

Does *Reasoning* Improve *Seeing*?

When Vision-Language Models Benefit from Thinking, and How to Steer It Without Retraining

Jing Bi¹, Luchuan Song¹, Dingxin Zhang², Pinxin Liu¹, Guangyu Sun³, Lianggong Bruce Wen⁴, Weidong Cai², Chen Chen³, Chenliang Xu¹
¹University of Rochester · ²The University of Sydney · ³University of Central Florida · ⁴Corning Incorporated

1 THE QUESTION Reasoning helps, but only sometimes

VLMs now ship both **Instruct** (direct) and **Thinking** (explicit-reasoning) modes. Extra reasoning is **instance-dependent**: it rescues hard, reasoning-dominant questions yet wastes tokens (and can hurt) on easy, perception-dominant ones. **When should a model think, and how much?**

Easy question

What is the color of the dog?

Perception heads activated

Hard question

What does the facial expression mean?

Reasoning heads activated

Exploit: see

Easy, perception-dominant items. Trust the visual evidence and answer directly; extra reasoning mostly burns tokens for no gain.

Explore: think

Hard, reasoning-dominant items. Weigh alternatives and spend more tokens; a fast, direct answer tends to miss.

Smarter, not longer

Selective reasoning beats always-on thinking: spend the extra tokens only where they actually change the answer.

2 ORACLE LABELS Three signals for adaptive inference

From **K = 10** paired stochastic generations (Instruct vs Thinking) on MMMU-val900, we derive per-instance oracle labels.

r_{help} Reasoning helpfulness

p_{think} - p_{instruct} ∈ [-1, 1]. Positive means explicit reasoning improves reliability.

ℓ_{opt} Optimal length

Minimal token budget for a correct answer from the winning mode: the compute target.

d Difficulty

Easy · Medium · Hard: an auxiliary label for attribution and diagnosis.

Input (I, X) → K=10 Instruct → K=10 Thinking → Label winner + shortest correct trace

We keep the questions where the decision matters

Only items whose Thinking-vs-Instruct win-rate clears **0.6** enter the study: the cases where choosing to reason actually changes the outcome. Labels are averaged over the 10 stochastic passes.

3 PROBE INTERFACE Read the plan from one prefill token

We train light **linear probes** on the final-layer **last-token embedding**, captured *before* any answer is decoded. If VLMs are meta-cognitive, the plan is already there.

VLM (reason) → VLM (instruct)

Probe → Desire Length, Reason Helpfulness, Question Difficulty

VLM → Last token embedding

The routing signal is essentially free

One frozen forward pass on the prompt: no extra generation, no fine-tuning. The decision is available before the first token is spent.

Frozen

The VLM backbone is never updated; we only read its activations.

Prefill

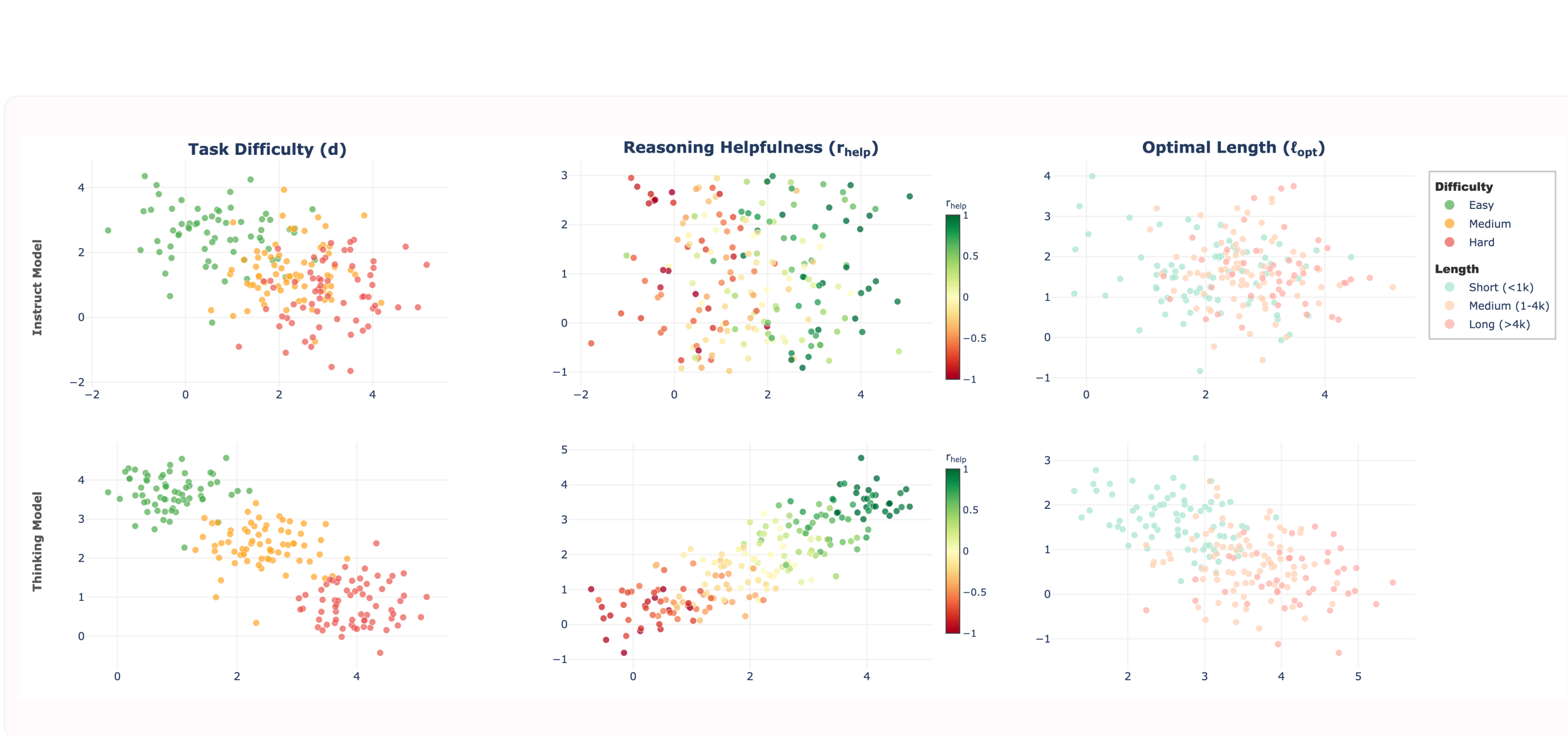
Captured at the last prompt token, before any answer is generated.

Linear

Two ridge regressors (r_{help}, ℓ_{opt}) and one difficulty classifier.

4 FINDING 1 · GEOMETRY Thinking models separate the controls linearly

Thinking embeddings form clean gradients for difficulty, helpfulness, and optimal length; **Instruct** embeddings stay entangled. Reasoning-oriented training **sharpens meta-cognitive structure**.



Projected last-token embeddings: Instruct (top) vs Thinking (bottom) across d, r_{help}, and ℓ_{opt}.

Reasoning training linearizes self-knowledge

The same axes that govern behavior (difficulty, helpfulness, budget) become linearly readable in Thinking models. Meta-cognition is not hidden; it is a direction.

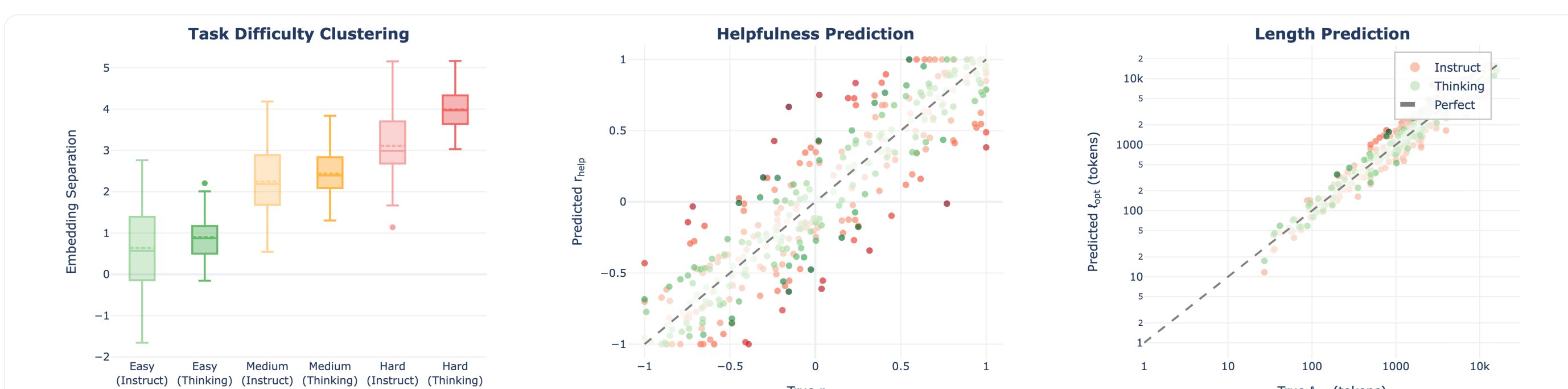
Thinking monotonic gradients along each control

Instruct diffuse, weakly structured

Readout controls lie on a linear direction

5 FINDING 2 · DECODING Probes recover difficulty, helpfulness & length

One linear probe, read three ways: it separates difficulty tiers by embedding distance and predicts both helpfulness and token budget close to the diagonal.



Left → right: difficulty clustering · helpfulness (predicted vs true) · optimal length (log scale, tokens).

Linear probes suffice: the structure is already there

No deep head is needed to recover difficulty, helpfulness, or length. The signal lives in the representation, so it can be read cheaply and steered directly.

3 tiers Easy · Medium · Hard separate by embedding distance

ρ ≤ 0.53 helpfulness tracks truth, best on Thinking

±32 tok optimal-length MAE across three decades

6 DECODABILITY Thinking beats Instruct on every probe

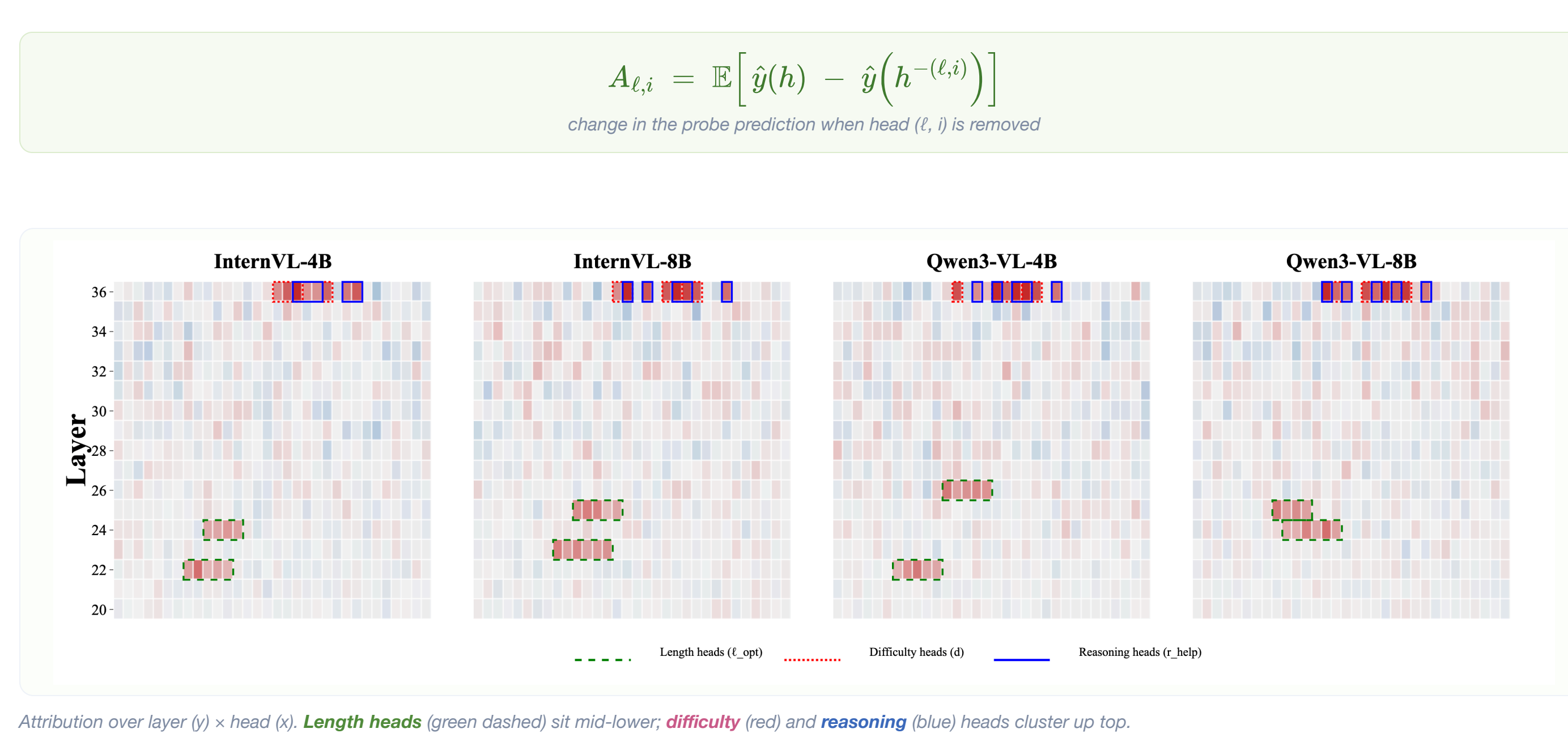
PROBE PERFORMANCE ON MMMU · THINKING / INSTRUCT

Model	r _{help} MSE ↓	r _{help} P ↑	ℓ _{opt} MAE ↓
InternVL-4B	.141 .182	.412 .324	38.7 42.3
InternVL-8B	.119 .168	.458 .341	35.2 40.1
Qwen3-VL-4B	.153 .194	.389 .298	41.2 45.8
Qwen3-VL-8B	.126 .173	.447 .328	36.8 42.9
Qwen3-VL-32B	.093 .157	.531 .362	31.8 39.1

Thinking-probes win **every metric on every model**; decodability keeps rising with scale (Qwen3-VL-32B is strongest), evidence that reasoning-tuned training builds localized meta-cognitive signals.

7 LOCALIZE · EXP 2 Two attention-head circuits emerge

Head-wise attribution ranks every head by how much ablating it shifts a probe. Two partially disentangled circuits emerge across all four models.



Attribution over layer (y) × head (x). **Length heads** (green dashed) sit mid-lower; **difficulty** (red) and **reasoning** (blue) heads cluster up top.

Length heads Mid-to-lower layers · a token-budget knob · Jaccard < 0.25 with the rest.

Reasoning + difficulty Upper layers · one shared circuit · Jaccard **0.68–0.76**.

The split is not a K artifact

Across K ∈ {8, 12, 16, 20} the reasoning and difficulty heads keep a Jaccard of **0.68–0.76**, while length heads stay disjoint (< 0.25); a stable, reproducible division of labor.

8 STEER · TRAINING-FREE Scale the heads, steer the compute

ACCURACY +5.3 Reasoning-dominant cohort, scaling upper reasoning heads

LENGTH ×0.5–2 Token budget spans a knob with <1.5% accuracy change

TRAINING 0 Pure inference-time head scaling, no fine-tuning

DEPTH-CONTROLLED REASONING STEERING (Δ ACCURACY)

Model	Perc. cohort low layers	Reas. cohort high layers	Len change
InternVL-4B-T	+3.2	+4.1	+28%
InternVL-8B-T	+4.1	+5.3	+31%
Qwen3-VL-4B-T	+2.8	+3.7	+25%
Qwen3-VL-8B-T	+3.6	+4.9	+29%

CROSS-DATASET GENERALIZATION (InternVL-8B-T)

Dataset	Base acc	ΔLen	low Δ	high Δ
MMMU	62.4	+31%	+4.1	+5.3
MathVista	54.7	+28%	+3.8	+4.9
ScienceQA	78.3	+25%	+2.9	+3.7
ChartQA	69.1	+33%	+3.5	+4.6
A-OKVQA	71.2	+29%	+3.2	+4.2

Depth-dependent steering holds on all five benchmarks: emphasizing *lower-layer* reasoning heads re-grounds perception-dominant items; emphasizing *upper-layer* heads sharpens abstract reasoning.

9 TAKEAWAYS What the prefill state already knows

Instance-dependent

Reasoning helps some inputs, hurts others, not a global switch.

Signalled early

VLMs encode when to think before the first token is decoded.

Linearly decodable

One prefill embedding predicts r_{help}, ℓ_{opt}, and d.

Two circuits

Length heads (lower) + reasoning heads (upper), localized.

Steer test-time compute without retraining

Scaling length heads is a token-budget knob (accuracy stable); depth-controlled reasoning-head scaling trades perception against reasoning, up to **+5.3%** accuracy across five vision-language benchmarks.