From Seeing to Thinking

Decoupling Perception and Reasoning Improves Post-Training of Vision-Language Models

Juncheng Wu *(presenter)*

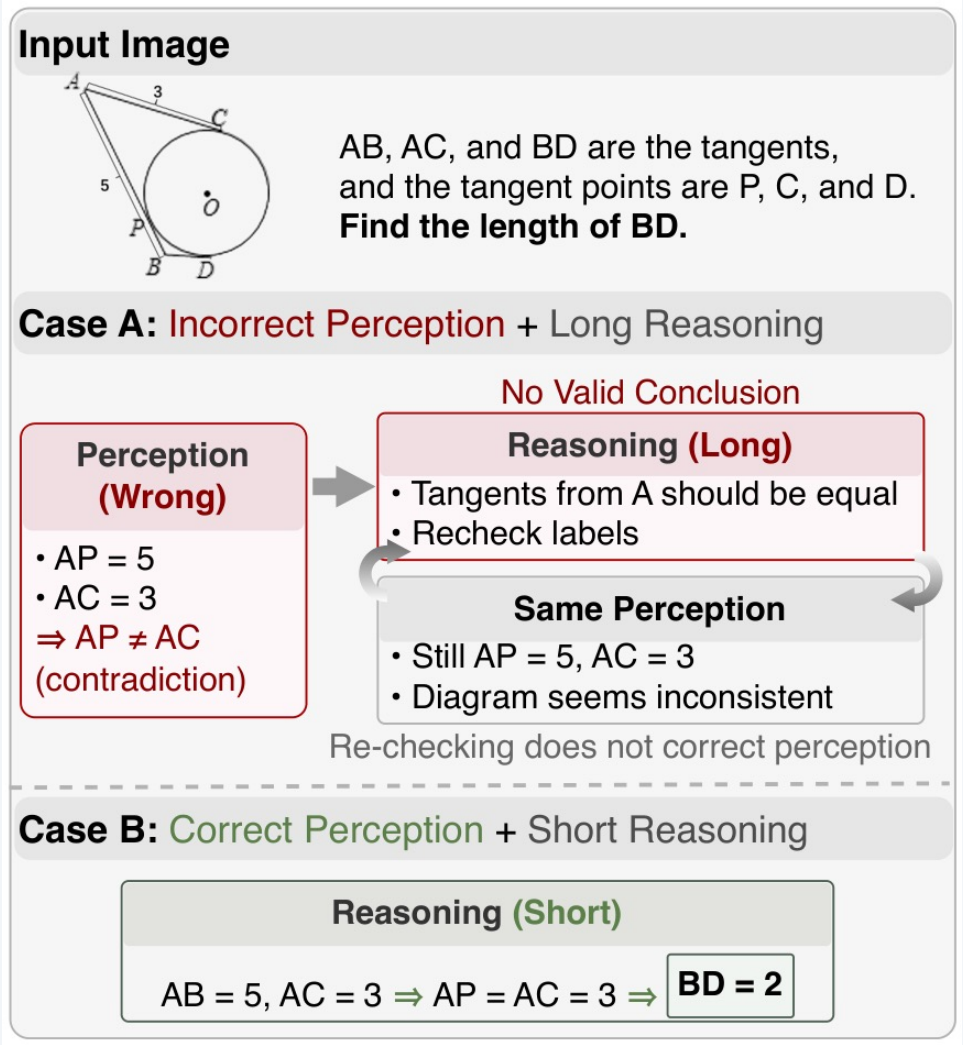
Hardy Chen · Haoqin Tu · Xianfeng Tang · Freda Shi · Hui Liu · Hanqing Lu · Cihang Xie · Yuyin Zhou

UC Santa Cruz · Amazon · University of Waterloo · Vector Institute

ICML 2026 · **ucsc-vlaa.github.io/VLM-CapCurriculum**

Longer thinking can't fix incorrect perception

- Recent VLMs push longer chain-of-thought reasoning via RL.
- **But on visual tasks the real limit is perception, not reasoning.**
  - Once the model mis-reads the image, re-checking repeats the same error and compounds it.
- We used Claude-Haiku-4.5 to label errors from Qwen3-VL-8B on 3 visual-math datasets:
  - **86.9% of wrong answers come from a visual perception error.**
- **Takeaway: longer reasoning does not compensate for wrong perception.**



OUR IDEA

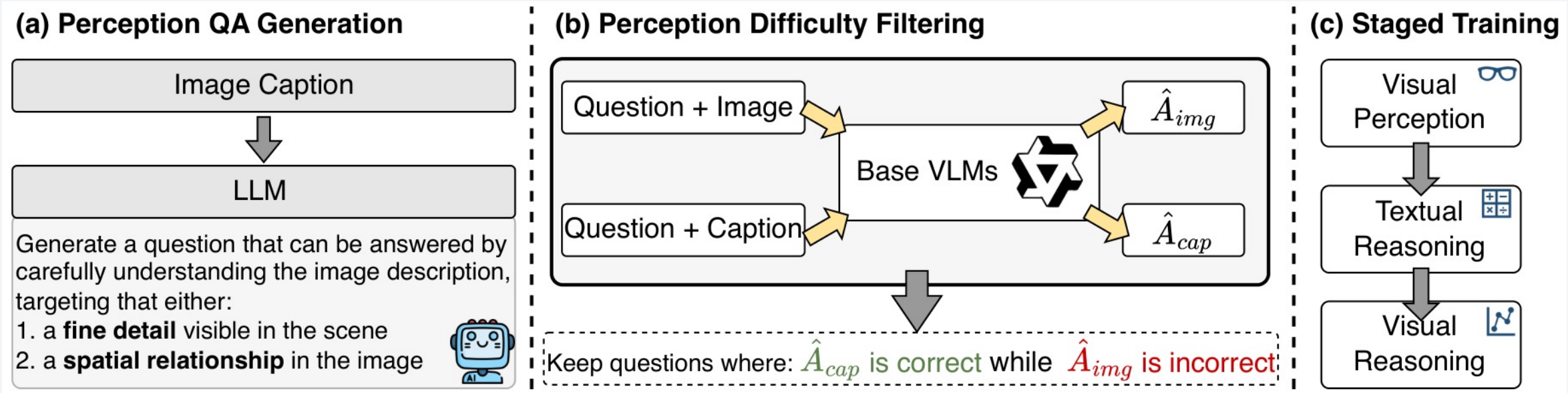
Treat perception as a separate, foundational capability

- **We decouple post-training into three capabilities and train them in stages:**

- **Perception is the scaffold — it must be solid before reasoning is refined.**
- Conceptually this is a new curriculum axis: capability, orthogonal to the usual difficulty axis.

METHOD

Building perception data & staged RL training



- Generate perception QA from DOCCI fine-grained captions with an LLM.
- Keep only hard cases: model is right from the caption but wrong from the image — isolates perception gaps.
- Train all stages with GRPO (verifiable-reward RL); reasoning data from ORZ-Math + open visual-math sets.

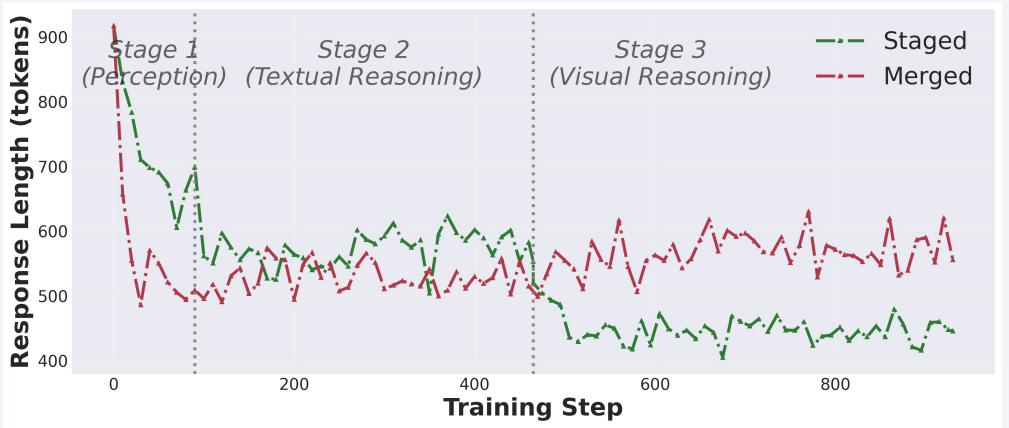
RESULTS

Finding 1 — Perception needs its own data

- Reasoning-only training boosts visual math but imposes a 'perceptual tax' — it can hurt perception (MMStar −1.6%).
- **Adding our perception data lifts visual math and protects perception:**
  - **RealWorldQA +3.0% (Qwen2.5-VL-7B), +3.6% (Qwen3-VL-8B)**
  - **WeMath +7.4 points over the 7B base model**
- **Perception is not a solved pre-training byproduct — it needs dedicated optimization.**

Finding 2 — Staged beats merged, and order matters

- **Staged training > merging all data: +1.5% math accuracy with 20.8% shorter reasoning traces.**
  - Strong perception means the model 'over-thinks' less.
- **Order is critical: perception → reasoning works;**
  - **reversing it drops Qwen2.5 visual-math from 42.3% to 37.7%.**
- Holds across 4 backbones (Qwen2.5/3-VL, InternVL3/3.5); wins 14/15 benchmarks over 3 runs.



Staged (green) converges to shorter responses in Stage 3

RESULTS

Finding 3 — RL > SFT, and curricula compose

- **For perception, RL with verifiable rewards beats caption-based SFT (+8.2% on WeMath, 7B).**
  - RL stays on-policy and penalizes hallucinated visual claims.
- **Capability + difficulty curricula are additive: 58.6 → 63.0 on Qwen3-VL-8B.**

Qwen3-VL-8B (Staged) — new best among 7–8B open VLMs

Benchmark	Base	OneThinker-8B	Ours (Staged)
MathVista	72.4	75.1	**75.9**
WeMath	50.9	54.6	**56.1**
RealWorldQA	70.9	71.5	**74.5**
Overall AVG	62.2	64.9	**65.8**

Takeaways

- Perception — not reasoning depth — is the dominant bottleneck in today's VLMs.
- Decouple post-training into stages: perception → textual reasoning → visual reasoning.
- Perception needs dedicated data and is best learned with RL, not SFT.
- Result: higher accuracy, shorter reasoning, and a new capability-curriculum axis.

Thank you! Project & code: ucsc-vlaa.github.io/VLM-CapCurriculum