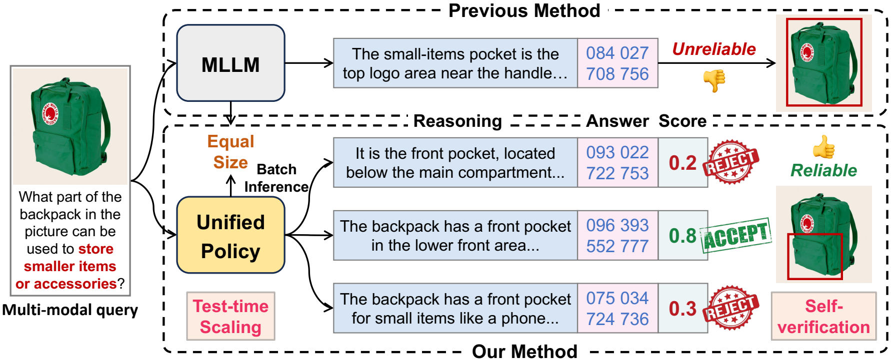
Unified Generation and Self-Verification for VLMs via Advantage Decoupled Preference Optimization

Motivation & Overview: Parallel Test-Time Scaling

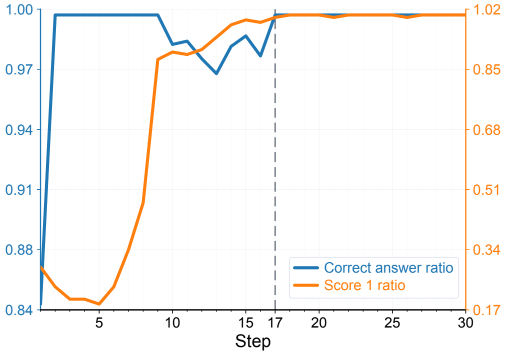
浙江大学 ZHEJIANG UNIVERSITY

- **Motivation**
  - Serial scaling (o1, R1): long chains, small multimodal gains.
  - Parallel scaling: best-of-N works but needs separate generator $+$ verifier.
  - Pain: two models $\Rightarrow$ double data, training, and inference cost.
  - Goal: one policy that generates and self-verifies for best-of-N.
- **Method: Advantage Decoupled Preference Optimization (ADPO)**
  - Unified framework: one policy that both generates answers and self-verifies.
  - Preference verification reward: cast verification as ranking to stay informative under severe class imbalance.
  - Advantage decoupled optimization: separate advantages and token masks to avoid reward hacking and disentangle generation vs. verification.
- **Contributions**
  - Preference verification reward
  - Advantage decoupled optimization
  - Comprehensive evaluation: $+2.8/+1.4$ acc. on MathVista/MMMU, $+1.9$ cIoU on ReasonSeg, and $+1.7/+1.0$ step success on AndroidControl/GUI Odyssey.

ADPO Overview



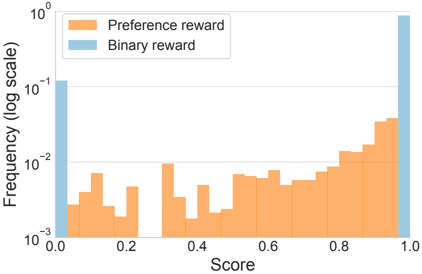
- Extends GRPO to unify answer generation and self-verification within one policy.
- Model outputs an answer plus a verification score $s \in [0, 1]$ per query.
- Inference: batch decode multiple candidates; pick the answer with the highest self-score.



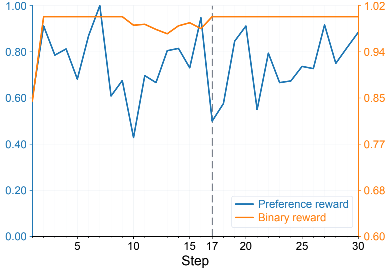
Domain	Entangled	GRPO	Gap
MathVista	61.5	62.2	-0.7
ReasonSeg	69.5	71.1	-1.6

- **Class imbalance**: binary verification reward collapses as more answers become correct; scores drift toward a single value, killing gradients.
- **Reward hacking**: summing answer and verification rewards lets the model output bad answers with low self-scores yet still get high total reward.

Preference Verification Reward



Score distribution collapse without preference reward



Positive/negative ratio imbalance over training

- Reframes verification as a ranking task to avoid collapse under imbalance.
- For sample i, contrastive set $\mathcal{C}_i = \{j \mid R_j^a \neq R_i^a\}$ (or $|R_j^a - R_i^a| > \gamma$ for continuous tasks).
- Reward:

$$R_i^p = \frac{1}{\max(|\mathcal{C}_i|, 1)} \sum_{j \in \mathcal{C}_i} \mathbf{1}\{(s_i - s_j)(R_i^a - R_j^a) > 0\}.$$

- Encourages higher scores for better answers and lower scores for worse ones; works for discrete and continuous rewards.

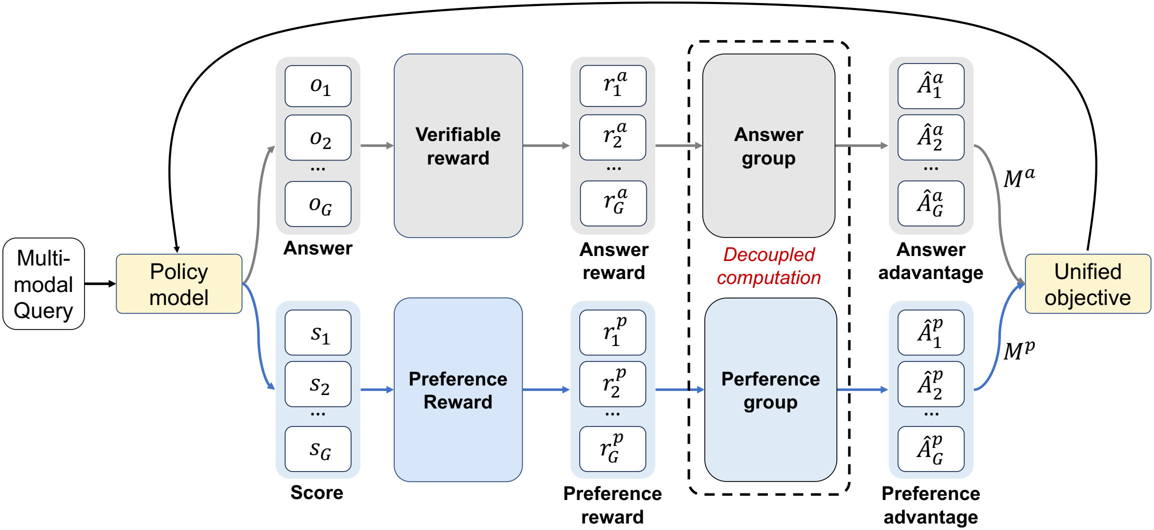
Advantage-Decoupled Optimization

- **Entangled advantage**
  Sum rewards $R^a + R^p$ and compute one advantage over all tokens.
  Verification gradients leak into generation tokens, enabling reward hacking: bad answers + low self-scores can still get positive total signal.
- **Decoupled advantage**
  Compute two advantages $\hat{A}^{(a)}$ (answer) and $\hat{A}^{(p)}$ (verification).
- **Unified training objective**
  Apply disjoint masks: M^a on answer tokens, M^p on the score token.

$$\mathcal{J}(\theta) = M^a \odot \mathcal{J}_\theta(\hat{A}^{(a)}) \; + \; M^p \odot \mathcal{J}_\theta(\hat{A}^{(p)}).$$

- **Effect**
  Isolates gradients, preserves pass@1 quality, and calibrates self-scores for best-of-N.

Training Pipeline



Main Results: Math & Grounding

Table 2. **Performance on MathVista [22] and MMMU [44].** We adopt Qwen2-VL-7B [34] as the base model and use majority voting for both the base and GRPO models. We report accuracy (%) and highlight the best results in **bold**.

Method	MathVista (In-domain)			MMMU (OOD)						
	GVQA	MVQA	ALL	ARD	BUS	HEM	HSS	SCI	TEN	ALL
Sample 1										
R1-VL-7B [46]	-	-	63.5	-	-	-	-	-	-	-
Base	68.9	48.5	57.9	**67.5**	39.1	49.3	69.0	33.9	36.7	47.1
GRPO	**69.8**	55.7	62.2	65.0	45.9	48.2	68.2	**35.9**	**39.8**	**48.7**
ADPO	68.7	**57.0**	**62.4**	63.1	**46.2**	**50.2**	**71.1**	33.3	35.3	47.7
Sample 4										
MM-Verifier [33]	67.0	53.7	59.8	-	-	-	-	-	-	-
Base	65.7	51.9	58.2	66.7	47.3	50.7	65.8	34.0	38.1	48.6
GRPO	69.8	58.0	63.4	65.8	44.7	50.0	**70.0**	**42.0**	36.7	49.4
ADPO	**71.3**	**59.3**	**64.8**	**68.3**	**48.0**	**52.0**	69.2	39.3	**39.5**	**50.8**
Sample 8										
MM-Verifier [33]	68.5	57.4	62.5	-	-	-	-	-	-	-
Base	68.0	53.3	60.1	**68.3**	50.0	53.3	68.3	32.7	36.7	49.4
GRPO	70.4	56.5	62.9	66.7	48.7	51.3	**74.2**	**42.7**	36.7	51.1
ADPO	**72.2**	**58.9**	**65.0**	65.8	**54.0**	**54.7**	66.7	40.7	**41.0**	**52.1**
Sample 12										
MM-Verifier [33]	70.4	58.7	64.1	-	-	-	-	-	-	-
Base	67.4	55.0	60.7	**69.2**	52.0	50.7	70.8	38.0	36.7	50.7
GRPO	70.7	57.2	63.4	64.2	50.0	51.3	73.3	**43.3**	39.5	51.7
ADPO	**71.7**	**59.8**	**65.3**	67.5	**53.3**	**54.0**	**71.7**	38.7	**40.5**	**52.3**

Table 3. **Performance on ReasonSeg [11].** We use Qwen2.5-VL-7B [1] as the base model.

Method	Short query			Long query			Overall		
	gIoU	cIoU	ACC	gIoU	cIoU	ACC	gIoU	cIoU	ACC
Sample 1									
LISA-7B [11]	47.1	48.5	-	49.2	48.9	-	48.7	48.8	-
SegLLM [37]	-	-	-	-	54.2	-	-	48.4	-
Seg-Zero-7B [18]	-	-	-	-	-	-	57.5	52.0	-
VLM-R1 [32]	-	-	-	-	-	-	-	-	63.1
Base	49.5	53.0	67.0	56.8	57.5	68.5	56.3	57.2	68.4
GRPO	51.8	**55.5**	67.9	59.1	**59.7**	71.3	**58.6**	**59.5**	71.1
ADPO	**51.7**	54.8	**68.0**	**60.2**	59.4	**71.9**	58.1	59.1	**71.7**
Sample 4									
Base	47.8	52.0	66.0	57.3	57.9	69.3	56.7	57.5	69.1
GRPO	**54.5**	**57.0**	**68.0**	58.8	59.5	72.1	58.5	59.4	71.8
ADPO	52.2	55.1	67.0	**61.0**	**61.5**	**73.3**	**60.5**	**61.1**	**72.9**
Sample 8									
Base	47.8	51.4	63.1	57.2	57.8	69.2	56.6	57.4	68.8
GRPO	52.0	55.6	**68.0**	59.2	59.9	72.0	58.7	59.6	71.7
ADPO	**53.2**	**56.0**	67.0	**60.9**	**61.5**	**73.7**	**60.4**	**61.2**	**73.5**
Sample 12									
Base	50.2	53.7	66.0	57.2	57.8	69.3	56.8	57.6	69.1
GRPO	**55.6**	**58.1**	**69.9**	58.8	59.5	72.2	58.6	59.4	72.0
ADPO	53.9	56.2	67.0	**61.3**	**62.0**	**73.6**	**60.9**	**61.6**	**73.2**

Main Results: Agents & Verifiers

浙江大學 ZHEJIANG UNIVERSITY

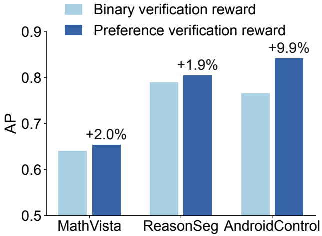
Table 4. **Performance on AndroidControl [13] and GUI Odyssey [23].** We adopt Qwen2.5-VL-7B as base model and report type accuracy, grounding accuracy and step success rate (SR).

Generator	AndroidControl			GUI Odyssey		
	Type	Grounding	**SR**	Type	Grounding	**SR**
Sample 1						
UI-TARS-7B [29]	83.7	-	72.5	86.1	-	67.9
SpiritSight-8B [10]	-	-	68.1	-	-	75.8
AgentCPM-GUI-8B [49]	77.7	-	69.2	90.8	-	75.0
Base	82.2	73.6	61.3	81.1	61.4	52.8
GRPO	**86.0**	**76.9**	**71.0**	93.1	**83.9**	**79.8**
ADPO	85.8	76.2	70.9	**94.2**	82.5	79.7
Sample 4						
Base	76.3	68.1	56.0	76.9	55.3	46.5
GRPO	85.5	77.2	71.0	94.7	83.9	81.3
ADPO	**86.3**	**79.5**	**72.7**	**94.7**	**84.5**	**81.6**
Sample 8						
Base	78.7	68.8	58.3	76.7	55.4	46.6
GRPO	85.6	76.9	70.8	94.6	84.4	81.5
ADPO	**86.4**	**78.7**	**72.7**	**94.8**	**84.7**	**81.7**
Sample 12						
Base	78.9	68.7	58.3	76.9	55.5	46.9
GRPO	85.6	77.4	71.1	**94.5**	84.0	81.1
ADPO	**86.3**	**78.9**	**72.9**	94.4	**84.5**	**81.4**

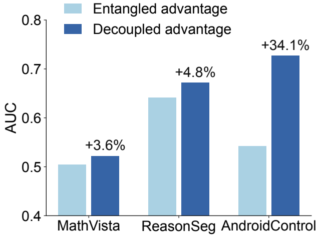
Table 5. **Performance of different generator–verifier settings on MathVista [22], ReasonSeg [11] and AndroidControl [13].**

Generator \ Verifier	MathVista			ReasonSeg			AndroidControl		
	Base	GRPO	ADPO	Base	GRPO	ADPO	Base	GRPO	ADPO
Sample 4									
Base	55.7	55.5	56.4	57.1	57.7	57.7	52.5	57.7	60.7
GRPO	62.4	62.1	62.0	60.2	59.5	60.9	71.0	70.8	71.2
ADPO	61.5	62.1	**64.8**	59.6	60.3	**61.1**	71.0	72.0	**72.7**
Sample 8									
Base	57.0	56.4	56.5	56.9	57.0	57.9	54.3	61.0	64.7
GRPO	60.7	60.8	60.5	60.4	60.4	61.1	71.0	70.9	71.4
ADPO	62.3	62.3	**65.0**	59.9	60.5	**61.2**	70.8	71.4	**72.7**
Sample 12									
Base	56.9	56.3	55.0	57.4	57.6	57.8	53.6	60.7	64.5
GRPO	62.5	62.5	61.8	59.7	59.6	61.3	71.4	70.9	71.5
ADPO	63.0	63.5	**65.3**	60.7	60.7	**61.6**	71.6	71.9	**72.9**

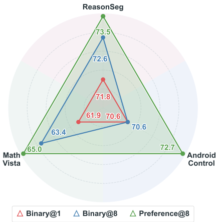
Ablations: Reward and Advantage



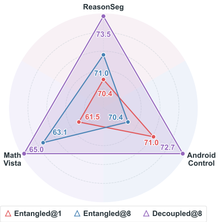
AP improvement (preference reward)



AUC improvement (decoupled advantage)



Preference reward vs. binary reward



Decoupled vs. entangled advantage

Table 6. **Ablation of the margin γ for preference verification reward on ReasonSeg.**

γ	Short query			Long query			Overall		
	gIoU	cIoU	ACC	gIoU	cIoU	ACC	gIoU	cIoU	ACC
0.025	**53.7**	56.5	**69.9**	58.1	58.9	71.3	57.8	58.8	71.1
0.050	52.6	54.4	63.1	60.2	61.0	73.3	59.8	60.5	72.7
0.100	53.2	56.0	67.0	**60.9**	**61.5**	**73.7**	**60.4**	**61.2**	**73.5**
0.200	53.2	55.7	66.0	59.9	60.7	72.7	59.6	60.4	72.3
0.250	**53.7**	**56.8**	68.9	59.7	60.4	72.5	59.3	60.2	72.3

Table 7. **Comparison of unified and separate verification.** *GRPO:* GRPO post-trained model as generator. +*Major:* majority voting as verifier. +*Judge:* GRPO post-trained model as verifier.

Method	MathVista Acc. ↑	Latency (s) ↓
GRPO+Major	62.9	**2.1**
GRPO+Judge	60.8	5.6
ADPO	**65.0**	2.6

Takeaways

- ADPO unifies generation and verification, enabling reliable parallel test-time scaling with one policy.
- Preference Verification Reward delivers stable, informative gradients under severe class imbalance.
- Advantage Decoupled Optimization isolates gradients and prevents reward hacking while preserving pass@1 quality.
- Stronger best-of-N performance and better-calibrated self-scores across math reasoning, visual grounding, and mobile agents with lower deployment overhead.