

TimeZero: Temporal Video Grounding with Reasoning-Guided LVLM

Ye Wang^{1*} Boshen Xu^{1*} Zihao Yue¹ Zihan Xiao² Ziheng Wang¹ Liang Zhang¹ Dingyi Yang¹
Wenxuan Wang¹³ Qin Jin^{1†}
¹ Renmin University of China
² Beijing University of Posts and Telecommunications
³ Hong Kong University of Science and Technology

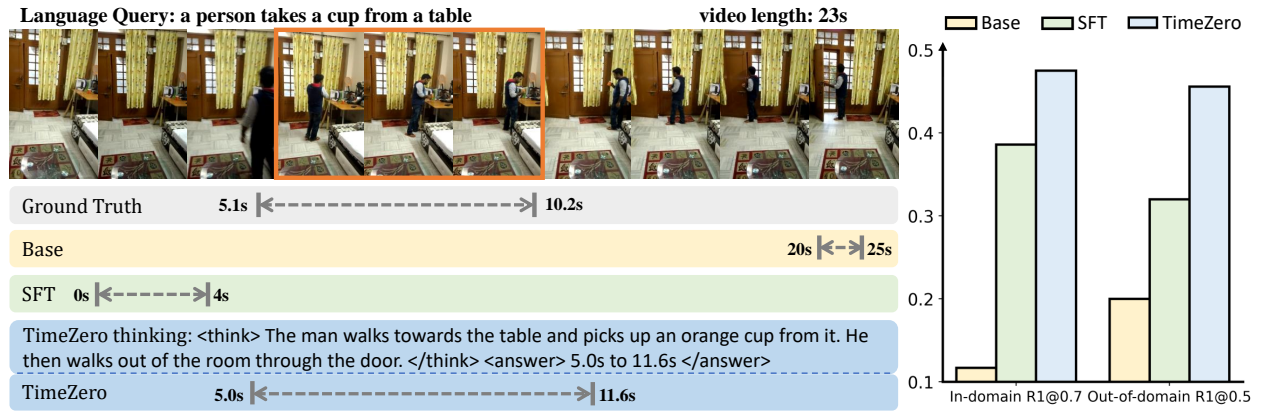


Figure 1. Illustration of TimeZero on temporal sentence grounding task. Given a long video and a textual query, TimeZero identifies the corresponding temporal segment within the video. It generates responses by first thinking before providing the timestamps. TimeZero trained with RL outperforms the model trained with SFT and the base model on both in-domain [31] and out-of-domain [5] tests.

Abstract

We introduce *TimeZero*, a reasoning-guided LVLM designed for the temporal video grounding (TVG) task. This task requires precisely localizing relevant video segments within long videos based on a given language query. *TimeZero* tackles this challenge by extending the inference process, enabling the model to reason about video-language relationships solely through reinforcement learning. To evaluate the effectiveness of *TimeZero*, we conduct experiments on two benchmarks, where *TimeZero* achieves state-of-the-art performance on *Charades-STA*.

1. Introduction

With the rapid development of large vision-language models (LVLM) in video understanding [13, 16, 32, 39], recent efforts [4, 17, 29, 37] have focused on extending their capabilities to understand long videos. One key task in

long video understanding is Temporal Video Grounding (TVG) [3, 10, 14, 19], requiring AI systems to localize visual content corresponding to a textual query within an untrimmed video. For example, a user may prompt the model to retrieve the time interval where “a person carries a cup of water.” Despite being pretrained on massive high-quality data that exceed domain-specific benchmarks [31] by at least 100×, LVLMs [37] with over 7B parameters surprisingly underperform compared to small models with as few as 9M parameters [12].

We attribute this phenomenon to the prevailing training paradigm of LVLMs. LVLMs are typically trained using Supervised Fine-Tuning (SFT) [7, 8, 23, 24], where they learn to regress timestamps directly. However, in TVG tasks, the visual content relevant to a given textual query often constitutes less than 20% of the total video duration [40], meaning that irrelevant visual information overwhelmingly dominates the input. Consequently, effective TVG requires additional reasoning steps to filter out redundant visual content and accurately localize the target video clip. Recently, the chain-of-thought post-training [1, 26] has shown that LLMs can enhance their reasoning abilities

† Qin Jin is the corresponding author.

* Equal Contribution.

by first engaging in a thinking process before generating answers, particularly through pure reinforcement learning (RL). Inspired by these advances, we propose increasing the number of reasoning steps to address the temporal challenges in the TVG task.

In this paper, we introduce TimeZero, a strong reasoning-guided LVLM for video grounding. TimeZero is trained by pure reinforcement learning with group relative policy optimization (GRPO) [1, 30]. Specifically, we design several rule-based rewards to facilitate RL training, containing format-related reward and the Intersection over Union (IoU) reward between predictions and ground truths. Extensive experiments demonstrate that TimeZero surpasses both specialized models and LVLMs on Charades, achieving state-of-the-art performance. Additionally, it also outperforms other LVLMs by a large margin on ActivityNet.

2. Related Works

Temporal Video Grounding. The Temporal Video Grounding (TVG) task [2, 9] requires localizing temporal segments within an untrimmed long video given a language query. Early works generally employed hand-designed mechanisms such as sliding window approaches [2, 9] or cross-modal alignment frameworks [15, 20–22, 38, 41]. For example, Liu *et al.* [22] propose iterative intra-modal and inter-modal fusion modules to align semantics between video and language. Despite improving benchmark performance, these approaches introduce increasingly complex architectures, adversely affecting their reproducibility and generalization. Recently, the LVLMs demonstrate promising results in temporal understanding tasks [17, 29, 37]. For instance, TimeChat introduces a time-aware frame encoder and a sliding video Q-former to compress video tokens, then conducts instruction tuning on time-sensitive tasks such as TVG. However, these models are both computationally intensive and present inferior performance compared with earlier approaches. In this work, we unleash the potential of LVLMs in the TSG task, establishing a new state-of-the-art performance, as shown in Figure 1.

Reasoning in LVLMs. In large language models (LLMs), Chain-of-Thought (CoT) [18, 34, 35, 43] is an important reasoning paradigm where the LLM first thinks before generating a response. Early implementations of CoT primarily rely on prompting or supervised fine-tuning (SFT) to develop this reasoning ability. For example, the classic work [34] is to prompt LLM with “think step by step” instructions. Recently, models such as OpenAI-o1 [26] and DeepSeek-r1 [1] have further advanced LLM reasoning by introducing reinforcement learning to post-train models for “slow thinking”, also termed as “inference-time scaling”. This method extends the length of the reasoning process at inference time, leading to consistent performance improvements across various tasks. Inspired by this paradigm, some

studies have explored its potential in the vision domain. For instance, R1-V [6] demonstrate that LVLMs trained with reinforcement learning exhibit superior generalization on image reasoning tasks. However, it remains unclear whether a similar phenomenon occurs in video understanding. To bridge this gap, we investigate the effectiveness of this paradigm in long video understanding, providing empirical evidence that models trained by RL with inference-time reasoning can enhance video grounding capability.

3. Method

The Temporal Video Grounding (TVG) task aims to temporally localize video segments corresponding to given textual queries in long-form videos. A video is represented as a sequence of T frames $\{x_1, \dots, x_T\}$, the language query is q , and the target segment is defined by its temporal boundaries $[t_s, t_e]$ where $t_s, t_e \in \mathbb{R}^+$.

Next, we introduce TimeZero to unleash the potential of an LVLM model for the TVG task through pure reinforcement learning. We first introduce the background of training with reinforcement learning in LLM in Section 3.1. Next, we describe how we train the TimeZero in Section 3.2.

3.1. Background of GRPO: RL for LLM

As a pioneer in open-sourced LLM excelling in reasoning, DeepSeek-R1 [1] adopts group relative policy optimization (GRPO) [30], a reinforcement learning algorithm, to train LLMs to incentivize reasoning capability at inference time. This method optimizes the policy model π_θ (i.e., the LLM) using a rule-based reward function, particularly suitable for tasks with well-defined answers such as mathematical reasoning. In the GRPO framework, given an input question p , the LLM samples G candidate responses $o = \{o_1, \dots, o_G\}$. A reward function $r(\cdot)$ is designed to compute the responses’ rewards $\{r(o_1), \dots, r(o_G)\}$. GRPO makes the LLM generate answers with a higher value of weighted summed reward $R(o)$, defined by:

$$R(o) = \sum_{i=1}^G \frac{\pi_\theta(o_i)}{\pi_{\theta_{\text{old}}}(o_i)} \cdot \frac{r(o_i) - \text{mean}(\{r(o_i)\}_{i=1}^G)}{\text{std}(\{r(o_i)\}_{i=1}^G)} \quad (1)$$

where $\pi_\theta(o)$ denotes the probability of LLM to generate the response o , and $\pi_{\theta_{\text{old}}}$ represents the parameters of the LLM model at a recently optimized state. The final training objective includes a KL-divergence term [1] $D_{\text{KL}}(\cdot \parallel \cdot)$ to prevent the optimized policy π_θ from deviating far from the original LLM parameters π_{ref} :

$$\max_{\pi_\theta} \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(p)} [R(o) - \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})] \quad (2)$$

where β is a hyperparameter.

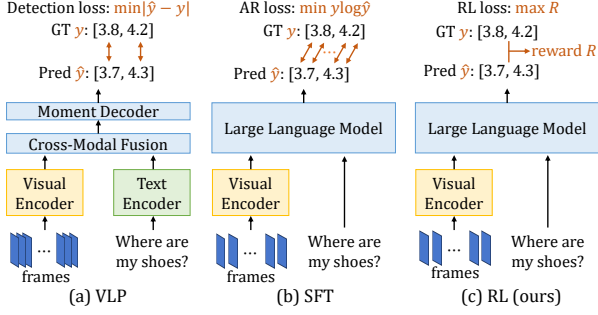


Figure 2. Comparison of different approaches for the TVG task, including Video-Language Pretraining (VLP) [12, 25], Supervised Fine-Tuning (SFT) [27, 37], and RL (ours).

3.2. TimeZero: RL for Temporal Video Grounding

In the video grounding problem, the video segments relevant to the textual query constitute only a small portion of the entire long video. For LVLMs, we argue that the model should not directly output timestamps but rather allocate computational resources to reasoning for fully understanding the video-language relationship before the final prediction. To achieve this goal, we adopt the GRPO-based RL training framework for the LVLM model to enable the model to reason before localization, as described below.

Reward Modeling. The definition of the reward r_i guides the model’s learning objective. To facilitate the VTG task with a reasoning process, we define the reward function based on the reasoning template r_{form} and the Intersection over Union (IoU) r_{IoU} required for the TVG task.

- Template Reward $r_{\text{form}}(\cdot)$: the reasoning template reward requires the LVLM to respond like “<think> ... </think> <answer>< t_s ><to/and>< t_e ></answer>”:

$$r_{\text{form}}(o) = \begin{cases} 0, & \text{if } o \text{ has wrong format} \\ 1, & \text{if } o \text{ has correct format} \end{cases} \quad (3)$$

- IoU Reward $r_{\text{IoU}}(\cdot)$: we require the model to generate output segment $[t_s, t_e]$ with high IoU on ground-truth segment t'_s, t'_e , computed as:

$$r_{\text{IoU}}(o) = \frac{[t_s, t_e] \cap [t'_s, t'_e]}{[t_s, t_e] \cup [t'_s, t'_e]} \quad (4)$$

where $A \cap B$ and $A \cup B$ denote the union and intersection between sets A and B, respectively.

The overall reward function is computed as their sum:

$$r(o) = r_{\text{form}}(o) + r_{\text{IoU}}(o) \quad (5)$$

GRPO Training. The LVLM $\mathcal{F}(\cdot)$ takes the given video and language query as input and outputs G responses $\{o_1, \dots, o_G\}$, where the i -th response is:

$$o_i = \mathcal{F}(x_1, \dots, x_t; q) \quad (6)$$

Table 1. Performance of temporal video grounding task on Charades-STA in fine-tune settings. The models marked with an asterisk (*) indicate results from zero-shot testing.

Method	Type	R1@0.3	R1@0.5	R1@0.7
SSRN [44]	VLP	-	65.5	42.6
SnAG [25]	VLP	-	64.6	46.2
EaTR [12]	VLP	-	68.4	44.9
ChatVTG* [27]	SFT	52.7	33.0	15.9
VideoChat-Flash* [17]	SFT	74.5	53.1	27.6
TimeChat [29]	SFT	-	46.7	23.7
HawkEye [33]	SFT	72.5	58.3	28.8
TRACE [11]	SFT	-	61.7	41.4
VideoChat-T [37]	SFT	79.4	67.1	43.0
TimeZero (ours)	RL	83.3	72.5	47.9

Then, we calculate the summed reward by Equation (1) and optimize the LVLM with the objective of Equation (2). During training, only the LLM parameters are optimized.

4. Experiments

4.1. Experimental Setup

Benchmarks. We evaluate the performance of our model on two widely used public datasets: (1) Charades-STA [31], which contains 6,672 long videos depicting indoor scenes of daily human activities. The official splits include 12,408 clip-query pairs for training and 3,720 for testing. (2) ActivityNet [5], comprising 20K long videos with an average of 3.65 clip-query pairs per video. Following previous work [12, 42], we adopt the standard dataset splits with 37,421 training, 17,505 validation, and 17,031 test samples.

Implementation Details. Our model is fine-tuned on Qwen2.5-VL-7B [4]. To balance training efficiency and memory usage, video frames are sampled at 2 FPS, and each video input is adaptively resized to 2.8 million pixels. Our training uses a batch size of 1 for 1 epochs. All experiments are conducted on 4×A800 GPUs.

Evaluation Metrics. Following [29, 37], we adopt the “R1@m” evaluation strategy, which measures the percentage of cases where the IoU of the top-1 prediction exceeds a given threshold m , where $m \in \{0.3, 0.5, 0.7\}$.

Method Comparison. TimeZero (RL) is compared with the following two types of approaches, including video-language pretraining (VLP) and supervised fine-tuning (SFT), as illustrated in Figure 2.

- VLP: Utilizing pretrained visual and text encoders like CLIP [28] to extract video and text features, these specialized models design cross-modal and moment decoder modules to generate time interval proposals. The timestamps are generated following a traditional video detection paradigm [25, 38, 42].
- SFT: Based on LVLMs with 7B sizes, these models [27, 29, 37] are supervised by an autoregressive loss to gener-

Table 2. Performance of temporal video grounding on ActivityNet.

Method	Type	R1@0.3	R1@0.5	R1@0.7
DRN [36]	VLP	-	45.45	24.36
2D-TAN [42]	VLP	-	46.16	29.21
SSRN [44]	VLP	-	54.49	33.15
SnAG [25]	VLP	-	48.55	30.56
EaTR [12]	VLP	-	58.18	37.64
HawkEye [33]	SFT	55.9	34.7	17.9
TRACE [11]	SFT	54.0	37.7	24.0
TimeZero (ours)	RL	68.6	47.3	26.9

ate the number of timestamps. These models typically use a large amount of video grounding data for pretraining.

4.2. Comparison with State of the Art

TimeZero outperforms VLP-based models on Charades.

TimeZero is the first LVLM model to surpass previous VLP-based models in TVG tasks. For instance, on Charades-STA in Table 1, TimeZero achieves an R1@0.7 score of 47.9, outperforming EaTR [12]’s 44.92 by +2.98. Additionally, at R1@0.5, our model exceeds EaTR by +4.1. Note that our model achieves superior performance despite using fewer pixels as input compared to VLP-based models. While the VLP typically processes video inputs with 64 frames of 224×224 resolution, totaling 3.2M pixels, TimeZero operates with only 2.8M pixels.

TimeZero is competitive with VLP-based models on ActivityNet. Table 2 shows that TimeZero outperforms some classic methods, such as surpassing DRN [36] by +2.54 in R1@0.7. Previous LVLM methods all lag behind classic methods. However, it falls behind approaches like EaTR. One reason is that the longer video durations in ActivityNet result in larger compression rates in TimeZero. While classic methods use a video input of 10M pixels, i.e., 200 frames with 224×224 resolution, ours is limited to 2.8M. In the future, we will align our input processing with traditional methods to achieve better performance.

TimeZero surpasses the SFT-based LVLMs. On the Charades, as shown in Table 1, TimeZero’s R1@0.7 is +4.9 percentage points higher than VideoChat-T [37], and its R1@0.3 improved by +3.9 percentage points. VideoChat-T collected 349K high-quality, temporally-related data samples for SFT. On the ActivityNet, TimeZero also outperforms the current state-of-the-art large video grounding model TRACE [11], improving the R1@0.7 by +2.9 and the R1@0.3 by +14.6 percentage points. This demonstrates that rule-based RL can effectively unlock the strong capabilities of the foundation VLM [4] for the TVG task.

4.3. Quantitative Analyses

Stronger in-domain performance of RL-based models.

Among different methods in Table 3, RL yields the highest

Table 3. Impact of incorporating CoT during test time across different training paradigms on Charades-STA. The “RL w/o CoT” denotes training models without thinking templates. We implement the test-time CoT as prompting models to think with the “<think>...</think>” template.

Method	Test CoT	R1@0.3	R1@0.5	R1@0.7
Base	✗	43.9	25.7	11.7
Base	✓	34.1	19.1	8.4
SFT	✗	72.3	60.3	38.6
SFT	✓	72.7	60.8	38.0
RL w/o CoT	✗	81.2	70.5	47.1
RL w/o CoT	✓	75.3	62.4	38.7
RL (ours)	✗	82.1	70.6	47.5
RL (ours)	✓	83.3	72.5	47.9

Table 4. Out-of-domain generalization test, where models are trained on Charades-STA and tested on ActivityNet.

Method	R1@0.3	R1@0.5	R1@0.7
Base	20.0	11.4	5.9
SFT	32.0	15.3	6.3
RL w/o CoT	42.2	25.8	14.3
RL (ours)	45.6	26.5	16.5

performance gain compared with base models, surpassing SFT by +9.9 in R1@0.7.

Superior OOD generalization of RL. As shown in Table 4, we train the model on Charades and test it on ActivityNet. RL demonstrates remarkable generalization capability, while SFT shows marginal improvement.

RL training with CoT provides further improvements.

Comparing the results of “RL” and “RL w/o CoT” in Tables 3 and 4, we observe that training the model with CoT consistently outperforms training without it. This suggests that reasoning before answering during training not only enhances the model’s in-domain performance but also improves its generalization ability. The model performs better even without CoT during inference.

RL benefits from reasoning at test time. As shown in the “Test CoT” column of Table 3, for models where RL training has enabled spontaneous CoT, using CoT during testing leads to further performance improvements. However, other models struggle to improve or even show performance degradation when using CoT.

5. Conclusion

In this paper, we introduce TimeZero, a chain-of-thought reasoning-guided LVLM trained for temporal video grounding by pure reinforcement learning. Experiments demonstrate that TimeZero achieves state-of-the-art performance on Charades, surpassing specialized models for the first time. Additionally, our analyses indicate that RL with CoT training is the key to generalizing TVG tasks.

References

- [1] Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1, 2
- [2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 2
- [3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 1
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 3, 4
- [5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–970, 2015. 1, 3
- [6] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-V>, 2025. 2
- [7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 1
- [8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1
- [9] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5267–5275, 2017. 2
- [10] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 1
- [11] Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024. 3, 4
- [12] Jinhyun Jang, Jungin Park, Jin Kim, Hyeonjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13846–13856, 2023. 1, 3, 4
- [13] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 1
- [14] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1
- [15] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 2
- [16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 1
- [17] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhang Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 1, 2, 3
- [18] 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 *The Twelfth International Conference on Learning Representations*, 2023. 2
- [19] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2794–2804, 2023. 1
- [20] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020. 2
- [21] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [22] Daizong Liu, Xiaoye Qu, and Pan Zhou. Progressively guide to attend: An iterative alignment framework for temporal sentence grounding. *arXiv preprint arXiv:2109.06400*, 2021. 2
- [23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1
- [24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Pro-*

- ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1
- [25] Fangzhou Mu, Sicheng Mo, and Yin Li. Snag: Scalable and accurate video grounding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18930–18940, 2024. 3, 4
- [26] OpenAI. Openai o1, 2024. 1, 2
- [27] Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. Chatvtg: Video temporal grounding via chat with video dialogue large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1847–1856, 2024. 3
- [28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3
- [29] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 1, 2, 3
- [30] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2
- [31] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1, 3
- [32] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023. 1
- [33] Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text llms for grounding text in videos, 2024. 3, 4
- [34] 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. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2
- [35] 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. *Advances in neural information processing systems*, 36:11809–11822, 2023. 2
- [36] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10287–10296, 2020. 4
- [37] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. Timesuite: Improving MLLMs for long video understanding via grounded tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 3, 4
- [38] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, 2020. 2, 3
- [39] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1
- [40] Qi Zhang, Sipeng Zheng, and Qin Jin. No-frills temporal video grounding: Multi-scale neighboring attention and zoom-in boundary detection. *arXiv preprint arXiv:2307.10567*, 2023. 1
- [41] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 2
- [42] Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo. Multi-scale 2d temporal adjacency networks for moment localization with natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3, 4
- [43] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. 2
- [44] Jiahao Zhu, Daizong Liu, Pan Zhou, Xing Di, Yu Cheng, Song Yang, Wenzheng Xu, Zichuan Xu, Yao Wan, Lichao Sun, and Zeyu Xiong. Rethinking the video sampling and reasoning strategies for temporal sentence grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022. 3, 4