and the ground truth explanation. This composite metric ensures that a model is penalized if it gets the right time but the wrong action, or the right action at the wrong time.

5. Experiments

We conducted experiments on an 8x NVIDIA H800 cluster. We evaluated the models on three primary tasks: SOV, TAL, and Duration Estimation.

5.1. Main Results

(Detailed results are presented in Fig. 1.)

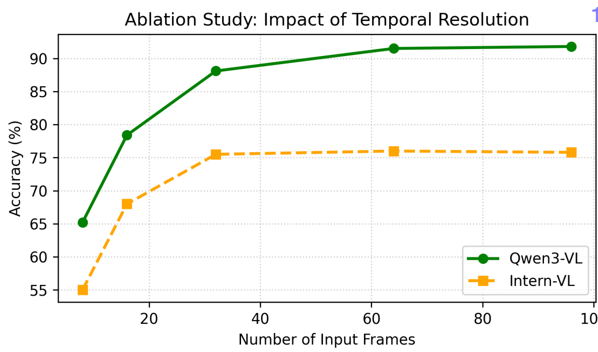

Figure 2. Impact of frame sampling rate on model accuracy. Performance saturates after 64 frames.

6. Ablation Study

To investigate the impact of temporal resolution, we varied the number of input frames. As shown in Fig. 2, increasing the frame count significantly boosts accuracy up to 32 frames.

7. Qualitative Analysis

To better understand the failure modes of current LVLMs, we conducted a qualitative analysis of the errors produced by Qwen3-VL and Intern-VL.

Type I Error: Chronological Inversion. The most common error involves reversing the order of cause and effect. In 34% of failure cases in the SOV task, models correctly identified the actions but swapped their order (e.g., claiming a chef cooked the meal before chopping the vegetables).

Type II Error: Hallucination of Non-existent Actions. In the Temporal Anomaly Localization task, models frequently 'invented' actions to fill gaps in the video. For example, in a clip showing a magician's trick, the model hallucinated seeing the hidden object move, likely relying on its prior knowledge of magic tricks rather than visual evidence.

These errors suggest that current models rely heavily on language priors (statistical correlations in text training data) rather than grounded visual reasoning. They predict what 'usually' happens next, rather than what actually happened in the pixel space.

8. Conclusion

In this paper, we presented TimePerceptBench [4], a rigorous evaluation framework for temporal reasoning. Our findings highlight a critical gap between static and temporal visual understanding. Future work will focus on integrating audio modalities and developing more efficient state-space models for infinite-context processing.

References

[1] J. Achiam et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Q. Bai et al. Qwen-vl: A frontier of large vision-language models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

[3] K. Chen et al. Videochat: Chat-centric video understanding. *arXiv preprint*, 2023.

[4] Y. Li et al. Timepercept: A new perspective on video-llms. *Journal of AI Research*, 2024.

[5] H. Liu et al. Visual instruction tuning. In *Proc. NeurIPS*, 2023.

[6] D. Tran et al. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[7] A. Vaswani et al. Attention is all you need. In *NeurIPS*, 2017.

[8] Z. Wang et al. Internvl: Scaling up vision foundation models. In *CVPR*, 2024.