

Watching Itself Watch

Self-Auditing Visual Reliance for Video Reasoning

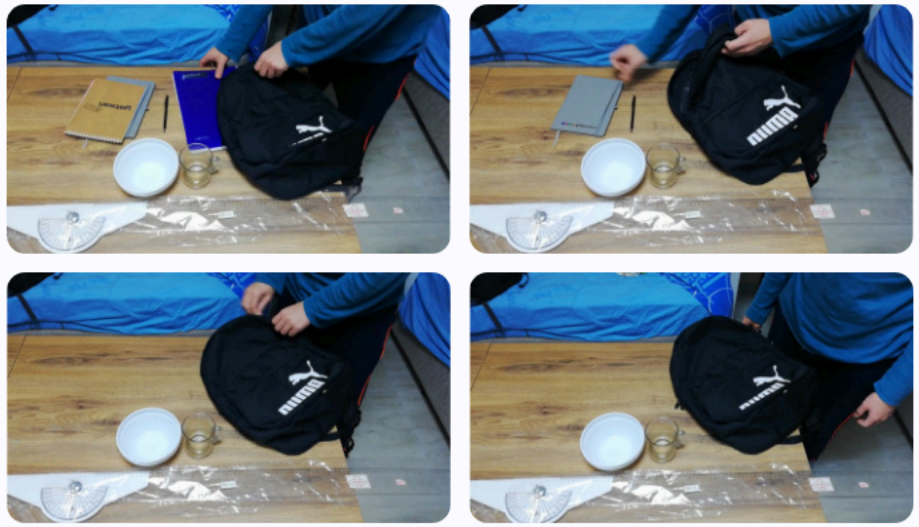
Rohan Dugad Mi Luo Kristen Grauman

The University of Texas at Austin

NeurIPS 2026

MOTIVATION

Visually decisive tokens



How many objects were taken out from the backpack?

A. 0 ✓ B. 1 C. 2

A model's reasoning (excerpt)

The person in the video is **packing** a backpack, and I need to determine how many objects were taken out from it. First, let's observe the objects on the table: - There is a **notebook** with a **pen**. - There is a **white bowl**. - There is a **glass mug**. - There is a **protractor**. - There is a **plastic sheet**. ... From what I can see, the person is **not taking anything out** of the backpack; instead, they are **putting items inside** it. ... The person is not taking anything out of the backpack, so the number of objects taken out is **zero**.

PRIOR WORK

Post-training for video reasoning



Recent video reasoners built on GRPO

Video-R1 Feng et al., 2025	spatio-temporal rewards	
VideoChat-R1 Li et al., 2025	spatio-temporal rewards	
Video-R2 Maaz et al., 2025	spatio-temporal rewards	
VideoRFT Wang et al., 2025	visual-evidence rewards	
Video-VER Luo et al., 2025	visual-evidence rewards	
VideoChat-R1.5 Yan et al., 2025	iterative perception	

relies on a larger judge model

PRIOR WORK

Limitations of GRPO

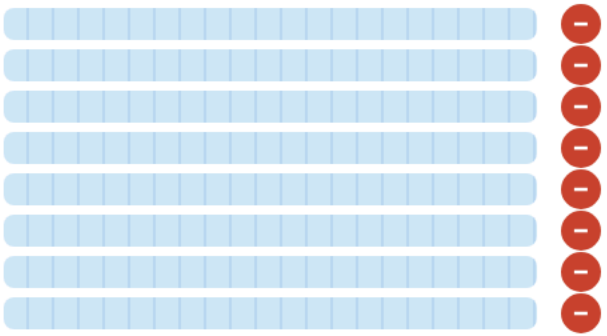
One reward for the whole trajectory

... The person appears to be packing or unpacking the backpack. ... it seems that only the notebook and pen were taken out from the backpack. The other objects remain on the table. ... Based on the analysis, only two objects were taken out: the notebook and the pen.

where it went wrong where it was right

Answer: C (2) ✗ → reward − on every token alike

Same reward for every rollout



→ no learning signal

30% of GRPO groups in our runs had identical rewards

PRIOR WORK

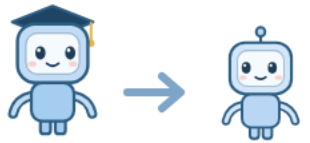
Other post-training methods



Supervised fine-tuning

Chu et al., 2025

- ✗ **Off-policy:** trains on prefixes it never produces itself
- ✗ **Distribution shift:** never sees its own mistakes
- ✗ **Memorization:** copies surface patterns, generalizes poorly



On-policy self-distillation

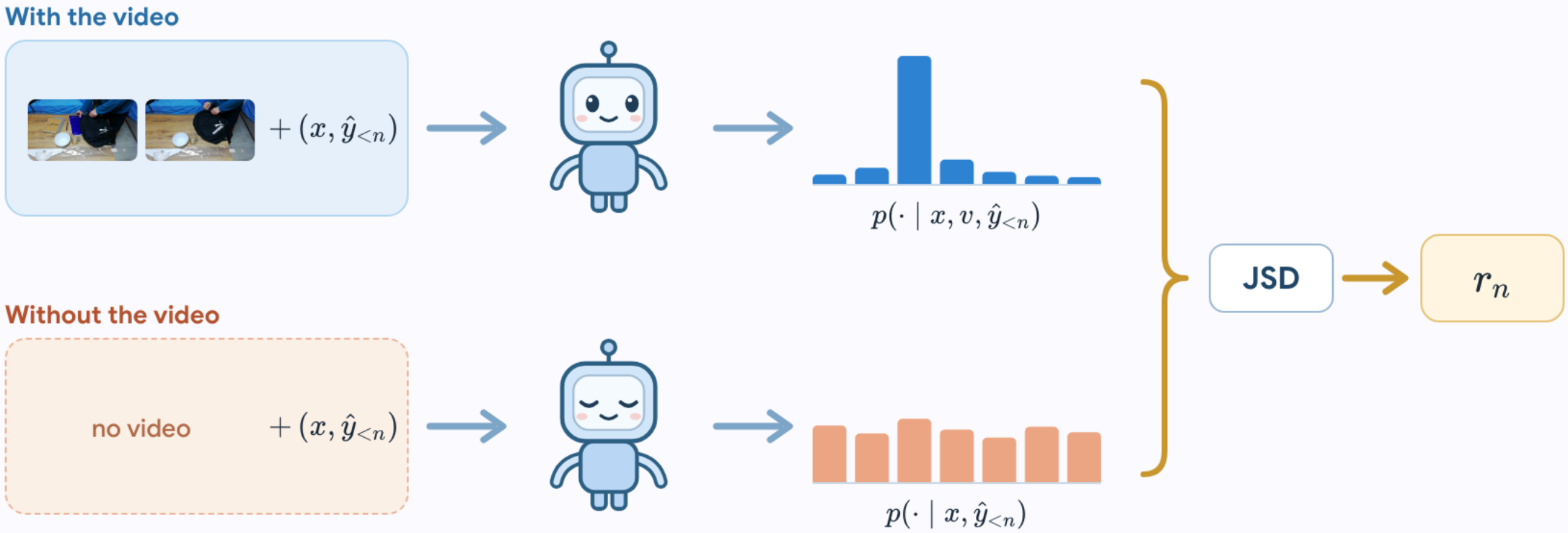
Zhao et al., 2026

- ✗ **Frozen teacher:** fixed at the base model, never improves
- ✗ **Needs ground-truth chains of thought** per example
- ✗ **Text only:** built for language, not visual input

None of this tells the model where it failed to use the video.

KEY IDEA

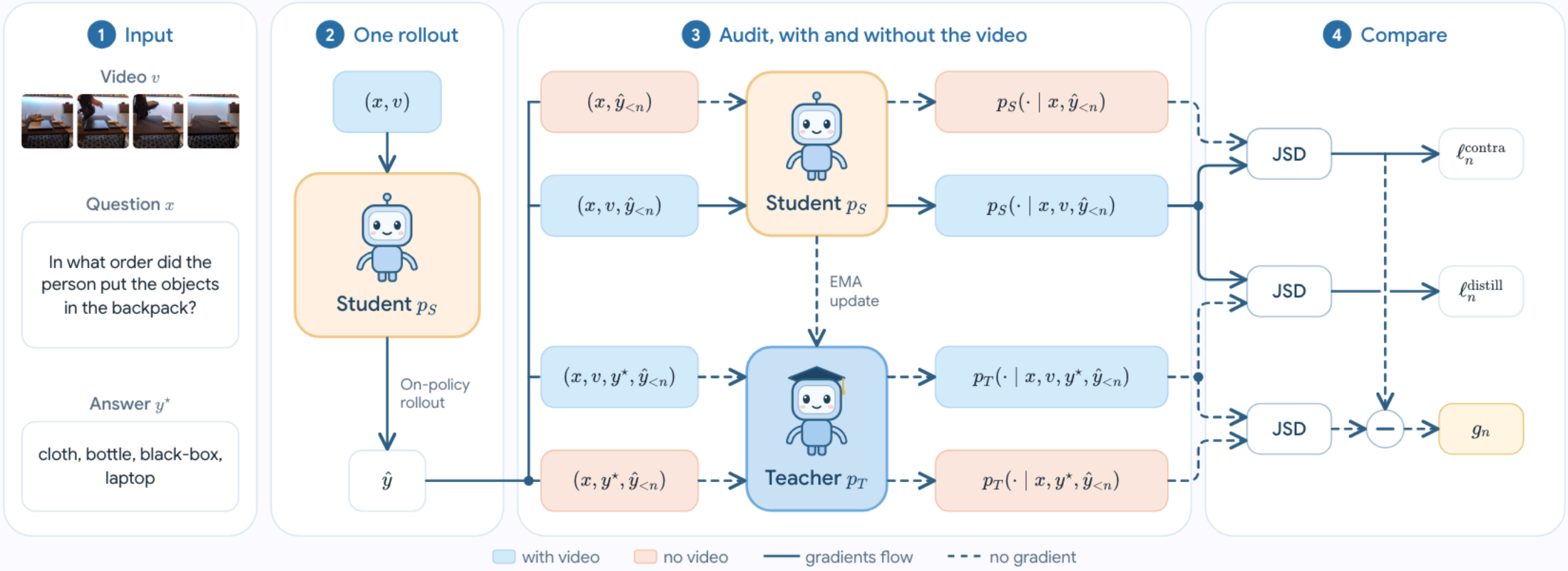
A model can audit its own visual reliance



$$r_n = \mathrm{JSD}\big(p(\cdot \mid x, v, \hat{y}_{<n}) \,\big\|\, p(\cdot \mid x, \hat{y}_{<n}) \big)$$

METHOD

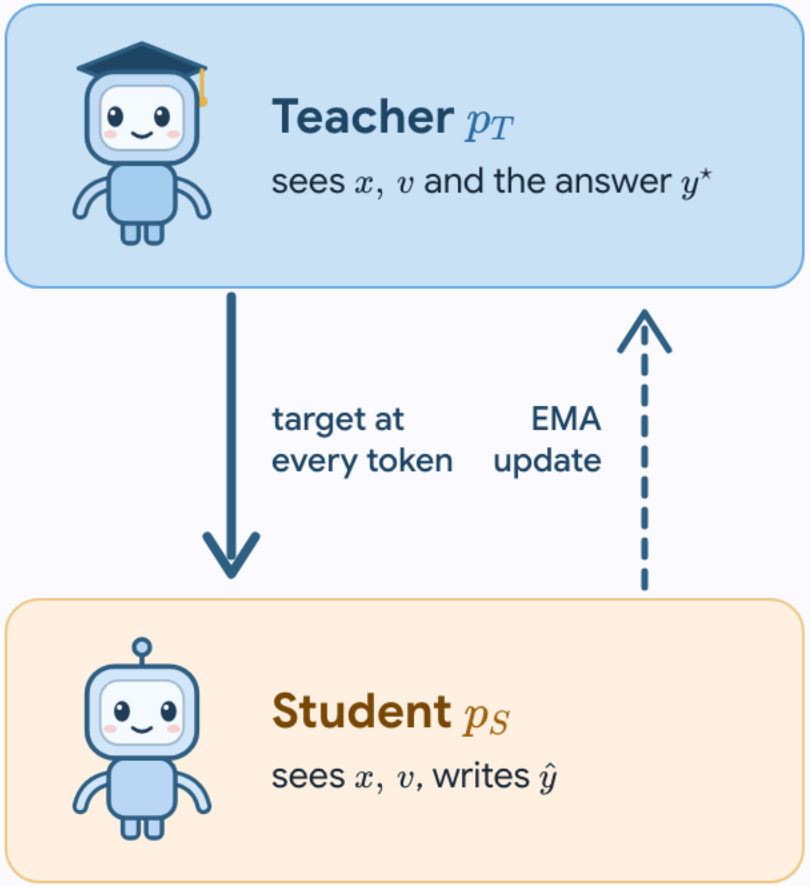
VIOS: Visually-aware On-policy Self-distillation



Per-token objective $\ell_n = g_n^\star \left(\ell_n^{\text{distill}} + \lambda \, \ell_n^{\text{contra}} \right)$

METHOD · 1 OF 3

Self-distillation toward a co-evolving teacher



$$\mathcal{L}_{\text{distill}} = \frac{1}{|\hat{y}|} \sum_{n=1}^{|\hat{y}|} \mathrm{JSD}\big(p_T(\cdot \mid x, v, y^\star, \hat{y}_{<n}) \,\big\|\, p_S(\cdot \mid x, v, \hat{y}_{<n})\big)$$

1. **Dense and on-policy:** a target at every token of the student's own rollout
2. **Jensen-Shannon divergence:** symmetric and bounded, between mode-covering and mode-seeking KL
3. **Co-evolving teacher:** an exponential moving average of the student, $\bar{\theta} \leftarrow \alpha\,\bar{\theta} + (1-\alpha)\,\theta$, so the target improves with the student

METHOD · 2 OF 3

Contrastive term: predictions must depend on the video

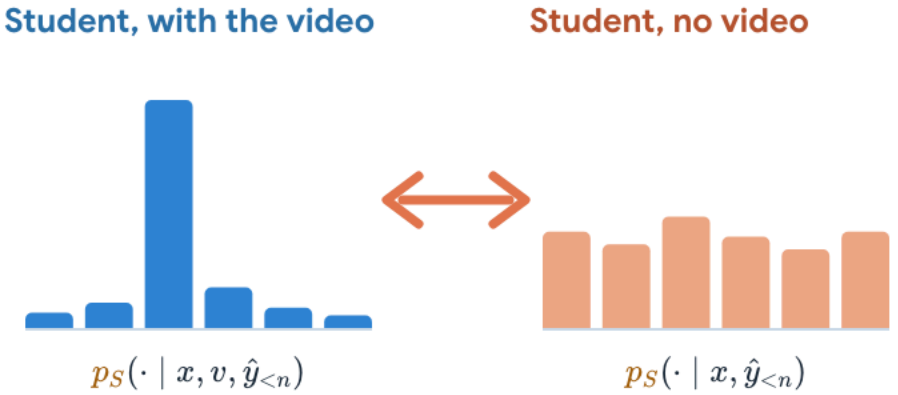
Matching the teacher ≠ using the video

Question: In what order did the person put the objects in the backpack?

Teacher also sees the answer: **cloth, bottle, black-box, laptop**



"...laptops usually go in last, so the laptop was packed last..."



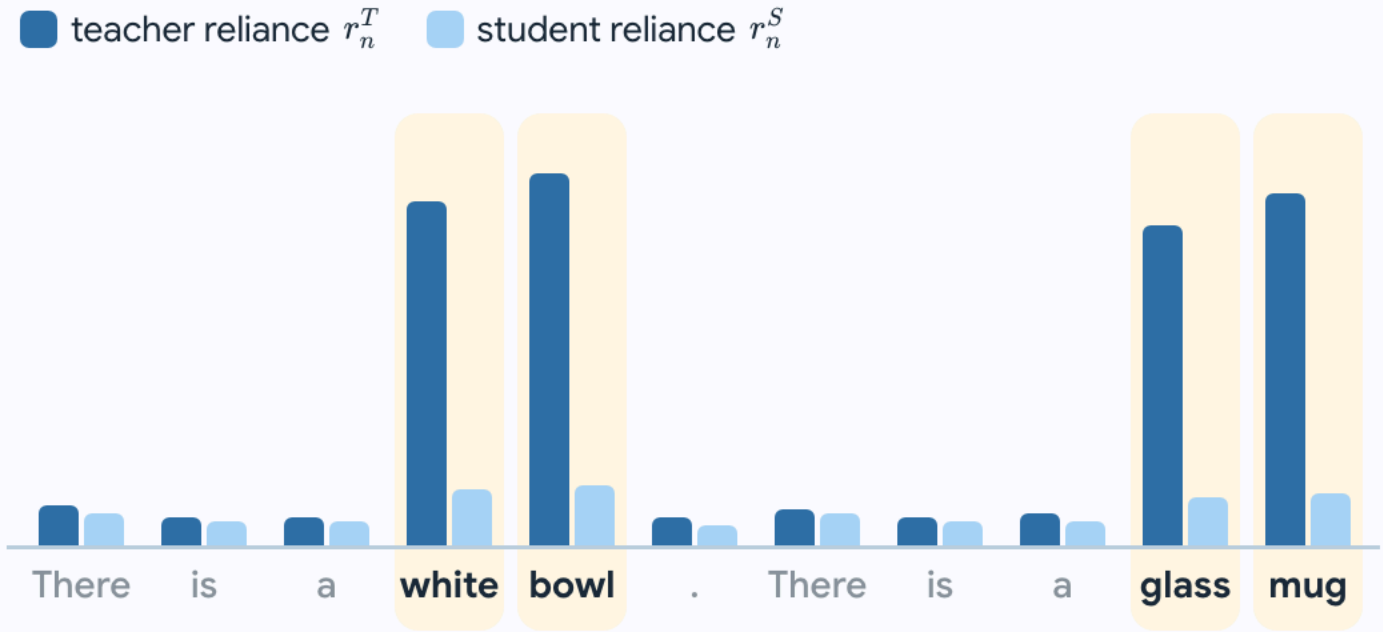
$$\mathcal{L}_{\text{contra}} = -\frac{1}{|\hat{y}|}\sum_{n=1}^{|\hat{y}|} \text{JSD}\left(p_S(\cdot \mid x, v, \hat{y}_{<n}) \,\middle\|\, \text{sg}\left[p_S(\cdot \mid x, \hat{y}_{<n})\right]\right)$$

Distillation: what to predict

Contrastive: predict it from the video

METHOD · 3 OF 3

Reliance-gap gate: focus where the video is under-used



$$g_n = r_n^T - r_n^S$$

Normalize along the rollout

$$\tilde{g}_n = \frac{g_n - \min_k g_k}{\max_k g_k - \min_k g_k}$$

$$g_n^\star = \tilde{g}_n \Big/ \Big(\frac{1}{|\hat{y}|} \sum_k \tilde{g}_k \Big)$$

Full objective

$$\mathcal{L}(\theta) = \mathbb{E}\Big[\frac{1}{|\hat{y}|} \sum_n g_n^\star \big(\ell_n^{\text{distill}} + \lambda\, \ell_n^{\text{contra}} \big) \Big]$$

METHOD

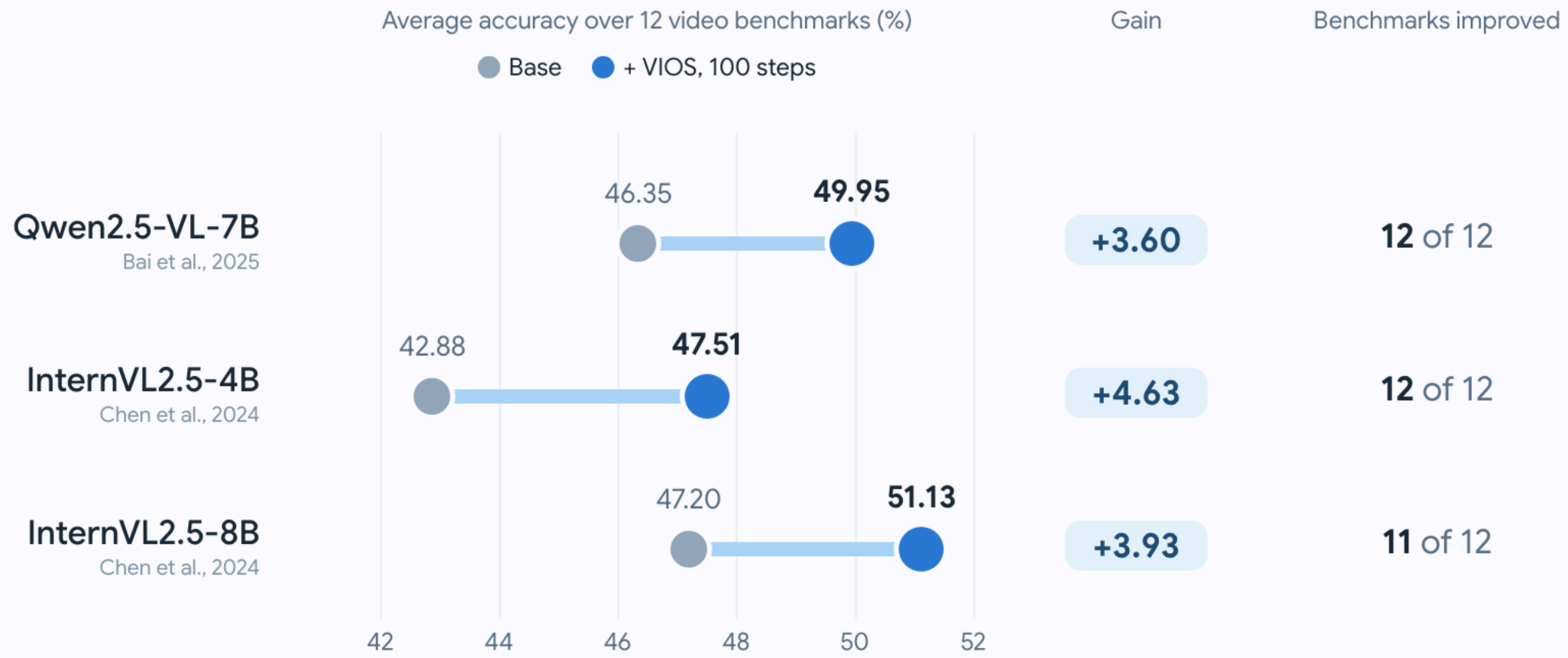
All from a single rollout of a single model

One rollout. One model.

- ✕ no external reward
- ✕ no judge
- ✕ no larger teacher

RESULTS

Across backbones



RESULTS

Main results

largest gain on each backbone

	General				Knowledge			Temporal		Long video		Spatial	
	MVBench	Video-MME	MINERVA	MMVU	Video-MMMU	VideoMath QA	SciVideo Bench	Temp Compass	TVBench	LongVideo Bench	MLVU	VSI-Bench	Avg.
Video-R1 Feng et al., 2025	64.42	58.37	30.76	63.04	**49.22**	23.57	24.50	71.58	50.81	55.42	**63.11**	35.25	49.17
VideoRFT Wang et al., 2025	62.25	58.26	28.26	64.96	47.44	24.76	24.10	72.65	**53.35**	54.08	61.82	35.10	48.92
VideoChat-R1 Li et al., 2025	62.40	57.07	29.47	63.68	47.67	23.33	23.30	72.43	52.16	53.85	57.82	33.08	48.02
VideoChat-R1.5 Yan et al., 2025	62.17	56.89	28.56	63.20	**49.22**	25.00	23.20	72.28	48.40	54.82	58.05	31.95	47.81
Video-R2 Maaz et al., 2025	62.08	56.04	27.42	64.80	48.33	20.00	**25.40**	72.64	51.68	50.19	58.00	32.56	47.43
Qwen2.5-VL-7B-SFT Bai et al., 2025; Feng et al., 2025	60.72	55.15	29.55	59.36	47.22	23.10	22.40	67.25	49.82	52.51	55.70	33.44	46.35
+ VIOS (ours)	**65.50**	**59.59**	**31.36**	**65.76**	47.44	**25.48**	24.90	**72.77**	52.24	**55.87**	62.74	**35.73**	**49.95**
Gain over base	+4.78	+4.44	+1.81	+6.40	+0.22	+2.38	+2.50	+5.52	+2.42	+3.36	+7.04	+2.29	+3.60
InternVL2.5-4B Chen et al., 2024	61.27	53.19	27.12	52.96	34.89	19.52	15.90	62.25	52.40	43.68	57.64	33.72	42.88
+ VIOS (ours)	**68.42**	**57.15**	**30.91**	**57.44**	**40.44**	**21.19**	**19.80**	**68.01**	**57.43**	**51.61**	**63.57**	**34.14**	**47.51**
Gain over base	+7.15	+3.96	+3.79	+4.48	+5.55	+1.67	+3.90	+5.76	+5.03	+7.93	+5.93	+0.42	+4.63
InternVL2.5-8B Chen et al., 2024	62.42	57.93	30.38	56.48	43.56	19.29	17.90	70.12	57.11	51.16	61.96	**38.04**	47.20
+ VIOS (ours)	**69.95**	**61.52**	**33.18**	**60.96**	**45.89**	**23.81**	**22.50**	**72.21**	**61.74**	**58.71**	**66.79**	36.33	**51.13**
Gain over base	+7.53	+3.59	+2.80	+4.48	+2.33	+4.52	+4.60	+2.09	+4.63	+7.55	+4.83	−1.71	+3.93

EFFICIENCY

Training efficiency

		General				Knowledge			Temporal		Long video		Spatial	
	Steps	MVBench	Video-MME	MINERVA	MMVU	Video-MMMU	VideoMath QA	SciVideo Bench	Temp Compass	TVBench	LongVideo Bench	MLVU	VSI-Bench	Avg.
Video-R1 Feng et al., 2025	1,000	64.42	58.37	30.76	63.04	**49.22**	23.57	24.50	71.58	50.81	**55.42**	**63.11**	35.25	49.17
VIOS† (ours) from Qwen2.5-VL-7B-SFT	110	**64.45**	**58.63**	**31.67**	**65.12**	46.33	**23.60**	**25.24**	**73.25**	**52.00**	54.00	61.68	**37.05**	**49.42**

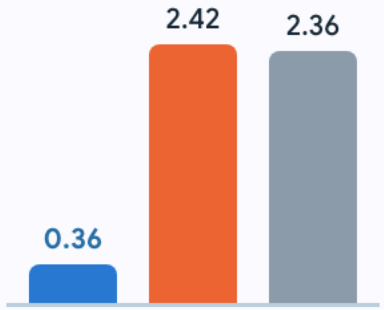
VIOS† Video-R1 GRPO

PER STEP

Training FLOPs

$\times 10^{15}$

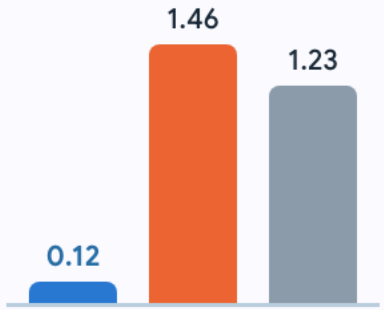
6.7× fewer



Generated tokens

$\times 10^{4}$

12× fewer

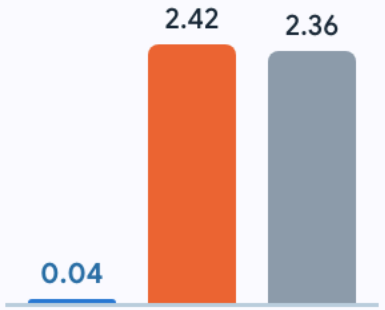


IN TOTAL

Training FLOPs

$\times 10^{18}$

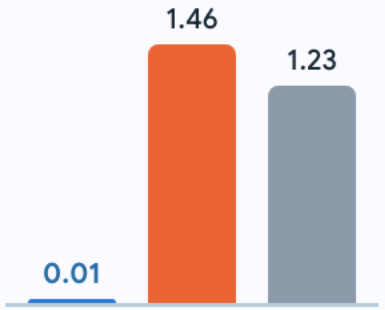
60× fewer



Generated tokens

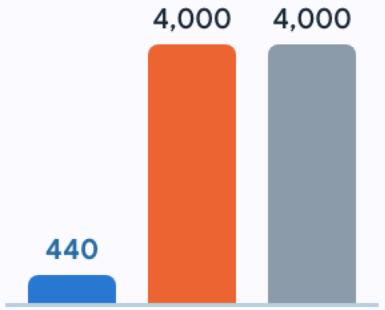
$\times 10^{7}$

111× fewer



Training samples

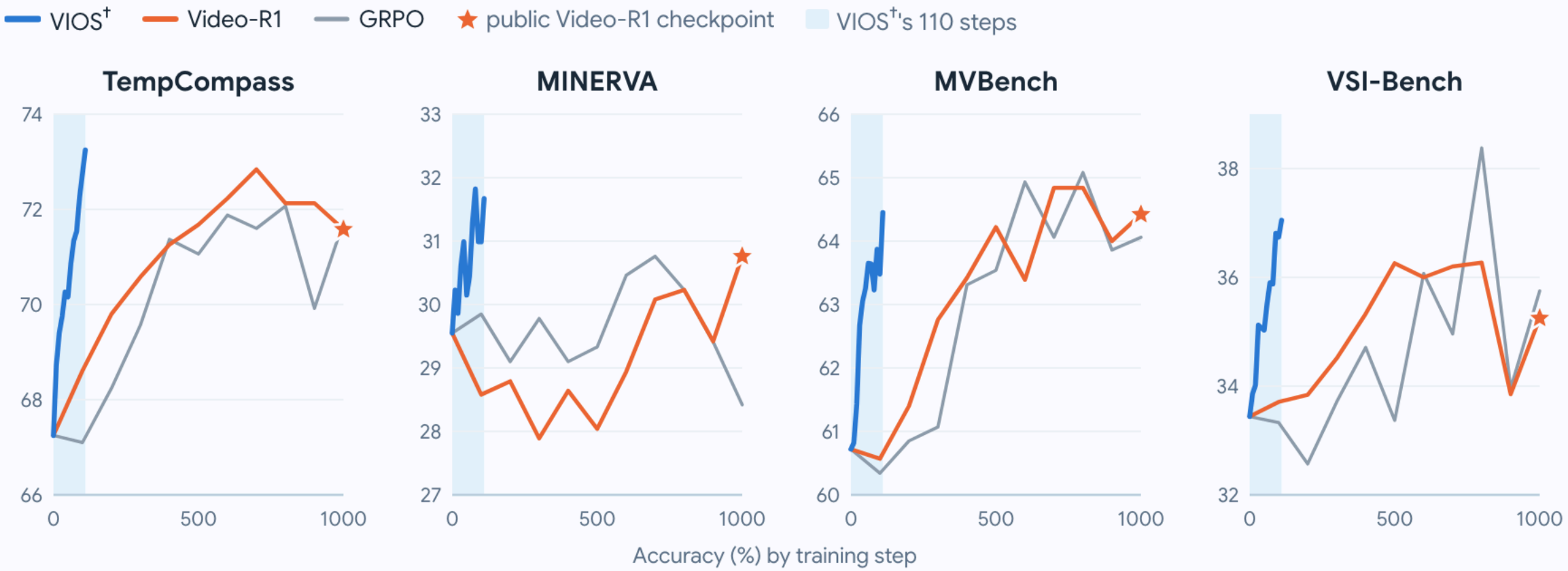
9.1× fewer



† VIOS trained under Video-R1's per-step training configuration.

EFFICIENCY

Training curves



† VIOS trained under Video-R1's per-step training configuration.

ANALYSIS

Qualitative example



How many objects were taken out from the backpack?

A. 0 B. 1 C. 2

Qwen2.5-VL-7B-SFT (base)

"...it seems that only the notebook and pen were taken out from the backpack. ... only two objects were taken out: the notebook and the pen."

C (2) ✗

VIOS · shaded by per-token training loss, darker is higher

The person in the video is packing a backpack, and I need to determine how many objects were taken out from it. First, let's observe the objects on the table: - There is a notebook with a pen. - There is a white bowl. - There is a glass mug. - There is a protractor. - There is a plastic sheet. ... From what I can see, the person is not taking anything out of the backpack; instead, they are putting items inside it.

A (0) ✓

OUTLOOK

Future work

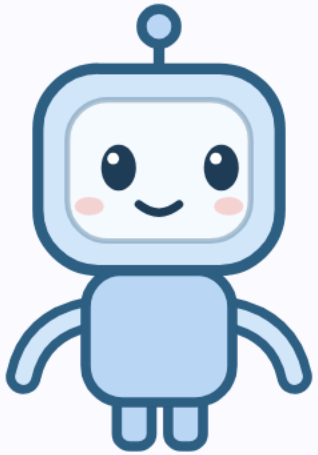
EXTENDING VIOS

- → Scaling to larger models
- → Open-ended video tasks, beyond ground-truth answers
- → Relaxing the need for a base model that already produces chains of thought

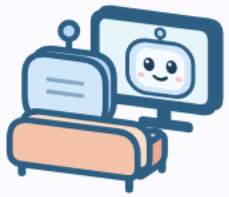
BEYOND VIDEO

- → Self-auditing other modalities: audio, depth, and sensor streams
- → Multimodal agents whose actions must depend on what they perceive
- → The reliance gap as a diagnostic: does a model actually use its inputs?

The supervision a model needs to use what it sees is already inside it.



It just has to watch itself watch.



Thank you

Watching Itself Watch: Self-Auditing Visual Reliance for Video Reasoning

vision.cs.utexas.edu/projects/VIOS

Rohan Dugad Mi Luo Kristen Grauman

The University of Texas at Austin · NeurIPS 2026