# StreamBridge:

## Turning Your Offline Video Large Language Model into a Proactive Streaming Assistant

# Streaming Video Understanding

## · Multi-turn Real-time

Incoming frames

00:00 00:24 01:26 05:45

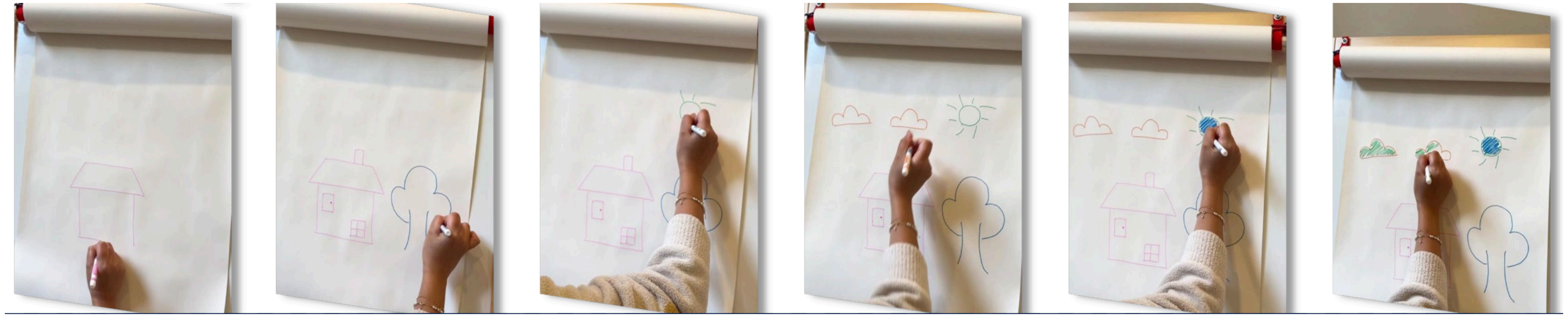
00:24 **USER:** What words are <u>currently</u> shown.
**ASSISTANT:** UNIVERSIAL STUDIO HOLLYWOOD

01:26 **USER:** What words are <u>currently</u> shown.
**ASSISTANT:** STUDIO STORE

05:45 **USER:** What's the white rabbit doing?
**ASSISTANT:** The White rabbit is nodding its head and swaying on the …

## · Proactive Responses

**USER:** How do I draw this picture?



Incoming frames

**ASSISTANT:** Draw a house and a tree.

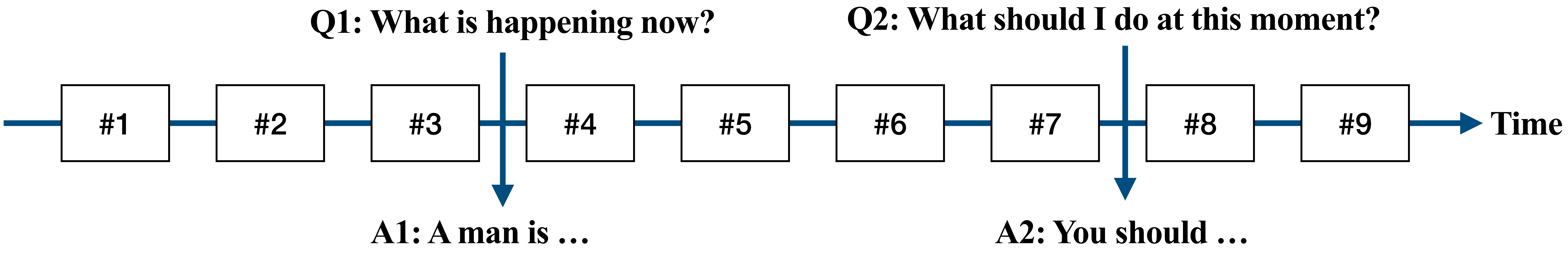
**ASSISTANT:** Draw a sun and two clouds.

**ASSISTANT:** Color the sun blue and the clouds green.

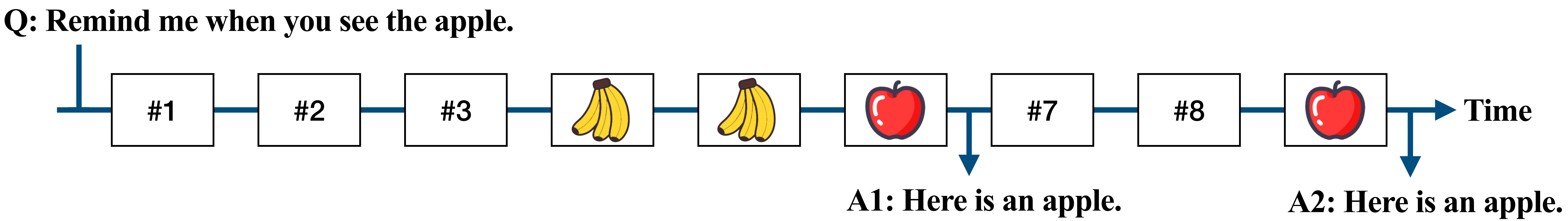
*Junming Lin, Zheng Fang, Chi Chen, etc. "STREAMINGBENCH: ASSESSING THE GAP FOR MLLMS TO ACHIEVE STREAMING VIDEO UNDERSTANDING"*

# Preliminary Analysis from the Perspective of Input Sequences

- **Multi-turn dialogue with interleaved video-text:** can be adapted from offline models w/o training.
  - **Format:** <V1> <Q1> <A1>, <V2> <Q2> <A2>, …




- **Proactive Responses:** a time interval between user query and assistant response.
  - **Format**: <Q>, <V1> <A1>, <V2> <A2>, …



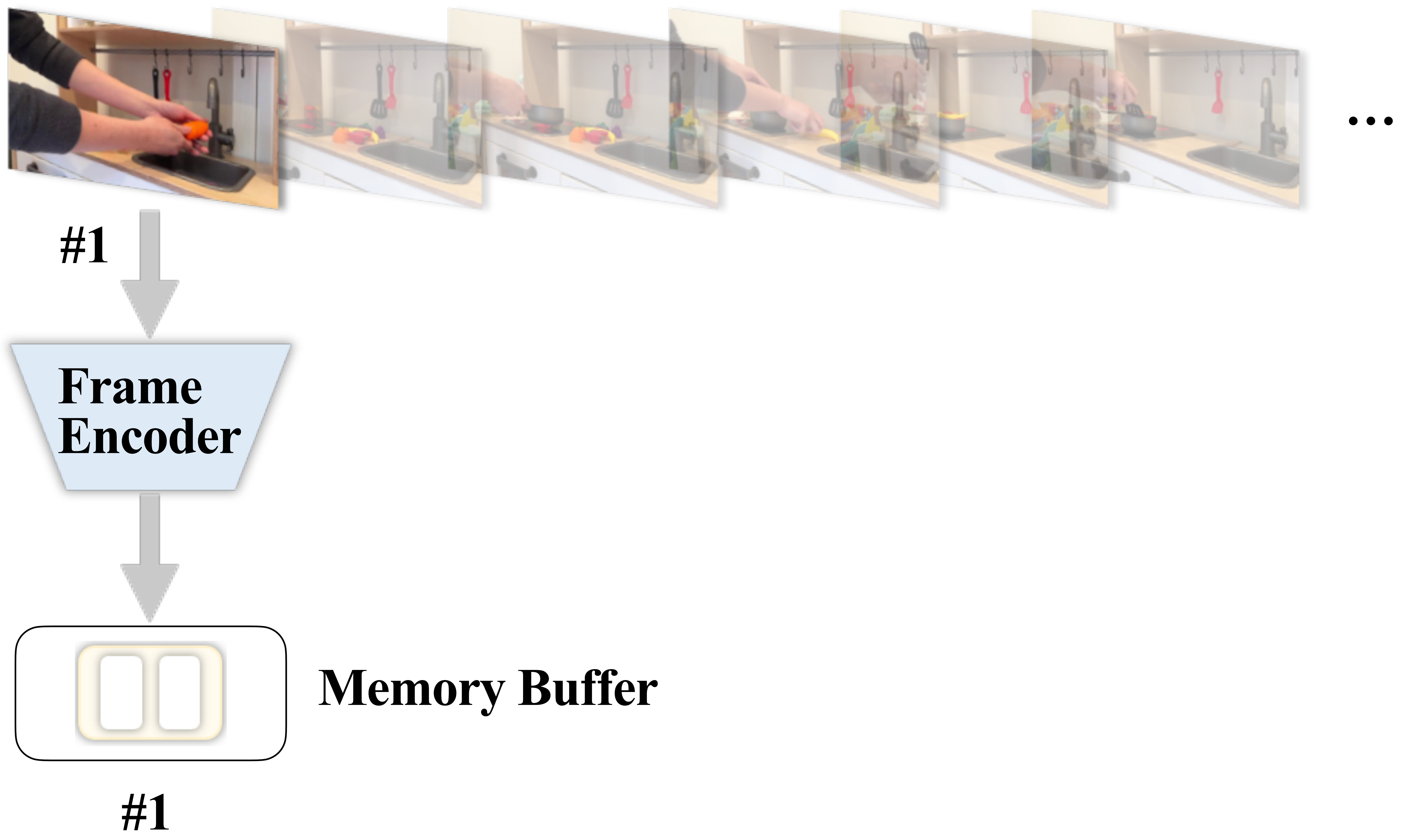
# Motivation

## How to equip current offline Video-LLMs with streaming interactive capabilities?

- A buffer to manage incoming frames and historical dialogues in streaming videos
- A compression strategy for infinite streaming contexts and real-time understanding
- An activation mechanism for proactive responses
- A dataset tailored for SFT with diverse streaming scenarios

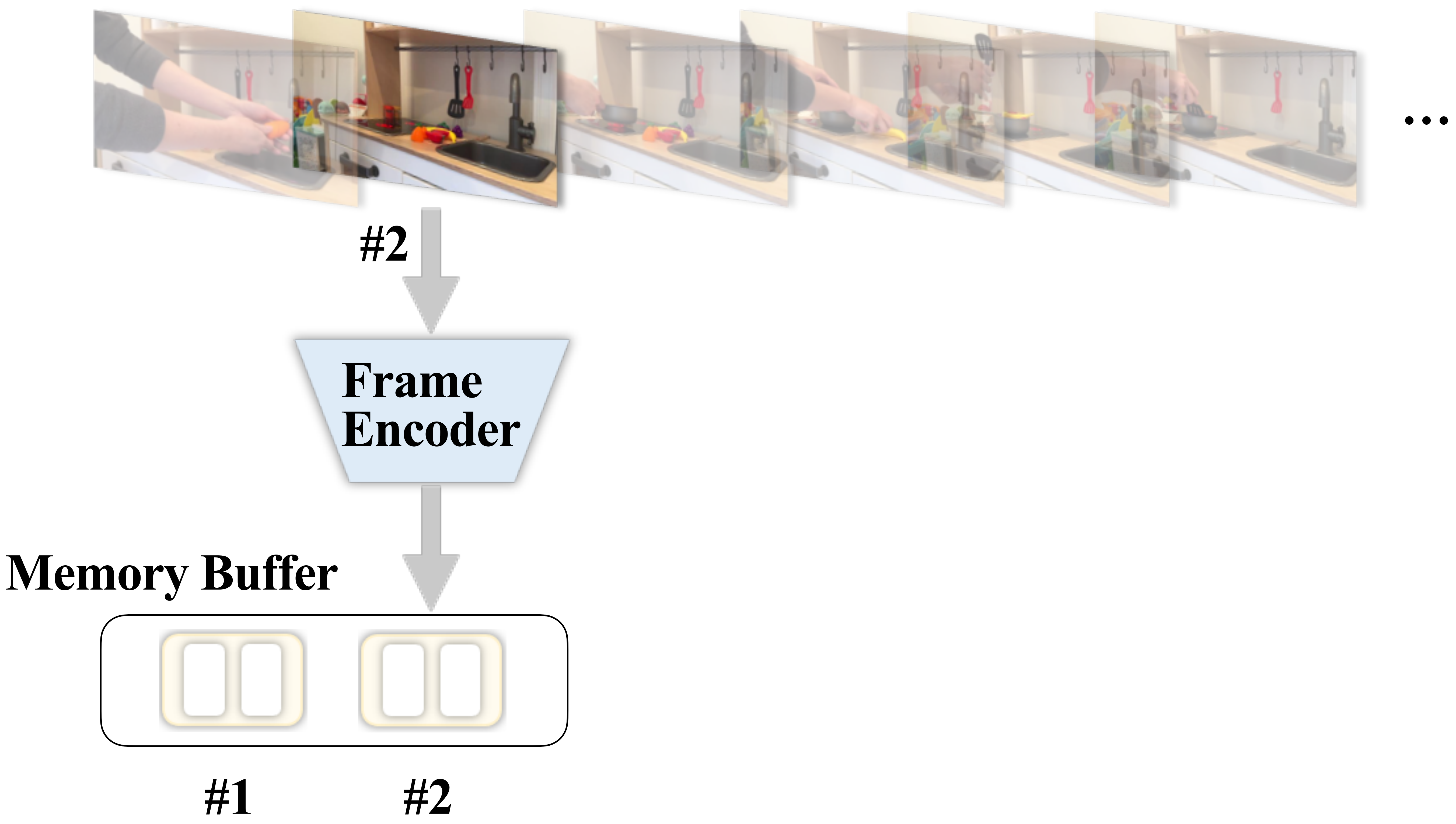
# StreamBridge-Framework

- In streaming scenarios, incoming frames are encoded one by one, and stored into the memory buffer



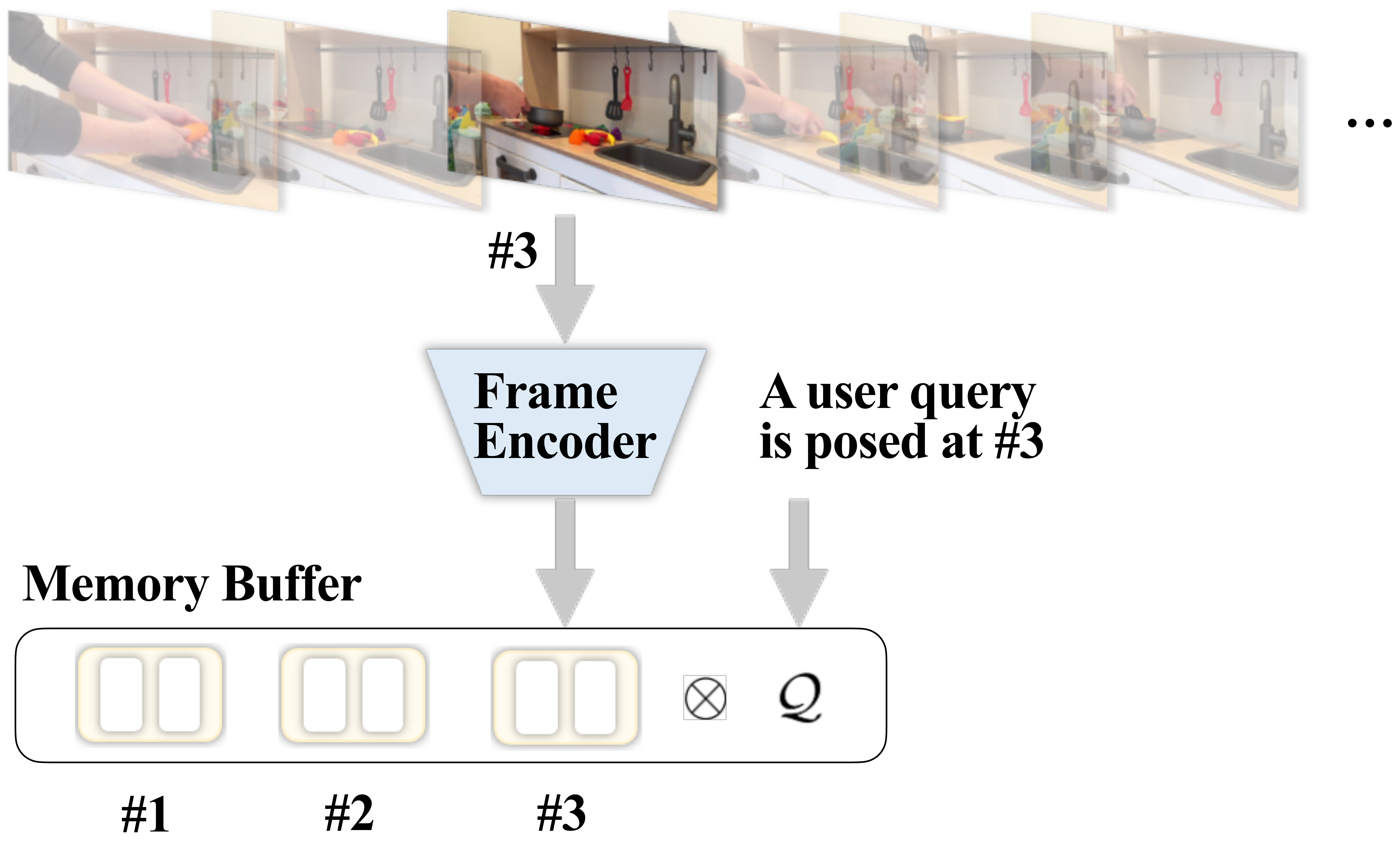
# StreamBridge-Framework

- In streaming scenarios, incoming frames are encoded one by one, and stored into the memory buffer



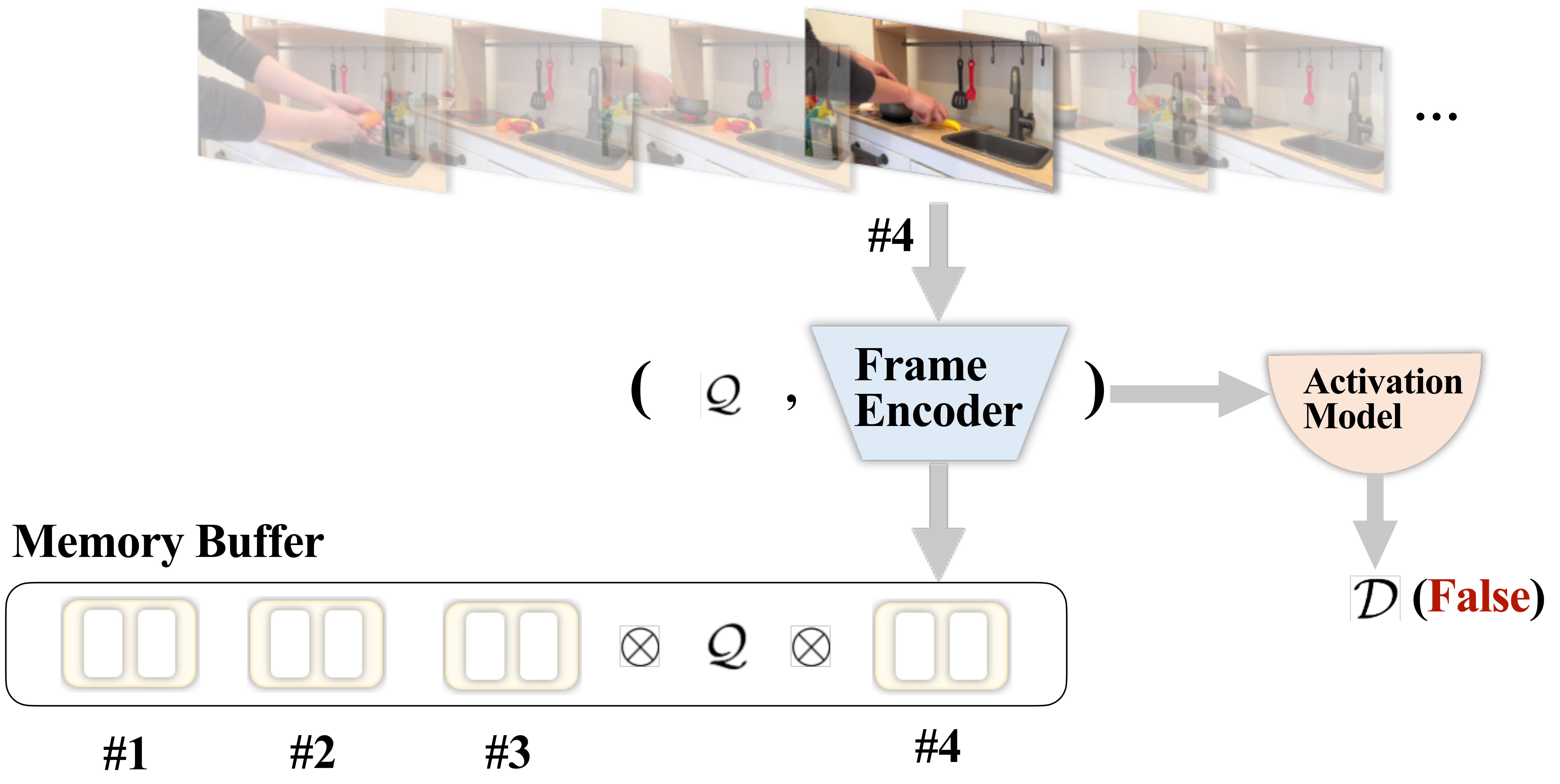
# StreamBridge-Framework

- In streaming scenarios, incoming frames are encoded one by one, and stored into the memory buffer



# StreamBridge-Framework

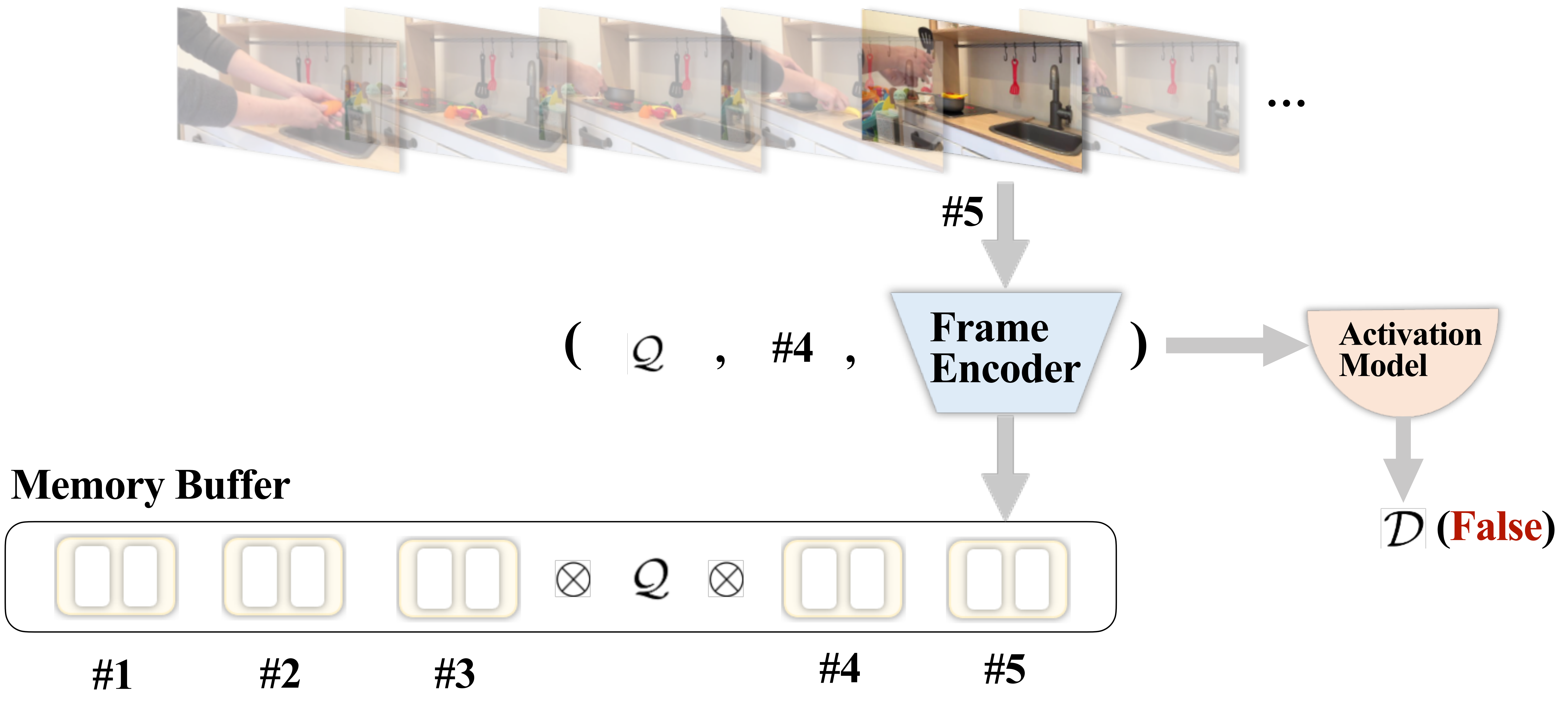
- When the user query is posed, an activation model makes decision on whether response or not




$\mathcal{D}$ is False at #4 and keeps buffering

# StreamBridge-Framework

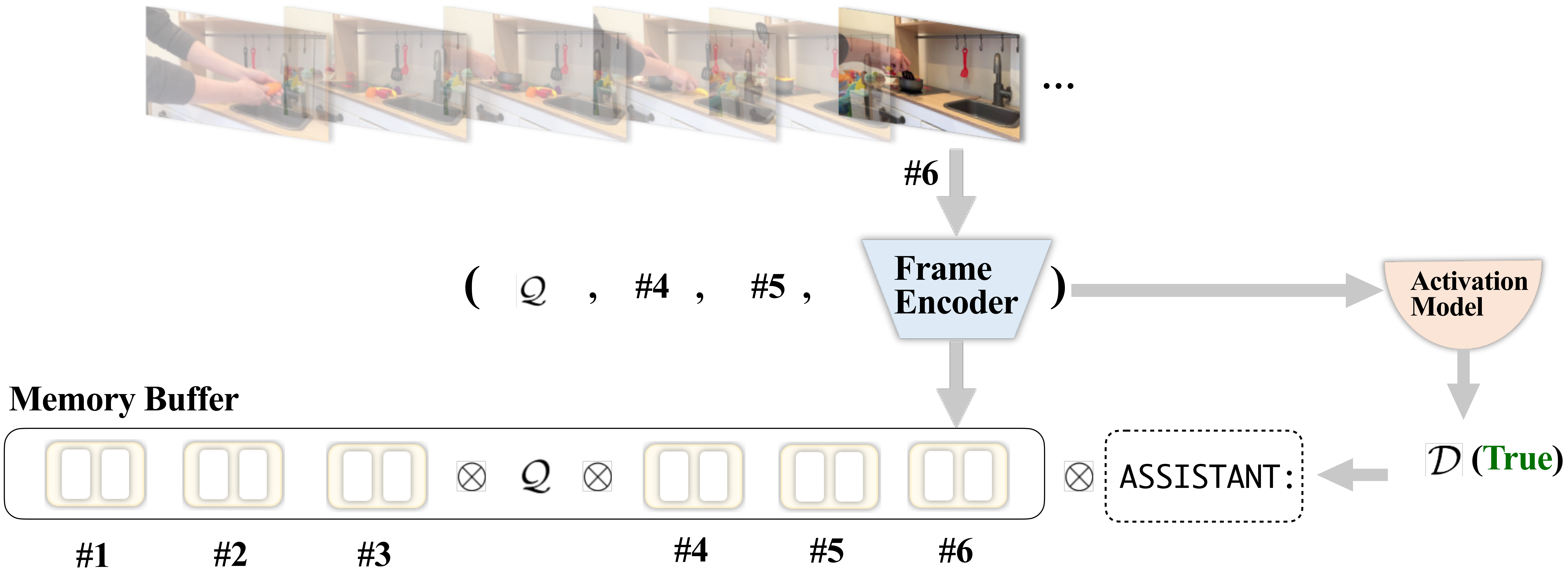
- When the user query is posed, an activation model makes decision on whether response or not




$\mathcal{D}$ is False at #5 and keeps buffering

# StreamBridge-Framework

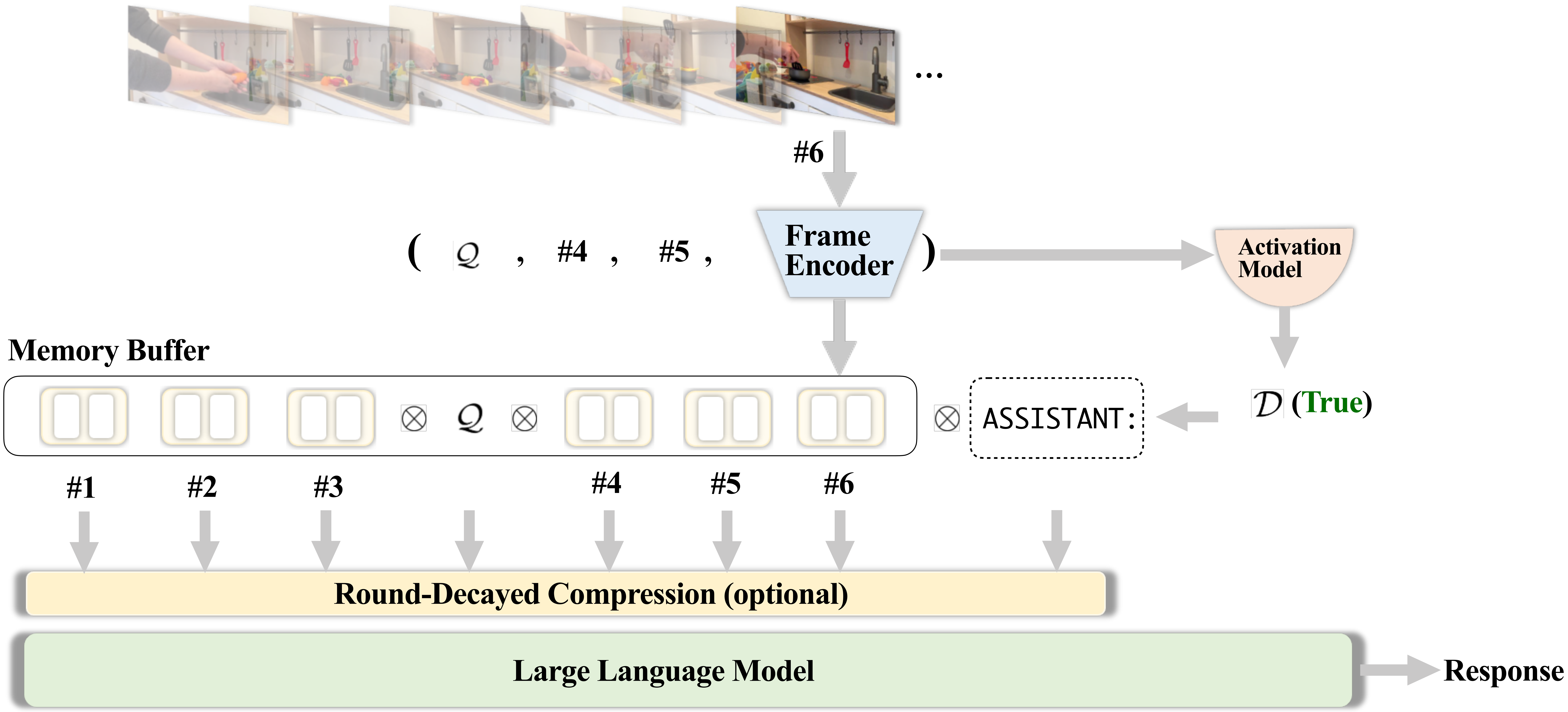
- When the user query is posed, an activation model makes decision on whether response or not




D is True at #6 and requires a response

# StreamBridge-Framework

- When the user query is posed, an activation model makes decision on whether response or not



# StreamBridge-Framework

- The memory buffer is also updated when generating new responses to preserve complete historical dialogues (line 17).

**Algorithm 1:** StreamBridge Framework

1 **Inputs:** incoming frames $[F_1, F_2, \ldots, F_t, \ldots]$;
2 **Initializations:** $\mathcal{I}(\cdot)$, $\mathcal{LLM}(\cdot)$, $\mathcal{ACT}(\cdot)$, $\mathcal{COM}(\cdot)$, $\mathcal{MB} = [\cdot]$, MaxLen, $t_{\mathcal{Q}}$=None;
3 **while** $F_t$ **do**
4 $\quad \mathcal{MB} \leftarrow \mathcal{I}(F_t)$ ; // store the frame feature $\mathcal{I}(F_t)$ into the Memory Buffer
5 $\quad$ **if** $\mathcal{Q}$ *at* timestamp $t$ **then**
6 $\quad\quad \mathcal{MB} \leftarrow \mathcal{Q}$
7 $\quad\quad t_{\mathcal{Q}} \leftarrow t$ ; // $t_{\mathcal{Q}}$ is the timestamp when $\mathcal{Q}$ is posed
8 $\quad$ **if** $t_{\mathcal{Q}}$ is not None **then**
9 $\quad\quad \mathcal{D} \leftarrow \mathcal{ACT}(\mathcal{Q}, F_{t_{\mathcal{Q}}:t-1}, F_t)$ ; // $\mathcal{D}$ denotes whether response or not at timestamp $t$
10 $\quad$ **else**
11 $\quad\quad \mathcal{D} \leftarrow$ False ; // not response if there is no $\mathcal{Q}$
12 $\quad$ **if** $\mathcal{D}$ **then**
$\quad\quad$ // $\mathcal{D}$ is true at timestamp $t$, and should return a response $\mathcal{R}$
13 $\quad\quad$ InputEmbeds $\leftarrow$ Flatten($\mathcal{MB}$)
14 $\quad\quad$ **if** Len(InputEmbeds) $>$ MaxLen **then**
15 $\quad\quad\quad$ InputEmbeds $\leftarrow \mathcal{COM}$(InputEmbeds) ; // compress redundant visual tokens
16 $\quad\quad \mathcal{R} \leftarrow \mathcal{LLM}$(InputEmbeds) ; // return a response $\mathcal{R}$
17 $\quad\quad \mathcal{MB} \leftarrow \mathcal{R}$ ; // update $\mathcal{MB}$
18 $\quad t$ += 1 ; // receive subsequent frames

# StreamBridge-Round Decayed Compression

- Streaming scenarios usually feature real-time understanding on the most recent video segment

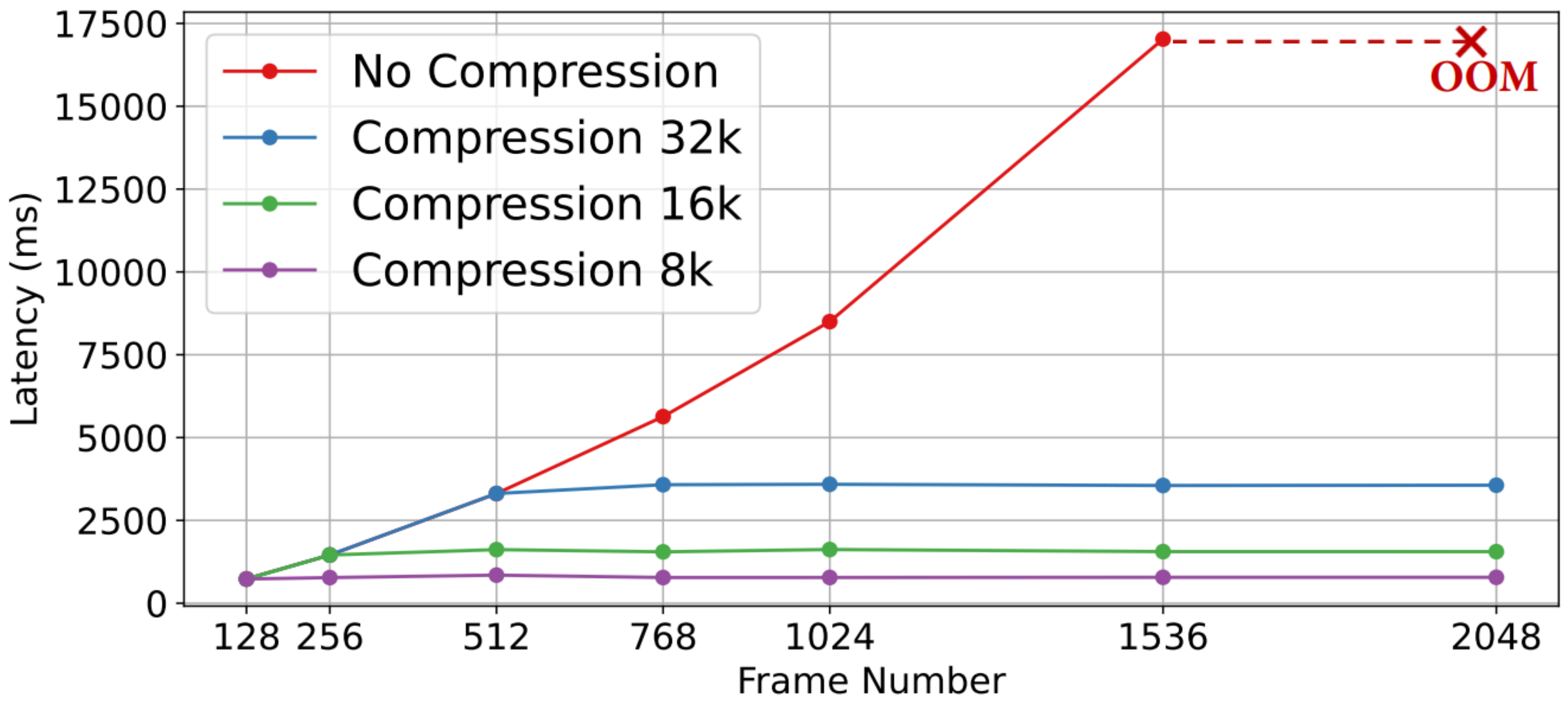
**USER:** What is written on the sign?

**ASSISTANT:** “Change the direction of all escalators and travelators”.

Compress frame tokens in previous rounds

Preserve most recent frame tokens

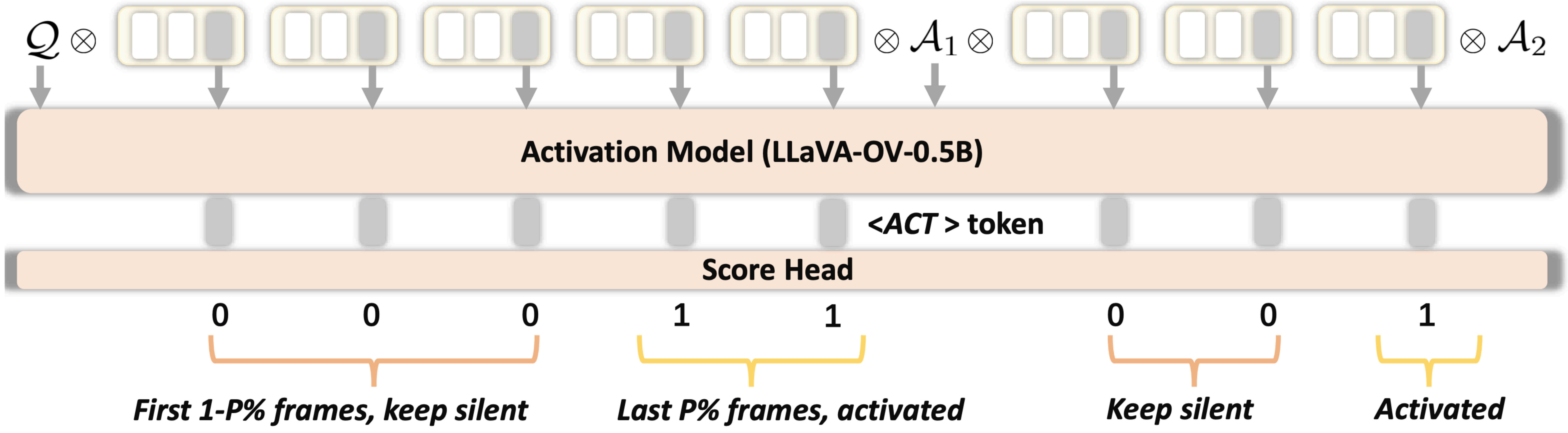
- Rounded Decayed Compression
  - Apply token merging (average pooling over adjacent frame tokens) in a round-by-round manner, starting from the earliest rounds, until the total embedding length below the predefined ***MaxLen***
  - Maintain near-constant latency and memory usage when the number of tokens exceeds ***MaxLen***




Inference Latency (y-axis) vs. Frame Number (x-axis) on a single A100-80G with different ***MaxLen***.

# StreamBridge-Activation Model

- A parallel pipeline by decoupling the activation mechanism into a compact activation model, to prevent the potential interference with the language modeling capacity of the main Video-LLM




- A Plug-and-play Activation Model
  - Replace the LM head with a score head for binary classification.
  - A learnable activation token <ACT>.
  - Collect a diverse set of temporally annotated video datasets to train the model.
  - Only the last P% of frames of each video segment <Vi> are labeled as positive (i.e., response-worthy), while earlier frames are treated as negatives.

# StreamIT Dataset

- Datasets for proactive understanding:
  - Format: <Q>, <V1> <A1>, <V2> <A2>, …
  - Tasks: dense video caption, sequential step recognition, grounded VideoQA
- StreamingQA-120K: Multi-turn, long-form QA construction:
  - Format: <V1> <Q1> <A1>, <V2> <Q2> <A2>, …
  - Clip concatenation: iteratively compute pairwise similarity between videos and concatenate highly similar clips from 1.28m video pools.
  - QA construction: Prompt GPT-4o to generate QA pairs spanning 8 task types
  - Data augmentation:
    - Random QA drop: randomly drops some QA pairs in a concatenated video.
    - QA interval shit: randomly transforms sequences from <Vi> <Qi> <Ai> to <Qi> <Vi> <Ai>

| **Task** | **# of Samples** | **Datasets** | **Average duration** |
|---|---|---|---|
| Dense Video Captioning | ∼54k | ActivityNet [57] (∼10k)<br>Shot2Story [58] (∼36k)<br>ViTT [69] (∼8k) | ∼180s<br>∼16s<br>∼210s |
| Sequential Step Recognition | ∼22k | YouCook2 [60] (∼1.3k)<br>COIN [59] (∼11k)<br>HowToStep [70] (∼10k) | ∼317s<br>∼145s<br>∼190s |
| Grounded Video Question Answering | ∼69k | MovieChat [72] (∼0.8k)<br>EgoTimeQA [71] (∼10k)<br>QAEgo4D [61] (∼15k)<br>FineVideo [73] (∼43k) | ∼10k frames<br>∼150s<br>∼495s<br>∼280s |
| Multi-turn Real-time Question Answering | ∼120k | StreamingQA-120K (∼120k)<br>(Sourced from Webvid-10M[19], Panda-70M[18], InternVid-10M[17]) | ∼150s |

# Multi-turn Real-time Understanding Results

| Method | # of Frames | OVO-Bench Real-Time | | | | | | | Streaming-Bench Real-Time | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | OCR | ACR | ATR | STU | FPD | OJR | AVG. | OP | CR | CS | ATP | EU | TR | PR | SU | ACP | CT | AVG. |
| **Human** | | | | | | | | | | | | | | | | | | | |
| Human | - | 93.96 | 92.57 | 94.83 | 92.70 | 91.09 | 94.02 | 93.20 | 89.47 | 92.00 | 93.60 | 91.47 | 95.65 | 92.52 | 88.00 | 88.75 | 89.74 | 91.30 | 91.46 |
| **Proprietary Models (Offline), Single-Turn Evaluation** | | | | | | | | | | | | | | | | | | | |
| Gemini 1.5 pro [23] | 1 FPS | 85.91 | 66.97 | 79.31 | 58.43 | 63.37 | 61.96 | 69.32 | 79.02 | 80.47 | 83.54 | 79.67 | 80.00 | 84.74 | 77.78 | 64.23 | 71.95 | 48.70 | 75.69 |
| GPT-4o [22] | 64 | 69.80 | 64.22 | 71.55 | 51.12 | 70.3 | 59.78 | 64.46 | 77.11 | 80.47 | 83.91 | 76.47 | 70.19 | 83.80 | 66.67 | 62.19 | 69.12 | 49.22 | 73.28 |
| **Open-Source Models (Offline), Single-Turn Evaluation** | | | | | | | | | | | | | | | | | | | |
| Qwen2-VL-72B [2] | 64 | 65.77 | 60.55 | 69.83 | 51.69 | 69.31 | 54.35 | 61.92 | - | - | - | - | - | - | - | - | - | - | - |
| LLaVA-Video-7B [15] | 64 | 69.13 | 58.72 | 68.83 | 49.44 | 74.26 | 59.78 | 63.52 | - | - | - | - | - | - | - | - | - | - | - |
| LLaVA-OV-7B [3] | 64/32 | 66.44 | 57.80 | 73.28 | 53.37 | 71.29 | 61.96 | 64.02 | 80.38 | 74.22 | 76.03 | 80.72 | 72.67 | 71.65 | 67.59 | 65.45 | 65.72 | 45.08 | 71.12 |
| Qwen2-VL-7B [2] | 64/1 FPS | 60.40 | 50.46 | 56.03 | 47.19 | 66.34 | 55.43 | 55.98 | 75.20 | 82.81 | 73.19 | 77.45 | 68.32 | 71.03 | 72.22 | 61.19 | 61.47 | 46.11 | 69.04 |
| InternVL-V2-8B [74] | 64/16 | 67.11 | 60.55 | 63.79 | 46.07 | 68.32 | 56.52 | 60.39 | 68.12 | 60.94 | 69.40 | 77.12 | 67.70 | 62.93 | 59.26 | 53.25 | 54.96 | 56.48 | 63.72 |
| **Open-Source Models (Streaming), Single-Turn Evaluation** | | | | | | | | | | | | | | | | | | | |
| Flash-VStream-7B [11] | 1 FPS | 24.16 | 29.36 | 28.45 | 33.71 | 25.74 | 28.80 | 28.37 | 25.89 | 43.57 | 24.91 | 23.87 | 27.33 | 13.08 | 18.52 | 25.20 | 23.87 | 48.70 | 23.23 |
| VideoLLM-Online-8B [10] | 2 FPS | 8.05 | 23.85 | 12.07 | 14.04 | 45.54 | 21.20 | 20.79 | 39.07 | 40.06 | 34.49 | 31.05 | 45.96 | 32.40 | 31.48 | 34.16 | 42.49 | 27.89 | 35.99 |
| Dispider [13] | 1 FPS | 57.72 | 49.54 | 62.07 | 44.94 | 61.39 | 51.63 | 54.55 | 74.92 | 75.53 | 74.10 | 73.08 | 74.44 | 59.92 | 76.14 | 62.91 | 62.16 | 45.80 | 67.63 |
| **Models under *StreamBridge* (Offline $\rightarrow$ Streaming), Multi-Turn Evaluation** | | | | | | | | | | | | | | | | | | | |
| Oryx-1.5-7B† [1] | 1 FPS | 60.40 | 52.29 | 69.83 | 50.00 | 65.35 | 57.61 | 59.25 | 78.47 | 77.17 | 83.86 | 80.20 | 71.07 | 66.98 | 79.63 | 61.38 | 66.29 | 40.93 | 70.59 |
| + *Stream-IT* | 1 FPS | 84.56 | 75.23 | 70.69 | 50.56 | 74.26 | 71.74 | 71.17 | 82.29 | 77.95 | 87.98 | 86.47 | 77.99 | 81.31 | 76.85 | 69.92 | 71.96 | 35.23 | 74.79 |
| LLaVA-OV-7B† [3] | 1 FPS | 58.39 | 59.63 | 69.82 | 44.38 | 76.23 | 61.41 | 61.64 | 76.84 | 77.17 | 82.60 | 75.25 | 64.15 | 64.17 | 75.00 | 61.38 | 61.19 | 46.11 | 68.39 |
| + *Stream-IT* | 1 FPS | 74.50 | 77.06 | 70.69 | 54.49 | 73.27 | 69.57 | 69.93 | 82.29 | 72.44 | 92.09 | 80.86 | 71.07 | 74.46 | 75.00 | 62.20 | 70.26 | 28.50 | 70.92 |
| Qwen2-VL-7B† [2] | 1 FPS | 65.10 | 64.22 | 64.66 | 46.63 | 74.26 | 65.22 | 63.35 | 80.38 | 78.74 | 83.22 | 79.86 | 74.21 | 69.47 | 77.78 | 63.41 | 69.97 | 43.01 | 72.01 |
| + *Stream-IT* | 1 FPS | 84.56 | 71.56 | 74.14 | 49.44 | 75.25 | 72.83 | **71.30** | 84.74 | 82.68 | 88.92 | 89.77 | 77.36 | 85.36 | 84.26 | 69.92 | 71.67 | 35.75 | **77.04** |

Table 1: Results on real-time understanding tasks on OVO-Bench and Streaming-Bench. † means models under *StreamBridge* framework, and + *Stream-IT* means finetuned on *Stream-IT*.

- *Models pre-trained on rich multi-modal interleaved data benefit from our framework w/o training. For example, Qwen2-VL (55.98 -> 61.64, 69.04 -> 72.01 ).*
- *Models fine-tuned on StreamIT demonstrates notable improvements, even surpassing Gemini-1.5 Pro and GPT-4o.*

# General Video Understanding Results

| Model | MVBench | PerceptionTest | TempCompass | EgoSchema | LongVideoBench | MLVU | VideoMME (w/o subs) |
|---|---|---|---|---|---|---|---|
| | Avg | Val | MC | Test | Val | M-Avg | Avg |
| Avg. Duration | 16s | 23s | 12s | 180s | 473s | 651s | 1010s |
| **Proprietary Models** | | | | | | | |
| Gemini 1.5 pro [23] | 60.5 | - | 67.1 | 71.2 | 64.0 | - | 75.0 |
| GPT-4o [22] | 64.6 | - | 70.9 | 72.2 | 66.7 | 64.6 | 71.9 |
| **Open-Source Models** | | | | | | | |
| Kangaroo-8B [76] | 61.0 | - | 62.5 | - | 54.8 | 61.0 | 56.0 |
| LongVILA-7B [77] | - | - | - | 67.7 | - | - | 57.5 |
| LongVU-7B [78] | 66.9 | - | - | 67.6 | - | 65.4 | 60.6 |
| Apollo-7B [4] | - | 67.3 | 64.9 | - | 58.5 | 70.9 | 61.3 |
| NVILA-8B [79] | 68.1 | 65.4 | 69.7 | - | 57.7 | 70.1 | 64.2 |
| SF-LLaVA-1.5-7B [5] | - | 69.6 | 68.8 | - | 62.5 | 71.5 | 63.9 |
| InternVL2.5-8B [80] | 72.0 | 68.2 | 68.3 | 51.5 | 60.0 | 68.9 | 64.2 |
| VideoChat-Flash-7B [81] | 74.0 | 76.2 | - | - | 64.7 | 74.7 | 65.3 |
| VideoLLaMA3-7B [37] | 69.7 | 72.8 | 68.1 | 63.3 | 59.8 | 73.0 | 66.2 |
| Oryx-1.5-7B [1] | 67.6 | 70.0 | 58.8 | - | 56.3 | 67.5 | 58.8 |
| Oryx-1.5-7B (ours) ‡ | 68.0 (↑0.4) | 71.0 (↑1.0) | 69.0 (↑10.2) | 61.2 | 58.9 (↑2.6) | 71.4 (↑4.0) | 65.5 (↑6.7) |
| LLaVA-OV-7B [3] | 56.7 | 57.1 | 64.8 | 60.1 | 56.3 | 64.7 | 58.2 |
| LLaVA-OV-7B (ours) ‡ | 59.4 (↑2.7) | 63.9 (↑6.8) | 67.7 (↑2.9) | 67.0 (↑6.9) | 54.3 (↓2.0) | 68.2 (↑3.5) | 61.2 (↑3.0) |
| Qwen2-VL-7B [2] | 67.0 | 62.3 | 67.9 | 66.7 | - | - | 63.3 |
| Qwen2-VL-7B (ours) ‡ | 64.4 (↓2.6) | 69.9 (↑7.6) | 71.1 (↑3.2) | 66.9 (↑0.2) | 59.1 | 69.6 | 64.4 (↑1.1) |

Table 2: Results on general video understanding benchmarks. ‡ means models under *Stream-Bridge* framework and fine-tuned on *Stream-IT*.

- *Models under our framework and fine-tuned on StreamIT maintain competitive or achieve better performance on standard general video understanding tasks. For example, Oryx-1.5-7B achieves 65.5 (+6.7) on VideoMME.*

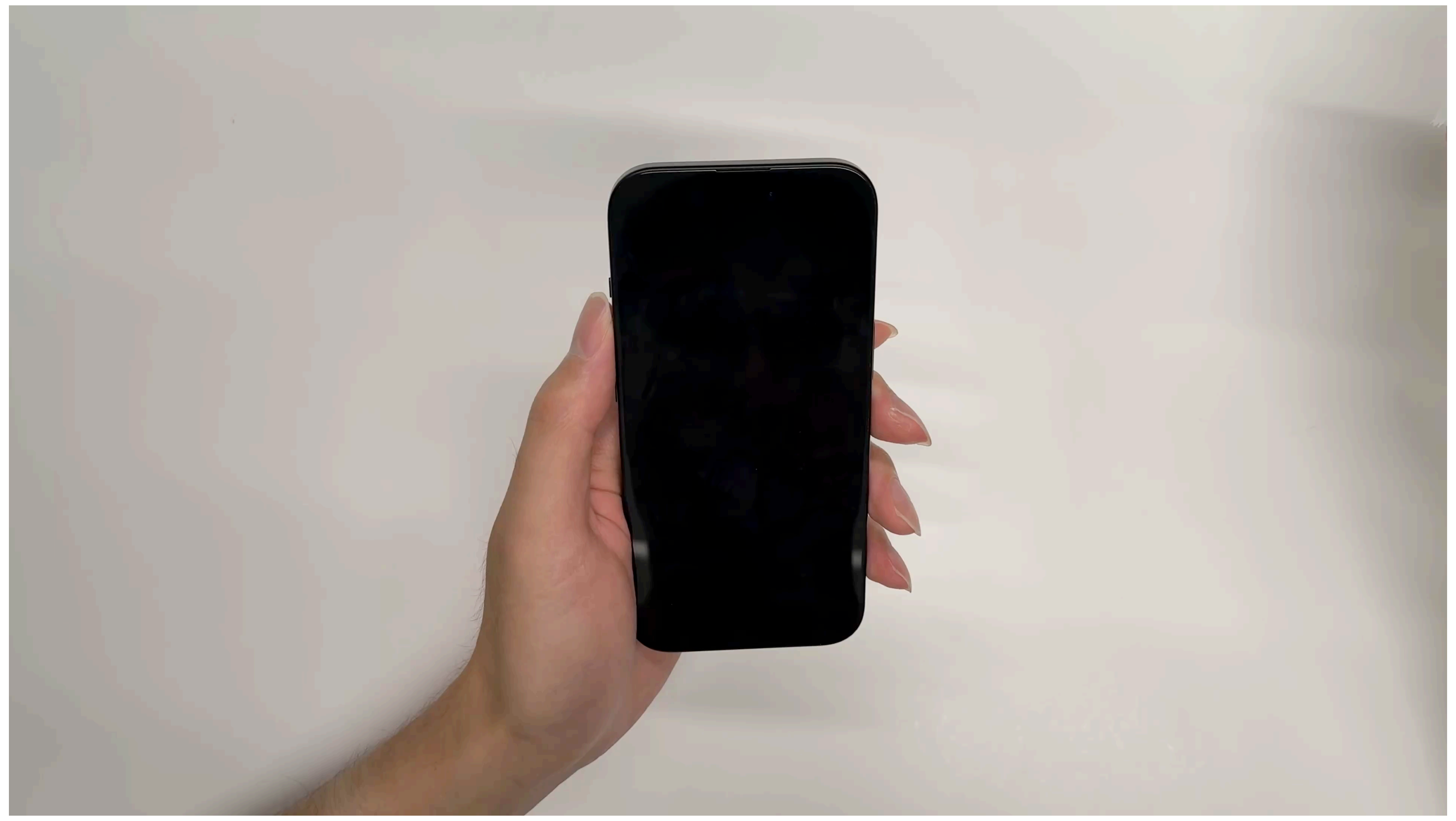
# Proactive Results

| Method | # of Frames | ET-Bench | | | | | |
|---|---|---|---|---|---|---|---|
| | | $TVG_{F1}$ | $TAL_{F1}$ | $DVC_{F1}$ | $DVC_{Sim}$ | $SLC_{F1}$ | $SLC_{Sim}$ |
| VideoLLM-Online [10] | 2 FPS | 13.2 | 9.1 | 24.0 | 13.4 | 9.9 | 10.1 |
| Dispider [13] | 1 FPS | **36.1** | **27.3** | 33.8 | 18.9 | 18.8 | 12.4 |
| **Models under *StreamBridge* Framework** | | | | | | | |
| Oryx-1.5 (ours)‡ | 1 FPS | 34.3 | 24.3 | 37.8 | 24.0 | 22.5 | **17.3** |
| LLaVA-OV (ours)‡ | 1 FPS | 34.3 | 24.3 | 37.9 | 24.2 | **22.8** | 16.2 |
| Qwen2-VL (ours)‡ | 1 FPS | 34.3 | 24.3 | **38.3** | **25.1** | 22.6 | 17.1 |

Table 3: Results on ET-Bench. ‡ denotes models under *StreamBridge* framework and fine-tuned on *Stream-IT*. $TVG_{F1}$ and $TAL_{F1}$ scores are identical across *StreamBridge* models due to sharing the same activation model.

- *In all tasks, the question is presented at the beginning of the video, and the model must autonomously decide when to respond.*
- *StreamBridge outperforms across generation-based tasks such as DVC (Dense Video Captioning) and SLC (Step Localization and Captioning), achieving higher similarity scores by producing more accurate and context aware descriptions in streaming scenarios.*
- *We attribute this to the decoupled nature of the activation model, which enables the main Video-LLM to focus solely on video understanding and language generation, free from the burden of proactive decision-making.*

# Qualitative Results




**Prompt:** What am I doing with the phone?

**StreamBridge:**

00:06: Wipe the screen.

00:13: Wipe the back of the phone.

00:19: Wipe the camera lens.

**Prompt:** Guide me through the cooking steps.

**StreamBridge:**

00:20: Drain the pasta.

00:34: Cut the garlic and lemon.

00:58: Cut the parsley.

01:18: Squeeze the lemon juice into a measuring cup.

01:42: Season the chicken with salt and pepper.

02:07: Add oil to the pan.

02:18: Add the chicken to the pan.

02:32: Flip the chicken over.

02:46: Add oil to the pan.

03:03: Add the lemon slices to the pan.

03:26: Remove the lemon slices from the pan.

03:49: Add the olives to the pan.

04:10: Add the chicken broth to the pan.

04:20: Cook the meat.

04:32: Fry the meat.

04:43: Add seasoning.

04:54: Cut the meat.

05:05: Add decoration.

05:16: Add seasoning.

05:26: Taste and enjoy.

05:37: Cut the olives.