Diana: A CoT-Driven Edge-Cloud Collaborative System for Long Video Understanding

Tianyi Qian♦, Shengyuan Ye♦, Bei Ouyang♦, and Xu Chen♦▲ *
♦ School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
▲ Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China
{qianty, yeshy8, ouyb9}@mail2.sysu.edu.cn, *Corresponding author: chenxu35@mail.sysu.edu.cn

Introduction

Long video understanding in real-world edge scenarios faces a critical trade-off among privacy, bandwidth, and computational resources. Existing edge-cloud collaborative frameworks typically rely on single-pass, open-loop retrieval, which often fails to extract sufficient evidence for complex reasoning due to semantic ambiguity. To address these challenges, we propose **Diana**, a CoT-driven edge-cloud collaborative system for long video understanding. The main contributions of this paper are as follows:

- We design a content-aware perception pipeline at the edge to construct a hierarchical multimodal memory, efficiently preserving visual semantics and temporal dependencies.
- We introduce a cloud-based CoT-driven reasoning framework with a predictor and a control module, enabling autonomous ambiguity resolution via a think-and-act loop.
- We implement **Diana** on realistic edge-cloud testbeds and evaluate it on NExT-QA and MVBench. The results show that **Diana** achieves state-of-the-art accuracy while reducing end-to-end latency by over 10× compared to cloud-centric baselines.

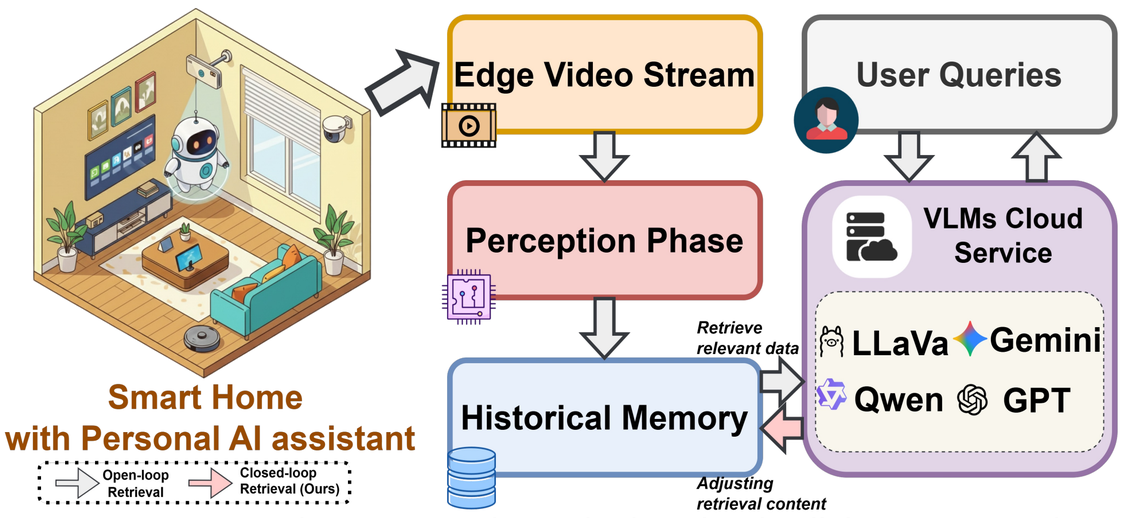

Fig. 1. Comparison between traditional edge-cloud collaborative paradigms and our proposed Diana system.

Diana System Design

■ Edge-Side Perception

The goal of edge-side perception is to decouple perception from cognition and convert redundant raw video streams into compact, queryable semantic representations.

Content-Aware Adaptive Stream Sampling

Processing all frames of a continuous video stream is costly on edge devices. **Diana** therefore measures the semantic change between the current frame F_t and the previous keyframe F_{prev} using normalized pixel difference:

$$\mathcal{D}(F_t, F_{prev}) = \frac{1}{H \times W} \sum_{i,j} \left| I(F_t)_{i,j} - I(F_{prev})_{i,j} \right|$$

A frame F_t is retained as a keyframe only if

$$\mathcal{D}(F_t, F_{prev}) > \tau_{change}.$$

Multimodal Memory Construction

For each retained keyframe, **Diana** builds a multimodal representation by combining visual embeddings with auxiliary metadata such as detected objects and recognized text. The fused representation is written as:

$$v_t = \mathcal{E}_{multimodal}(F_t, T_t) \in \mathbb{R}^d$$

where F_t is the keyframe and T_t is the textual metadata extracted by auxiliary modules.

■ Cloud-Side Active Cognition

The cloud side performs reasoning over the edge memory through an active reasoning framework rather than one-shot retrieval.

Probabilistic Keyframe Sampling

To retrieve evidence, **Diana** uses probabilistic sampling instead of deterministic Top-K.

Difficulty-Aware Query Routing

Diana uses a lightweight predictor to classify query complexity. Level One queries follow direct single-pass retrieval, while Level Multi queries trigger deeper multi-round reasoning. This routing improves the trade-off between efficiency and accuracy.

Chain-of-Thought Driven Iterative Retrieval

For complex queries, **Diana** uses a control module to perform CoT-driven iterative retrieval. The system maintains a reasoning state H_t and an accumulated evidence set $\mathcal{E}_t$ .If current evidence is insufficient, the module generates a targeted query q_t' ,retrieves more evidence from edge memory, and updates the evidence set. The loop continues until a high confidence answer is reached or the iteration limit is met, enabling better ambiguity resolution and long-video reasoning than single-pass retrieval.

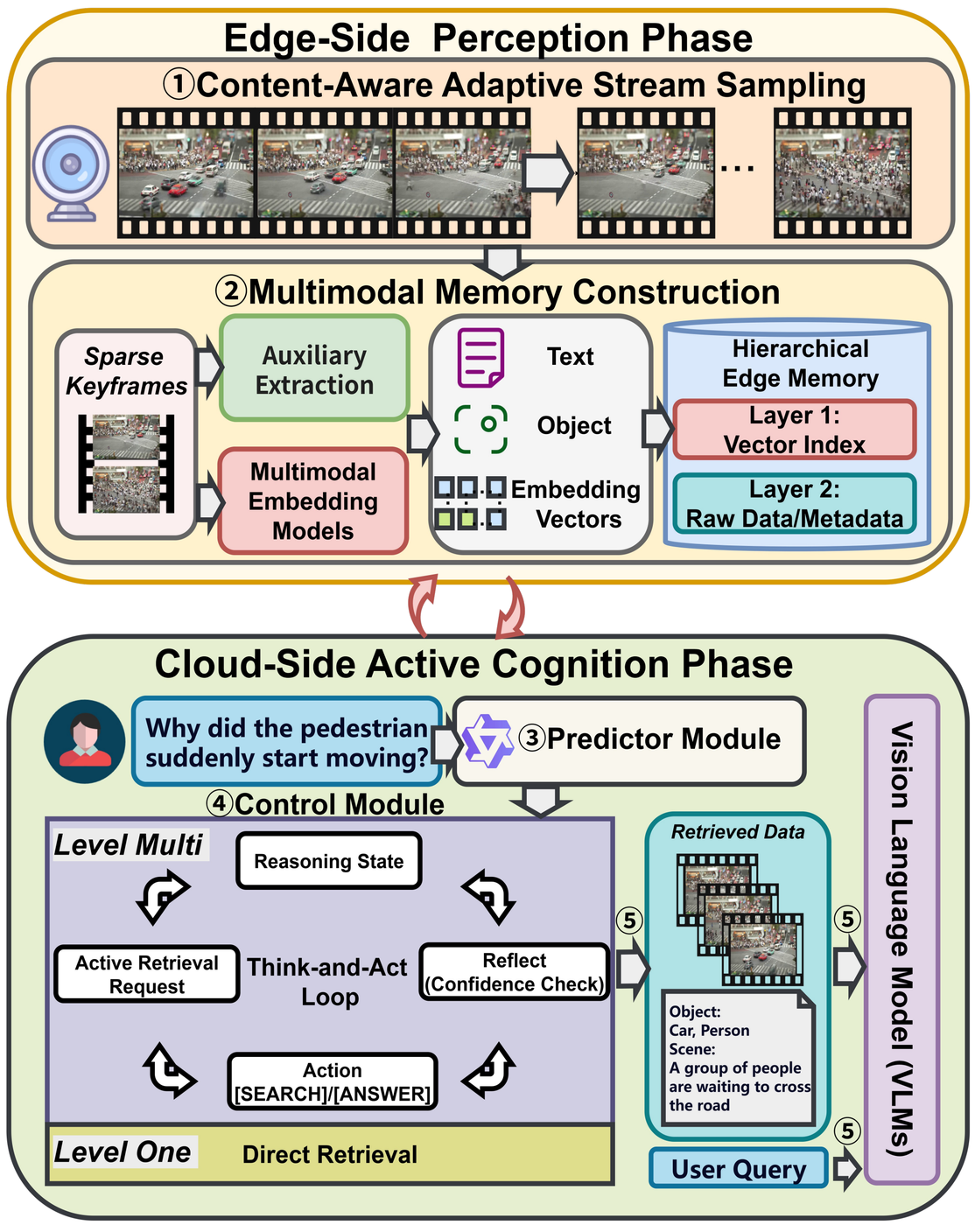

Fig. 2. System overview of **Diana**, comprising an edge-side perception phase and a cloud-side CoT-driven cognition phase.

Experiment

- **Diana** consistently outperforms Video-RAG and Venus on both NExT-QA and MVBench, demonstrating stronger long-video reasoning ability across different VLM backbones. With Qwen2-VL-7B, **Diana** achieves 78.42% accuracy on NExT-QA and 67.1% on MVBench, showing clear gains over existing edge-cloud baselines.
- In addition to higher accuracy, **Diana** significantly improves efficiency: compared with cloud-only methods, it reduces end-to-end latency by over 10×, while avoiding the memory limitations of edge-only deployment.
- The predictor further improves system efficiency by routing most queries (70-77%) through the fast path and reserving iterative reasoning only for complex cases.

TABLE I
ACCURACY COMPARISON ON THE NEXT-QA BENCHMARK.

Model	Method	NExT-QA (Frames=16)			
		Acc@C	Acc@T	Acc@D	Acc@All
Qwen2-VL-7B	Video-RAG	77.70	69.57	80.90	75.65
	Venus	78.13	72.30	82.01	76.93
	Diana	**79.18**	**74.02**	**84.47**	**78.42**
LLaVA-OV-7B	Video-RAG	72.66	67.28	79.95	72.13
	Venus	74.15	69.05	80.30	73.54
	Diana	**75.20**	**69.98**	**81.69**	**74.60**

TABLE II
PERFORMANCE COMPARISON ON MVBENCH ACROSS DIFFERENT FRAME SAMPLING BUDGETS.

Model	Method	MVBench		
		Frames=8	Frames=16	Frames=32
Qwen2-VL-7B	Video-RAG	59.8	62.6	63.8
	Venus	61.0	64.1	64.7
	Diana	**61.9**	**65.3**	**67.1**
LLaVA-OV-7B	Video-RAG	59.4	61.1	62.5
	Venus	58.6	62.3	63.4
	Diana	**60.7**	**62.9**	**64.5**

TABLE III
AVERAGE ELAPSED TIME PER QUERY AND DISTRIBUTION OF PREDICTED COMPLEXITY LEVELS FROM THE PREDICTOR ON MVBENCH AND NEXT-QA.

Dataset	Level	Comm. Latency (s)	Infer. Time (s)	Percentage (%)
MVBench	One	2.62	3.35	77.19
	Multi	3.56	12.80	22.81
NExT-QA	One	2.71	3.42	70.64
	Multi	3.47	13.21	29.36

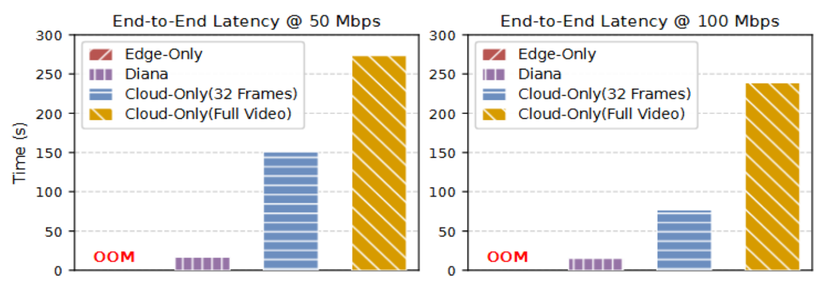

Fig. 3. End-to-end latency comparison measured on an NVIDIA AGX Orin (Edge) and an NVIDIA L40S GPU (Cloud) using a 1-minute, 30 fps video from NExT-QA. For *Diana*, the probabilistic sampling cap is set to 32 frames. We compare against Cloud-Only baselines with both full-video input and a 32-frame sampled input.

This paper was supported by Longgang District Shenzhen's "Ten Action Plan" for Supporting Innovation Projects under Grant LGKCSDPT2025001.