Fast-Slow Thinking GRPO for Large Vision-Language Model Reasoning

Balancing Reasoning Length and Accuracy in LVLMs

Wenyi Xiao Leilei Gan

School of Software Technology
Zhejiang University

NeurIPS 2025

Outline

Background: Rise of Slow-Thinking Reasoning

- **Slow-thinking models** show remarkable capabilities
  - OpenAI o1, DeepSeek-R1, Qwen QwQ
  - Solve complex tasks through deliberate reasoning
- **Slow-thinking in LVLMs**
  - SFT-RL two-stage methods
  - RL-only methods
- **Key Challenges**
  - Limited reasoning length scaling (-20% to +10%)
  - Overthinking phenomenon
  - Marginal accuracy improvements

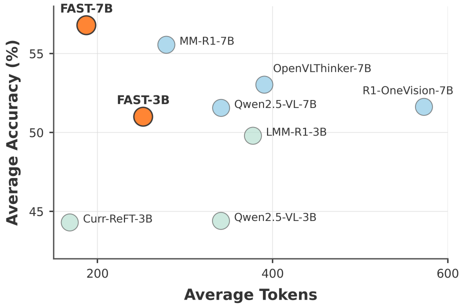

Figure: FAST achieves higher accuracy with shorter reasoning

Problem: Overthinking Phenomenon

Table: Comparison of accuracy and response length on Geometry 3K test set

Test	Qwen2.5-VL		R1-OneVision		FAST	
	Acc.	Len.	Acc.	Len.	Acc.	Len.
Easy	72.7	318	69.5	623	**78.2**	**189**
Med	33.9	406	40.4	661	**49.2**	**220**
Hard	5.5	412	10.2	835	**12.3**	**304**
All	37.7	378	40.3	731	**46.4**	**239**

Key Findings

- R1-OneVision produces $2\times$ longer reasoning chains
- Overthinking on simple questions degrades accuracy (69.5% vs. 72.7%)
- Need for adaptive fast-slow thinking mechanism

Length Rewards Analysis

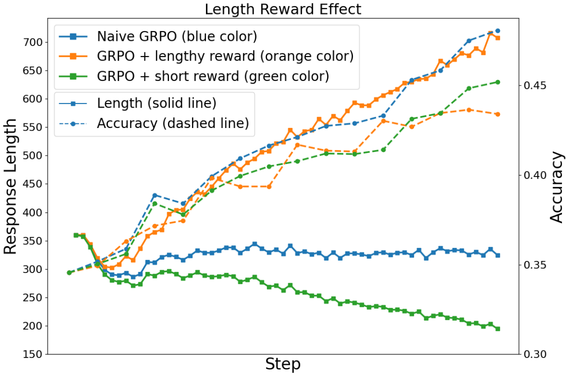

Figure: Effect of length rewards on reasoning

Experimental Setup

- Base model: Qwen2.5-VL
- Dataset: Geometry 3K
- Three strategies:
  - GRPO + lengthy reward
  - GRPO + short reward
  - Naive GRPO

Observation 1

LVLMs can produce significantly different reasoning lengths (180-700 tokens) via length rewards with modest accuracy changes ($\pm 3\%$)

Data Distribution Analysis

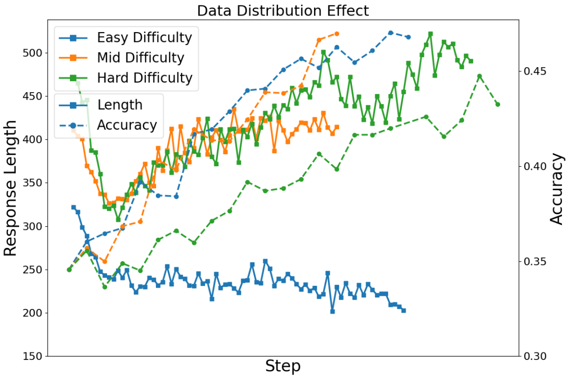

Figure: Effect of data distribution on reasoning

Difficulty Stratification

- Easy: $0.75 \leq$ passrate@8
- Medium: $0.25 <$ passrate@8 < 0.75
- Hard: passrate@8 ≤ 0.25

Observation 2

Question difficulty acts as implicit regulator of reasoning length, suggesting strategic data distribution for adaptive thinking

Method Overview

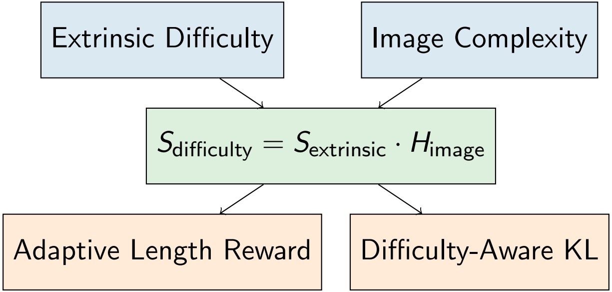

Figure: FAST-GRPO Framework

Core Components

1. **Multimodal Question Difficulty Estimation**: Extrinsic + Intrinsic
2. **Slow-to-Fast Sampling**: Dynamic training data distribution
3. **FAST-GRPO Algorithm**: Adaptive rewards + difficulty-aware regularization

Difficulty Estimation

Extrinsic Difficulty

$$S_{\text{extrinsic}} = 1 - \text{passrate@k}$$

- Reflects model's current capability
- Computed online during training

Intrinsic Difficulty

$$H_{\text{image}} = -\frac{1}{P}\sum_{p=1}^{P} H(g_p) - H(v)$$

- GLCM entropy: texture complexity
- ViT entropy: semantic complexity

Combined Difficulty Metric

$$S_{\text{difficulty}} = S_{\text{extrinsic}} \cdot H_{\text{image}}$$

Slow-to-Fast Sampling

- **Early Epochs**: Exclude easy samples ($S_{\text{extrinsic}} \leq 0.25$)
- **Later Epochs**: Exclude hard samples ($S_{\text{extrinsic}} \geq 0.75$)

First develop slow thinking, then learn adaptive fast thinking

FAST-GRPO Algorithm

Adaptive Length Reward

$$r_t = \begin{cases} 1 - \frac{L}{L_{\text{avg}}} & \text{if } S_d < \theta, r_a = 1 \\ \min(\frac{L}{L_{\text{avg}}} - 1, 1) & \text{if } S_d \geq \theta, r_a = 0 \\ 0 & \text{otherwise} \end{cases}$$

- Encourage brevity for simple correct
- Encourage detail for complex incorrect
- Cap reward at 1 to prevent verbosity

Difficulty-Aware KL Regularization

$$\beta_d = \beta_{\min} + (\beta_{\max} - \beta_{\min})(1 - S_{\text{ext}})$$

- High difficulty: $\beta_d \to \beta_{\min}$ (explore)
- Low difficulty: $\beta_d \to \beta_{\max}$ (exploit)

Gradient Coefficient

$$GC = \hat{A}_i + \beta_d \left(\frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)} - 1 \right)$$

Main Results: Reasoning Accuracy

Table: Accuracy comparison across 7 benchmarks (See full results in Paper)

Method	MathVis.	MathVer.	MathVista	MM-Math	WeMath	DynaMath	MM-Vet	Avg.
GPT-4o	30.4	49.9	63.8	31.8	69.0	63.7	80.8	55.6
Claude-3.5	37.9	46.3	67.7	–	–	64.8	68.7	–
Qwen2.5-VL-7B	25.6	46.9	68.2	34.1	61.0	58.0	67.1	51.6
R1-OneVision	29.9	46.4	64.1	34.1	61.8	53.5	71.6	51.6
OpenVLThinker	29.6	47.9	70.2	33.1	64.5	57.4	68.5	53.0
FAST-7B	**30.6**	**50.6**	**73.8**	**44.3**	**68.8**	**58.3**	**71.2**	**56.8**

Key Achievements

- SOTA on MathVista: 73.8 (surpassing GPT-4o)
- 10%+ average improvement over base model
- Strong performance on challenging benchmarks

Main Results: Reasoning Length (See full results in Paper)

Table: Average reasoning length (tokens) across 7 benchmarks

Method	MathVis.	MathVer.	MathVista	MM-Math	WeMath	DynaMath	MM-Vet	Avg.
Qwen2.5-VL-7B	340	378	318	412	356	324	298	346
R1-OneVision	689	731	623	835	756	712	680	718
OpenVLThinker	402	415	389	456	420	398	375	408
MM-R1	512	556	489	623	578	534	502	542
FAST-7B	**189**	**239**	**175**	**304**	**256**	**220**	**195**	**225**

Key Findings

- **67.3% reduction** compared to R1-OneVision
- **Adaptive reasoning**: Harder problems get more tokens automatically
- **Efficiency gain**: Better accuracy with shorter responses

Ablation Study

Table: Component contributions

Model	MathVista	MathV.	MathVer.	Len.
Qwen-2.5-VL-7B	68.2	25.6	46.9	340
FAST	**73.8**	**30.6**	**50.6**	**176**
w/o Data Samp.	69.9	27.2	48.4	257
w/o Think. Rew.	73.6	31.5	45.9	302
w/o Diff. Aware	72.0	29.5	49.2	172
Naive GRPO	67.2	25.3	47.6	205

Key Findings

- **Data Sampling**: Critical for all benchmarks
- **Thinking Reward**: 42% relative length reduction
- **Difficulty-Aware**: 1.8 points gain on MathVista

Sampling Strategy Effect

- Fast-to-Slow: Performance degradation
- Dynamic: 80% longer responses
- Slow-to-Fast: Optimal balance

Error Analysis

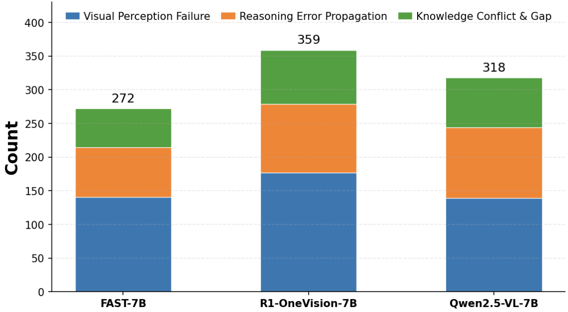

Figure: Error type distribution

Three Main Error Types

1. **Visual Perception Failures** (~50%)
   - Incorrect visual cue extraction
   - Spatial relation misinterpretation
2. **Reasoning Error Propagation** (~27%)
   - Mid-chain mistakes
   - Logic chain breakage
3. **Knowledge Conflict & Gap** (~19%)
   - Language priors override visual evidence
   - Domain knowledge insufficiency

Case Study: Visual Perception Failure (More cases & insights in Paper)

Figure: Angle measurement problem

Problem

Find the angle ∠ABC in the figure

Model Responses

- **Ground Truth**: 50°
- **Base Model**: "I see 40°" (incorrect reading)
- **R1-OneVision**: Long reasoning but misread angle
- **FAST**: Correctly identifies 50°

Key Issue

Visual perception errors propagate through entire reasoning chain

Main Contributions

1. **Identified and analyzed overthinking in LVLMs**
   - First systematic study of LVLM reasoning length
   - Revealed decoupling between length and accuracy
2. **Proposed FAST-GRPO framework**
   - Multimodal question difficulty estimation
   - Adaptive fast-slow thinking mechanism
   - Difficulty-aware reinforcement learning
3. **Achieved dual improvements**
   - 10%+ accuracy improvement
   - 32.7-67.3% reasoning length reduction
   - SOTA on multiple benchmarks

Limitations and Future Work

Current Limitations

- Computational constraints: Evaluated up to 32B parameters
- Visual perception bottleneck: 50%+ errors from visual misinterpretation
- Data scale: 18K training samples relatively small

Future Directions

1. **Scale to larger models** (70B+ parameters)
2. **Improve visual perception**
   - Fine-grained OCR
   - Accurate chart value extraction
   - Robust spatial grounding
3. **Explore other modalities** (audio, video)
4. **Online learning and continuous adaptation**

Key Insights

Adaptive Fast-Slow Thinking: The Future of LVLM Reasoning

Traditional Approach

- **One-size-fits-all** reasoning
- Fixed reasoning depth
- Either too brief or too verbose
- Inefficient resource utilization

FAST Approach

- **Problem-aware** reasoning
- Adaptive depth based on difficulty
- Simple $\rightarrow$ Fast thinking
- Complex $\rightarrow$ Slow thinking

Why This Matters

1. **Efficiency**: 67% shorter responses with better accuracy
2. **Intelligence**: Mimics human cognitive patterns
3. **Scalability**: Better resource allocation for real applications

Thank You!

Questions & Discussion

Code & Models: `https://github.com/Mr-Loevan/FAST`

Contact: wenyixiao@zju.edu.cn
leileigan@zju.edu.cn