

InternVideo3: Agentify Foundation Models with Multimodal Contextual Reasoning

Ziang Yan*, Sheng Xia*, Jiashuo Yu*, Yue Wu* et al.

Corresponding author: Yi Wang

Shanghai AI Laboratory · Shanghai Innovation Institute · Nanjing University

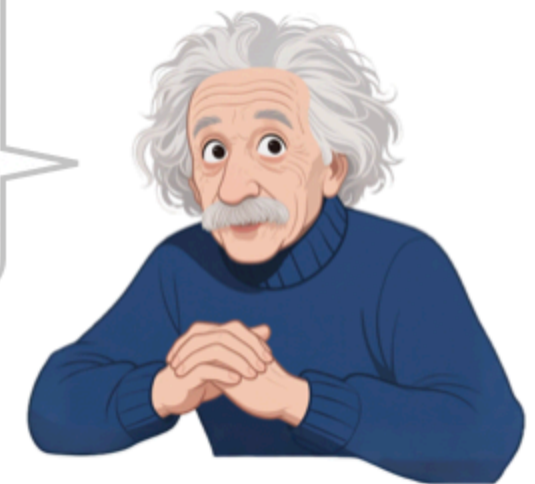
github.com/OpenGVLab/InternVideo/tree/main/InternVideo3

Where Did the Suspect Go?

Police have footage from every camera in a neighborhood.



How do we connect
scattered sightings
into one continuous
route?



From Video QA to a Visual Agent

The model decides what to inspect next and what evidence to retain.

Search

Find a candidate sighting; retrieve nearby times and camera views.

Remember

Keep timestamps, locations, observations, and unresolved uncertainty.

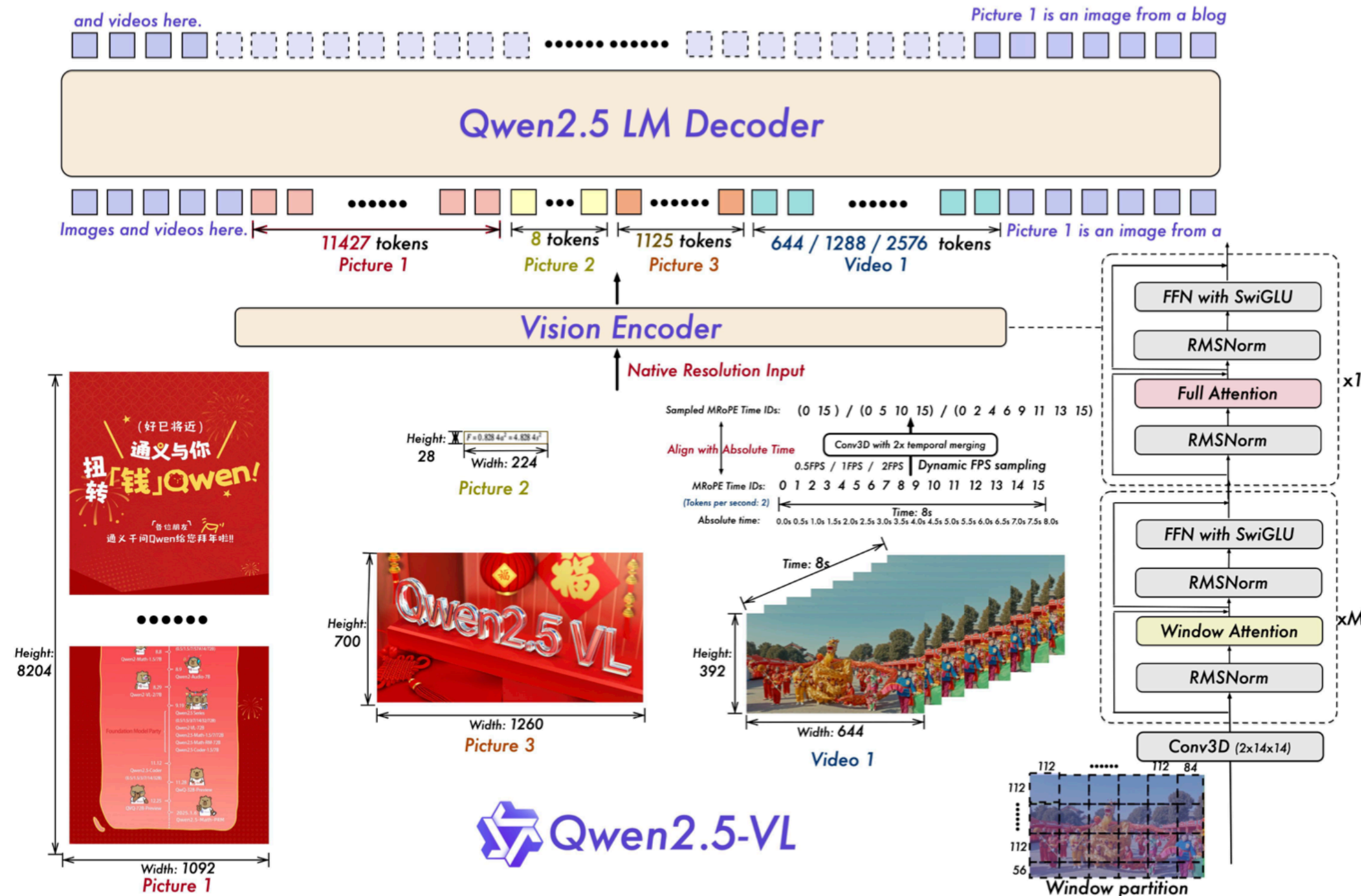
Reason & Verify

Link sightings, test possible routes, and revisit gaps in the evidence.

A visual agent alternates between perception, reasoning, and action.

Video QA: Qwen2.5-VL

A vision encoder connects images and videos to a language decoder.



Visual encoding

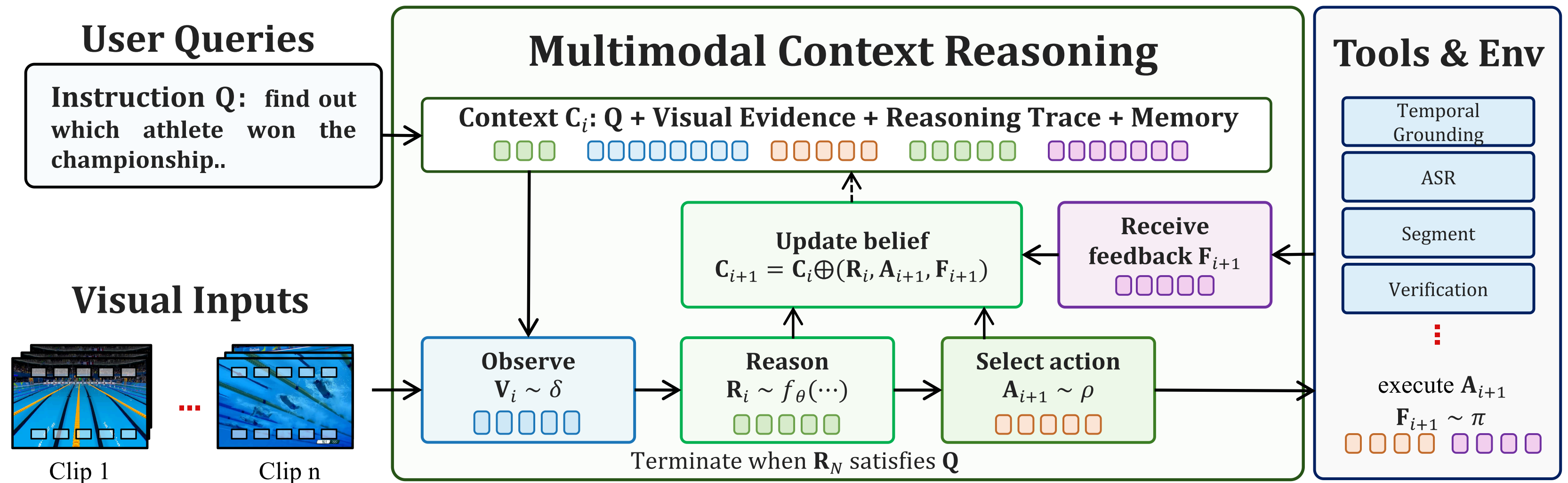
Images and video frames become visual tokens.

Language decoding

The decoder combines visual tokens with text to generate a response.

Visual Agent: A Shared, Evolving Context

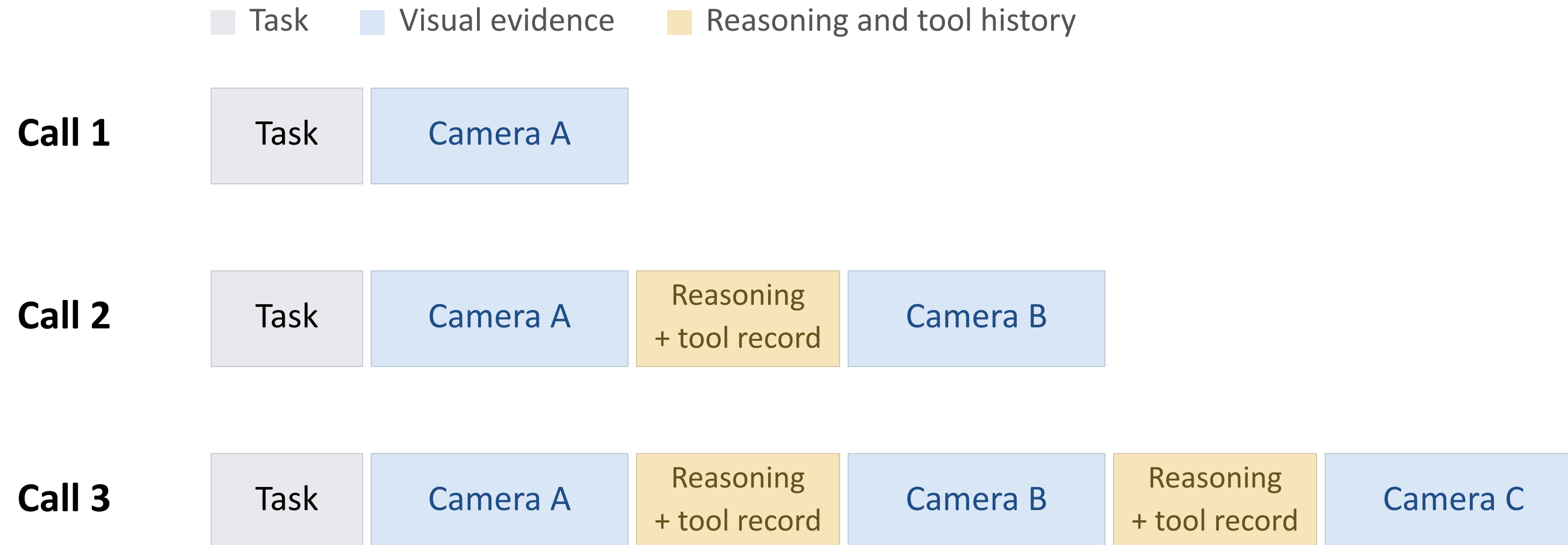
Keep observations, hypotheses, actions, and feedback together.



Observe $\rightarrow$ Reason $\rightarrow$ Act $\rightarrow$ Receive feedback $\rightarrow$ Update context

How the Context Grows

Each new inspection adds evidence to the history the agent already holds.



More retained tokens require more KV-cache memory during decoding.

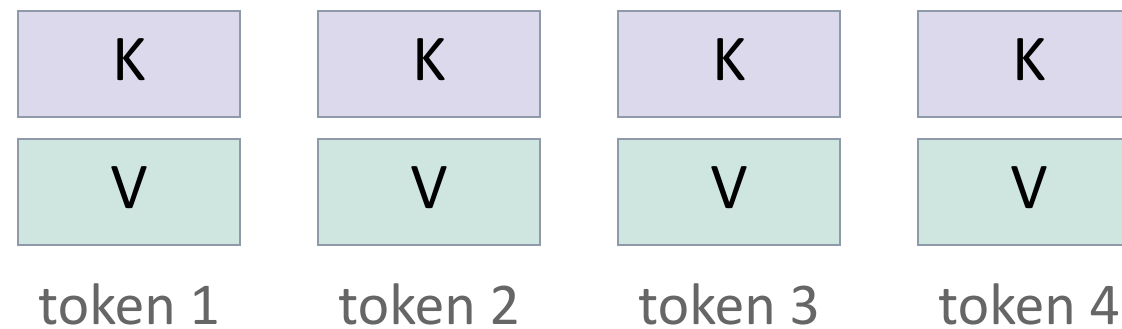
Why Does the KV Cache Grow?

Stored keys and values let each new token attend to earlier evidence.

Reuse past K/V instead of recomputing them at every decoding step.

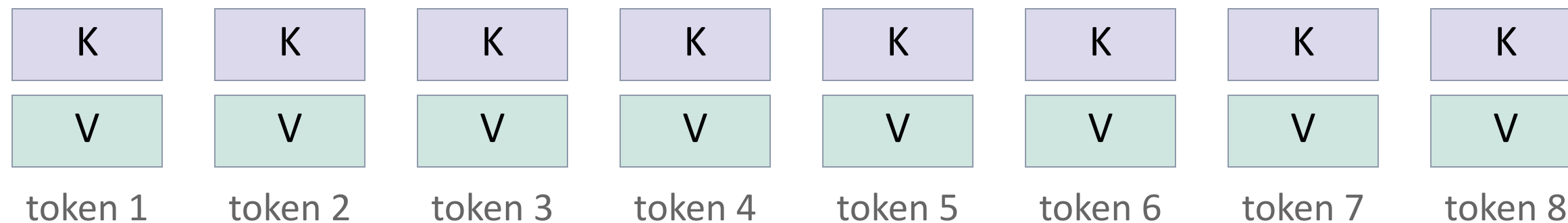
4 tokens

1× cache



8 tokens

2× cache



$\text{KV memory} = \text{tokens} \times \text{layers} \times \text{K/V state per token per layer}$

With a fixed model and precision, doubling tokens doubles KV-cache memory.

KV Cache and Inference GPU Memory

A smaller KV cache lowers GPU memory use at the same context length.

Inference memory = model weights + KV cache + runtime workspace

■ Weights ■ KV cache ■ Runtime workspace ■ Free memory

Standard cache

Same context



Compact cache

Same context



Compact cache

Longer context



The saved memory can hold more tokens within the same GPU budget.

How Cached K/V Predict the Next Token

With tokens x_1 and x_2 given, the last position predicts x_3 .

1. Project

Token x_2
hidden state h_2



q_2

$$q_2 = W_Q h_2$$

h_2 also produces k_2 and v_2 ,
which join the cache.



2. Attend

Visible K/V, including token x_2

Token	Key	Value	Weight
x_1	k_1	v_1	0.3
x_2	k_2	v_2	0.7

$$\text{weights} = \text{softmax}(q_2 K^T / \sqrt{d})$$



3. Predict

$$\text{Weighted sum } o_2 \\ = 0.3v_1 + 0.7v_2$$



Remaining layers
+ output head



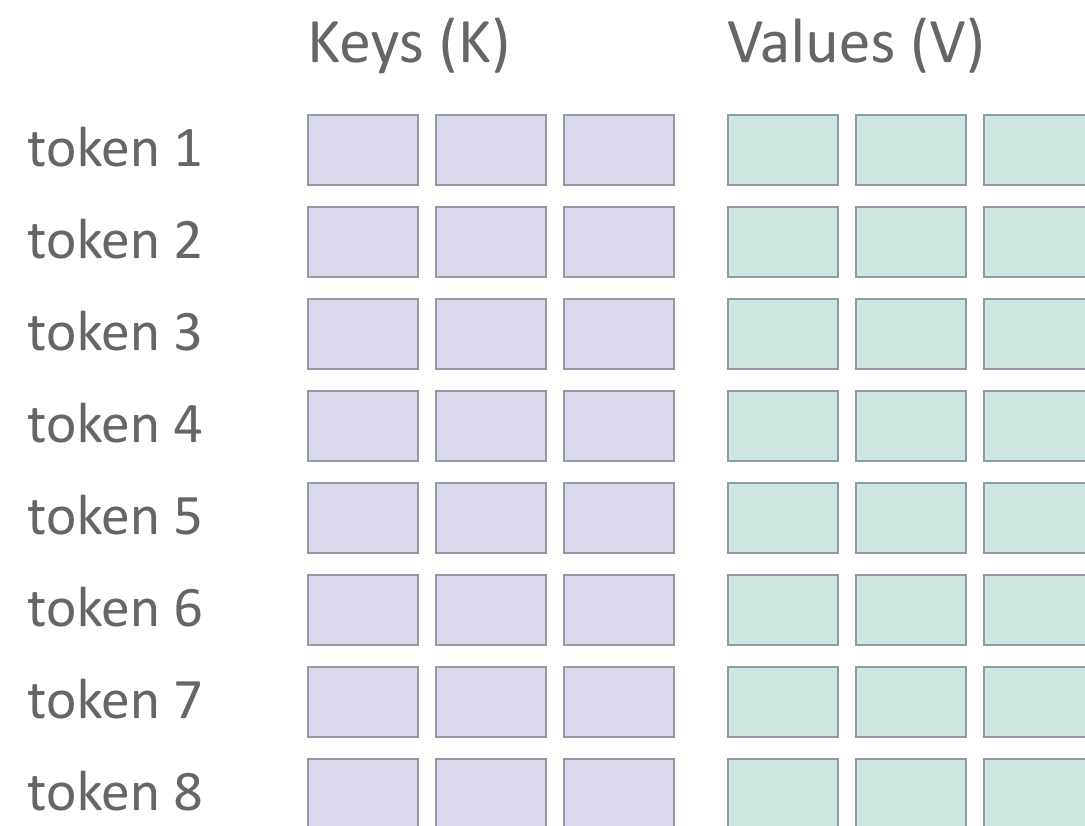
$$p(x_3 \mid x_1, x_2)$$

The last observed token supplies the query used to predict the next token.

M²LA: A Compact State for Every Token

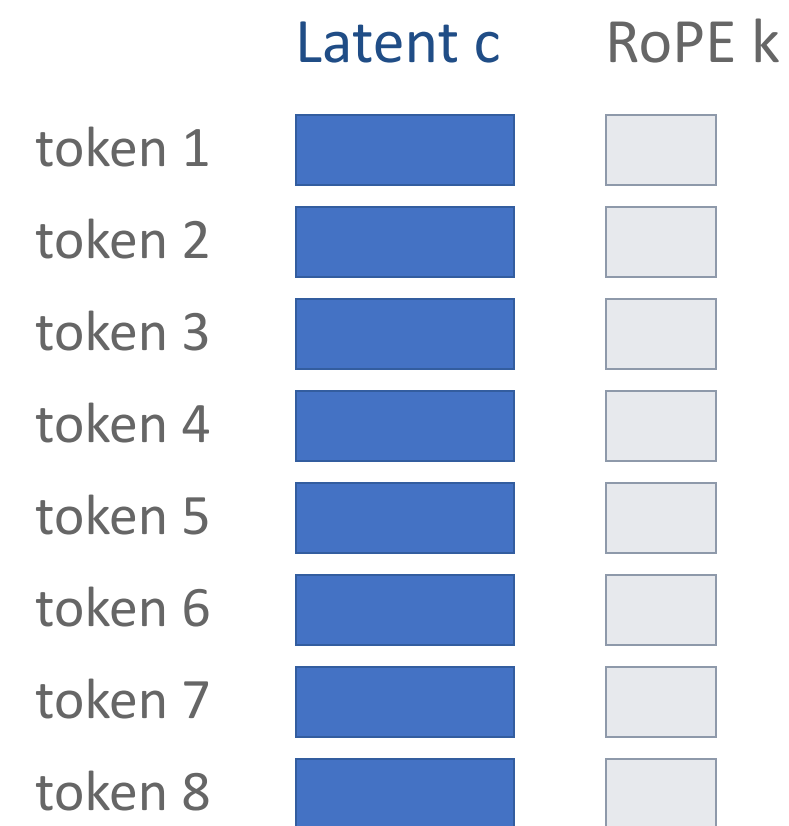
Cache a low-dimensional latent and a shared positional key for each token.

Pretrained GQA cache



Full K/V across KV heads

M²LA cache



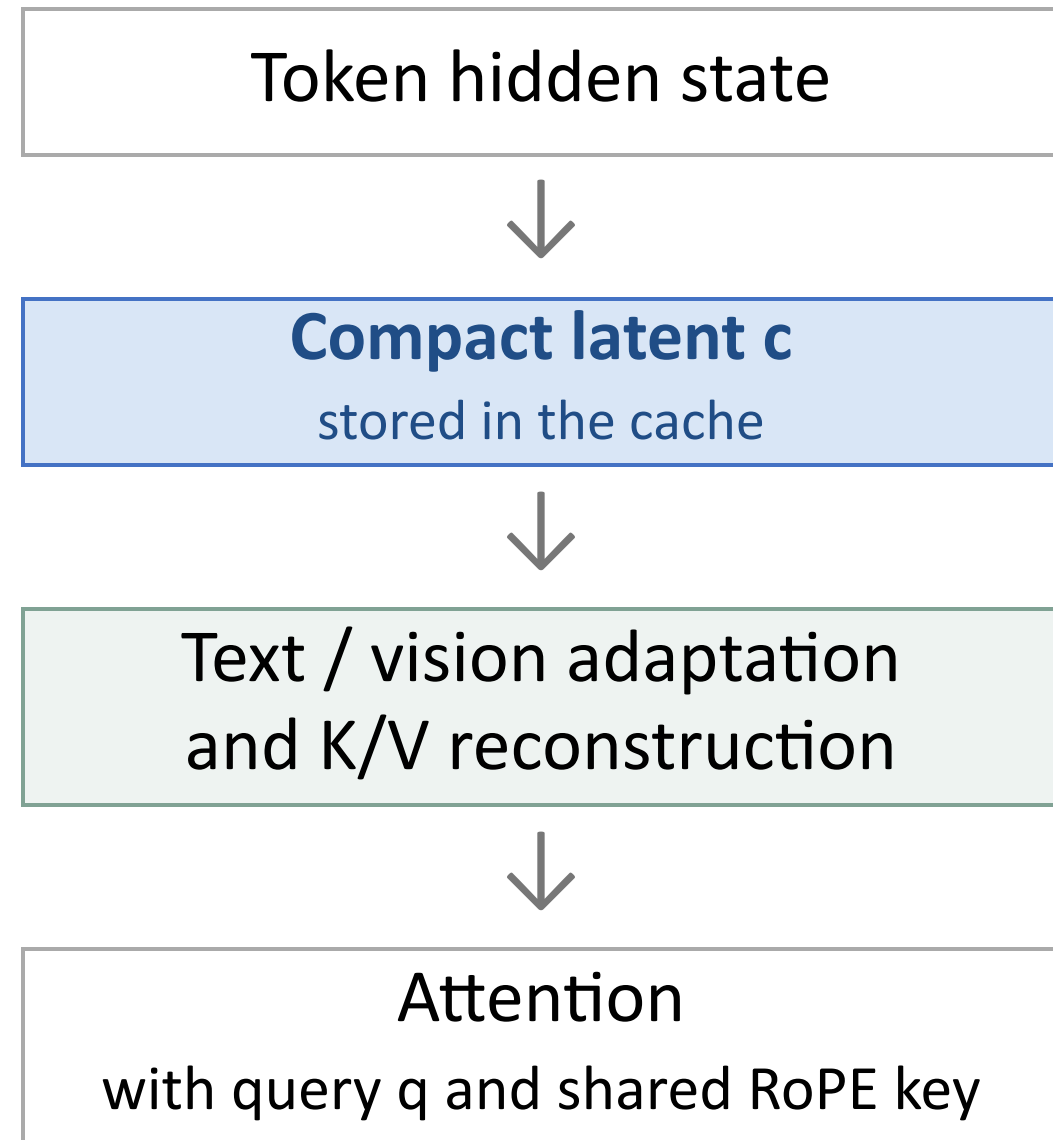
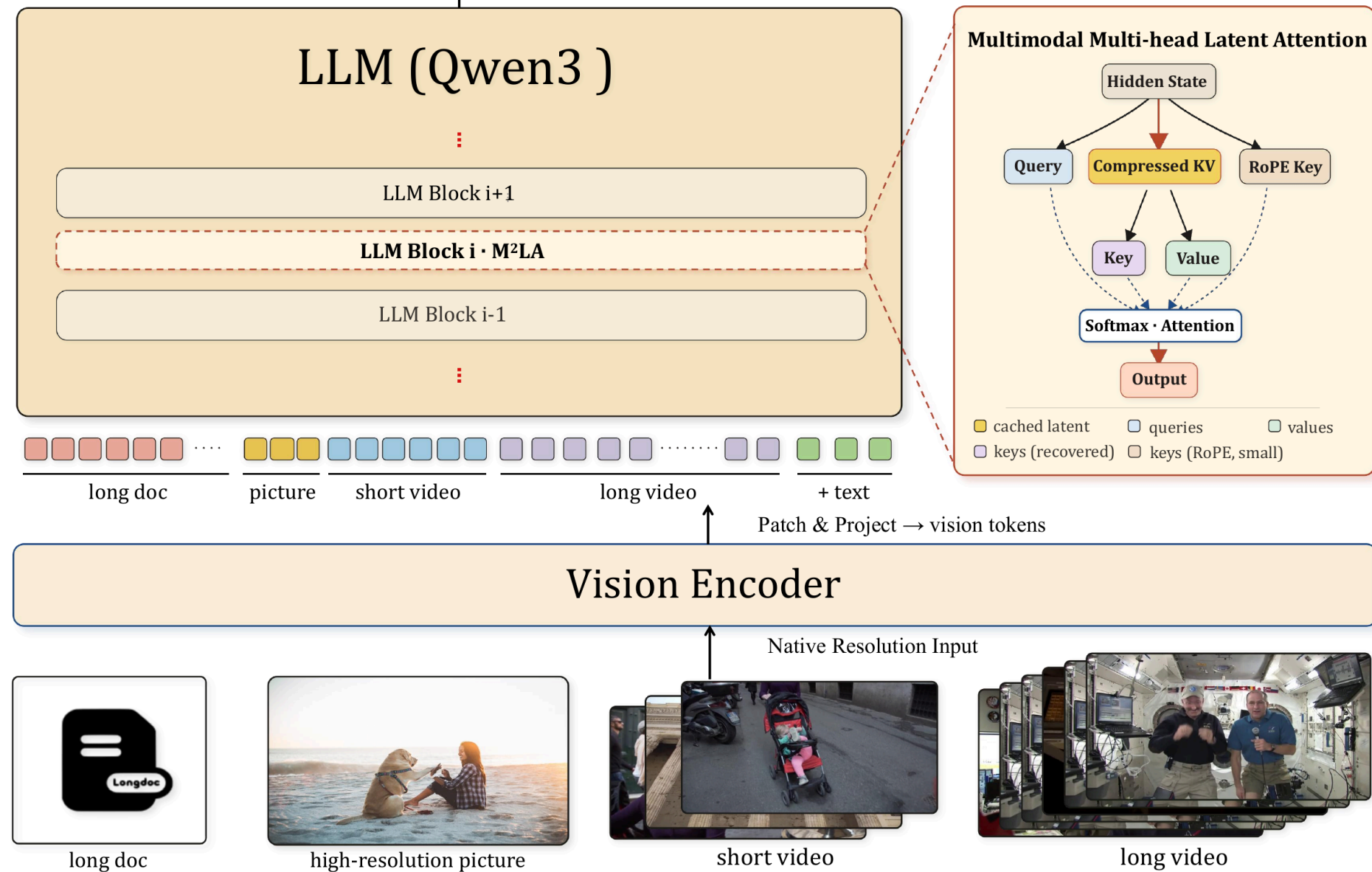
Compact state per token

Every token position remains available with a smaller cached state.

How M²LA Reconstructs Attention

The attention layer recovers head-specific keys and values from cached latents.

Response: The video opens with a man standing in a mission control room, speaking to the camera. He is surrounded by multiple computer screens displaying various data and maps, including one that shows a global map with orbital paths. The room is filled with technical equipment, including monitors and control panels, indicative of a high-tech operational environment.



The latent budget can vary by layer and modality.

Making Qwen3 Compatible with M²LA

M²LA converts pretrained attention using fixed linear projections.

1. What gets in the way?

Each Q/K head uses a scale that changes with the input. A fixed linear map cannot represent this normalization exactly.

2. What does the paper observe?

The measured scales cluster near a typical value μ .

$$k / \text{RMS}(k) \approx k / \mu$$

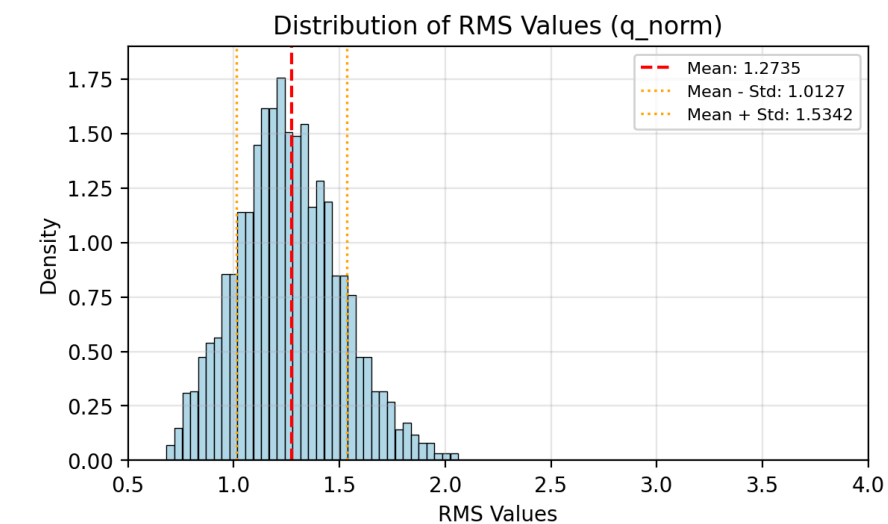
3. How is it converted?

Fit a linear substitute on calibration data, then recover capability with continued pretraining.

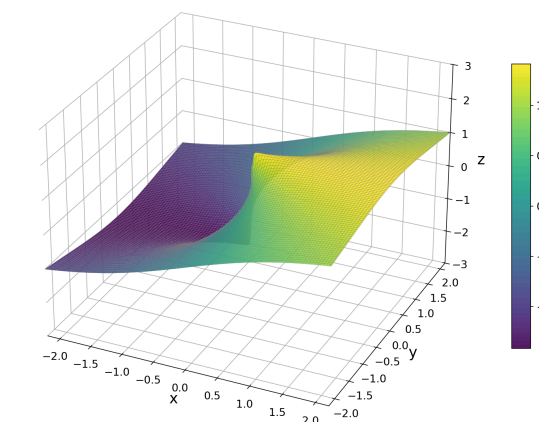
Calibrate a linear approximation, then train to recover capability.

RMS(k): vector scale; μ : typical scale. Simplified normalization intuition; Global-Norm-Linear is approximate.

Measured RMS values concentrate



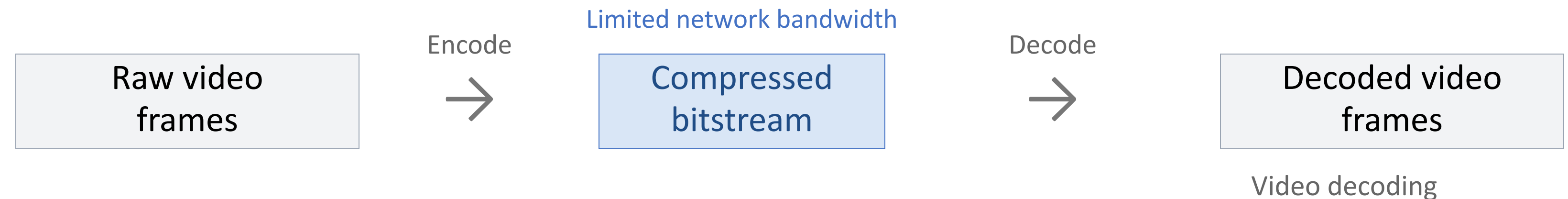
Local linear approximation is plausible



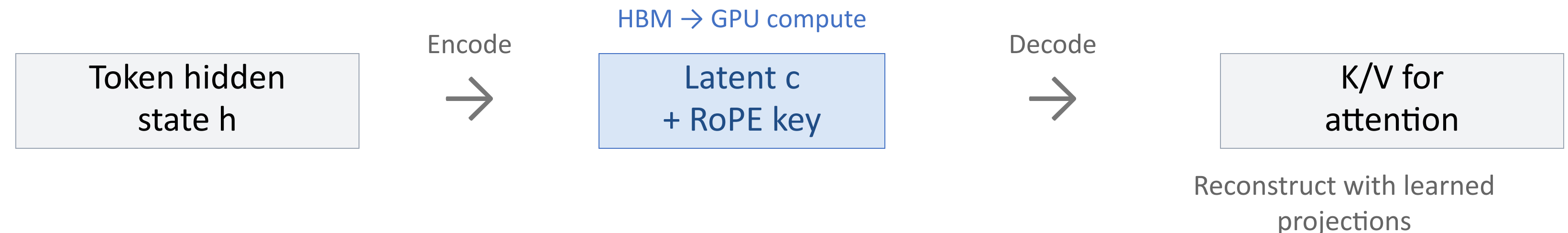
M²LA as a Codec for the KV Cache

Compression trades extra computation for fewer bytes moved.

Video over a network



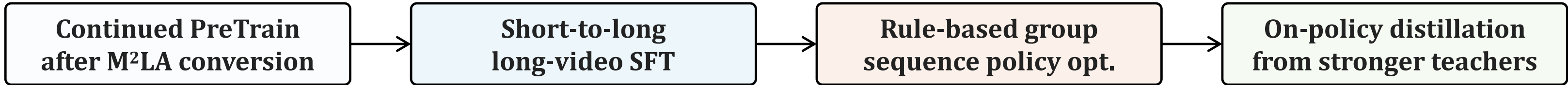
M²LA during inference



Less data movement can outweigh the extra encoding and decoding work.

Training After Attention Conversion

Restore the pretrained model, then extend its temporal and reasoning abilities.



CPT	Short-to-long SFT	Rule-based RL	On-policy Distill.
Restore language and visual alignment after attention conversion.	Increase temporal detail and context length in stages.	Reward correct QA and temporal grounding.	Guide the student on its own generated trajectories.
16M samples ~13.5B tokens	~7.2M samples 32K to 256K context	5K grounding 10K video QA samples	Qwen3-235B teacher

Efficient attention needs recovery training and long-video supervision.

Source: Training Recipe; Data Curation; train_recipe.pdf. RL uses GSPO; OPD uses reverse KL.

Teaching the Model to Read Longer Videos

Increase frame coverage and the context budget through a staged curriculum.

Stage 1

2 fps

512 frames



≈32K tokens

Stage 2

4 fps

2,048 frames

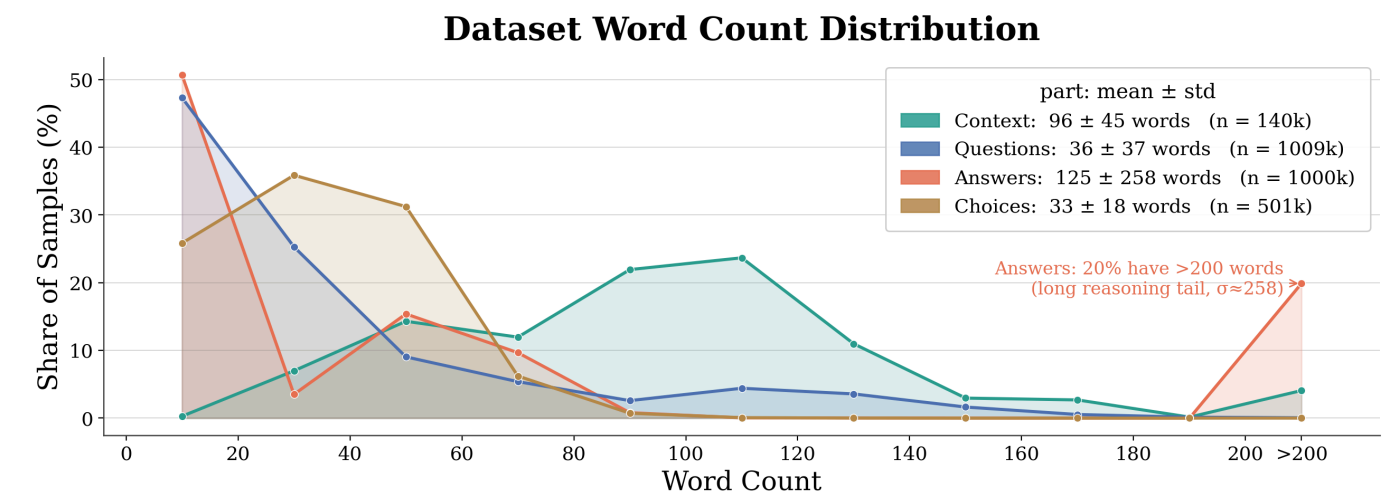
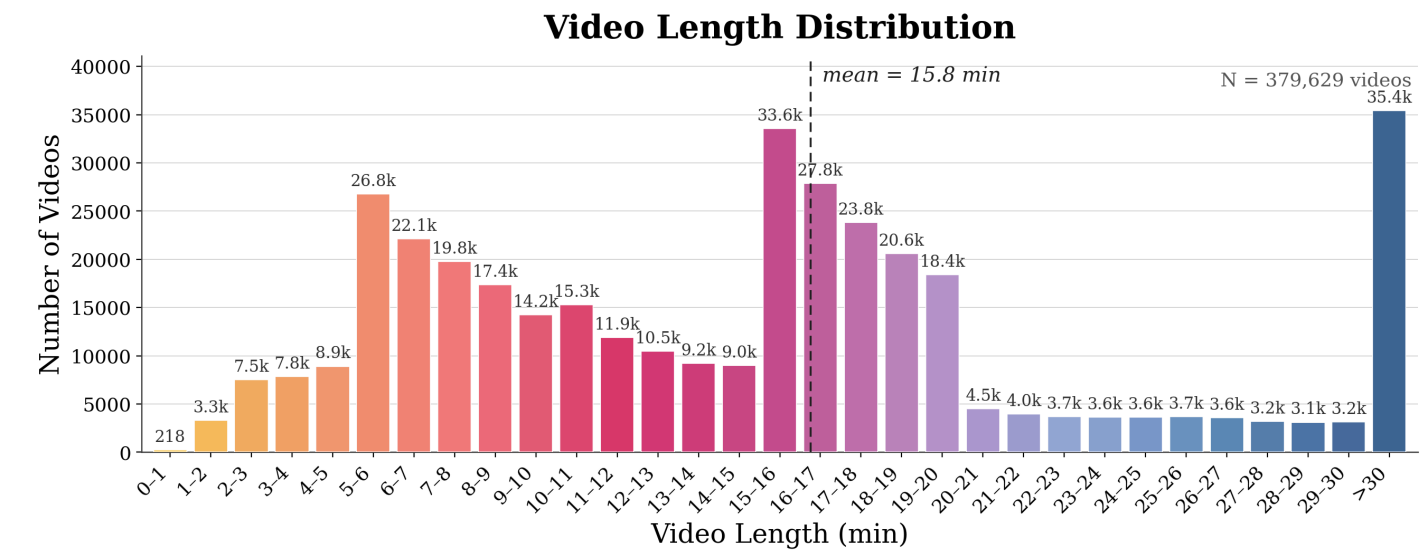
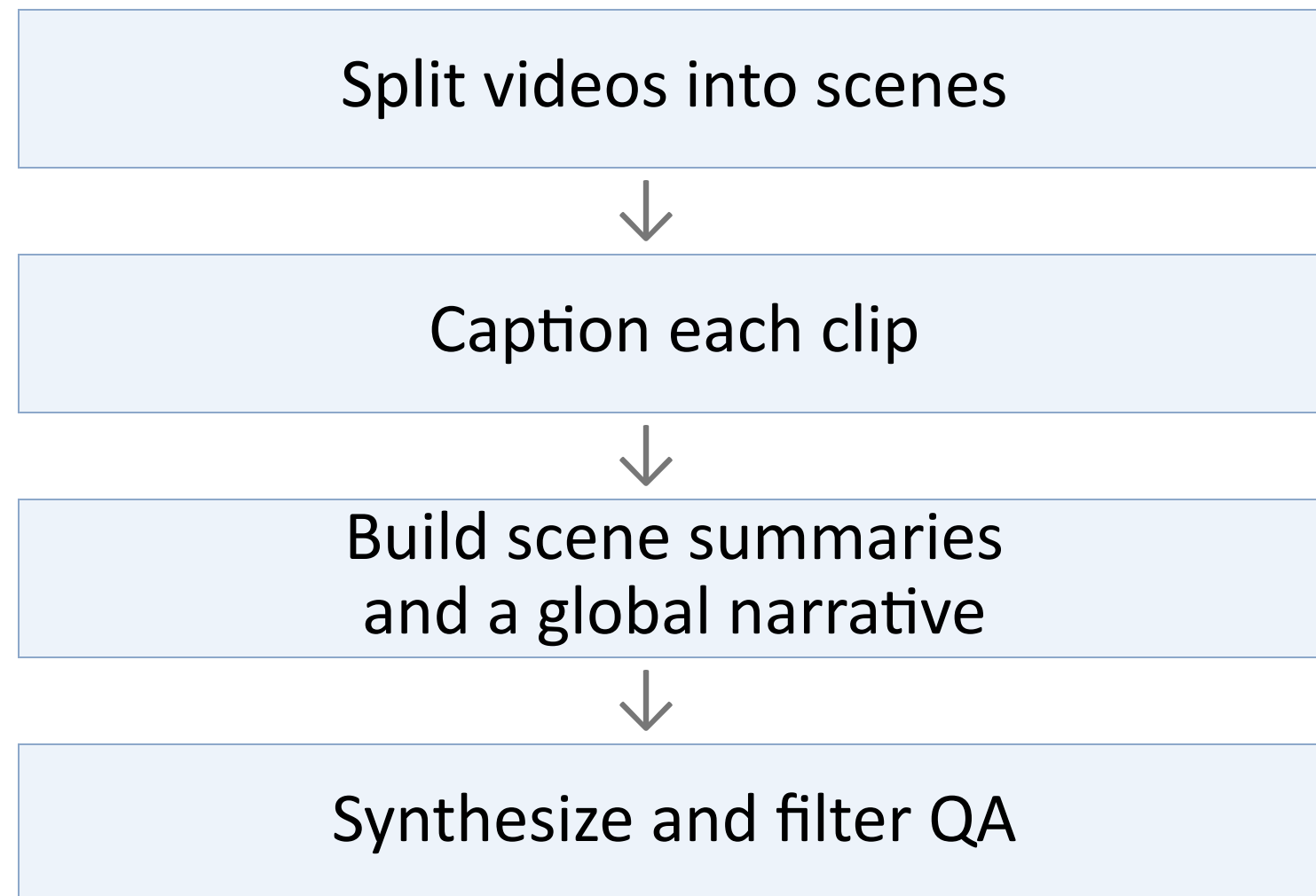


Up to 256K tokens

More frames provide temporal detail; longer contexts connect distant events.

From Long Videos to Training Signals

~380K long videos, averaging 15.8 minutes, cover about 100K hours.



Local captions preserve details; global narratives connect distant scenes.

Long-Video Understanding

Official scores versus the base family and prior model.

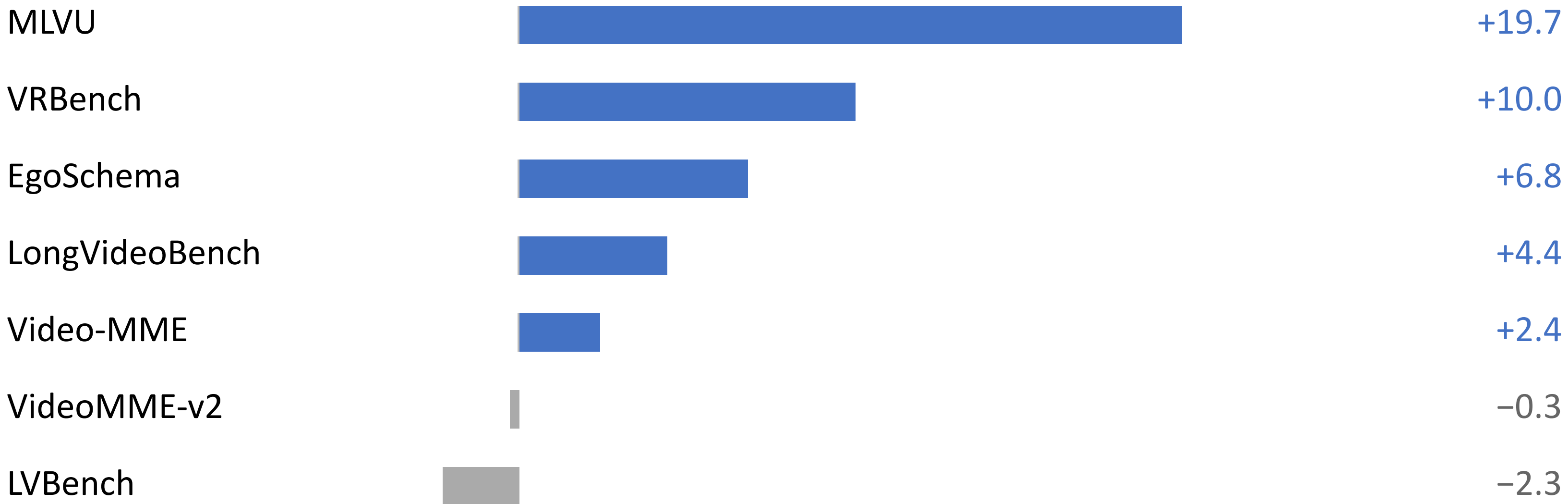
Model	Video-MME	VideoMME-v2	LongVideo Bench	MLVU	LVBench	VRBench	Ego Schema
Qwen3-VL-8B	71.4	27.9	62.4	57.6	58.0	59.4	69.8
InternVideo2.5-7B	65.1	26.4	60.6	72.8	46.4	51.9	63.9
InternVideo3	73.8	27.6	66.8	77.3	55.7	69.4	76.6

Improves on **5 of 7 benchmarks** relative to Qwen3-VL-8B.

VideoMME-v2 and LVBench remain below the base model.

Where Do the Gains Come From?

Score changes: InternVideo3 versus Qwen3-VL-8B



Gains vary by task; detail retrieval and data distribution still matter.

Short-Video Understanding

Short QA Avg.: **65.3 → 69.0** (versus Qwen3-VL-8B)

Benchmark	Qwen3-VL-8B	InternVideo2.5-7B	InternVideo3
NExT-QA	83.4	84.9	85.5
PerceptionTest	72.7	74.9	81.4
MVBench	68.7	75.7	75.0
Tomato	35.7	32.9	37.4
MotionBench	56.9	60.8	60.6
TempCompass	74.3	70.1	74.0

Source: Short-video results, selected models. 69.0 is the highest average among the listed open-weight models.

Spatiotemporal Intelligence

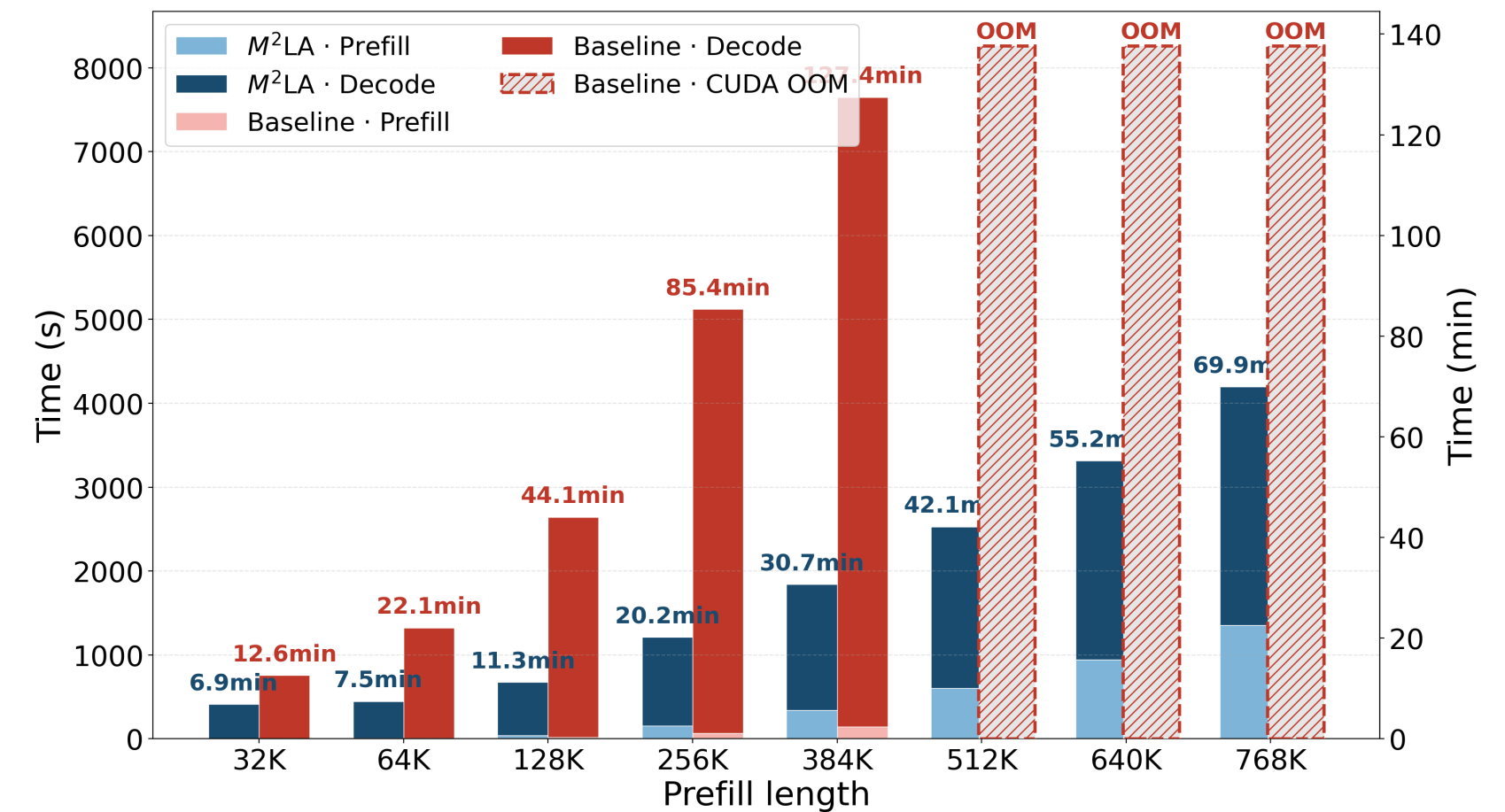
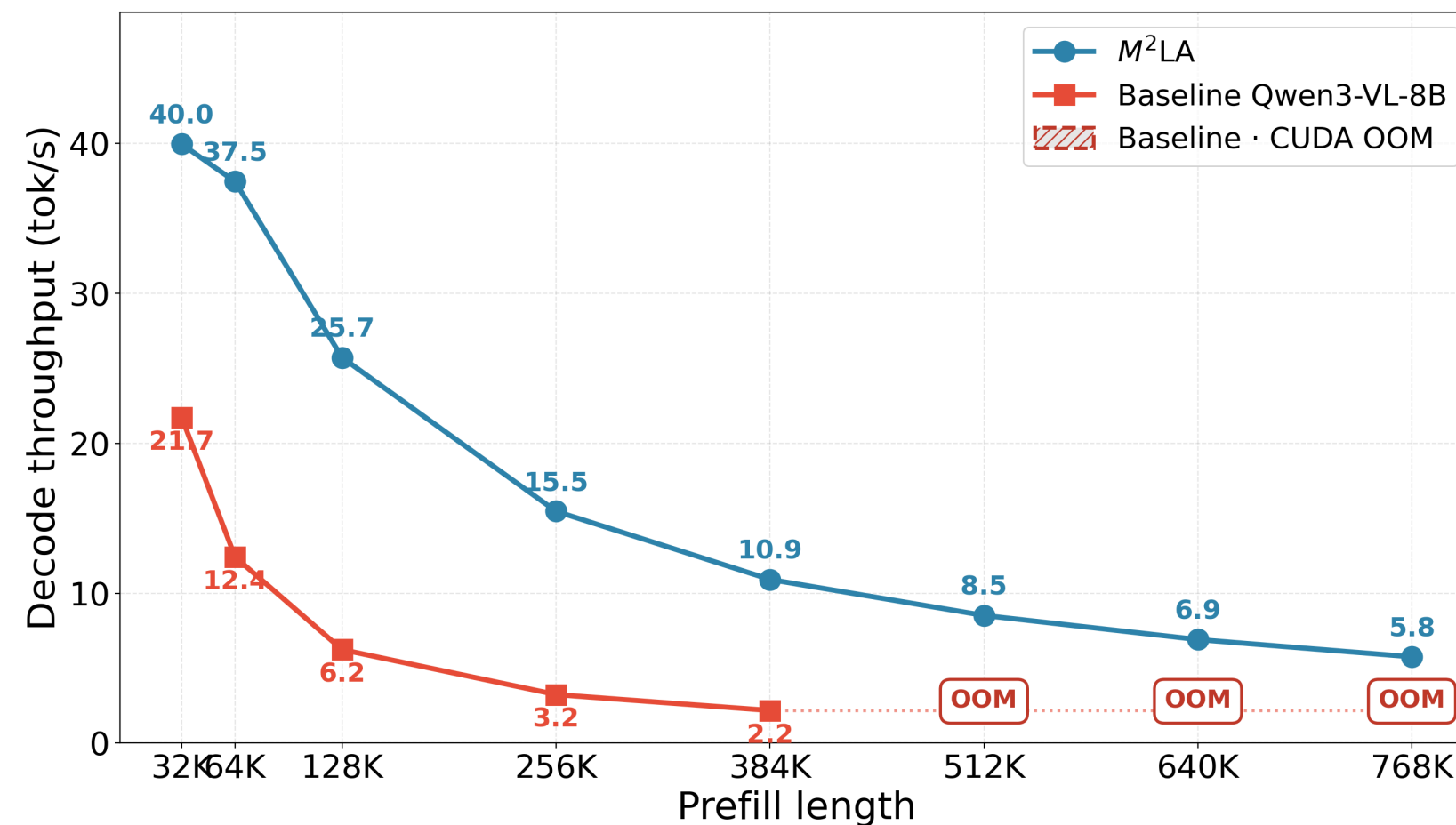
Identify when events occur and reason about spatial relationships.

Task	Benchmark	Qwen3-VL-8B	InternVideo3
Temporal Grounding	QVHighlights	59.4	59.9
	Charades-STA	48.3	50.4
	ActivityNet Captions	46.8	47.9
Spatial Intelligence	VSIBench	59.1	68.1
	MMSIBench	27.0	27.6
	MMSIBench-Video	27.6	30.7

Source: Spatiotemporal intelligence results, selected models. Each benchmark uses its official metric.

Long-Context Inference Efficiency

Single H200 · 16K decode tokens · 64K chunked prefill for inputs $\geq 256K$



384K: 5.01× decode speedup **512K:** 8.52 tok/s; base model OOM

Inference stress test reaches 768K; the training context limit is 256K.

Ablations: Architecture and Training

Removing CPT, long-context training, or long-video data lowers the average.

Variant	Video-MME	LongVideo Bench	MLVU	LVBench	Avg.
Full	73.1	64.6	75.4	54.2	66.8
w/o CPT	68.9	62.6	72.4	54.7	64.7
w/o Long Context	70.4	61.6	75.1	54.5	65.4
w/o Long-video Data	70.9	64.6	72.1	51.4	64.8

Average drops versus Full: **2.1 / 1.4 / 2.0**, respectively.

Agentic Inference: Gains and Evidence

Video-MME: Direct QA **73.1** → Agentic inference **75.8** (+2.7)

Question Why is Angelo, the main character in the video, matching his customized suit?

Video input



Key visual evidence



The video depicts a character named Angelo who is navigating complex personal and professional relationships. He is seen in various settings, including social gatherings, domestic environments, and formal settings, suggesting a rich tapestry of interactions.



Angelo is shown interacting with several characters, including a woman in period clothing, a man in a priest's attire, and another man in a suit, indicating a mix of familial and professional ties. His interactions often involve themes of loyalty, family responsibilities, and financial matters, hinting at a narrative deeply rooted in these themes.



The video also includes scenes where Angelo is depicted in formal attire, interacting with other characters in settings that suggest a blend of personal and professional contexts. This further supports the idea that his relationships and decisions are influenced by both his personal and professional life.



The presence of themes such as money, loyalty, and marriage, as well as the interactions between characters, suggest that Angelo's decisions and actions are deeply tied to his relationships and responsibilities. The scenes involving money and family responsibilities indicate that his choices are shaped by both personal and professional obligations.

Output

Because he took the glasses guy's advice about wanting to be a banker, he had two professional designers match custom suits to him in order to look more like a banker.

Conclusions and Discussion

Takeaway

Design evidence accumulation with cache efficiency.
Recovery training and video supervision make long contexts useful.

Scope

Focus: GQA conversion and video understanding.
Agent evaluation remains preliminary and video-centric.

Open Questions

How do tool errors affect later reasoning?
Does this extend to GUI, browser, embodied, and multi-session tasks?

Questions?