iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.80091

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Comparative Study of Single-Pass and Coarse-to-Fine Temporal Question Answering Models for Egocentric Video Analysis

Aswin A[1], Feona Teresa Fly[1], Krishnapriya Prasannan[2], Prof. Rotney Roy Meckamalil[3], Prof. Sumi Joy[4], Prof. Richu Shibu[5]
Department of Computer Science and Engineering, Mar Athanasius College of Engineering (Autonomous), Kothamangalam

Abstract: ***Understanding and retrieving information from long egocentric videos is a challenging problem due to their unstructured nature and the presence of complex temporal events. The project presents a comparative study between two Temporal Question Answering (Temporal-QA) frameworks designed for egocentric daily activity videos. The first model employs a single-pass Audio-Visual framework with Gaussian Contrastive Grounding (GCG) to localize relevant temporal segments and generate answers using multimodal reasoning. While effective, the approach is limited by sparse temporal sampling, which can miss short-duration events. The second model introduces an enhanced Coarse-to-Fine (C2F) hierarchical grounding strategy. The model performs a global coarse scan to identify candidate segments, followed by a fine-grained local analysis to capture detailed temporal information. It also extends the grounding mechanism using a Gaussian Mixture Model (GMM) and im-proves multimodal integration through bidirectional audiovisual fusion. The comparative analysis evaluates both models on the NExT-QA dataset using metrics such as accuracy and reasoning capability. Experimental results demonstrate that the C2F model significantly outperforms the single-pass baseline, achieving im-proved temporal localization and better handling of complex causal and short duration events.***

I. INTRODUCTION

The rapid proliferation of wearable devices and smart sensing technologies has led to an exponential growth in egocentric video data, capturing continuous first-person per-spectives of daily human activities [1]. These videos provide rich contextual information about real-world environments, including object interactions, human behavior, and temporal event sequences. However, the inherently unstructured and long-duration nature of such data makes manual analysis inefficient and impractical, particularly for applications such as healthcare monitoring, personal memory assistance, and intelligent surveillance systems.

Temporal Question Answering (TQA) has emerged as a promising paradigm to address these challenges by enabling users to query video content using natural language [2], [3]. Unlike traditional video analysis techniques, TQA systems aim not only to understand visual content but also to reason over temporal relationships between events. This involves identifying relevant temporal segments within a video and gen-erating contextually accurate answers. By integrating computer vision, natural language processing, and temporal reasoning, TQA systems facilitate effective retrieval of information from complex and unstructured video streams.

Recent advances in multimodal learning and Large Multi-modal Models (LMMs) have significantly enhanced the ca-pabilities of Video Question Answering (VideoQA) systems
[4]–[6]. In particular, temporal grounding plays a crucial role in linking textual queries to specific segments in a video [7]. Among existing approaches, the Single-Pass Gaussian Contrastive Grounding (GCG) framework [8] provides an efficient mechanism for temporal localization by modeling relevance using a Gaussian distribution. However, its reliance on uniform frame sampling can limit its ability to capture short-duration events and fine-grained temporal details.

Alternatively, hierarchical grounding strategies such as the Coarse-to-Fine (C2F) framework adopt a multi-stage analysis by identifying coarse temporal regions followed by fine-grained refinement. While this approach enables improved temporal resolution and more detailed analysis, it introduces additional computational complexity and design overhead.

Thus, both Single-Pass and Coarse-to-Fine frameworks ex-hibit distinct strengths and limitations in terms of efficiency, temporal precision, and scalability. Rather than positioning one approach as a direct solution to the limitations of the other, it is important to analyze them comparatively to understand the trade-offs involved in their design and performance.

In this work, we present a comparative study of Single-Pass and Coarse-to-Fine Temporal Question Answering frameworks for egocentric daily activity videos. The study systematically examines their architectural design, temporal grounding strate-gies, and multimodal integration techniques. Furthermore, we evaluate their performance on the NExT-QA benchmark [2] using metrics such as accuracy and reasoning capability.

Rather than proposing a single superior solution, this work emphasizes understanding the trade-offs between simplicity and complexity, efficiency and precision, and global versus hierarchical temporal reasoning. The insights derived from this study contribute to a deeper understanding of TQA systems and provide guidance for designing more effective and balanced video understanding frameworks.

II. RELATED WORKS

A. *Video Question Answering*

Video Question Answering (VideoQA) has evolved significantly with the advancement of vision-language learning. Early approaches relied on recurrent neural networks (RNNs) combined with attention mechanisms to capture temporal dependencies in video sequences [9], [10]. To better model long-term temporal relationships, memory-based architectures such as Motion-Appearance Co-Memory Networks [11] and Hierarchical Interactive Memory Networks [12] were intro-duced, enabling the storage and retrieval of temporal context. More recent approaches leverage Transformer-based archi-tectures [3], [13], which provide improved scalability and contextual representation learning. Large Multimodal Models such as InstructBLIP [6] and Video-LLaMA [14] integrate powerful language models like LLaMA [5] to perform reason-ing across modalities. These models demonstrate strong gen-eralization capabilities and improved performance in complex VideoQA tasks. Benchmark datasets such as NExT-QA [2] and Ego4D [1] have further driven research toward temporal and causal reasoning in long-form egocentric videos. Re-cent advancements have also explored alternative architectures for VideoQA beyond transformer-based designs. Progressive Attention Memory Networks (PAMN) [15] improve tempo-ral reasoning by iteratively refining attention over memory representations. Similarly, MovieQA [16] introduced story-level reasoning in videos, emphasizing narrative understanding rather than frame-level analysis. These approaches highlight the importance of long-term temporal coherence and struc-tured reasoning in video-based question answering tasks.Self-chained reasoning approaches [17] improve the model's abil-ity to handle complex queries by decomposing them into sequential reasoning steps across temporal segments. Spatio-temporal rationale extraction methods [18] identify the most relevant video regions and temporal segments that contribute to a model's prediction. This enhances transparency by pro-viding visual justifications for generated answers, which is particularly important for real-world applications requiring explainability.

B. *Temporal Grounding and Weak Supervision*

Temporal grounding is a critical component of VideoQA, aiming to localize relevant segments in a video corresponding to a given query [7]. Traditional methods rely on densely annotated temporal boundaries [19], which are expensive and difficult to scale. As a result, weakly supervised approaches have gained popularity, utilizing only video-level annotations for training [20], [21].

Gaussian Contrastive Grounding (GCG) [8] introduces a weakly supervised approach that models temporal relevance using a Gaussian distribution while leveraging CLIP-based pseudo-labels. This method improves temporal continuity and reduces dependence on manual annotations. However, its single-pass design may limit its ability to capture fine-grained temporal details in long-duration videos.Symbolic re-play methods [22] enhance temporal reasoning by reconstruct ing structured event sequences, enabling better understanding of long-duration activities.

Hierarchical grounding strategies such as the Coarse-to-Fine (C2F) framework provide an alternative approach by perform-ing multi-stage analysis. These methods first identify coarse candidate segments and subsequently refine them through fine-grained processing. While this improves temporal precision and enables better localization of short-duration events, it introduces additional computational complexity.

C. *Audio-Visual Learning*

Most traditional VideoQA systems primarily focus on visual features, often neglecting the complementary role of audio. However, audio signals provide important contextual information such as speech, environmental sounds, and event-specific cues that may not be visually observable. Recent multimodal frameworks, including Flamingo [23], demonstrate the effectiveness of integrating multiple modal-ities for improved reasoning. Despite this, many existing VideoQA approaches either ignore audio or utilize basic audio representations.

Advanced audio models such as Whisper enable the extraction of rich semantic features, improving the understanding of auditory context. Large-scale visual represen-tation models such as EVA [24] have significantly improved feature extraction capabilities for video understanding tasks. These models leverage masked visual pretraining to capture high-level semantic features, which can enhance downstream tasks such as VideoQA. Additionally, multimodal architectures that combine strong visual encoders with language mod-els continue to demonstrate improved reasoning performance across complex video scenarios.

The integration of audio and visual modalities through cross-attention mechanisms has shown significant improve-ments in temporal reasoning and question answering perfor-mance. This is particularly beneficial in scenarios involving off-screen events, ambiguous visual cues, or sound-driven interactions, highlighting the importance of multimodal fusion in egocentric video understanding.

III. METHODOLOGY

A. Single-Pass GCG Framework

The proposed Temporal Question Answering Framework follows a modular and hierarchical pipeline designed to trans-form raw multimodal video data into temporally grounded natural-language answers. The methodology consists of six major components: input acquisition, preprocessing (Visual & Audio), pseudo-label generation, Gaussian Contrastive Grounding (GCG), Audio-Visual Fusion, and final answer generation.

1) Input Acquisition

The system receives two inputs: (i) an egocentric video file captured using wearable devices, and (ii) a natural-language query. Videos are normalized for resolution and frame rate to handling varying recording formats. Frame timestamps are synchronized to ensure strict temporal correspondence across the visual and auditory processing stages.

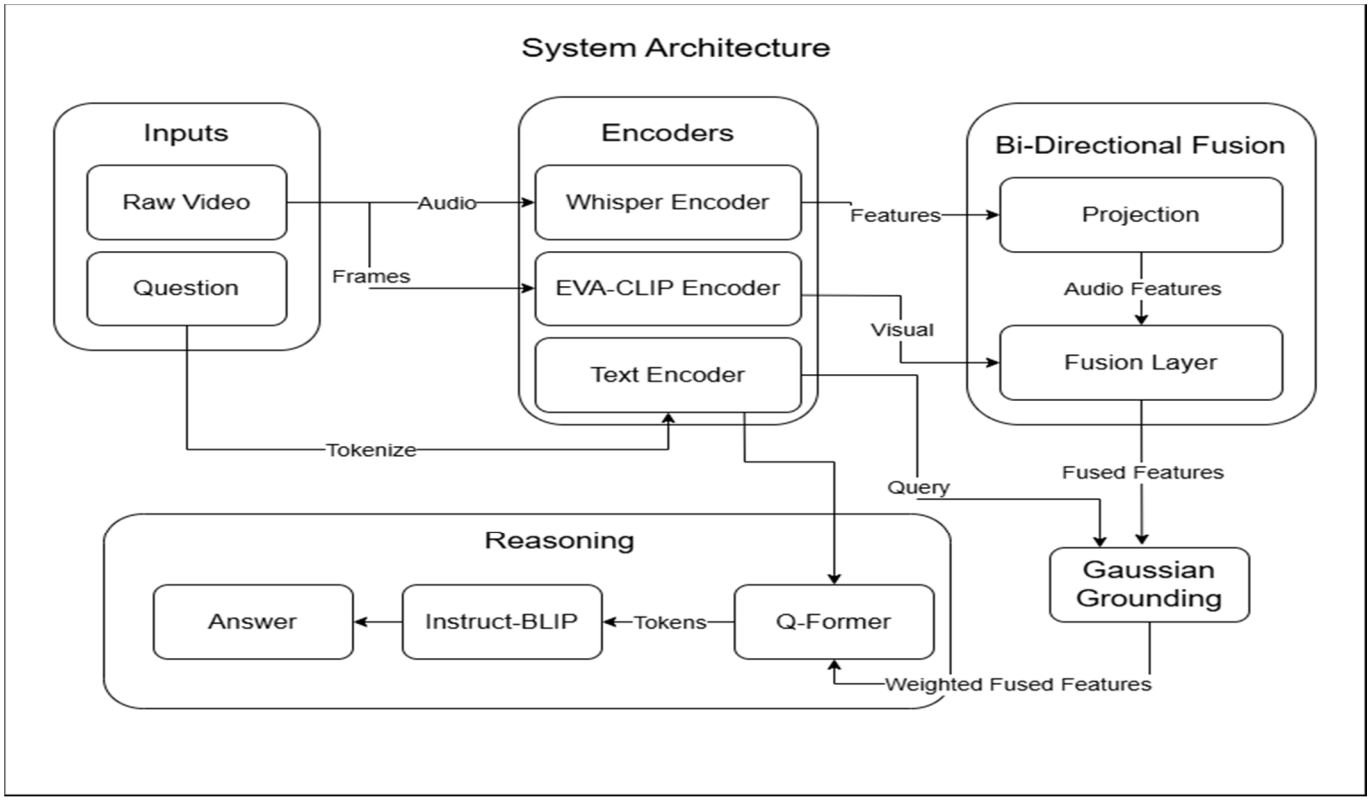


Fig. 1: Comprehensive System Architecture. Validated flow: Encoders → Bi-Directional Fusion → Gaussian Grounding → Reasoning

2) Preprocessing And Feature Extraction

Feature extraction is the fundamental process of transform-ing raw high-dimensional data such as video pixels, audio waveforms, and textual queries into compact, informative vector representations that are computationally efficient for the model to process. In our proposed framework, we extract a comprehensive set of multimodal features to capture the complex dynamics of egocentric videos. Specifically, visual features (derived from EVA-CLIP) encode high-level spatial semantics, including object identities, hand-

object interactions, scene layouts, and visual appearance cues. Audio features (derived from Whisper) capture the acoustic environment, encompassing speech content, distinct sound events, ambient background noise, and temporal rhythm patterns. Finally, tex-tual features (derived from Flan-T5) encapsulate the semantic intent of the user's question, identifying key entities and the type of reasoning (causal, temporal, or descriptive) required. These extracted features serve as the foundational inputs for our cross-modal reasoning and grounding modules.

a) Visual Stream

We employ the EVA-CLIP (ViT-g/14) vision encoder as our visual backbone. Key frames are sampled uniformly from the video (T = 32 frames) and resized to 224 224 pixels. The encoder processes these frames to generate dense frame-level embeddings f_v R^{T} $^{\times d}v$, where d_v = 1408. These embed-dings capture high-level spatial semantics, object presence, and scene context.

b) Audio Stream

A distinct audio pipeline is implemented to handle the auditory modality. $\in$

- Extraction: Raw audio is separated from the video container using FFmpeg and resampled to 16kHz mono.
- Feature Encoding: We utilize the OpenAI Whisper (Base) model as a frozen feature extractor. The raw waveform is converted into a log-Mel spectrogram and processed by the Whisper Transformer encoder. We ex-tract the output of the final hidden layer, capturing rich semantic acoustic features.
- Interpolation: The extracted audio features are tem-porally interpolated to strictly align with the T visual frames, resulting in a synchronized sequence f_a $\mathrm{R}^{T\times d}a$, where d_a = 512.

c) Text Encoding $\in$

The user's natural language query is tokenized and encoded using the Flan-T5-XL tokenizer and text encoder (part of the InstructBLIP architecture), producing contextual query embeddings $q \in \mathrm{R}^{M\times d}t$.

Collectively, these feature streams provide a semantically rich and temporally synchronized representation of the ego-centric experience. By projecting visual and textual embed-dings into a shared latent space and strictly aligning auditory cues with visual frames, the framework establishes a solid foundation for the subsequent cross-modal reasoning stages. This alignment is critical for enabling the model to not only recognize objects and sounds but to understand their causal relationships and temporal ordering in response to the user's query.

3) Pseudo-Label Generation

To eliminate the need for manually annotated temporal bound-aries, we employ a weak supervision strategy. We compute the cosine similarity between each visual frame embedding f_t and the global query embedding q:

$$s_t = \frac{f_t \cdot q}{\|f_t\| \|q\|} \quad (1)$$

Frames exhibiting similarity scores above an adaptive thresh-old τ_{sim} are treated as pseudo-positive labels. Temporal smoothing is applied to ensure continuity, grouping relevant frames into coherent event segments.

4) Gaussian Contrastive Grounding (GCG)

The GCG module is responsible for temporally localizing the query-relevant segments. We model the temporal attention distribution as a Gaussian kernel parameterized by a learnable mean μ (center) and variance σ^2 (width). The Gaussian weight for frame t is computed as:

$$w_t = \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \quad (2)$$

The weighted video representation is then aggregated:

$$v_{weighted} = \sum_{t=1}^{T} w_t \cdot f_t \quad (3)$$

A contrastive learning objective aligns this weighted video representation with the query embedding while separating it from negative samples:

$$L_{contrastive} = -\log \frac{\exp(\text{sim}(v_{weighted}, q)/\tau)}{\sum_j \exp(\text{sim}(v_{weighted}, q_j)/\tau)}$$

Through backpropagation, the network learns to adjust μ and σ to focus on the most relevant temporal window for the given question.

5) Bi-Directional Audio-Visual Fusion

To integrate the audio and visual representations effectively, we propose a symmetric fusion architecture. First, the lower-dimensional audio features ($d_a = 512$) are projected to the visual dimension ($d_v = 1408$) using a learnable linear projection network , yielding projected audio features $\hat{f}a$. We then employ dual Multi-Head Cross-Attention layers:

a) Vision-to-Audio (V2A) Attention

The visual features act as the Query, attending to the audio features (Key/Value) to identify sounds relevant to the visible action:

$$A_{v2a} = \text{softmax}\left(\frac{f_v (f_a)^T}{\sqrt{d_k}}\right) f_a \qquad (5)$$

b) Audio-to-Vision (A2V) Attention

The audio features act as the Query, attending to the visual features (Key/Value) to ground sounds to specific visual regions:

$$A_{a2v} = \text{softmax}\left(\frac{f_a (f_v)^T}{\sqrt{d_k}}\right) f_v \qquad (6)$$

c) Fusion

The outputs of these two attention streams are concatenated and passed through a fusion layer (Linear + LayerNorm + GELU) to produce a unified multimodal representation f_{final}, enriched with both visual and acoustic context. A residual connection preserves the original visual information.

6) Answer Generation (InstructBLIP)

The fused features f_{final} are fed into the Q-Former, a lightweight transformer that compresses the variable-length video sequence into a fixed set of query-aware tokens. These tokens are then concatenated with the encoded text query and provided as soft prompts to the frozen Flan-T5-XL Large Language Model. The LLM performs auto-regressive decoding to generate the final natural language answer, leveraging the grounded audio-visual evidence.

B. Coarse-to-Fine (C2F) Temporal Grounding Framework

1) Coarse-Level Localization

In the first stage, the model performs a global scan of the video by uniformly sampling frames to obtain a holistic representation. Visual features f_t are compared with the query embedding q using cosine similarity:

$$s_t = \frac{f_t \cdot q}{\|f_t\| \, \|q\|} \qquad (7)$$

Frames with high similarity scores are grouped into coarse candidate segments. This stage prioritizes coverage of the entire video to avoid missing relevant events.

2) Fine-Level Temporal Refinement

In the second stage, the model focuses on the selected coarse segments and performs dense frame sampling to capture fine- grained temporal details. This improves the detection of short-duration and complex actions that may be missed during coarse analysis.

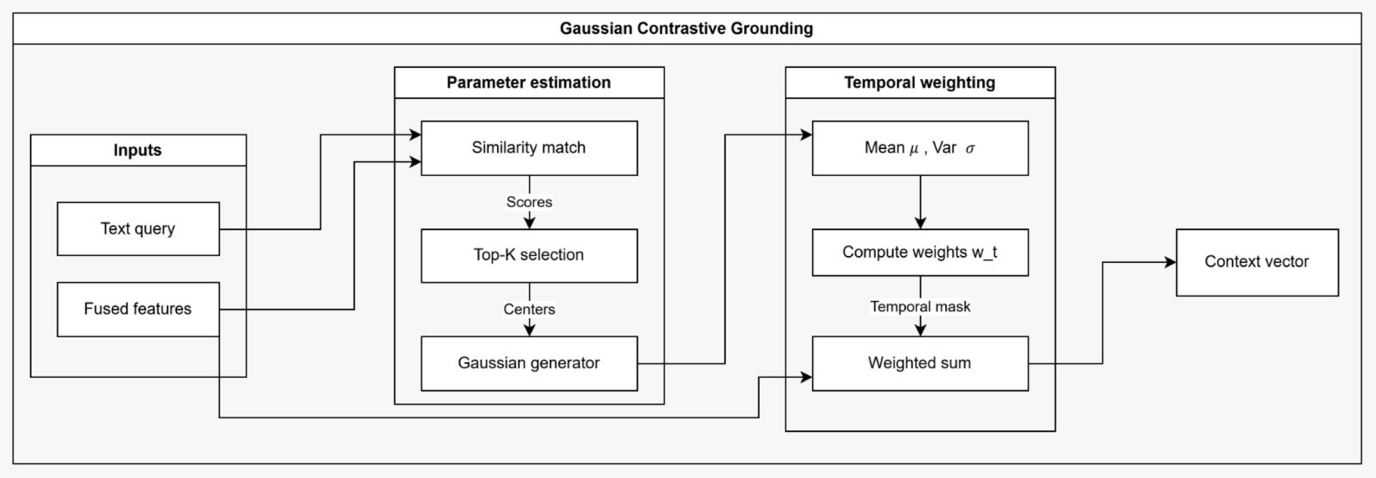


Fig. 2: Gaussian Contrastive Grounding (GCG) Mechanism.

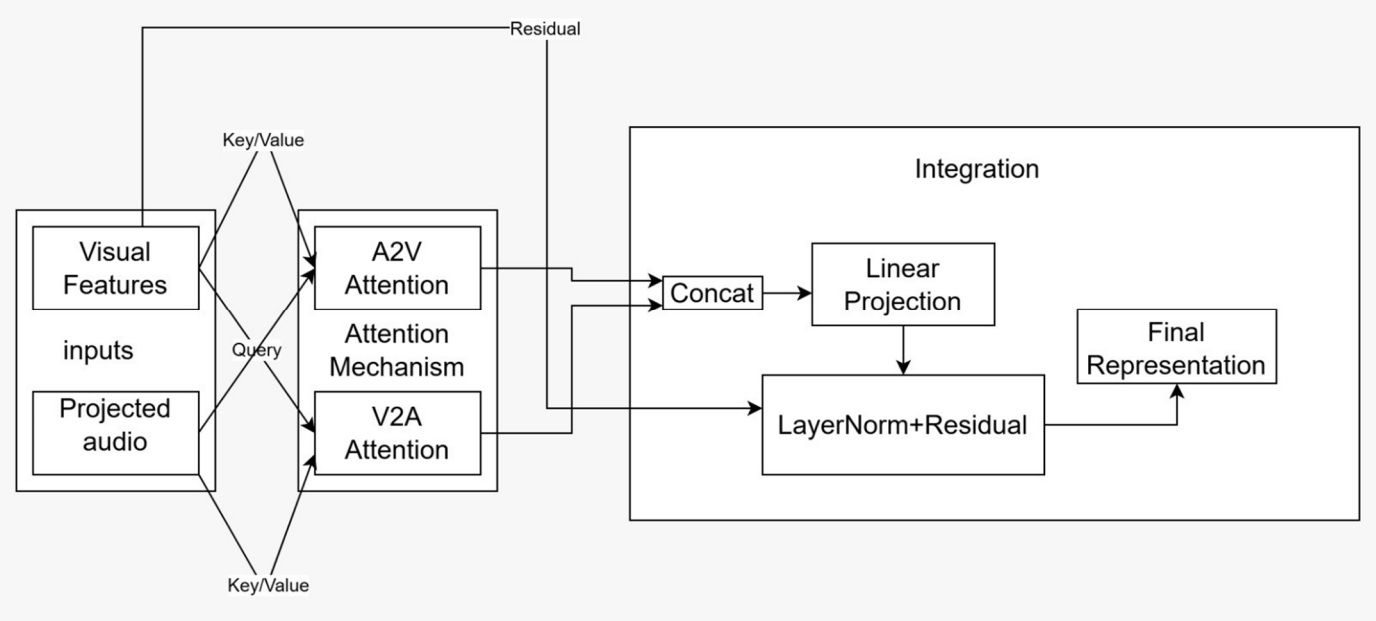


Fig. 3: Bi-Directional Audio-Visual Fusion Module.

3) Gaussian Mixture-Based Grounding

To capture multiple relevant temporal regions within a video, the framework extends single Gaussian grounding to a Gaus-sian Mixture Model (GMM). Instead of assuming a single peak of relevance, the temporal attention is modeled as a weighted combination of multiple Gaussian components:

$$w_t = \sum_{k=1}^{K} \pi_k \exp\left(-\frac{(t-\mu_k)^2}{2\sigma_k^2}\right) \qquad (8)$$

where K denotes the number of Gaussian components, μ_k and σ_k represent the mean and variance of each component, and π_k corresponds to the mixture weights. This formulation enables the model to localize multiple non-contiguous events and improves temporal grounding accuracy for complex activities.

4) Audio-Visual Fusion

The model integrates audio and visual features using bi-directional cross-attention mechanisms. In Vision-to-Audio (V2A) attention, visual features attend to audio representations to identify relevant acoustic cues associated with observed actions. Conversely, in Audio-to-Vision (A2V) attention, audio features attend to visual frames to ground sound events within the visual context. These complementary interactions enable the model to effectively capture both visual and auditory dependencies, resulting in a richer and more informative multimodal representation.

The fused representation is computed as:

$$f_{final} = \text{Fusion}(A_{v2a}, A_{a2v}) \quad (9)$$

5) Answer Generation

The final multimodal representation is processed by the Q-Former, which extracts a compact set of query-relevant tokens

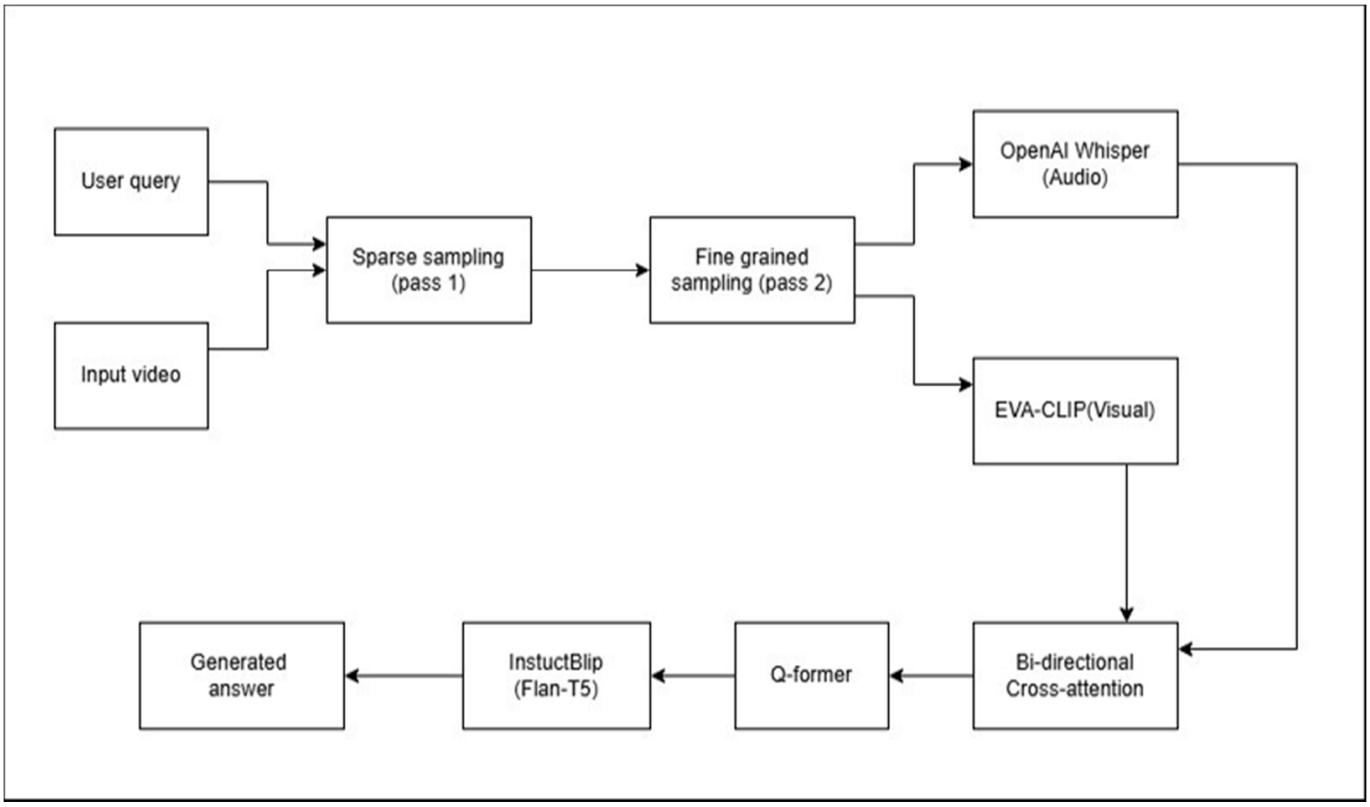


Fig. 4: Coarse-to-Fine Temporal Grounding Framework

from the grounded features. These tokens are combined with the encoded textual query and passed to the Flan-T5 language model through the InstructBLIP framework. The model then performs autoregressive decoding to generate the final natural language answer:

$$\text{Answer} = \text{LLM}(Q\text{-Former}(f_{\text{final}}, q)) \quad (10)$$

This step ensures that the generated response is both contex-tually accurate and grounded in the relevant temporal segments of the video.

IV. EXPERIMENTAL ANALYSIS

A comprehensive evaluation of the proposed framework, focusing on a detailed comparison between the Single-Pass Gaussian Contrastive Grounding (GCG) model and the Two-Pass Coarse-to-Fine (C2F) model. The analysis covers learning behaviour, multimodal fusion, temporal grounding, reasoning capability, efficiency trade-offs, and qualitative performance.

A. *Learning Behaviour of the Model*

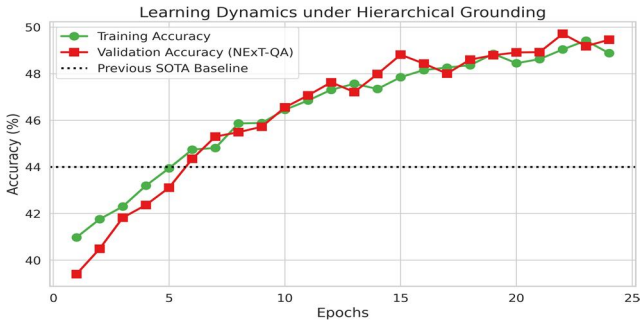


Fig. 5: Training and validation accuracy over epochs.

The proposed Two-Pass C2F model exhibits stable conver-gence and surpasses the Single-Pass baseline early in training. This indicates that the C2F grounding strategy enables more efficient learning by focusing on relevant temporal regions.

Both training and validation accuracy increase steadily dur-ing early epochs, demonstrating effective multimodal represen-tation learning. The small gap between training and validation accuracy suggests minimal overfitting.

Compared to the Single-Pass model, the Two-Pass C2F model converges faster and achieves higher final accuracy, highlighting the advantage of hierarchical grounding. The smooth learning curve further indicates stable optimization and well-tuned hyperparameters.

B. *Performance Across Question Types*

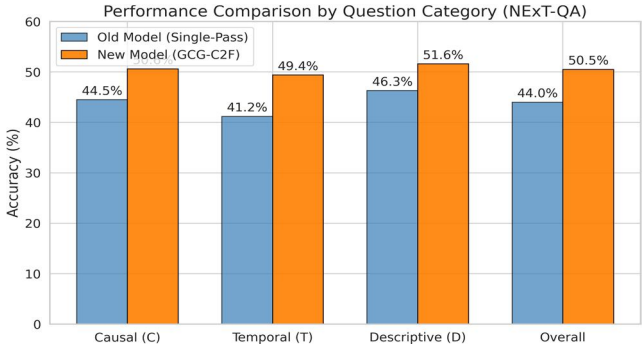


Fig. 6: Performance comparison across descriptive, temporal, and causal question types.

Both models perform well on descriptive questions; how-ever, the Single-Pass model struggles with temporal depen-dencies due to sparse frame sampling. In contrast, the Two-Pass C2F model refines relevant seg-ments, improving sequence understanding and causal infer-ence. This demonstrates the effectiveness of coarse-to-fine grounding in capturing event transitions and interactions.

C. *Multimodal Fusion Analysis*

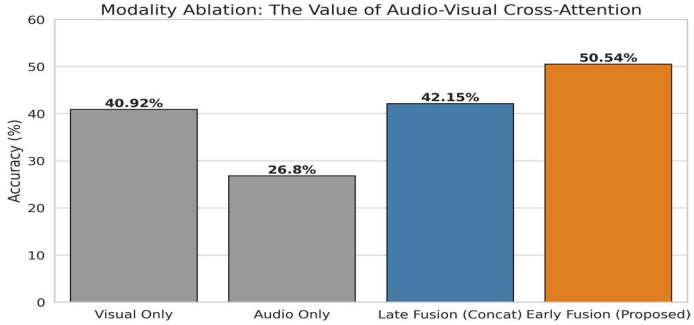


Fig. 7: Performance improvement due to audio-visual fusion.

The inclusion of audio significantly enhances performance in both models. The Single-Pass model shows moderate im-provement due to coarse alignment, whereas the Two-Pass C2F model achieves better fusion through refined temporal grounding.

TABLE I: Comparison of Multimodal Fusion

Aspect	Single-Pass GCG	Two-Pass C2F
Fusion Strategy	Single-stage	Multi-stage refinement
Alignment Quality	Coarse	Fine-grained
Off-screen Event Detection	Limited	Effective
Context Understanding	Moderate	Strong

D. Robustness to Video Duration

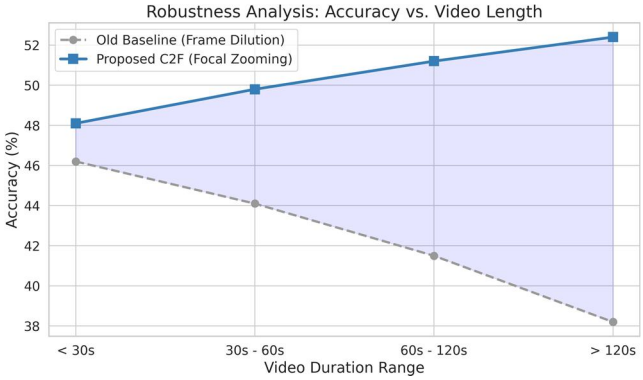


Fig. 8: Effect of video length on model performance.

The Single-Pass model suffers from the frame dilution problem as video length increases. In contrast, the Two-Pass C2F model maintains stable accuracy by focusing on relevant temporal segments.

E. Temporal Grounding Analysis

The Single-Pass model produces a single dominant attention peak, limiting its ability to capture multiple events. The Two-Pass C2F model identifies multiple relevant segments, improving temporal localization.

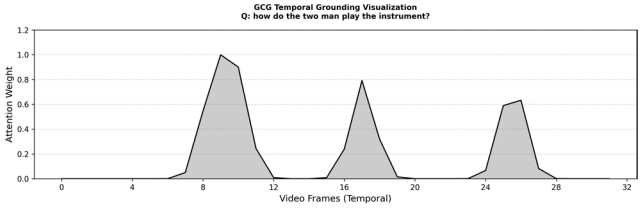


Fig. 9: GCG temporal grounding visualization.

TABLE II: Temporal Grounding Comparison

Feature	**Single-Pass GCG**	**Two-Pass C2F**
Grounding Method	Single Gaussian	Gaussian Mixture
Temporal Resolution	Low	High
Multi-event Detection	Limited	Strong
Localization Accuracy	Moderate	High

F. Reasoning Capability Analysis

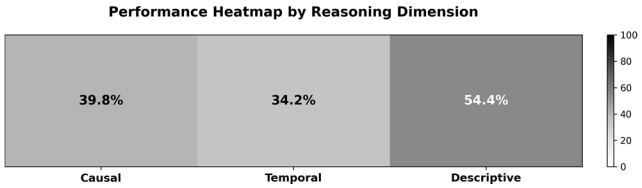


Fig. 10: Performance heatmap across reasoning dimensions.

The Single-Pass model shows limitations in temporal and causal reasoning, while the Two-Pass C2F model sig-nificantly improves performance through refined ground-ing.Recent works have also emphasized validating generated answers to ensure reliability. Visually grounded answer verifi-cation methods [25] assess whether predictions are supported by actual visual evidence, reducing hallucinations and improv-ing trustworthiness in VideoQA systems.

TABLE III: Reasoning Capability Comparison

Aspect	Single-Pass GCG	Two-Pass C2F
Descriptive Questions	Good	Excellent
Temporal Reasoning	Moderate	Strong
Causal Reasoning	Weak	Improved
Multi-step Reasoning	Limited	Effective
Answer Precision	Moderate	High

G. Ablation Study and Sampling Strategy

TABLE IV: Effect of Different Sampling Strategies

Strategy	Frames	Accuracy (%)
Single-Pass	32	44.00
Single-Pass	64	45.21
Coarse-to-Fine	**32 + 32**	**50.54**

The results show that increasing frames in the Single-Pass model provides only marginal gains. In contrast, the Two-Pass C2F model significantly improves accuracy by focusing on relevant segments rather than increasing global sampling.

H. Qualitative Comparison Between Models

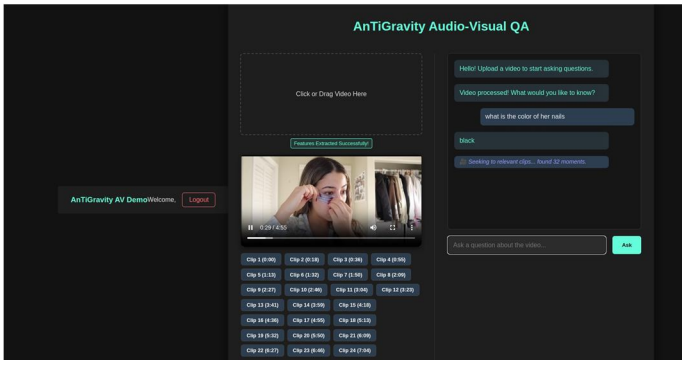


Fig. 11: Single-Pass model output interface.

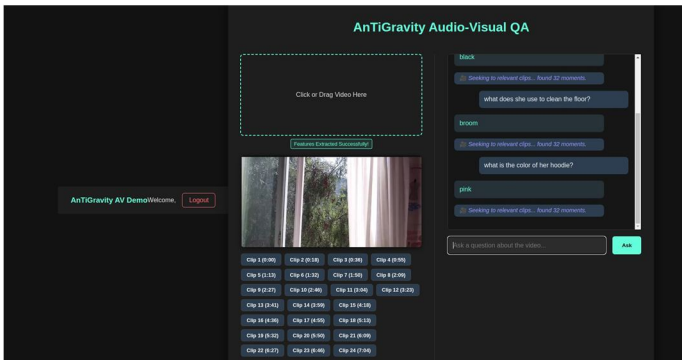


Fig. 12: Additional Single-Pass output.

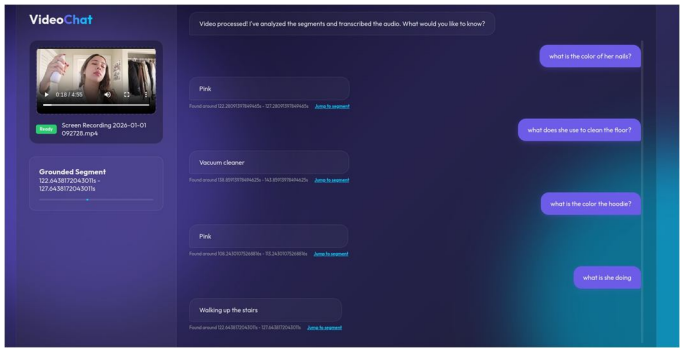


Fig. 13: Two-Pass model with grounded segments.

TABLE V: Qualitative Comparison of Model Responses

Query	Single-Pass	Two-Pass
Clean floor	broom	vaccum cleaner
Hoodie color	pink	pink
Nail color	black	pink
Action	Painting on cup	Walking up the stairs
Tree	Christmas tree	Christmas tree

I. Observations

- The Two-Pass C2F model demonstrates improved accu-racy and contextual understanding.
- It resolves ambiguities better, producing more precise action descriptions compared to the Single-Pass model.
- Fine-grained grounding enables improved detection of visual attributes.
- Both models perform similarly for simple queries, but diverge significantly on complex tasks.
- The Two-Pass C2F model shows superior temporal rea-soning and consistency.

J. Efficiency vs Performance Trade-off

TABLE VI: Efficiency vs Performance Comparison

Metric	Single-Pass	Two-Pass
Computation Cost	Low	Higher
Speed	Fast	Moderate
Accuracy	Moderate	High

V. CONCLUSION

The work presented a Temporal Question Answering frame-work for egocentric video understanding, with a detailed com-parative analysis between a Single-Pass Gaussian Contrastive Grounding (GCG) model and a Two-Pass Coarse-to-Fine (C2F) model. The study demonstrates that while the Single-Pass approach provides a computationally efficient baseline, it suffers from limitations in temporal precision, multimodal alignment, and contextual reasoning.

Through extensive experimental analysis, including quan-titative evaluation and qualitative comparisons, the Two-Pass C2F model consistently outperforms the Single-Pass model across multiple dimensions. The hierarchical grounding strat-egy enables more accurate localization of relevant temporal segments, effectively addressing the frame dilution problem in longer videos. Additionally, the integration of audio-visual features enhances the model's ability to capture both visible and non-visible events, leading to improved contextual under-standing.

The qualitative results further validate these findings, show-ing that the Two-Pass model produces more precise and context-aware responses, particularly for action-based and reasoning-intensive queries. The inclusion of explicit tempo-ral grounding also improves interpretability, allowing users to trace answers back to specific video segments, thereby increasing transparency and reliability.

Despite these improvements, the Two-Pass model introduces additional computational overhead compared to the Single-Pass approach. This highlights a trade-off between efficiency and performance, suggesting that model selection should be guided by application requirements.

The proposed framework provides a robust foundation for real-world video question answering systems, particularly in applications such as assistive technologies, intelligent video analysis, and memory augmentation systems. Future work will focus on optimizing the model for real-time deployment, improving reasoning capabilities, and extending the framework to incorporate additional modalities for richer contextual un-derstanding.

REFERENCES

[1] K. Grauman, A. Westbury, M. Chavis, and Others, "Ego4d: Around the world in 3,000 hours of egocentric video," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[2] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, "Next-qa: Explaining temporal actions via question answering," arXiv preprint, 2021.
[3] J. Lei, T. L. Berg, and M. Bansal, "Mart: Memory-attended recurrent transformers for long-term video question answering," in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2020.
[4] A. Radford, J. W. Kim, C. Hallacy, and Others, "Clip: Connecting text and images," OpenAI Technical Report, 2021.
[5] H. Touvron, L. Albert, T. Bresson, and Others, "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.
[6] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in Advances in Neural Information Processing Systems (NeurIPS), 2023.
[7] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Temporal sentence grounding in videos: A survey and future directions," arXiv preprint, 2023.
[8] H. Wang, C. Lai, Y. Sun, and W. Ge, "Weakly supervised gaussian contrastive grounding with large multimodal models for video question answering," in Proceedings of the ACM International Conference on Multimedia (ACM MM), Melbourne, Australia, 2024.
[9] Z. Yang, X. He, and J. Gao, "Stacked attention networks for visual question answering," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[10] J. Xu, T. Liu, C. Ou, and Others, "Activitynet-qa: A dataset for under-standing complex web videos via question answering," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[11] J. Gao, R. Ge, K. Chen, and R. Nevatia, "Motion-appearance co-memory networks for video question answering," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2018.
[12] Z. Fan, B. Jiang, and D. Lin, "Hipnet: Hierarchical interactive memory network for video question answering," in Proceedings of the European Conference on Computer Vision (ECCV), 2020.
[13] X. Li, Y. Song, and S. Fan, "Beyond rnns: Positional self-attention for video question answering," in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2019.
[14] W.-Y. Jin, S. Lee, Y. Cho, and Others, "Video-llama: An instruction-following video large language model," arXiv preprint, 2022.
[15] M. Peng, Y. Wu, A. Wang, and Others, "Progressive attention memory network for video question answering," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
[16] M. Tapaswi, Y. Zhu, and R. Stiefelhagen, "Movieqa: Understanding stories in movies through question answering," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[17] S. Yu, J. Cho, P. Yadav, and M. Bansal, "Self-chained image-language model for video localization and question answering," in Advances in Neural Information Processing Systems (NeurIPS), 2023.
[18] Y. Li, J. Xiao, C. Feng, X. Wang, and T.-S. Chua, "Discovering spatio-temporal rationales for video question answering," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
[19] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "Tvqa+: Spatio-temporal grounding for video question answering," in IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
[20] Z. Shen, L. Li, Z. Lin, and Others, "Weakly supervised dense event captioning in videos," in Advances in Neural Information Processing Systems (NeurIPS),

2017.

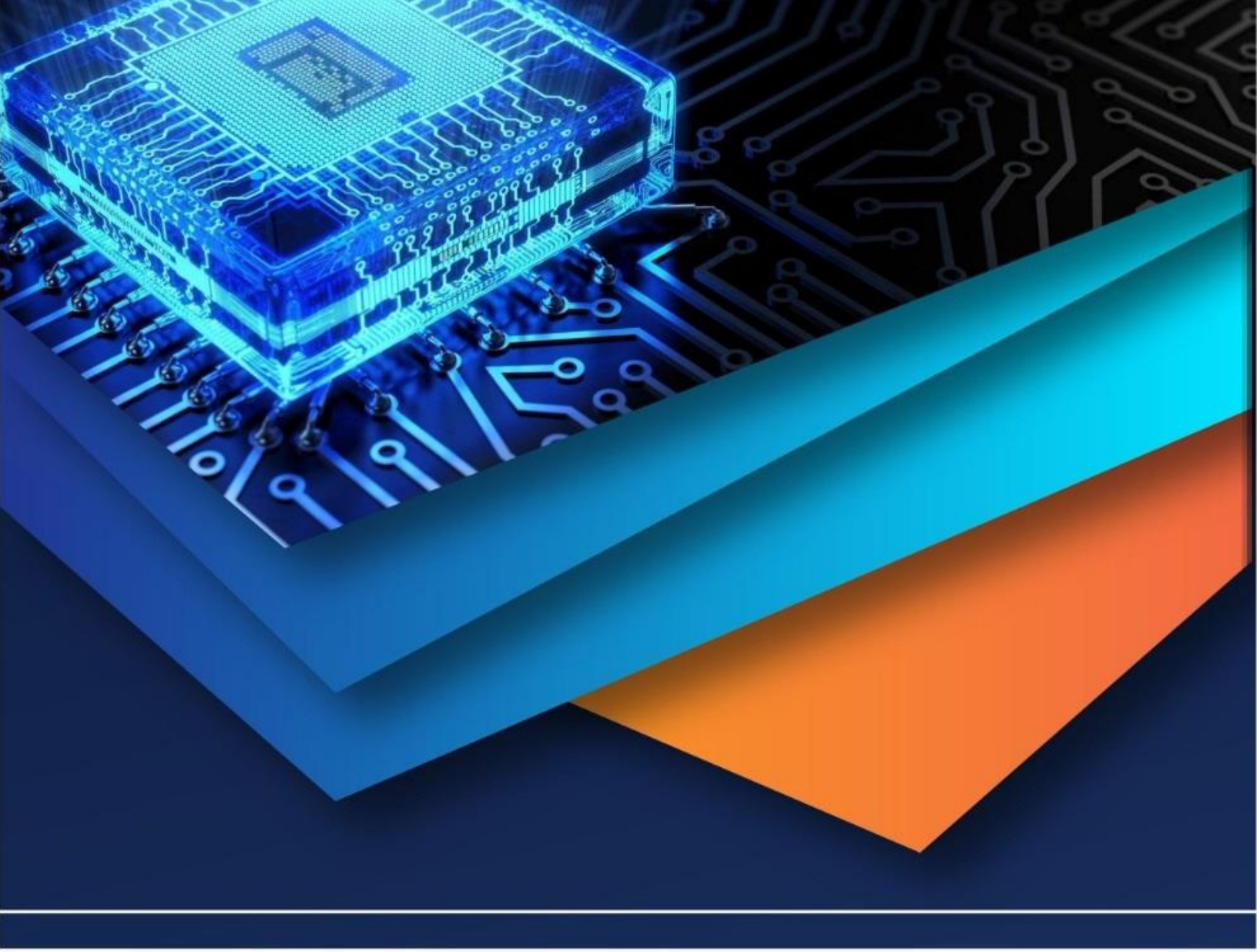
[21] P. Seo, A. Nagrani, A. Arnab, and Others, "Attentive moment retrieval in videos," in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2020.

[22] Y. Lei, H. Tan, and M. Bansal, "Symbolic replay for long video question answering," in Proceedings of the European Conference on Computer Vision (ECCV), 2022.

[23] J.-B. Alayrac, J. Donahue, P. Liu, and Others, "Flamingo: A visual lan-guage model for few-shot learning," in Advances in Neural Information Processing Systems (NeurIPS), 2023.

[24] Y. Fang, P. Sun, X. Chen, and Others, "Eva: Exploring the limits of masked visual representation learning at scale," arXiv preprint, 2023.

[25] J. Xiao, A. Yao, Y. Li, and T.-S. Chua, "Can i trust your answer? visually grounded video question answering," arXiv preprint arXiv:2309.01327, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY