VideoMind: A Chain-of-LoRA Agent for Temporal-Grounded Video Reasoning

Ye Liu[1*], **Kevin Qinghong Lin**[2*], **Chang Wen Chen**[1†], **Mike Zheng Shou**[2†]
[1] The Hong Kong Polytechnic University [2] National University of Singapore
https://videomind.github.io/

Abstract

Videos, with their unique temporal dimension, demand precise grounded understanding, where answers are directly linked to visual, interpretable evidence. Despite significant breakthroughs in text-based reasoning with large language models, multi-modal reasoning – especially for videos – remains limited. In this work, we fill this gap by introducing **VideoMind**, a novel video-language agent for temporal-grounded video reasoning. Our method involves two key innovations: (1) We identify four essential capabilities for grounded video reasoning and propose a role-based agentic workflow, comprising a `planner` to coordinate roles, a `grounder` for temporal event localization, a `verifier` to assess event candidates, and an `answerer` for question answering. (2) To efficiently integrate these roles during inference, we propose a novel **Chain-of-LoRA** mechanism, where a unified base model with multiple LoRA adapters is leveraged to enable seamless role switching, balancing efficiency and flexibility. Extensive experiments on 14 benchmarks across 3 tasks, including Grounded VideoQA, Video Temporal Grounding, and General VideoQA, demonstrate the effectiveness of the proposed scheme in advancing video agent, test-time scaling, and long-form video reasoning.

1 Introduction

Recent advancements in large language models (LLMs) have demonstrated remarkable success in text-based reasoning [72, 82, 57], significantly improving both accuracy and interpretability in complex problem-solving scenarios [83]. Following these breakthroughs, efforts have been devoted to extending these reasoning capabilities to multi-modal domains [92, 76, 63] such as vision-centric science [39] and math [42] understanding.

Compared with images or text, videos pose a unique challenge due to their temporal dimension. Effective video reasoning requires not only recognizing visual appearances but also understanding how they evolve over time [75, 7, 4]. While recent visual Chain-of-Thought (CoT) methods [92, 76, 63] excel at generating detailed thoughts for static images, they struggle with long videos as they cannot explicitly localize or revisit earlier parts of the sequence. Humans, by contrast, can reason over long videos with ease: they break down complex problems, identify relevant moments, revisit them to confirm details, and synthesize their observations into coherent answers. This natural proficiency motivates the development of an AI agent that emulates this process – flexibly coordinating multiple capabilities to achieve advanced, vision-centric reasoning.

In this work, we introduce **VideoMind**, a video-language agent with enhanced temporal-grounded reasoning capabilities. To meet the demands of diverse tasks, we define four essential roles for understanding complex long-form videos: (1) a `planner` to decompose tasks and coordinate other roles, (2) a `grounder` for precise moment localization, (3) a `verifier` for moment candidates assessment, and (4) an `answerer` for moment-aware response generation. Each role is carefully

[*] *Equal contribution.* [†] *Corresponding authors.*

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop.

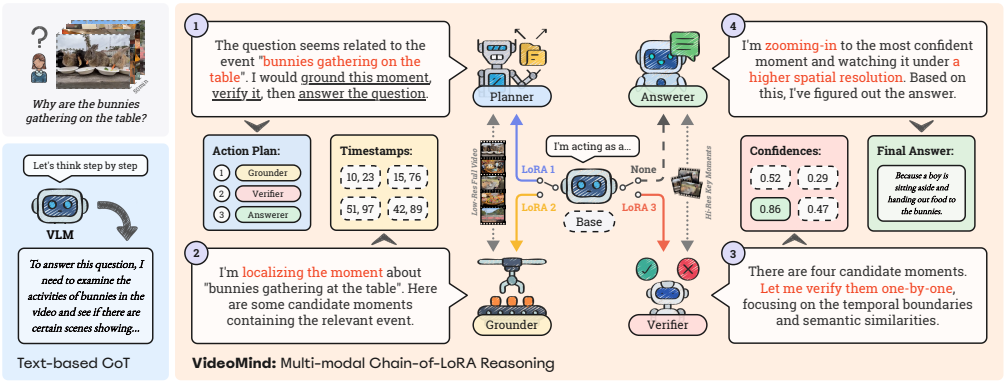

Figure 1: Illustration of VideoMind's Chain-of-LoRA reasoning strategy. The problem is decomposed by the planner and distributed to grounder, verifier, and answerer.

designed to deliver strong performance. To enable efficient integration of these roles, we also propose a novel **Chain-of-LoRA** mechanism, where all the roles are implemented based on the same base MLLM but with different LoRA adapters [16]. During inference, all the LoRA parameters are cached into the memory, so that each role could be activated by simply switching to the corresponding LoRA, as shown in Figure 1 (right). This approach facilitates seamless transitions and interactions among roles without incurring the memory overhead of maintaining multiple full models.

We conduct extensive experiments on 14 public benchmarks, including 3 on Grounded VideoQA, 6 on Video Temporal Grounding, and 5 on General VideoQA, to evaluate the effectiveness of our approach. VideoMind exhibits strong adaptability in addressing diverse reasoning tasks by jointly providing accurate responses and temporal-grounded evidence. Notably, our 2B model surpasses GPT-4o [48] and Gemini-1.5-Pro [54] on several long video benchmarks [4, 94, 66].

2 Method

Figure 2 provides an overview of VideoMind. It derives from Qwen2-VL [65]. Given a video input $\mathcal{V}$ and a text query $\mathcal{Q}$, the model adaptively performs step-by-step reasoning by calling individual roles.

2.1 Planner

An agent should be flexible enough to meet various demands and determine efficiently which function to call. To this end, we introduce the **Planner**, which dynamically coordinates all other roles for each query. We utilize a JSON-style object `{"type": "<role>", "value": "<argument>"}` to trigger a function call. In this way, a sequence of roles can be succinctly represented as a list of such objects. We define three reasoning plans illustrated below.

(1) Grounding & Verifying & Answering: This plan requires the agent to generate both a response and a temporal moment for grounded question-answering tasks [74] such as "*What is the boy doing when the baby is crying?*". **(2) Grounding & Verifying:** Designed for tasks focusing solely on grounding such as moment retrieval like "*When does the woman go downstairs?*". **(3) Answering Only:** When the question is straightforward (*e.g.*, "*Summarize this video*"), the model may not need to localize moments. Instead, it can watch the entire video and answer the question directly.

2.2 Grounder

Timestamp Decoder Rather than directly predicting timestamps through language modeling [55] or special tokens [17, 38], we develop a timestamp decoder on top of the LMM. We introduce a `<REG>` token to facilitate the timestamp decoding process. When this token is generated, the last-layer hidden states of `<REG>` and all the visual tokens will be sent into the decoder for timestamp prediction, obtaining a tuple $[t_{start}, t_{end}]$ representing the normalized start and end timestamps.

The decoder accepts hidden states of the visual tokens $\mathbf{h}_v \in \mathbb{R}^{(T\times H\times W)\times D_L}$ and the `<REG>` token $\mathbf{h}_r \in \mathbb{R}^{1\times D_L}$ as inputs, where T, H, W, D_L are the down-sampled number of frames, height, width,

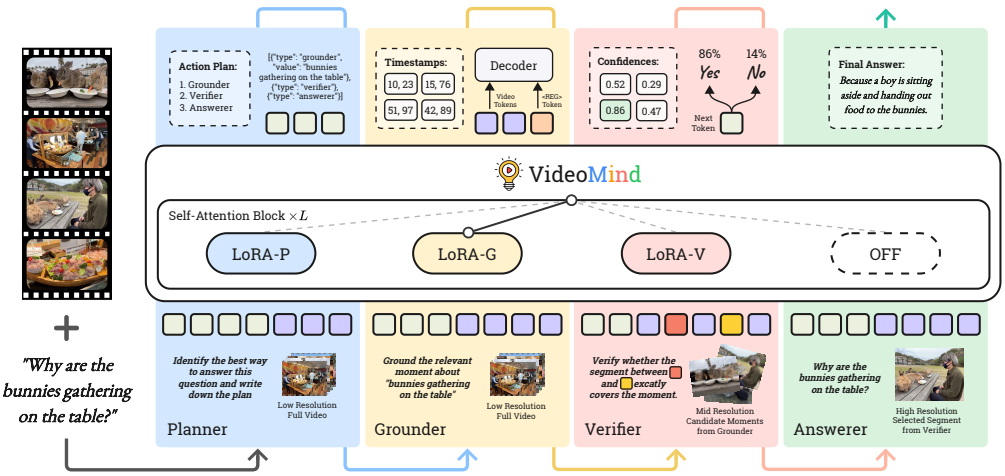

Figure 2: The overall workflow of VideoMind. Given a video and a query, it adaptively activates different roles and performs step-by-step reasoning by calling individual modules.

and hidden dimensions of the LLM, respectively. We apply a 1D average pooling to compress the visual tokens to one token per frame, denoted as $\mathbf{h}'_v$. Then, $\mathbf{h}'_v$ and $\mathbf{h}_r$ are projected by two linear layers E_v and E_r to reduce the hidden dimension to D. The resulting $\mathbf{e}_v$ and $\mathbf{e}_r$ serve as consolidated representations of the video frames and the query, respectively. To effectively integrate their information, we add them with learnable modality embeddings, concatenate them along the sequence dimension, and encode them with a transformer encoder. The output sequence is split into $\mathbf{e}'_v$ and $\mathbf{e}'_r$, indicating the contextualized frame and query embeddings, respectively. We then map $\mathbf{e}'_v$ into a four-level temporal feature pyramid to enhance its adaptability to varying moment lengths.

Prediction Heads **(1) A classification head** is adopted for frame-level foreground-background classification. This is instantiated by a two-layer `Conv1D` module (`kernel size=3, padding=1`) followed by a Sigmoid activation. **(2) A boundary regression head** is utilized to predict the 1D offset with the temporal boundaries $\{[\hat{b}_i^s, \hat{b}_i^e]\}_{i=0}^{L}$ for each frame. This is a two-layer convolution block, with an output dimension of 2 and an exponential function as activation.

2.3 Verifier

A key moment is crucial for providing visual cues, yet it may be imprecise due to its sensitivity. We let the grounder generate top-5 predictions, then apply the verifier to select the most reliable one.

Recap by Zoom-in For each candidate moment, we apply a zoom-in strategy by expanding the boundaries by 50% on both sides, cropping, and enlarging the resolution. The resulting video segment, together with the original text query, is then sent to the verifier to assess whether the queried event occurs within the segment. To enhance boundary awareness, we introduce two special tokens, `<SEG_START>` and `<SEG_END>`, to explicitly mark the beginning and end of the moment. These tokens are inserted among the visual tokens at the corresponding frames.

Boolean Judgement The verifier's responses are designed to be binary – either "*Yes*" or "*No*". To train the verifier, we sample predictions from the grounder on its training datasets and assign binary labels based on an IoU threshold of 0.5. The model is then fine-tuned via SFT to predict these labels. During inference, for each candidate moment, we employ teacher forcing to obtain the likelihoods of the `<Yes>` and `<No>` tokens, denoted as L_y and L_n, respectively. The confidence score for a moment is defined as $\text{Sigmoid}(L_y - L_n)$. We select the moment with the highest confidence score.

2.4 Answerer

The answerer is responsible for answering the given question based on the cropped video segment. Since the objective of this role is strictly aligned with existing LMMs, we employ the pre-trained model directly *without any fine-tuning or architectural modifications.*

To meet the diverse demands of different roles, we introduce a **Chain-of-LoRA** strategy to enable flexible switching. All the roles are built on top of the same backbone LMM and augmented with

Table 1: Performance comparison on Grounded VideoQA on CG-Bench [4].

Method	Size	long-acc.	mIoU	rec.@IoU	acc.@IoU
GPT-4o [48]	–	**45.2**	5.62	8.30	4.38
Gemini-1.5-Pro [54]	–	37.2	3.95	5.81	2.53
Video-LLaVA [29]	7B	16.2	1.13	1.96	0.59
VILA [30]	8B	28.7	1.56	2.89	1.35
LongVA [90]	7B	28.7	2.94	3.86	1.78
LLaVA-OV [24]	7B	31.1	1.63	1.78	1.08
VITA [10]	8×7B	33.3	3.06	3.53	2.06
Qwen2-VL [65]	72B	41.3	3.58	5.32	3.31
InternVL2 [61]	78B	42.2	3.91	5.05	2.64
VideoMind (Ours)	2B	31.0	5.94	8.50	4.02
VideoMind (Ours)	7B	38.4	**7.10**	**9.93**	**4.67**

Table 2: VideoQA performance on Video-MME [9], MLVU [94], and LVBench [66].

Method	Size	Video-MME		MLVU	LVBench
		All	Long	M-Avg	Overall
Gemini-1.5-Pro [54]	–	75.0	67.4	–	33.1
GPT-4o [48]	–	71.9	65.3	54.5	30.8
Video-LLaVA [29]	7B	41.1	37.8	29.3	–
TimeChat [55]	7B	34.3	32.1	30.9	22.3
MovieChat [59]	7B	38.2	33.4	25.8	22.5
PLLaVA [77]	34B	40.0	34.7	53.6	26.1
VideoChat-TPO [79]	7B	48.8	41.0	54.7	–
LongVA [90]	7B	52.6	46.2	56.3	–
VideoMind (Ours)	2B	53.6	45.4	58.7	35.4
VideoMind (Ours)	7B	**58.2**	**49.2**	**64.4**	**40.8**

Table 3: Comparison with different test-time scaling strategies. Mem means peak GPU memory.

Method	Mem	NExT-GQA		Charades-STA		Video-MME	
		mIoU	Acc	R@0.5	mIoU	All	Long
Qwen2-VL-2B	4.1G	–	69.6	–	–	49.7	43.1
+ CoT [72]	4.1G	–	69.7	–	–	49.6	43.2
+ All-in-One	4.2G	28.0	70.5	47.8	42.1	52.8	44.6
+ All-Distributed	16.6G	28.6	71.4	51.1	45.2	53.6	45.4
+ **Chain-of-LoRA**	**4.2G**	**28.6**	**71.4**	**51.1**	**45.2**	**53.6**	**45.4**

Table 4: Effects of individual roles. G% is the percentage of samples processed by grounder.

VideoMind Roles					ReXTime		Charades-STA		
Ans	Gnd	Ver	Pla	G%	mIoU	Acc	R@0.5	R@0.7	mIoU
✓				0%	–	68.0	–	–	–
✓	✓			100%	24.5	68.8	–	–	–
✓	✓	✓		100%	24.8	69.1	–	–	–
✓	✓	✓	✓	100%	24.7	69.2	–	–	–
✓	✓	✓	✓	40%	**26.7**	**70.0**	–	–	–
	✓				–	–	47.2	21.7	42.0
	✓	✓			–	–	**51.1**	**26.0**	**45.2**

additional LoRA adapters [16] and a lightweight timestamp decoder (for grounder only). The model dynamically activates role-specific LoRA adapters during inference via self-calling, allowing for maximizing role-specific capabilities while minimizing architectural modifications.

3 Experiments

We conduct experiments across various benchmarks. Some key results are presented here. Details about implementation, training data, benchmarks, and discussions are in the supplementary material.

Grounded Video Question-Answering In Table 1, we report results on CG-Bench [4], a challenging benchmark with an average duration of 27 minutes. In grounding metrics, our lightweight 2B model outperforms all compared models (including InternVL2-78B [61] and most closed-source models such as Gemini-1.5-Pro [54]), with the exception of GPT-4o [48], while our 7B model surpasses it and achieves competitive overall performance.

General Video Question-Answering We are also interested in whether our temporally augmented design can improve general VideoQA tasks. In Table 2, we evaluate our model on three widely used benchmarks to determine if the Chain-of-LoRA design generalizes to common settings. Our designs effectively help the model localize cue segments before answering the question.

Ablation Study We summarize some ablation study results in Table 3 and Table 4. (1) Naive text-based CoT does not improve the base model, highlighting the need for vision-centric reasoning. (2) Chain-of-LoRA achieves identical performance as `all-distributed`, but without multiple copies (4×) of the model weights. (3) All roles contribute to the final performance, where the grounder is crucial on long videos and the verifier consistently enhances temporal grounding accuracy. (4) Coordinating roles through the planner enables the model to flexibly adaptive to different context – performing grounding on only 40% samples yields higher accuracy but with less compute.

4 Conclusion

We introduced **VideoMind**, a novel video-language agent designed for temporal grounded video reasoning. Our approach employs an agentic workflow consisting of a Planner, Grounder, Verifier, and Answerer, along with a **Chain-of-LoRA** strategy to efficiently switch among these roles. Extensive experiments demonstrate the effectiveness and significance of VideoMind, particularly in long-form video reasoning tasks by providing precise, evidence-based answers. We hope this work inspires future advancements in multi-modal video agents and reasoning.

Acknowledgements

This study was supported by The Hong Kong RGC General Research Fund (15229423). We also acknowledge The University Research Facility in Big Data Analytics (UBDA) at The Hong Kong Polytechnic University for providing computing resources that have contributed to the research results reported within this paper.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv:2404.14219*, 2024.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023.

[3] Souradip Chakraborty, Soumya Suvra Ghosal, Ming Yin, Dinesh Manocha, Mengdi Wang, Amrit Singh Bedi, and Furong Huang. Transfer q star: Principled decoding for llm alignment. *arXiv:2405.20495*, 2024.

[4] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv:2412.12075*, 2024.

[5] Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Frank Wang. Rextime: A benchmark suite for reasoning-across-time in videos. In *NeurIPS*, pages 28662–28673, 2024.

[6] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. In *NeurIPS*, pages 19472–19495, 2024.

[7] Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. *arXiv:2312.06505*, 2023.

[8] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. *arXiv:2403.11481*, 2024.

[9] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv:2405.21075*, 2024.

[10] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv:2408.05211*, 2024.

[11] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. In *CVPR*, pages 22898–22909, 2023.

[12] Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv:2306.08640*, 2023.

[13] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017.

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022.

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv:2501.12948*, 2025.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[17] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *CVPR*, pages 14271–14280, 2024.

[18] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. *arXiv:2403.19046*, 2024.

[19] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv:2405.01483*, 2024.

[20] Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. Language repository for long video understanding. *arXiv:2403.14622*, 2024.

[21] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017.

[22] Hugo Laurencon, Leo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *NeurIPS*, pages 87874–87907, 2024.

[23] Jie Lei, Tamara L Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, 2021.

[24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv:2408.03326*, 2024.

[25] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv:2305.06355*, 2023.

[26] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206, 2024.

[27] Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. Momentdiff: Generative video moment retrieval from random to real. In *NeurIPS*, pages 65948–65966, 2023.

[28] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *ICLR*, 2023.

[29] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv:2311.10122*, 2023.

[30] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024.

[31] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *NeurIPS*, pages 7575–7586, 2022.

[32] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *CVPR*, pages 2794–2804, 2023.

[33] Kevin Qinghong Lin, Pengchuan Zhang, Difei Gao, Xide Xia, Joya Chen, Ziteng Gao, Jinheng Xie, Xuhong Xiao, and Mike Zheng Shou. Learning video context as interleaved multimodal sequences. In *ECCV*, pages 375–396. Springer, 2024.

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pages 34892–34916, 2023.

[35] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In *ECCV*, pages 1–18, 2024.

[36] Ye Liu, Jixuan He, Wanhua Li, Junsik Kim, Donglai Wei, Hanspeter Pfister, and Chang Wen Chen. r^2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In *ECCV*, 2024.

[37] Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, pages 3042–3051, 2022.

[38] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang W Chen. E.t. bench: Towards open-ended event-level video-language understanding. In *NeurIPS*, pages 32076–32110, 2024.

[39] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, pages 2507–2521, 2022.

[40] Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training. In *CVPR*, pages 23045–23055, 2023.

[41] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv:2306.07207*, 2023.

[42] Zongyang Ma, Yuxin Chen, Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Shaojie Zhu, Chengxiang Zhuo, Bing Li, Ye Liu, Zang Li, Ying Shan, and Weiming Hu. Visionmath: Vision-form mathematical problem-solving. In *ICCV*, 2025.

[43] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv:2406.09418*, 2024.

[44] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv:2306.05424*, 2023.

[45] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019.

[46] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *CVPR*, pages 23023–23033, 2023.

[47] OpenAI. Gpt-4v(ision) system card, 2023.

[48] OpenAI. Gpt-4o system card, 2024.

[49] OpenAI. Openai o1 system card, 2024.

[50] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv:2402.11435*, 2024.

[51] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *NeurIPS*, pages 119336–119360, 2024.

[52] Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. Chatvtg: Video temporal grounding via chat with video dialogue large language models. In *CVPR*, pages 1847–1856, 2024.

[53] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

[54] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024.

[55] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, pages 14313–14323, 2024.

[56] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

[57] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*, pages 8634–8652, 2023.

[58] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[59] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv:2307.16449*, 2023.

[60] Didac Suris, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*, pages 11888–11898, 2023.

[61] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024.

[62] Siyu Teng, Xuemin Hu, Peng Deng, Bai Li, Yuchen Li, Yunfeng Ai, Dongsheng Yang, Lingxi Li, Zhe Xuanyuan, Fenghua Zhu, et al. Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Transactions on Intelligent Vehicles*, 8(6):3692–3711, 2023.

[63] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv:2501.06186*, 2025.

[64] Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Dian Yu, Haitao Mi, Jinsong Su, and Dong Yu. Litesearch: Efficacious tree search for llm. *arXiv:2407.00320*, 2024.

[65] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024.

[66] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv:2406.08035*, 2024.

[67] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *arXiv:2403.10517*, 2024.

[68] Xiyao Wang, Linfeng Song, Ye Tian, Dian Yu, Baolin Peng, Haitao Mi, Furong Huang, and Dong Yu. Towards self-improvement of llms via mcts: Leveraging stepwise knowledge with curriculum preference learning. *arXiv:2410.06508*, 2024.

[69] Xiyao Wang, Ruijie Zheng, Yanchao Sun, Ruonan Jia, Wichayaporn Wongkamjan, Huazhe Xu, and Furong Huang. Coplanner: Plan to roll out conservatively but to explore optimistically for model-based rl. *arXiv:2310.07220*, 2023.

[70] Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text llms for grounding text in videos. *arXiv:2403.10228*, 2024.

[71] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *AAAI*, volume 36, pages 2613–2623, 2022.

[72] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, pages 24824–24837, 2022.

[73] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS*, pages 28828–28857, 2024.

[74] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021.

[75] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *CVPR*, pages 13204–13214, 2024.

[76] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *ICCV*, 2023.

[77] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv:2404.16994*, 2024.

[78] Yuancheng Xu, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. Genarm: Reward guided generation with autoregressive reward model for test-time alignment. *arXiv:2410.08193*, 2024.

[79] Ziang Yan, Zhilin Li, Yinan He, Chenting Wang, Kunchang Li, Xinhao Li, Xiangyu Zeng, Zilei Wang, Yali Wang, Yu Qiao, et al. Task preference optimization: Improving multimodal large language models with vision task alignment. *arXiv:2412.19326*, 2024.

[80] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022.

[81] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv:2303.11381*, 2023.

[82] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, pages 11809–11822, 2023.

[83] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

[84] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *NeurIPS*, pages 76749–76771, 2023.

[85] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *CVPR*, pages 23056–23065, 2023.

[86] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv:2312.17235*, 2023.

[87] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. *arXiv:2406.03816*, 2024.

[88] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv:2306.02858*, 2023.

[89] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv:2004.13931*, 2020.

[90] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv:2406.16852*, 2024.

[91] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, pages 12870–12877, 2020.

[92] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv:2302.00923*, 2023.

[93] Yue Zhao, Ishan Misra, Philipp Krahenbuhl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, pages 6586–6597, 2023.

[94] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv:2406.04264*, 2024.

Appendix

A Related Work

A.1 Temporal-grounded Video Understanding

Significant advances in video understanding have propelled tasks such as video captioning [93, 33], video question answering [74, 86], and video-text retrieval [45, 31], which emphasize instance-level understanding, yet these models often lack *visual-grounded correspondence* and interpretability, particularly for long-form video streams. The task of Video Temporal Grounding [13, 21] tackles this by requiring precise temporal localization for diverse queries, though regression-based models [37, 36] excel at localization but fall short in providing textual interpretability. Recent benchmarks [75, 4] intensify this challenge, demanding both reasoning for complex questions and fine-grained temporal correspondence. Previous baselines for these tasks typically rely on multi-task objectives or modular agents composed of distinct components [86, 84, 67, 8], often yielding sub-optimal performance (*e.g.*, LLM-based approaches for temporal grounding) or overly complex systems, which constrain their efficiency and flexibility. In this work, our proposed VideoMind introduces an agentic workflow built upon a unified backbone, seamlessly integrating multiple functionalities while enhancing localization and interpretability, thus surpassing the limitations of prior methods.

A.2 Multi-modal Reasoning

Large Multi-modal Models [34], trained with supervised instruction-tuning (SFT), exhibit generalized capabilities such as free-form dialogue and question answering; however, they fall short in addressing complex challenges that often require the reasoning abilities of LLMs [72]. One approach to overcome this is to develop agent-based interfaces [86, 20], which integrates textual outputs from multiple visual tools to enable language reasoning via LLMs. Advanced methods [60, 81, 12] leverage strategies like Codex or ReAct [83] to invoke visual APIs (*e.g.*, detectors, captioners) through progressive execution and reasoning. Alternatively, pure text-based reasoning [49, 15] has been a dominant paradigm in LLMs [72, 86], exemplified by training with long CoT processes using Reinforcement Learning, which provides detailed, step-by-step readable reasoning, with some works [92**?**] extending this to the visual domain for complex mathematical or scientific problems. Despite these advances, extending reasoning to videos across temporal dimensions remains an open challenge. Given the long-context nature of informative videos, we think that *a video-centric CoT* should incorporate a human-like re-watch strategy and self-validation of intermediate observations, leading us to introduce a novel Chain-of-LoRA framework for video reasoning.

A.3 Inference-time Searching

Inference-time searching has emerged as a critical technique for tackling complex reasoning and planning challenges in domains like robotics [69], games [58], and navigation [62], distinct from training-time strategies as it optimizes model behavior during inference rather than model parameters during training. The advent of OpenAI o1 [49] has advanced these inference-time techniques within LLMs by integrating sampling strategies such as controlled decoding [3, 78], Best-of-N sampling [28], and Monte Carlo Tree Search (MCTS) [68, 87, 64], allowing LLMs to iteratively refine outputs and achieve superior performance without altering their underlying weights. However, the potential of inference-time searching remains largely untapped in video understanding, where temporal reasoning introduces unique challenges. In our framework, we explore how MCTS can be tailored for video temporal reasoning, observing that models are highly sensitive to the selection of temporal segments, often producing unreliable predictions when segment choices are sub-optimal. To address this, we propose a *moment-level* searching approach where a grounder generates multiple candidates, followed by a verifier that evaluates and determines the correct correspondence. This strategy significantly enhances temporal grounding accuracy and robustness across diverse scenarios.

B Model Details

B.1 Query Rephrasing

When the user query lacks sufficient detail for accurate localization, the planner is allowed to **rephrase** the question into a more descriptive version. For instance, the question "*What is the person sitting on the bed doing as the baby plays?*" might confuse the grounder as it contains multiple instances (person and baby). It can be rephrased to "*the baby is playing*" for accurate scene description. We leverage GPT-4o mini [48] generated question-query pairs for training.

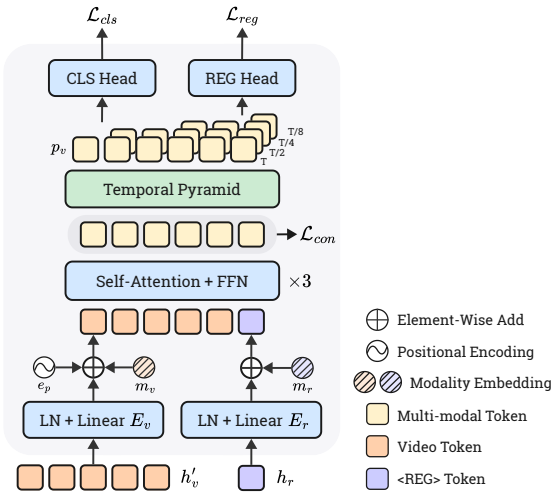

Figure 3: Detailed architecture of the timestamp decoder with temporal feature pyramid.

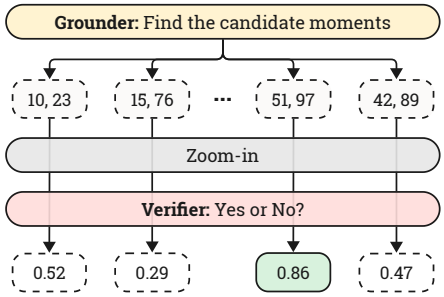

Figure 4: Schematic illustration of the grounding and verifying process.

Table 5: Training datasets for different roles. The planning dataset is repurposed from NExT-QA [74] and QVHighlights [23]. Verify datasets are generated from the pre-trained grounder's predictions. *mr* and *step* denote the moment retrieval and step localization subsets of HiREST [85], respectively.

Role	#Samples	Datasets
Planner	39K	NeXT-QA-Plan (34K), QVHighlights-Plan (5K)
Grounder	210K	QVHighlights (5K), DiDeMo (33K), TACoS (9K), InternVid-VTime (54K), CosMo-Cap (87K), QuerYD (19K), HiREST$_{mr}$ (8K), HiREST$_{step}$ (4K)
Verifier	232K	DiDeMo-Verify (165K), TACoS-Verify (43K), QVHighlights-Verify (24K)

B.2 Temporal Feature Pyramid

To enhance adaptability to varying lengths of videos and moments, we map $\mathbf{e}'_v$ into a four-level temporal feature pyramid with multiple temporal resolutions. This is achieved by applying `Conv1D` $\rightarrow$ `LayerNorm` $\rightarrow$ `SiLU` blocks for each pyramid level, where the `Conv1D` employs a kernel size and stride of 2 (down-sampling the sequence by $1/2$ along the temporal dimension). In practice, the four levels retain 1, $1/2$, $1/4$, and $1/8$ of the original sequence length, respectively. They can be denoted as $\mathbf{p}_v^n \in \mathbb{R}^{(T/n^2)\times D}$ where $n \in [1, 4]$ is the index of the pyramid level. To accelerate the prediction process, we concatenate the sequences from all pyramid levels along the temporal dimension into $\mathbf{p}_v$ with length $L = T + T/2 + T/4 + T/8$, such that the prediction can be made in parallel.

C Experiments

We conduct extensive experiments across various benchmarks to evaluate our VideoMind. Specifically, we study the following research questions.

- **Q1.** Whether VideoMind is flexible and effective on different video temporal reasoning tasks compared to baselines with task-specific designs?
- **Q2.** Compared with training a single agent on multiple tasks, what advantages does Chain-of-LoRA offer?
- **Q3.** What effects does each individual design contribute?

C.1 Benchmarks and Settings

The experiments are extensively designed across 14 diverse benchmarks, where the statistics are listed in Table 6. We mainly evaluate VideoMind on grounded VideoQA and video temporal grounding scenarios, and also study its generalizability on general long VideoQA benchmarks. The information about major benchmarks is introduced below.

Table 6: Statistics of evaluation benchmarks. The datasets encompass three representative tasks – Grounded VideoQA, Video Temporal Grounding, and General VideoQA, with video durations ranging from several seconds to more than 1 hour.

Dataset	Duration	Domain	Main Metrics
Grounded Video Question-Answering (Grounding + QA)			
CG-Bench [4]	1624.4s	Diverse	rec.@IoU, acc.@IoU
ReXTime [5]	141.1s	Vlog, News, Activity	mIoU, Acc (IoU $\geqslant 0.5$)
NExT-GQA [75]	39.5s	Reasoning	mIoP, Acc@GQA
Video Temporal Grounding (Grounding only)			
Charades-STA [13]	30.1s	Indoor	R@$\{0.3 \sim 0.7\}$, mIoU
ActivityNet-Captions [21]	111.4s	Activity	R@$\{0.3 \sim 0.7\}$, mIoU
QVHighlights [23]	150s	Vlog, News	R@$\{0.5, 0.7\}$, mAP
TACoS [53]	358.2s	Cooking	R@$\{0.3 \sim 0.7\}$, mIoU
Ego4D-NLQ [14]	379.0s	Egocentric	R@$\{0.3 \sim 0.7\}$, mIoU
ActivityNet-RTL [18]	111.4s	Reasoning	P@0.5, mIoU
General Video Question-Answering (QA only)			
Video-MME [9]	1017.9s	Diverse	Acc (w/o subs)
MLVU [94]	930s	Diverse	Acc
LVBench [66]	4101s	Diverse	Acc
MVBench [26]	15s	Diverse	Acc
LongVideoBench [73]	473s	Diverse	Acc

Table 7: Performance comparison on Grounded VideoQA on ReXTime [5]. FT indicates whether fine-tuned on the downstream training set.

Method	Size	FT	R@0.3	R@0.5	mIoU	Acc	Acc@IoU
VTimeLLM [17]	7B		28.84	17.41	20.14	36.16	–
TimeChat [55]	7B		14.42	7.61	11.65	40.04	–
LITA [18]	13B		29.49	16.29	21.49	34.44	–
VTimeLLM [17]	7B	✓	43.69	26.13	29.92	57.58	17.13
TimeChat [55]	7B	✓	40.13	21.42	26.29	49.46	10.92
VideoMind (Ours)	2B		34.31	22.69	24.83	69.06	17.26
VideoMind (Ours)	7B		**38.22**	**25.52**	**27.61**	**74.59**	**20.20**

Table 8: Performance comparison on Grounded VideoQA on NExT-GQA [75].

Method	Size	IoU			IoP			Acc@GQA
		R@0.3	R@0.5	mIoU	R@0.3	R@0.5	mIoP	
FrozenBiLM NG+ [80]	890M	13.5	6.1	9.6	28.5	23.7	24.2	17.5
VIOLETv2 [11]	–	4.3	1.3	3.1	25.1	23.3	23.6	12.8
SeViLA [84]	4B	29.2	13.8	21.7	34.7	22.9	29.5	16.6
LangRepo [20]	8×7B	–	12.2	18.5	–	28.7	31.3	17.1
VideoStreaming [51]	8.3B	–	13.3	19.3	–	31.0	32.2	17.8
LLoVi [86]	1.8T	–	15.3	20.0	–	36.9	37.3	24.3
HawkEye [70]	7B	37.0	19.5	25.7	–	–	–	–
VideoChat-TPO [79]	7B	41.2	23.4	27.7	47.5	32.8	35.6	25.5
VideoMind (Ours)	2B	45.2	23.2	28.6	51.3	32.6	36.4	25.2
VideoMind (Ours)	7B	**50.2**	**25.8**	**31.4**	**56.0**	**35.3**	**39.0**	**28.2**

CG-Bench [4] is designed for long video grounded question-answering, featuring a diverse domain and various evaluation metrics. It includes 1.2K manually curated videos, ranging from 10 to 80 minutes, with a total of 12K QA pairs. The dataset is categorized into perception, reasoning, and hallucination question types, and introduces clue-based evaluation methods like white box and black box assessments to ensure models provide answers based on accurate video reasoning.

ReXTime [5] tests models on complex temporal reasoning, using an automated pipeline for QA pair generation, significantly reducing manual effort. It includes 921 validation and 2,1K test samples, each manually curated for accuracy, and highlights a 14.3% accuracy gap between SoTA models and human performance. This benchmark is crucial for evaluating models on cause-and-effect relationships across video segments, driving advancements in video understanding research.

NExT-GQA [75] aims to challenge models to reason about causal and temporal actions, supporting both multi-choice and open-ended tasks. This is an extension of NExT-QA [74] comprising 10.5K manually labeled

Table 9: Performance of video temporal grounding on Charades-STA [13].

Method	Size	FT	R@0.3	R@0.5	R@0.7	mIoU
Moment-DETR [23]	–	✓	65.8	52.1	30.6	45.5
UniVTG [32]	–	✓	70.8	58.1	35.6	50.1
R^2-Tuning [36]	–	✓	70.9	59.8	37.0	50.9
VTimeLLM [17]	13B		55.3	34.3	14.7	34.6
TimeChat [55]	7B		51.5	32.2	13.4	–
Momentor [50]	7B		42.6	26.6	11.6	28.5
HawkEye [70]	7B		50.6	31.4	14.5	33.7
ChatVTG [52]	7B		52.7	33.0	15.9	34.9
VideoChat-TPO [79]	7B		58.3	40.2	18.4	38.1
E.T. Chat [38]	4B		65.7	45.9	20.0	42.3
VideoMind (Ours)	2B		67.6	51.1	26.0	45.2
VideoMind (Ours)	7B		**73.5**	**59.1**	**31.2**	**50.2**

Table 10: Performance of video temporal grounding on ANet-Captions [21].

Method	Size	FT	R@0.3	R@0.5	R@0.7	mIoU
2D-TAN [91]	–	✓	60.4	43.4	25.0	42.5
MMN [71]	–	✓	64.5	48.2	29.4	46.6
VDI [40]	–	✓	–	48.1	28.8	–
VideoChat [25]	7B		8.8	3.7	1.5	7.2
Video-LLaMA [88]	7B		6.9	2.1	0.8	6.5
Video-ChatGPT [44]	7B		26.4	13.6	6.1	18.9
Valley [41]	7B		30.6	13.7	8.1	21.9
ChatVTG [52]	7B		40.7	22.5	9.4	27.2
Momentor [50]	7B		42.9	23.0	12.4	29.3
E.T. Chat [38]	4B		24.1	12.8	6.1	18.9
VideoMind (Ours)	2B		44.0	26.5	12.6	30.1
VideoMind (Ours)	7B		**48.4**	**30.3**	**15.7**	**33.3**

Table 11: Video temporal grounding on TACoS [53]. Note that our method was co-trained on this dataset during pre-training.

Method	Size	FT	R@0.3	R@0.5	R@0.7	mIoU
Non-LLM-based Specialists						
2D-TAN [91]	–	✓	40.0	28.0	12.9	27.2
Moment-DETR [23]	–	✓	38.0	24.7	12.0	25.5
UniVTG [32]	–	✓	51.4	35.0	17.4	33.6
R^2-Tuning [36]	–	✓	49.7	38.7	25.1	35.9
LLM-based Models						
VideoMind (Ours)	2B	✓	38.6	26.9	15.5	27.4
VideoMind (Ours)	7B	✓	**49.5**	**36.2**	**21.4**	**34.4**

Table 12: Performance of video temporal grounding on Ego4D-NLQ [14].

Method	Size	FT	R@0.3	R@0.5	R@0.7	mIoU
Non-LLM-based Specialists						
2D-TAN [91]	–	✓	4.3	1.8	0.6	3.4
VSLNet [89]	–	✓	4.5	2.4	1.0	3.5
Moment-DETR [23]	–	✓	4.3	1.8	0.7	3.5
UniVTG [32]	–	✓	7.3	4.0	1.3	4.9
R^2-Tuning [36]	–	✓	7.2	4.5	2.1	4.9
UniVTG [32]	–		6.5	3.5	1.2	4.6
LLM-based Models						
VideoMind (Ours)	2B		5.9	2.9	1.2	4.7
VideoMind (Ours)	7B		**7.2**	**3.7**	**1.7**	**5.4**

video QA pairs with temporal segments. The samples in this benchmark are from "causal" and "temporal" classes, while the "descriptive" questions in NExT-QA are discarded.

Charades-STA [13] contains 10K in-door videos, averaging 30.1 seconds each, with 16K temporal annotations spanning daily activity, alongside free-text descriptions. These rich annotations make Charades-STA particularly suitable for evaluating temporal grounding models under indoor environments.

ActivityNet-Captions [21] is a large-scale benchmark with 20K untrimmed YouTube videos totaling 849 hours, covering diverse activities from personal care to sports. This dataset contains high-quality dense video captioning annotations (3.65 temporally localized sentences per video), which we use as queries for video temporal grounding. Each query has an average length of 13.5 words.

C.2 Implementation Details

We leverage the 2B and 7B versions of Qwen2-VL [65] as our base model, and apply LoRA adapters with rank 64 and alpha 64 to planner, grounder, and verifier. The hidden size of the timestamp decoder in the grounder is 256. The maximum number of tokens per frame and maximum number of frames for planner, grounder, verifier, and answerer are set as [64, 100], [64, 150], [64, 64], and [256, 32], respectively. We train different roles separately on different datasets (listed in Table 5) and load them together by setting different names for LoRA adapters, so that the model can efficiently switch roles by actively setting different LoRAs. During training, we set the global batch size as 32, and utilize the AdamW optimizer with learning rates of 2e-5, 1e-4, and 5e-5 for planner, grounder, and verifier, respectively. All the roles were trained for 1 epoch on their specific datasets, with a linear warmup in the first 3% steps. During inference, we apply NMS with an IoU threshold of 0.75 to reduce duplicated moments from the grounder.

C.3 Q1: Comparison with State-of-the-Arts

Grounded Video Question-Answering In Table 7, we show the results on ReXTime [5]. Despite the challenge posed by the temporal and causal event relationships, our model successfully identifies the correct event and zooms in on the relevant moment. Notably, our zero-shot model outperforms all zero-shot baselines by

Table 13: Fine-tuned video temporal grounding results on QVHighlights [23].

Method	Size	R1		mAP		
		@0.5	@0.7	@0.5	@0.75	Avg.
Non-LLM-based Specialists						
Moment-DETR [23]	–	59.78	40.33	60.51	35.36	36.14
UMT [37]	–	60.83	43.26	57.33	39.12	38.08
MomentDiff [27]	–	58.21	41.48	54.57	37.21	36.84
QD-DETR [46]	–	62.40	44.98	62.52	39.88	39.86
UniVTG [32]	–	65.43	50.06	64.06	45.02	43.63
R^2-Tuning [36]	–	68.03	49.35	69.04	47.56	46.17
LLM-based Models						
VideoMind (Ours)	2B	75.42	59.35	74.11	55.15	51.60
VideoMind (Ours)	7B	**78.53**	**61.09**	**76.07**	**58.17**	**54.19**

Table 14: Performance of reasoning temporal localization on ActivityNet-RTL [18]. Our zero-shot 7B model outperforms the fine-tuned baseline LITA [18] by a considerable margin.

Method	Size	FT	P@0.5	mIoU
LITA [18]	7B	✓	21.2	24.1
LITA [18]	13B	✓	25.9	28.6
VideoMind (Ours)	2B		20.1	22.7
VideoMind (Ours)	7B		**28.0**	**31.3**

Table 15: Performance of VideoQA on LongVideoBench [73] `val` split.

Method	Size	Acc	Acc @ Duration Groups			
			(8, 15]	(15, 60]	(180, 600]	(900, 3600]
GPT-4o [48]	–	66.7	71.4	76.7	69.1	60.9
GPT-4 Turbo [2]	–	59.0	65.2	68.2	62.4	50.5
Gemini-1.5-Pro [54]	–	64.0	67.4	75.1	65.3	58.6
Gemini-1.5-Flash [54]	–	61.6	68.3	76.2	62.6	54.0
Idefics2 [22]	8B	49.7	59.8	65.7	47.8	42.7
Phi-3-Vision [1]	4B	49.6	59.3	61.6	46.8	44.7
Mantis-Idefics2 [19]	8B	47.0	56.6	55.8	45.6	42.2
Mantis-BakLLaVA [19]	7B	43.7	53.4	57.6	40.3	38.7
VideoMind (Ours)	2B	48.8	59.3	59.3	49.3	41.7
VideoMind (Ours)	7B	**56.3**	**67.7**	**67.4**	**56.8**	**48.6**

a significant margin and yields comparable performance to several fine-tuned variants in grounding, while also achieving higher accuracy. This demonstrates its strong generalization capability.

We further present results on NExT-GQA [75] in Table 8. Compared to text-rich, agent-based solutions such as LLoVi [86] and LangRepo [20] – which leverage additional tools and chain-of-thought, and SeViLA [84] – a self-chained video agent with a similar design, our 2B model matches the performance of state-of-the-art 7B models across both agent-based and end-to-end approaches. Moreover, our 7B model significantly outperforms all other models.

Video Temporal Grounding Since the performance of the grounder and verifier is essential for VideoMind, we evaluate these modules on temporal grounding datasets. In Table 9 and Table 10, we validate the zero-shot grounding capabilities of VideoMind. Benefiting from (1) the timestamp decoder design of the grounder, and (2) a verifier that refines the results by focusing on critical segments, our model achieves significant zero-shot performance – surpassing all LLM-based temporal grounding methods and yielding competitive results compared to fine-tuned temporal grounding experts.

We additionally compare VideoMind with representative methods on the challenging TACoS [53], Ego4D-NLQ [14], and QVHighlights [23] datasets in Table 11, Table 12, and Table 13, respectively. Our model performs better than the strong task-specific baseline UniVTG [32] on TACoS but slightly worse than it on Ego4D-NLQ, because UniVTG was originally pre-trained on egocentric videos. Even without egocentric pre-training, VideoMind can still produce comparable results on Ego4D-NLQ. To our best knowledge, VideoMind is **the first LLM-based grounding model that supports multi-moment outputs**, thereby being able to be evaluated on QVHighlights. Compared with task-specific experts, our VideoMind-2B significantly outperforms all previous methods on this challenging dataset.

Reasoning Temporal Localization. We also evaluate the generalization ability of grounder and verifier on the more challenging reasoning temporal localization [18] task, which is similar to video temporal grounding, but the queries are not directly describing the moment. The models are required to infer the actual event using their world knowledge. The results in Table 14 show that VideoMind can successfully generalize its strong zero-shot grounding capability to complex scenarios.

General Video Question-Answering For long VideoQA, we provide evaluations on LongVideoBench [73] in Table 15, which further verifies the effectiveness of VideoMind on videos scaling to one-hour long. Table 16

Table 16: Performance comparison on general VideoQA on MVBench [26].

Model	Size	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI	Avg.
GPT-4V [47]	–	55.5	63.5	72.0	46.5	73.5	18.5	59.0	29.5	12.0	40.5	83.5	39.0	12.0	22.5	45.0	47.5	52.0	31.0	**59.0**	11.0	43.5
Video-ChatGPT [44]	7B	23.5	26.0	62.0	22.5	26.5	54.0	28.0	40.0	23.0	20.0	31.0	30.5	25.5	39.5	48.5	29.0	33.0	29.5	26.0	35.5	32.7
Video-LLaMA [88]	7B	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	32.5	45.5	32.5	40.0	30.0	21.0	37.0	34.1
VideoChat [25]	7B	33.5	26.5	56.0	33.5	40.5	53.0	40.5	30.0	25.5	27.0	48.5	35.0	20.5	42.5	46.0	26.5	41.0	23.5	23.5	36.0	35.5
Video-LLaVA [29]	7B	46.0	42.5	56.5	39.0	53.5	53.0	48.0	41.0	29.0	31.5	82.5	45.0	26.0	53.0	41.5	33.5	41.5	27.5	38.5	31.5	43.0
TimeChat [55]	7B	40.5	36.0	61.0	32.5	53.0	53.5	41.5	29.0	19.5	26.5	66.5	34.0	20.0	43.5	42.0	36.5	36.0	29.0	35.0	35.0	38.5
PLLaVA [77]	7B	58.0	49.0	55.5	41.0	61.0	56.0	61.0	36.0	23.5	26.0	82.0	39.5	42.0	52.0	45.0	42.0	53.5	30.5	48.0	31.0	46.6
ShareGPT4Video [6]	7B	49.5	39.5	79.5	40.0	54.5	82.5	54.5	32.5	50.5	41.5	84.5	35.5	62.5	75.0	51.0	25.5	46.5	28.5	39.0	51.5	51.2
ST-LLM [35]	7B	66.0	53.5	**84.0**	44.0	58.5	80.5	73.5	38.5	42.5	31.0	86.5	36.5	56.5	78.5	43.0	44.5	46.5	34.5	41.5	58.5	54.9
VideoGPT+ [43]	3.8B	69.0	60.0	83.0	48.5	66.5	85.5	75.5	36.0	44.0	34.0	89.5	39.5	71.0	90.5	45.0	53.0	50.0	29.5	44.0	60.0	58.7
VideoChat2 [26]	7B	75.5	58.0	83.5	**50.5**	60.5	87.5	74.5	**45.0**	47.5	44.0	82.5	37.0	64.5	87.5	**51.0**	**66.5**	47.0	35.0	37.0	72.5	60.4
VideoMind (Ours)	2B	**77.0**	**78.0**	77.0	46.5	70.5	87.0	71.5	33.0	48.0	39.5	**91.0**	**53.0**	78.0	89.0	43.5	53.5	61.5	37.5	49.5	53.0	61.9
VideoMind (Ours)	7B	74.0	71.5	81.0	50.0	**77.0**	**93.0**	**75.0**	38.0	**48.5**	**46.0**	**91.0**	39.0	**80.0**	**94.5**	49.5	55.5	**70.0**	**40.5**	57.0	**61.0**	**64.6**

Table 17: Effect of the temporal feature pyramid on Charades-STA [13].

#Pyramid Levels	Charades-STA			
	R@0.3	R@0.5	R@0.7	mIoU
1	60.55	44.57	15.82	38.13
2	61.51	46.90	19.36	40.43
3	62.62	47.02	20.08	41.27
4	**63.55**	**47.23**	**21.69**	**42.02**

Table 18: Effect of different verifier styles on Charades-STA [13].

Verifier Type	Charades-STA			
	R@0.3	R@0.5	R@0.7	mIoU
Direct	60.42	45.28	19.32	39.84
Expand	65.10	48.70	23.15	43.57
Textual	65.24	49.33	23.89	44.01
Special Token	**67.63**	**51.05**	**25.99**	**45.22**

presents more results of VideoMind on MVBench [26], which is a benchmark with very short videos (around 15s). Our model can still achieve good performance on these short video scenarios.

C.4 Q2: The Advantages of Chain-of-LoRA

Table 3 investigates the impact of integrating different modules. First, naive CoT does not improve the base model, highlighting the need for a visual-centric test-time scaling strategy. Second, although the all-distributed approach achieves the best performance, it requires multiple copies ($4\times$) of the model weights. In contrast, Chain-of-LoRA maintains top performance while being efficient.

C.5 Q3: Ablation Studies

Effect of Individual Roles The contribution of each roleis studies in Table 4. Our observations are as follows: **(1) Grounder:** By identifying visual cues, the grounder can slightly improve QA Acc, indicating that the grounder is especially effective on long videos. **(2) Verifier:** Selecting the best candidate with the verifier improves grounded moments, yielding a consistent gain of 3.2 mIoU on the pure grounding metrics for Charades-STA. **(3) Planner:** Coordinating roles via the planner – even when performing grounding on only 40% samples (the remaining 60% are directly processed by the answerer) – boosts the accuracy from 69.2 to 70.0. This highlights the model's flexibility under different temporal contexts.

Effect of the Temporal Feature Pyramid Table 17 studies the effectiveness of the temporal feature pyramid. Our baseline model directly makes predictions on the last-layer transformer outputs. When adding more pyramid levels, the performance of video temporal grounding consistently improves under all metrics on the Charades-STA [13] dataset under zero-shot setting, suggesting the effectiveness of improving the robustness of the model when facing moments with different lengths.

Design Choices of Verifier In Table 18, we examine various design choices for the verifier. The term "Direct" refers to the method where the grounded moment is directly sent into the model without any expansion. "Expand" denotes expanding the temporal boundaries by 50%, while "Textual" involves adding supplementary textual information to indicate the length of the target event. "Special Token" represents our final approach, utilizing special tokens to denote the grounded start and end timestamps. The comparison demonstrates that expanding the temporal boundaries effectively broadens the verifier's perceptual range, and the use of special tokens enhances the model's awareness of precise moment boundaries.

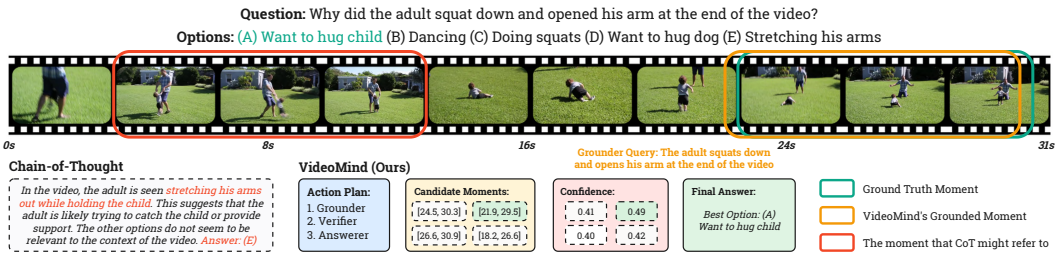

Figure 5: Visualization of VideoMind's reasoning process. Through chaining the grounder, verifier, and answerer, our model accurately localizes the critical moment and selects the correct answer, avoiding confusion from incorrect segments.

C.6 Visualization

In Figure 5, we illustrate how VideoMind applies all roles to progressively derive the correct answer while avoiding potential mistakes or confusion. The Planner determines what roles are needed, then calls the grounder to generate candidate moments. The verifier selects the most relevant segment (highlighted in yellow), which is then zoomed in and passed to the answerer for further reasoning.

D Limitations & Future Work

The proposed framework represents our initial attempt at a multi-agent-based solution for interactive video understanding. In this work, we identify two limitations that warrant deeper exploration. First, the current planner design and role distribution are based on heuristic decisions, which might be sub-optimal and can potentially be more flexible. For example, a possible enhancement of planner would be allowing it to call the answerer for an initial answer generation, then generate the detailed plan based on the preliminary answer. Second, the entire inference pipeline lacks an explicit reflection design. Therefore, the overall performance of each role is largely affected by the previous role. Enabling the reflection capability is essential for building a real interactive content understanding agent. To address these limitations, a promising direction for future work is to integrate the planner and answerer into a unified agent that possesses reflection capabilities. This would allow the model, upon detecting that a predicted answer may be incorrect, to re-trigger the grounder and/or verifier dynamically for correction and improvement. Modern reinforcement learning training techniques such as PPO [56] and GRPO [15] might also be utilized.