To Think or Not to Think: Task-Dependent Reasoning for Event Prediction

Michael Chen
Computing + Mathematical Sciences
California Institute of Technology
Pasadena, CA 91126
mhchen@caltech.edu

Gavin Yuliu Hua
Computing + Mathematical Sciences
California Institute of Technology
Pasadena, CA 91126
ghua@caltech.edu

Edward Ju
Computing + Mathematical Sciences
California Institute of Technology
Pasadena, CA 91126
eju@caltech.edu

Taekyung Lee
Department of Electrical Engineering
California Institute of Technology
Pasadena, CA 91126
tlee2@caltech.edu

Abstract

Anticipating future events from dynamic visual scenes, such as predicting potential traffic accidents or driver braking behavior from dashcam footage, is a critical capability for intelligent systems but poses significant challenges in temporal reasoning and uncertainty handling. Multimodal Large Language Models (MLLMs) offer a promising avenue by integrating rich visual perception with sophisticated language-based reasoning. In this work, we systematically investigate MLLM performance for event prediction tasks, focusing on the relationship between prediction accuracy, time-to-event horizon, and model reasoning capacity (thinking budget). We use two complementary dashcam datasets: 1) the Berkeley DeepDrive Attention (BDD-A) dataset for brake timing prediction, and 2) the Car Crash Dataset (CCD) for binary crash prediction. We evaluate Google DeepMind's Gemini 2.5 Flash across different temporal segments and reasoning budgets. Contrary to expectations, with no reasoning, MLLMs achieved optimal performance for timing prediction, with larger reasoning budgets degrading accuracy. However, longer reasoning improved crash predictions. Furthermore, we introduce iterative feedback refinement, demonstrating that providing prior reasoning traces as context for the upcoming scene significantly improves performance, even without including the previous video segment. We challenge conventional scaling assumptions and provide empirical insights into optimally allocating reasoning resources and trends in reasoning efficiency for MLLM-based temporal prediction systems. Finally, to encourage reproducibility and transparency, we release all code and dataset segmentation publicly: github.com/math-ysics/EventPrediction

1 Introduction

The ability to anticipate future events from dynamic visual scenes is crucial for intelligent systems operating in the real world, particularly in safety-critical domains like autonomous driving, robotics, and surveillance [6–8]. Predicting events such as potential traffic accidents seconds before they occur allows for proactive intervention, enhancing both safety and efficiency. Multimodal Large Language Models (MLLMs) have emerged as powerful tools capable of integrating rich perceptual information (vision, audio, etc.) with the sophisticated reasoning and world knowledge inherent in LLMs [9–11].

Preprint. Under review.

This fusion promises systems that not only perceive the present but also reason about and predict the immediate future, potentially providing explanatory justifications for their predictions [12].

The emergence of MLLMs represents a paradigm shift from traditional computer vision approaches that typically rely on specialized architectures trained on domain-specific datasets [13, 14]. Unlike conventional event prediction systems that operate as black boxes, MLLMs offer the potential for interpretable predictions through natural language reasoning, making them particularly valuable for safety-critical applications where understanding the rationale behind predictions is essential [15]. Recent advances in vision-language integration have demonstrated promising capabilities in static scene understanding and simple temporal reasoning tasks, yet their effectiveness for complex temporal prediction scenarios, particularly those involving multiple interacting agents and long-range dependencies, remains largely unexplored [16].

The temporal nature of event prediction introduces unique challenges that distinguish it from static visual understanding tasks. Traditional approaches often struggle with the inherent ambiguity in temporal sequences, where similar visual patterns can lead to vastly different outcomes depending on subtle contextual factors [17]. Furthermore, the evaluation of temporal prediction systems requires sophisticated metrics that can account for both accuracy and temporal precision, as predictions that are correct in nature but incorrect in timing can have catastrophic consequences in real-world deployment [18].

Despite their potential, effectively leveraging MLLMs for future event prediction presents significant challenges. Temporal dynamics pose a fundamental difficulty, requiring the capture and reasoning about complex, long-range temporal dependencies, motion patterns, and the evolution of interactions between multiple agents, which is particularly challenging within the context-length limitations of current LLMs [19]. Reasoning under uncertainty represents another key challenge since the future is inherently uncertain, requiring models to predict plausible outcomes based on incomplete and often ambiguous information while ideally quantifying this uncertainty [20]. Visual grounding is critical as predictions and reasoning must be firmly grounded in the visual evidence, yet MLLMs can sometimes hallucinate or rely too heavily on priors from their pre-training, leading to predictions disconnected from actual scene dynamics [21]. The distinction between causality and correlation is essential, as models must distinguish true causal precursors of events (e.g., reckless driving causing a crash) from statistical correlations observed in training data for generalizable prediction [22]. Finally, evaluation presents non-trivial challenges, as evaluating the quality of predictions, particularly when they involve explanatory reasoning, requires metrics beyond simple accuracy.

We aim to investigate the capabilities and limitations of MLLMs for event prediction with reasoning. Our contributions are:

(i) Empirical characterization of the effects of reasoning length for MLLM event prediction across different temporal horizons.

(ii) Evidence for the effectiveness of iterative temporal reasoning approaches in improving prediction accuracy and consistency.

(iii) Insights into optimal reasoning budget allocation and temporal context integration strategies for practical MLLM deployment in anticipation systems.

2 Related Work

2.1 Transformer-Based Visual Anticipation

Early approaches often treated event prediction as a sequence modeling problem purely on visual data. Vision Transformers (ViTs) and autoregressive models were applied directly to sequences of video frames or sensor inputs (e.g., dashcam footage, driver gaze) to predict future frames, actions, or event labels like accidents. For example, DrivingGPT [13] demonstrates strong performance for short-horizon accident prediction but lacks explicit reasoning explanations and struggles with longer-term predictions.

2.2 Video-Language Models for Understanding and Reasoning

More recent work integrates visual perception with the reasoning capabilities of LLMs. Architectures like BLIP-2 [10] and Flamingo [9] are adapted for video via adapters or specialized modules (e.g., AccidentBlip2 [1]) to feed sequences of frame features into an LLM. These models can generate textual descriptions, answer questions about video content, and increasingly, predict future events with justifications (e.g., SeeUnsafe [15]). A key challenge is effectively representing temporal information for the LLM, leading to techniques such as time tokens (Vid2Seq [19]) or hierarchical processing.

2.3 Iterative Refinement and Feedback

Techniques involving iterative processing, self-reflection, and feedback are being explored to enhance MLLM outputs. This includes preference optimization with retrospection for better visual grounding (Iterative Self-Retrospective DPO [2]), iterative prompt/output refinement (PhyT2V [3]), and agent-based frameworks with search (CoMCTS [12]). These methods leverage prior processing steps or evaluations to guide subsequent ones, aiming for improved accuracy, consistency, and robustness (e.g., TEMPLE [16]).

3 Datasets

We utilize two complementary dashcam video datasets that provide rich temporal sequences with precise event annotations, enabling systematic analysis of prediction accuracy across different time-to-event horizons.

3.1 Car Crash Dataset (CCD)

The Car Crash Dataset [4] is a specialized collection focusing on traffic accident prediction from dashcam footage. This dataset is particularly valuable for our research as it provides precise temporal annotations of accident occurrences, enabling systematic evaluation of prediction accuracy at various time offsets before critical events.

Dataset Composition. The CCD contains 4,500 dashcam video clips in total, with 1,500 videos containing actual crash events and 3,000 normal driving scenarios serving as negative examples. Each crash video is temporally annotated with the exact moment of collision, providing ground truth for our time-to-event analysis. The dataset spans diverse driving conditions including urban intersections, highway scenarios, and various weather conditions. Videos in the CCD are 50-frame clips of 5 seconds duration at 10 FPS, with crash events occurring at varying temporal positions within each clip. This temporal structure enables systematic extraction of video segments at consistent time offsets before the actual collision across different scenarios. Each crash event is manually annotated with precise frame-level timestamps, enabling accurate extraction of pre-event segments. Additionally, the dataset includes contextual metadata such as day/night labels and primary collision factors.

3.2 Berkeley DeepDrive Attention Dataset (BDD-A)

The BDD-A dataset [5] extends the large-scale Berkeley DeepDrive dataset with attention annotations, providing a complementary perspective on critical driving event anticipation through driver attention modeling. Notably, BDD-A focuses on braking events and critical driving situations rather than direct collision events. We leverage this to ask the LLM to predict when the driver will brake, testing its inference and observational capabilities.

Dataset Composition. BDD-A comprises 1,232 videos (approximately 3.5 hours total) collected following procedures that specifically selected video clips including braking events in busy areas. Each video includes 6.5 seconds prior to and 3.5 seconds after each braking event, at 30 FPS, yielding clips of about 10 seconds in length. The dataset's diversity across geographic regions, driving environments, and participant demographics supports generalization analysis.

4 Methodology

For each event, we extract video segments at specific temporal intervals. We extract six consecutive 1-second segments for brake prediction (BDD-A) and two critical pre-crash windows for accident prediction (CCD). Our experimental design systematically varies model reasoning capacity and temporal context to characterize scaling behaviors.

4.1 Experimental Design

Datasets and Tasks. Our experiments utilize two datasets: the Berkeley DeepDrive Attention (BDD-A) dataset where the task is to predict when a driver will apply brakes with an output of `prediction_seconds` and a `confidence_level` (0-10). The Car Crash Dataset (CCD) involves binary crash prediction using 250 crash and 250 normal videos, with the task being to predict whether a crash will occur within five seconds, outputting `will_crash_in_five_seconds` and a `confidence_level` (0-10).

Video Segmentation. To enable systematic time-to-event analysis, video segments are extracted with specific temporal structures. For BDD-A, this involves six consecutive 1-second segments relative to brake onset, specifically (0,1), (1,2), (2,3), (3,4), (4,5), and (5,6) seconds. For CCD, we focus on two critical pre-crash windows: early segments (T-2.5s to T-1.5s) and late segments (T-1.5s to T-0.5s) before a crash. We supplement this dataset with random 1-second clips from normal driving videos as a control.

Model and Reasoning Variations. We choose `gemini-2.5-flash-preview-05-20` as the MLLM. Model reasoning capacity is systematically varied using the Google Gemini API. For both tasks, we explore different thinking budgets (0, 2000, 4000, 8000, 16000 tokens). All predictions use structured function calling for consistent output formats.

Temporal Context Manipulation. Two distinct prediction approaches are implemented to evaluate iterative reasoning. The baseline approach involves independent analysis of individual video segments, without incorporating prior temporal context. The iterative approach conducts sequential analysis, where later segments receive reasoning outputs from all preceding temporal segments as context, thereby enabling cumulative temporal understanding that builds comprehensive behavioral patterns leading to target events. However, previous video is not supplemented to the model input; it can only access previous reasoning.

4.2 Implementation Pipeline

We developed a parallel processing framework using Google's Gemini API to systematically evaluate MLLM performance across the experimental conditions. The implementation leverages `ProcessPoolExecutor` for concurrent API calls with resume functionality to handle interruptions and avoid redundant processing.

API Integration and Architecture. Our pipeline processes video uploads via Gemini's File API with structured function calling ensuring consistent output parsing across all experimental conditions. The system generates CSV output files for all predictions.

Iterative Context Management. For iterative predictions, we implemented a system that maintains reasoning chains across temporal segments. The iterative approach builds cumulative context by providing previous segment analyses to subsequent predictions, enabling temporal understanding that spans multiple time intervals leading to target events. For BDD-A, each segment receives context from all previous segments in the sequence. For CCD, early segments provide context to late segments in crash videos.

5 Results

Our experimental evaluation systematically investigates the scaling behaviors of MLLMs for temporal event prediction. We analyze the model's performance across different time-to-event horizons and

thinking budgets, comparing a baseline approach against an iterative refinement method. The results are summarized in Figure 1 and Figure 9 below, addressing our core research questions.

5.1 Prediction Accuracy

Effects of Reasoning Capacity. Contrary to the common assumption that more computation yields better results, we observed a non-monotonic and often inverse relationship between the allocated thinking budget and prediction accuracy. As illustrated in the heatmap in Figure 1 and plots in the appendix, the model achieved the lowest prediction error with a thinking budget of 0 across nearly all temporal horizons. Increasing the budget significantly increases prediction error. This suggests that for this task, additional reasoning may introduce noise or over-analysis, leading to less accurate predictions.

Error Distribution. The distribution of relative errors, shown in Figure 1, is heavily right-skewed. The majority of predictions have low relative error, indicating that the model is often reasonably accurate. However, the long tail signifies the presence of significant outliers where the model makes large, catastrophic errors, a challenge for reliability in safety-critical applications.

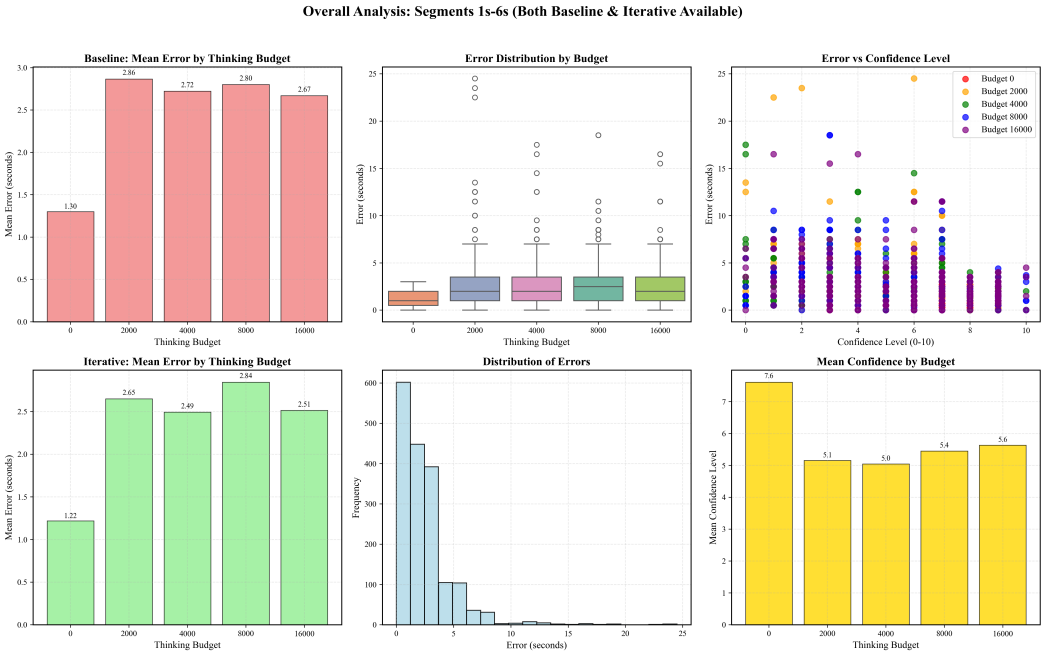

Figure 1: Overall analysis of MLLM prediction performance across 1s–6s pre-braking segments for both baseline and iterative reasoning approaches. Top-left and bottom-left: Mean absolute error (s) by thinking budget for baseline and iterative methods, showing lowest error at budget 0 and degradation with increased budget. Top-center: Box plot showing error distribution widening with larger budgets. Top-right: Scatter plot of prediction error vs. self-reported confidence, revealing weak calibration. Bottom-center: Overall error distribution, highlighting a strong right skew with most errors under 5 s. Bottom-right: Mean confidence by budget, peaking at budget 0 and dropping with higher budgets, consistent with error trends.

Crash Clip Accuracy. As shown in Figure 9 (top-left), the model's ability to detect an upcoming crash improves from the early window to the late window across all thinking budgets. The modest but consistent gains with larger budgets suggest that additional reasoning helps the model identify imminent collision cues, although absolute detection rates remain low.

Normal Clip Accuracy. Figure 9 (top-right) reports near-perfect accuracy on normal (non-crash) clips for all budgets. This indicates the model reliably avoids false alarms, a critical requirement for safety-critical deployment.

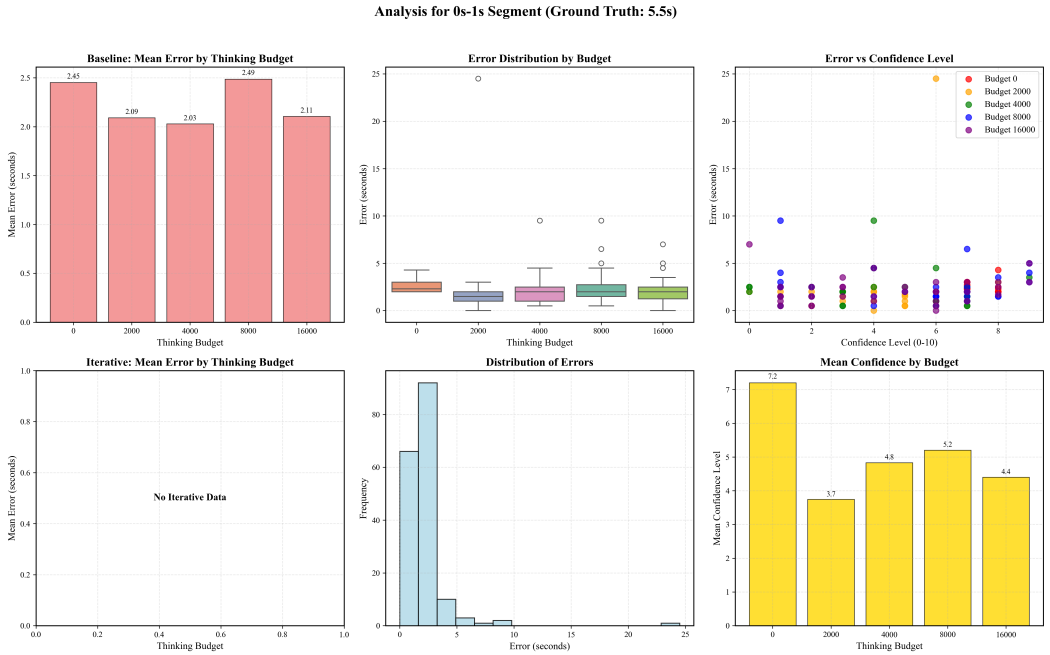

Figure 2: Overall analysis of MLLM prediction performance across 0s-1s pre-braking segments for both baseline and iterative reasoning approaches.

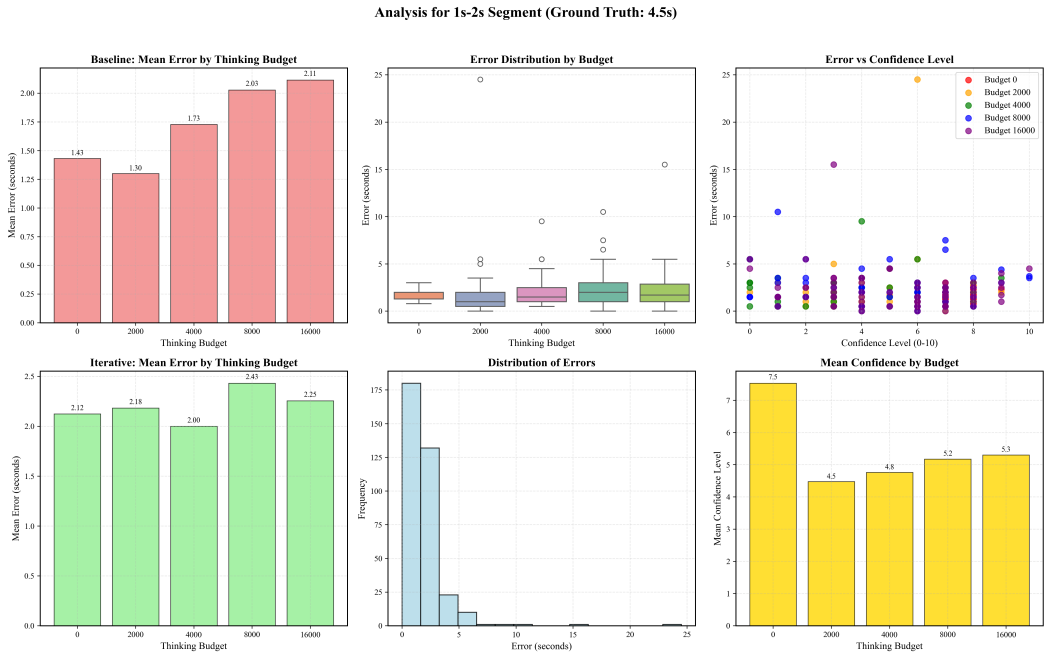

Figure 3: Overall analysis of MLLM prediction performance across 1s-2s pre-braking segments for both baseline and iterative reasoning approaches.

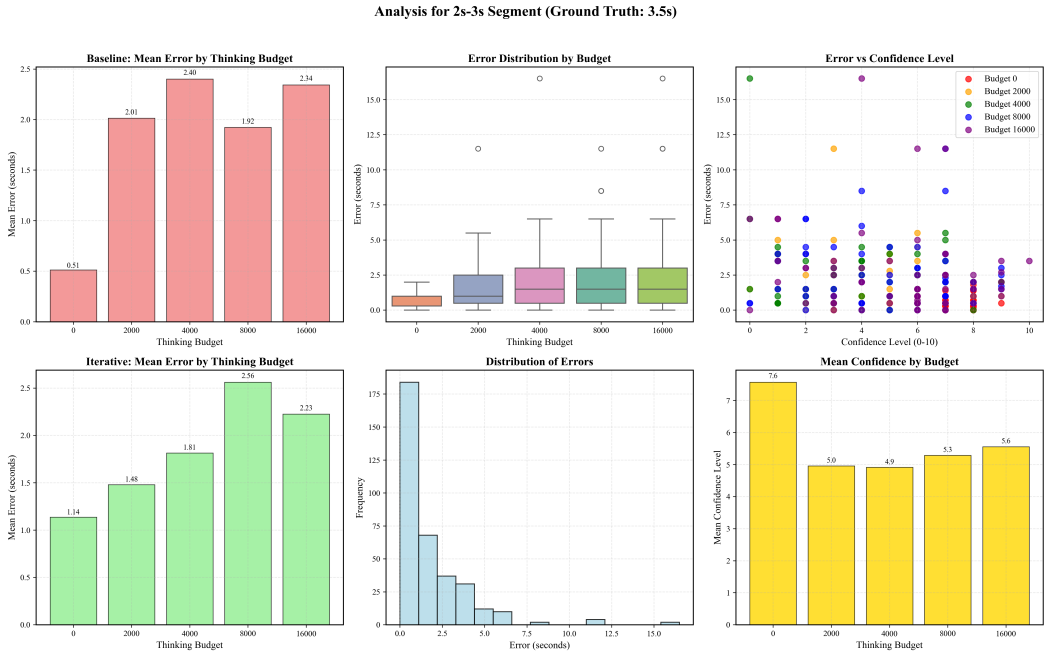

Figure 4: Overall analysis of MLLM prediction performance across 2s-3s pre-braking segments for both baseline and iterative reasoning approaches.

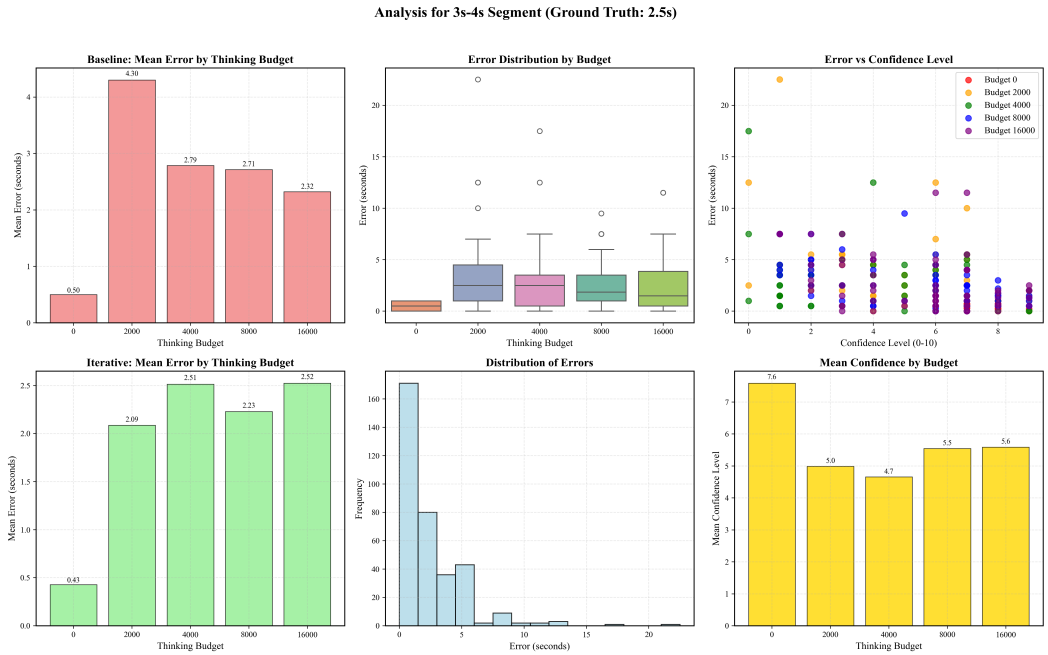

Figure 5: Overall analysis of MLLM prediction performance across 3s-4s pre-braking segments for both baseline and iterative reasoning approaches.

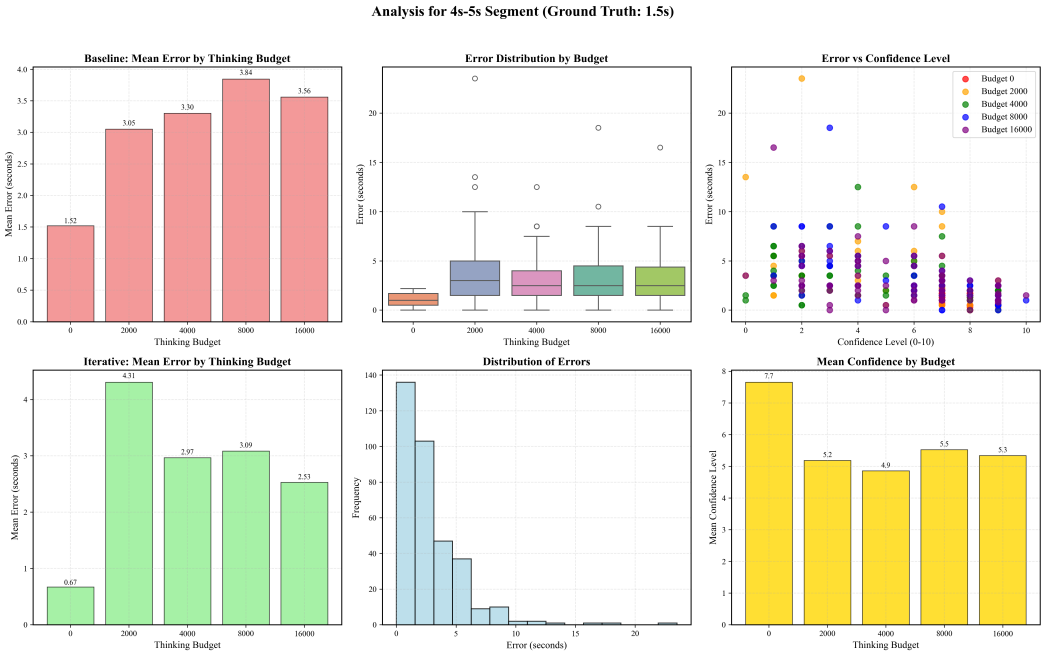

Figure 6: Overall analysis of MLLM prediction performance across 4s-5s pre-braking segments for both baseline and iterative reasoning approaches.

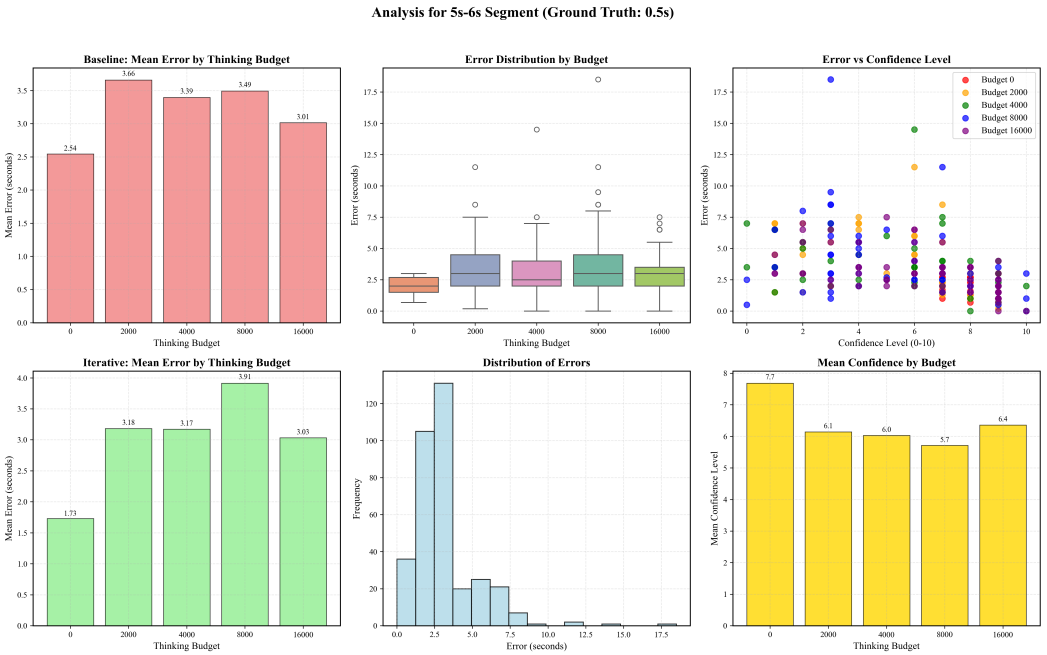

Figure 7: Overall analysis of MLLM prediction performance across 5s-6s pre-braking segments for both baseline and iterative reasoning approaches.

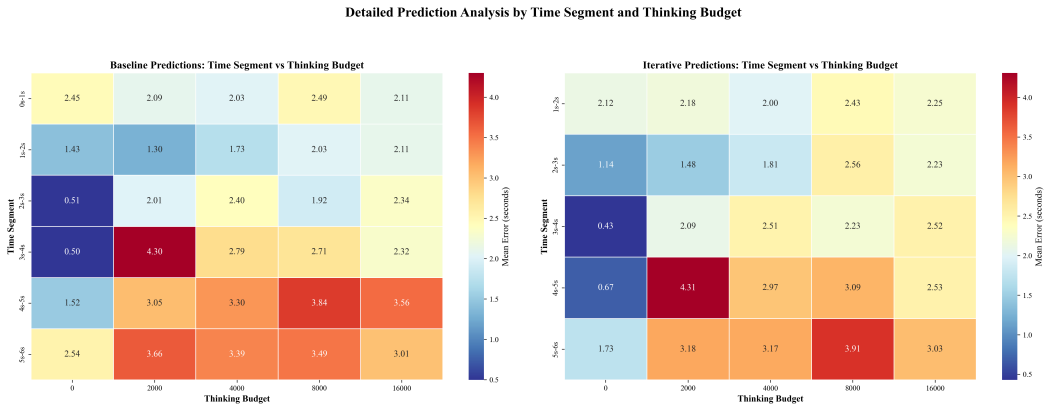

Figure 8: Heatmaps of mean absolute prediction error (in seconds) for each 1-second time segment before the braking event (rows) across different thinking budgets (columns). The left panel shows baseline predictions made independently on each segment, while the right panel shows iterative predictions that incorporate reasoning outputs from all prior segments. Cooler (blue) cells indicate lower error, and warmer (red) cells indicate higher error. Notably, minimal reasoning (budget 0) often attains the lowest error, and the iterative approach yields modest error reductions (especially in the later pre-event intervals) compared to the baseline.

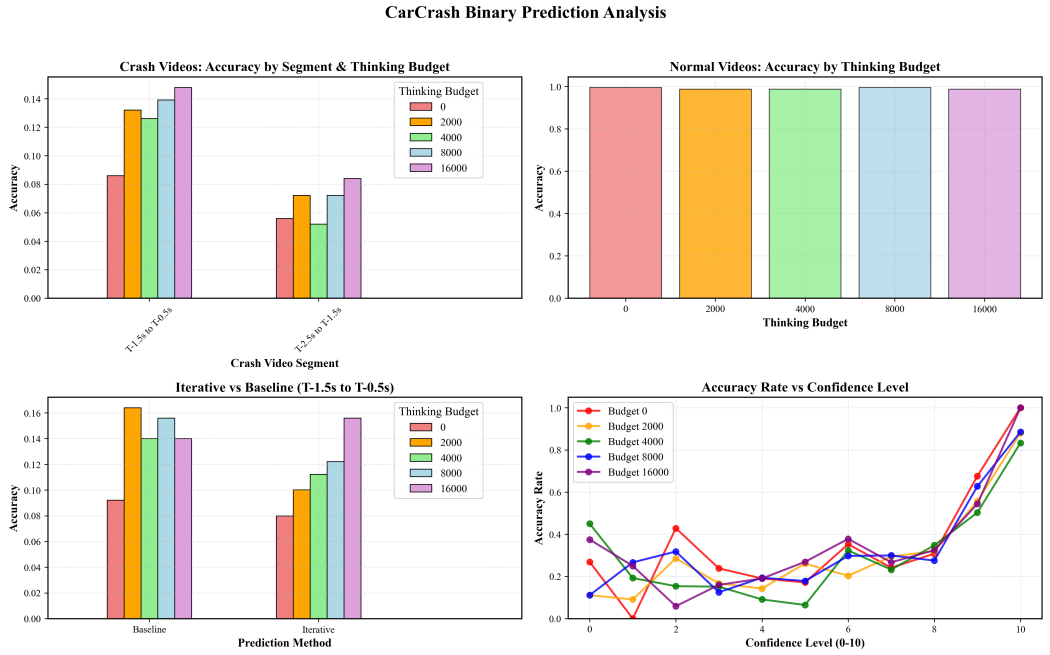

Figure 9: Binary-crash prediction performance on the CarCrash dataset. Top-left: Crash-clip accuracy in early and late windows across thinking budgets. Top-right: Normal-clip accuracy remains $> 95\%$ for all budgets. Bottom-left: Increased thinking budget with iterative reasoning yields a consistent accuracy boost in the late window. Bottom-right: Accuracy vs. self-reported confidence, showing low reliability at confidence < 4 but improved accuracy curves for confidence ≥ 8 independent of budget.

Iterative Refinement Benefits. Providing prior segment reasoning via the iterative method boosts accuracy over the baseline (Figure 9, bottom-left). This mirrors the improvements seen on braking prediction and validates the general benefit of temporal context in binary crash detection. Furthermore, it demonstrates strong scaling behavior with thinking budget.

Confidence Calibration. The calibration curve (Figure 9, bottom-right) shows that low-confidence predictions (< 4) are unreliable, whereas high-confidence predictions (≥ 8) improve on reliability independent of budgets. Thus, self-reported confidence is a potential indicator of prediction reliability in the crash-detection task, unlike for the braking task.

6 Discussion

6.1 Implications for MLLM Scaling

Contrary to expectations, increased thinking budget initially degraded prediction accuracy, with the shortest reasoning time (none) achieving optimal performance in the braking prediction task. This suggests that excessive computational resources may lead to overthinking, increased uncertainty, or reasoning artifacts that harm performance. The performance degradation with longer thinking suggests that brief, intuitive responses may be superior to extensive deliberation for certain temporal prediction scenarios. We recognize that performing reinforcement learning with verifiable rewards (RLVR) on LLMs typically focuses on mathematics and computer science tasks, so scaling computational budget for reasoning may not transfer very efficiently to out-of-distribution tasks like vehicle braking prediction.

6.2 Temporal Reasoning Capabilities

The temporal horizon effects confirm that MLLMs can effectively leverage visual cues that evolve over time leading up to critical events. The model's ability to achieve higher accuracy for segments closer to crashes demonstrates sophisticated understanding of temporal dynamics in driving scenarios.

6.3 Iterative Refinement Benefits

The substantial scaling from iterative refinement validates the hypothesis that reasoning chains can significantly enhance temporal prediction accuracy. This improvement demonstrates that prior context helps the model build more coherent temporal understanding across sequential segments, suggesting that temporal reasoning benefits from explicit memory and context propagation.

The magnitude of improvement indicates that iterative reasoning is a promising direction for enhancing MLLM temporal capabilities. The successful maintenance of reasoning chains across all video segments shows the robustness of this approach, contradicting concerns about context brittleness in sequential prediction tasks.

6.4 Limitations and Failure Modes

The high variance in results indicates inconsistent performance across different driving scenarios. We were not able to provide error bars due to the increased costs. The model exhibited sensitivity to scale, particularly struggling with segments very close to events, which highlights challenges in ultra-short temporal predictions. Our analysis was limited to a single MLLM architecture, restricting the generalizability of findings across other model families.

6.5 Broader Implications

These findings have significant implications for deploying MLLMs in autonomous systems and in safety-critical domains. The scaling relationships between computational resources and accuracy provides a framework for balancing performance requirements against computational constraints in real-time applications. The temporal scaling effects suggest that MLLMs could be useful for intermediate-term prediction tasks where visual cues are informative.

The implications extend beyond autonomous driving to other domains requiring temporal anticipation, including robotics, surveillance, and human-computer interaction. As MLLMs continue to improve,

understanding their scaling behaviors and limitations becomes increasingly critical for responsible deployment in applications where prediction accuracy directly impacts safety and reliability.

7 Conclusion

We presented a systematic investigation of event prediction in multimodal large language models, focusing on temporal reasoning capabilities for critical driving events. Our key contributions include the discovery of non-monotonic performance-to-reasoning returns; a temporal horizon analysis revealing strong scaling, with accuracy significantly improving for segments closer to critical events; an evaluation of iterative reasoning showing that incorporating prior context led to substantial accuracy gains, supporting temporal memory approaches; and the development of a robust experimental framework for assessing MLLM temporal reasoning using structured function calling and relative error metrics.

7.1 Future Directions

Several promising directions emerge from this work. Extending the investigation to a broader range of MLLM architectures and sizes could help establish more general and robust empirical patterns. There is an opportunity to develop more sophisticated techniques for temporal context integration that go beyond simple reasoning chain propagation. Analyzing how MLLMs distinguish causal relationships from mere correlations in temporal prediction tasks could deepen our understanding of their reasoning capabilities. Exploring the trade-offs between accuracy and latency is essential for enabling real-time deployment in autonomous systems. Advancing methods for uncertainty quantification and calibration would improve the reliability of MLLM predictions in temporal reasoning contexts.

References

[1] Shao, Yihua et al. (2024) AccidentBlip: Agent of Accident Warning based on MA-former. *arXiv preprint arXiv:2404.12149*.

[2] Ahn, Daechul et al. (2024) ISR-DPO: Aligning Large Multimodal Models for Videos by Iterative Self-Retrospective DPO. *arXiv preprint arXiv:2406.11280*.

[3] Xue, Qiyao, Yin, Xiangyu, Yang, Boyuan & Gao, Wei (2024) PhyT2V: LLM-Guided Iterative Self-Refinement for Physics-Grounded Text-to-Video Generation. *arXiv preprint arXiv:2412.00596*.

[4] Shah, Ankit, Davuluri, Srikar, Horowitz, Roberto, Bajcsy, Ruzena (2018) CrashNet: Collision Detection for Autonomous Vehicles. *arXiv preprint arXiv:1801.04090*.

[5] Xia, Yao, Zhang, Danqing, Kim, Jinkyu, Nakayama, Ken, Zipser, Karl, Whitney, David (2018) Predicting Driver Attention in Critical Situations. *Asian Conference on Computer Vision (ACCV)*.

[6] Dosovitskiy, Alexey, et al. (2017). CARLA: An Open Urban Driving Simulator. *Conference on Robot Learning (CoRL)*.

[7] Sadat, Arsalan, et al. (2020). Perceive, Predict, and Plan: Safe Motion Planning through Interpretable Semantic Representations. *ECCV*.

[8] Singh, Amanpreet, et al. (2023). Remember What You See: Language-guided Temporal Localization in Videos. *CVPR*.

[9] Alayrac, Jean-Baptiste, et al. (2022). Flamingo: A Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198*.

[10] Li, Junnan, et al. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*.

[11] Zhu, Deyao, et al. (2023). MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*.

[12] Yao, Huanjin, et al. (2024). Mulberry: Empowering MLLM with o1-like Reasoning and Reflection via Collective Monte Carlo Tree Search. *arXiv preprint arXiv:2412.18319*.

[13] Chen, Yuntao, et al. (2024). DrivingGPT: Unifying Driving World Modeling and Planning with Multi-modal Autoregressive Transformers. *arXiv preprint arXiv:2412.18607*.

[14] Ji, Xiangyang, et al. (2022). A Survey on 3D Scene Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[15] Zhang, Ruixuan, et al. (2025). When Language and Vision Meet Road Safety: Leveraging Multimodal Large Language Models for Video-Based Traffic Accident Analysis. *arXiv preprint arXiv:2501.10604*.

[16] Li, Shicheng, et al. (2025). TEMPLE: Temporal Preference Learning of Video LLMs via Difficulty Scheduling and Pre-SFT Alignment. *arXiv preprint arXiv:2503.16929*.

[17] Yan, Xinchen, et al. (2021). Learning Predictive Models to Simulate Human Driving Behavior. *CVPR*.

[18] Sun, Chen, et al. (2020). On the Scalability of Video-based Action Recognition with a Large Number of Classes. *ECCV*.

[19] Yang, Antoine, et al. (2023). Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning. *arXiv preprint arXiv:2302.14115*.

[20] Amini, Alexander, et al. (2020). Deep Evidential Regression. *NeurIPS*.

[21] Zellers, Rowan, et al. (2021). MERLOT: Multimodal Neural Script Knowledge Models. *NeurIPS*.

[22] Pearl, Judea. (2009). Causality: Models, Reasoning and Inference. *Cambridge University Press*.