# StreamReady: Learning *What* to Answer and *When* in Long Streaming Videos

**Shehreen Azad**[1], Vibhav Vineet[2], Yogesh S Rawat[1]

[1] Center for Research in Computer Vision, University of Central Florida;

[2] Microsoft Research

**CVPR 2026**

CVPR JUNE 3-7, 2026 DENVER COLORADO

# Streaming Video Understanding

**Causal understanding**

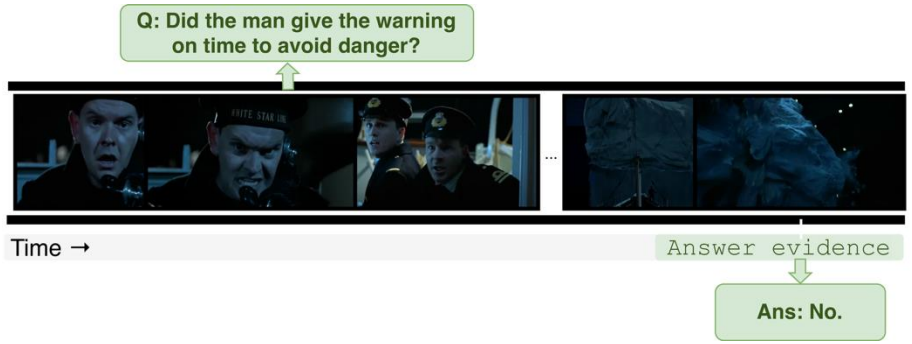
- Current or past-dependent
- Example: “Which team’s player started the match?”

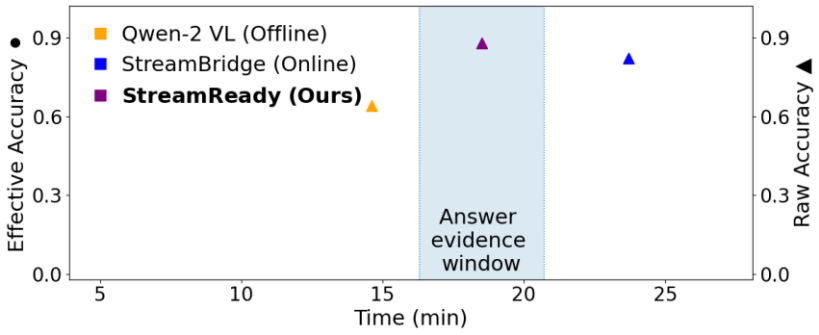
**Proactive understanding**

- Future-dependent
- Example: “Output GOAL when a goal has been scored.”
- When is there enough evidence to answer?

# Current Methods Ignore Answer Timing

- Answer content focused → ignore timing.
- No notion of evidence sufficiency.
- Cannot distinguish between mistimed and on-timed answer.

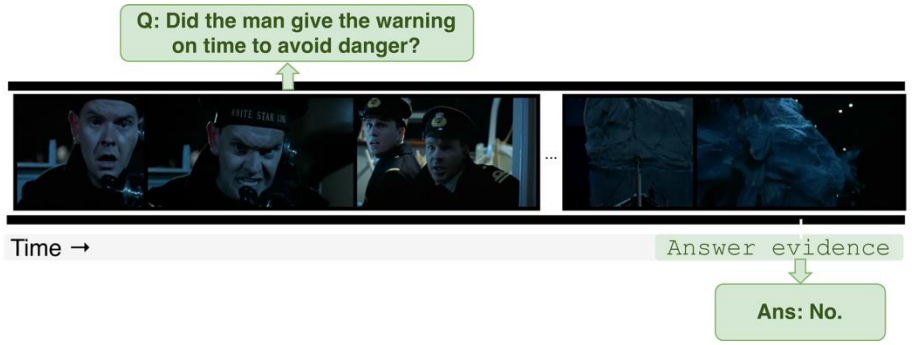


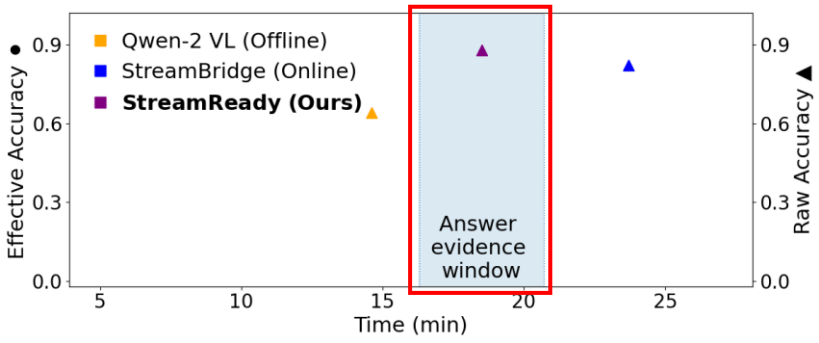



# Towards Timing-Aware Answering

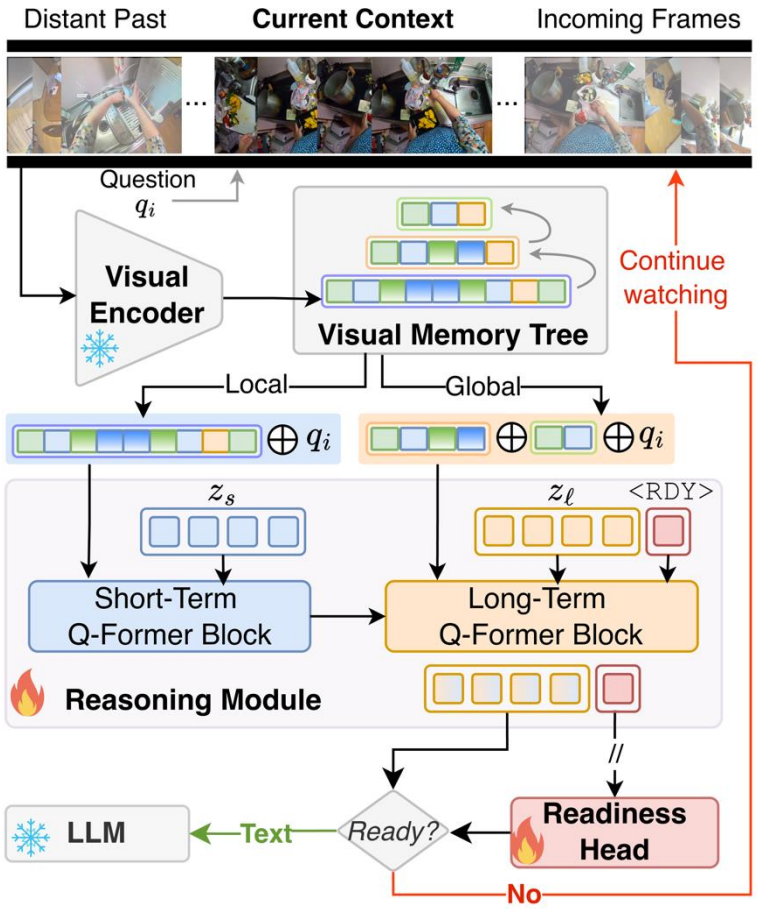
- Readiness-aware streaming understanding.

- Models should answer on-time based on visual evidence.



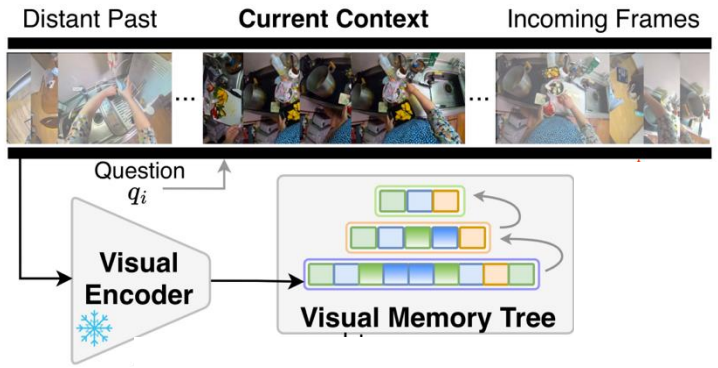



# Framework Overview




- Task-agnostic storage.
- Task-aware reasoning.
- Evidence sufficiency measured throughout the video.
- Answers when readiness exceeds threshold.

# Task-Agnostic Storage

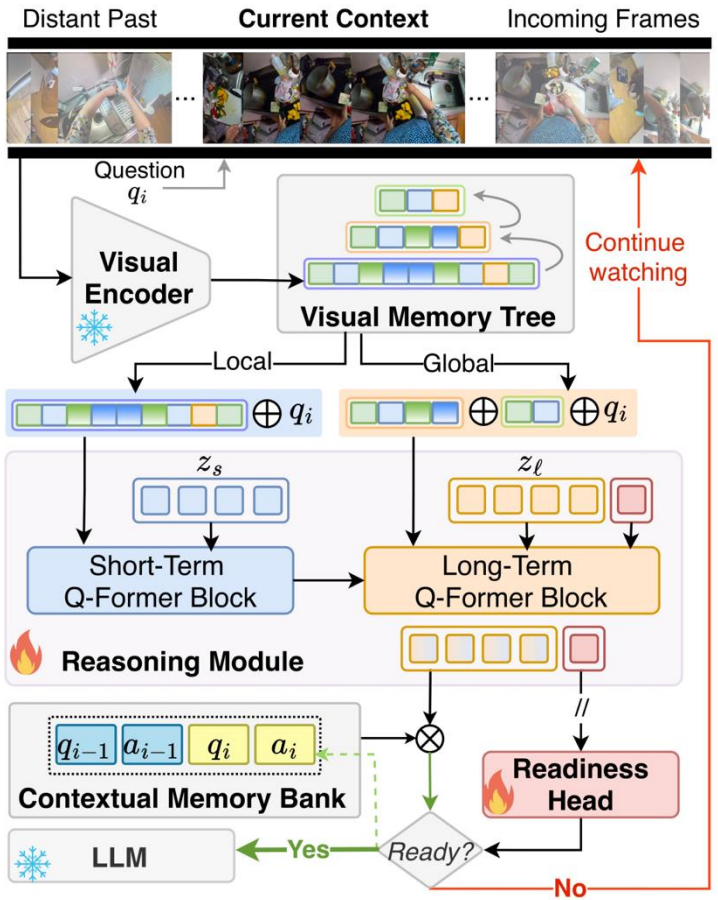



$$c_j \leftarrow \begin{cases} (1-\alpha)c_j + \alpha f_o, & \text{if } \mathrm{sim}(f_o, c_j) \geq \tau_t \\ \text{new centroid}, & \text{otherwise} \end{cases}$$

**Visual Memory Tree**

- Multi-level with increasing abstraction.
- Lowest level: short term details.
  - FIFO update.
  - Evicted frames → next level.
- Higer levels: long term summary
  - Initialized by clustering (K-means)
  - Updated with EMA
- When question appears → task-aware reasoning.

# Readiness Monitoring




- Introduce learnable token in long-term reasoning block of HierarQ.
- Readiness head monitors token and assigns high score when sufficient evidence.
- If score exceeds threshold, trigger LLM.

# Learning Readiness

- No answer evidence ground truth
- Similarity to memory and current answer representation.
  - High similarity → pseudo-positive, Low similarity → pseudo-negative.
- Optimize with contrastive loss.

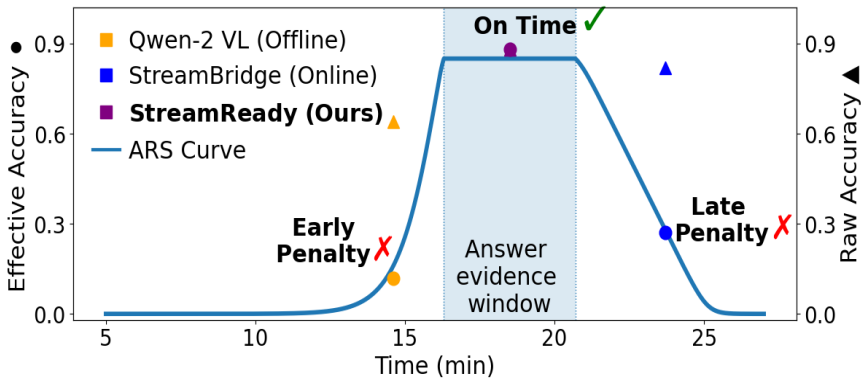

$$\mathcal{L}_{ctr} = -log\ \sigma\ \left(R_{pred}(t^{+}) - R_{pred}(t^{-})\right)$$

# Answer Readiness Score (ARS)

- Timing-aware evaluation metric to reward on-time answers

$$\text{ARS} = \frac{1}{N}\sum_{i=1}^{N}(\text{EP}_i \cdot \text{LP}_i); \quad \in [0,1]$$

- Harsh early penalty
- Mild late penalty
- Defines effective accuracy




$$\text{EP} = \text{softmin}\left(1, 2\,\sigma\left(\gamma_e \frac{t_a - t_s}{\tau + \epsilon}\right)\right)$$

$$\text{LP} = \text{softmin}\left(1, \text{softmax}\left(0, 1 - \gamma_\ell \frac{t_a - t_e}{\tau + \epsilon}\right)\right)$$

$$\text{Acc}_e = \text{Acc} \times \text{ARS}$$

# Proactive Readiness Benchmark


**ProReady-QA**

- 5 proactive understanding tasks with answer evidence window.
- 30 min – 1 hour long videos.
- 5000 QA pairs.

# Results

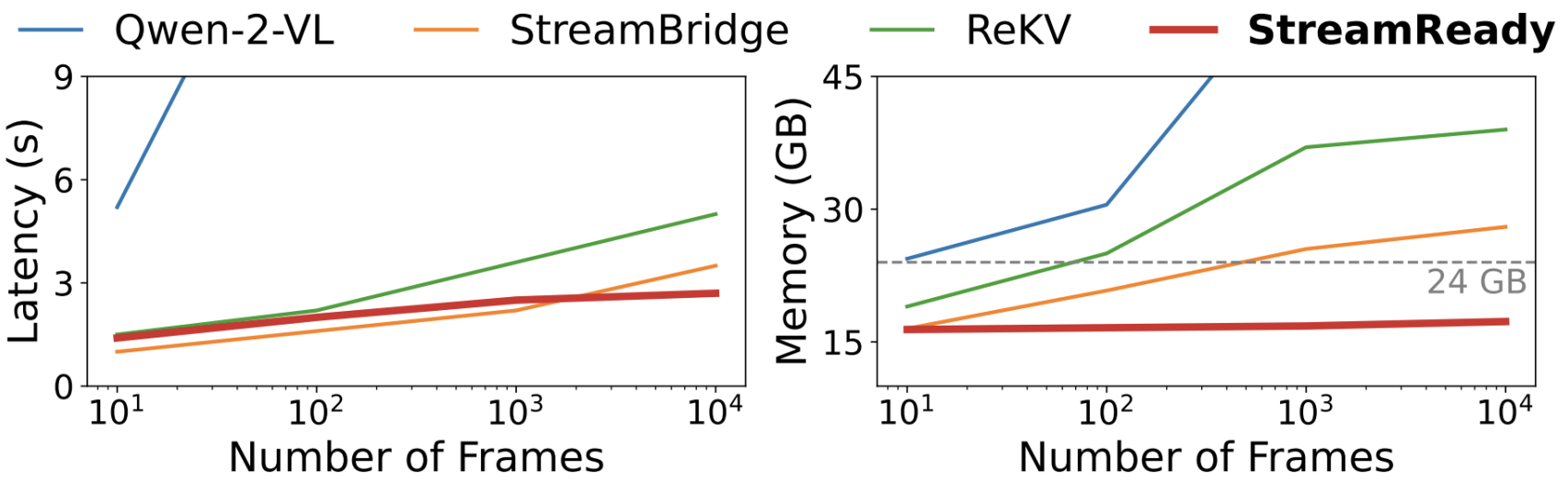
- 9 benchmarks spanning ~2 min to ~1 hour.
- Task: VQA (streaming understanding, offline understanding)
- Evaluation metric: Accuracy, ARS, Effective accuracy

| Method | StreamingBench (1-10 min) | ProReady-QA (30 min – 1 h) | | |
|---|---|---|---|---|
| | Acc | Acc | ARS | Eff. Acc |
| HierarQ (CVPR'25) | 35.1 | 46.0 | 0.40 | 27.0 Δ-19.0 |
| Previous Best StreamBridge(NeurIPS'25) | 43.9 | 53.1 | 0.60 | 42.3 Δ-10.8 |
| **StreamReady** | **48.2** +4.3 | **56.4** + 3.3 | **0.69** +.09 | **53.2** Δ-3.2 |

# Contribution of Each Component

| Method | ProReady-QA | |
|---|---|---|
| | Acc. | ARS |
| Baseline | 28.7 | 0.44 |
| + Task-Agnostic Storage | $32.4_{+3.7}$ | $0.46_{+0.02}$ |
| **+ Task-Aware Reasoning** | $39.4_{+7.0}$ | $0.48_{+0.02}$ |
| **+ Readiness Monitoring** | $\mathbf{39.6}_{+0.2}$ | $\mathbf{0.68}_{+0.20}$ |

# Cost Analysis




Lightweight readiness monitoring system and task-agnostic storage keeps cost minimum.

# Thank You