Behavior Pack Optimization for Video MLLM Post-Training

Zhaolu Kang[1,2*], Shiyu Liu[1*], Tailong Luo, Wei Zhang, Yingjie He[1], Guangyuan Dong, Lei Wei[1], Guansu Wang[3], Liang He, Siheng Wang, Jiaqi Su, Shuang Chen, Haoyu Ji, Qishi Zhan, Kaiyue Zhou[2†]

[1]Peking University [2]Chengdu Minto Tech [3]The University of Melbourne *Equal contribution †Corresponding author

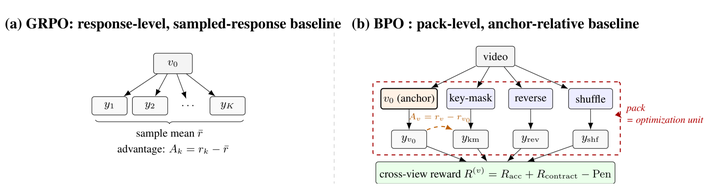
Paper
Abstract & paper

Email
zlkang25@stu.pku.edu.cn

TL;DR **Video MLLMs ignore the temporal evidence their answers should depend on. BPO trains on a behavior pack of counterfactual views and beats budget-matched GRPO by +4.7 pp accuracy and +20.0 pp abstention F1.**

1 Problem and idea

- Video MLLMs keep climbing QA benchmarks, yet **shuffling frames**, **masking the key segment** or **occluding the target** barely changes their answers.
- Accuracy rests on **appearance and language priors**, not on the temporal evidence the question asks for.
- Root cause: RL rewards grade **one response on one clip**, so the model is never asked to behave consistently across views.



GRPO centers same-clip rollouts on their mean; BPO scores one rollout per view against the anchor.

- **Idea**: a behavior pack holds the original clip (**anchor**) plus **question-typed counterfactual views**, decoded and **scored jointly**.
- **Not data augmentation**: the views share one optimization unit, so the same unchanged answer can be rewarded after an irrelevant edit and penalized after the evidence is removed.

2 Question type × view family

Trained on	Q1 mask	Q1 temp.	Q2 mask	Q2 temp.
Q1 + mask	**64.3**	**56.1**	**58.4**	**54.7**
Q1 + temporal	57.8	55.9	56.2	55.1
Q2 + mask	58.9	54.3	57.5	55.8
Q2 + temporal	**57.4**	**55.2**	**56.9**	**62.1**

Temporal-hard accuracy. Q1: event, binding, count; Q2: order, direction, trend.

- Mask views test **evidence presence**; reverse and shuffle test **temporal order**.
- Matched pairs win by **12.8 pp**, 3.2× the best mismatch.

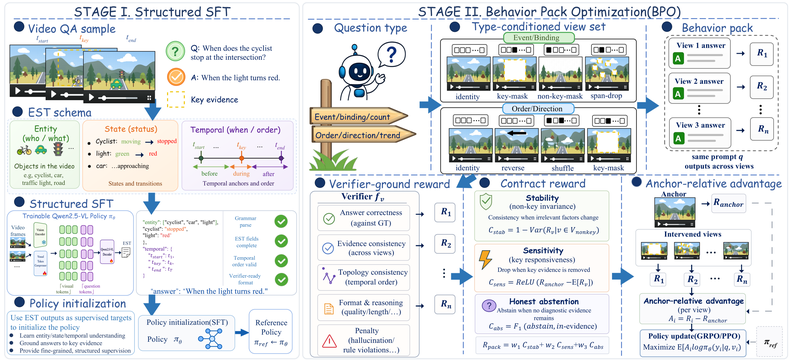
3 Ablations (3 seeds)

Remove	Macro	Temp.	Key-s.	Rev-s.	Abst.
nothing (full)	**67.4**	**61.2**	**54.6**	**48.9**	**0.67**
evidence reward	65.8	59.4	50.2	47.1	0.58
stability	66.1	59.7	52.8	47.5	0.65
sensitivity	65.4	57.3	**38.4**	**35.2**	0.55
abstention	66.2	60.1	53.1	47.8	**0.43**
learned verifier	64.1	56.8	46.7	43.5	0.52
type conditioning	65.8	58.9	48.3	44.7	0.58

Remove one component from full BPO.

- Each contract controls **its own behavior**: no sensitivity contract, key-sens 54.6 → 38.4 and rev-sens 48.9 → 35.2.
- No abstention contract: F1 drops by 0.24, so abstention is not a side effect of accuracy.
- A rule-only verifier still beats GRPO.

4 Behavior Pack Optimization



Stage I installs an Entity-State-Temporal (EST) trace by SFT; Stage II freezes the vision encoder and trains the LLM on behavior packs.

View	Expected behavior
non-key-mask	**Stability**: answer should not drift
key-mask, span-drop, reverse, shuffle	**Sensitivity**: change answer, flip order or lower confidence
evidence removed	**Honest abstention**

Behavior contracts relative to the anchor answer.

- **Question-typed views.** Event/binding/count: key-mask, non-key-mask, span-drop. Order/direction/trend: reverse, shuffle, key-mask. A rule classifier routes each question by its interrogative template and verb frame.
- **Verifier reward.** Answer correctness, evidence support, EST consistency and format, plus the contract term; a penalty catches parse failures and uniform confident answers across views.
- **Anchor-relative advantage.** Each view is compared with the original-view rollout instead of a mixed group mean. The anchor answer itself gets no RL term, so it cannot move the reference it provides.

$$\hat{A}^{(v)} = R_{\text{view}}(v, y_v;\, y_{v_0}) - R_{\text{view}}(v_0, y_{v_0};\, y_{v_0})$$

The anchor is graded against the ground truth, so it stays grounded.

5 Why the anchor helps

$$\text{Var}[A_{\text{anchor}}] = 2\sigma^2(1-\rho) \;<\; \frac{\sigma^2}{m}[1+(m-1)\rho] = \text{Var}[A_{\text{batch}}] \;\Leftrightarrow\; \rho > \frac{2m-1}{3m-1}$$

ρ: correlation between view rewards of one prompt; m: views per pack.

- Packs share prompt difficulty, so ρ is high: measured **0.83** vs. the threshold 0.64 for m = 4.
- Same packs and rewards: advantage variance drops **0.341 → 0.142** and the early KL spikes disappear.
- **Robust to reward noise**: 10% label flips in non-anchor rewards cost 0.7 pp (batch-norm: 3.2 pp).

6 Results (Qwen2.5-VL-7B, 5 seeds)

Setting	Macro	Temp.-hard	Key-sens	Abst.-F1
B0 zero-shot	55.8	48.2	12.4	0.31
B1 + EST-SFT	58.1	50.6	13.1	0.34
B2 vanilla GRPO	62.7	53.4	31.5	0.47
B2.5 Video-R1	63.4	54.8	35.2	0.49
B2.6 TPO	63.1	54.2	33.8	0.51
B3 pack, batch-norm	64.5	56.1	42.7	0.55
B3 single-view control	62.9	53.7	32.1	0.47
B4 + anchor-relative	65.8	58.9	48.3	0.58
B5 BPO (full)	**67.4**	**61.2**	**54.6**	**0.67**

TempCompass, MVBench, NExT-QA. RL rows: 4 rollouts per prompt, FLOPs within 5%.

+4.7 pp macro accuracy

+7.8 pp temporal-hard

+20.0 pp abstention F1

- **Reading the ladder**: EST-SFT alone barely changes behavior; packs (B3) add 11.2 pp key-sensitivity; the anchor (B4) lifts accuracy and behavior; typed views (B5) add 2.3 pp temporal-hard.

GRPO → BPO	Macro	Temp.-hard
Video-MME	54.3 → 58.1	46.8 → 52.9
LongVideoBench	46.2 → 49.8	41.5 → 47.3
LLaVA-Video-7B	61.3 → 65.9	51.7 → 58.3

Same budget; LLaVA-Video-7B swaps the backbone.

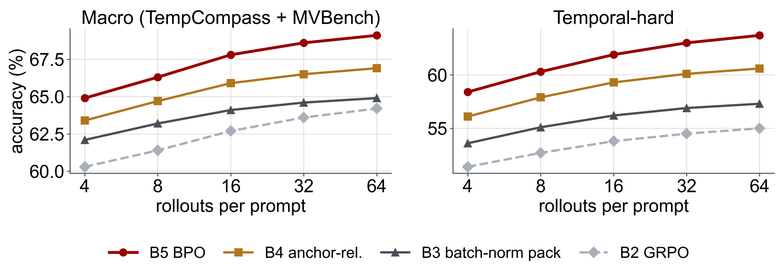
- **3B**: +4.1 macro, +4.9 temporal-hard.
- **Router**: oracle vs. predicted question types differ by 1.4 pp.
- **No test leakage**: benchmark answers never drive masks, spans or routing.

View set	Views	Macro	Q1-hard	Q2-hard
minimum	3	65.4	57.2	55.8
mask family	4	66.8	**61.4**	56.9
temporal family	4	66.5	57.8	**60.7**
full oracle	6	67.6	62.1	60.9

Matched compute. Minimum: anchor, key-mask, reverse.

- Each 4-view family comes within **0.7 pp** of the 6-view oracle on its own question type.

7 Budget frontier and checks



Rollout budget from 4 to 64 per prompt; macro over TempCompass and MVBench.

- BPO with **4 rollouts** (64.9) beats GRPO with **64 rollouts** (64.2): more same-view samples do not close the gap.
- **Verifier audit**: 0.93 accuracy on 200 human-labeled samples; reward hacking below 2.3%.

zlkang25@stu.pku.edu.cn