Video-R4: Reinforcing Text-Rich Video Reasoning with Visual Rumination

Yolo Tang[1] Daiki Shimada[2] Hang Hua[3] Chao Huang[1] Jing Bi[1] Rogerio Feris[3] Chenliang Xu[1]

[1]University of Rochester [2]Sony Group Corporation [3]MIT-IBM Watson AI Lab

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

Understanding **text-rich videos** (news, driving, UI demos, slides) requires reading small, transient textual cues that appear only in specific frames or localized regions.

Most video QA models rely on **single-pass perception** over fixed frames with text-only chain-of-thought, leading to **hallucinations** about content never actually observed.

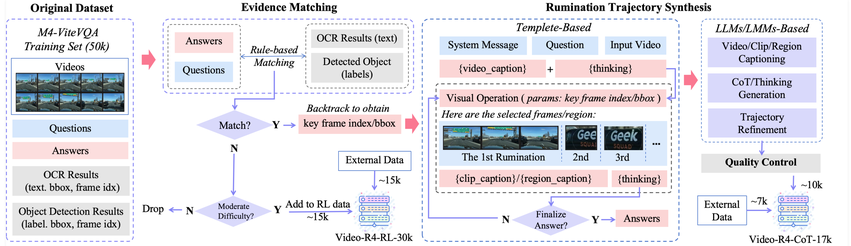
Key Insight: Humans adopt an iterative *"pause–zoom–check"* strategy. We equip LMMs with the same ability: **select frames → zoom into regions → re-encode pixels → revise understanding**.

Contributions

- Two curated datasets: **Video-R4-CoT-17k** (SFT) and **Video-R4-RL-30k** (RL) with executable rumination trajectories.
- **Video-R4**: a video reasoning LMM performing iterative visual rumination via frame selection, spatial zooming, and pixel re-encoding.
- **Multi-stage rumination learning**: DRP → RL → CRP → RL with GRPO-based RL and novel reward design (diversity + representativeness + curiosity).
- **State-of-the-art on M4-ViteVQA** with strong zero-shot transfer to document QA, slides QA, and general video QA.

Data Curation Pipeline

Starting from M4-ViteVQA, we apply **rule-based evidence matching** (OCR + object detection), synthesize **rumination trajectories** with clipping and cropping operations, and apply quality control via human annotation.



17k	30k	2
CoT Trajectories (SFT)	RL Samples (GRPO)	Visual Ops (Clip+Crop)

Reward Design

$$R' = R + \lambda_{\text{div}} R_{\text{div}} + \lambda_{\text{rep}} R_{\text{rep}} + \lambda_{\text{cur}} R_{\text{cur}}$$

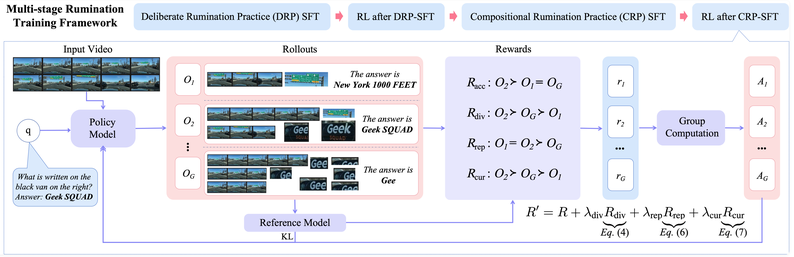
Diversity Reward R_{div}
Encourages selected regions to be mutually dissimilar in feature space via average pairwise cosine distance.

Representativeness Reward R_{rep}
Ensures selected frames cover the global video in feature space by minimizing average nearest-neighbor distance.

Curiosity Reward R_{cur}
Encourages visual tool use when under-utilized; penalizes excessive calls to prevent over-reliance.

Multi-Stage Training Framework

We train Video-R4 with a four-stage curriculum: **DRP-SFT → RL$_d$ → CRP-SFT → RL$_c$**.



DRP-SFT: Learn each atomic operation (crop/clip) in isolation on 7k trajectories. **RL$_d$:** Refine with GRPO on 15k single-tool samples.

CRP-SFT: Interleave both operations in 10k multi-step trajectories. **RL$_c$:** Final GRPO stage (15k) to sharpen decisions on when to stop, re-zoom, and explore.

GRPO (Group Relative Policy Optimization): For each query, sample a group of G rollouts, compute rewards, and normalize advantages within the group: $A_i = (R_i - \text{mean}) / \text{std}$. The policy is updated with a clipped objective plus KL regularization to prevent deviation from the reference model. This group-relative formulation reduces variance and encourages relative ranking over absolute reward magnitudes.

State-of-the-Art on M4-ViteVQA

Models	LMM	Grnd.	RL	Task 1-S1		Task 1-S2		Task 2	
				Acc.	ANLS	Acc.	ANLS	Acc.	ANLS
Qwen2-VL-7B	✓	✗	✗	35.22	45.84	27.25	38.45	21.23	28.79
GAT-L	✗	✓	✗	38.30	48.23	30.90	41.81	22.13	30.75
Video-R1-7B	✓	✗	✓	37.10	48.25	33.67	44.94	43.16	53.37
Pixel-Reasoner	✓	✓	✓	52.91	61.44	48.88	58.23	58.97	65.32
Video-R4-7B (Ours)	✓	✓	✓	**56.17**	**65.22**	**52.69**	**61.89**	**64.21**	**69.99**
Human	–	–	–	*85.27*	*89.30*	*78.41*	*82.80*	*82.26*	*85.10*

Finding 1
Longer rumination with more pixel-grounded steps consistently boosts accuracy, demonstrating a clear **test-time scaling effect**.

Finding 3
RL encourages a preference for **cropping over clipping**. Zooming provides more informative and less redundant evidence, mirroring how humans pause and inspect frames.

Ablation Study

Training Framework	Task 1-S1		Task 1-S2		Task 2	
	Acc.	ANLS	Acc.	ANLS	Acc.	ANLS
DRP→RL$_d$→CRP→RL$_c$ (full)	**56.17**	**65.22**	**52.69**	**61.89**	**64.21**	**69.99**
w/o R$_{\text{cur}}$	54.35	63.92	50.90	61.10	61.92	69.20
DRP→CRP→RL$_c$	54.98	63.74	51.50	60.80	60.44	68.59
CRP-SFT only	46.76	62.67	40.47	59.68	49.47	66.33
Base (Qwen2.5-VL-7B)	26.53	44.91	24.34	39.60	32.81	50.82

Finding 2
DRP → RL → CRP → RL schedule is optimal. Curiosity reward is critical for stable training.

Generalization (Zero-Shot Transfer)

Trained only on text-rich video QA. **No dataset-specific tuning** for benchmarks below.

General Video QA

Models	MVBench	Video-MME	Video-MMMU
Qwen2.5-VL-7B	57.4	53.1	47.8
Video-R1-7B	62.7	**57.4**	49.8
Pixel-Reasoner	**65.4**	54.6	47.7
Video-R4-7B (Ours)	**64.5**	**54.5**	**52.2**

MP-DocVQA (val)

Models	ZS	Acc.	ANLS
Big-Bird	✗	41.06	49.29
Hi-VT5	✗	48.28	62.01
Video-R4 (Ours)	✓	**53.21**	**62.22**

SlidesVQA (test)

Models	ZS	EM	F1
M3D	✗	33.5	41.7
Video-R4 (Ours)	✓	**43.0**	**52.2**
Human	–	*89.8*	*93.0*

Finding 4: Training on text-rich videos transfers well to **multi-page documents, slides, and general video QA**, with strong gains on the text-heavy Video-MMMU benchmark.

Visualization Example

Iterative visual rumination in action: the model selects frames, zooms into text regions, re-reads, and refines its answer.

Conclusion

Video-R4 acquires evidence through iterative visual rumination, decomposing video understanding into **frame selection, spatial zooming, and re-encoding cycles**. Multi-stage training with GRPO-based RL proves essential. Achieves **state-of-the-art on M4-ViteVQA** with broad zero-shot generalization to documents, slides, and general video QA.

Project Page & Code

Acknowledgements: Supported by Sony Group Corporation. Thanks to Sayaka Nakamura and Jerry Jun Yokono.