Optimizing Multimodal LLMs for Egocentric Video Understanding: A Solution for the HD-EPIC VQA Challenge

Sicheng Yang[1,⋆], Yukai Huang[1,⋆], Shitong Sun[1], Weitong Cai[2],
Songcen Xu[1], Jiankang Deng[3], Jifei Song[1], Zhensong Zhang[1,✉]
[1]Huawei Noah's Ark Lab [2]Queen Mary University of London [3]Imperial College London
{sicheng.yang, yukai.huang}@h-partners.com, weitong.cai@qmul.ac.uk
{shitong.sun, xusongcen, jifeisong, zhangzhensong}@huawei.com, j.deng16@imperial.ac.uk

Abstract

Multimodal Large Language Models (MLLMs) struggle with complex video QA benchmarks like HD-EPIC VQA due to ambiguous queries/options, poor long-range temporal reasoning, and non-standardized outputs. We propose a framework integrating query/choice pre-processing, domain-specific Qwen2.5-VL fine-tuning, a novel Temporal Chain-of-Thought (T-CoT) prompting for multi-step reasoning, and robust post-processing. This system achieves 41.6% accuracy on HD-EPIC VQA, highlighting the need for holistic pipeline optimization in demanding video understanding. We will make the code, and fine-tuned models available to the public in the future.

1. Introduction

Visual question answering (VQA) is a video understanding task that studies how to answer questions about video content based on its temporal visual information [10, 16]. Compared to VQA for short, static or exocentric videos, implementing egocentric VQA for long videos [12, 15–17] is significantly more challenging, as it requires robust temporal reasoning over long durations, inferring human intent, taking into account temporal information, history memory, and complex multistep reasoning for nuanced queries.

To rigorously evaluate model performance on these specific, challenging egocentric VQA tasks, we focus on the VQA benchmark from the HD-EPIC dataset [16]. HD-EPIC provides highly detailed long egocentric videos of complex kitchen activities, making its VQA benchmark a vital testbed for the aforementioned challenges. This benchmark is notable for its comprehensive coverage, featuring 30 distinct question prototypes that generate 26K questions specifically designed to test intricate temporal reasoning, object interaction, and fine-grained action understanding, which are crucial for evaluating models in this domain.

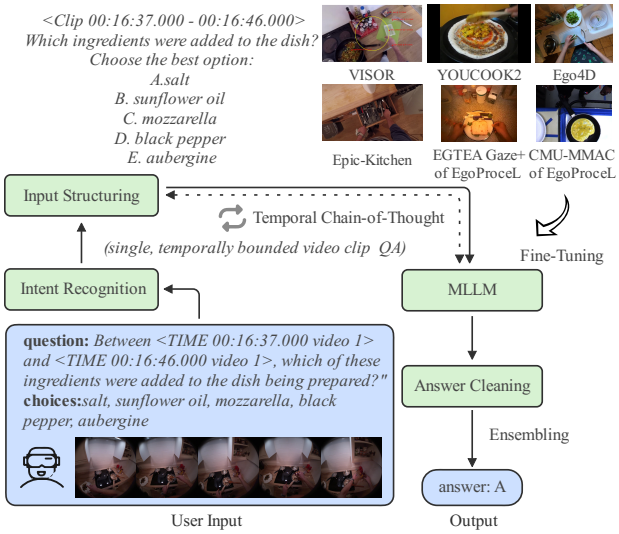

Figure 1. Overview of the proposed VQA system.

Despite the remarkable performance of many multimodal large language models (MLLMs) on general visual question answering benchmarks [3], including notable closed-source models like Gemini 2.5 Pro [11], GPT-4o [13], and Grok-3 [20], and prominent open-source counterparts such as DeepSeek-v3 [9], InternVL3 [22], and Qwen2.5-VL [1], they often exhibit limited performance on such specific, challenging tasks involving egocentric perspectives, complex kitchen scenarios, long temporal contexts, and intricate reasoning-based questions.

In this work, using Qwen-2.5 VL-7B [1] as our base model, we analyze several factors critical to improving performance on this task as shown in Figure 1: (1) Careful data pre-processing and optimization to better align user intent with MLLM comprehension. (2) Fine-tuning on extensive and densely annotated egocentric kitchen video data. (3) Implementing a two-stage reasoning process for temporal questions, specifically for long videos and keyframes,

enabling iterative processing and inference by the MLLM instead of a single-pass output. (4) Robust post-processing, including handling anomalous outputs, result cleaning, and integrating responses from multiple prompts, which collectively led to our model exceeding the performance of previous baseline approaches.

2. Method and Experiments

2.1. Question Intent Recognition and Clarification

To enhance MLLM performance on specialized benchmarks like HD-EPIC VQA, meticulous pre-processing of queries and choices is paramount. We address the inherent ambiguity and varied formatting of the 30-question prototypes through a multi-faceted strategy for intent recognition and clarification. First, we conduct a rigorous input modality analysis for each query, by classifying its visual context to the following 4 types: (1) A single static image (often a keyframe, e.g., for `3d perception fixture location`). (2) Multiple static images (e.g., for comparative tasks like `nutrition image nutrition estimation`). (3) A single, temporally bounded video clip (e.g., `fine grained action recognition`). (4) Multiple distinct video segments (e.g., `recipe prep localization`). This classification critically informs subsequent processing tailored to the MLLM.

Secondly, we implement task-specific prompt refinement. Original questions are strategically transformed into clearer, structured formats that are more conducive to MLLM comprehension. Using regular expression-based parsing, we extract core entities, temporal markers, and relational constraints. These are then re-synthesized into improved prompts, for instance, by rephrasing concise queries (e.g., "where is X?") into more explicit, viewpoint-grounded questions (e.g., "Based on this image of my current viewpoint, determine the direction of X.").

Thirdly, based on the observation that MLLM is highly sensitive to the choice presentation [16], we standardize and optimize the structure of multiple-choice options in the following ways. Converting numeric enumerators to alphabetic enumerators (e.g., A., B.) yielded an initial +1.6% accuracy gain. Further subtle formatting—inter-option spacing or semicolons—improved accuracy by +1.8% and +2.0% respectively. Critically, clear newline delineation (\n) for each option contributed a significant +2.4% uplift. Complex temporal options with multiple time segments are also reformatted for clarity (e.g., A. [V1] HH:MM:SS.sss - HH:MM:SS.sss). These refinements, with uniform delimiters, reduce parsing ambiguity.

These pre-processings enhance the MLLM's task comprehension by reducing cognitive load and aligning the input structure with the model's processing strengths. As shown in Table 1, these targeted refinements alone yield a discernible accuracy improvement (+3.5%), highlighting robust input engineering as a vital precursor to advanced model capabilities.

Model	Recipe	Ingredient	Nutrition	Action	3D	Motion	Gaze	Avg.
VideoLlama 2 [16]	30.8	25.7	32.7	27.2	25.7	28.5	21.2	27.4
LongVA [16]	29.6	30.8	33.7	30.7	32.9	22.7	24.5	29.3
LLaVA-Video [16]	36.3	33.5	**38.7**	**43.0**	27.3	18.9	29.3	32.4
Gemini Pro [16]	60.5	**46.2**	34.7	39.6	32.5	20.8	28.7	37.6
Qwen2.5 VL 7B In. [1]	40.6	35.8	32.0	37.3	35.0	23.9	29.7	33.5
Qwen2.5 VL 32B In. [1]	59.0	37.0	33.0	40.3	35.6	19.8	**33.6**	36.9
Ours	**64.8**	43.3	37.0	42.0	**40.9**	**29.9**	33.0	**41.6**
w/o Pre-Processing	62.0	40.3	35.7	39.0	35.4	25.0	29.4	38.1
w/o Fine-tuning / T-CoT	62.6	41.5	32.0	38.3	35.2	23.5	29.1	37.5
w/o Post-Processing	65.0	43.0	35.0	40.3	38.4	26.5	31.6	40.0

Table 1. VQA Results per Category (% Acc.). 'w/o' is short for 'without' in ablation study.

2.2. Model Fine-Tuning and Temporal Domain Based Thinking

Given that HD-EPIC [16] is a first-person video dataset focused on kitchen scenarios, domain-specific fine-tuning is crucial for optimal performance [15]. We fine-tuned the Qwen2.5-VL-7B-Instruct [1] model on a diverse collection of egocentric kitchen video datasets, including EPIC-KITCHENS [4–6], CMU-MMAC [8] and EGTEA Gaze+ [14] of EgoProceL [2] subsets, YOUCOOK2 [21], VISOR [7], and selected portions of Ego4D [12] relevant to object interaction and procedural understanding. Fine-tuning was configured with a learning rate of 2×10^{-7}, batch size 2, 1 gradient accumulation step, and 1 epoch. Only LLM components were tuned, freezing the vision tower and MLP projector. We utilized bfloat16 precision, the AdamW optimizer, and a maximum sequence length of 131072 tokens. Video processing involved a maximum of 768 frames/sample (minimum 4), with dynamic total pixel adjustment per video (3136 to 846720).

Despite fine-tuning, Qwen2.5-VL, like many MLLMs, exhibited challenges with long-term temporal relationship understanding. For instance, on `Multi-Step Localization`, our fine-tuned model achieved 26% accuracy (vs. 22% pre-tuning), substantially below Gemini Pro's 88%. Similar disparities occurred in `Step Localization` (25% vs. 70%) and `Rough Step Localization` (28% vs. 74%), and tasks like `Ingredients Order`. We hypothesize this stems partly from the model's training sequence lengths (8192/32768) and its video processing strategy, capping analyzed frames at 768 (total video tokens $\leq$ 24576). For videos >12 minutes (at 1 FPS, >720 frames), dynamic resolution scaling or token limits may yield insufficient effective input frames for fine-grained temporal analysis.

To address these temporal limitations, we developed a Temporal Chain-of-Thought (T-CoT) prompting strategy [18, 19], guiding the MLLM through intermediate reasoning steps to isolate and comprehend relevant temporal context, rather than directly posing complex temporal

	Recipe Recognition	Multi-Recipe Recognition	Multi-Step Localization	Step Localization	Prep Localization	Rough Step Localization	Step Recognition	Following Activity Recognition	Ingredient Retrieval	Ingredient Weight	Ingredients Order	Ingredient Adding Localization	Exact Ingredient Recognition	Ingredient Recognition	Image Nutrition Estimation	Video Nutrition Estimation	Nutrition Change	Action Recognition	How Recognition	Why Recognition	Action Localization	Fixture Location	Object Location	Object Contents Retrieval	Fixture Interaction Counting	Object Movement Counting	Object Movement Itinerary	Stationary Object Localization	Gaze Estimation	Interaction Anticipation
VideoLlama 2 [16]	22.0	52.0	18.0	38.0	13.0	13.0	21.0	64.0	19.0	30.0	20.0	27.0	26.0	32.0	24.0	20.0	54.0	30.9	25.2	32.2	20.7	18.8	31.0	35.5	17.7	11.0	44.0	30.5	30.0	12.4
LongVA [16]	14.0	44.0	36.0	18.0	18.0	26.0	19.0	62.0	25.0	24.0	44.0	42.0	30.0	20.0	25.0	22.0	54.0	36.9	28.4	37.0	20.5	26.6	41.2	31.5	32.3	10.2	34.5	23.5	36.0	13.0
LLaVA-Video [16]	28.0	68.0	44.0	20.0	21.0	23.0	24.0	62.0	22.0	36.0	38.0	41.0	**36.0**	28.0	**28.0**	**26.0**	**62.0**	**58.6**	**41.4**	**51.2**	20.9	21.8	30.6	40.5	16.3	9.8	20.0	27.0	47.5	11.1
Gemini Pro [16]	**42.0**	**76.0**	**88.0**	70.0	35.0	45.0	**74.0**	54.0	49.0	**46.0**	**56.0**	**62.0**	**36.0**	28.0	26.0	16.0	**62.0**	49.3	35.6	43.2	**30.3**	20.8	32.4	41.5	**35.3**	**18.0**	13.0	**31.5**	36.5	**20.8**
Qwen2.5 VL 7B In. [1]	30.0	64.0	20.0	20.0	20.0	78.0	21.0	72.0	71.0	28.0	34.0	22.0	30.0	30.0	20.0	20.0	56.0	50.1	33.8	48.4	16.9	27.0	40.0	46.5	26.3	11.2	38.0	22.5	45.4	14.0
Qwen2.5 VL 32B In. [1]	28.0	58.0	62.0	**72.0**	35.0	72.0	65.0	**80.0**	66.0	26.0	30.0	38.0	24.0	**38.0**	23.0	16.0	60.0	51.3	37.8	48.8	23.1	26.2	46.3	45.2	24.7	7.8	22.0	29.5	49.8	17.4
Ours	34.0	72.0	70.0	68.0	**50.0**	**81.0**	73.0	70.0	**76.0**	32.0	36.0	48.0	30.0	**38.0**	25.0	**26.0**	60.0	55.6	36.6	51.0	24.9	**34.2**	**49.8**	**50.5**	29.0	14.2	**44.5**	31.0	**51.4**	14.5
w/o Pre-Processing	30.0	64.0	66.0	66.0	52.0	79.0	67.0	72.0	70.0	30.0	36.0	42.0	26.0	38.0	25.0	20.0	62.0	51.5	34.2	48.0	22.4	26.4	44.6	45.0	25.7	12.4	38.5	24.0	45.5	13.2
w/o Fine-tuning / T-CoT	26.0	66.0	66.0	70.0	53.0	79.0	69.0	72.0	70.0	36.0	40.0	45.0	22.0	36.0	22.0	18.0	56.0	49.8	33.6	47.6	22.2	27.2	43.8	45.0	24.7	12.4	36.5	21.5	44.5	13.7
w/o Post-Processing	40.0	68.0	66.0	70.0	52.0	81.0	69.0	74.0	74.0	30.0	40.0	44.0	30.0	40.0	27.0	22.0	56.0	52.2	35.4	48.6	24.8	30.0	47.6	46.5	29.3	14.0	40.0	25.5	47.4	15.8

Table 2. Model results per question prototype. 'w/o' is short for 'without' in ablation study.

queries. This strategy encompasses: (1) Explicit Temporal Cue Exploitation: For tasks with localized visual information or specified time points/segments (e.g., `3d perception` with bounding box (BBOX), `gaze` with time segments), we first process these cues. BBOX information is resolved by prompting the MLLM to generate a noun phrase for the object within the BBOX, which then replaces the BBOX placeholder in the question. For specified timestamps or segments, we extract the relevant clip and prompt the MLLM to analyze or narrate its content. (2) Focused Temporal Windowing: For questions implicitly tied to a narrow temporal window around a key event (e.g., `3d perception object location` implying "now"), we dynamically segment the video to a shorter duration (e.g., $\pm$ 10s around the relevant point), focusing MLLM attention and reducing irrelevant processing. (3) Multi-Video Synchronization: When questions or options involve multiple distinct video clips (e.g., `recipe prep localization`), these are programmatically concatenated. All timestamps in the question and options are then re-normalized relative to this new unified timeline, enabling the MLLM to process a single, coherent video stream. (4) Hierarchical Processing for Long Videos: For tasks requiring detailed understanding of extended videos exceeding the MLLM's single-pass capacity (e.g., `ingredient ingredients order`), we employ a chunking strategy. The video is divided into manageable, non-overlapping segments (e.g., 10-min, $<$ 768 frames). The MLLM generates a concise narration for each chunk. These temporally ordered narrations are then aggregated and prepended to the original question, providing rich, summarized contextual background for the final MLLM reasoning.

Our proposed two-stage T-CoT process—initially extracting, segmenting, or summarizing temporal/spatial context, followed by addressing the VQA query with this refined input—substantially reduces MLLM cognitive load and irrelevant information. Our T-CoT approach yielded a +3.0% overall accuracy improvement across all tasks compared to direct VQA with only initial pre-processing, demonstrating its efficacy in enhancing MLLM reasoning for complex temporal video understanding. We report results for the 7 categories HD-EPIC VQA scores (Table 1) and 30 task details (Table 2), and found that the fine-tuning and T-CoT strategies had the greatest impact on the results.

2.3. Answer Cleaning and Ensembling

MLLMs, despite explicit instructions for single-letter outputs (e.g., A-E), may generate verbose responses, hindering automated evaluation. We introduced a robust post-processing step via an answer cleaning module. This module employs regular expressions to parse raw MLLM textual outputs, extract the most probable single-letter choice, and convert it to a zero-based index for ground truth comparison. This cleaning procedure proved crucial, enabling automated scoring, mitigating misinterpretations of verbose outputs, and improving accuracy by +0.8% over evaluating raw outputs.

To enhance prediction robustness and accuracy for the multiple-choice HD-EPIC VQA benchmark, we implemented an ensembling strategy. This involved generating five distinct, semantically equivalent prompts per question by subtly varying the phrasing while preserving core semantic elements (entities, temporal information, relational constraints from Section 2.1). The MLLM processed each prompt independently, and the final answer was determined via majority voting over the five cleaned predictions.

3. Discussion and Conclusion

Our comprehensive strategy—unifying input pre-processing, domain-specific fine-tuning, Temporal Chain-of-Thought (T-CoT) prompting, answer cleaning, and ensembling—demonstrably elevates MLLM

performance on the HD-EPIC VQA benchmark. Employing Qwen2.5-VL-7B-Instruct, we observed that direct scaling to its 32B variant offered no proportional performance gains within our current pipeline. This is attributed to larger models' increased verbosity or uncertainty when constrained, and their extended T-CoT generations introducing noise detrimental to reasoning (e.g., an overabundance of detailed short video segments proved less effective than fewer, longer ones). Targeted fine-tuning of these larger 32B or 72B models is posited as a more promising path. Despite significant baseline improvements, our multi-stage architecture incurs latency, posing a critical performance-efficiency trade-off, particularly for real-time applications. Optimizing this balance is imperative for future work. Furthermore, a substantial gap persists towards human-level cognition, notably in tasks requiring deep reasoning and robust long-term memory for extended videos, underscoring a crucial research trajectory.

In conclusion, we presented a comprehensive methodology that demonstrably enhances MLLM efficacy for complex egocentric video understanding. Our results underscore the collective importance of structured input/output processing, domain adaptation, and guided temporal reasoning. Despite persistent challenges in computational efficiency and attaining human-level cognition, our work provides a robust baseline and crucial insights for the future development of advanced and practical AI systems targeting egocentric video analysis.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025. 1, 2, 3

[2] Siddhant Bansal, Chetan Arora, and C. V. Jawahar. My view is the best view: Procedure learning from egocentric videos. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIII*, pages 657–675. Springer, 2022. 2

[3] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. 1

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, et al. Scaling egocentric vision: The EPIC-KITCHENS dataset. *CoRR*, abs/1804.02748, 2018. 2

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, et al. The EPIC-KITCHENS dataset: Collection, challenges and baselines. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43 (11):4125–4141, 2021.

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, et al. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *Int. J. Comput. Vis.*, 130(1):33–55, 2022. 2

[7] Ahmad Darkhalil, Dandan Shan, Bin Zhu, et al. EPIC-KITCHENS VISOR benchmark: Video segmentations and object relations. In *Advances in Neural Information Processing Systems 35: Conference on NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 2

[8] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. *Robotics Institute*, 2009. 2

[9] DeepSeek-AI, Aixin Liu, Bei Feng, et al. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024. 1

[10] Chaoyou Fu, Yuhan Dai, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *CoRR*, abs/2405.21075, 2024. 1

[11] Google DeepMind. Gemini 2.5 pro preview model card. Technical report, Google, 2025. Technical report (preview release). 1

[12] Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4d: Around the world in 3, 000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18973–18990. IEEE, 2022. 1, 2

[13] Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024. 1

[14] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, pages 639–655. Springer, 2018. 2

[15] Baoqi Pei, Guo Chen, Jilan Xu, et al. Egovideo: Exploring egocentric foundation model and downstream adaptation. *CoRR*, abs/2406.18070, 2024. 1, 2

[16] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, et al. HD-EPIC: A highly-detailed egocentric video dataset. *CoRR*, abs/2502.04144, 2025. 1, 2, 3

[17] Heqian Qiu, Zhaofeng Shi, Lanxiao Wang, et al. Egome: Follow me via egocentric view in real world. *CoRR*, abs/2501.19061, 2025. 1

[18] Chen Wang, Fei Xia, Wenhao Yu, Tingnan Zhang, Ruohan Zhang, C. Karen Liu, Li Fei-Fei, Jie Tan, and Jacky Liang. Chain-of-modality: Learning manipulation programs from multimodal human videos with vision-language-models. *arXiv preprint arXiv:2504.13351*, 2025. 2

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 2

[20] xAI. Grok 3 beta — the age of reasoning agents. https://x.ai/blog/grok-3, 2025. 1

[21] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7590–7598. AAAI Press, 2018. 2

[22] Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1