Time-R1: Post-Training Large Vision Language Model for Temporal Video Grounding

Ye Wang[1*] Ziheng Wang[1*] Boshen Xu[1*‡] Yang Du[1] Kejun Lin[1] Zihan Xiao[1]
Zihao Yue[1] Jianzhong Ju[2] Liang Zhang[1] Dingyi Yang[1] Xiangnan Fang[1] Zewen He[2]
Zhenbo Luo[2] Wenxuan Wang[1] Junqi Lin[2] Jian Luan[2] Qin Jin[1†]

[1]AIM3 Lab, Renmin University of China [2]MiLM Plus, Xiaomi Inc

AI·M³
www.ruc-aim3.com

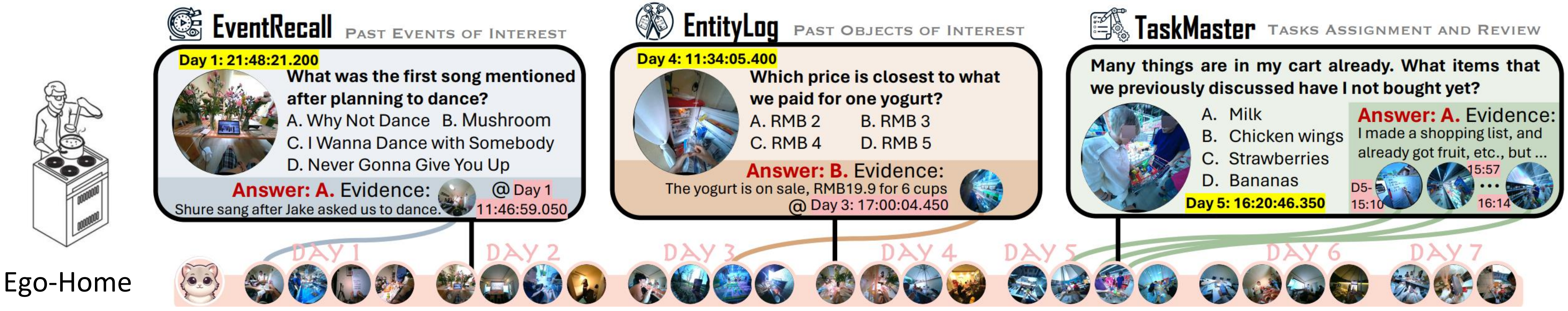
中国人民大學
RENMIN UNIVERSITY OF CHINA

NeurIPS

Ambitions for Pursuing Long Video Understanding

- Explosion of long videos—from minutes to hours, days, and even months

- Applications: smart home devices, video retrieval systems on platforms



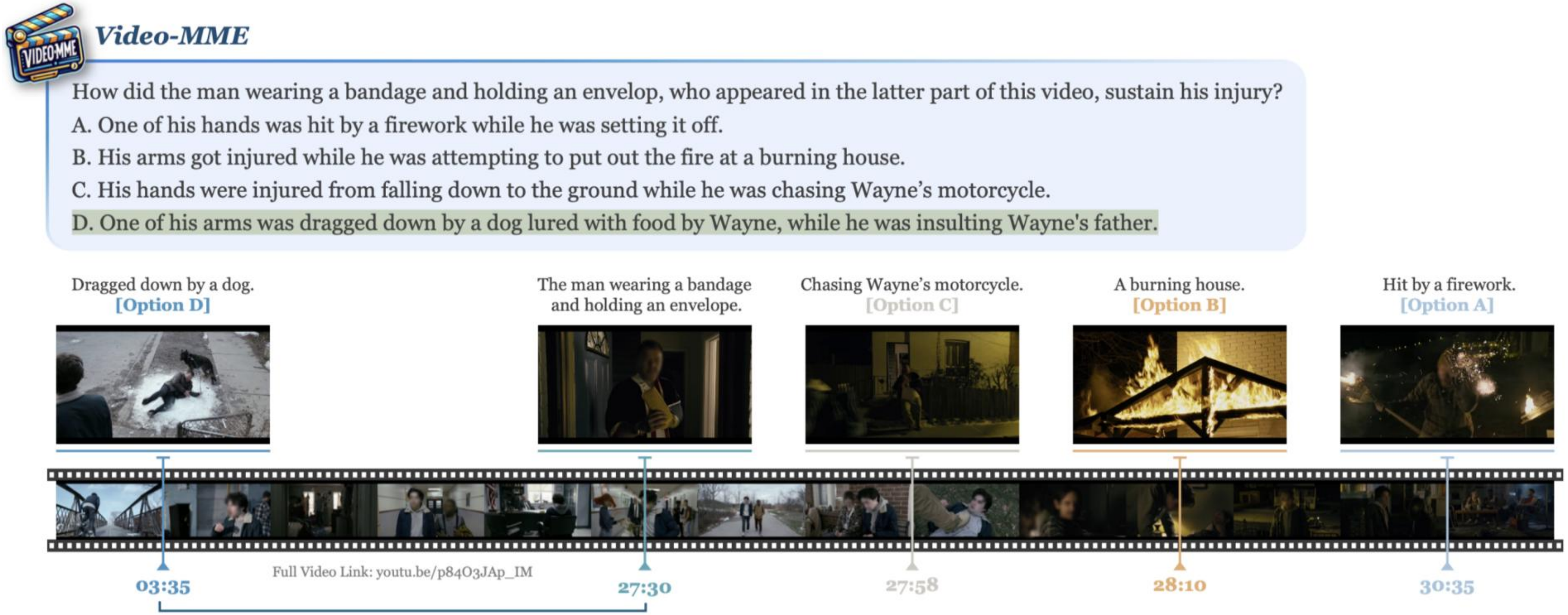
Fundamental Task: Temporal Video Grounding (TVG)

- Input long videos V and language query Q, localize corresponding segment $[t_s, t_e]$.

Question: Many things are in my cart already. What items that we previously discussed have I not bought yet?

Answer: **Clue: Day 5: 15:10:00-16:20:46**

- TVG is a fundamental temporal retrieval task for solving complex temporal reasoning



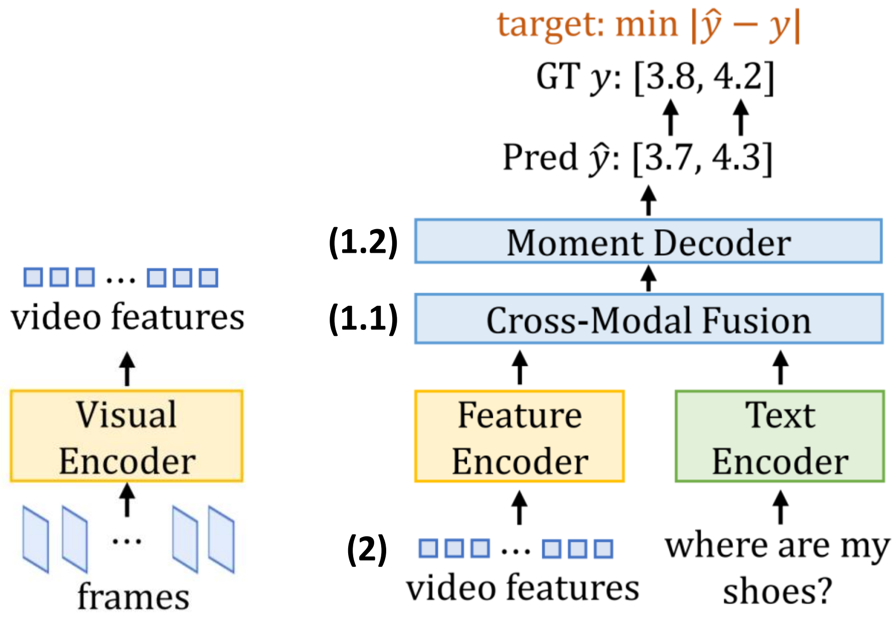
Core Challenges in the TVG Task

① Understanding query-event, event-timestamp correspondence

(1.1) Query-event: video-language alignment

(1.2) Event-timestamp: proper decoding/regression strategy (e.g., proposal-based, proposal-free, SFT, RL)

② Receiving longer video as input, i.e., hundreds of frames as input.



(a) VLP

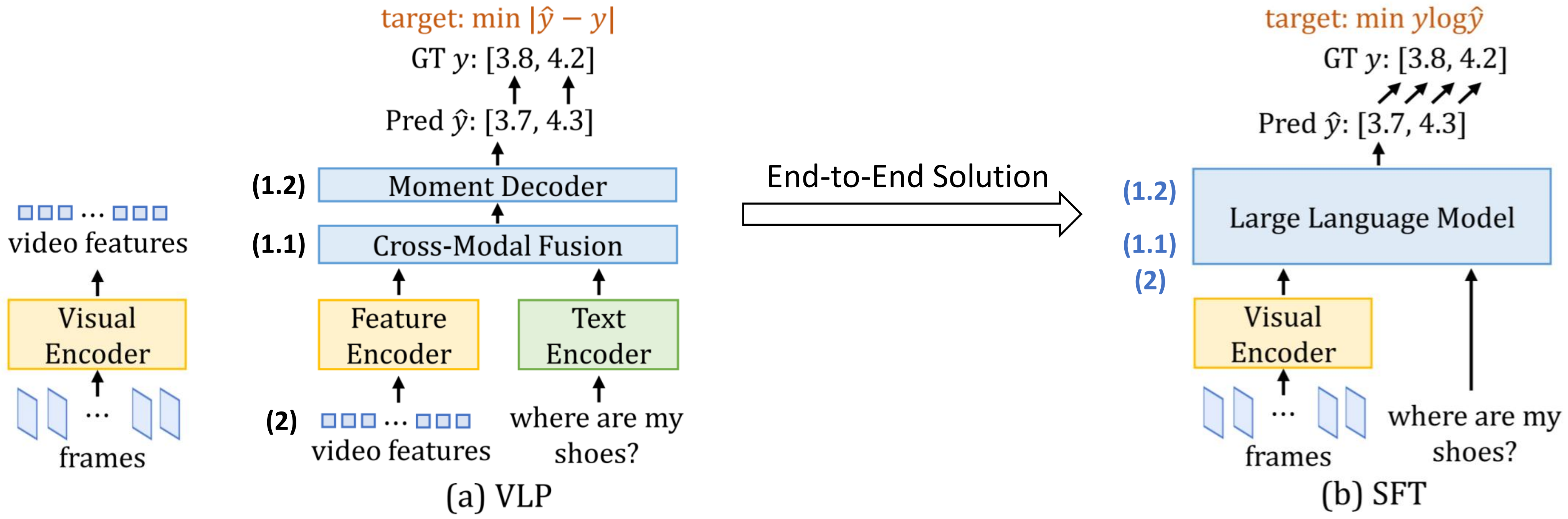
Paradigm Shift to End-to-End LVLM Solution

① Understanding query-event, event-timestamp correspondence

(1.1) Query-event: video-language alignment

(1.2) Event-timestamp: proper decoding/regression strategy (e.g., proposal-based, proposal-free, SFT, RL)

② Receiving longer video as input, i.e., hundreds of frames as input.



Paradigm Shift to End-to-End LVLM Solution

① Understanding query-event, event-timestamp correspondence

(1.1) Query-event: video-language alignment

However, the LVLM solution consistently underperforms VLP methods even on the simplest benchmark, despite LVLM being pretrained on 10(0)× more data and equipped with significantly more parameters (7B vs. 9M),

For example:

TimeSuite-7B [ICLR'25]:| Millions of IT data + 349K TVG | Charades-R1@0.7: 24.0 (ZS) / 43.0 (FT)

EaTR-9M [CVPR'23]: | 12K TVG | Charades-R1@0.7: 44.9 (FT)

(a) VLP

(b) SFT

Reasons that LVLM Falls Behind VLP

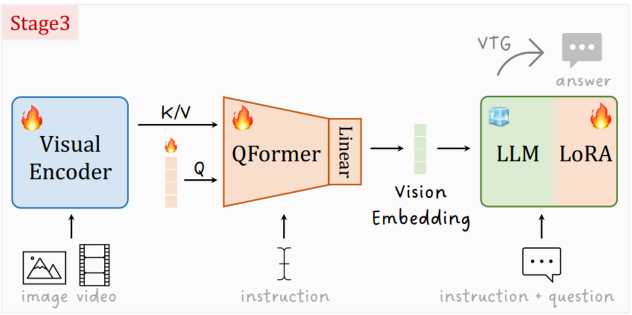
① The over-penalization of false negatives by SFT

Suppose a reasonable prediction: [1.9s, 3.9s], GT: [2s, 4s], tokenization:

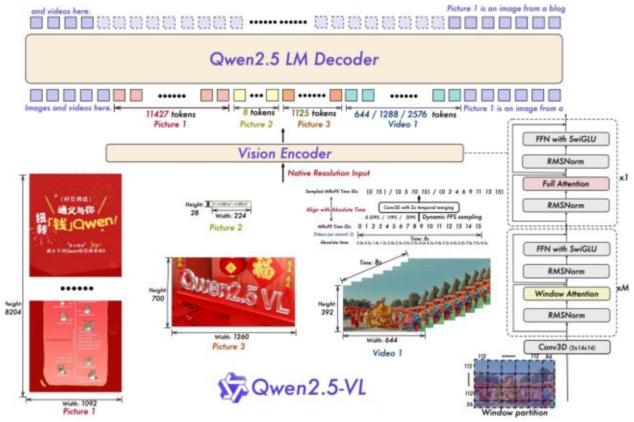
tokenize([1.9s, 3.9s]) = [58, 16, 13, 24, 82, 11, 220, 18, 13, 24, 82, 60]
tokenize([2s, 4s]) = [58, 17, 82, 11, 220, 19, 82, 60]

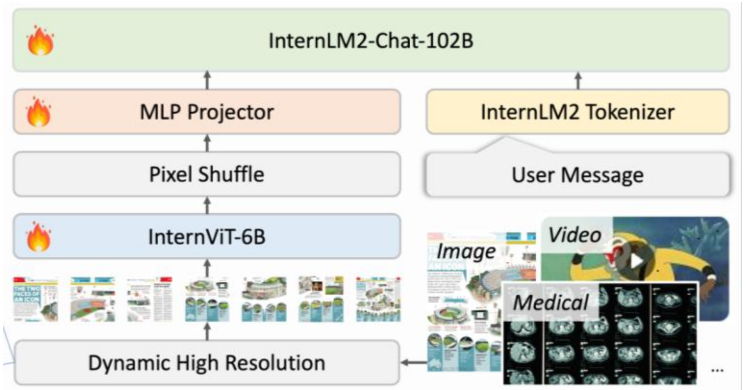
Easy to overfitting and lead to poor generalization

② LLM learns both query-event and event-timestamp correspondence, requiring more data for video-text alignment and timestamp prediction, i.e., a stronger base model



✗ Videochat, LLaVA-Video, …

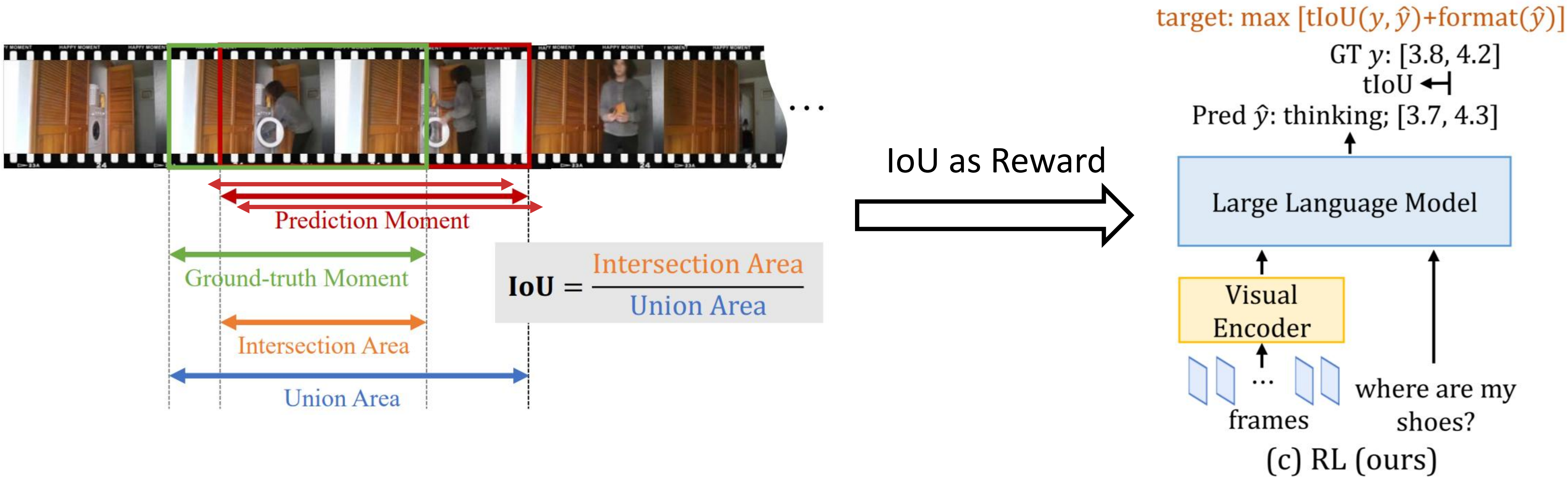




✓ VideoChat-Flash, Qwen2.5-VL, Qwen2-VL, InternVL2, InternVL3, …

Introduce RL and Stronger Base Model for the TVG Task

- Reinforcement learning with verifiable reward relieves the problem ① posed by SFT
  - metric-oriented (IoU)
  - fault tolerant
  - Joint event-timestamp learning

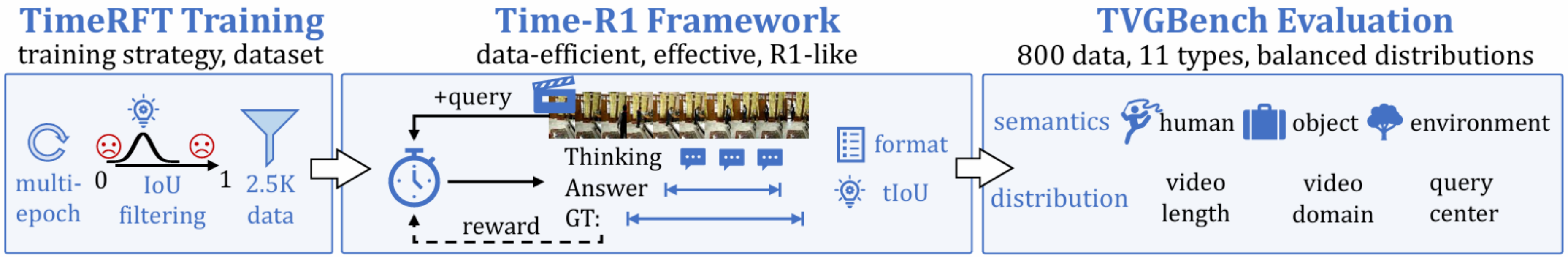


- ② Initialize with a stronger Base Model, e.g., Qwen2.5-VL

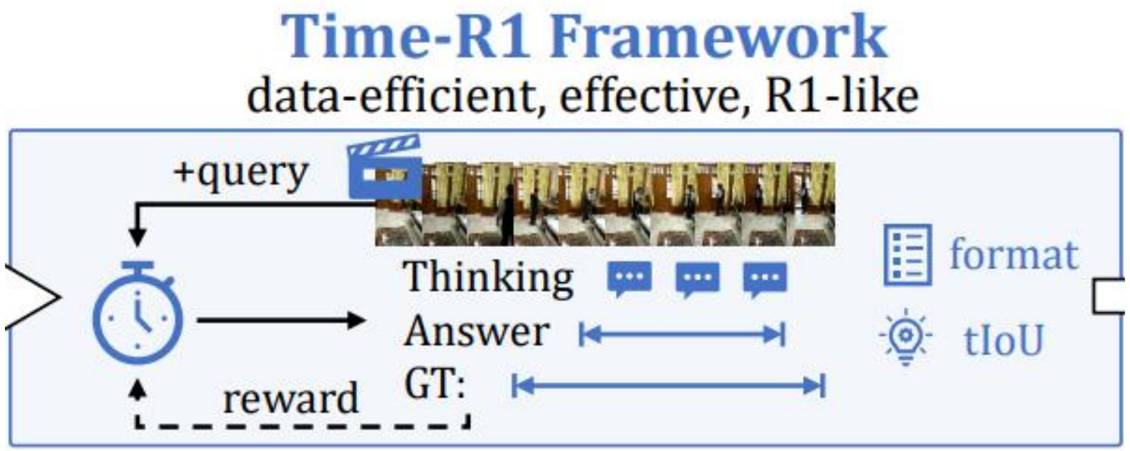
TL; DR

We systematically explore RLVR for TVG with LVLM along these direction:

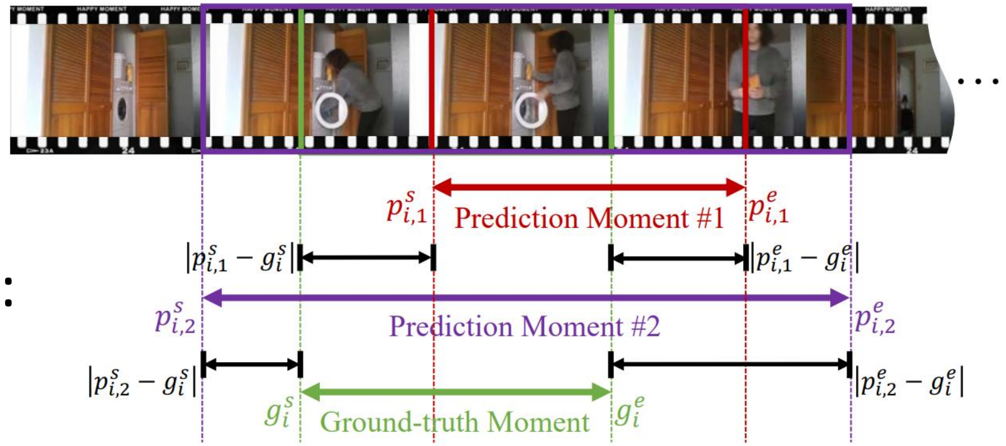
- **Time-R1 framework**: a reasoning-guided post-training framework via RL with verifiable reward to enhance the capabilities of LVLMs on the TVG task.
- **TimeRFT training**: we explore data-efficient post-training strategies on our curated RL-friendly dataset.
- **TVGBench evaluation**: we carefully construct a small yet comprehensive benchmark for LVLM evaluation.
- **SOTA Performance**: Time-R1 achieves state-of-the-art performance across multiple downstream datasets using only 2.5K training data, while improving its general video understanding capabilities.



Time-R1: Training LVLM with RLVR for the TVG Task



tIoU:



- Reward function: $r(o) = r_{\text{tIoU}}(o) + r_{\text{form}}(o)$
  - Timestamp-aware IoU (tIoU):

$$r_{\text{tIoU}}(o) = \text{IoU} \cdot \left(1 - \frac{|t_s - t'_s|}{t}\right) \cdot \left(1 - \frac{|t_e - t'_e|}{t}\right)$$

  - Thinking template: “<think>$\cdots$</think> <answer><t_s to t_e></answer>”,

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

- GRPO Loss: $\max_{\pi_\theta} \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(p)} [R(o) - \beta \mathrm{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})]$

TimeRFT: Time-Aware RL-Friendly Fine-Tuning

- **Data Filtering:** To learn relatively hard samples

➢ **Source Data Collection:**

  ➢ Video source: public YT-Temporal, DiDeMo, QuerYD, InternVid, HowTo100M

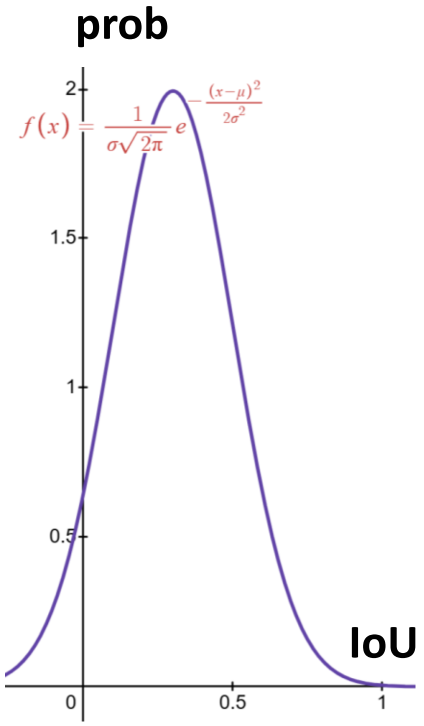
  ➢ TVG annotation: VTG-IT, TimeIT, TimePro, HTStep, LongVid, 339K data.

➢ **Collection Steps:**

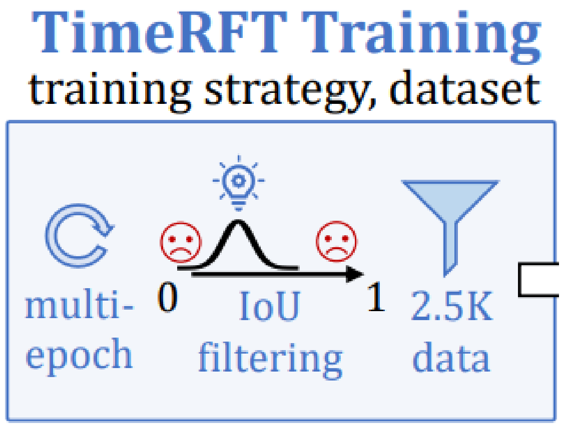
1. **Scoring by base model** Qwen2.5-VL: predicts the IoU for all samples, which serves as a "difficulty score."

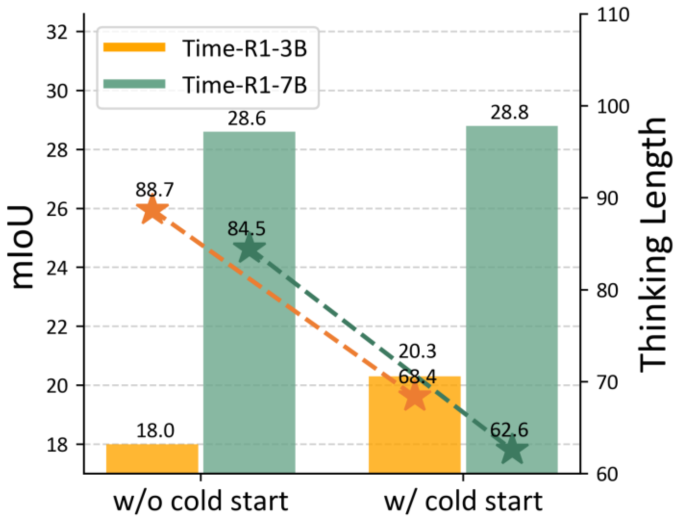
2. **Gaussian distribution sampling**: samples are drawn using a Gaussian distribution centered at 0.3 (mean)-0.2(var).

➢ **Result:** A total of 2.5K sample



TimeRFT: Time-Aware RL-Friendly Fine-Tuning





- **Training Strategy**: data-efficient, RL-Friendly

➢ **Dynamic Hard Sampling:** Multi-epoch Training + Per-epoch Sample Filtering

- For each epoch:

1. **Evaluate Difficulty** of current training set.
2. **IoU Filter Easy Samples**: Remove “easy" samples (e.g., IoU > 0.7) from the training set for the next epoch.

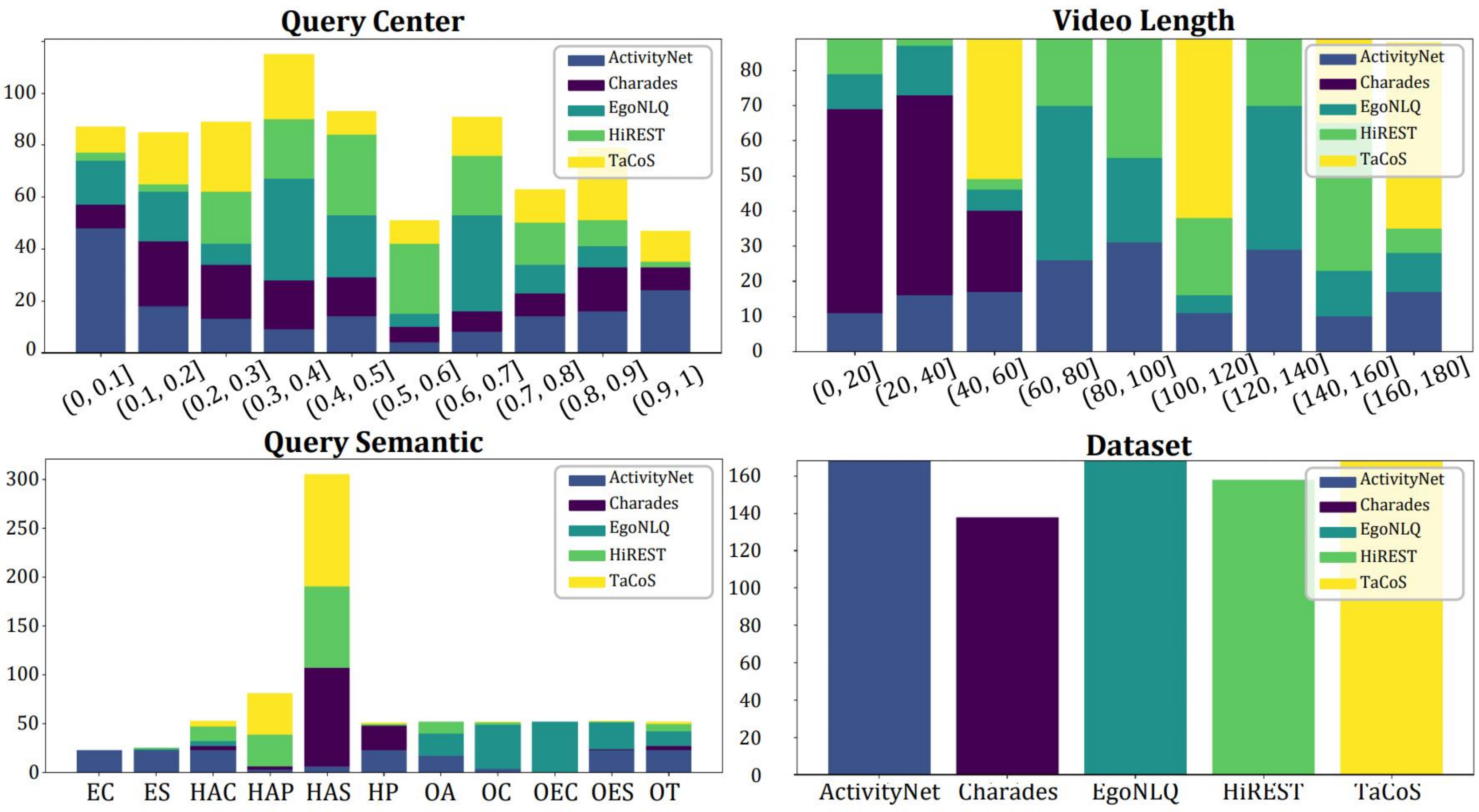
➢ **Cold Start with Few CoT Data:** Reduce hallucinations, control thinking length

$$< \text{think} >< t_{s_1} \text{ to } t_{e_1} : C_1;\ t_{s_2} \text{ to } t_{e_2} : C_2 >< /\text{think} >< \text{answer} > t_s \text{ to } t_e < /\text{answer} >$$

TVGBench: Evaluation Benchmark for LVLM on TVG

- Source from benchmarks: EgoNLQ, TaCoS, Charades, HiREST, ActivityNet
- 800 samples, 11 semantic types of Human/Object/Environment

human pose	The old wolman’s face.
human action (simple)	A person walks in.
human action (complex)	He gets down on the ground and kicks.
human action (procedural)	The person takes out a knife from the drawer, rinses it, then cuts off the stem of the fig.
object attribute	What color was the bag?
object counting	Fourteen desert are here.
object transition	Meet is placed into a pot.
object existence (simple)	Where is the bicycle?
object existence (complex)	Where is the footmat after I moved it with my leg?
environment change	The screen goes black.
environment state	A call center is shown.



SOTA TVG Performance on Charades and ActivityNet

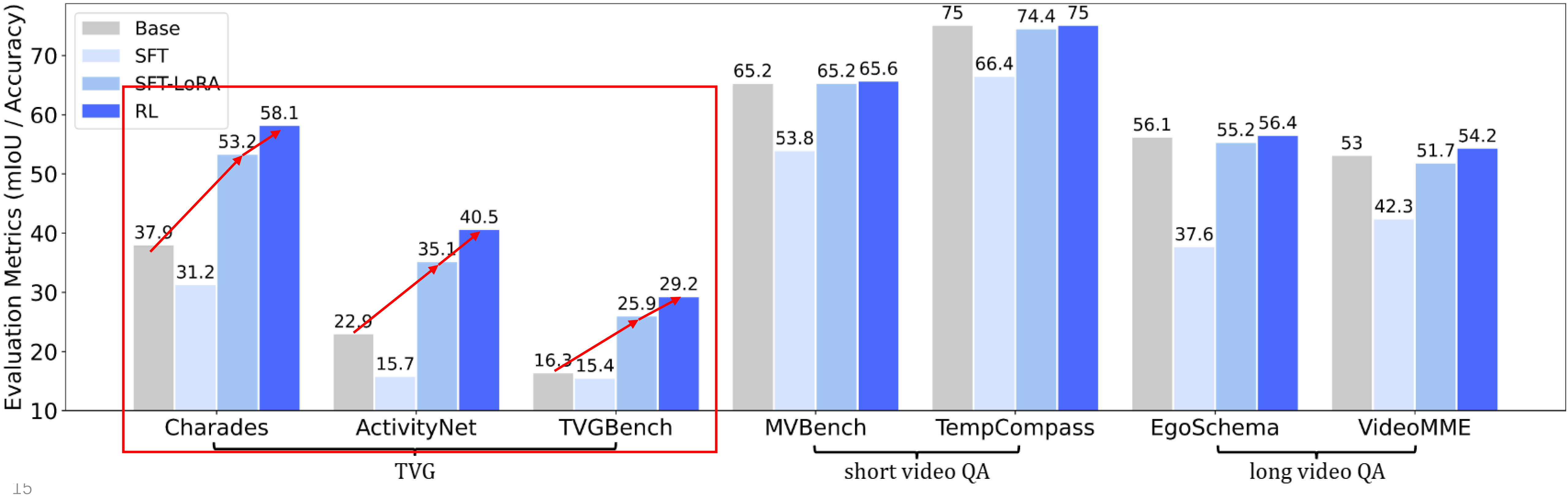
- Improvement over prev SOTA (R1@0.7):

Charades: 27.8%↑, ActivityNet: 49.6%↑, TVGBench: 12.3%↑

Type	Method	Charades-STA			ActivityNet			TVGBench		
		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
VLP	2D-TAN* [63]	57.3	45.8	27.9	60.4	43.4	25.0	-	-	-
	UniVTG* [30]	72.6	60.2	38.6	56.1	43.4	24.3	-	-	-
	SSRN* [66]	-	65.5	42.6	-	54.5	33.2	-	-	-
	SnAG* [37]	-	64.6	46.2	-	48.6	30.6	-	-	-
	EaTR* [22]	-	68.4	44.9	-	58.2	37.6	-	-	-
	Gemini-2.5-Pro [10]	-	-	-	-	-	-	39.1	24.4	12.8
SFT	ChatVTG [41]	52.7	33.0	15.9	40.7	22.5	9.4	-	-	-
	TimeChat [44]	-	32.2	13.4	36.2	20.2	9.5	22.4	11.9	5.3
	HawkEye [50]	50.6	31.4	14.5	49.1	29.3	10.7	-	-	-
	VTimeLLM [21]	51.0	27.5	11.4	44.0	27.8	14.3	-	-	-
	TimeSuite [60]	69.9	48.7	24.0	-	-	-	31.1	18.0	8.9
	VideoChat-Flash [27]	74.5	53.1	27.6	-	-	-	32.8	19.8	10.4
	TRACE [18]	-	40.3	19.4	-	37.7	24.0	37.0	25.5	14.6
	HawkEye* [50]	72.5	58.3	28.8	55.9	34.7	17.9	-	-	-
	TimeSuite* [60]	79.4	67.1	43.0	-	-	-	-	-	-
RL	Time-R1 (ours)	**78.1**	**60.8**	**35.3**	**58.6**	**39.0**	**21.4**	**41.8**	**29.4**	**16.4**
	Time-R1 (ours)*	82.8	72.2	50.1	73.3	55.6	34.0	-	-	-
	Time-R1-3B	74.6	53.1	26.0	40.0	21.0	8.7	33.5	21.0	10.5
	Time-R1-3B*	78.7	64.1	36.9	66.8	46.8	24.7	-	-	-

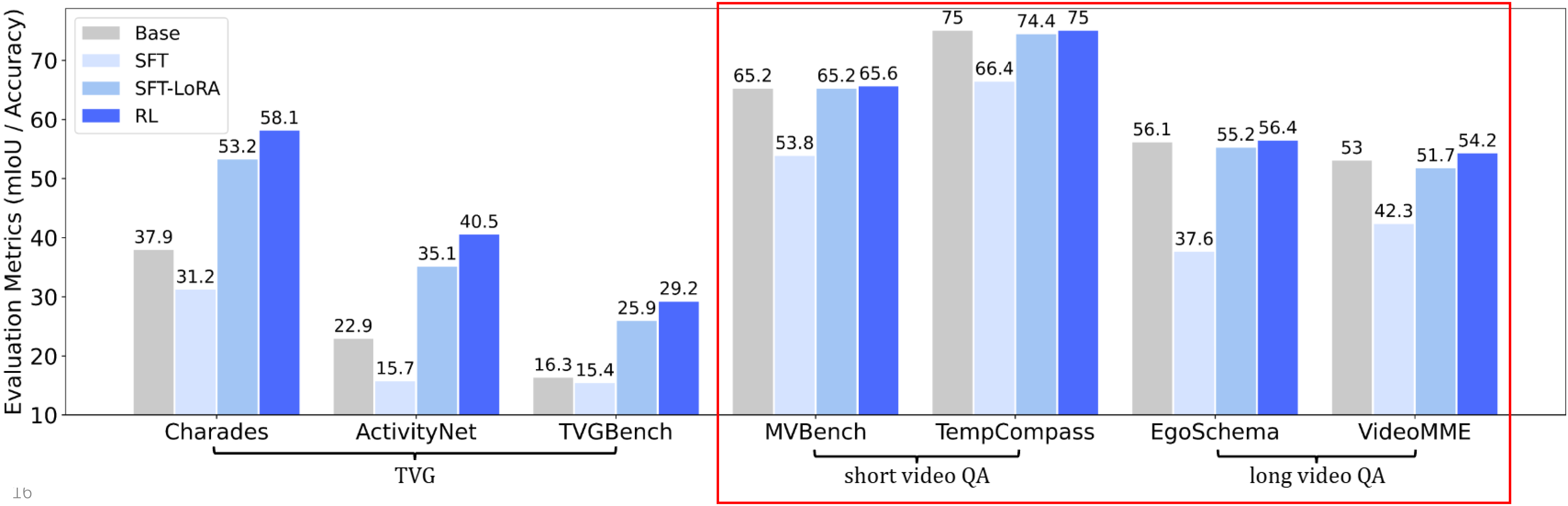
Comparison between SFT and RL on TVG and VideoQA

- TVG:
  - Both SFT-LoRA and RL boost performance over Base model (Qwen2.5-VL-7B)
  - RL consistently outperforms SFT-LoRA by ~5 points.
- VideoQA: RL improves, while SFT-LoRA lose some VideoQA performance



Comparison between SFT and RL on TVG and VideoQA

- TVG:
  - Both SFT-LoRA and RL boost performance over Base model (Qwen2.5-VL-7B)
  - RL consistently outperforms SFT-LoRA by ~5 points.
- VideoQA: RL improves, while SFT-LoRA lose some VideoQA performance



Ablation Study of Time-R1 Training

- Each component matters

Table 2: Ablation of Time-R1-7B trainning. GF, ME, SF refers to Gaussian Filtering, Multi-Epoch, and Sample Filtering per epoch, respectively.

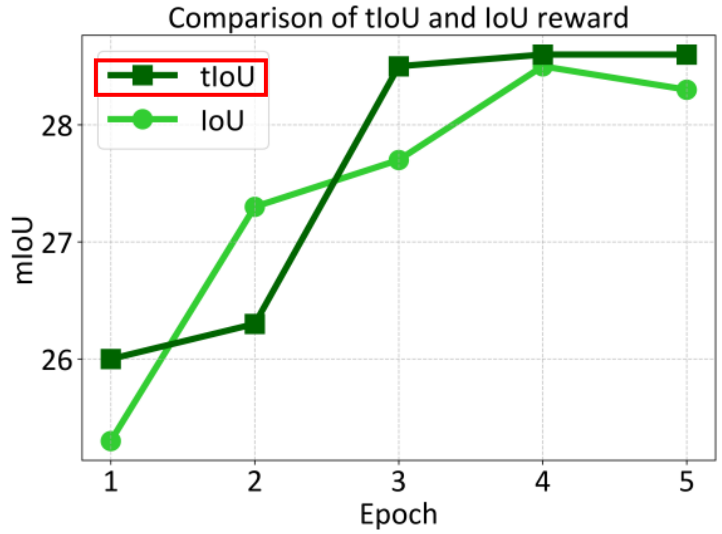
	tIoU	GF	ME	SF	TVGBench R1@0.3	R1@0.5	R1@0.7
1	✗	✗	✗	✗	38.0	24.8	13.2
2	✓	✗	✗	✗	36.0	23.6	12.9
3	✗	✓	✗	✗	37.2	25.0	13.4
4	✗	✗	✓	✗	39.9	26.0	14.2
5	✓	✓	✗	✗	38.4	25.6	14.1
6	✓	✗	✓	✗	39.4	26.5	16.4
7	✓	✓	✓	✗	41.6	28.5	15.6
8	✓	✓	✓	✓	41.8	29.4	16.4

Table 4: Ablation of data filtering strategies.

Method	R1@0.3	R1@0.5	R1@0.7	mIoU
random	39.4	26.5	16.4	27.4
gaussian (0.3)	41.6	28.5	15.6	28.6
gaussian (0.5)	40.6	28.2	16.0	28.3
gaussian (0.7)	37.2	26.9	15.5	26.5
uniform	40.4	28.5	15.9	28.3

Table 5: Ablation of KL and CoT in GRPO.

KL	CoT	R1@0.3	R1@0.5	R1@0.7	mIoU
✗	✗	40.4	29.1	14.9	28.1
✓	✗	40.8	27.4	15.0	27.7
✗	✓	42.9	29.5	15.0	29.1
✓	✓	41.6	28.5	15.6	28.6



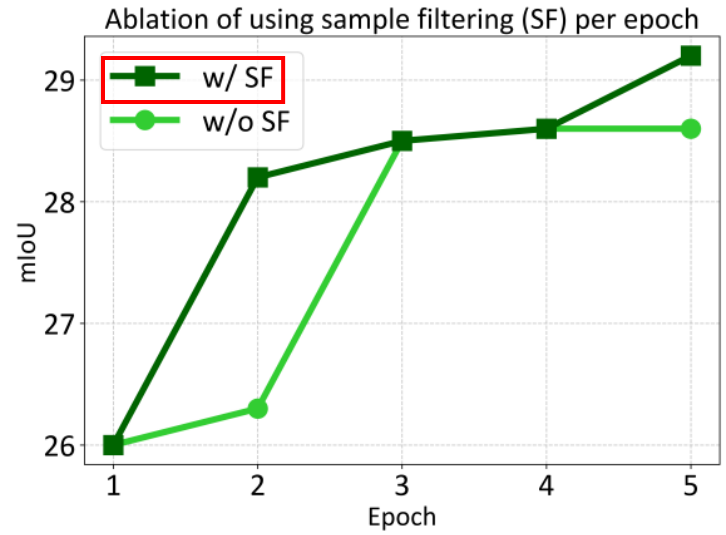


Table 6: Comparison of the token-level loss design used by DAPO [56] and the sample-level loss design used by GRPO [45].

Loss	Charades-STA R1@0.3	R1@0.5	R1@0.7	mIoU	ActivityNet R1@0.3	R1@0.5	R1@0.7	mIoU	TVGBench R1@0.3	R1@0.5	R1@0.7	mIoU
GRPO	76.7	59.8	34.4	57.0	55.9	37.1	20.3	37.8	40.8	28.0	16.5	28.4
DAPO	77.4	60.0	34.1	57.2	56.2	37.4	20.4	38.0	41.6	28.5	15.6	28.6

Ablation Across Different Base Models and Model Sizes

- Significant improvement across all models

Model	Method	Charades				ActivityNet				TVGBench			
		R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
Qwen-2.5-VL-3B	Base	24.2	15.5	8.1	16.3	13.0	7.1	3.3	9.8	11.5	6.5	3.8	8.3
	Time-R1	74.6	53.1	26.0	51.2	40.0	21.0	8.7	23.2	33.5	21.0	10.5	21.7
	Time-R1*	78.7	64.1	36.9	59.9	66.8	46.8	24.7	46.1	-	-	-	-
Qwen-2.5-VL-7B	Base	58.7	38.3	16.6	37.9	34.3	21.6	12.9	22.9	24.9	16.0	8.0	16.3
	Time-R1	78.1	60.8	35.5	58.1	58.1	39.0	21.4	40.5	41.8	29.4	16.4	29.2
	Time-R1*	82.8	72.2	50.1	60.9	73.3	55.6	34.0	52.1	-	-	-	-
MiMo-VL-7B	Base	48.5	27.0	12.1	31.7	31.3	19.7	12.1	24.2	22.4	12.6	6.6	15.7
	Time-R1	79.9	63.9	33.4	53.9	45.6	27.2	14.2	31.9	41.2	27.8	15.1	27.4
InternVL-2B	Base	20.9	7.8	1.9	15.4	18.6	8.5	3.1	14.2	16.3	6.3	2.3	11.7
	Time-R1	24.0	11.5	3.5	15.7	20.6	9.5	3.9	14.2	21.8	9.5	4.1	14.8
InternVL-8B	Base	27.8	11.9	3.7	20.6	33.1	18.4	10.3	24.0	17.4	8.3	3.4	11.8
	Time-R1	70.0	45.1	18.3	44.1	46.8	25.9	11.7	31.1	38.0	22.5	9.2	24.2

Ablation of Reward Design

- Both the tIoU and format reward matters

- Standard IoU reward $r_{\text{IoU}}(\cdot)$. The standard Intersection over Union between the predicted segment $[t_s, t_e]$ and the ground-truth segment $[t'_s, t'_e]$, computed as:

$$r_{\text{IoU}} = \frac{\max(0, \min(t_e, t'_e) - \max(t_s, t'_s))}{\max(t_e, t'_e) - \min(t_s, t'_s)} \tag{8}$$

- Timestamp-aware IoU reward $r_{\text{tIoU}}(\cdot)$. The timestamp-aware IoU reward augments the standard IoU with a center alignment term that penalizes discrepancies between the centers of the predicted and ground-truth segments:

$$r_{\text{tIoU}} = r_{\text{IoU}} + r_{\text{center}}, \quad \text{where} \quad r_{\text{center}} = 1 - \frac{|(t_s + t_e)/2 - (t'_s + t'_e)/2|}{t'_e - t'_s} \tag{9}$$

This modification provides a more fine-grained grounding signal by encouraging both boundary alignment and temporal center consistency.

- Exact matching reward $r_{\text{em}}(\cdot)$. A sparse binary reward that is 1 only if the predicted timestamps exactly match the ground truth, and 0 otherwise:

$$r_{\text{em}} = \mathbb{I}(t_s = t'_s \wedge t_e = t'_e) \tag{10}$$

- Absolute error reward $r_{\text{abs}}(\cdot)$. The negative L1 distance between the predicted and ground-truth boundaries:

$$r_{\text{abs}} = -(|t_s - t'_s| + |t_e - t'_e|) \tag{11}$$

- RMSE reward $r_{\text{rmse}}(\cdot)$. The negative Root Mean Square Error, which penalizes larger boundary errors more heavily:

$$r_{\text{rmse}} = -\sqrt{\frac{(t_s - t'_s)^2 + (t_e - t'_e)^2}{2}} \tag{12}$$

Reward Design	R1@0.3	R1@0.5	R1@0.7
$r_{\text{tIoU}} + r_{\text{format}}$ (Ours)	**41.8**	**29.4**	**16.4**
r_{format} only	27.1	18.0	10.1
r_{tIoU} (w/o format)	40.5	27.6	15.4
$r_{\text{IoU}} + r_{\text{format}}$	41.4	28.0	15.8
$r_{\text{em}} + r_{\text{format}}$	26.5	16.8	9.1
$r_{\text{abs}} + r_{\text{format}}$	39.1	27.8	14.8
$r_{\text{rmse}} + r_{\text{format}}$	38.9	27.0	15.8
$r_{\text{center}} + r_{\text{format}}$	37.6	25.9	15.0

Case Study of the TVG Task

- ActivityNet

Question: Where is the event "The man goes back to the pink bucket to wash another dark item clothing, he wrings it out in the pink bucket, throws it in the yellow bucket, then wrings his hands together a few times while staring at the camera"?

(video length: 74s)

Ground Truth	33.1s	73.5s
TimeSuite	20.0s	30.0s
Gemini-2.5-Pro	57.0s	73.0s
VideoChat-Flash	0.0s	17.2s
Trace	54.9s	74.1s
Time-R1 (ours)	35.0s	75.0s

Time-R1 thinking (ours): The event "The man goes back to the pink bucket to wash another dark item clothing, he wrings it out in the pink bucket, throws it in the yellow bucket, then wrings his hands together a few times while staring at the camera." occurs between 35.00 and 75.00 seconds in the video.

- Charades

Question: Where is the event "a person is seen sitting on a couch"?

(video length: 31s)

Ground Truth	0.0s	12.8s
TimeSuite	0.0s	5.0s
Gemini-2.5-Pro	0.0s	30.0s
VideoChat-Flash	0.0s	5.0s
Trace	0.0s	31.6s
Time-R1 (ours)	0.0s	10.0s

Time-R1 thinking (ours): The event "a person is seen sitting on a couch" occurs at the beginning of the video. The person is seated on a couch, holding a phone, and appears to be engaged with it. This aligns with the initial frames of the video.

Case Study on the VideoQA Task for Both Short and Long Videos

(video length: 38m) Question: What do heroes of legend use to defeat the enemy based on the video?

(A) Their wisdom ✗
(B) A big robot
(C) Their superpower
(D) Power of music ✓

(video length: 17s) Question: What's wrong with this car?

(A) It doesn't have a left rear wheel. ✗
(B) It doesn't have a right front wheel. ✓
(C) Its headlamp is broken.
(D) Its right door is broken.

VideoMME

Limitations

Weakness of MLLM-based solution

- Inference speed.
- Low fps, losing motion information.
- Unable to handle ultra-long videos (e.g., > 1hours)

Thank You!

If any questions, feel free to contact
boshenx@ruc.edu.cn,
or visit
https://xuboshen.github.io/
to check details in our paper!