IEEE INFOCOM

IEEE ComSoc IEEE Communications Society

IEEE

# Venus: An Efficient Edge Memory-and-Retrieval System for VLM-based Online Video Understanding


Shengyuan Ye[1], Bei Ouyang[1], Tianyi Qian[1], Liekang Zeng[2], Mu Yuan[2]
Xiaowen Chu[3], Weijie Hong[4], Xu Chen[1]

[1] Sun Yat-sen University
[2] The Chinese University of Hong Kong
[3] The Hong Kong University of Science and Technology (GZ)
[4] Shenzhen Smart City Communications Co., Ltd.


中山大學 SUN YAT-SEN UNIVERSITY

香港中文大學 The Chinese University of Hong Kong

香港科技大学(广州) THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

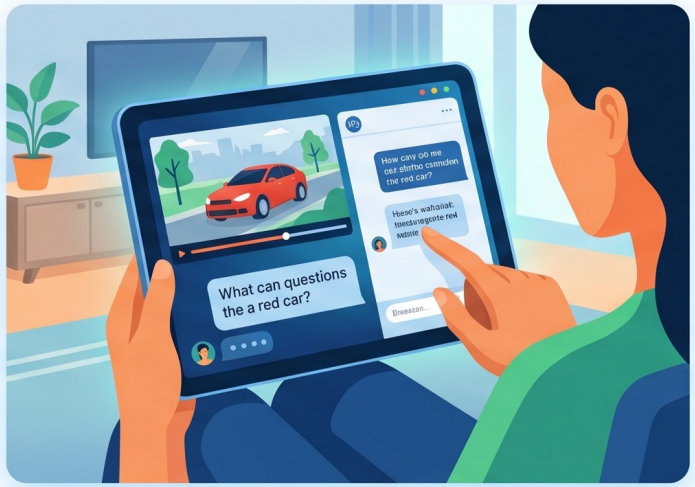
# Online Video Understanding Applications

- Online video understanding is an essential edge intelligence application.

## Intelligent Surveillance

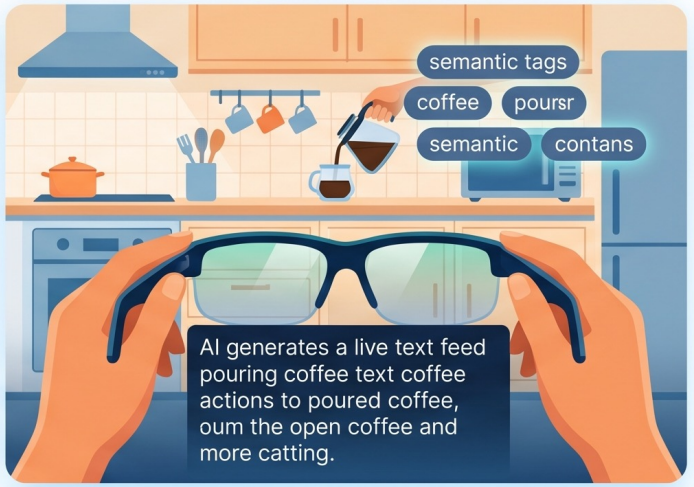

Real-time video analytics
& Event alerting

## Interactive Video QA




Semantic search &
Time-segment analysis

## Live Scene Description



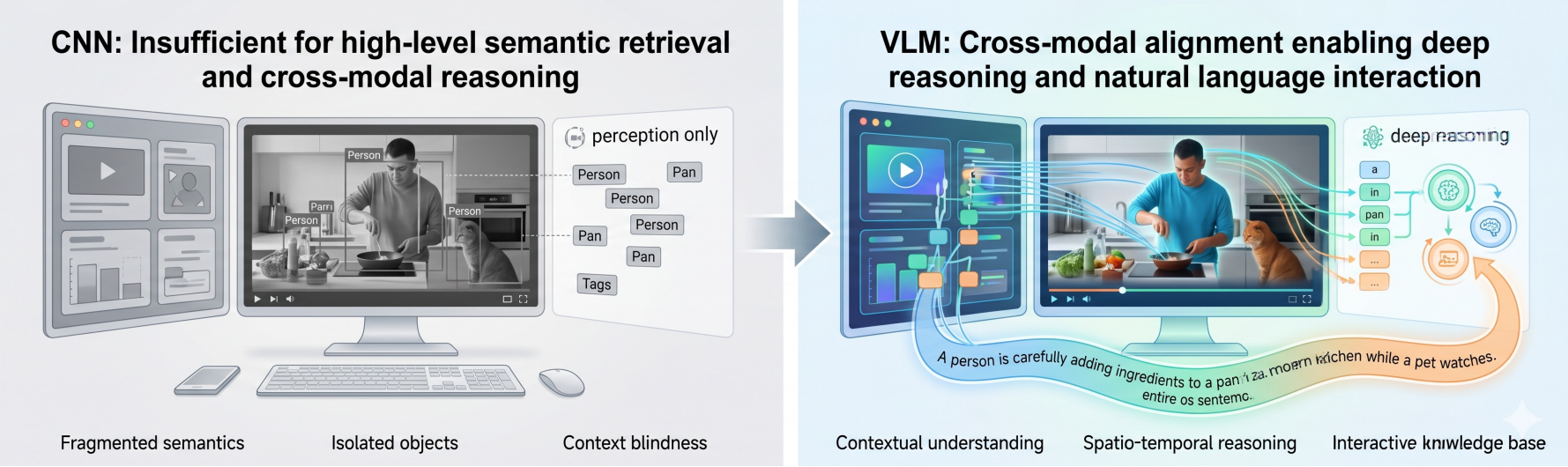

Continuous event streaming
& Vision-to-text

# Online Video Understanding Applications

- CNNs had long served as the backbone of online video understanding.
  - **Strengths:** Lightweight, low latency, edge-friendly.
  - **Limitations:** Insufficient for high-level semantic retrieval and cross-modal reasoning; struggles with natural language interaction and control.

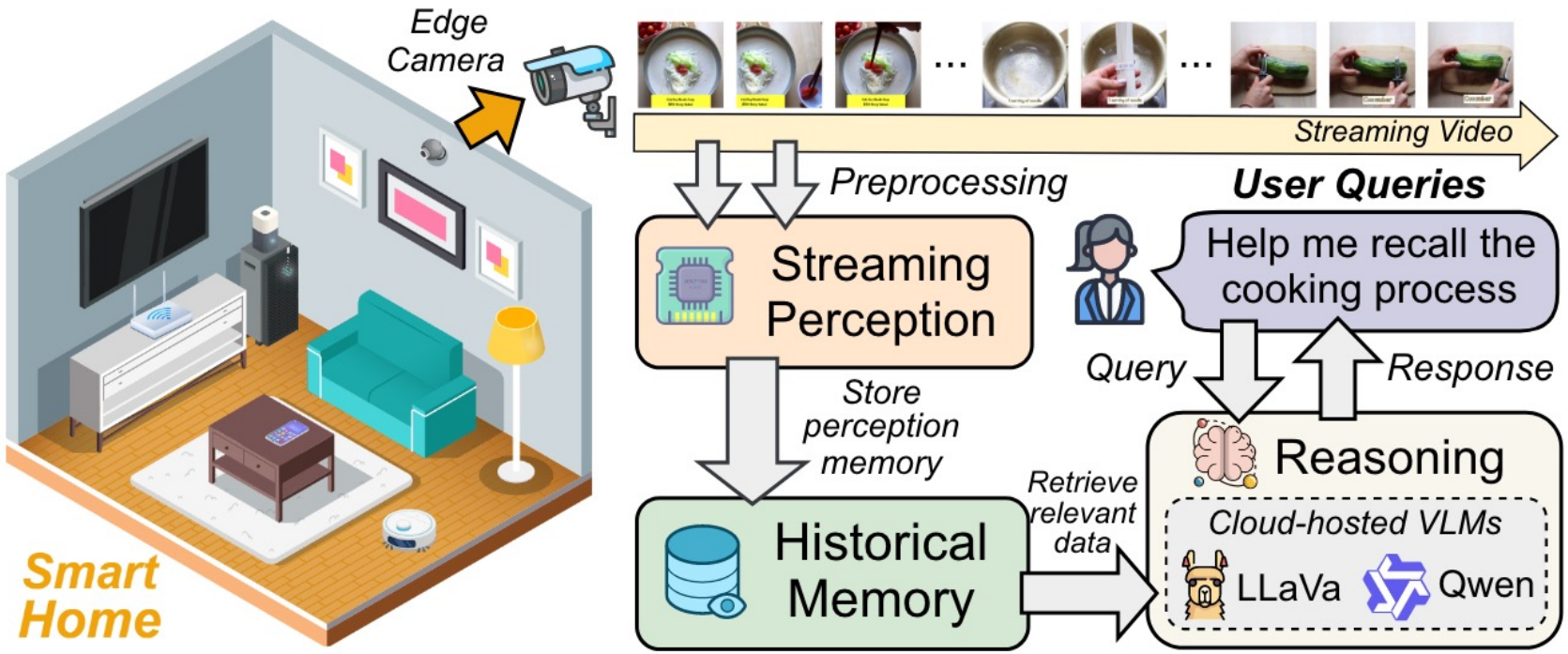
# Online Video Understanding Applications

- CNNs had long served as the backbone of online video understanding.
  - **Strengths:** Lightweight, low latency, edge-friendly.
  - **Limitations:** Insufficient for high-level semantic retrieval and cross-modal reasoning; struggles with natural language interaction and control.
- Modern Paradigm: Transformer-based vision-language models (VLMs).
  - Cross-modal alignment enabling deep reasoning and natural language interaction.



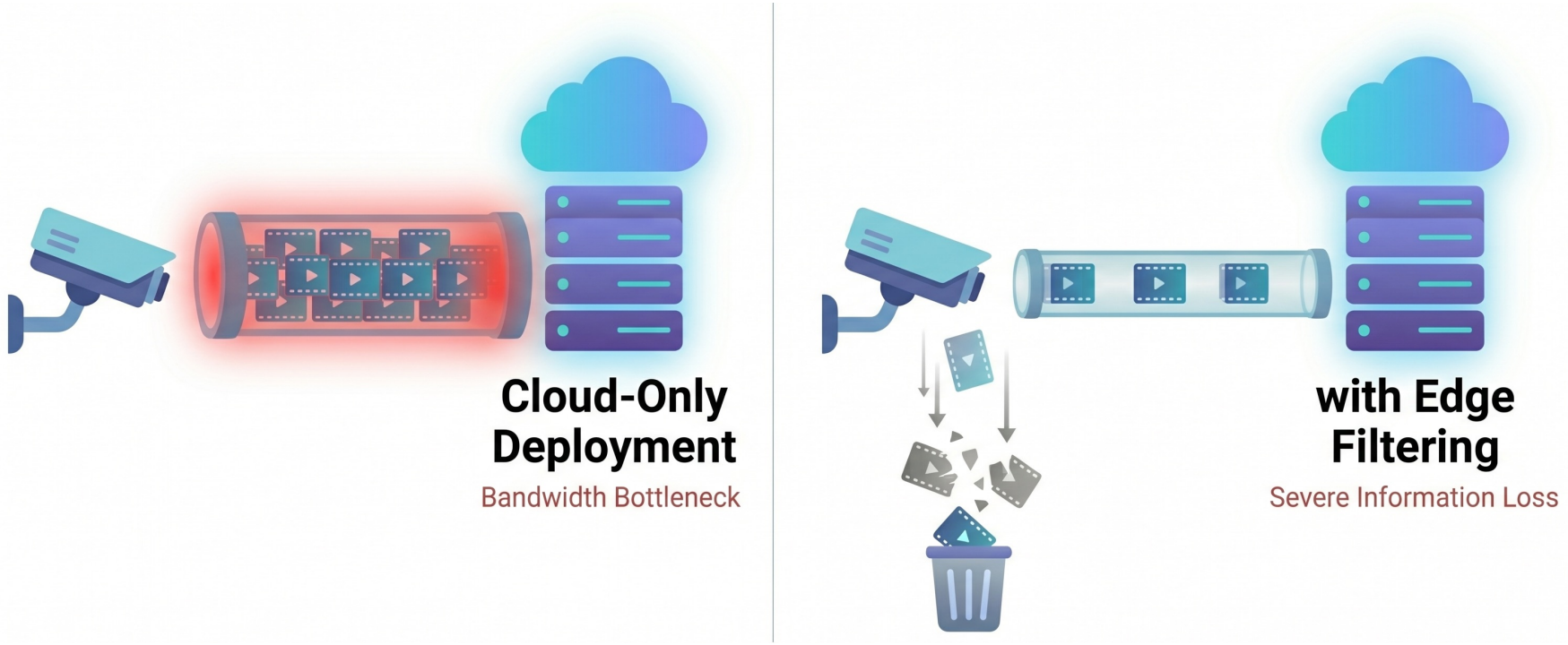
# VLMs-based Online Video Understanding

- An online video understanding application empowered by VLMs.
  - **Three core modules**: Streaming Perception Module, Historical Memory Module, and Reasoning Module.



# The Deployment Dilemma

- **Cloud-Only Deployment:** Uploading raw video streams directly to the cloud incurs prohibitive communication latency. *(**Bandwidth Bottleneck**)*
- **With Edge Filtering:** Simple on-device frame dropping reduces data transmission but discards critical visual evidence. *(**Severe Information Loss**)*

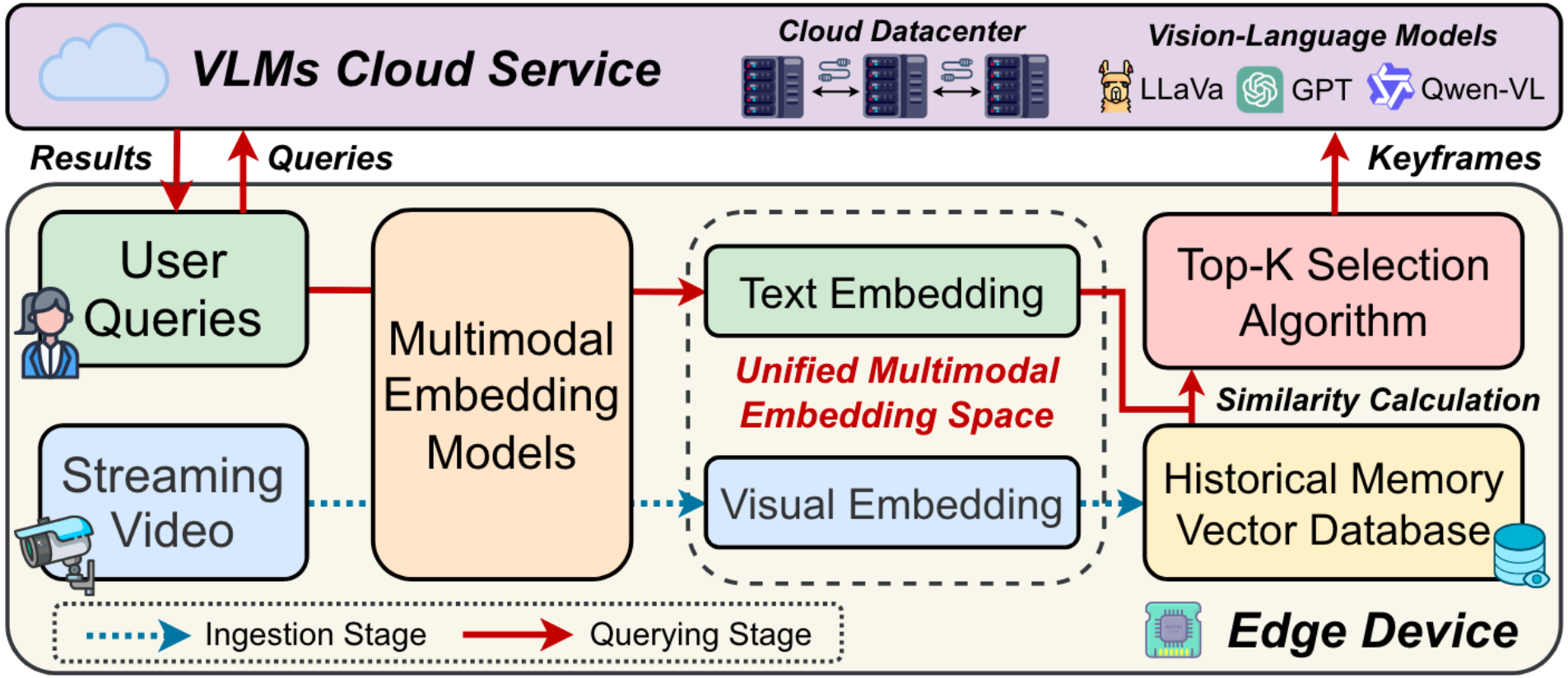


# Our System Design Goal

- A system supporting efficient online video understanding applications.
  1. **Communication Efficiency:** Alleviate edge-cloud bandwidth bottlenecks by eliminating redundant full-frame video uploads.
  2. **Real-time Real-time Online Querying:** Support real-time video processing and enable low-latency reasoning over both live streams and historical data.
  3. **Natural Language Interaction:** Support intuitive, user-friendly semantic queries to ensure accessible and accurate video understanding.

# Edge-cloud Disaggregated Architecture

- The edge for building video memory and keyframe retrieval, while the cloud side is dedicated to the heavy-duty VLM reasoning.

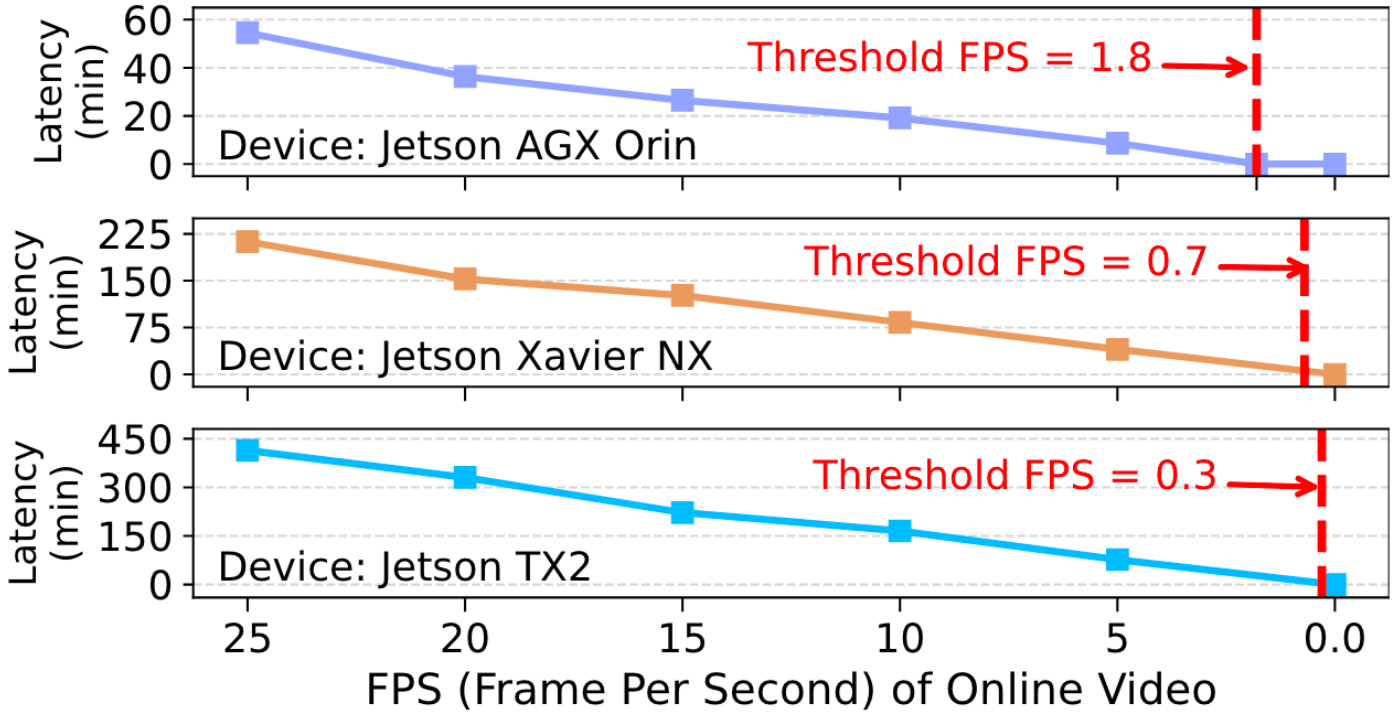


# Technical Challenges in Practical Deployment

- Despite its simplicity, this foundational architecture faces critical technical challenges for real-world deployment.

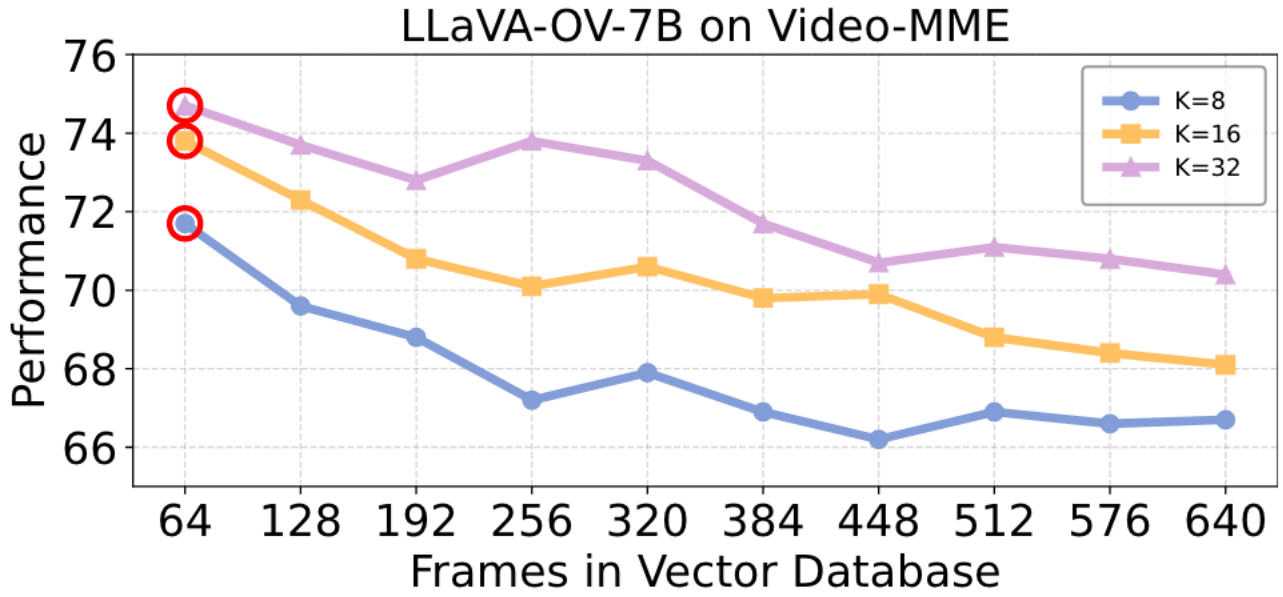
# Technical Challenges in Practical Deployment

- Despite its simplicity, this foundational architecture faces critical technical challenges for real-world deployment.
  - **Challenge 1**: High latency hinders real-time embedding and ingestion.



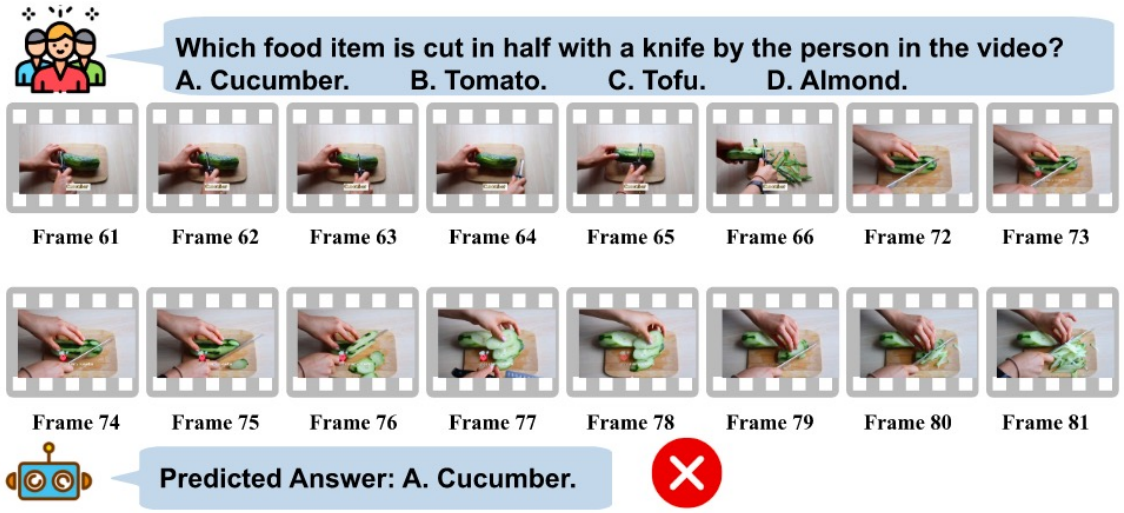
# Technical Challenges in Practical Deployment

- Despite its simplicity, this foundational architecture faces critical technical challenges for real-world deployment.
  - **Challenge 2:** Excessively redundant frames overwhelm the memory database and degrade retrieval performance.



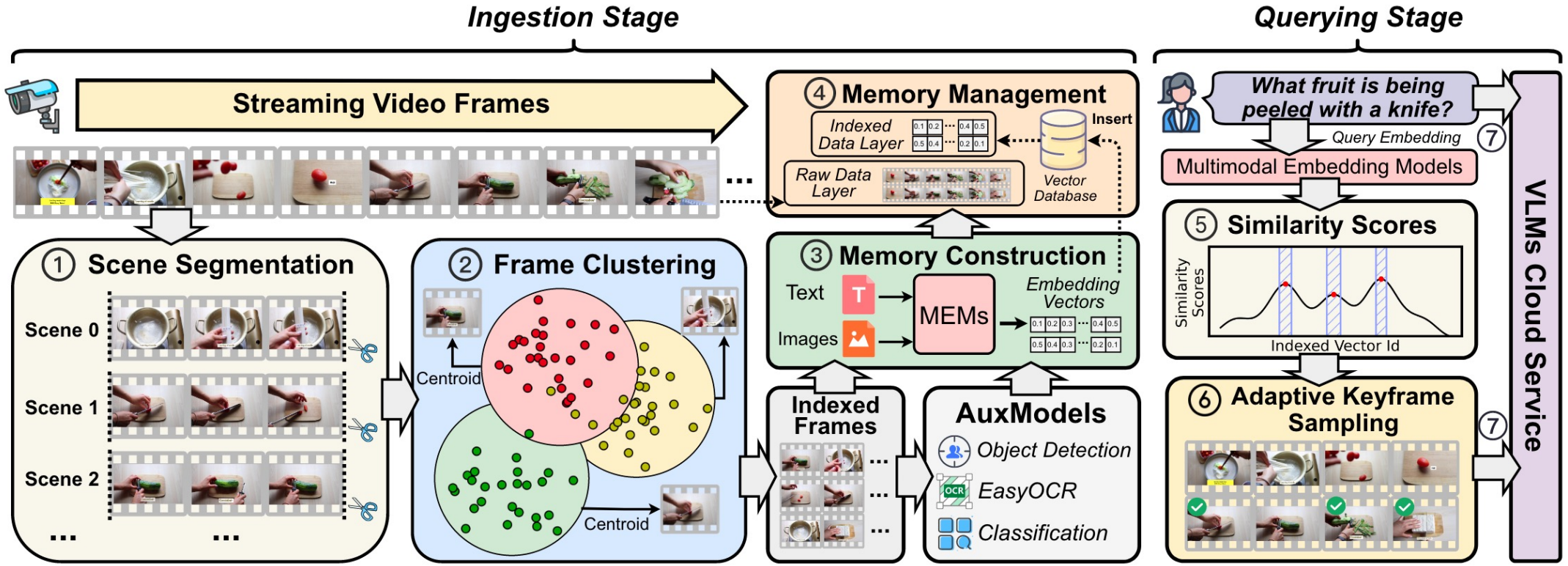
# Technical Challenges in Practical Deployment

- Despite its simplicity, this foundational architecture faces critical technical challenges for real-world deployment.
  - **Challenge 3**: Lack of diversity and adaptivity in Top-K selection algorithm.

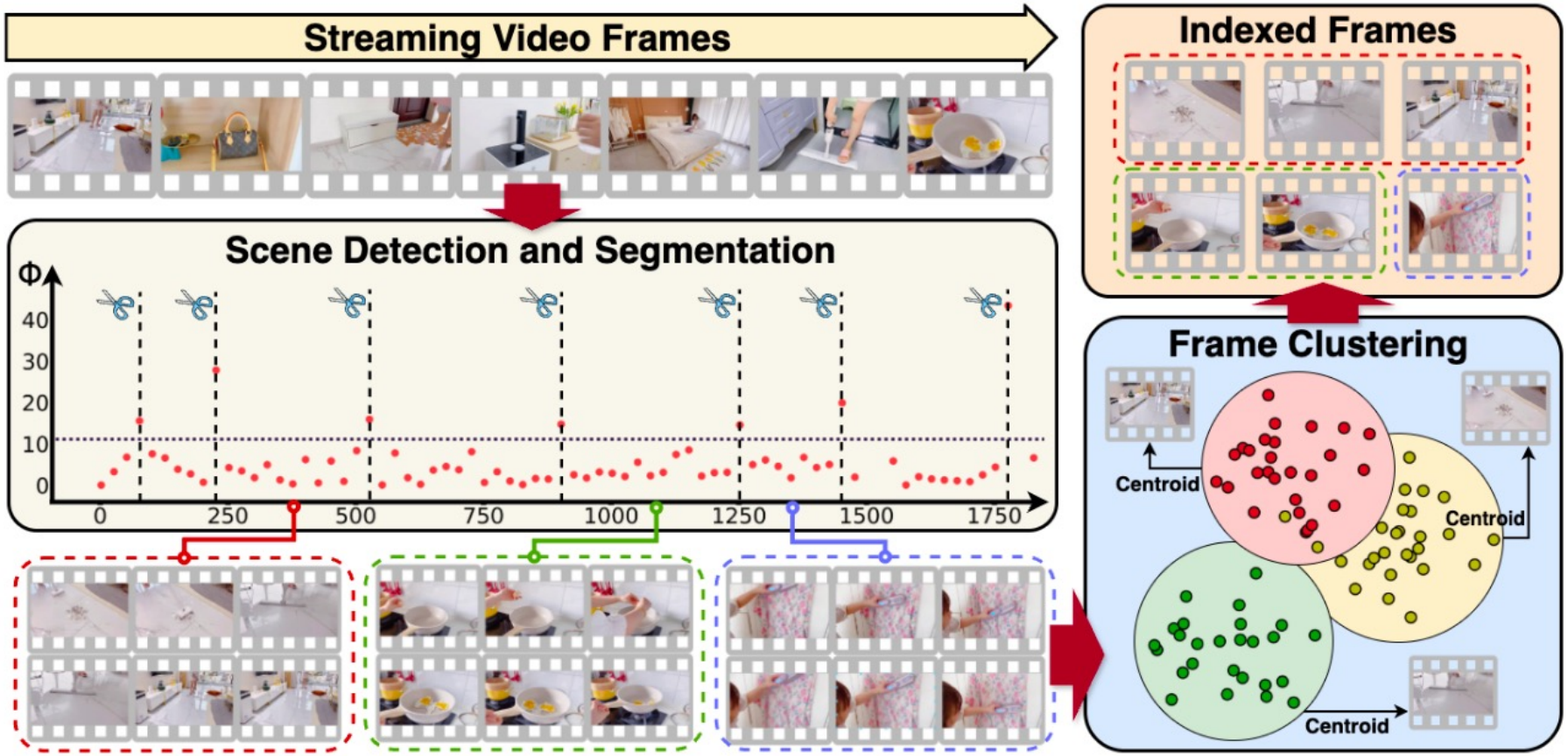
# Venus System Overview

- Venus builds upon previously proposed edge–cloud disaggregated architecture to further address the technical challenges.



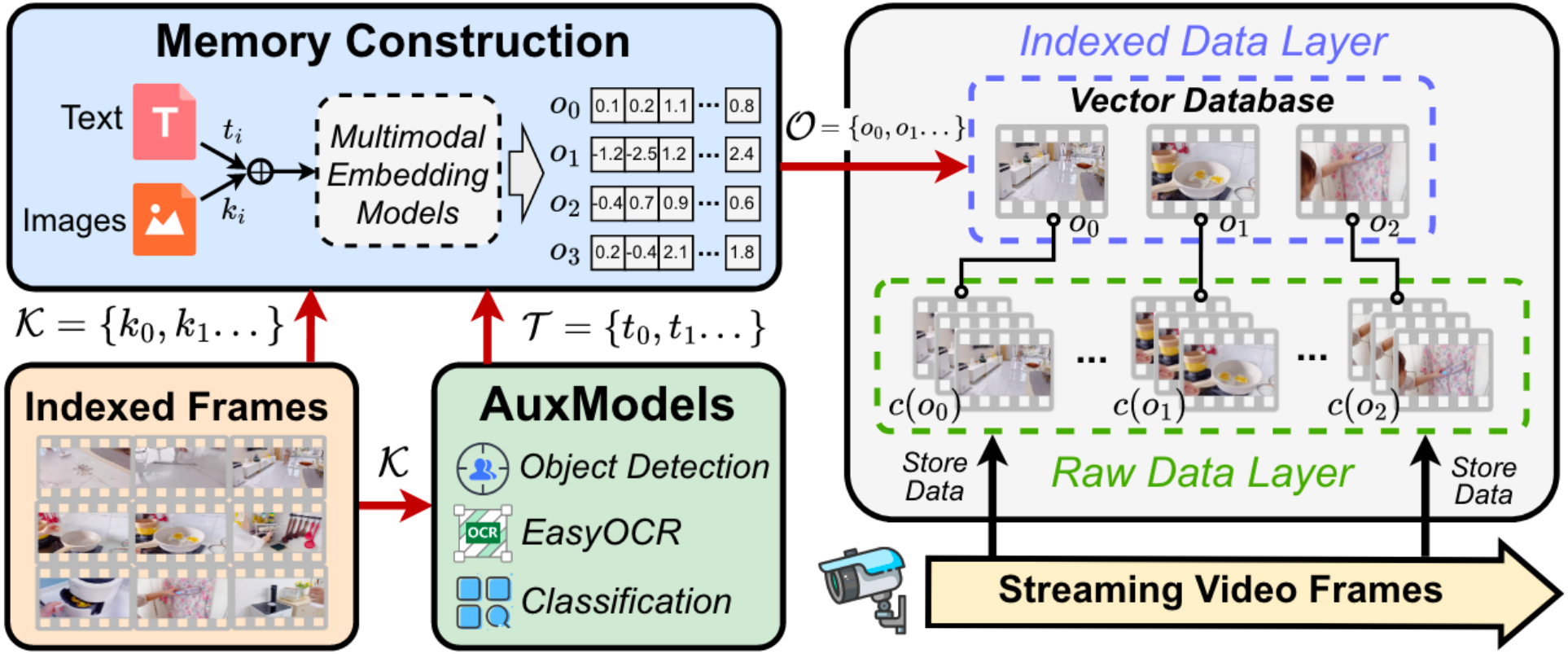
# Segmentation and Clustering for Frame Filtering

- Extracting representative centroids via segmentation and clustering to eliminate visual redundancy.



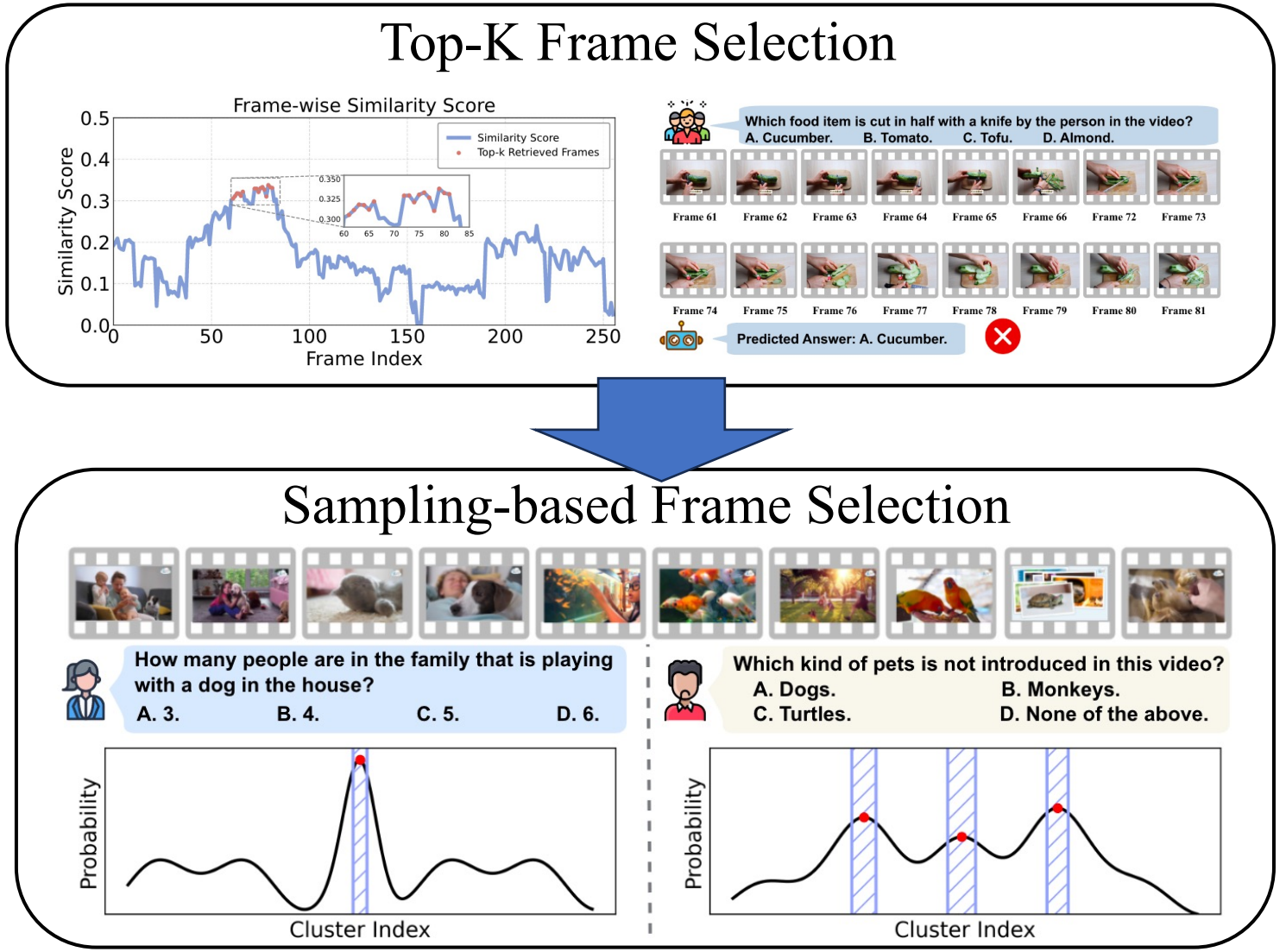
# Hierarchical Memory Construction

- Our memory management architecture is organized into two hierarchical layers: the **raw data layer** and the **index data layer**.



# Sampling-Based Frame Retrieval

- Upon receiving a user query, Venus enters the querying stage, retrieving semantically relevant frames from on-device memory and forwarding them to the cloud-hosted VLMs for reasoning.



# Evaluation

- **Hardware Setup:**
  - Edge Side: NVIDIA Jetson AGX Orin equipped with an NVMe SSD for external storage.
  - Cloud Side: A server with NVIDIA L40S GPUs (48 GB memory) to emulate cloud compute resources.
- **Datasets and Models:**
  - We evaluate Venus on two widely used VQA benchmarks, Video-MME and EgoSchema, which simulate realistic online video understanding scenarios.
  - We deploy two popular open-source VLMs on the cloud server: LLaVA-OV-7B and Qwen2-VL-7B.

# Evaluation

- **Baselines:**
  - Uniform Sampling samples frames at fixed intervals.
  - AKS proposes an adaptive keyframe selection method and employs an optimization algorithm to ensure comprehensive coverage of the selected keyframes.
  - BOLT designs an inverse transform sampling method to prioritize query-relevant frames, improving performance on multi-source retrieval VQA tasks.
  - Vanilla refers to the naive edge–cloud disaggregated architecture introduced in Section III, without any optimization.

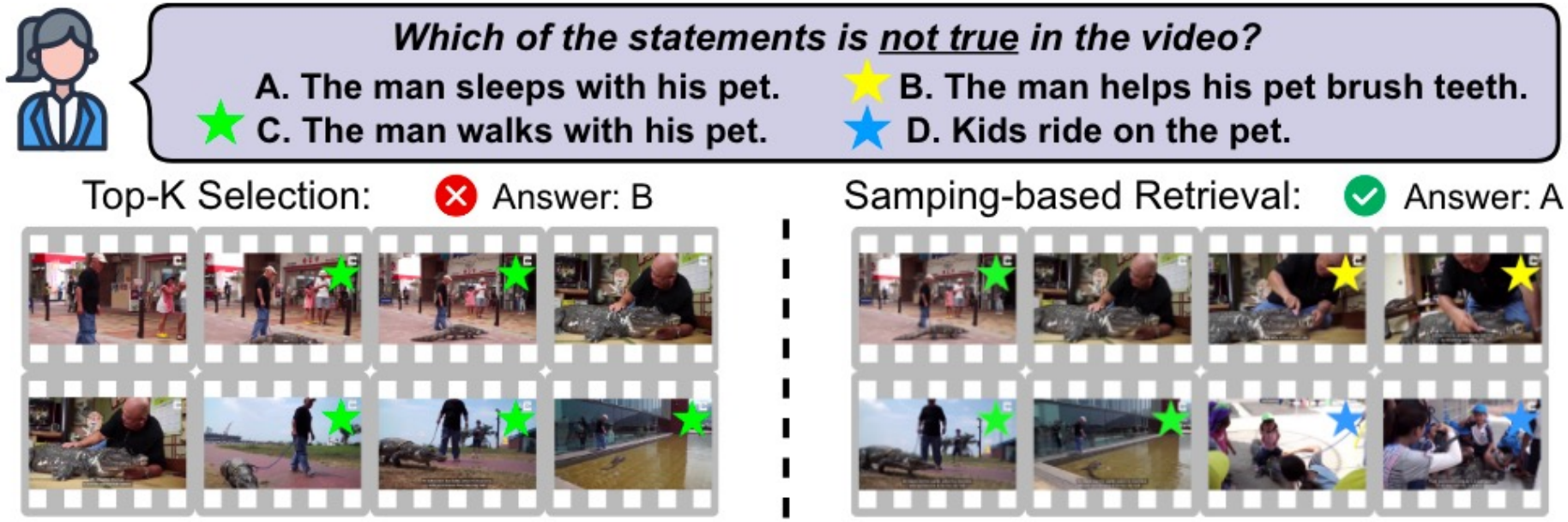
# Evaluation

- Venus achieves comparable reasoning accuracy to state-of-the-art baselines while significantly reducing query response latency.

| Model | Method | Video-MME (Short) | | Video-MME (Medium) | | Video-MME (Long) | | EgoSchema (Full) | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | Latency | Accuracy | Latency | Accuracy | Latency | Accuracy | Latency |
| LLaVA-OV-7B | AKS (*Cloud-Only*) | 65.9 | 46.8s | 52.1 | 2.7min | 52.2 | 11.2min | 51.6 | 78.3s |
| | AKS (*Edge-Cloud*) | 65.9 | 419.1s | 52.1 | 43.7min | 52.2 | 212.1min | 51.6 | 924.0s |
| | BOLT (*Cloud-Only*) | **67.9** | 43.9s | 55.9 | 2.5min | 56.2 | 10.7min | 52.4 | 70.1s |
| | BOLT (*Edge-Cloud*) | **67.9** | 398.1s | 55.9 | 41.8min | 56.2 | 206.7min | 52.4 | 896.9s |
| | Vanilla | 63.6 | 379.0s | 52.5 | 39.6min | 51.0 | 192.2min | 50.8 | 852.7s |
| | Venus | 67.4 | **4.7s** | **59.6** | **4.9s** | **56.4** | **5.1s** | **53.3** | **4.8s** |
| Qwen2-VL-7B | AKS (*Cloud-Only*) | 72.4 | 46.8s | 62.0 | 2.8min | 52.1 | 11.4min | 66.5 | 82.1s |
| | AKS (*Edge-Cloud*) | 72.4 | 417.1s | 62.0 | 44.8min | 52.1 | 214.8min | 66.5 | 950.2s |
| | BOLT (Cloud-Only) | **75.1** | 44.8s | 63.3 | 2.7min | 52.8 | 11.2min | 67.4 | 75.9s |
| | BOLT (Edge-Cloud) | **75.1** | 418.9s | 63.3 | 43.0min | 52.8 | 212.8min | 67.4 | 959.1s |
| | Vanilla | 71.9 | 391.0s | 58.7 | 41.5min | 50.9 | 190.9min | 64.0 | 894.3s |
| | Venus | 74.3 | **4.8s** | **63.8** | **5.1s** | **53.6** | **5.4s** | **69.5** | **5.0s** |

# Evaluation

- Case Study: Sampling-based retrieval provides broader visual coverage, ensuring the VLM has sufficient context for accurate reasoning.



# Thanks for listening

**Shengyuan Ye**[1], Bei Ouyang[1], Tianyi Qian[1], Liekang Zeng[2], Mu Yuan[2]
Xiaowen Chu[3], Weijie Hong[4], Xu Chen[1]

[1] Sun Yat-sen University
[2] The Chinese University of Hong Kong
[3] The Hong Kong University of Science and Technology (GZ)
[4] Shenzhen Smart City Communications Co., Ltd.