EchoVision: A Decoupled Audio-Visual Retrieval Framework for Video Question Answering

Subhajit Sarkar*, Joyita Chakraborty†, Sagnik Nandi‡, Suman Samui§, and Subrata Nandi*
*Dept. of CSE, National Institute of Technology Durgapur, India
†BlueVerse Research, LTM Limited, Kolkata, India
‡Dept. of CSE, Kalinga Institute of Industrial Technology, Bhubaneswar, India
§Dept. of ECE, National Institute of Technology Durgapur, India
{subhajit4422}@gmail.com, joyita.ckra@gmail.com, sagnik.nandi25@gmail.com,
ssamui.ece@nitdgp.ac.in, subrata.nandi@gmail.com

***Abstract*—Multimodal video question answering (VideoQA) remains largely vision-centric, underutilizing auditory signals that are essential for real-world understanding. We propose *EchoVision*, a retrieval-based audio-visual framework that abandons explicit cross-modal fusion in favor of a decoupled dual-pipeline design with implicit alignment. Independent audio and visual streams are processed using frozen pretrained models—Whisper for speech transcription, PANNs-CNN14 for audio event detection, and keyframe-based captioning with temporal reasoning traces for visual understanding and unified through a shared CLIP embedding space. We introduce a unified multimodal evidence representation that integrates transcripts, sound events, captions, and reasoning traces into a single vector-indexed knowledge base. At inference, query-conditioned retrieval selects top-k evidence, and a large language model generates answers via late fusion, enabling implicit Audio-Visual Segment Matching (AVSM) without any learned fusion or alignment parameters. Experiments on ActivityNet-QA, MovieChat1K, and the Office-Lab benchmark (490 QA pairs), newly curated by us to target audio-dependent scenarios, demonstrate consistent gains, with Answer Relevancy improving from 0.22 to 0.46 and Context Recall from 0.10 to 0.16.**

***Index Terms*—Audio Visual Retrieval, Multimodal RAG, Video Understanding, Video Question Answering,**

I. Introduction

Video understanding has evolved from early hand-crafted spatiotemporal features to modern deep learning approaches, including CNNs, two-stream networks, Vision Transformers, and 3D CNNs [1]. In recent times, vision-language models and video-LLMs [2] have also shown promise in video question answering tasks. Such systems typically fall into one of two paradigms: (i) *Video Analyzer* $\times$ *LLM*, where structured semantic descriptions (e.g., captions or events) are extracted and provided to the LLM for reasoning [3] [4]; and (ii) *Video Embedder* $\times$ *LLM*, where dense visual embeddings are directly fed into the LLM [2] [5]. Both paradigms share a fundamental limitation: they are *vision-centric*.

In real-world situations, the discriminative cue is usually provided by auditory signals. This gives rise to three common failure modes: (i) *off-screen events*, in which important information is audible but not visible; (ii) *sound-based counting*, which cannot be determined from visual frames; and (iii) *speech semantics*, in which meaning is encoded in speech rather than appearance. However, audio signals in existing benchmarks and models are still underutilized [6] [7].

Apart from the audio gap, most multimodal VideoQA models are *tightly coupled*, utilizing joint audio-visual encoders. This restricts modularity, does not allow modality components to be improved separately, and reduces robustness when a modality is missing or noisy. By contrast, EchoVision utilizes *decoupled modality-specific pipelines*, which only interact at the retrieval step to facilitate multimodal reasoning. We present EchoVision with the following contributions:

- **Decoupled dual-pipeline architecture**: Two completely independent audio and visual processing pipelines that are combined through a shared CLIP embedding space [8] and late fusion retrieval. No shared encoder; both pipelines rely on their own pre-trained models.
- **Unified multimodal evidence construction:** We have shown a method of Unified multimodal evidence construction by using a structured evidence set combining speech transcripts, audio event labels, short captions, long-form captions, and temporal reasoning traces, all indexed in a single vector store.
- **Implicit AVSM mechanism**: In our framework, there is no explicit fusion or alignment of audio and visual information in the pipeline. The fusion happens implicitly in the LLM reasoning/generation stage. There are no weights being learned in any stage. Despite the fact that there is no explicit Audio-Visual Segment Matching (AVSM) in our framework, there is an improvement in the ability to reason about the temporal order and context of events in videos, which enables answering complex queries about when and where events occur in long-duration videos.
- **Benchmark and experimental validation**: We introduce a new multimodal VideoQA dataset, "Office-Lab," with 490 QA pairs, focusing on audio-dependent scenarios. We evaluate EchoVision on ActivityNet-QA [9], MovieChat1K [10], and the proposed Office-Lab benchmark, demonstrating improved performance on audio-dependent queries.

The rest of the paper is organized as follows. In Section II,

we review some related work. Then, in Section III, we introduce our EchoVision framework. In Section IV, we present our experimental setup. Finally, in Section V, we present our results, and we conclude in Section VI.

II. Related Work

The development of video understanding has evolved from short-term action recognition to VideoQA and multimodal retrieval-augmented generation (Video-RAG) [11] [12]. Although recent work relies on Multimodal LLMs (MLLMs) to capture spatio-temporal information [13], it remains largely vision-centric and struggle with long-range temporal coherence and audio–visual synchronization.

Training-Free VideoQA: Recently, it has attracted much attention with LLoVi [3], which incorporates dense captioning and zero-shot LLM. VideoAgent [4] uses agentic retrieval based on CLIP, while DrVideo [14] adopts document retrieval-based representation of frame captions. However, this method is limited to visual inputs and does not incorporate any audio signals, which is a drawback in multimodal reasoning.

Multimodal RAG for Video: Retrieval-Augmented Generation (RAG) [15] alleviates the issue of hallucination using evidence-based outputs. VideoRAG [16] retrieves relevant video clips, while DrVideo relies on caption-based retrieval. Current AVQA approaches (QA-TIGER) prove that question-informed temporal processing via continuous mechanisms is critical; however, these methods are limited to supervised answer prediction without retrieval. On the other hand, EchoVision forms a unified multimodal evidence set $\mathcal{D}$ that includes both audio and visual features for late audio-visual fusion.

Research Gaps: The existing literature faces the following limitations: (i) audio is not a primary modality in training-free video QA; (ii) video RAG does not have audio-visual joint retrieval units; (iii) the tight coupling of multimodal encoders makes it hard to scale; and (iv) temporal and semantic alignment of the modalities is not jointly modeled in the video RAG for retrieval-based reasoning.

III. Proposed EchoVision Framework

EchoVision is based on two primary principles:

- *Decoupled independence*: Audio and visual processing are completely independent, with no shared parameters or encoders, and only rely on their own frozen pretrained models.
- *Late fusion via retrieval*: Audio and visual processing are combined only in the retrieval stage, where all evidence, regardless of type, is represented as L2-normalized vectors in a shared CLIP space and is retrieved based on cosine similarity.

None of the models used in our framework is fine-tuned or trained on any task-specific data. All the models use publicly available pretrained models. Fig. 1 shows an overview of our pipeline.

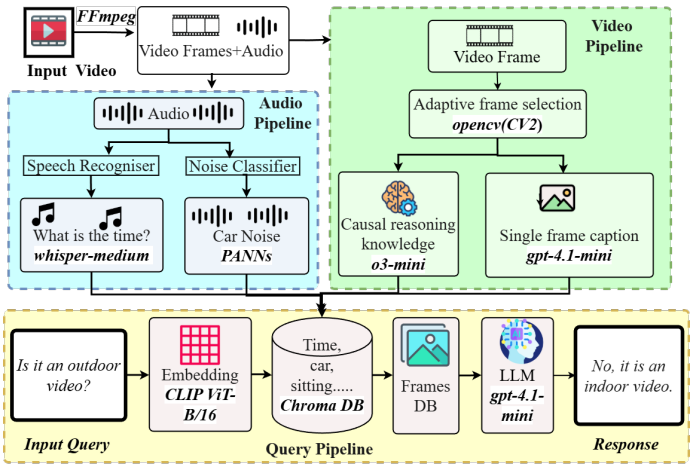

Fig. 1: The proposed EchoVision(EV) System framework

A. Problem Formulation:

Definition 1 (Multimodal VideoQA)**.** Let $X = (V, A)$ be a video consisting of visual stream V and audio stream A. Given a natural language query q, the goal is to predict an answer a^*:

$$a^* = \arg\max_a P(a \mid q, X). \tag{1}$$

Direct end-to-end modeling of Eq. (1) on long videos is intractable from a computational viewpoint and is also likely to suffer from hallucinations. EchoVision takes a *retrieval-based formulation*:

$$a^* = \arg\max_a \; P(a \mid q, \mathcal{C}_k(q)) \,, \tag{2}$$

where $\mathcal{C}_k(q)$ denotes the top-k multimodal evidence chunks retrieved from the pre-built index $\mathcal{D}$. This reformulation has three benefits: (i) it substitutes $O(T)$ frame processing with $O(|\mathcal{C}_k|)$ context; (ii) it provides LLM generation with verified retrieved evidence to reduce hallucination; (iii) it splits the task of evidence building and answer generation,

B. Multimodal Evidence Construction

The evidence set $\mathcal{D}$ is constructed offline from the two independent pipelines described below. Fig. 1 illustrates the flow.

1) Audio Pipeline: The raw audio track is obtained using FFmpeg with metadata probing to check the validity of the codec. Videos without a valid audio track are passed through the visual pipeline only. The audio is split into two parallel paths.

a) Automatic Speech Recognition (ASR).: Whisper-medium [17] is applied with frozen pretrained weights in zero-shot mode, producing word-level timestamped transcript chunks:

$$T = \{T_i\}, \qquad T_i = \left(t_s^i,\ t_e^i,\ x_i^{(T)}\right), \quad i = 1, \ldots, N_T \tag{3}$$

where $x_i^{(T)}$ is the transcribed text and $[t_s^i, t_e^i]$ is the temporal interval.

b) Audio Event Classification (AEC).: PANNs-CNN14 [18] is applied with AudioSet pretrained weights using a 1-second sliding window (0.5-second stride), producing a sound event labels for non-speech content:

$$E = \{E_j\}, \qquad E_j = \left(t_s^j,\ t_e^j,\ x_j^{(E)}\right), \quad j = 1, \ldots, N_E, \tag{4}$$

where $x_j^{(E)}$ is the predicted event class (e.g., "Thunderstorm", "Telephone bell ringing").

2) Visual Pipeline:

a) Adaptive Keyframe Selection.: Consecutive frames f_t, f_{t+1} are compared by their Bhattacharyya distance over 256-bin normalised grayscale histograms H_t:

$$d_B(f_t, f_{t+1}) = -\ln\left(\sum_{b=1}^{256} \sqrt{H_t(b) \cdot H_{t+1}(b)}\right). \tag{5}$$

A keyframe is selected when $d_B > \delta = 0.45$, corresponding to a Bhattacharyya coefficient $\rho_B = e^{-0.45} \approx 0.638$. This retains approximately one keyframe every 8 seconds for typical office videos, avoiding redundant near-duplicate frames while preserving all visually distinct scene transitions.

b) Dual-Granularity Captioning.: Each of the K selected keyframes is captioned by frozen GPT-4.1-mini at two levels of granularity:

$$C^{(s)} = \{C_m^{(s)}\}_{m=1}^K, \qquad C^{(l)} = \{C_m^{(l)}\}_{m=1}^K, \tag{6}$$

where $C_m^{(s)}$ is a short caption (1–2 sentences, max 30 words) capturing immediate visual content of keyframe $f_{k_m} \in \mathcal{K}$, and $C_m^{(l)}$ is a long-form caption (up to 120 words) including spatial relationships, visible text, and interaction semantics. Here k_m is the frame index of the m-th selected keyframe in temporal order.

c) Temporal Reasoning Traces.: EchoVision uses a frozen reasoning LLM (OpenAI o3-mini) to incrementally model temporal and semantic dependencies across keyframes. Given the sequence of long-form captions $\{C_m^{(l)}\}_{m=1}^M$, the reasoning trace at keyframe m is defined as

$$R_m = g\left(C_1^{(l)}, C_2^{(l)}, \ldots, C_m^{(l)}\right), \tag{7}$$

where $g(\cdot)$ denotes the frozen o3-mini model conditioned on all preceding captions. Each trace R_m captures cumulative temporal and causal relations, such as *before*, *after*, and *because*, while remaining anchored to keyframe m. This produces frame-aligned reasoning traces that support temporally grounded retrieval and improve causal and semantic understanding over long video sequences.

3) Unified Evidence Set: All evidence from both pipelines is collected into a single unified set:

$$\mathcal{D} = T \cup E \cup C^{(s)} \cup C^{(l)} \cup R. \tag{8}$$

Each element of $\mathcal{D}$ carries its modality type, temporal boundaries (t_s, t_e), source video ID, and raw content. The set $\mathcal{D}$ serves as the multimodal knowledge base from which query-relevant evidence is retrieved. Table I summarises the five evidence types.

TABLE I: Evidence types in $\mathcal{D}$ and their source models.

Symbol	Content	Source model	Modality
T	Speech transcripts	Whisper-medium (frozen)	Audio
E	Sound event labels	PANNs-CNN14 (frozen)	Audio
$C^{(s)}$	Short keyframe captions	GPT-4.1-mini (frozen)	Visual
$C^{(l)}$	Long-form captions	GPT-4.1-mini (frozen)	Visual
R	Temporal reasoning traces	o3-mini (frozen)	Visual

C. Embedding and Indexing

Each element $\mathcal{D}_n \in \mathcal{D}$ is first converted into a textual representation (event labels and captions are inherently textual, while images are represented via their captions) and mapped into a d-dimensional vector (d=512) space using the frozen CLIP ViT-B/16 text encoder $f_{\text{text}}(\cdot)$.

We define a generic L2-normalised embedding operator:

$$\phi(x) = \frac{f(x)}{\|f(x)\|_2} \in \mathbb{S}^{d-1}, \tag{9}$$

where $f(\cdot)$ denotes the corresponding encoder.

Thus, the textual representation is obtained as $\hat{u}_n = \phi_{\text{text}}(\mathcal{D}_n)$, where $f = f_{\text{text}}$. Similarly, for each keyframe f_k, the visual representation is computed as $\hat{v}_k = \phi_{\text{img}}(f_k)$ using the CLIP image encoder $f_{\text{img}}(\cdot)$. These representations are indexed separately to enable hybrid retrieval across modalities. All representations are stored in a Chroma vector database with deterministic identifiers and associated metadata (source type, video ID, timestamps). CLIP is employed due to its ability to perform cross-modal alignment. The pretraining of CLIP on both text and images in an aligned manner, as proposed in [8], ensures that text and images are embedded in the same space.

D. Response generation module

During inference, the user query is first encoded using the same frozen CLIP text encoder employed during indexing. Based on this representation, the top-k relevant evidence elements are retrieved from the dataset $\mathcal{D}$ using cosine similarity. These retrieved elements are then organized into a structured context string, $\text{ctx}(\mathcal{C}_k(q))$, and passed to the frozen GPT-4.1-mini model along with a system prompt sys[1] that instructs the model to rely only on the provided evidence and include timestamps in its response. The final input is constructed as a single string, $(\text{sys}||\text{ctx}(\mathcal{C}_k(q))||q)$, and processed in one API call. Notably, this design employs a late-fusion strategy: audio and visual information are not explicitly combined within the pipeline. Instead, their integration happens implicitly during the reasoning stage of the large language model. The complete inference procedure is summarized in Algorithm 1.

IV. Experimental Setup

We use two benchmarks [9], [10] and one curated audio-focused dataset. The questions span multiple categories (refer

[1]sys is a fixed *system prompt* passed as the first message to GPT-4.1-mini at every inference call. It contains three fixed instructions: (i) answer using *only* the provided retrieved evidence; (ii) cite timestamps in `[MM:SS]` format when referring to specific moments; and (iii) respond with "Insufficient evidence" if the retrieved context does not support a definitive answer.

Algorithm 1 EchoVision: Inference (per query)

Require: Query q; CLIP index $\{\hat{u}_n\}, \{\hat{v}_k\}$
Ensure: Answer a^*
// Step 1: embed query into shared CLIP space
$\hat{q} \leftarrow f_{\text{text}}(q) \,/\, \|f_{\text{text}}(q)\|_2$ ▷ $\hat{q} \in \mathbb{S}^{d-1}$, frozen CLIP text encoder
// Step 2: score all indexed evidence by cosine similarity
$s_n \leftarrow \hat{q}^\top \hat{u}_n$ **for all** n ▷ search both $\{\hat{u}_n\}$ (text) and $\{\hat{v}_k\}$ (image) indices
// Step 3: retrieve top-k evidence chunks
$\mathcal{C}_k(q) \leftarrow \text{TopK}_n\{s_n\}$, merge and re-rank by s_n
// Step 4: serialise and generate answer (late fusion)
ctx $\leftarrow$ serialise$(\mathcal{C}_k(q))$ ▷ ordered by s_n; includes modality, timestamps, content
$a^* \leftarrow$ GPT-4.1-mini(sys $\|$ ctx $\|$ q)
return a^*

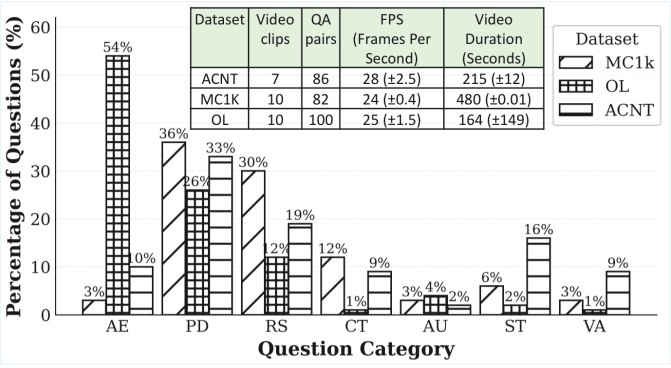

Fig. 2: Distribution of question categories across datasets.

to Fig. 2), including Audio Events (AE), Presence Detection (PD), Reasoning (RS), Counting (CT), Action Understanding (AU), Spatial-Temporal (ST), and Visual Attributes (VA), with open-ended answers.

ActivityNet-QA (ACNT) [9]: real-world activities in diverse settings; 86 questions (mainly PD, RS, ST) from 7 videos (∼3.3 min avg). **MovieChat1K (MC1k)** [10]: cinematic multimodal reasoning; 82 questions (mainly PD, RS) from 10 videos (∼8 min avg). **Office-Lab (OL)**: our curated audio–visual VideoQA dataset for indoor office/lab settings, with 490 annotated questions emphasizing audio events, speech understanding, and scene semantics (Table II). We evaluate on 10 videos (∼2.55 min avg) and 100 open-ended questions; The dataset used in this study is not publicly available but can be obtained from the corresponding author upon request.

Baselines: We compare with *LLoVi (LL)* [3] and *VideoAgent (VA)* [4]. Both are predominantly visual-caption driven: LL uses frame captions/narrations; VA augments captions with image embeddings and self-evaluates to retrieve additional frames when confidence is low. These primarily target MCQs (where options give implicit hints), whereas EchoVision targets open-ended QA via unified multimodal evidence and decoupled audio-visual processing.

V. Results and Analysis

We evaluate using ROUGE-L (structural similarity), BERTScore (semantic alignment), and the RAGAS-style retrieval metrics Answer Relevancy, Faithfulness, Context Precision/Recall/Relevancy (Table III).

TABLE II: Sample responses from frameworks on audio-centric questions.

Q, GT, and Answers from LLoVi (LL), VideoAgent (VA), EchoVision (EV)
Q: What sound is heard when the cabinet is opened? GT: A metallic clanking sound. LL: silence. VA: unknown. *EV: A metallic clanking sound is heard when the cabinet is opened.*
Q: Does the paper slide before or after you see it come out? GT: Sound begins before paper is fully visible. LL: No sound. VA: Silence. *EV: The sound begins before the paper is fully visible due to internal roller friction.*
Q: Does the footstep sound like a slap or a hard clack? GT: A sharp 'clack'. LL: No audio. VA: No clack present. *EV: It is a sharp "clack".*

Fig. 3: Sample keyframes from the Office-Lab (OL) dataset illustrating typical indoor activities used in audio-visual reasoning queries.

Audio-centric dataset (OL). EV achieves the strongest ROUGE-L (0.51 vs. 0.23/0.29) and BERTScore-F1 (0.93 vs. 0.89/0.90), and the highest absolute retrieval scores (Context Precision 0.05, Context Recall 0.16). On Answer Relevancy, VA narrowly exceeds EV (0.52 vs. 0.46) but with much lower Context Recall (0.06 vs. 0.16), indicating answer-likeness without supporting evidence; LL's near-perfect Faithfulness (0.99) coexists with very low Answer Relevancy (0.22), reflecting frequent "insufficient evidence" refusals on audio queries. The absolute Context Precision (0.05) is low across all systems—an open challenge for retrieval-based audio-visual QA—but EV's $5\times$ margin over baselines on OL shows that the unified evidence index recovers far more relevant audio chunks than caption-only pipelines. Together, these results indicate EV trades a small Answer Relevancy gap for substantially better grounding and retrieval coverage on audio-dependent queries.

Vision-dominant datasets (MC1k, ACNT). On MC1k, VA's MCQ-tuned design boosts ROUGE-L (0.70) and BERTScore, but EV remains competitive in open-ended QA (BERTScore-F1 0.90). On ACNT—where queries are largely visual—EV underperforms LL/VA on lexical and BERTScore metrics, consistent with the dataset's low audio dependency; the gap narrows on Answer Relevancy (EV 0.47 vs. VA 0.50). This confirms EV's gains are concentrated where audio matters, with no advantage when discriminative information lies in the visual stream alone.

G-Eval (LLM-as-a-judge). On OL, EV reaches a mean correctness of 3.46 ($\pm$1.71), substantially exceeding LL (2.11) and VA (2.26); on ACNT, EV again leads (2.50 vs. LL 1.76, VA 2.12). The distribution (Fig. 4) shows EV producing the highest frequency of "5" (perfect) ratings and the lowest count of "1" ratings, while LL peaks sharply at the lowest tier (>160 questions). EchoVision thus raises both the ceiling and the floor of response quality on audio-centric reasoning.

TABLE III: Performance comparison of different frameworks across datasets.

Framework (Dataset)	ROUGE-L	BERTScore-P	BERTScore-R	BERTScore-F1	Answer Relevancy	Faithfulness	Context Precision	Context Recall	Context Relevancy
EV (MC1k)	0.21 (±0.15)	0.87 (±0.02)	0.93 (±0.03)	0.90 (±0.02)	0.60 (±0.46)	0.89 (±0.30)	0.15 (±0.27)	0.31 (±0.46)	0.04 (±0.09)
LL (MC1k)	0.41 (±0.39)	0.93 (±0.04)	0.93 (±0.05)	0.93 (±0.04)	0.68 (±0.38)	0.90 (±0.28)	0.11 (±0.24)	0.27 (±0.44)	0.05 (±0.12)
VA (MC1k)	0.70 (±0.47)	0.96 (±0.07)	0.96 (±0.06)	0.96 (±0.06)	0.61 (±0.49)	0.94 (±0.24)	0.56 (±0.52)	0.29 (±0.46)	0.00 (±0.01)
EV (ACNT)	0.12 (±0.14)	0.84 (±0.03)	0.87 (±0.03)	0.85 (±0.03)	0.47 (±0.50)	0.90 (±0.31)	0.07 (±0.20)	0.25 (±0.43)	0.07 (±0.21)
LL (ACNT)	0.16 (±0.30)	0.90 (±0.04)	0.88 (±0.03)	0.89 (±0.03)	0.64 (±0.48)	1.00 (±0.00)	0.23 (±0.42)	0.41 (±0.50)	0.09 (±0.18)
VA (ACNT)	0.27 (±0.45)	0.92 (±0.05)	0.90 (±0.07)	0.91 (±0.06)	0.50 (±0.50)	0.94 (±0.24)	0.08 (±0.24)	0.23 (±0.42)	0.07 (±0.22)
EV (OL)	**0.51 (±0.37)**	**0.92 (±0.07)**	**0.93 (±0.04)**	**0.93 (±0.05)**	0.46 (±0.46)	0.96 (±0.21)	**0.05 (±0.17)**	**0.16 (±0.37)**	**0.05 (±0.11)**
LL (OL)	0.23 (±0.34)	0.92 (±0.04)	0.87 (±0.03)	0.89 (±0.03)	0.22 (±0.42)	0.99 (±0.10)	0.01 (±0.17)	0.10 (±0.30)	0.02 (±0.10)
VA (OL)	0.29 (±0.46)	0.91 (±0.07)	0.89 (±0.08)	0.90 (±0.07)	0.52 (±0.50)	1.00 (±0.00)	0.01 (±0.07)	0.06 (±0.24)	0.01 (±0.10)

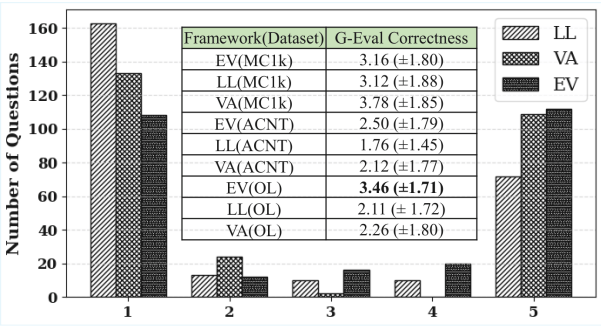

Fig. 4: G-Eval score distribution across frameworks.

VI. Conclusion

EchoVision is a training-free multimodal VideoQA framework that elevates audio alongside vision through a decoupled retrieval architecture, unifying transcripts, sound events, captions, and reasoning traces in a single RAG-indexed evidence base. We acknowledge that all modalities are aligned via CLIP *text* embeddings, making the present design effectively audio–visual–text rather than purely audio–visual; cross-modal interaction between raw audio and visual features is left implicit to the LLM stage. Future work targets (i) adaptive modality weighting and dynamic prioritization, (ii) unimodal (audio-only, vision-only) ablations to isolate per-modality contribution, and (iii) multi-video retrieval and inference efficiency for real-world deployment.

References

[1] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[2] M. Maaz, H. Rasheed, S. Khan, and F. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 12 585–12 602.

[3] C. Zhang, T. Lu, M. M. Islam, Z. Wang, S. Yu, M. Bansal, and G. Bertasius, "A simple llm framework for long-range video question-answering," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 21 715–21 737.

[4] X. Wang, Y. Zhang, O. Zohar, and S. Yeung-Levy, "Videoagent: Long-form video understanding with large language model as agent," in *ECCV*, 2024.

[5] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[6] Y. Yang, J. Zhuang, G. Sun, C. Tang, Y. Li, P. Li, Y. Jiang, W. Li, Z. Ma, and C. Zhang, "Audio-centric video understanding benchmark without text shortcut," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 6580–6598.

[7] A. Sheikh, S. Samui, and N. Chattaraj, "Exploring the potential of state-of-the-art speaker diarization frameworks for multilingual multi-speaker conversations," in *2025 2nd International Conference on Circuits, Power and Intelligent Systems (CCPIS)*. IEEE, 2025, pp. 1–6.

[8] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[9] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9127–9134.

[10] E. Song *et al.*, "Moviechat: From dense token to sparse memory for long video understanding," in *CVPR*, 2024.

[11] R. Raja and A. Vats, "Multimedia-aware question answering: A review of retrieval and cross-modal reasoning architectures," in *Proceedings of the 2nd ACM Workshop in AI-powered Question & Answering Systems*, 2025, pp. 28–35.

[12] S. Sarkar, S. Mandal, J. Chakraborty, and S. Nandi, "Query based deep contextual video understanding," in *Intelligent Human Computer Interaction: 17th International Conference, IHCI 2025, Jaipur, India, November 14–16, 2025*, ser. Lecture Notes in Computer Science, vol. 16437. Cham: Springer, 2026, in press.

[13] Y. Tang, J. Bi, S. Xu, L. Song, S. Liang, T. Wang, D. Zhang, J. An, J. Lin, R. Zhu *et al.*, "Video understanding with large language models: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[14] Z. Ma, C. Gou, H. Shi, B. Sun, S. Li, H. Rezatofighi, and J. Cai, "Drvideo: Document retrieval based long video understanding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 18 936–18 946.

[15] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *NeurIPS*, 2020.

[16] S. Jeong *et al.*, "Videorag: Retrieval-augmented generation over video corpus," in *EMNLP*, 2024.

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[18] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.