Seminar

2025.08.22
Nanxing Hu

DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning

Ziwei Zheng[1,2*], **Michael Yang**[1*], **Jack Hong**[1*], **Chenxiao Zhao**[1*†], **Guohai Xu**[1‡], **Le Yang**[2‡], **Chao Shen**[2], **XingYu**[1]

[1]Xiaohongshu Inc., [2] Xi'an Jiaotong University
[*] Equal contribution, Random order [†] Main Code Contributor [‡] Corresponding Author

`Project Homepage`

`{chenxiao2, xuguohai}@xiaohongshu.com, yangle15@xjtu.edu.cn,`
`ziwei.zheng@stu.xjtu.edu.cn, {yangminghao199,jaaackhong}@gmail.com`

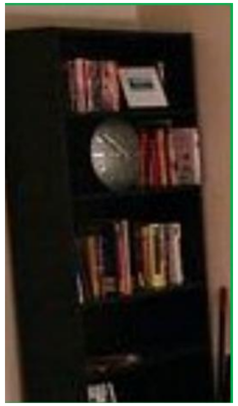
Setting Clarification

Setting: (high resolution) image understanding by VLM

Example:

Q: Is the clock to the left of the laptop?

A: ???????

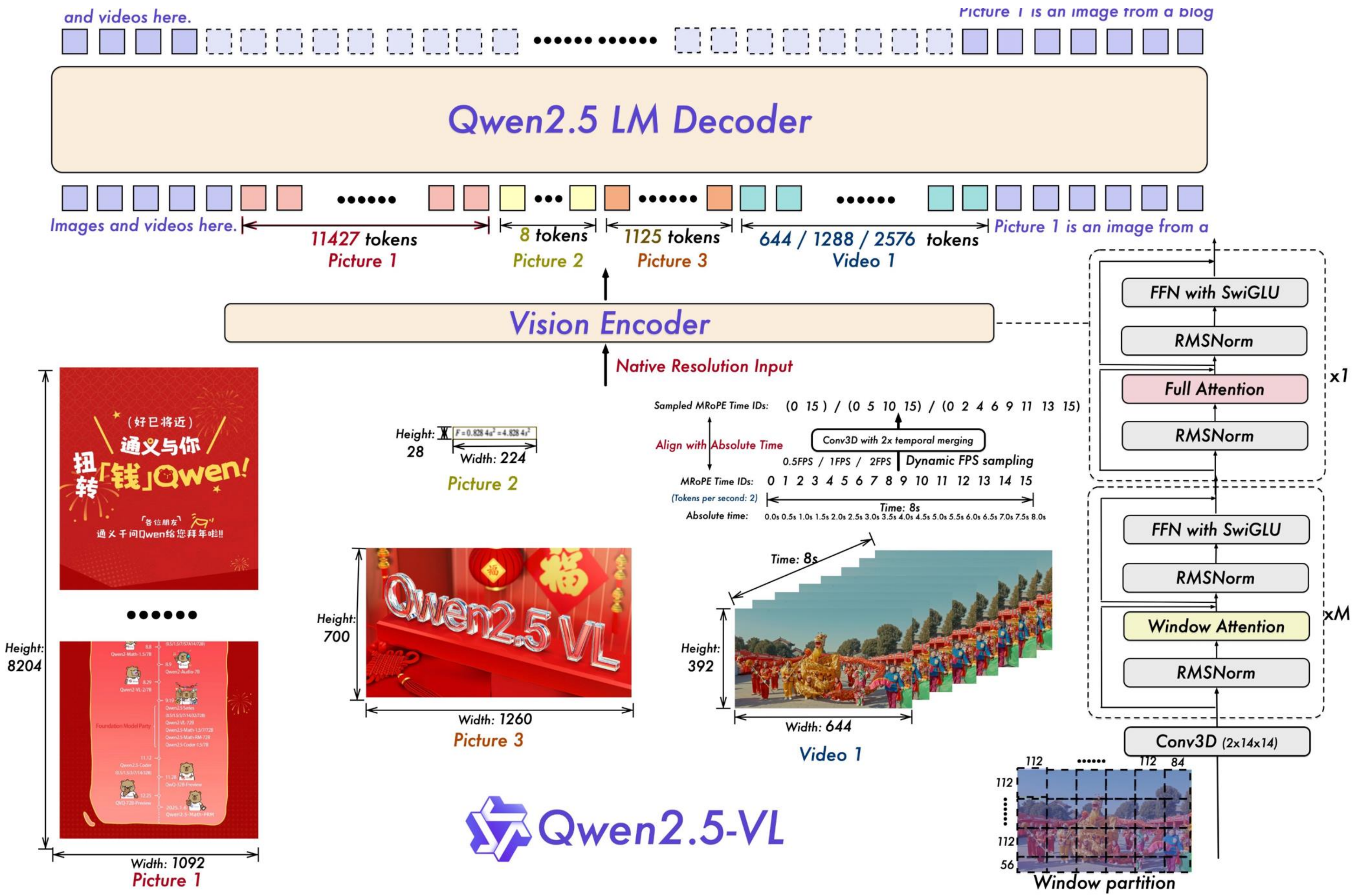


A: Yes !

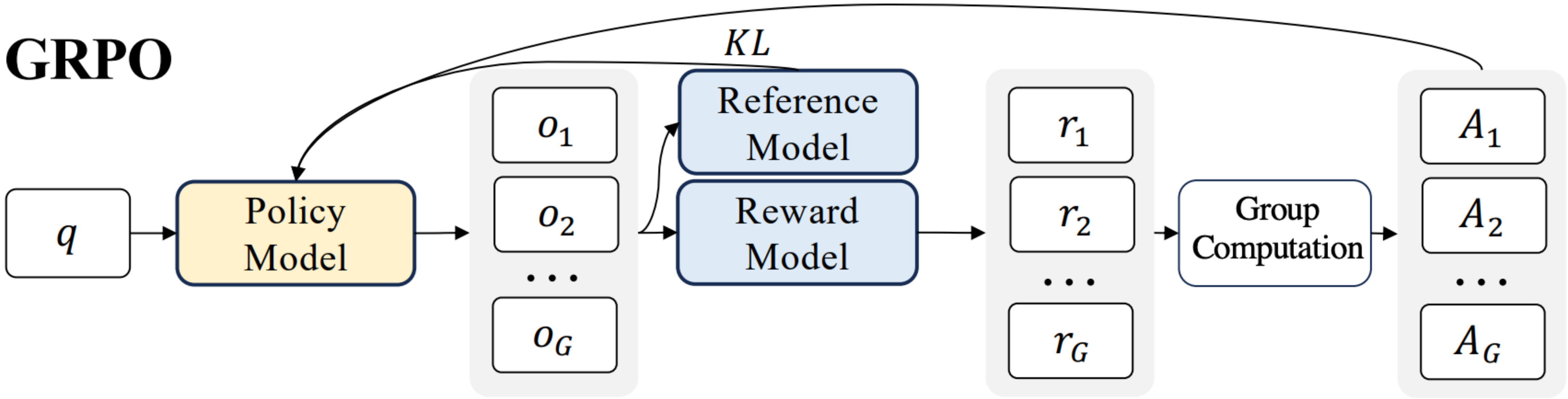
If we can make VLM recall local visual information in the reasoning process, the reasoning ability will be improved.

Approach: Reinforcement Learning

Preliminary: VLM



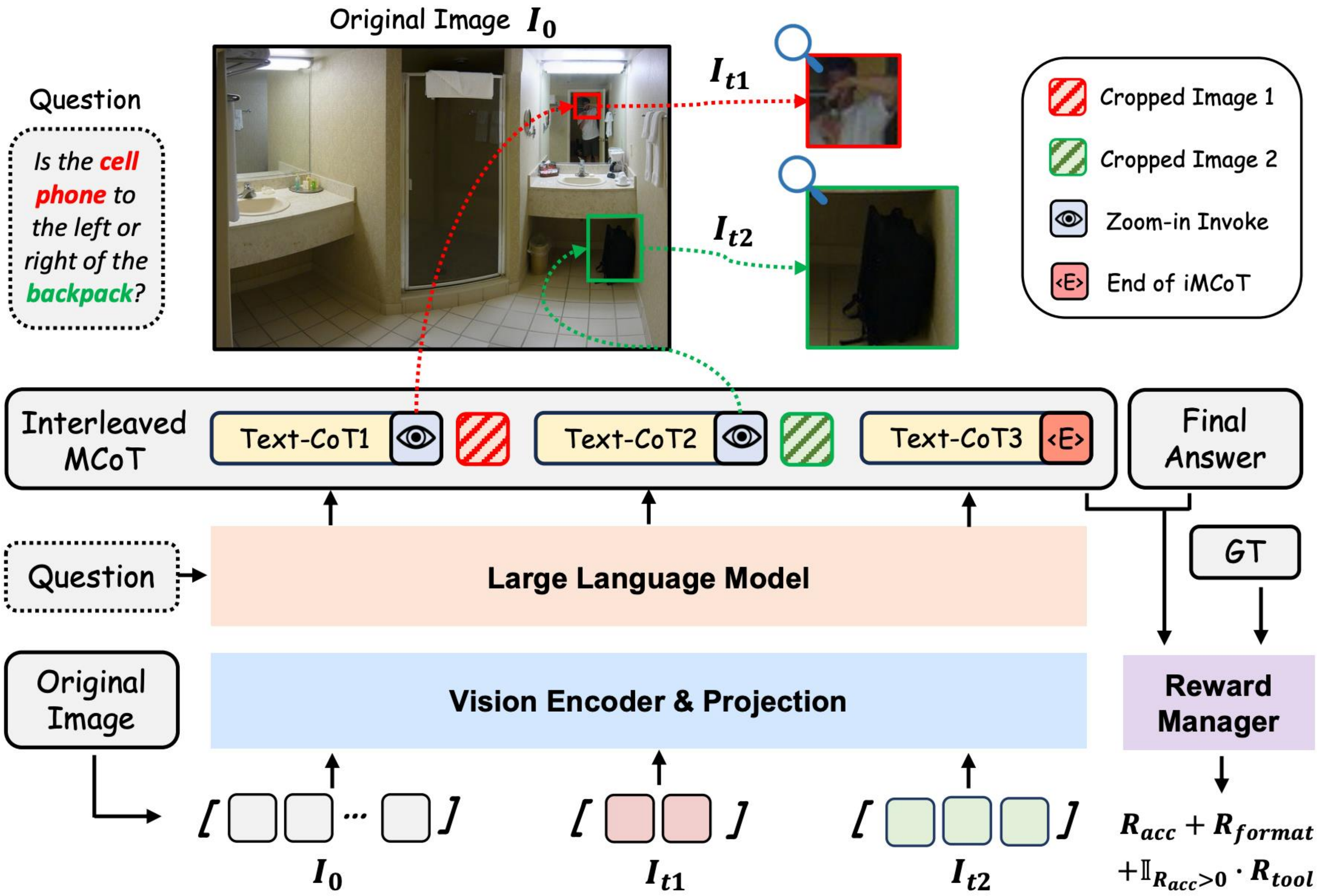
Preliminary: GRPO



$$\hat{A}_{i,t} = \widetilde{r}_i = \frac{r_i - \mathrm{mean}(\mathbf{r})}{\mathrm{std}(\mathbf{r})},$$

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$
$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \mathrm{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta || \pi_{ref}\right]\right\}, \quad (3)$$

Pipeline



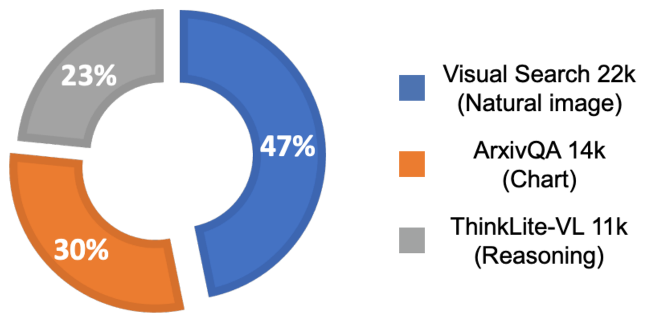
Original Image I_0
Question
Is the cell phone to the left or right of the backpack?
I_{t1}
I_{t2}
Cropped Image 1
Cropped Image 2
Zoom-in Invoke
End of iMCoT
Interleaved MCoT
Text-CoT1
Text-CoT2
Text-CoT3
Final Answer
GT
Question
Large Language Model
Original Image
Vision Encoder & Projection
Reward Manager
I_0
I_{t1}
I_{t2}
$R_{acc} + R_{format} + \mathbb{I}_{R_{acc}>0} \cdot R_{tool}$

Data Construction

Principle:

- Diverse Tasks and Image Distribution
- Reasoning Ability Enhancement
- Tool Effectiveness

Data Collection:



fine-grained data, chart data, and reason data.

Data Selection:

Managing Difficulties: generate 8 responses per question excluded as they are either too easy or too hard.

+

Facilitating Tool Integration: select instances achieves correct results when utilizing ground-truth crop regions.

Reward design

$$R(\tau) = R_{\text{acc}}(\tau) + R_{\text{format}}(\tau) + \mathbb{I}_{R_{\text{acc}}(\tau)>0} \cdot R_{\text{tool}}(\tau),$$

$R_{acc}(\tau)$: accuracy reward assesses whether the final answer is correct.

$R_{format}(\tau)$: formatting reward penalizes poorly structured outputs.

$R_{tool}(\tau)$: Tool usage bonus is awarded only when the model produces a correct answer and invokes at least one external perception tool during the trajectory

Experiment results

High-Resolution Benchmarks

Model	E2E	Param Size	V^* Bench [41] Attr	V^* Bench [41] Spatial	V^* Bench [41] Overall	HR-Bench 4K [59] FSP	HR-Bench 4K [59] FCP	HR-Bench 4K [59] Overall	HR-Bench 8K [59] FSP	HR-Bench 8K [59] FCP	HR-Bench 8K [59] Overall
GPT-4o [60]	✓	-	-	-	66.0	70.0	48.0	59.0	62.0	49.0	55.5
o3 [8]	✓	-	-	-	95.7	-	-	-	-	-	-
SEAL [41]	✗	7B	74.8	76.3	75.4	-	-	-	-	-	-
DyFo [44]	✗	7B	80.0	82.9	81.2	-	-	-	-	-	-
ZoomEye [61]	✗	7B	93.9	85.5	90.6	84.3	55.0	69.6	88.5	50.0	69.3
LLaVA-OneVision [62]	✓	7B	75.7	75.0	75.4	72.0	54.0	63.0	67.3	52.3	59.8
Qwen2.5-VL* [58]	✓	7B	73.9	67.1	71.2	85.2	52.2	68.8	78.8	51.8	65.3
Qwen2.5-VL* [58]	✓	32B	87.8	88.1	87.9	89.8	58.0	73.9	84.5	56.3	70.4
DeepEyes	✓	7B	91.3	88.2	90.1	91.3	59.0	75.1	86.8	58.5	72.6
Δ (*vs* Qwen2.5-VL 7B)	-	-	+17.4	+21.1	+18.9	+6.1	+6.8	+6.3	+10.0	+6.8	+7.3

Grounding and Hallucination Benchmarks

Model	Param Size	refCOCO	refCOCO+	refCOCOg	ReasonSeg	POPE Adversarial	POPE Popular	POPE Random	POPE Overall
LLaVA-OneVision [62]	7B	-	-	-	-	-	-	-	88.4
Qwen2.5-VL [58]	7B	90.0	84.2	87.2	-	-	-	-	-
Qwen2.5-VL* [58]	7B	89.1	82.6	86.1	68.3	85.9	86.5	87.2	85.9
DeepEyes	7B	89.8	83.6	86.7	68.6	84.0	87.5	91.8	87.7
Δ (*vs* Qwen2.5-VL 7B)	-	+0.7	+1.0	+0.6	+0.3	-1.9	+1.0	+4.6	+1.8

Multimodal Reasoning Benchmarks

Model	Param Size	MathVista [64]	MathVerse [65]	MathVision [66]	WeMath [67]	DynaMath [68]	LogicVista [69]
LLaVA-OneVision [62]	7B	58.6†	19.3†	18.3†	20.9†	-	33.3†
Qwen2.5-VL [58]	7B	68.2	49.2	25.1	35.2†	-	44.1†
Qwen2.5-VL* [58]	7B	68.3	45.6	25.6	34.6	53.3	45.9
DeepEyes	7B	70.1	47.3	26.6	38.9	55.0	47.7
Δ (*vs* Qwen2.5-VL 7B)	-	+1.9	+1.7	+1.0	+4.3	+1.7	+1.8

Ablations

Table 4: **Ablation Study on iMCoT.** We compare the results of RL training using text-only CoT and iMCoT on the same datasets.

Model	V^* Bench Attr	V^* Bench Spatial	V^* Bench Overall	HR-Bench 4K FSP	HR-Bench 4K FCP	HR-Bench 4K Overall	HR-Bench 8K FSP	HR-Bench 8K FCP	HR-Bench 8K Overall
Qwen2.5-VL [58]	73.9	67.1	71.2	85.2	52.2	68.8	78.8	51.8	65.3
RL w. Text-only CoT	90.4	85.5	88.5	92.3	58.5	75.4	69.3	52.3	60.8
DeepEyes	91.3	88.2	90.1	91.3	59.0	75.1	86.8	58.5	72.6

Multi-modal CoT make a difference

Table 5: **Impact of Training Data.** Fine represents the fine-grained data. HR denotes HR-Bench. Row #0 is the origin score of Qwen2.5-VL 7B.

#	Fine	Reason	Chart	High-Resolution V^* Bench	High-Resolution HR-4K	High-Resolution HR-8K	Basic VL Capability ReasonSeg	Basic VL Capability POPE	Reasoning MathVista	Reasoning MathVerse
0				71.2	68.8	65.3	68.3	85.9	68.2	45.6
1	✓			86.9	68.9	67.3	69.0	86.6	67.0	42.9
2	✓			91.6	74.1	71.0	69.1	88.1	64.7	41.3
3	✓	✓		91.6	73.8	70.5	68.6	88.8	67.7	43.8
4	✓		✓	90.1	74.6	74.6	68.5	87.9	64.6	38.1
5	✓	✓	✓	90.1	75.1	72.6	68.6	87.7	70.1	47.3

#1 denote training with unfiltered data

- data selection is necessary
- reasoning data is necessary for maintain reasoning ability
- Chart data can benefit the Math problem

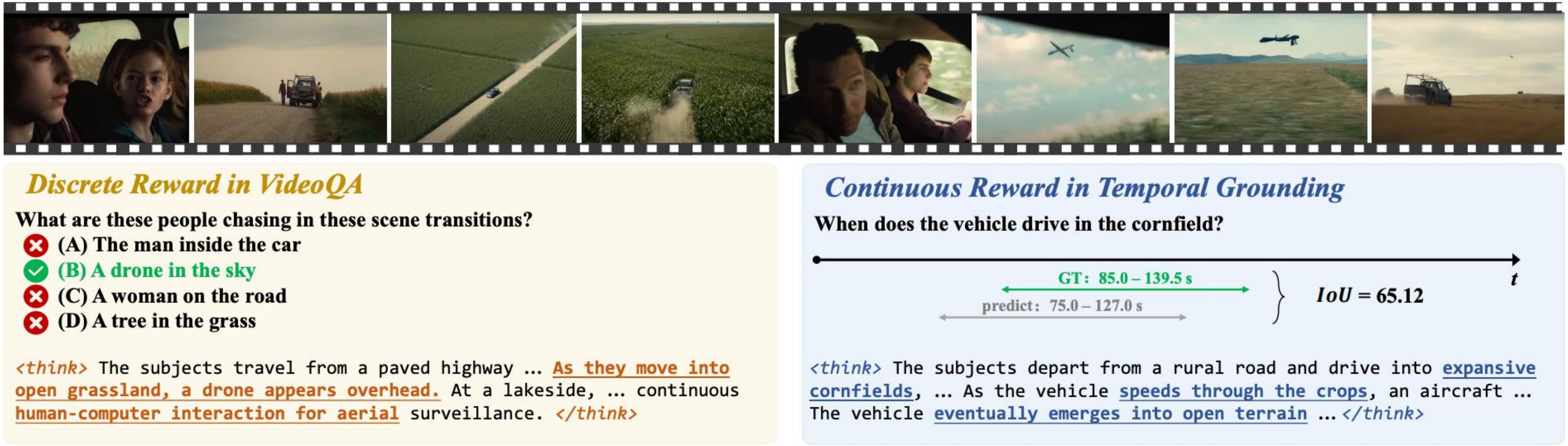
Reinforcement Learning Tuning for VideoLLMs: Reward Design and Data Efficiency

Hongyu Li[1]*, Songhao Han[1]*, Yue Liao[2]*†, Junfeng Luo[3], Jialin Gao[3], Shuicheng Yan[2], Si Liu[1]‡

[1] BUAA [2] NUS [3] Meituan

Setting Clarification

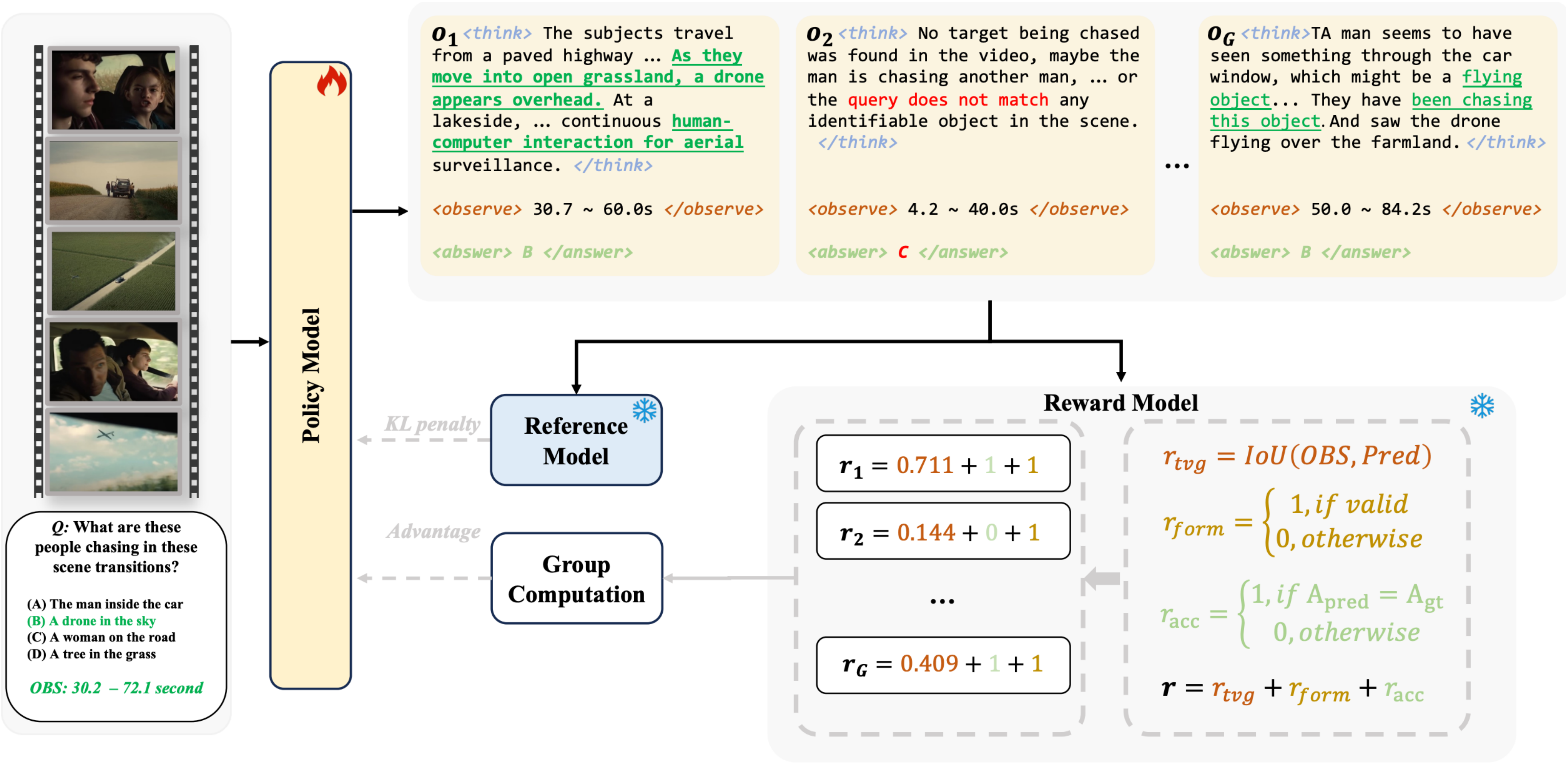
Setting: Reinforcement learning for video-specific reasoning capabilities of MLLMs



Video QA

Temporal Grounding

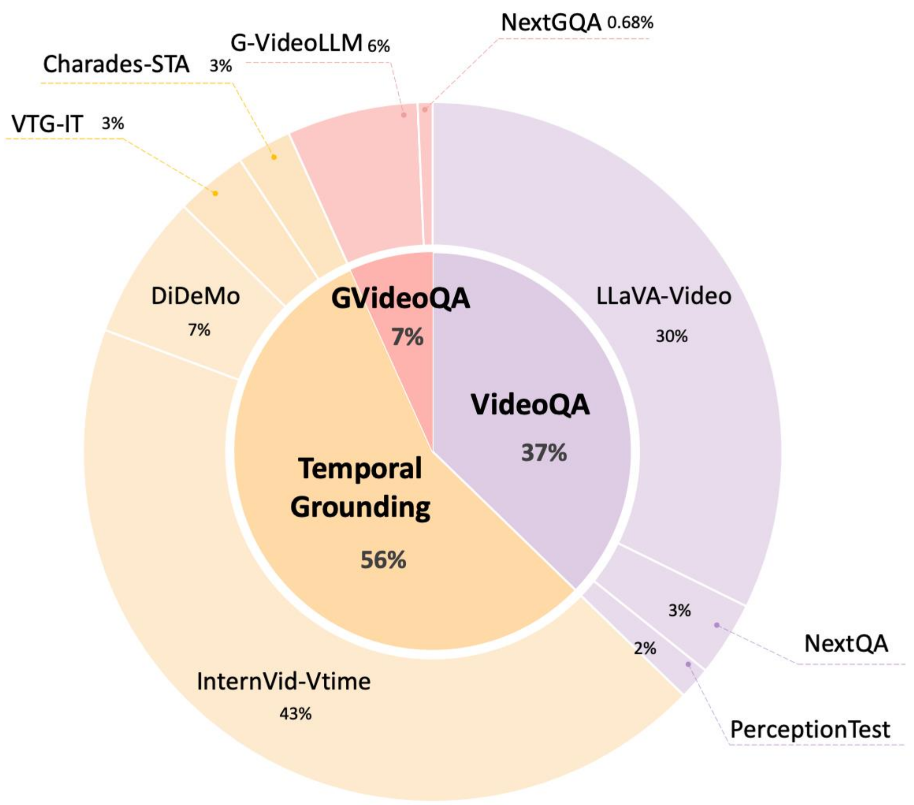
Pipeline



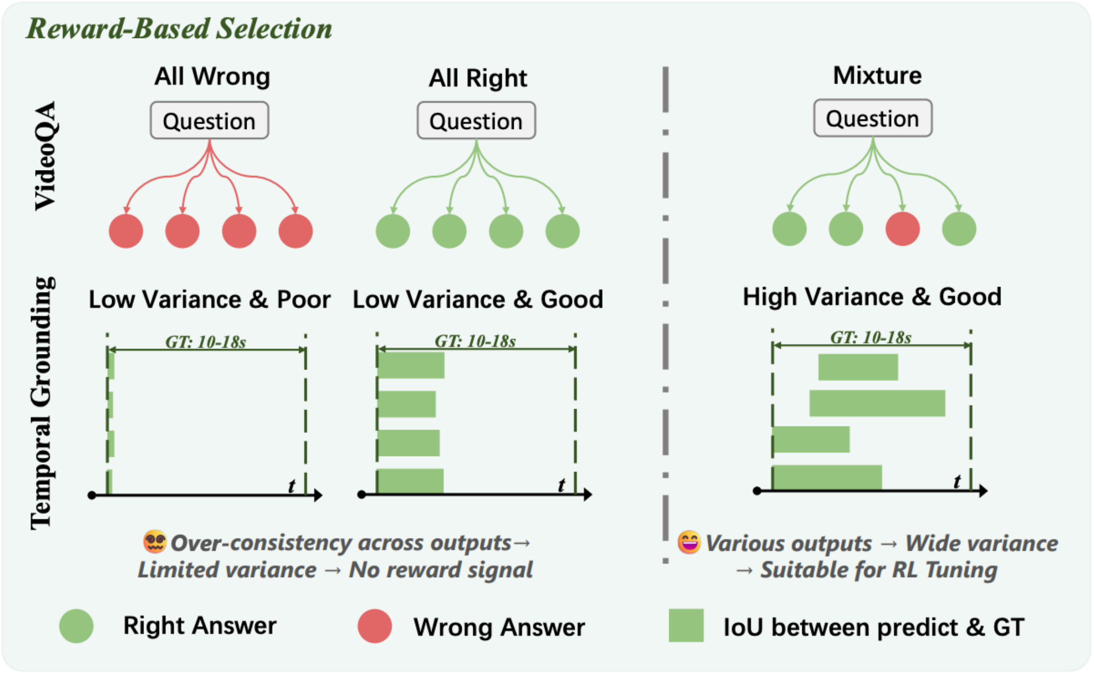
Q: What are these people chasing in these scene transitions?
(A) The man inside the car
(B) A drone in the sky
(C) A woman on the road
(D) A tree in the grass
OBS: 30.2 – 72.1 second
Policy Model
o1 <think> The subjects travel from a paved highway … As they move into open grassland, a drone appears overhead. At a lakeside, … continuous human-computer interaction for aerial surveillance. </think>
<observe> 30.7 ~ 60.0s </observe>
<abswer> B </answer>
o2 <think> No target being chased was found in the video, maybe the man is chasing another man, … or the query does not match any identifiable object in the scene. </think>
<observe> 4.2 ~ 40.0s </observe>
<abswer> C </answer>
…
oG <think>TA man seems to have seen something through the car window, which might be a flying object… They have been chasing this object. And saw the drone flying over the farmland. </think>
<observe> 50.0 ~ 84.2s </observe>
<abswer> B </answer>
KL penalty
Reference Model
Advantage
Group Computation
Reward Model
r1 = 0.711 + 1 + 1
r2 = 0.144 + 0 + 1
…
rG = 0.409 + 1 + 1
r_tvg = IoU(OBS, Pred)
r_form = 1, if valid; 0, otherwise
r_acc = 1, if A_pred = A_gt; 0, otherwise
r = r_tvg + r_form + r_acc

Data Construction

Data construction:



Data Selection :



VQA:

$$\text{Easy if } c \geq \tau_{\text{easy}}, \quad \text{Hard if } c \leq \tau_{\text{hard}}, \quad \text{otherwise Medium},$$

$\tau_{easy} = 1; \tau_{hard} = 7$

VTG:

$$\Delta_{\text{IoU}} = \max_i \text{IoU}_i - \text{mean}_i(\text{IoU}_i), \qquad \Delta_{IoU} = 0.3$$

Reward design

Multi-Choice VideoQA (MC-QA) : $R_{\text{mc}} = R_{\text{format}} + R_{\text{acc}},$

Temporal Video Grounding (TVG) : $R_{\text{tvg}} = R_{\text{format}} + R_{\text{IoU}},$

Grounded VideoQA (GQA) : $R_{\text{gqa}} = R_{\text{format}} + \frac{1}{2}(R_{\text{acc}} + R_{\text{IoU}}),$

Experiment results

Method	Temporal Video Grounding			General VideoQA			Reasoning QA	Grounded QA	
	Charades	ANet	ANet-RTL	MVBench	TempCompass	VideoMME	MMVU	NextGQA	
	mIoU	mIoU	mIoU	Avg	Avg	Avg (wo sub)	Avg	mIoU	acc
General VideoLLM									
LLaMA-VID[18]	-	-	-	41.9	45.6	-	-	-	-
VideoLLaMA2[3]	-	-	-	54.6	-	47.9	44.8	-	-
LongVA-7B[39]	-	-	-	-	56.9	52.6	-	-	-
Video-UTR-7B[35]	-	-	-	58.8	59.7	52.6	-	-	-
LLaVA-OV-7B[14]	-	-	-	56.7	-	58.2	49.2	-	-
Kangeroo-7B[19]	-	-	-	61.1	62.5	56.0	-	-	-
GRPO-based Method and Baseline									
Qwen-VL-2.5[2]	28.0	24.0	6.0	65.3	70.9	56.1	61.3	20.2	77.2
Qwen-VL-2.5-SFT	43.0	24.3	18.1	62.0	68.7	49.6	52.5	28.3	70.6
Video-R1[4]	-	-	-	62.7	72.6	57.4	64.2	-	-
Temporal-RLT (ours)	**57.0**	**39.0**	**27.6**	**68.1**	**73.3**	**57.6**	**65.0**	**37.3**	**78.7**

Ablations

Table 4: Ablation Studies: Video QA and TVG Data Selection.

Easy: Middle: Hard	General VideoQA			Reasoning QA
	MVBench	TempCompass	VideoMME	MMVU
4 : 4 : 2	64.3	70.0	52.8	59.5
2 : 4 : 4	65.9	70.3	55.9	63.0
2 : 6 : 2	67.2	71.3	56.8	62.1
1 : 8 : 1	68.1	**73.4**	57.1	**63.4**
0 : 10 : 0	**68.1**	72.5	**58.6**	63.1

(a) Ablation for Video QA Data Selection

Δ_{IoU}	Charades-STA			
	Recall@0.3	Recall@0.5	Recall@0.7	mIoU
0	78.2	63.9	37.4	54.7
0.1	78.0	64.8	38.9	54.9
0.2	**78.8**	63.7	38.5	55.0
0.3	78.6	**64.5**	**39.9**	**55.5**

(b) Ablation for TVG Data Selection

<u>Diversity and training efficacy data makes a difference.</u>

Table 5: Temporal Video Grounding OOD Evaluation.

Tuning Type	Charades-STA				ActivityNet				ActivityNet-RTL			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
✗	42.4	29.8	14.0	28.0	34.4	22.5	11.6	24.0	7.9	2.6	2.9	6.0
SFT	73.9	61.6	38.5	52.8	33.4	18.9	9.0	23.1	24.0	14.8	7.4	17.8
RLT	80.2	68.3	44.5	57.9	56.9	38.4	20.2	39.1	40.2	22.7	10.9	26.3

only trained on Charades-STA dataset

<u>RLT performs significantly better than SFT on OOD task.</u>