

AdaLoop: Thinking with the Vision Encoder

Xinran Nicole Han¹ Yiyang Lu² Katelyn Gan² Todd Zickler¹

¹ Harvard University ² MIT

Oct. 13th

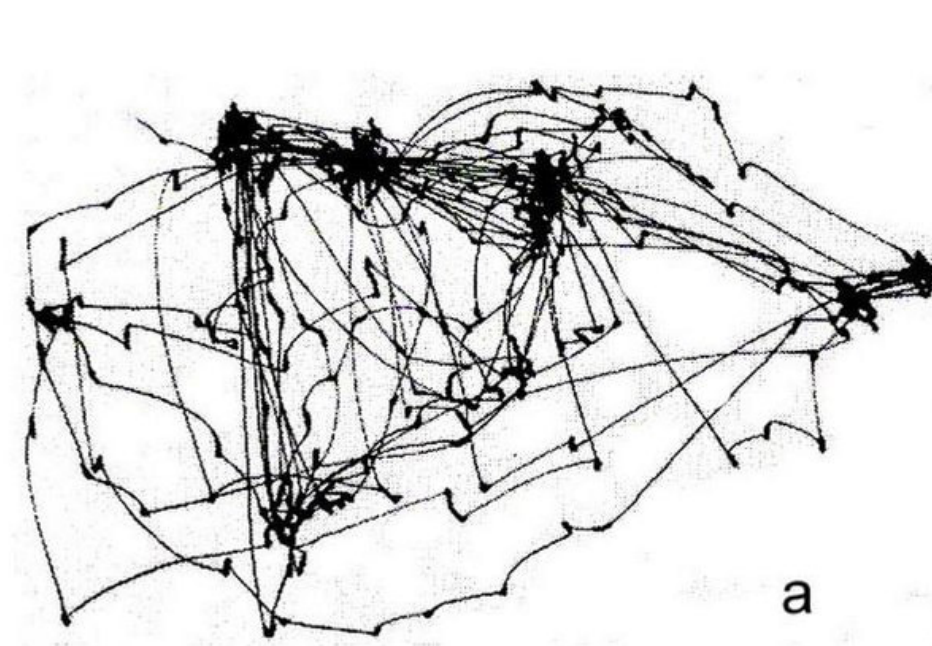
How do humans answer visual questions?

Visual attention is guided by the task.

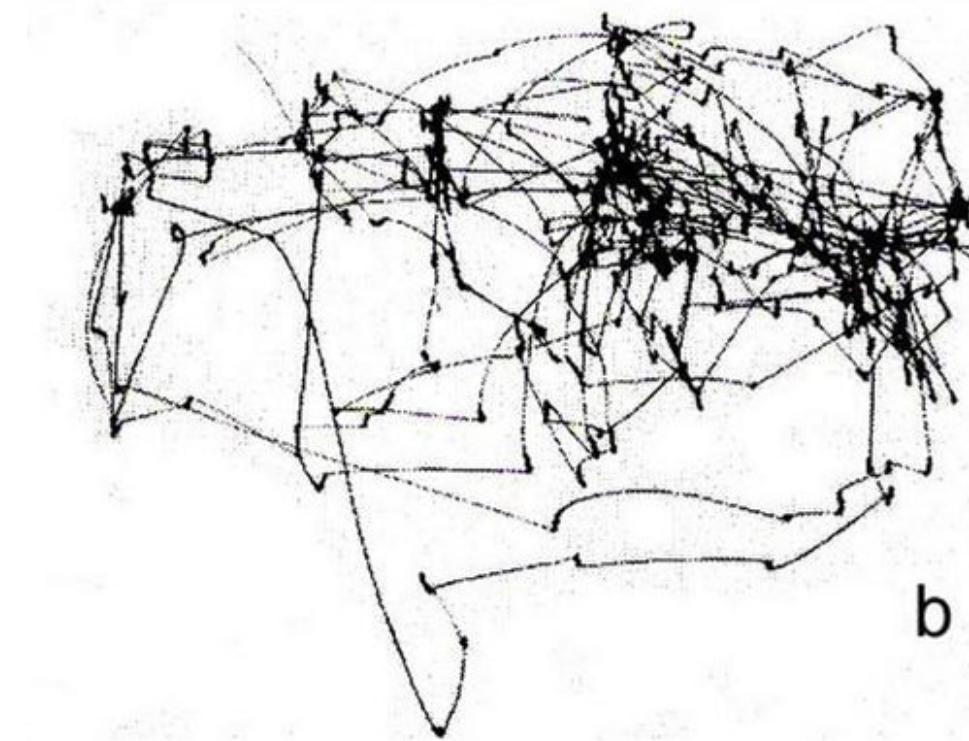
The same painting



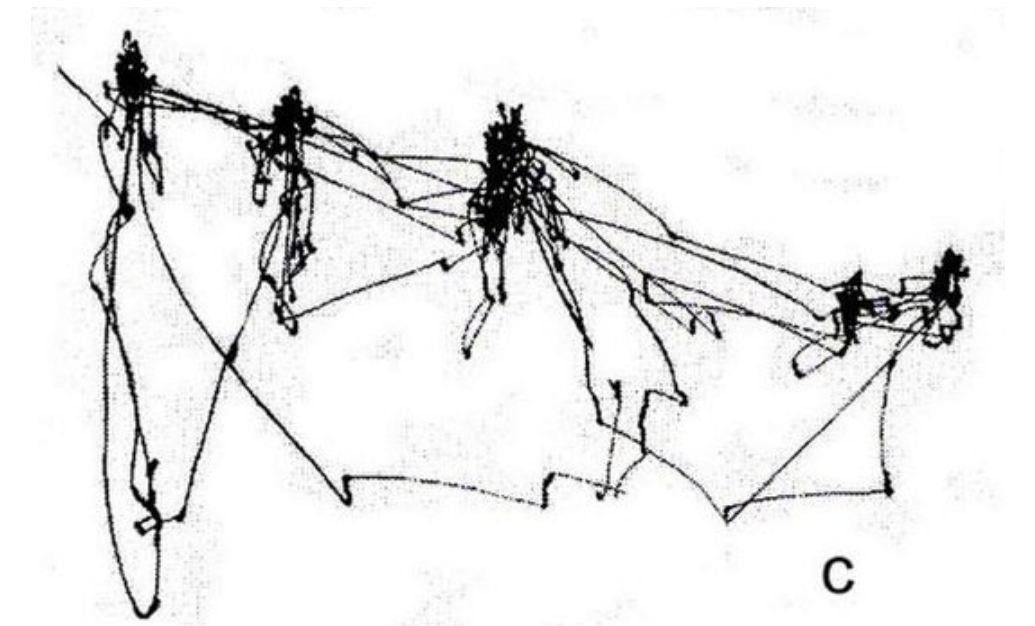
Free viewing



Estimate wealth



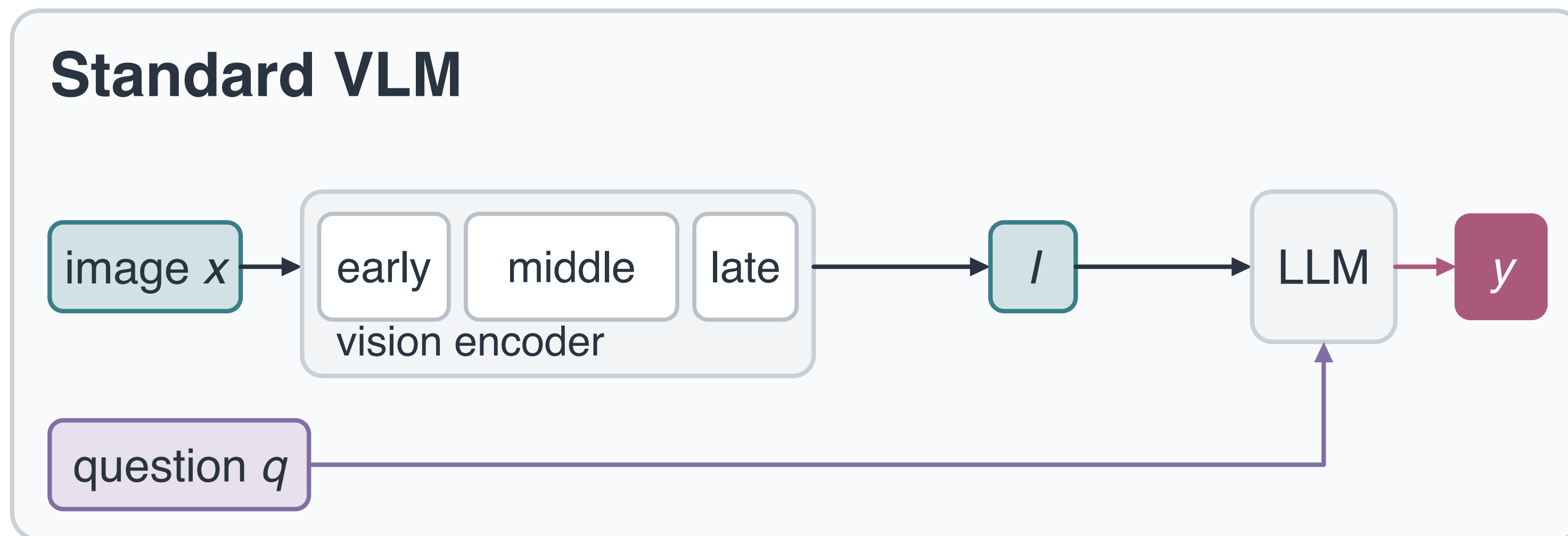
Estimate ages



Same observer · Same image · 3 min per task

How do VLMs answer visual questions?

Same image, same visual tokens—regardless of the question.



The question reaches the language model after visual encoding.

A common training recipe

1 Image-text pretraining

Vision encoder learns visual features.

2 Caption-based alignment

Connect image features to the LLM.

3 Visual instruction tuning

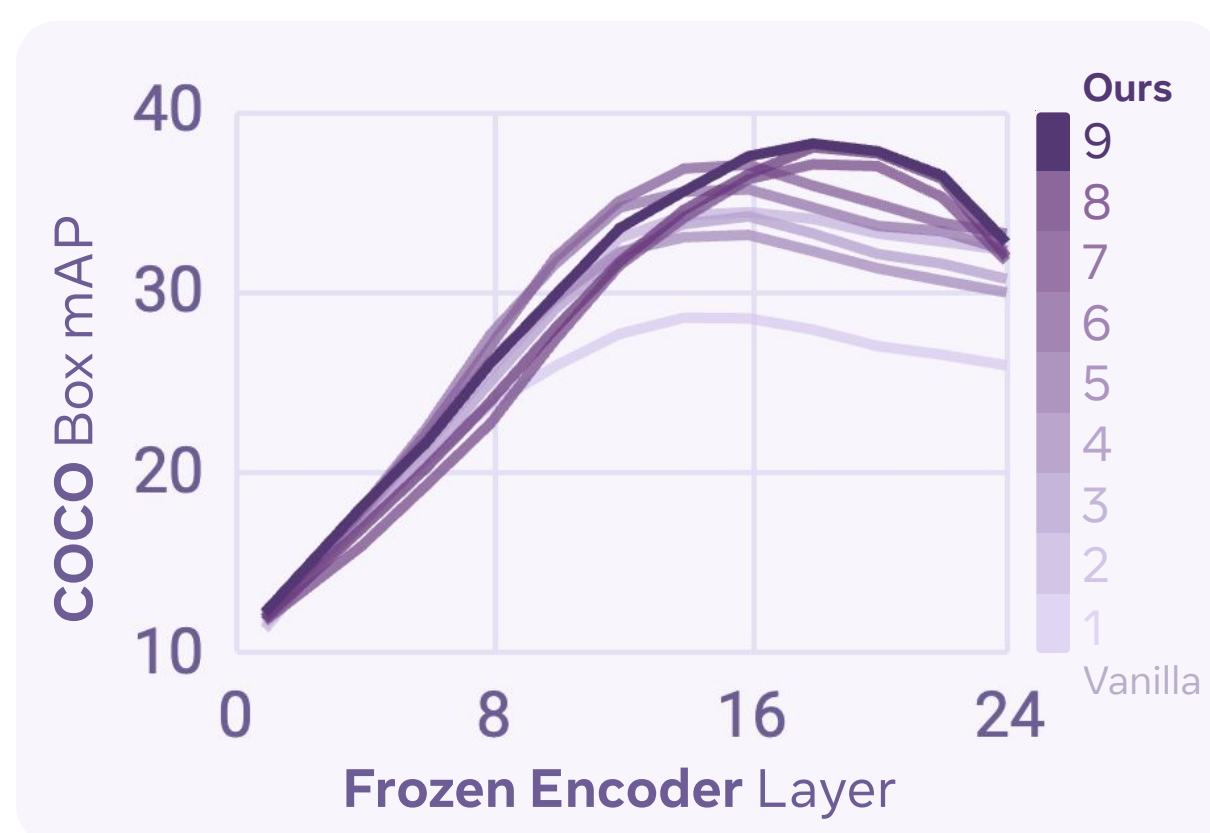
Learn to answer image questions.

Are these features enough for visual reasoning?

The final visual features may not expose what the question needs.

Perception Encoder: The best visual embeddings are not at the output of the network

Bolya et al., NeurIPS 2025



Detection features peak before the final encoder layer.

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

Tong et al., CVPR 2024



CLIP similarity: 0.95

Visually different images can have very similar CLIP embeddings.

How can we make the vision encoder more effective for visual reasoning?

Our approach

Data and evaluation · VisRO

Scalable training data and diagnostic evaluation.

Controlled difficulty and distribution shifts.

Architecture · AdaLoop

Adapt visual computation with question conditioning and recurrence.

What do current benchmarks ask?

Selected examples of scene description and recognition

COCO Captions

Describe the scene



A batter prepares to hit a pitch.

MMStar

Identify the scene



View from the deck? → Forest

MMBench

Recognize the environment

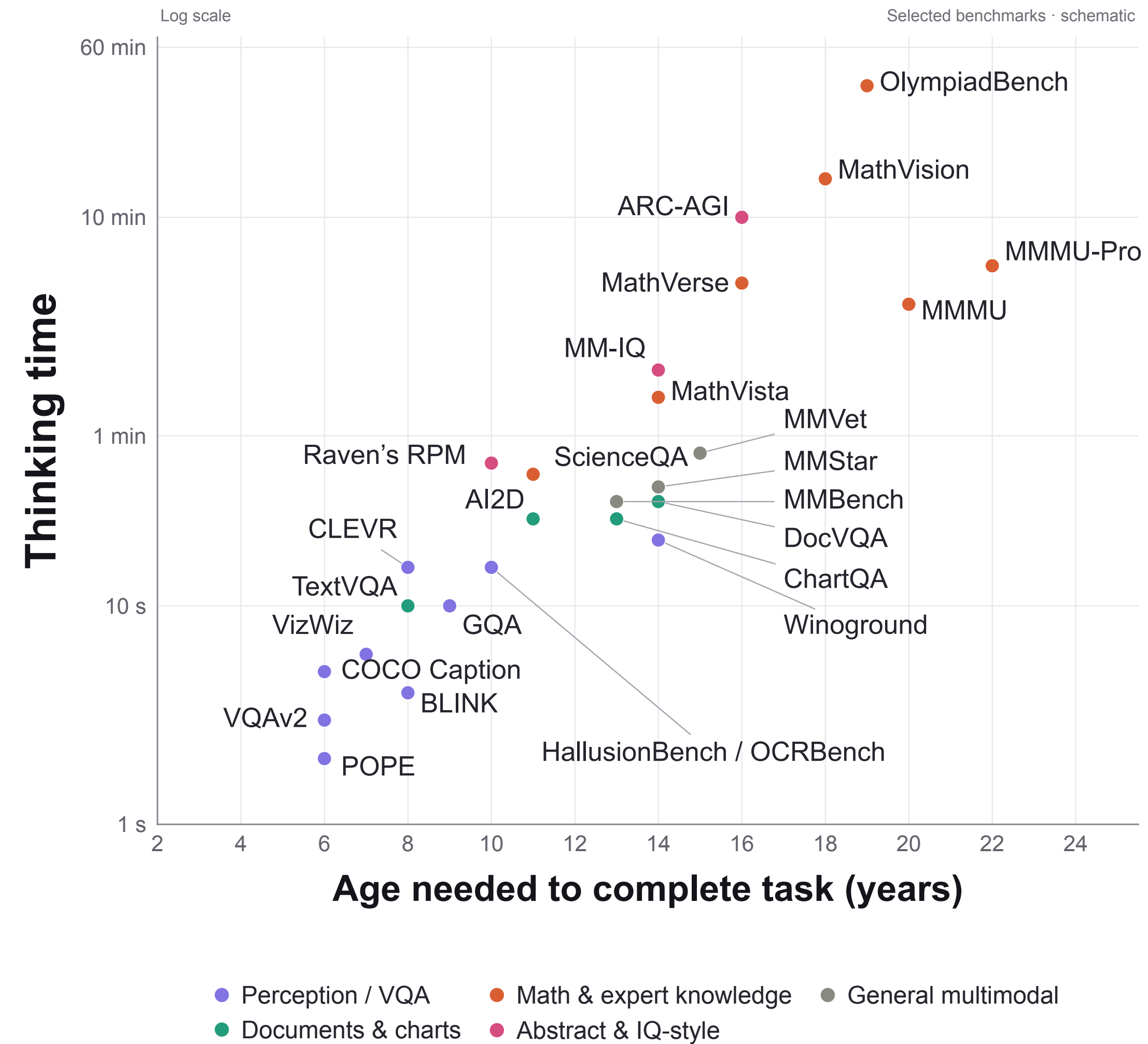


Which setting is shown? → Home

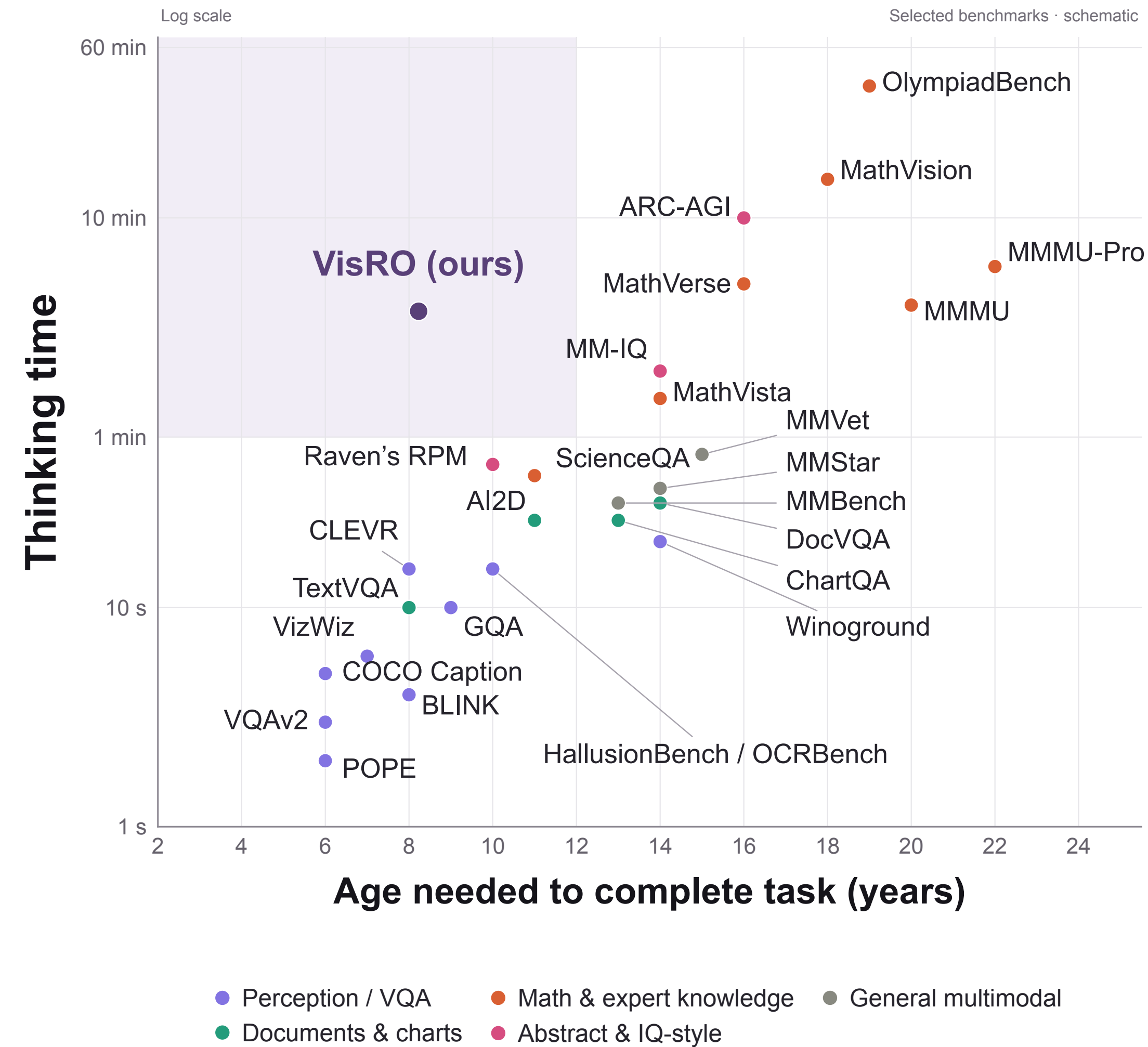
These questions ask for facts a short caption could contain.

Examples adapted from Chen et al. (2015), Chen et al. (2024), and Liu et al. (2024).

Multimodal benchmarks



Multimodal benchmarks



VisRO: beyond a single glance

Six core tasks · 1D examples · 600k training examples · All images 448 × 448

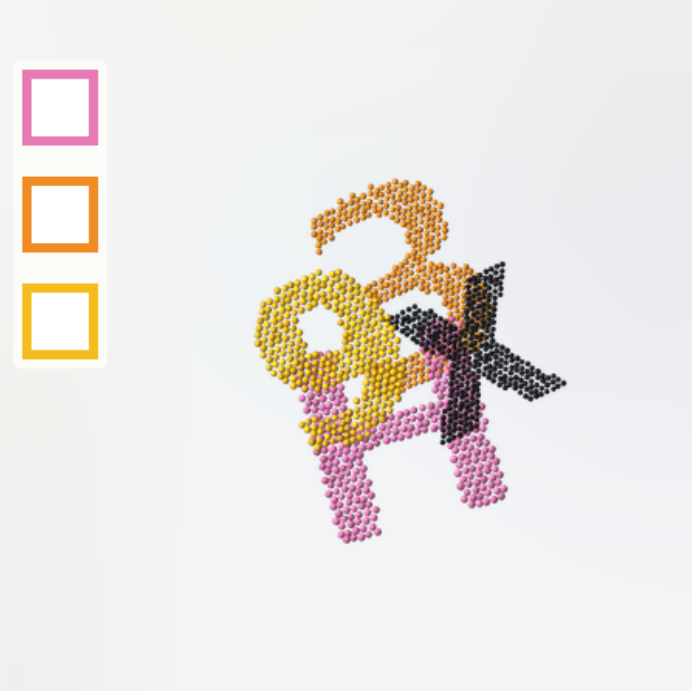


Count Clusters

How many objects overlap another?

medium

Ans. 5

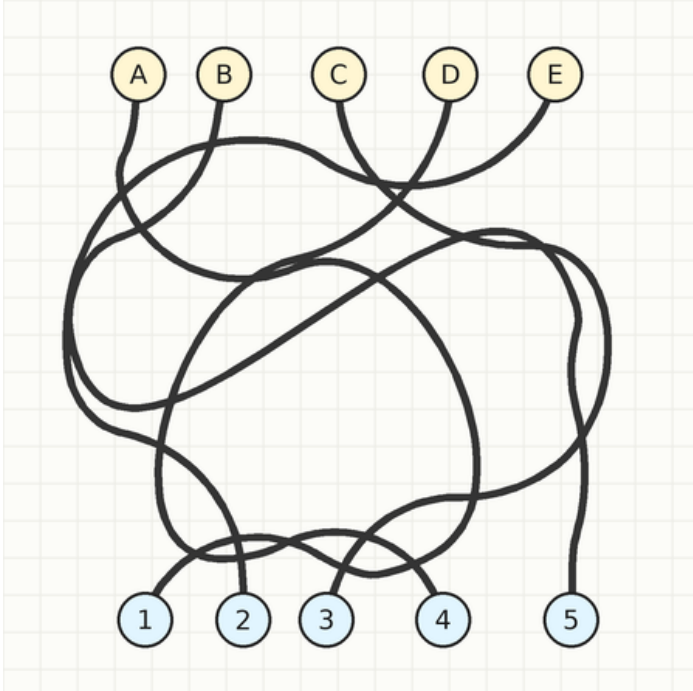


Overlapping Glyphs

Read left-box characters, top to bottom.

medium

Ans. H, 3, 9

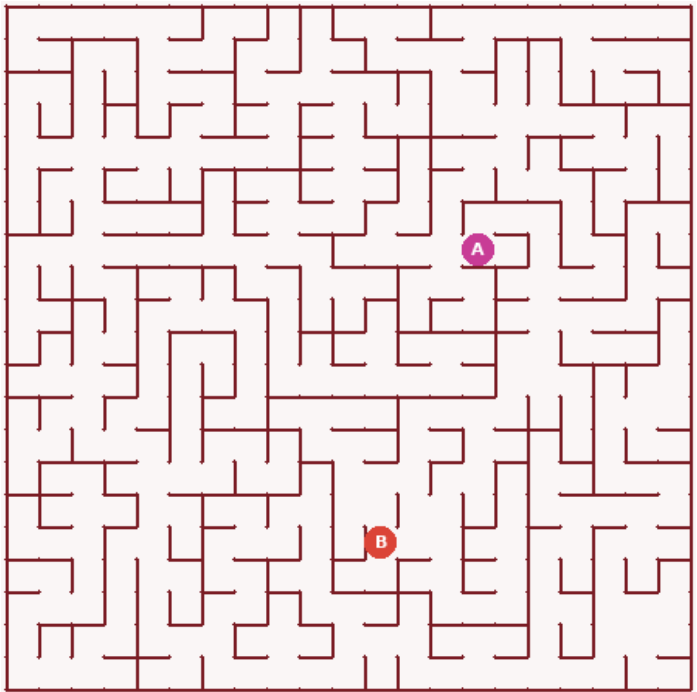


Path Tracing

What is ABCDE connected to, in order?

medium

Ans. 15342

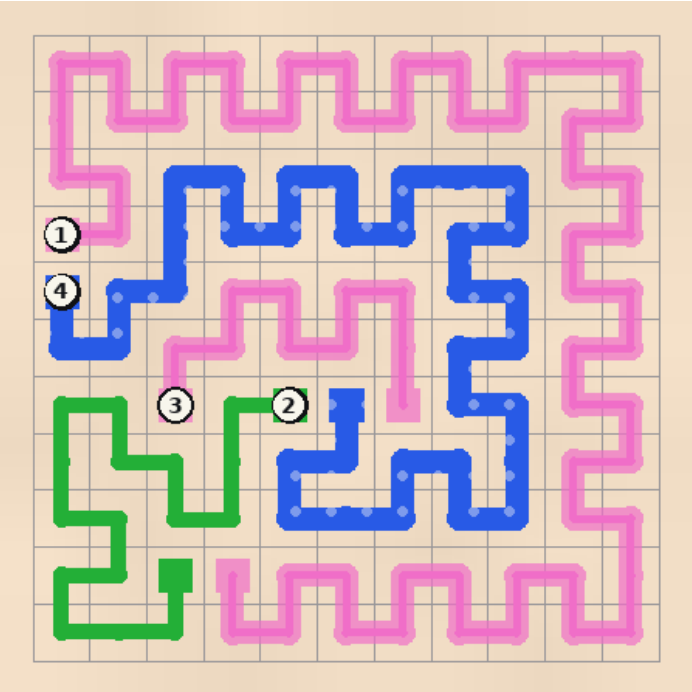


Maze

Does a route connect A and B?

hard

Ans. True

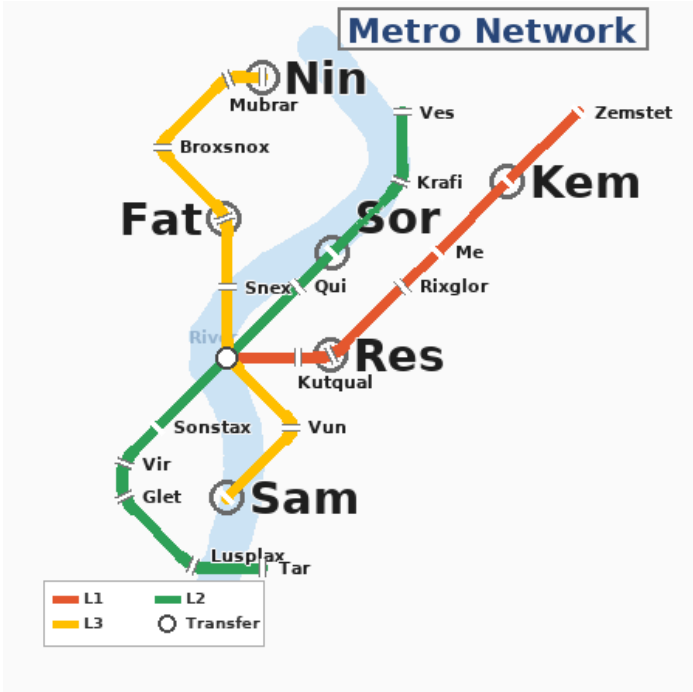


Flow Free

Squares per pipe 1–4, including endpoints?

easy

Ans. 56, 18, 11, 36



Metro Map

Count all stations on L3.

easy

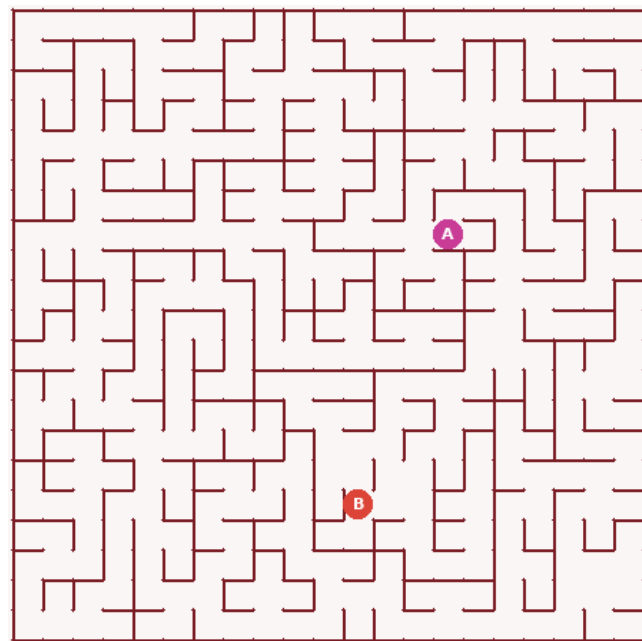
Ans. 8

VisRO: in-distribution and OOD

OOD changes complexity, generation rules, or visual appearance.

Maze

ID

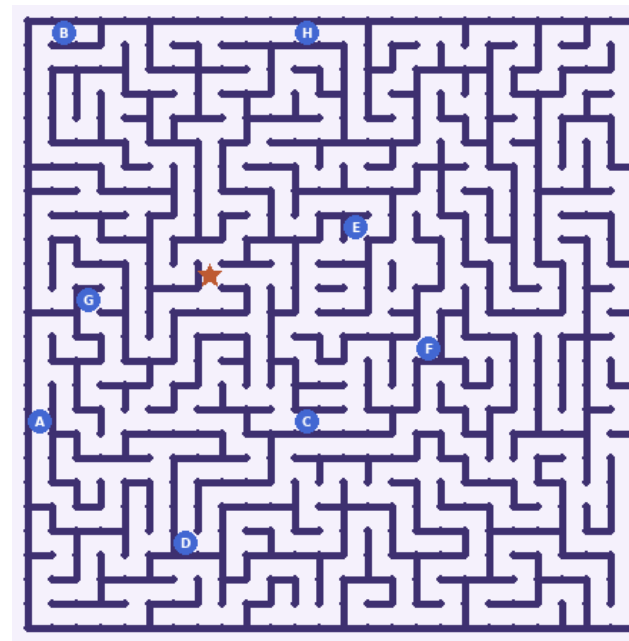


Does a route connect A and B?

Ans. True

Larger maze grids

OOD

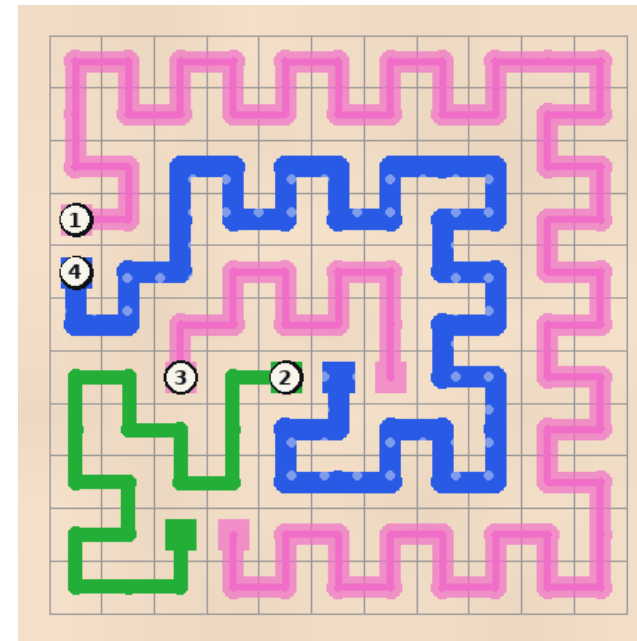


Starts within 20 corridor squares of target?

Ans. 1

Flow Free

ID

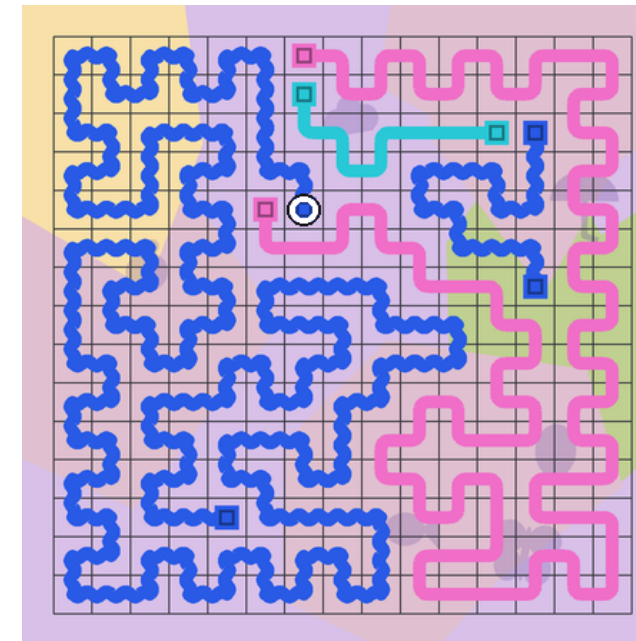


Squares per pipe 1–4, including endpoints?

Ans. 56,18,11,36

Different path construction

OOD

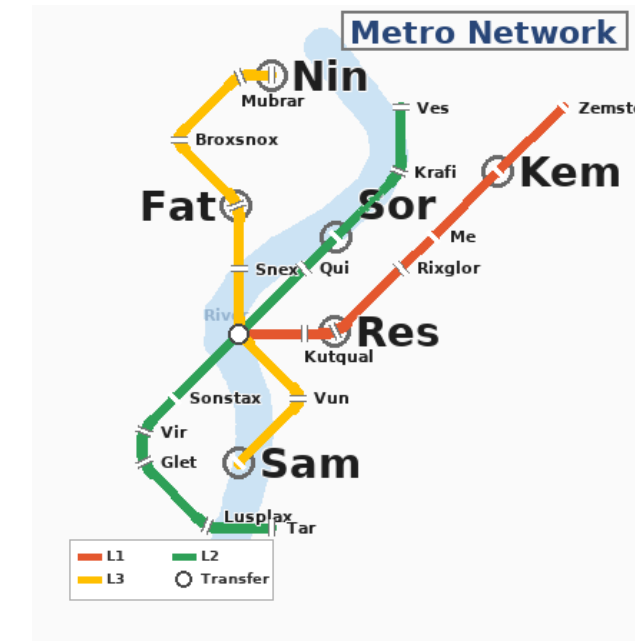


Grid turns on ringed pipe (exclude endpoints)?

Ans. 95

Metro Map

ID



Count all stations on L3.

Ans. 8

Different map style

OOD



Min. Kor–Rol stops (no ends; transfers OK)?

Ans. 8

Our approach

Data and evaluation · VisRO

Scalable training data and diagnostic evaluation.

Controlled difficulty and distribution shifts.

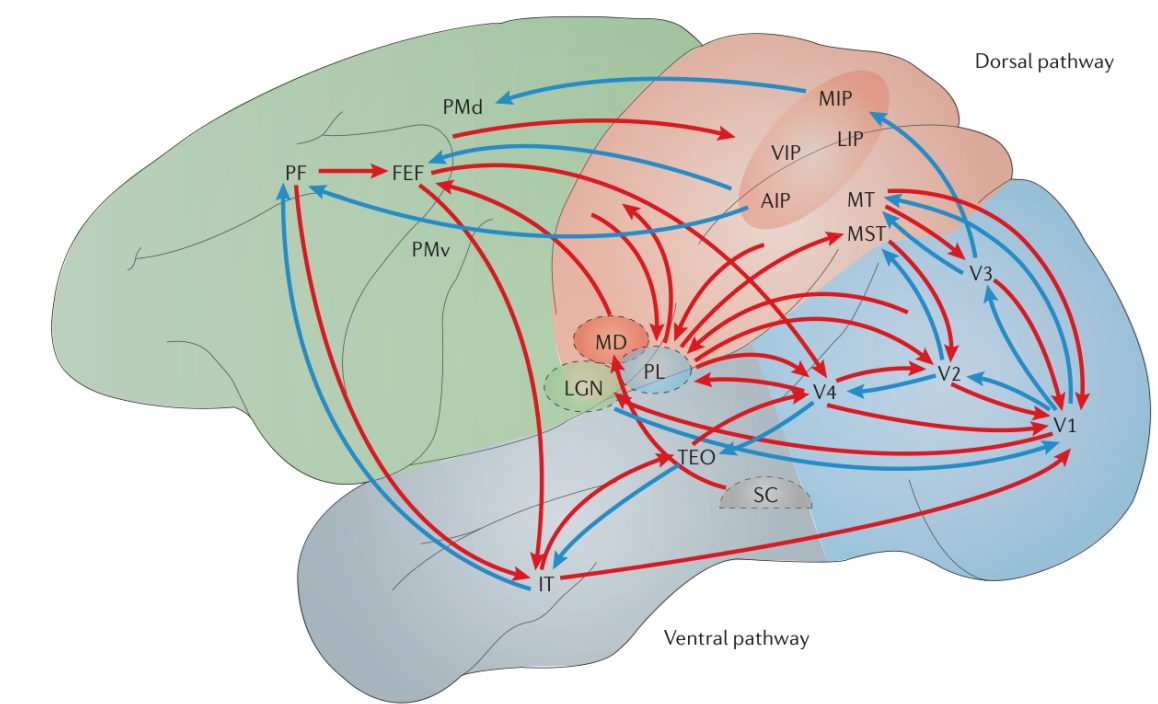
Architecture · AdaLoop

Adapt visual computation with question conditioning and recurrence.

Can the vision encoder do more?

- **Pretrained encoders already learn useful visual transformations.**
 - Adapt and reuse the existing encoder.
- **Different questions call for different visual features.**
 - Condition visual computation on the question.
- **Harder tasks may benefit from more visual processing.**
 - Reapply the same layers to update the visual state.

Recurrent visual processing

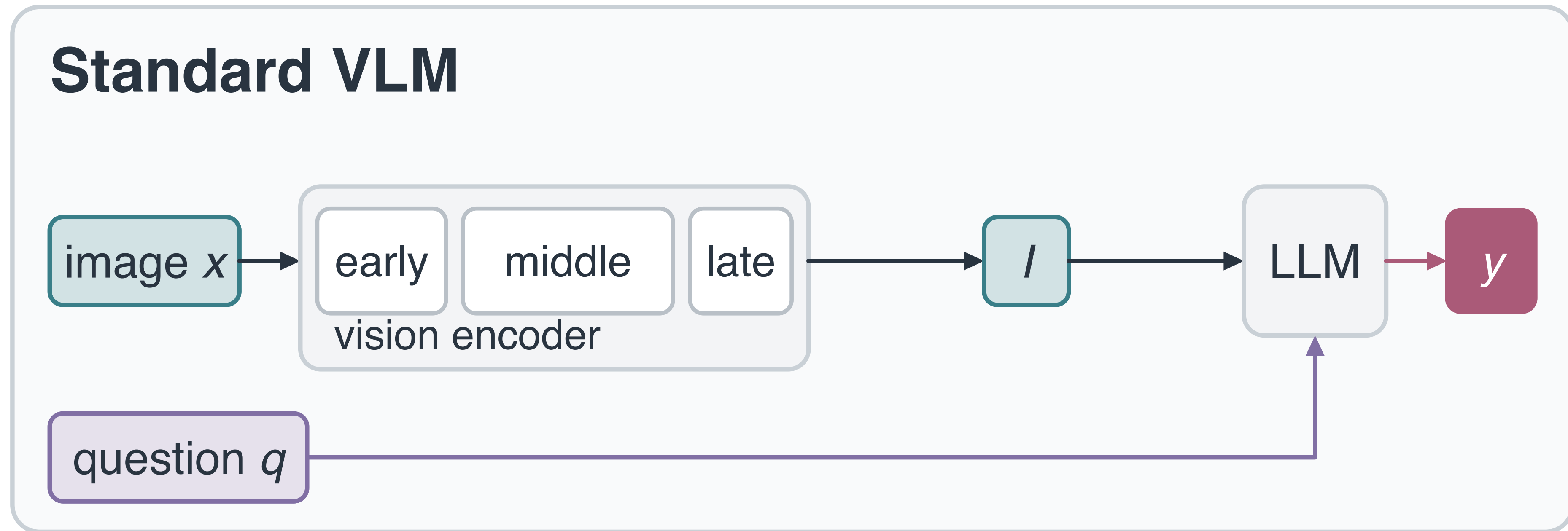


Feedforward

Feedback

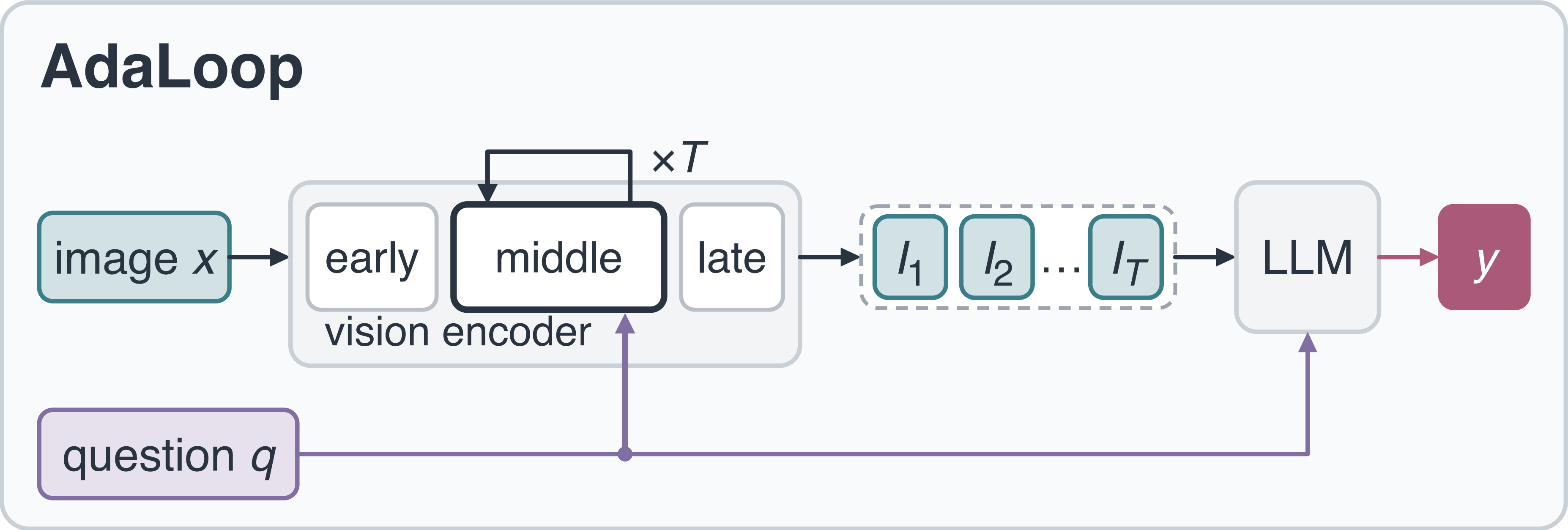
Gilbert & Li (2013), Fig. 1
Nature Reviews Neuroscience

What does a visual loop do?



One visual pass. The question enters at the language model.

What does a visual loop do?



Prelude

Early layers · run once

$$z_0 = \mathcal{P}(x)$$

Loop body

Middle span · repeated

$$z_t = \mathcal{R}(z_{t-1}; q)$$

Coda

Late layers + connector

$$I_t = \mathcal{C}(z_t)$$

Two design questions

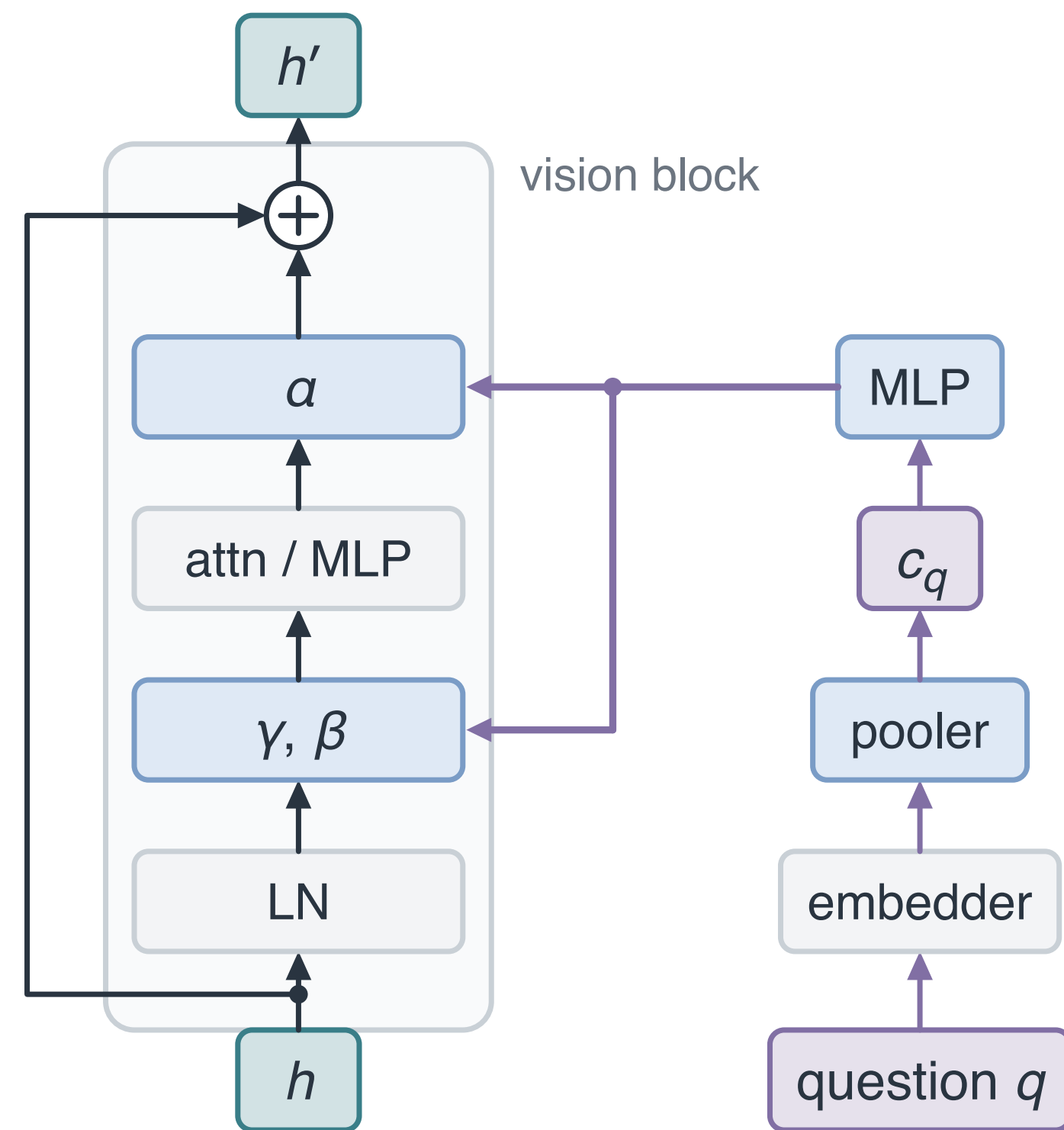
1. How should question signal adapt pretrained layers?

The vision encoder was trained for a single question-agnostic pass.

2. How should the visual loop connect to language?

Each iteration produces a visual state. Which states should the language model use?

Conditioning Visual Computation on the Question



- Reuse a span of the pretrained vision encoder.
- The same question modulates each iteration through AdaLN.

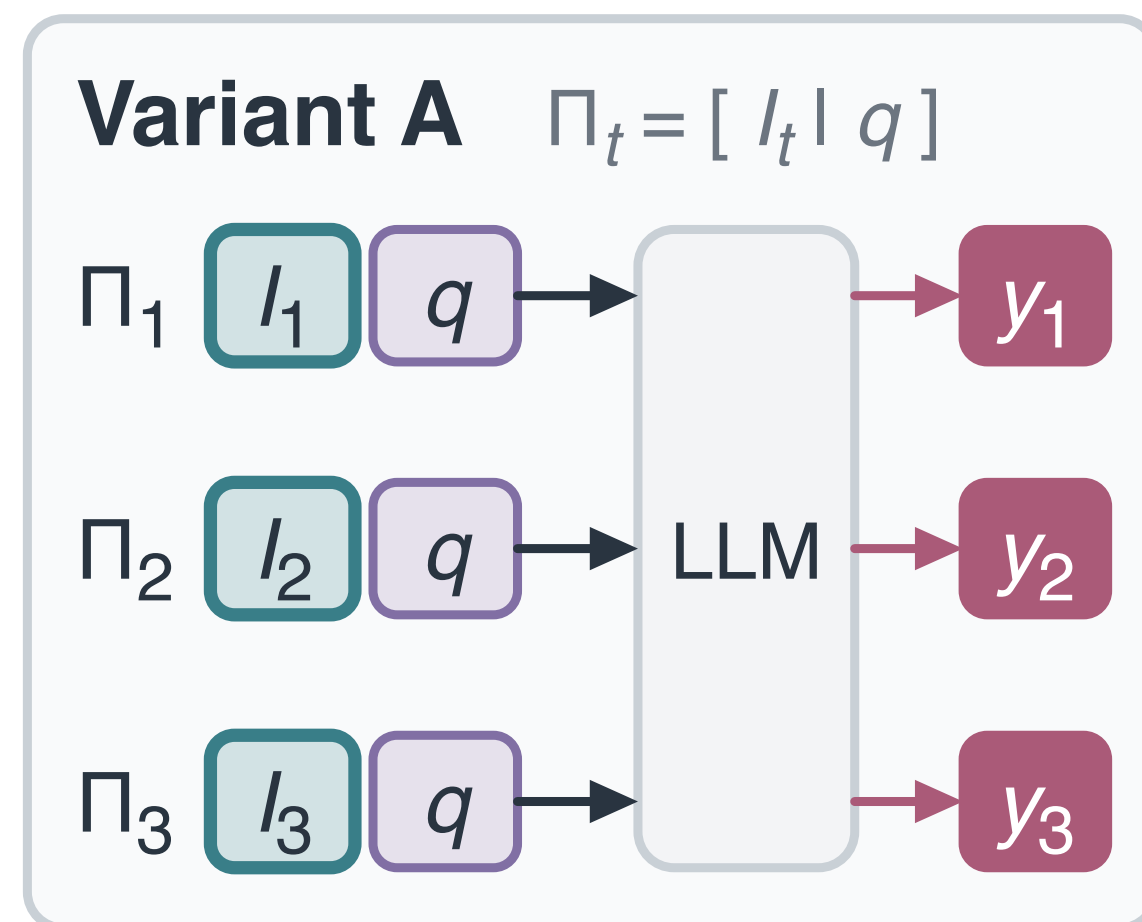
$$(\alpha, \beta, \gamma) = \text{MLP}(c_q)$$

c_q : pooled question embedding

$$h \leftarrow h + \underbrace{(1 + \alpha)}_{\text{gate}} \odot f\left(\underbrace{(1 + \gamma)}_{\text{scale}} \odot \text{LN}(h) + \underbrace{\beta}_{\text{shift}}\right)$$

How should language read the loop?

The language context can retain different amounts of history.



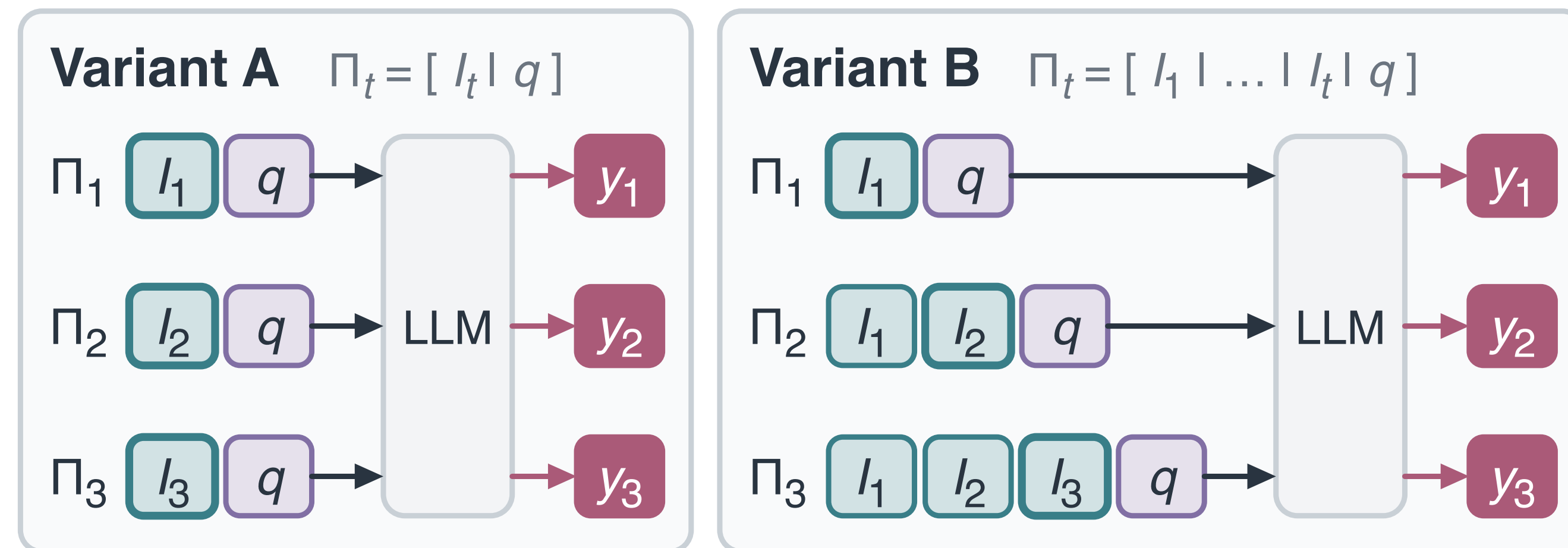
A • Current state (default)

Fixed-length visual context.

All variants share the same question-conditioned visual loop.

How should language read the loop?

The language context can retain different amounts of history.



A • Current state (default)

Fixed-length visual context.

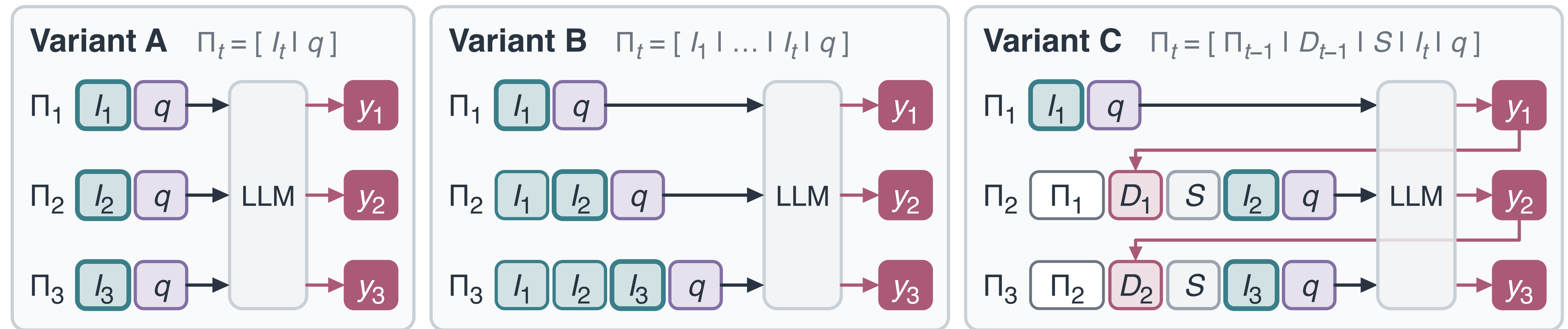
B • Visual history

Keep earlier visual states.

All variants share the same question-conditioned visual loop.

How should language read the loop?

The language context can retain different amounts of history.



A • Current state (default)

Fixed-length visual context.

B • Visual history

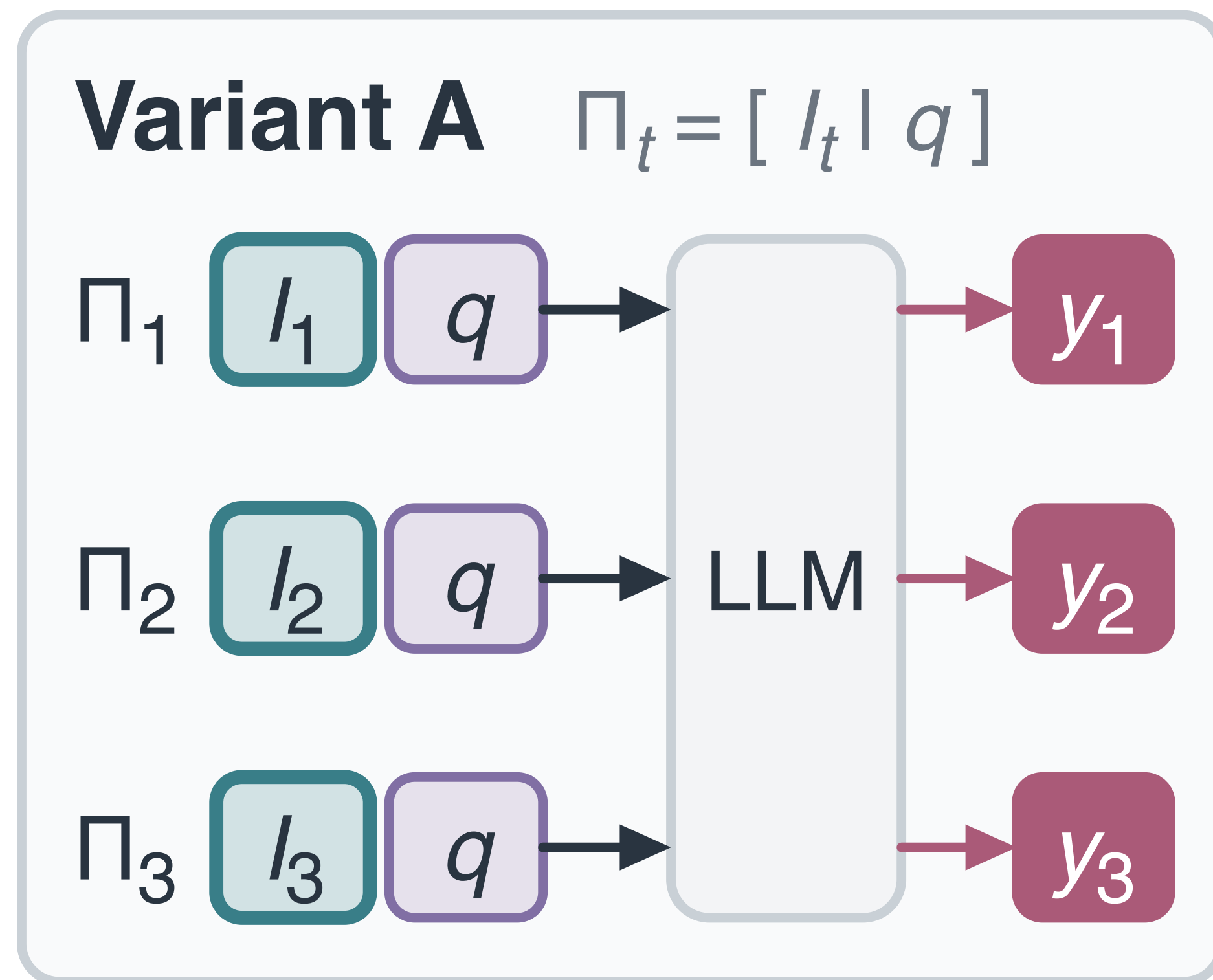
Keep earlier visual states.

C • History + answer drafts

Retain drafts in language context.

All variants share the same question-conditioned visual loop.

Learning to answer at every iteration



Answer loss

$$\ell_t = -\frac{1}{|y|} \log p_G(y \mid \Pi_t)$$

Deep supervision

$$\mathcal{L} = \sum_{t=1}^{T_{\text{train}}} w_t \ell_t$$

The same final answer supervises every iteration.

Fixed budget

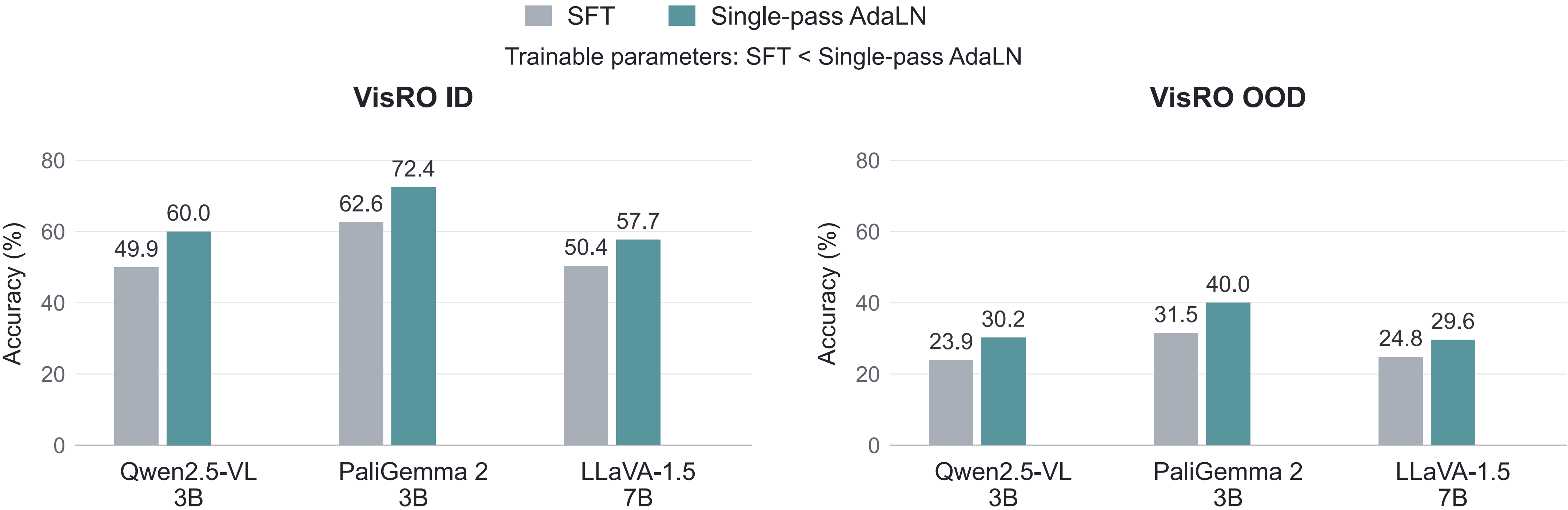
Run for the chosen iteration budget and return the final answer.

Adaptive stopping

Stop when consecutive answers agree or the iteration budget is exhausted.

Does looping improve reasoning?

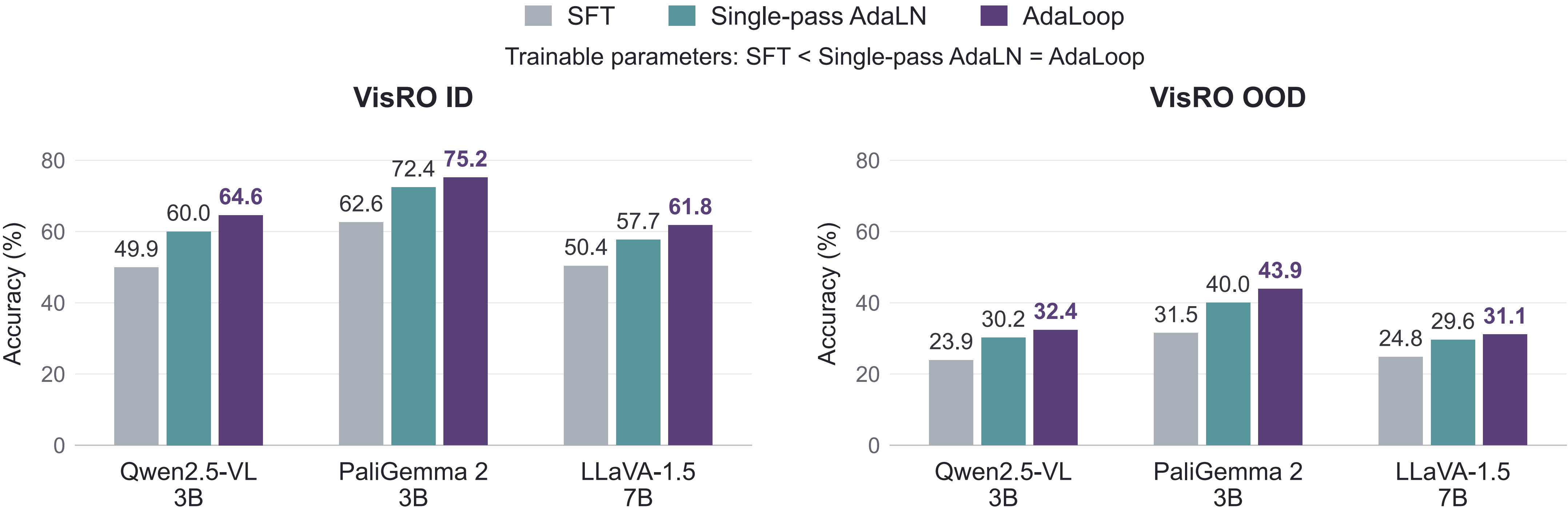
Matched training data and optimizer updates within each backbone



Question conditioning helps.

Does looping improve reasoning?

Matched training data and optimizer updates within each backbone



Question conditioning helps; looping adds gains without adding trainable parameters.

Why do we need both Ada and Loop?

Qwen2.5-VL · VisRO accuracy (%) · Trainable parameters in millions

Modulation	Params	Single pass		Loop	
		ID	OOD	ID	OOD
None	66	49.9	23.9	50.9	23.8
AdaLN (constant)	82	57.8	29.1	60.9	30.0
AdaLN (question)	82	60.0	30.2	64.6	32.4

Question conditioning and recurrence each add gains beyond extra parameters.

Harder problems, more iterations

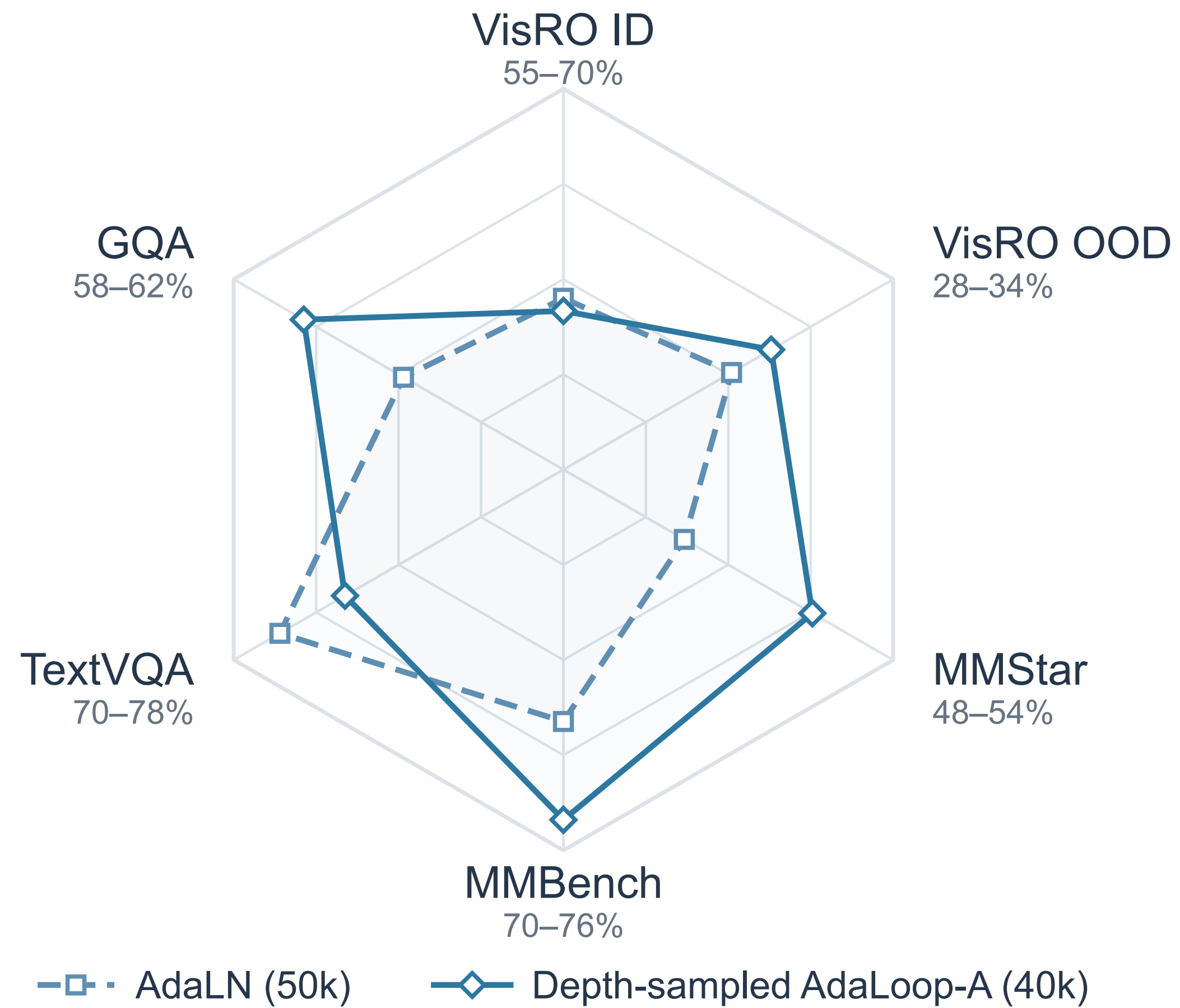
Adaptive stopping vs. one iteration of the same checkpoint

Difficulty	Qwen2.5-VL		PaliGemma 2		LLaVA-1.5	
	Rel. gain	$\bar{T}$	Rel. gain	$\bar{T}$	Rel. gain	$\bar{T}$
Easy	+3.33	2.11	+0.87	2.06	+3.05	2.11
Medium	+5.87	2.17	+0.00	2.09	−0.17	2.15
Hard	+4.09	2.21	+1.93	2.14	+4.87	2.21
XHard	+4.46	2.26	+5.43	2.21	+7.85	2.28
OOD	+6.39	2.30	+4.49	2.23	+6.57	2.29

Rel. gain: relative accuracy gain (%) $\bar{T}$: mean stopping depth

More visual iterations for harder inputs, without difficulty labels.

Two ways to supervise the loop



Qwen2.5-VL · Training depth = 3

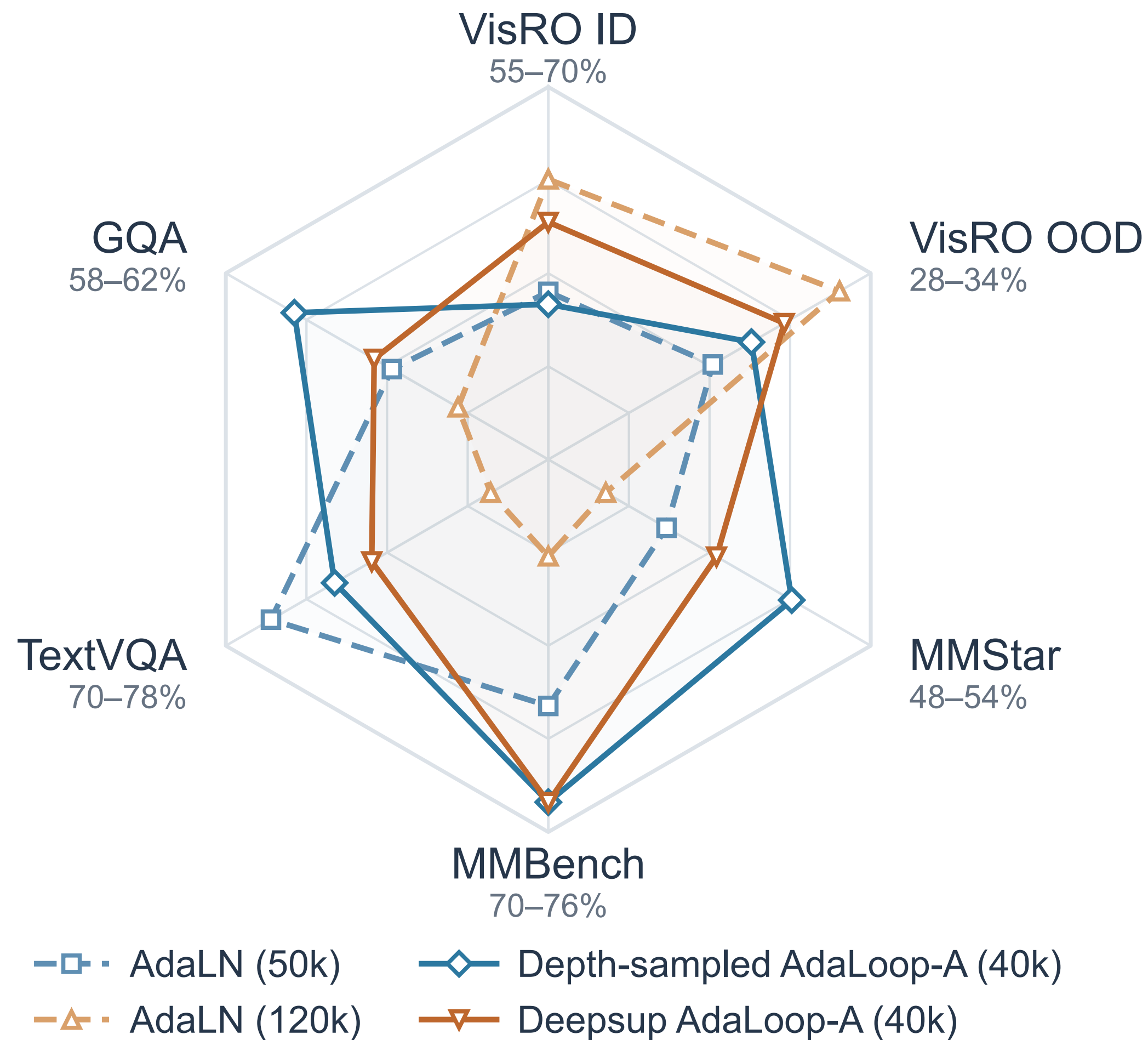
Each pair uses a similar training-FLOP budget.

Depth-sampled

Sample one depth per example and apply the final-answer loss at that depth.

AdaLoop (40k) vs. single-pass AdaLN (50k)

Two ways to supervise the loop



Qwen2.5-VL · Training depth = 3

Each pair uses a similar training-FLOP budget.

Depth-sampled

Sample one depth per example and apply the final-answer loss at that depth.

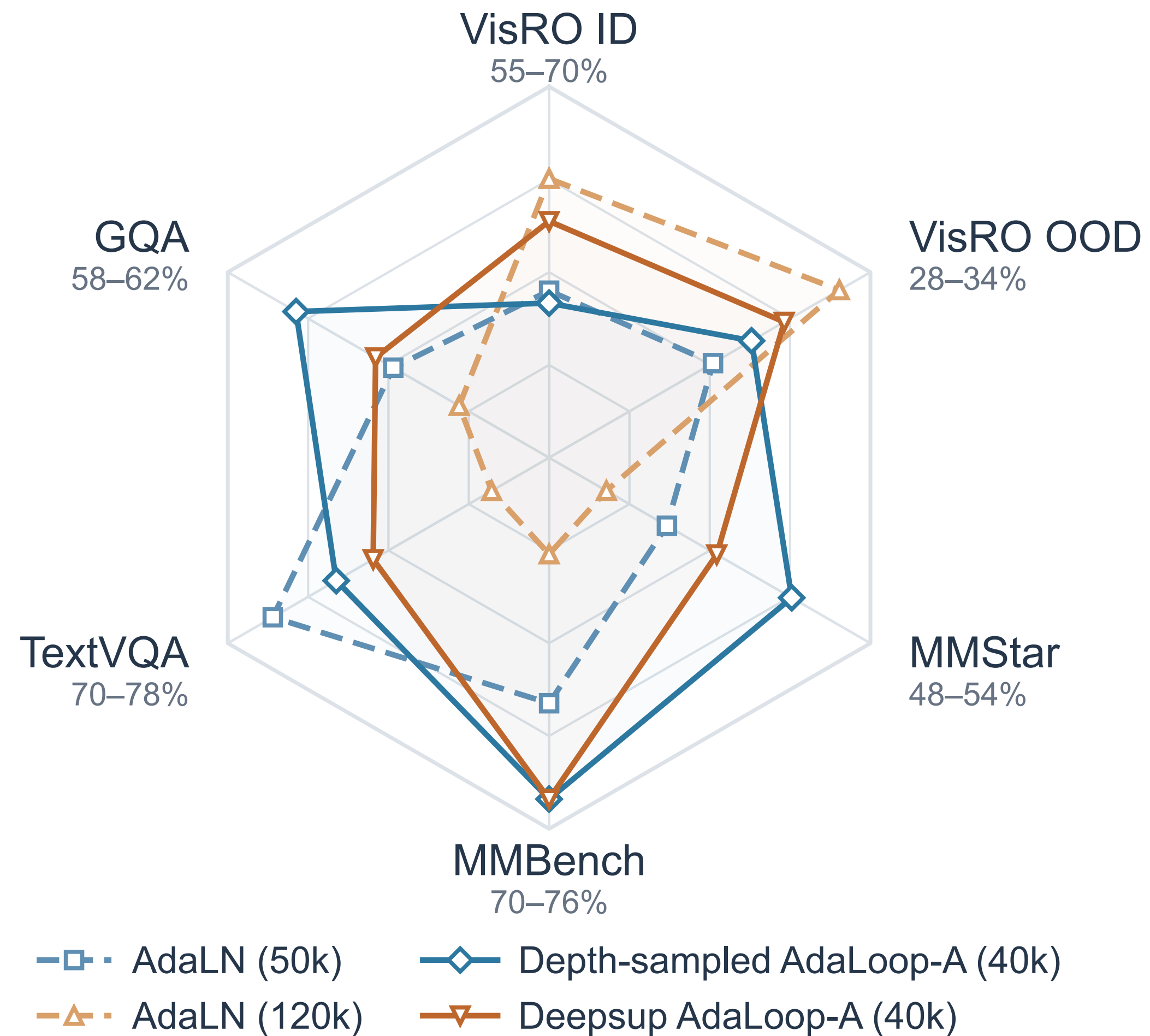
AdaLoop (40k) vs. single-pass AdaLN (50k)

Deep supervision

Apply the same final-answer loss at all three depths for every example.

AdaLoop (40k) vs. single-pass AdaLN (120k)

What does extra compute buy?



Our finding · Qwen2.5-VL

At similar training FLOPs, AdaLoop achieves competitive VisRO accuracy with fewer updates and better general VQA performance.

Takeaways

- **Vision encoders can support iterative reasoning.**
AdaLN and looping provide complementary gains across three VLM families.
- **Visual computation can adapt to problem difficulty.**
Harder examples receive more iterations on average, without difficulty supervision.
- **VisRO supports research beyond AdaLoop.**
Scalable training data, verifiable answers, and controlled difficulty and distribution shifts.

Thank you!

- Questions?