GridProbe: Posterior-Probing for Adaptive Test-Time Compute in Long-Video VLMs

Mohamed Eltahir[1,2] Lama Ayash[1] Ali Habibullah[1]

Tanveer Hussain[3] Naeemullah Khan[1]

[1] KAUST Academy, King Abdullah University of Science and Technology (KAUST),
[2] Islamic University of Madinah (IUM),
[3] Edge Hill University

PROBLEM STATEMENT

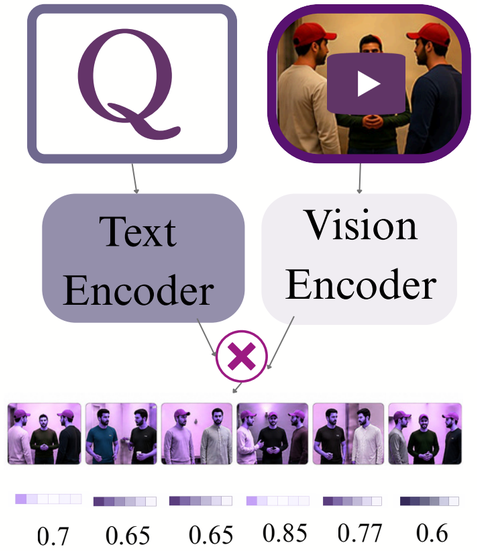
Long-video VLMs are expensive because they process thousands of frames in a single monolithic forward pass → $\mathcal{O}(n^2)$



PROBLEM STATEMENT

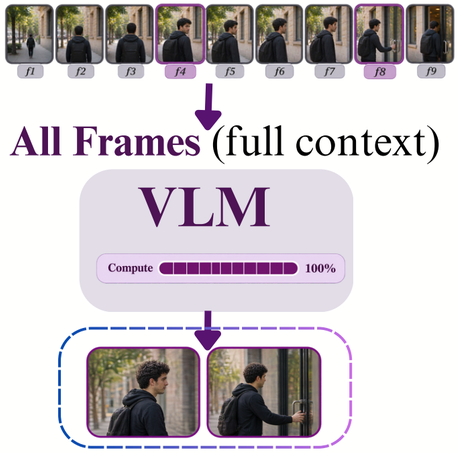
① CLIP-based

Similarity, not reasoning. Each frame is scored independently.



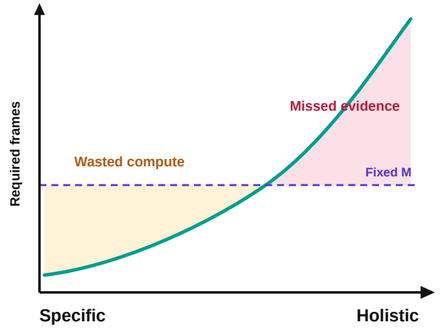
② VLM-based

Long context again... and no per-frame score.



③ Fixed budget M

One budget cannot fit all questions: specific questions waste compute, holistic questions miss evidence.



The Question

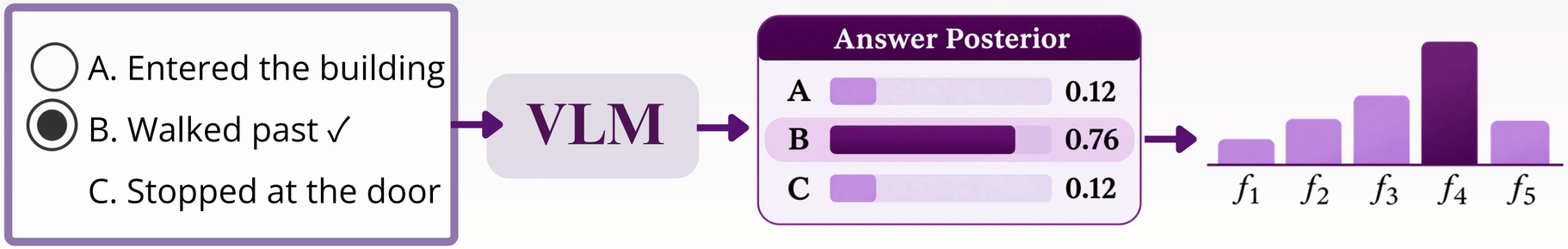
Can we **reason like the VLM**, capture both **local and global** context without paying the long-context cost, and **adaptively** select how many frames each question needs?

CONTRIBUTIONS

1 Answer-Space Posterior Probing

A Jev-like decision model!!

Use VLM confidence to score a sequence of frames (a band)



CONTRIBUTIONS

2 Local-Global Grid Factorization

like LoRA, on the input!!

Grid (bands) scoring → Outer Product (mix contexts) → per-frame score!

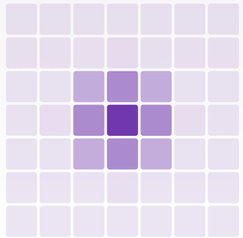


CONTRIBUTIONS

3 Shape-Adaptive Test-Time Compute

Track **1)** frames required **2)** question difficulty using the map's distribution shape!

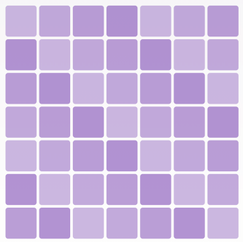
Sparse evidence



M_{eff}

LOW

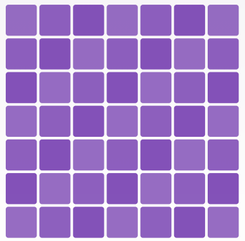
Distributed evidence



M_{eff}

HIGH

Redundant evidence



M_{eff}

LOW

METHODOLGY



Where is the man's family most likely to be when he crosses the canyon in the video?

Q

A. Waiting in the car.

B. Camping in the canyon.

C. Staying in the hotel.

D. At home.

Standard Inference

GridProbe Inference

K^2 *Frames*

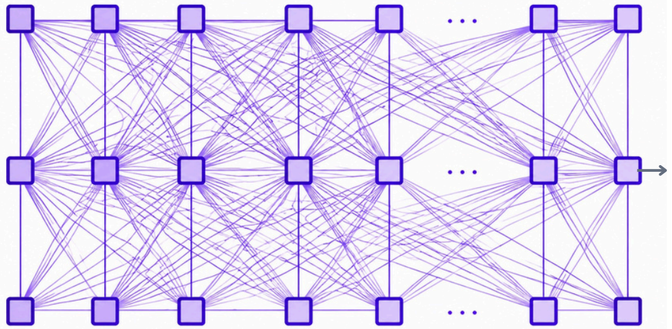


$\mathcal{O}(N^2)$

VLM

1 *Pass*

Full temporal self-attention



Answer: C. Staying in the hotel.

K^2 *Frames*

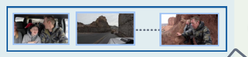


Input Grid

$K * K$

Stage 1: Probe

K Row Passes



K Column Passes



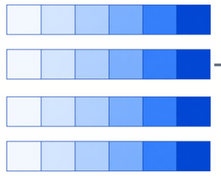
$\mathcal{O}(N^{1.5})$

VLM

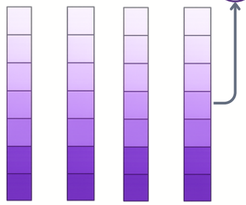
2K *Passes*

K *Frames Each*

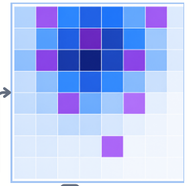
Rows Confidence



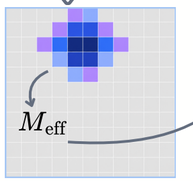
Columns Confidence



Importance Map M



$\sigma(M)$



M_{eff}

Stage 2: Focused

$\mathcal{O}(M_{\text{eff}}^2)$

VLM



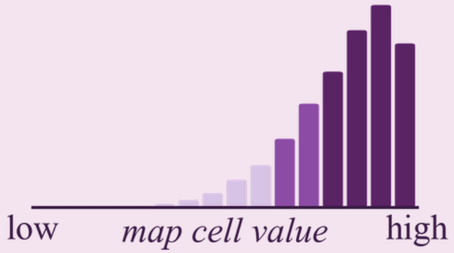
Top M_{eff} *Frames*

Answer: C. Staying in the hotel.

METHODOLGY

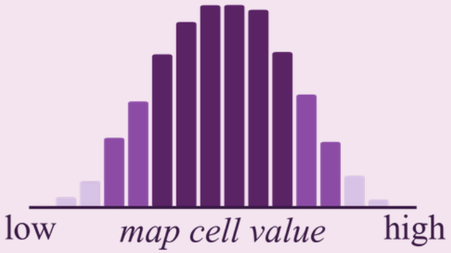
The Shape Sets the Budget

Redundant



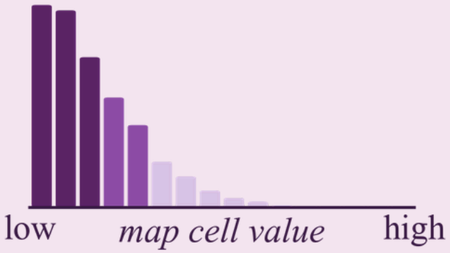
skew $< 0 \rightarrow M_{\text{eff}}$ low

Holistic



skew $\approx 0 \rightarrow M_{\text{eff}}$ high

Specific



skew $> 0 \rightarrow M_{\text{eff}}$ low

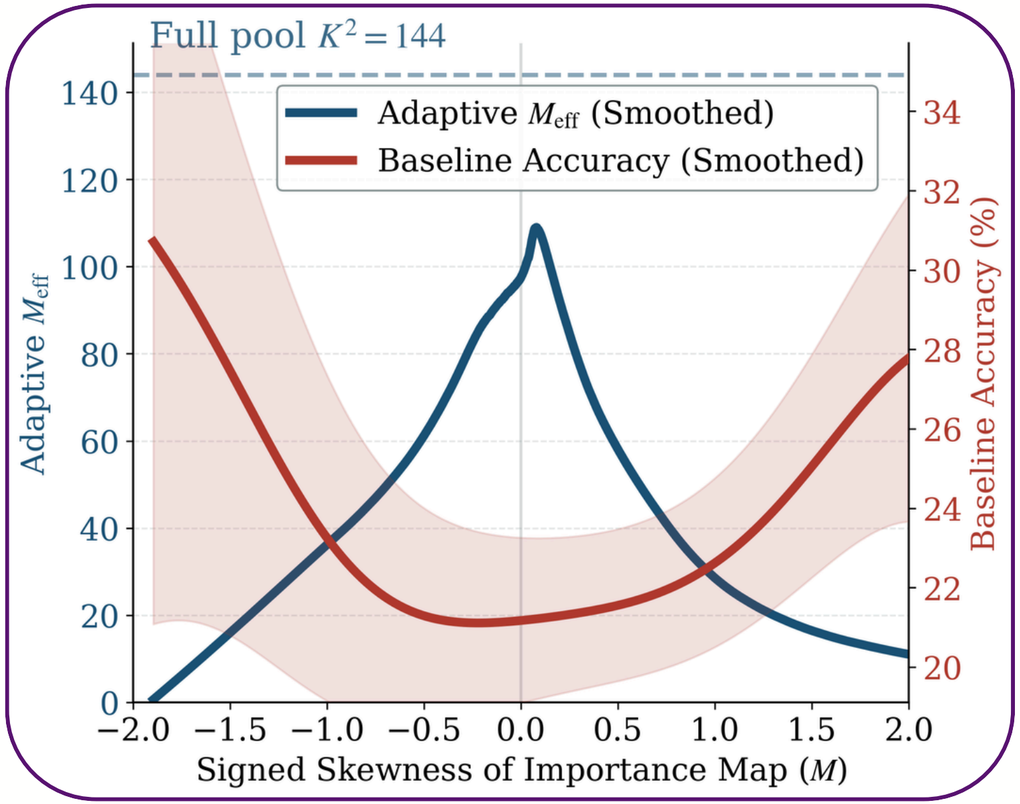
Shape-Adaptive Selection:
skewness and kurtosis capture the map's shape, and the shape sets each question's frame budget.

$$\sigma = |skew(M)| + \lambda \cdot kurt^{+}(M)$$

$$M_{\text{eff}} = \left\lceil \frac{K^2}{1 + \gamma_0 K \sigma(M)} \right\rceil$$

METHODOLGY

The Budget Follows Difficulty

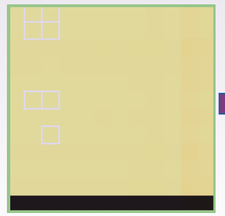


METHODOLGY

Redundant

Where does the video take place? Q

Importance Map M



Top-1 Frame

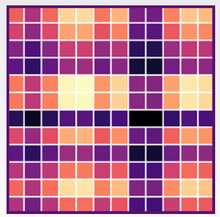


$\sigma = 6.6$ / $M_{\text{eff}} = 7$

Holistic

What is the underlying meaning of this animation? Q

Importance Map M



Top-1 Frame

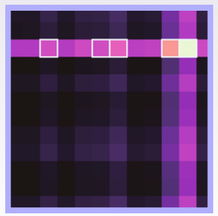


$\sigma \approx 0$ / $M_{\text{eff}} = 140$

Specific

According to the video, what is the price of an organic tomato? Q

Importance Map M



Top-1 Frame



$\sigma = 9.6$ / $M_{\text{eff}} = 5$

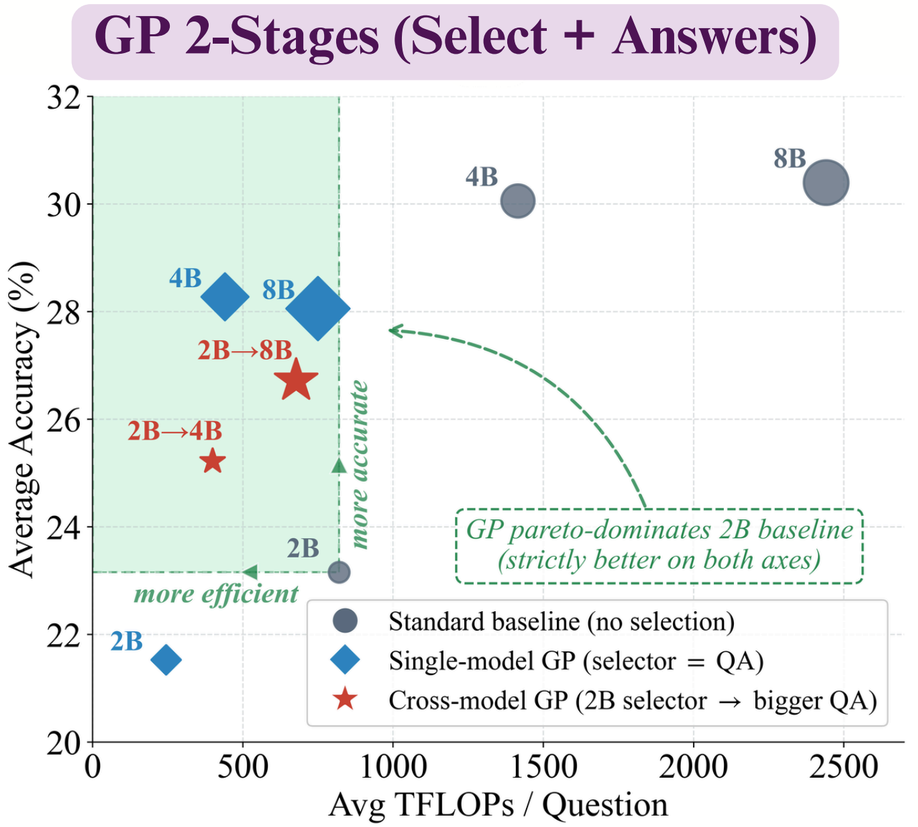
RESULTS

Same model:

~3× less compute, within 3 points!

Cross model (2B picks, 8B answers):

+3.6 points at 0.83× the 2B's compute!



RESULTS

Accuracy-Latency Trade-off

Method	Accuracy (%) ↑	TTFT (s) ↓
Uniform (144 frames)	44,6	1,9
HERMES	42,4	**1,4**
AVP	43,6	81,3
GridProbe (Ours)	**47,0**	5,4

RESULTS

Question-Conditioned Evidence Ranking

Query Condition	Mean Normalized Rank ↓
Its own question	0,44
Chance	0,5
Another video's question	0,58

RESULTS

Paper

Code

What happens when we extend this to images and text?👀

Thank You