北京大学
PEKING UNIVERSITY

ICML
International Conference
On Machine Learning

Temporal-Aware Reasoning Optimization for Video Temporal Grounding

Minghang Zheng [1] **Zihao Yin** [2] **Yi Yang** [2] **Yuxin Peng** [1] **Yang Liu** [1 3]

[1]Wangxuan Institute of Computer Technology, Peking University
[2]Central Media Technology Institute, Huawei Technologies Ltd
[3]State Key Laboratory of General Artificial Intelligence, Peking University

Introduction

北京大学 PEKING UNIVERSITY

Task: Video Temporal Grounding

- **Inputs:** Video + Sentence query
- **Outputs:** Target video clip (start and end timestamps)

Challenges: Temporal-Sensitive reasoning

Recent Advances: Post-train MLLMs with **RL** to reason over videos and complex queries

Query: The guy plays the saxophone **again**.

Outputs: from 55s to 93s.

Introduction

北京大学 PEKING UNIVERSITY

Limitations of Existing RL-based Methods:

- Generate **superficial reasoning** (e.g. generally describe the video)
- The superficial reasoning **contributes little** to the final grounding

Reasons:

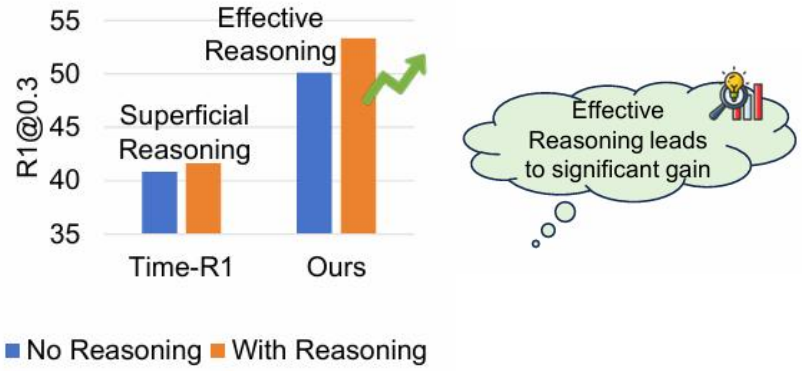
- **Random rollout** during RL blindly explores the vast reasoning space **without guidance**
- **Reward** focus on the final answer, ignoring the **quality of the reasoning process**

Effective Reasoning (Ours)

From 0s to 12s, he plays the saxophone for the first time
From **55s** to **93s**, he plays the saxophone again

Superficial Reasoning

The video shows a guy plays the saxophone



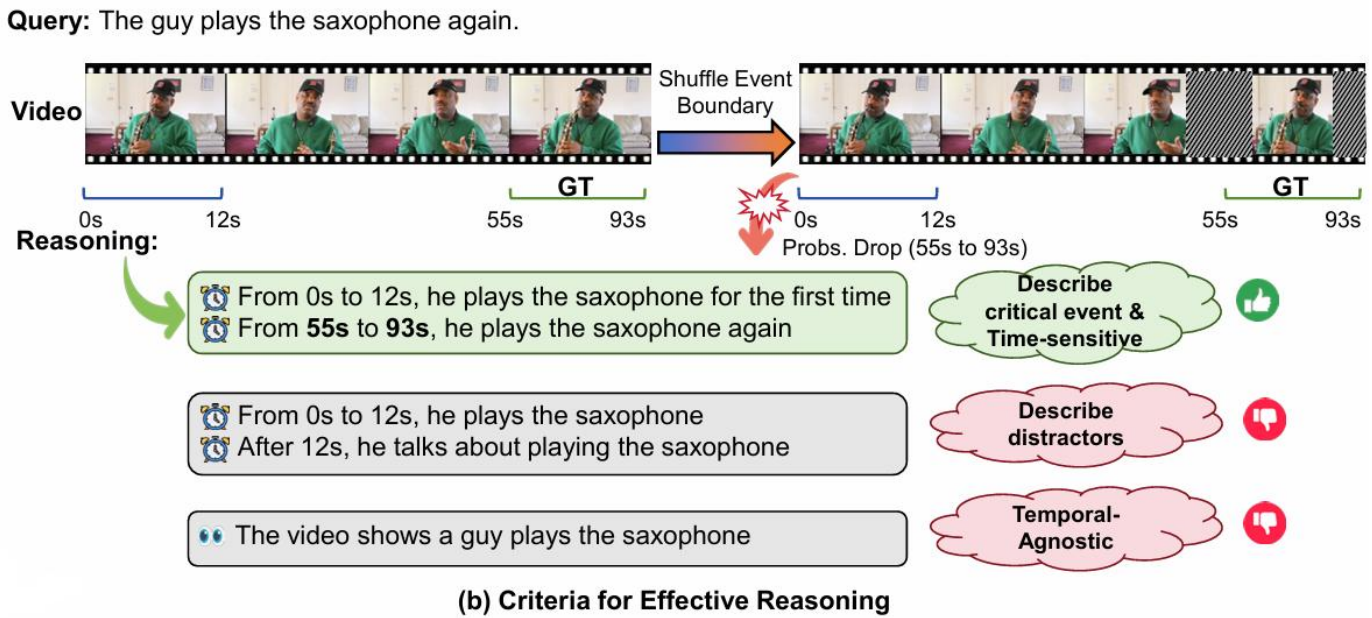
Introduction

Effective Reasoning: selectively attend to **critical visual cues** and be **temporally sensitive**, anchoring these cues to **specific timestamps**

How to assess the quality of reasoning?

Good reasoning (attend to critical visual cues) — shuffle frames near GT → **Large confidence drop** of the reasoning

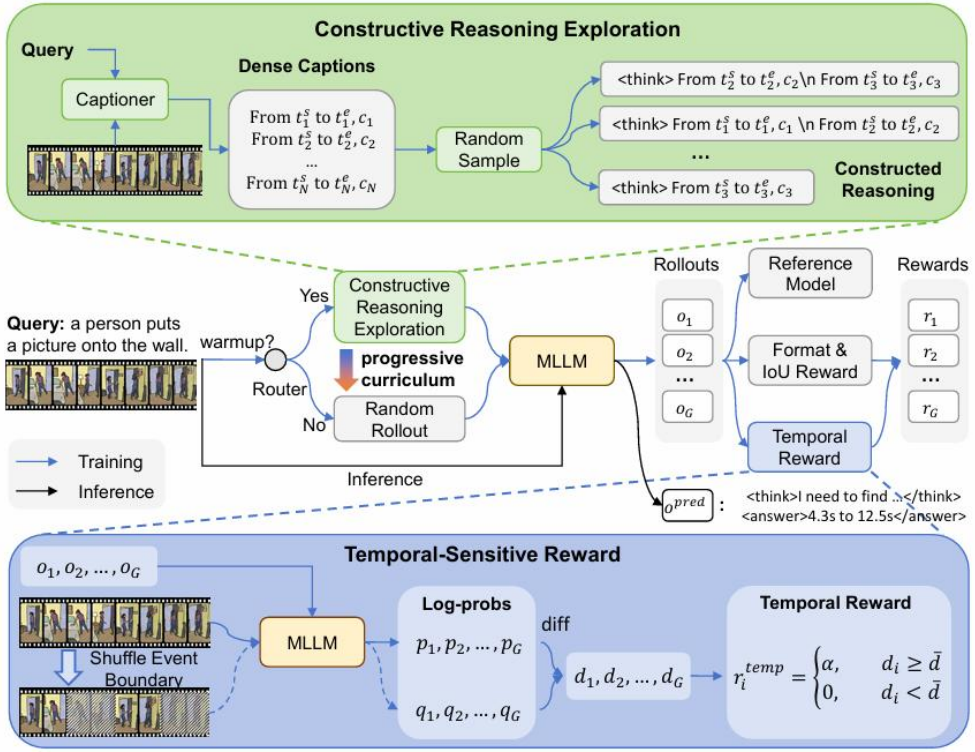
Bad reasoning (not attend to critical visual cues) → **Small confidence drop** of the reasoning



Introduction

Temporal-Aware Reasoning Optimization

- **Constructive Reasoning Exploration:** teach model to **identify critical visual cues** and adopt the **thinking with time** paradigm
- **Temporal-Sensitivity Reward:** assess **reasoning quality** by the **confidence drop** after shuffling near GT
- **Progressive Curriculum:** first guided by the **constructive reasoning**, then transition to **self-exploration**



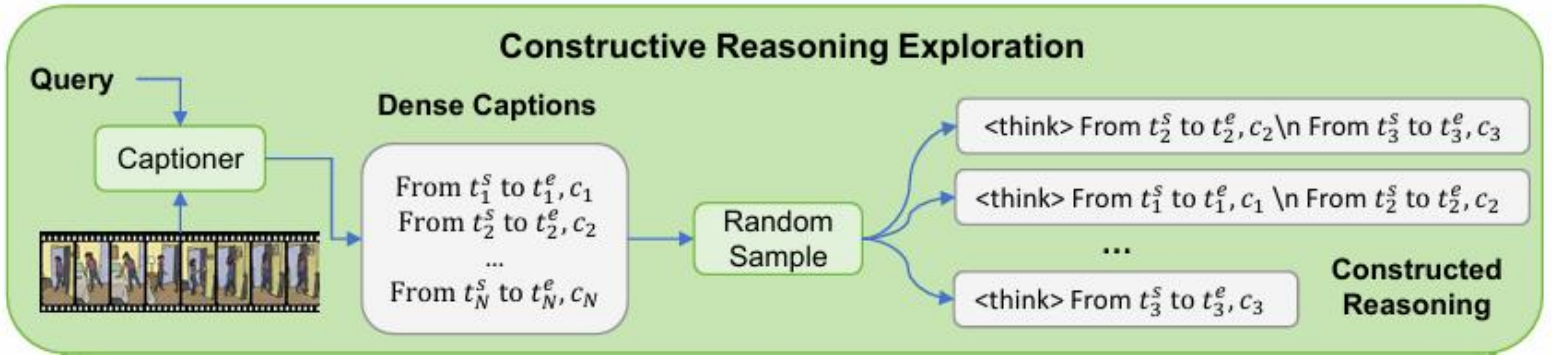
Method

Constructive Reasoning Exploration

Objective: provide high-quality reasoning initialization

- Generate dense captions (**potential visual cues** in the video)
- Random sample dense captions as reasoning (**selectively** attend to different visual cues)
- Learn from reasoning with high rewards (teach model to **identify critical visual cues**)

$$\mathcal{L}_{AW-BC} = -\frac{1}{G}\sum_{i=1}^{G}\mathbb{I}(A_i > 0)\cdot A_i \cdot \log \pi_\theta(o_i|V,Q).$$



Method

北京大学 PEKING UNIVERSITY

Temporal-Sensitivity Reward

Objective: assess the quality of reasoning by the **confidence drop** after shuffling near GT

- Compute the average log-probability of the reasoning tokens on the **original video**

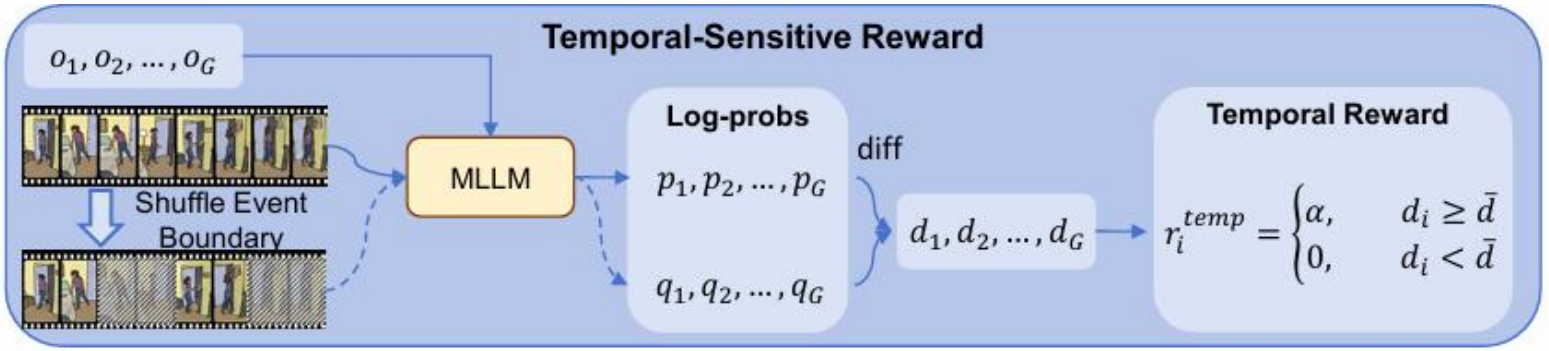
$$p_i = \frac{1}{|r_i|} \sum_{k=1}^{|r_i|} \log \pi(r_{i,k} | V, Q, r_{i,<k}),$$

- Randomly shuffle frames near GT, and recompute the log-probability on **shuffled video**

$$q_i = \frac{1}{|r_i|} \sum_{k=1}^{|r_i|} \log \pi(r_{i,k} | V', Q, r_{i,<k}).$$

- Use the **drop in confidence** as the reward

$$d_i = p_i - q_i. \quad r_i^{\text{temp}} = \begin{cases} \alpha, & \text{if } d_i > \bar{d} \\ 0, & \text{otherwise} \end{cases}$$

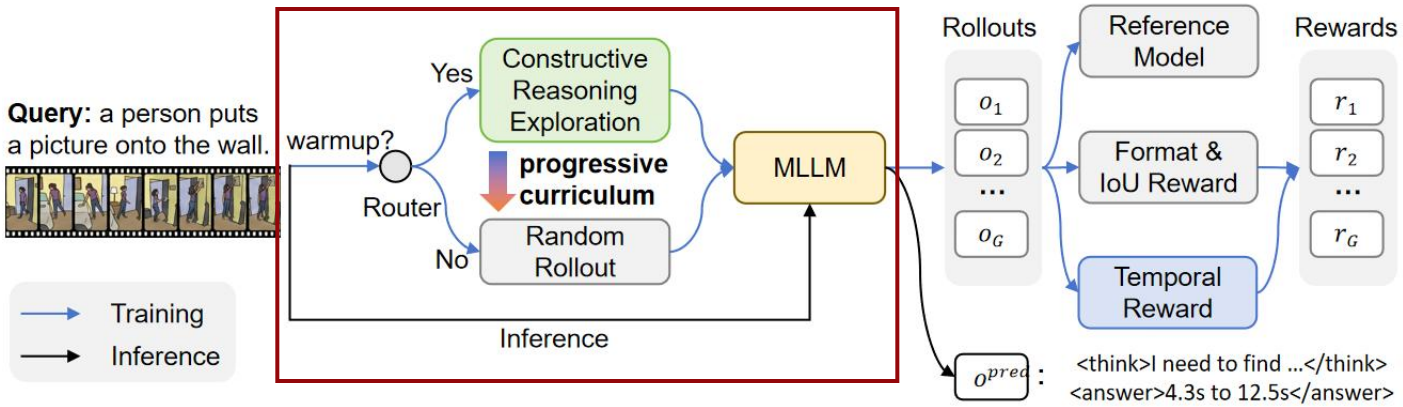


Method

Progressive Curriculum

Objective: bridge the gap between using constructed reasoning and autonomously generating robust reasoning

- **Warm-up with Constructive Reasoning:** teach the model which visual cues to select and how to ground them temporally
- **Self-Exploration:** transition to generates its own reasoning without external constructions



Experiments

- **Trainin Data**
  - Time-R1 Dataset, 2500 samples
- **Datasets**
  - 4 public datasets: ActivityNet Captions, Charades-STA, QVHighlights, TVGBench
- **Metrics**
  - R1@m (m=0.3, 0.5, 0.7)

Query: A man in a red tank top is crossing the monkey bars

Query: Person runs to a table

Experiments

北京大学 PEKING UNIVERSITY

- SOTA performance across **4 VTG datasets**
- Consistent improvements over **different base models**

Method	Size	Charades-STA (Gao et al., 2017)			ActivityNet (Krishna et al., 2017)			QVHighlights (Lei et al., 2021)			TVGBench (Wang et al., 2025)		
		R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7
ChatVTG (Qu et al., 2024)	7B	52.7	33.0	15.9	40.7	22.5	9.4	-	-	-	-	-	-
TimeChat (Ren et al., 2024)	7B	-	32.2	13.4	36.2	20.2	9.5	-	8.32	4.26	22.4	11.9	5.3
HawkEye (Wang et al., 2024)	7B	50.6	31.4	14.5	49.1	29.3	10.7	-	-	-	-	-	-
VTimeLLM (Huang et al., 2024)	7B	51.0	27.5	11.4	44.0	27.8	14.3	-	26.1	11.1	-	-	-
TimeSuite (Zeng et al., 2025a)	7B	69.9	48.7	24.0	-	16.6	9.28	-	12.3	9.16	31.1	18.0	8.9
VideoChat-Flash (Li et al., 2024)	7B	74.5	53.1	27.6	-	-	-	-	-	-	32.8	19.8	10.4
TRACE (Guo et al., 2025)	7B	-	40.3	19.4	-	-	-	-	-	-	37.0	25.5	14.6
Qwen2.5-VL-7B-Instruct as base model													
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	7B	72.5	53.6	28.5	24.4	13.6	6.7	15.93	7.10	4.19	35.3	20.0	12.5
UniTime (Li et al., 2025b)	7B	-	59.1	31.9	-	22.8	14.1	-	41.0	31.5	-	-	-
VideoChat-R1.5 (Yan et al., 2025)	7B	-	-	-	52.4	32.3	16.8	71.4	55.8	38.4	-	-	-
Time-R1 (Wang et al., 2025)	7B	78.1	60.8	35.3	58.6	39.0	21.4	80.3	66.2	44.8	41.8	29.4	16.4
TaRO (Ours)	7B	**79.7**	**64.8**	**38.4**	**60.6**	**39.8**	**21.4**	**82.6**	**69.4**	**48.8**	**54.6**	**37.8**	**20.0**
Qwen2.5-VL-3B-Instruct as base model													
Qwen2.5-VL-3B-Instruct (Bai et al., 2025)	3B	62.1	42.0	20.1	26.5	15.2	7.2	16.8	9.9	3.1	26.1	17.6	10.3
Time-R1 (Wang et al., 2025)	3B	74.6	53.1	26.0	46.2	26.0	11.4	40.8	19.7	5.9	40.1	24.4	11.8
TaRO (Ours)	3B	**75.1**	**55.2**	**28.6**	**52.3**	**32.7**	**14.2**	**65.6**	**43.1**	**20.8**	**44.3**	**28.0**	**13.4**
Qwen3-VL-8B-Instruct as base model													
Qwen3-VL-8B-Instruct (Yang et al., 2025a)	8B	72.7	49.8	21.3	45.9	31.7	19.8	57.2	45.9	34.5	37.3	26.5	15.0
Qwen3-VL-8B-Think (Yang et al., 2025a)	8B	72.7	57.9	32.6	44.0	30.5	19.2	69.4	56.3	42.7	38.1	25.9	13.7
TaRO (Ours)	8B	**83.2**	**67.8**	**41.9**	**59.7**	**40.6**	**23.9**	**82.7**	**68.5**	**51.7**	**54.3**	**39.1**	**20.3**

Ablation Study

北京大学 PEKING UNIVERSITY

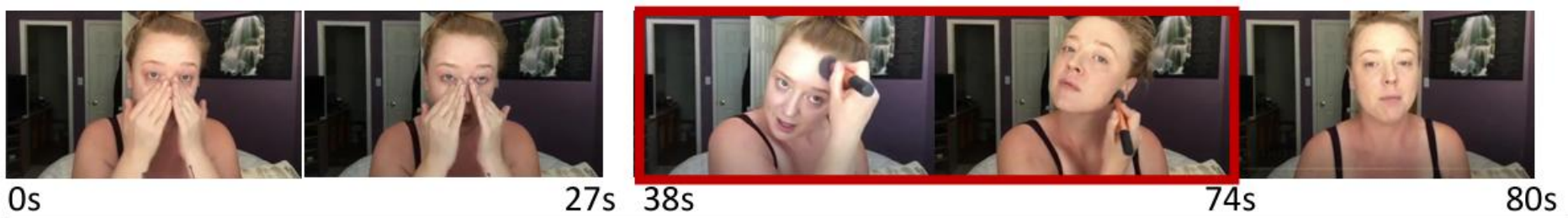
- **Temporal Reward (TR) alone** brings aclear improvement
- **Constructive Reasoning (CRE) alone** leads to a drastic performance drop, as constructive reasoning cannot be obtained during testing
- Introducing **CRE** with **Progressive Curriculum (PC)** further improves the performance

CRE	TR	PC	Charades-STA		
			R1@0.3	R1@0.5	R1@0.7
×	×	×	78.2	61.1	35.2
×	✓	×	78.6	63.1	36.1
✓	×	×	70.4	51.6	25.5
✓	✓	×	71.1	53.0	26.3
✓	×	✓	78.9	63.9	36.3
✓	✓	✓	**79.7**	**64.8**	**38.4**

Qualitative Results

北京大学 PEKING UNIVERSITY

- Our method generates reasoning **anchored to specific timestamps and visual details**



Query: Woman wipes her face with a brush.

GT: 38s — 74s

Time-R1: 0s — 27s

Thinking: The event 'Woman wipes her face with a brush.' occurs towards the beginning of the video.

Ours: 39s — 76s

Thinking: I need to find the specific moment Woman wipes her face with a brush.

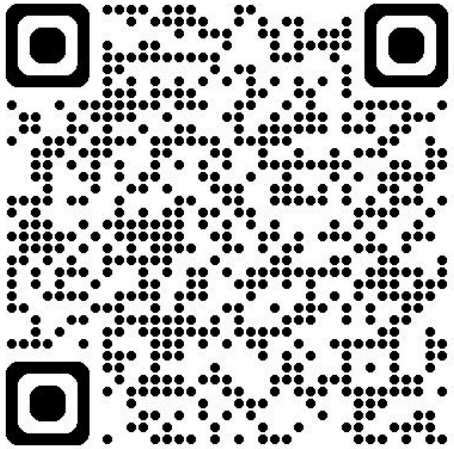
From 39.0s to 76.0s, A woman applies makeup to her face using a brush.

From 76.0s to 80.0s, The woman stops applying makeup and looks at the camera.

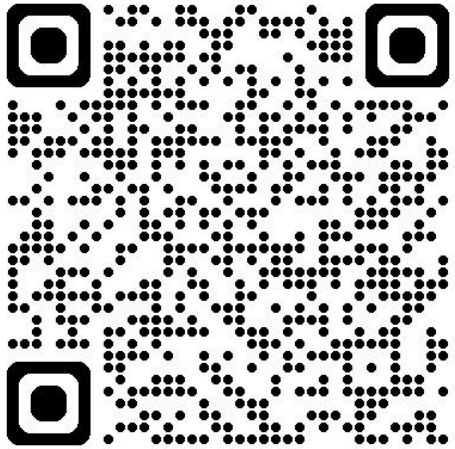
北京大学
PEKING UNIVERSITY

ICML
International Conference
On Machine Learning

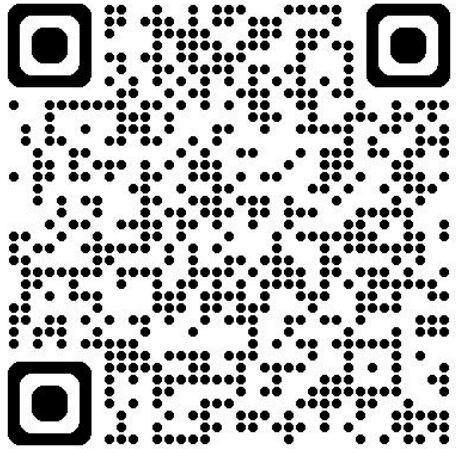
Thank you!



Paper



Code



Project Page