Diana: Edge-Cloud Collaborative Long Video Understanding via CoT-Driven Iterative Retrieval

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, chenxu35@mail.sysu.edu.cn

***Abstract*—The ubiquitous deployment of edge cameras and the emergence of multimodal large language models (MLLMs) have necessitated intelligent long video understanding at the network edge. However, current deployment paradigms face a critical trade-off. Cloud-centric approaches incur prohibitive bandwidth costs and privacy concerns, while edge-only solutions are hindered by limited computational resources. Furthermore, existing edge-cloud collaborative frameworks typically rely on single-pass, open-loop retrieval, often failing to extract sufficient evidence for complex reasoning due to semantic ambiguity. To address these challenges, we propose *Diana*, a chain-of-thought (CoT) driven edge-cloud collaborative system that strategically decouples perception from cognition. At the edge, *Diana* employs a lightweight, content-aware perception pipeline to construct a hierarchical multimodal memory for efficient video indexing. On the cloud, we introduce a dynamic reasoning framework featuring a predictor for difficulty-aware query routing and a control module for CoT-driven iterative retrieval. This architecture establishes a closed-loop reasoning mechanism, iteratively re-examining edge memory to resolve ambiguities. Extensive evaluations on the NExT-QA and MVBench benchmarks demonstrate that *Diana* achieves state-of-the-art accuracy (78.42% on NExT-QA), significantly outperforming baselines while reducing end-to-end latency by over $10\times$ compared to cloud-centric methods.**

***Index Terms*—Long Video Understanding; Edge-Cloud Collaboration; Multimodal Large Language Models; Chain-of-Thought**

I. Introduction

The advent of multimodal large language models (MLLMs) has fundamentally revolutionized the landscape of artificial intelligence, transitioning machine perception from passive sensory recording to active semantic understanding. Powered by massive cross-modal datasets, state-of-the-art VLMs [1], [2] have demonstrated unprecedented capabilities in reasoning about complex visual dynamics, identifying object relationships, and inferring intent from temporal context. These capabilities have driven the proliferation of standardized VLM APIs [3], which empower developers to seamlessly integrate advanced visual reasoning into diverse intelligent visual assistants. In smart home environments, for instance, such agents can monitor elderly activities to provide proactive healthcare services [4], while in intelligent transportation systems (ITS), they enable real-time interpretation of traffic irregularities beyond simple object detection [5]. As video sensors become ubiquitous, generating continuous high-resolution streams that account for the majority of global IP traffic, the demand for deploying these cognitive capabilities at the network edge has become increasingly critical.

This paper was supported by Longgang District Shenzhen's "Ten Action Plan" for Supporting Innovation Projects under Grant LGKCSDPT2025001.
*Corresponding author: Xu Chen.

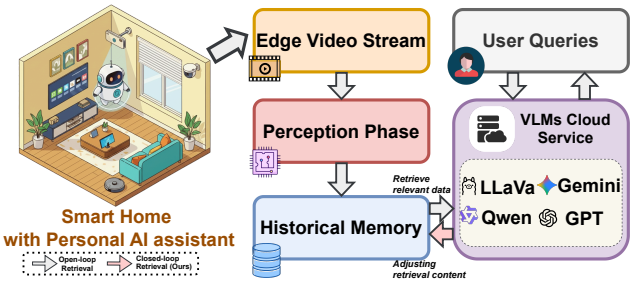

Fig. 1. Comparison between traditional edge-cloud collaborative paradigms and our proposed *Diana* system.

However, enabling long video understanding in real-world edge scenarios presents a multifaceted dilemma involving privacy, bandwidth, and computational resource constraints. Traditional cloud-centric paradigms, which necessitate uploading raw video streams for centralized processing, incur prohibitive bandwidth costs and raise severe privacy concerns, rendering them impractical for large-scale deployment. Conversely, deploying full-scale VLMs directly on edge devices, such as the NVIDIA Jetson Orin [6], is often infeasible due to hardware constraints. For instance, processing a single hour of 30 fps video yields over 100,000 frames. The resulting token sequences not only far exceed the context limits of local models like LLaVA [7] but also require massive GPU memory exceeding 80 GB to store the Key-Value (KV) cache, which is impossible for typical edge devices to accommodate. Moreover, the limited compute power at the edge leads to prohibitively long inference times. Therefore, reliance on cloud-based VLM services remains indispensable for ensuring high-performance video understanding.

To mitigate these issues, recent research has gravitated towards an edge-cloud collaborative paradigm [8], [9]. However, these pioneering works remain nascent and face significant challenges in practical deployment (as illustrated by the grey paths in Figure 1). A critical limitation lies in their reliance on single-pass, open-loop retrieval. These systems typically treat the edge memory as a static database, where the cloud issues a one-time query to fetch evidence. This mechanism is

inherently limited in robustness because if the initial retrieval fails due to semantic ambiguity or visual occlusion, the cloud model is forced to reason on incomplete or irrelevant premises, lacking the agency to re-examine the video data. Consequently, the potential of edge-cloud collaboration remains constrained by the lack of an iterative, self-correcting retrieval mechanism, hindering its maturation into a robust solution for real-world applications.

To address these limitations, we propose *Diana*, a CoT-driven edge-cloud collaborative system that fundamentally rethinks video understanding by decoupling perception from cognition. Departing from passive retrieval models, *Diana* introduces an active reasoning framework inspired by human cognitive processes. At the edge, a lightweight perception module filters redundancy and constructs a structured multimodal memory, retaining only high-value semantic information. On the cloud, we design a difficulty-aware routing mechanism via a predictor and a control module that orchestrates chain-of-thought driven iterative retrieval. This autonomous think-and-act loop enables *Diana* to actively re-examine edge memory, effectively resolving semantic ambiguities and capturing long-term dependencies while minimizing bandwidth consumption.

In summary, 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 benchmarks. Our results show that *Diana* achieves state-of-the-art accuracy (78.42%) while reducing end-to-end latency by over $10\times$ compared to baselines.

II. Diana System Design

A. System Overview

In this paper, we introduce *Diana*, a CoT-driven edge-cloud collaborative system designed for iterative and precise long video understanding. As illustrated in Figure 2, the workflow of *Diana* comprises two coordinated phases. In the *Edge-Side Perception Phase*, *Diana* continuously processes streaming video via a content-aware adaptive stream sampling module (Step ①) to filter redundancy and retain semantically significant sparse keyframes. These keyframes are processed by a multimodal memory construction module (Step ②) to organize vectors into a hierarchical edge memory. This enables a bidirectional loop: the cloud acts as the cognitive brain issuing queries, and the edge serves as perceptual memory returning visual evidence. Subsequently, in the *Cloud-Side Active Cognition Phase*, the incoming user query is assessed by a predictor (Step ③) for complexity. Complex queries activate a control module (Step ④) that orchestrates a think-and-act loop to iteratively maintain a reasoning state. Finally, the retrieved evidence is forwarded to the VLM (Step ⑤) to produce an accurate response.

B. Edge-Side Perception

To decouple perception from cognition and mitigate bandwidth bottlenecks, *Diana* transforms edge devices into active information perceptors. This phase focuses on executing information extraction and structural memory organization, converting highly redundant raw video streams into compact, queryable semantic representations for efficient downstream reasoning.

1) Content-Aware Adaptive Stream Sampling: Processing continuous video frames at standard rates incurs prohibitive overhead on edge devices. To mitigate spatiotemporal redundancy, we employ a content-aware sampling strategy. Let $\mathcal{V} = \{F_1, F_2, ..., F_t\}$ denote the input stream, where each frame has resolution $H \times W$. We quantify the semantic shift between the current frame F_t and the previous keyframe F_{prev} using a normalized pixel difference function $\mathcal{D}$:

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

A frame F_t is retained as a keyframe only if $\mathcal{D}(F_t, F_{prev})$ exceeds an adaptive threshold τ_{change}. This mechanism effectively filters redundancy while preserving significant visual evolution for downstream processing.

2) Multimodal Memory Construction: Drawing inspiration from hierarchical memory management in cognitive systems, we design a lightweight embedding pipeline to ensure low latency on edge devices. Specifically, for each retained keyframe F_t, we utilize a lightweight visual embedding model, such as BGE-VL [10], to project the visual content into a high-dimensional semantic vector v_t. To enhance retrieval accuracy, auxiliary models (e.g., YOLO [11], OCR [12]) extract explicit metadata, which are formatted into textual prompts T_t and fused with the visual embedding.

$$v_t = \mathcal{E}_{multimodal}(F_t, T_t) \in \mathbb{R}^d, \quad (2)$$

where d denotes the embedding dimension. These vectors are organized into a Hierarchical Edge Memory, comprising a vector index (via FAISS [13]) for fast similarity retrieval and a temporal metadata store for persistent archiving. This dual-layer structure efficiently maps semantic queries to temporal segments, providing a robust foundation for active reasoning without the overhead of complex knowledge graphs.

C. Cloud-Side Active Cognition

To reconcile edge resource constraints with cloud-based reasoning demands, *Diana* adopts a disaggregated architecture. While the edge handles efficient perception, the cloud executes heavy-duty cognitive tasks. Departing from conventional one-pass RAG, we introduce an active reasoning framework featuring a predictor for difficulty-aware routing and a control module for CoT-driven iterative retrieval.

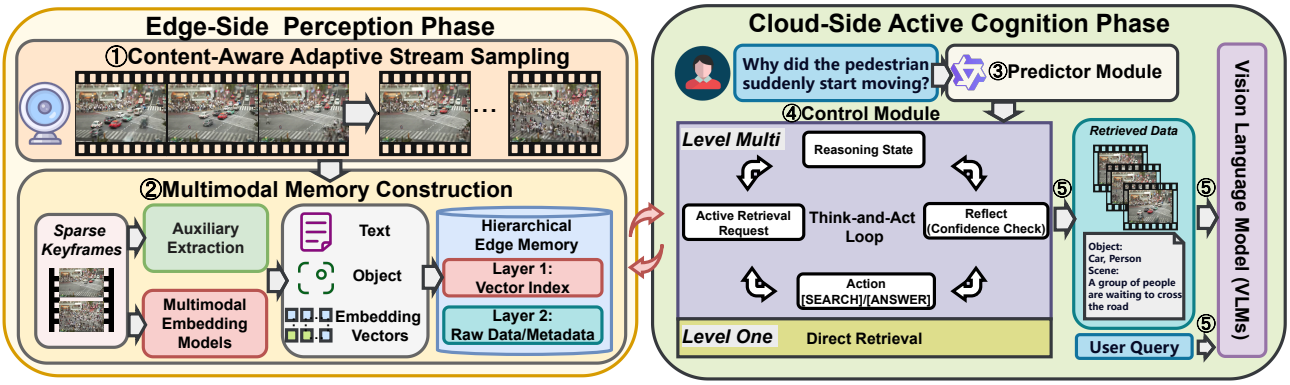

Fig. 2. System overview of *Diana*, comprising an edge-side perception phase and a cloud-side CoT-driven cognition phase.

1) Probabilistic Keyframe Sampling: Recent advances in video understanding, such as BOLT [14] and Venus [8], have demonstrated that probabilistic sampling strategies significantly outperform deterministic Top-K selection in capturing the semantic diversity of long videos. While Top-K methods prioritize the highest similarity scores, they tend to retrieve redundant frames from a single temporal cluster, thereby limiting the contextual scope available for reasoning. Aligning with this state-of-the-art paradigm, *Diana* incorporates a sampling-based retrieval mechanism to dynamically balance relevance with diversity.

Formally, for a given query q and a set of indexed vectors $\{v_1, \ldots, v_m\}$ in the edge memory, we construct a query-guided probability distribution $\Pi = \{\pi_i\}$ using a softmax function with a temperature parameter γ, which regulates the sharpness of the attention distribution to balance relevance with diversity:

$$\pi_i = \frac{\exp(\cos(q, v_i)/\gamma)}{\sum_{j=1}^{m} \exp(\cos(q, v_j)/\gamma)}. \tag{3}$$

The system then performs sampling based on Π. To adaptively control the volume of retrieved information, the sampling process accumulates frames sorted by their probability and automatically terminates when the cumulative probability mass of the selected set $\mathcal{S}$ exceeds a sufficiency threshold η (i.e., $\sum_{v_k \in \mathcal{S}} \pi_k > \eta$). This strategy ensures that the retrieved evidence covers the necessary semantic span without imposing a rigid frame count, effectively filtering redundancy while preserving long-tail context.

2) Difficulty-Aware Query Routing: Real-world video understanding tasks vary widely in cognitive demands, from simple recognition to complex reasoning. A uniform retrieval strategy is suboptimal, as heavy iterative retrieval adds latency for simple queries while single-pass methods lack depth for complex ones. To optimize the efficiency-accuracy trade-off, we introduce a difficulty-aware query routing mechanism.

This mechanism is orchestrated by the predictor module, which serves as the first line of cognitive processing. Upon receiving a user query Q, the predictor, instantiated as a lightweight LLM, analyzes its semantic complexity to determine the processing path. Queries classified as *Level One* represent simple tasks solvable via direct, single-pass retrieval, thus triggering the fast routing path to prioritize responsiveness. Conversely, queries classified as *Level Multi* denote complex tasks requiring multi-round iterative reasoning, which activate the control module to initiate a chain-of-thought driven retrieval process.

Algorithm 1 Cloud-Side Active Cognitive Reasoning Process

Require: User Query Q, Edge Memory $\mathcal{M}$, Max Iter. K_{max}
Ensure: Final Answer A
Stage 1: Difficulty-Aware Routing
$L_{type} \leftarrow \text{Predictor}(Q)$
if L_{type} is *Level One* **then**
  {**Fast Path: Direct Retrieval**}
  $\mathcal{E} \leftarrow \text{ProbabilisticSampling}(\mathcal{M}, Q)$
  return $\text{VLM}(Q, \mathcal{E})$
else
  {**Slow Path: CoT Iterative Retrieval**}
  Init: $\mathcal{E}_0 \leftarrow \emptyset, H_0 \leftarrow \emptyset$
  for $t = 1$ to K_{max} **do**
    // Reflective Reasoning (Think)
    $a_t, c_t \leftarrow \text{ControlModule}(Q, H_{t-1}, \mathcal{E}_{t-1})$
    if $a_t ==$ `[ANSWER]` **then**
      return c_t {Terminate and Answer}
    else if $a_t ==$ `[SEARCH]` **then**
      // Active Retrieval (Act)
      $q'_t \leftarrow c_t$ {Targeted Query}
      $\mathcal{E}_{new} \leftarrow \text{ProbabilisticSampling}(\mathcal{M}, q'_t)$
      $\mathcal{E}_t \leftarrow \mathcal{E}_{t-1} \cup \mathcal{E}_{new}; \quad H_t \leftarrow \text{Update}(H_{t-1}, \mathcal{E}_{new})$
    end if
  end for
  Fallback: Generate answer with accumulated evidence
  return $\text{VLM}(Q, \mathcal{E}_{K_{max}}, H_{K_{max}})$
end if

3) Chain-of-Thought Driven Iterative Retrieval: While pioneering edge-cloud collaborative paradigms [8], [9], [15] have made significant strides in efficient video understanding, they typically rely on a single-pass, open-loop retrieval mechanism. This approach, however, encounters inherent limitations when dealing with semantic ambiguity or visual occlusion, as reasoning is constrained by initial retrieval results. If critical evidence is missed in the first pass, the model may be forced to infer answers from incomplete premises. To address this

challenge, the control module implements a chain-of-thought driven iterative retrieval framework. Instead of relying on a static set of frames, this framework empowers the control module to actively maintain a dynamic reasoning state H_t, which tracks the history of logical deductions, and an evidence set $\mathcal{E}_t$, which accumulates visual frames and metadata retrieved from the hierarchical edge memory $\mathcal{M}$.

In each iteration t, the control module engages in a reflective reasoning phase based on the current state H_{t-1} and evidence $\mathcal{E}_{t-1}$. Specifically, it evaluates whether distinct information gaps, such as missing antecedents, unobserved consequences, or critical transitions, prevent a conclusive answer. Simultaneously, it assesses whether the current evidence is sufficient to converge to a high-confidence conclusion. Based on this dual assessment, the control module executes one of two distinct actions. If evidence is deemed sufficient and confidence exceeds a threshold, the agent generates a final answer (denoted as `[ANSWER]`), terminating the retrieval loop. Conversely, if significant ambiguity or missing links are detected, it outputs a search command (denoted as `[SEARCH]`) accompanied by a targeted natural language query q'_t. This query is specifically formulated to retrieve missing narrative elements and is dispatched to the edge memory $\mathcal{M}$ for focused vector retrieval. Newly retrieved frames are then aggregated into the evidence set $\mathcal{E}_t = \mathcal{E}_{t-1} \cup \text{ProbabilisticSampling}(\mathcal{M}, q'_t)$ for the next cycle. This iterative process continues until a high-confidence answer is derived or the iteration count reaches a limit K_{max}, ensuring a balance between reasoning depth and computational efficiency.

By actively investigating memory through this CoT-driven loop rather than passively accepting initial results, *Diana* significantly reduces hallucinations and improves reasoning accuracy for complex long video tasks.

III. Implementation and Evaluation

A. Experimental Setup

1) Hardware Setup: On the edge side, we use NVIDIA Jetson AGX Orin [6] for our experiments. On the cloud side, we use a server with NVIDIA L40S GPUs to emulate cloud compute resources, where the VLMs are deployed. The network bandwidth between the edge and cloud is tested at 50 Mbps and 100 Mbps to simulate different network conditions.

2) Datasets and Models: We evaluate *Diana* on MVBench [16] and the NExT-QA [17] validation set (5,000 pairs) in a zero-shot setting. NExT-QA challenges the system with diverse Descriptive (@D), Temporal (@T), and Causal (@C) reasoning tasks. MVBench provides a comprehensive assessment across 20 fine-grained temporal tasks, ranging from action sequence and object interaction to counterfactual inference, which cannot be effectively solved with a single frame.

For cloud-side reasoning, we deploy LLaVA-OneVision-7B [7] and Qwen2-VL-7B [2]. While 7B models are chosen for hardware feasibility, *Diana*'s modular design is backbone-agnostic and compatible with any cloud-hosted VLM. On the edge, we utilize BGE-VL-large [10] for multimodal embedding, complemented by YOLOv8 [11] and EasyOCR [12] to enrich the semantic context.

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**

B. Baselines

To rigorously validate the effectiveness of *Diana*, we compare it against four categories of representative baselines:

- *Video-RAG* [15] retrieves visually-aligned auxiliary texts to boost VLM performance, representing a traditional single-pass retrieval strategy without iterative optimization.
- *Venus* [8] utilizes scene segmentation and adaptive keyframe retrieval to balance efficiency and accuracy in an edge-cloud collaborative system.
- *Cloud-Only* directly uploads the entire captured video stream to the cloud server for processing, representing the most naive centralized scheme.
- *Edge-Only* executes the entire pipeline directly on the edge device, often suffering from limited reasoning capability and context windows for complex long videos.

C. Experimental Results

1) Reasoning Accuracy on Complex Benchmarks: We evaluate the reasoning accuracy of *Diana* against baselines on NExT-QA and MVBench benchmarks. As shown in Table I and Table II, *Diana* consistently outperforms both Video-RAG and Venus across different VLM backbones. Specifically, with the Qwen2-VL-7B backbone, *Diana* achieves an overall accuracy of 78.42% on NExT-QA and 67.1% on MVBench, surpassing the strongest baselines by a clear margin. In contrast to single-pass approaches that rely on static evidence, the control module enables *Diana* to perform a think-and-act loop, actively re-examining the edge memory to resolve semantic ambiguities and capture long-term temporal dependencies.

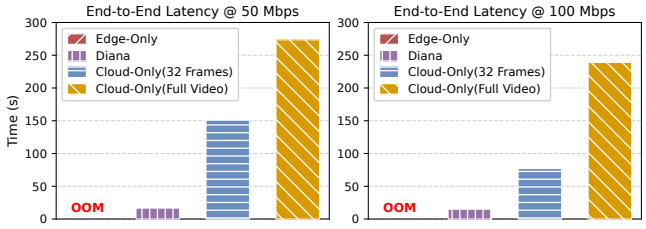

Fig. 3. End-to-end latency comparison measured on an NVIDIA AGX Orin (Edge) and an NVIDIA L40S GPU (Cloud) using a 1-minute, 30 fps video from NExT-QA. For *Diana*, the probabilistic sampling cap is set to 32 frames. We compare against Cloud-Only baselines with both full-video input and a 32-frame sampled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

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

2) Efficiency and Latency Analysis: Figure 3 illustrates the end-to-end latency comparison between *Diana* and baselines under varying bandwidths (50 Mbps and 100 Mbps). *Diana* consistently maintains low latency. In contrast, the Edge-Only baseline fails to complete tasks due to Out-Of-Memory (OOM) errors, highlighting the impracticality of deploying full-scale pipelines on resource-constrained edge devices. Meanwhile, Cloud-Only approaches suffer from severe transmission delays due to raw video uploading. Even when the VLM input is restricted to 32 frames, their total latency remains significantly higher than *Diana*, while using the full video as input incurs prohibitive latency. This demonstrates that *Diana*'s edge perception phase effectively balances edge resource constraints with cloud bandwidth bottlenecks.

3) Effectiveness of Predictor Mechanisms: Table III details the query complexity distribution, calculated as the percentage of total queries in NExT-QA and MVBench classified into each level by Qwen2-VL-7B. Under a 50 Mbps bandwidth and 32-frame budget, the predictor identifies the majority (70-77%) of queries as *Level One*, routing them through the fast path to ensure low latency. Conversely, the remaining 23-30% are designated as *Level Multi*, reserving computationally intensive iterative retrieval for these challenging instances. This selective mechanism effectively balances efficiency with reasoning depth. This distribution also elucidates the moderate overall accuracy gains, as advanced reasoning is specifically targeted at the smaller subset of complex queries.

IV. CONCLUSION

This paper presents *Diana*, a CoT-driven edge-cloud collaborative system for efficient long video understanding. *Diana* employs a decoupled architecture that enables content-aware edge perception and hierarchical memory construction, while empowering a cloud-based CoT-driven reasoning framework to autonomously resolve ambiguities via a CoT-driven think-and-act loop. Our extensive evaluation on NExT-QA and MVBench demonstrates that *Diana* achieves state-of-the-art accuracy (78.42%) and reduces end-to-end latency by over $10\times$ compared to cloud-centric baselines.

REFERENCES

[1] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)," *arXiv preprint arXiv:2309.17421*, 2023.

[2] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[3] OpenAI, "Openai api," https://openai.com/api/, 2025.

[4] Y. Chen and Y. Ren, "Analysis of artificial intelligence models for the smart home industry," *Applied and Computational Engineering*, vol. 77, no. 1, p. 117–123, Jul 2024. [Online]. Available: http://dx.doi.org/10.54254/2755-2721/77/20240664

[5] V. Hassija, T. Majumder, D. Roy, R. Piyush, and V. Chamola, "The role of large language models (llms) in enhancing intelligent transportation systems: A survey," *Vehicular Communications*, p. 100996, 2025.

[6] "Jetson orin," 2025. [Online]. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin

[7] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[8] S. Ye, B. Ouyang, T. Qian, L. Zeng, M. Yuan, X. Chu, W. Hong, and X. Chen, "Venus: An efficient edge memory-and-retrieval system for vlm-based online video understanding," *arXiv preprint arXiv:2512.07344*, 2025.

[9] Z. Xu, J. Zhang, Q. Wang, and Y. Liu, "E-vrag: Enhancing long video understanding with resource-efficient retrieval augmented generation," *arXiv preprint arXiv:2508.01546*, 2025.

[10] J. Zhou, Y. Xiong, Z. Liu, Z. Liu, S. Xiao, Y. Wang, B. Zhao, C. J. Zhang, and D. Lian, "Megapairs: Massive data synthesis for universal multimodal retrieval," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 19 076–19 095.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[12] JaidedAI, "Easyocr: Ready-to-use ocr with 80+ languages supported," https://github.com/JaidedAI/EasyOCR, 2023.

[13] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[14] S. Liu, C. Zhao, T. Xu, and B. Ghanem, "Bolt: Boost large vision-language model without training for long-form video understanding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3318–3327.

[15] Y. Luo, X. Zheng, G. Li, S. Yin, H. Lin, C. Fu, J. Huang, J. Ji, F. Chao, J. Luo *et al.*, "Video-rag: Visually-aligned retrieval-augmented long video comprehension," *arXiv preprint arXiv:2411.13093*, 2024.

[16] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo *et al.*, "Mvbench: A comprehensive multi-modal video understanding benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 195–22 206.

[17] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, "Next-qa: Next phase of question-answering to explaining temporal actions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9777–9786.