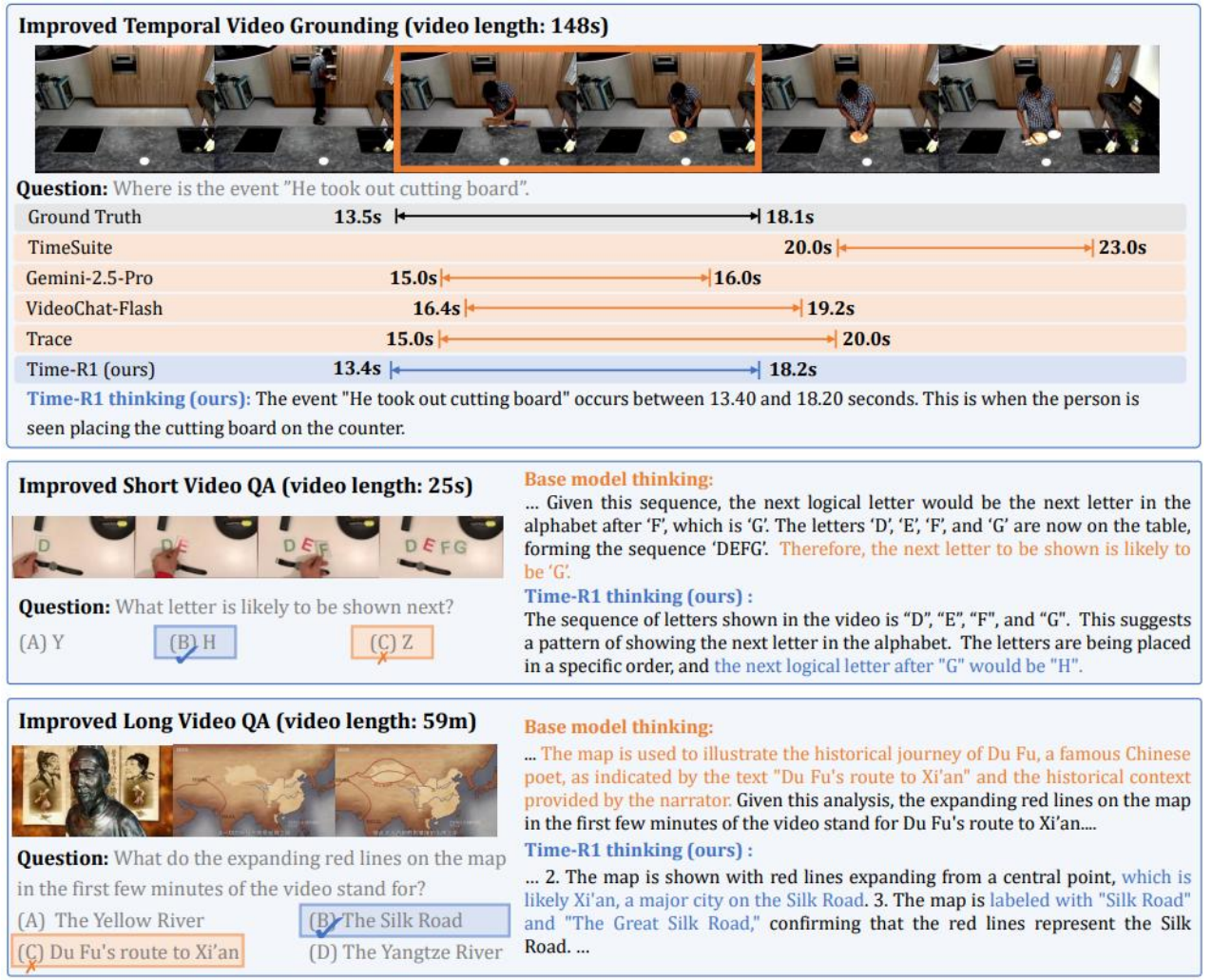
NEURAL INFORMATION PROCESSING SYSTEMS

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

- AIM3 Lab, Renmin University of China
- MiLM Plus, Xiaomi Inc

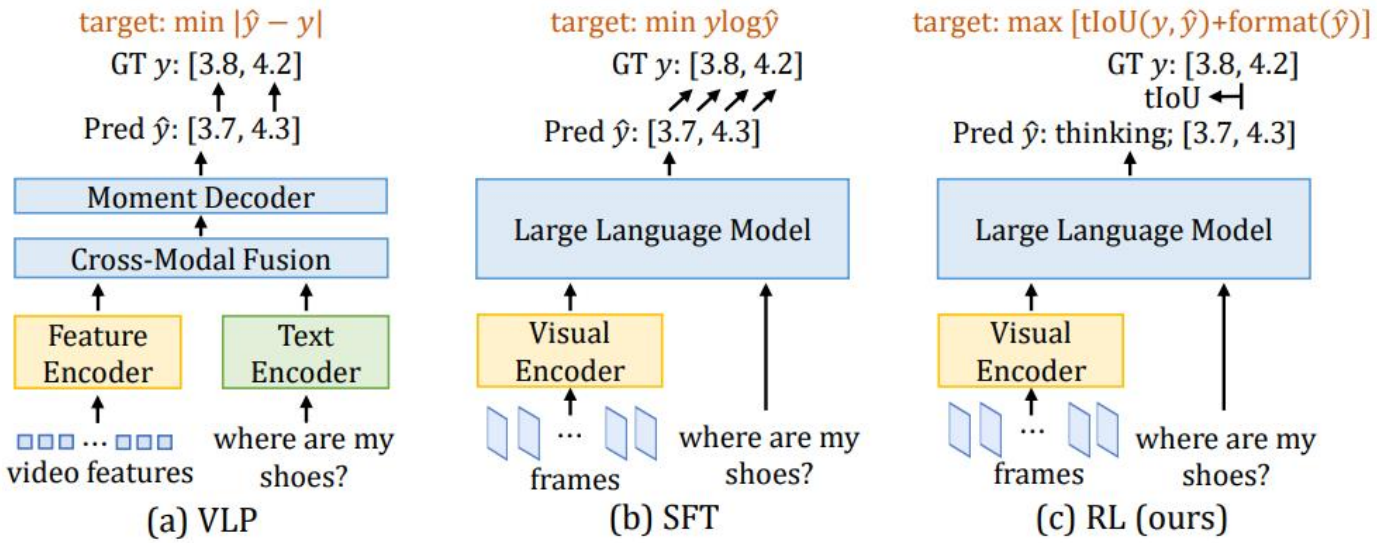
Introduction

- **Challenge:** Large Vision-Language Models (LVLMs) underperform on Temporal Video Grounding (TVG) because Supervised Fine-Tuning (SFT) over-penalizes reasonable timestamp deviations via autoregressive loss, leading to poor generalization.
- **Solution:** We propose Time-R1, a Reinforcement Learning (RL) based post-training framework that directly optimizes the task metric (IoU) to overcome SFT's limitations.
- **Contributions:**
  - **Time-R1:** An RL framework to enhance TVG capabilities.
  - **TimeRFT:** A data-efficient RL fine-tuning strategy and dataset.
  - **TVGBench:** A balanced benchmark designed for evaluating LVLMs on TVG.



Time-R1: Reinforcement Learning Framework

- **Core Idea:** Shift from token-matching SFT to outcome-driven RL.
- **Process:** Video/Text Input -> LVLM -> <think>Reasoning</think><answer>Timestamp</answer>
- **Reward Function:** Total Reward = r_tIoU (Timestamp-aware IoU) + r_form (Format Reward)
  - **r_tIoU:** Standard IoU plus a penalty for timestamp center deviation, ensuring more precise localization.
  - **r_form:** Encourages the structured "think-then-answer" output format.

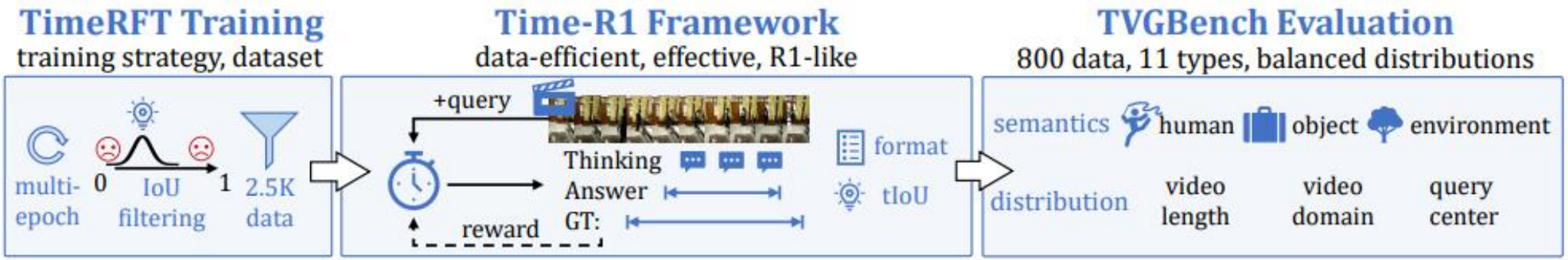


$$\text{IoU} = \frac{[t_s, t_e] \cap [t'_s, t'_e]}{[t_s, t_e] \cup [t'_s, t'_e]}$$

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

TimeRFT: Efficient RL Fine-Tuning

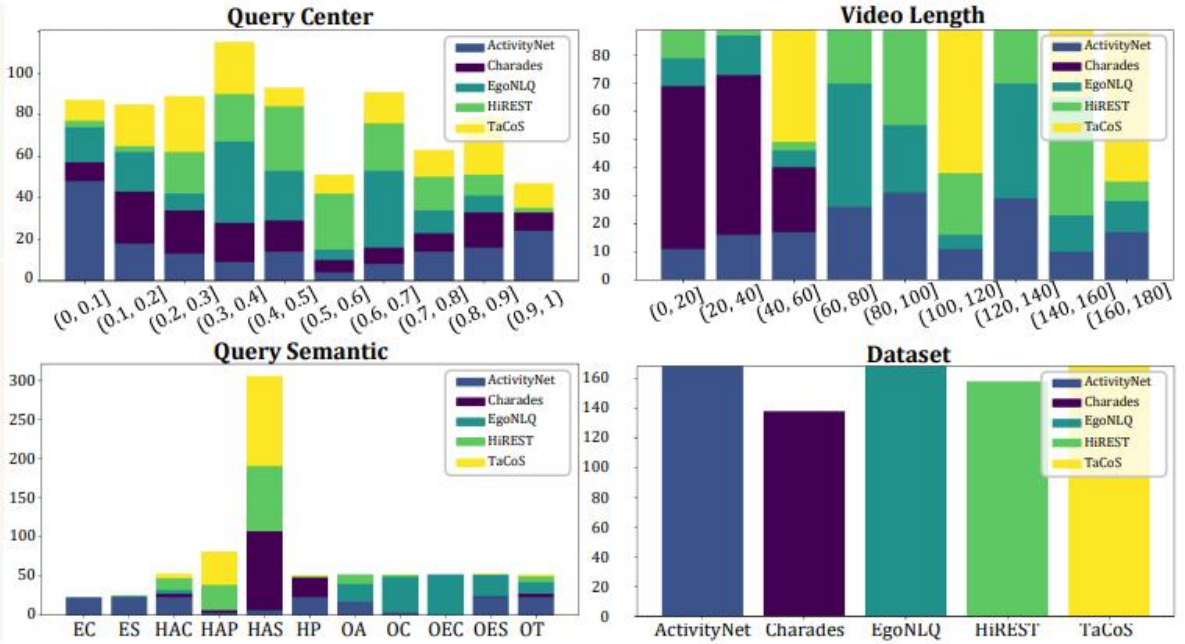
- **Data Efficiency:** Fine-tuned on only 2.5K samples selected from 339K, focusing on medium-difficulty instances (IoU≈0.3).
- **Dynamic Hard Sampling:** A multi-epoch strategy that filters out easy samples (IoU>0.7) after each epoch to focus on challenging cases.
- **Cold Start:** An initial SFT phase with a few CoT examples to suppress hallucinations, stabilize training, and reduce reasoning length.



TVGBench: A Balanced Benchmark

- **Features:** Lightweight (800 samples), with balanced distributions of data sources, video durations, and query temporal locations.
- **Innovation:** First to introduce 11 query semantic categories for fine-grained model analysis.

human action	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	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	environment change	The screen goes black.
	environment state	A call center is shown.



Experiments & Analysis

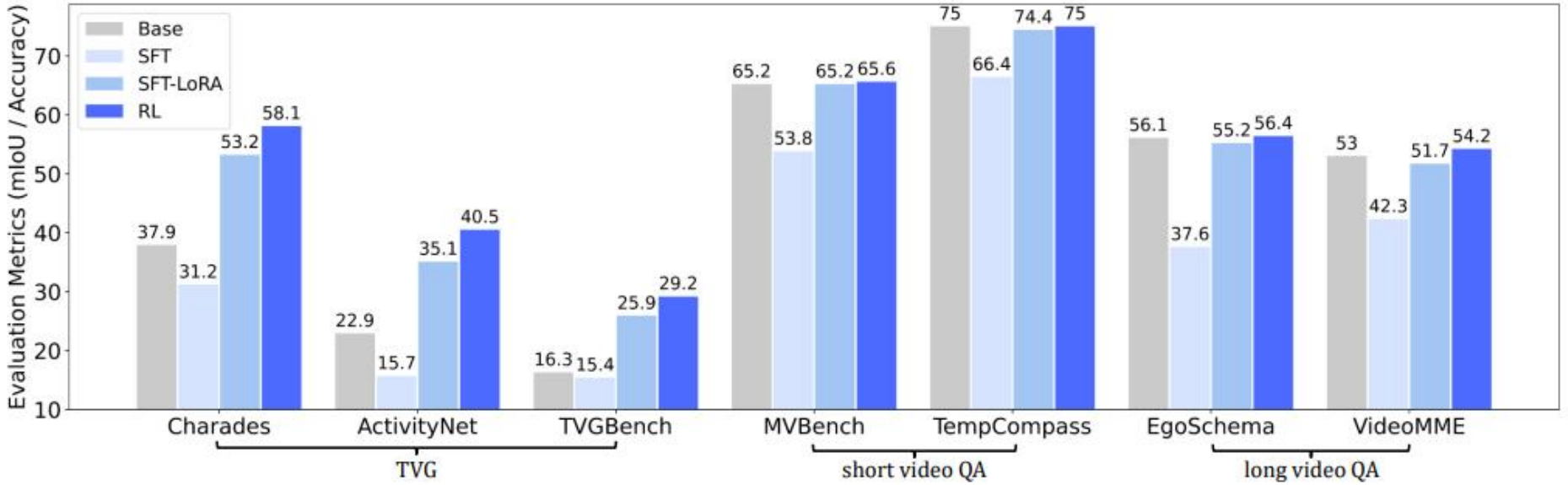
State-of-the-Art Performance

- **Zero-Shot:** Time-R1 (trained on 2.5K data) outperforms all SFT-based models and surpasses Gemini-2.5-Pro on TVGBench (41.8 vs. 39.1, R1@0.3).
- **Fine-Tuned:** Achieves top performance on Charades-STA and ActivityNet, exceeding all prior LVLMs and most classic VLP methods.

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* [68]	57.3	45.8	27.9	60.4	43.4	25.0	-	-	-
	UniVTG* [31]	72.6	60.2	38.6	56.1	43.4	24.3	-	-	-
	SSRN* [71]	-	65.5	42.6	-	54.5	33.2	-	-	-
	SnAG* [39]	-	64.6	46.2	-	48.6	30.6	-	-	-
	EaTR* [23]	-	68.4	44.9	-	58.2	37.6	-	-	-
	Gemini-2.5-Pro [10]	-	-	-	-	-	-	39.1	24.4	12.8
SFT	Momentor [43]	42.6	26.6	11.6	42.9	23.0	12.4	-	-	-
	ChatVTG [44]	52.7	33.0	15.9	40.7	22.5	9.4	-	-	-
	TimeChat [47]	-	32.2	13.4	36.2	20.2	9.5	22.4	11.9	5.3
	VTG-LLM [18]	-	33.8	15.7	-	-	-	-	-	-
	HawkEye [53]	50.6	31.4	14.5	49.1	29.3	10.7	-	-	-
	VTimeLLM [22]	51.0	27.5	11.4	44.0	27.8	14.3	-	-	-
	VideoChat-TPO [57]	58.3	40.2	18.4	-	-	-	-	-	-
	VideoExpert [70]	61.5	40.3	20.9	-	-	-	-	-	-
	TimeSuite [65]	69.9	48.7	24.0	-	-	-	31.1	18.0	8.9
	VideoMind [32]	73.5	59.1	31.2	48.4	30.3	15.7	-	-	-
	VideoChat-Flash [28]	74.5	53.1	27.6	-	-	-	32.8	19.8	10.4
	TRACE [19]	-	40.3	19.4	-	37.7	24.0	37.0	25.5	14.6
	HawkEye* [53]	72.5	58.3	28.8	55.9	34.7	17.9	-	-	-
	TimeSuite* [65]	79.4	67.1	43.0	-	-	-	-	-	-
	VideoChat-TPO* [57]	77.0	65.0	40.7	-	-	-	-	-	-
	VideoExpert* [70]	74.3	60.8	36.5	-	-	-	-	-	-
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	-	-	-

Experiments & Analysis

- **Generalization:** RL preserves and improves general video QA capabilities, whereas SFT causes catastrophic forgetting.
- **Data Efficiency:** 2.5K (RL) > 339K (SFT-LoRA).



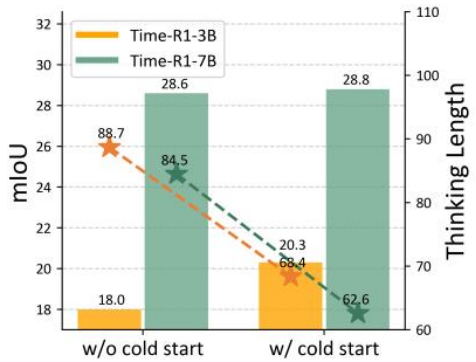
Experiments & Analysis

- **Framework Generality:** Consistently effective across various model architectures and sizes (Qwen-VL, MiMo-VL, InternVL).
- **Component Effectiveness:** Ablation studies confirm the critical roles of the TimeRFT strategy, rtIoU reward, and cold start in boosting performance.

Model	Type	R1@0.3	R1@0.5	R1@0.7
Qwen-2.5-VL-3B [4]	Base	11.5	6.5	3.8
	Time-R1	**33.5**	**21.0**	**10.5**
Qwen-2.5-VL-7B [4]	Base	24.9	16.0	8.0
	Time-R1	**41.6**	**28.5**	**15.6**
MiMo-VL-7B [56]	Base	22.4	12.6	6.6
	Time-R1	**41.2**	**27.8**	**15.1**
InternVL3-2B [72]	Base	16.3	6.3	2.3
	Time-R1	**21.8**	**9.5**	**4.1**
InternVL3-8B [72]	Base	17.4	8.3	3.4
	Time-R1	**38.0**	**22.5**	**9.2**

	tIoU	GF	ME	SF	TVGBench R1@0.3	TVGBench R1@0.5	TVGBench 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

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



Thank You