Spatio-Temporal Input Densification for Efficient and Robust Open-Domain Video Question Answering

Bao Tran[1,2,†], Thuyen Tran Doan[1,2], Tien Do[1,2], Tien-Dung Mai[1,2], Thanh Duc Ngo[1,2], Duy-Dinh Le[1,2], and Shin'ichi Satoh[3]

[1] University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
[3] National Institute of Informatics, Japan
[†] 22520121@gm.uit.edu.vn
{thuyentd, tiendv, dungmt, thanhnd, duyld}@uit.edu.vn
satoh@nii.ac.jp

Abstract. Video Question Answering (VQA) requires systems to jointly reason over visual, auditory, and linguistic cues, and remains challenging due to complex temporal dependencies and the diverse, open-ended nature of real-world queries. Recent approaches often depend on supervised finetuning of large vision-language models, which yields strong in-distribution performance but comes with substantial data and computational demands. Furthermore, finetuned systems can struggle with temporal reasoning, small-object recognition, and effective use of audio information, limiting their robustness in open-domain benchmarks such as TRECVID. In this work, we introduce a VQA framework that enhances existing multimodal models without task-specific finetuning. Its core component is a spatio-temporal input densification strategy that reorganizes video evidence using dense frame sampling and spatial tiling, enabling finer visual understanding and more reliable temporal inference. The framework also incorporates lightweight modules for textualized audio integration, question-type–aware prompting, and output normalization, contributing to improved robustness and answer consistency. Despite requiring no task-specific finetuning, the proposed system achieves strong results on the TRECVID 2025 VQA test set. Our Multiple Choice submission attains a Top-1 Accuracy of 0.774 and an MRR of 0.859, ranking as the top-performing run. For Answer Generation, the system reaches METEOR 0.173, BERTScore 0.887, STS 0.270, and $\text{NDCG}_{\text{BERTScore}}$ 0.996, placing it among the highest-ranked submissions. These results demonstrate the effectiveness of inference-time input densification as a scalable alternative to supervised finetuning.

1 Introduction

The TRECVID 2025 Video Question Answering (VQA) [14] [8] task evaluates systems on their ability to understand and reason over multimodal video content in order to answer natural-language questions. The task consists of two subtasks: Answer Generation (AG), where systems must produce free-form textual answers, and MC, where systems select the correct answer from predefined options. Both subtasks require joint reasoning over visual frames, audio signals, and language, and many queries involve non-trivial temporal relations between events distributed across the video. As a result, VQA remains a challenging problem in multimodal video understanding.

Recent work on video question answering has mainly focused on supervised finetuning of pretrained vision-language models on large annotated datasets. Typical approaches first extract visual representations using convolutional networks or video transformers, and then fuse them with text using cross-modal attention. Early models such as spatio-temporal attention networks and Co-Memory architectures showed that attending to relevant frames is crucial for reasoning over dynamic scenes. More recent methods exploit large-scale video–language pretraining in models such as VIOLET [4], InternVideo [11], or multimodal large language models (MLLMs) [12] such as InternVL2/3 [3] and LLaVA-Video [13]. These models achieve strong in-distribution performance and significantly push forward the state of the art.

However, finetuned systems still face several limitations in open-world video reasoning benchmarks such as TRECVID. First, supervised finetuning is expensive in terms of both data and computation, and it usually requires tens of thousands of annotated question–answer pairs to generalize well [6] [7]. Second, models trained on curated datasets often generalize poorly to the diverse and noisy queries in TRECVID, which cover a wide range of topics and linguistic styles [1]. Third, finetuning does not fully solve temporal reasoning: models frequently fail on questions about event order, repeated actions, or long-range dependencies [2]. Finally, many pipelines underutilize non-visual information such as speech transcripts, scene text, or other metadata, even though these sources can provide strong semantic cues.

In this work, we propose a *training-free* Video Question Answering framework that aims to overcome these limitations by reusing large multimodal models without any task-specific finetuning. Our key idea is to improve

how videos are presented to large vision-language models, rather than modifying the models themselves. The main technical contribution is a *Spatio-Temporal Input Densification* scheme that reorganizes video inputs in both time and space before feeding them to the model. In the temporal dimension, we sample a denser set of frames to cover more stages of the video and capture short-lived events. In the spatial dimension, we tile each frame into local regions and submit these tiles as separate visual inputs, acting as a structured zoom-in mechanism for small or peripheral details. This densification strategy enhances temporal coverage and spatial granularity while keeping the underlying model architecture unchanged.

On top of this core module, we introduce several lightweight, training-free extensions that further improve robustness: (i) *textualized audio integration*, where speech is converted to text using Whisper and injected into the prompt as auxiliary context; (ii) *question-type–aware prompting*, which adapts the reasoning instructions to the type of question (e.g., temporal, counting, descriptive); and (iii) *answer normalization* for AG, which enforces concise and consistent answer formats aligned with the TRECVID metrics. We instantiate our framework on two strong backbone models. For the AG subtask, we use Aria3x8B [5] as a unified multimodal reasoner, combined with Whisper transcripts and answer normalization to produce short, evaluation-friendly responses. For the MC subtask, we build on InternVL3.5–Instruct [10], enhanced with our spatio-temporal densification and question-type–aware prompting.

The proposed system was evaluated on the official TRECVID 2025 VQA test set. For the Answer Generation subtask, our submission achieved a METEOR score of 0.173, a BERTScore of 0.887, and an STS score of 0.270. In terms of ranking-based evaluation, the system obtained $\mathrm{NDCG}_{\mathrm{METEOR}} = 0.825$ and $\mathrm{NDCG}_{\mathrm{BERTScore}} = 0.996$. For the Multiple Choice (MC) subtask, our submission attained a Top-1 Accuracy of 0.774 and an MRR of 0.859 on the test set. These results demonstrate that the proposed training-free framework is capable of producing competitive performance across both subtasks, despite operating without task-specific finetuning.

2 Training-Free Spatio-Temporal Input Densification for Video QA

This section describes our training-free Video Question Answering (VQA) framework for the TRECVID 2025 evaluation. The system is designed to support both AG and MC without any task-specific finetuning. Instead of updating model parameters, we adapt large multimodal models through: (i) a spatio-temporal input densification scheme that serves as the core contribution of this work; and (ii) auxiliary, training-free components for audio integration, prompting, and output normalization, as shown in Figure 1.

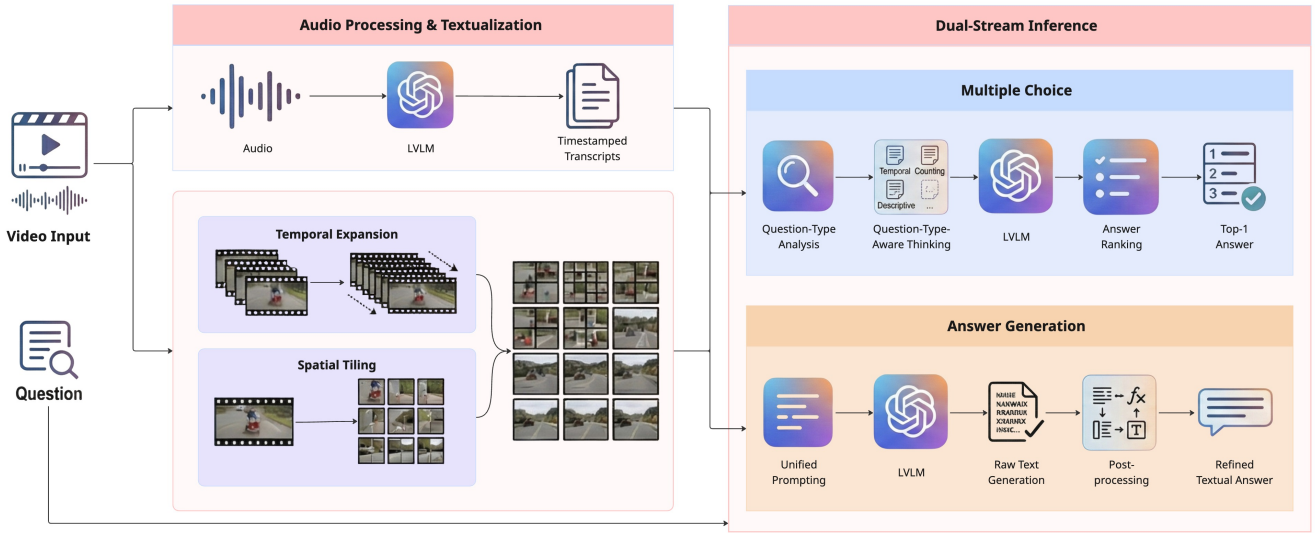

Fig. 1: Overview of the training-free Video Question Answering (VQA) framework. The system integrates a spatio-temporal input densification scheme, auxiliary components for audio integration, prompting, and output normalization, enabling efficient AG and MC tasks without task-specific finetuning.

2.1 Overall Training-Free VQA Framework

Given a video clip and a natural-language question, our system operates as follows. First, we select a set of frames from the video and apply spatio-temporal input densification to construct a fine-grained visual evidence set. In par-

allel, we extract the audio track and transcribe it into text using Whisper, obtaining timestamp-aligned transcripts that capture spoken content.

For the AG subtask, the question, densified visual inputs, and transcripts are passed to the Aria3x8B model through a carefully designed prompt. The model produces an open-ended textual answer, which is then refined by an answer normalization module to match the style and length expected by the TRECVID evaluation metrics.

For the MC subtask, we use InternVL3.5–Instruct as the backbone large vision-language model (LVLM). The model receives the densified visual evidence, textualized audio, the question, and the candidate options in a structured prompt that explicitly asks it to rank the options. A question-type–aware prompting strategy adjusts the instructions depending on whether the question is temporal, counting-based, yes/no, or descriptive. In both subtasks, the LVLM weights remain frozen; all adaptation is achieved by modifying the inputs and prompts.

2.2 Spatio-Temporal Input Densification

Standard LVLMs typically treat videos as a small set of global frames. This leads to two main limitations for VQA: (i) small or distant objects can be overlooked when the model sees only low-resolution global views; and (ii) temporal reasoning is weak when only a few frames are sampled. These issues are especially problematic in TRECVID, where questions often depend on subtle visual cues or on the order and repetition of events.

We address these limitations with a *Spatio-Temporal Input Densification* scheme that increases both temporal coverage and spatial granularity, while leaving the internal model unchanged. Instead of modifying the network architecture, we change how the video is decomposed into inputs.

2.2.1 Temporal Expansion: In the temporal dimension, we sample a larger number of frames that are roughly uniformly distributed over the video. In practice, we use twelve frames per video in our main configuration. This expanded sampling captures more intermediate states of the scene, including transitions and short-lived events that would be missed by a sparse strategy.

Although the LVLM does not explicitly model long frame sequences, exposing it to multiple time points in a single inference call still provides a useful approximation of the underlying dynamics. The model can observe how objects and agents appear, move, and disappear across frames, which is beneficial for questions about action order, repeated behaviors, and the presence or absence of entities at different times.

2.2.2 Spatial Tiling as Structured Zoom-In: In the spatial dimension, each selected frame is partitioned into a grid of tiles (for example, 2×2 or 3×3). Each tile is then sent to the LVLM as an individual visual input. This tiling acts as a structured zoom-in mechanism: instead of relying on the model to attend to small regions within a single global frame, we explicitly crop and enlarge local regions for the model.

This strategy improves the model's sensitivity to small or localized details such as handheld objects, facial expressions, or background elements near the borders of the frame. It is particularly important for the MC subtask, where candidate options may differ only in a small visual attribute or in the presence of a specific object. By providing tiles separately, we effectively increase the resolution of the model's perception without changing the backbone architecture.

Figure 2 illustrates the concept of spatial tiling, where each frame is divided into smaller tiles to enhance the model's ability to focus on fine-grained visual details. This approach significantly contributes to improved performance, particularly in tasks like Multiple Choice.

2.3 Textualized Audio as Auxiliary Tri-Modal Context

A considerable portion of TRECVID questions reference spoken content, conversational cues, or other auditory events that are not visually observable. Purely visual models are therefore unable to access information conveyed through the audio stream. Conventional multimodal systems address this issue by incorporating dedicated audio encoders and audio-visual fusion modules, but such architectures typically require substantial training data and introduce non-trivial model complexity.

To avoid these limitations, we adopt a lightweight, training-free strategy that converts audio into natural-language text and incorporates it directly into the model's linguistic input. Specifically, we use Whisper [9] to transcribe the audio track of each video into timestamped textual segments. These transcripts are injected into the prompt alongside the question and, when available, a brief scene description.

This *textualized audio* design leverages the strong language reasoning capabilities of modern LVLMs. By representing audio as text, the model can process auditory information through its most reliable modality without

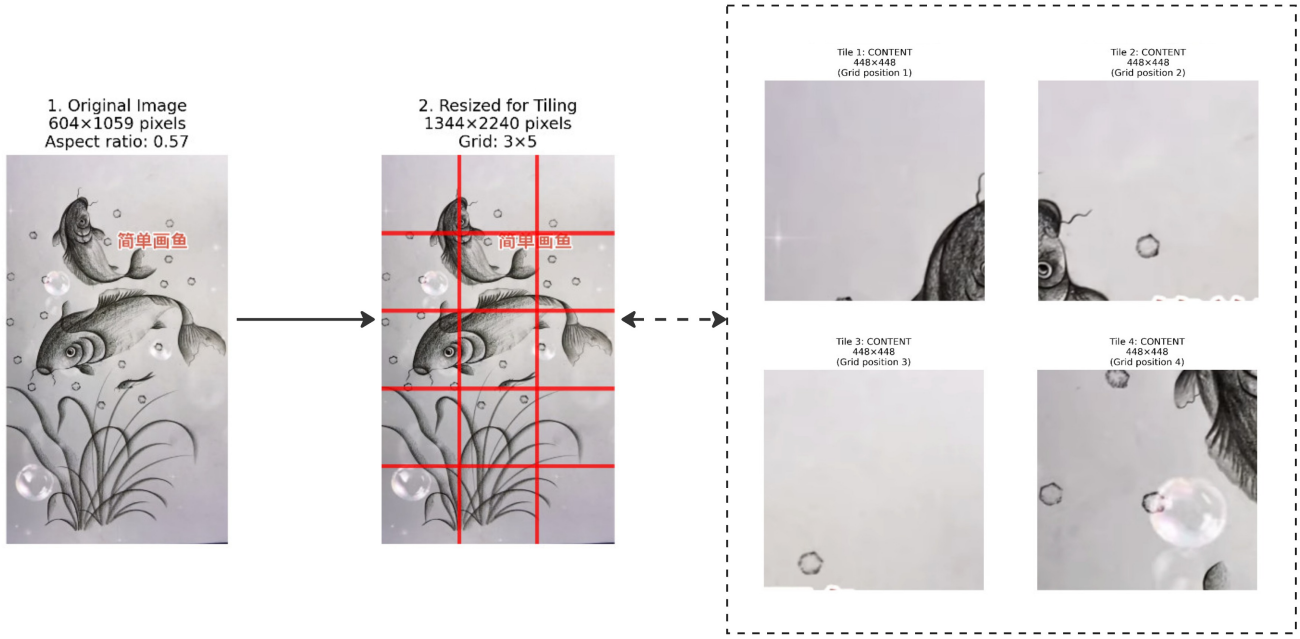

Fig. 2: Illustration of spatial tiling, where each frame is divided into smaller tiles to improve the model's focus on fine-grained visual details, enhancing performance in tasks like Multiple Choice.

requiring any architectural changes. In the AG subtask, transcripts provide additional semantic cues such as names, spoken descriptions, or non-visual events. In the MC subtask, they supply discriminative evidence that helps differentiate between visually similar answer options. As with our spatio-temporal densification module, this component is fully model-agnostic and operates without any finetuning.

2.4 Question-Type–Aware Prompting and Answer Normalization

Large multimodal models do not employ a uniform reasoning strategy across different types of questions. Temporal and counting queries typically require multi-step inference, whereas descriptive or yes/no questions are more effectively answered with brief, direct responses. Consequently, a single fixed prompt can be suboptimal when applied to the heterogeneous query set found in TRECVID.

To address this challenge, we introduce a *question-type-aware thinking* mechanism. Prior to inference, each question is assigned to one of several coarse categories-temporal, counting, yes/no, or descriptive. Each category is associated with a tailored prompting profile. For temporal and counting questions, the prompt encourages step-by-step reasoning and explicitly instructs the model to consider the sequence or magnitude of events. In contrast, prompts for yes/no and descriptive questions emphasize conciseness and discourage unnecessary elaboration.

This targeted prompting strategy serves as a lightweight alternative to task-specific finetuning: instead of modifying model parameters, we adjust the inference behavior through structured natural-language instructions. When combined with the proposed spatio-temporal densification, it enables the model to more effectively leverage the enriched visual evidence, particularly for queries that require detailed temporal analysis.

For the Answer Generation subtask, we further apply a simple answer normalization module to the raw outputs of Aria3x8B. This post-processing step removes redundant wording, standardizes numerical and enforces short, declarative responses that align better with TRECVid's reference style. Although training-free, this normalization consistently improves output consistency and evaluation scores.

3 Experiments results

This section presents a comprehensive empirical study of the proposed VQA system on the official TRECVid VQA training dataset, which serves as the sole source of ground-truth supervision available during development. Because TRECVid does not release labels for evaluation queries prior to the workshop, all analyses reported here rely exclusively on internal validation using the training set. Our experiments aim to quantify performance on both subtasks, AG and MC, and to assess the incremental benefits of each component within the training-free pipeline.

3.1 Datasets

All experiments were conducted using the TRECVid 2025 VQA training dataset, which contains **2000** short video clips. Each video is paired with one of two types of questions: (1) open-ended natural-language questions with textual answers, or (2) multiple-choice questions with four candidate answers. The dataset covers a wide variety of scenes, human activities, interactions, and object configurations. It includes videos with both meaningful audio, such as dialogue and environmental sounds, as well as videos with minimal sound content.

Although primarily intended for model training, the dataset provides a sufficiently challenging benchmark due to its complexity. The videos exhibit large visual and semantic variability, requiring the model to process diverse content. Additionally, many questions necessitate reasoning over multi-frame dynamics, especially for temporal queries. The dataset also includes questions that depend on audio or speech content, requiring models to utilize both visual and auditory information. Moreover, the open-ended answer formats demand semantic alignment with the reference answers, rather than relying on literal matching.

These characteristics make the dataset an appropriate and challenging benchmark for evaluating training-free multimodal reasoning approaches, such as the one proposed in this work. Table 1 provides representative examples from the dataset, highlighting the diversity of video content and question types included in the benchmark.

Table 1: Representative examples from the TRECVid 2025 VQA dataset, illustrating the diversity of video content and question types.

Video Source	**Question**	**Correct Answer**
B24dV_uori4	What causes shampoo to lather best when washing hair?	Washing your hair twice gives you cleaner hair and better lather.
BADi9YBwUNc	Why do the men keep riding their motorcycles through water?	The motorcycles are made for water.
BDKkVYmWqdM	What does a bicyclist do before landing his bike on a tennis court?	The bicyclist drives his bike across a tennis net.
Bdyiy2P1WrA	What was being cooked during the food demonstration?	A woman was demonstrating how to cook sausages and eggs.
BfcR9WtJQBQ	What goes in the pan after a layer of chocolate?	A layer of marshmallow goes on top.
BGdwSxRiXzw	What are the two people doing in this video?	The man and the woman are flirting while engaging in.
BkdaXjkp_h0	What happens at the end of the video?	A man holds up a dress and a woman describes it.
BLQQH7JQ2cE	What happens at the end of the video?	The man sitting on the sofa gets up and leaves.
BrGdWP7lia0	What happens when a man is playing football with a young boy?	The boy runs between the man's legs instead of kicking the ball.

3.2 Evaluation Metrics

3.2.1 Answer Generation: In alignment with TRECVid standards, we evaluate performance on the AG task using several complementary metrics. The **Semantic Textual Similarity (STS)** metric assesses the degree of semantic alignment between the generated answers and the ground-truth responses. **BERTScore** is employed to measure embedding-based similarity by comparing contextualized token representations between the system's output and the reference answers. **METEOR** evaluates the quality of phrasing by considering synonym matching, recall, and word-order structure, ensuring both linguistic accuracy and fluency in the generated responses. Additionally, the **Normalized Discounted Cumulative Gain (NDCG)** metric, which is the official TRECVid metric, is used to evaluate the quality of ranked answer lists, giving higher weight to higher-ranked answers and thus measuring ranking fidelity.

3.2.2 Multiple Choice: For the MC task, we employ standard evaluation metrics to assess model performance. **Top-1 Accuracy** measures the percentage of questions for which the highest-ranked candidate matches the ground truth. To further evaluate the quality of the ranking, we use the **Mean Reciprocal Rank (MRR)**, which emphasizes the position of the correct answer in the ranking, with higher importance placed on earlier correct predictions. This metric provides a comprehensive assessment of the model's ability to correctly prioritize candidates.

3.3 Module-Level Experimental Results

To better understand the contribution of each component within our training-free framework, we conducted a series of controlled ablation studies across both subtasks of the TRECVid VQA 2025. Instead of treating the system as a

monolithic entity, we isolated the effects of four key modules introduced in our methodology: (i) textualized audio integration, (ii) spatio-temporal input densification, (iii) question-type-aware thinking, and (iv) structured output normalization for answer generation. Unless otherwise specified, all MC experiments utilized the InternVL3.5–38B-Instruct backbone model, while AG experiments employed the Aria-8×3.5B model.

3.3.1 Multiple-Choice: Table 2 presents the incremental improvements observed with the addition of each module. The baseline performance of the InternVL3.5 model yields a Top-1 Accuracy of 0.746 and an MRR of 0.850, demonstrating strong zero-shot capabilities but also revealing certain weaknesses in speech-dependent queries and temporally structured questions.

Table 2: Ablation study on the TRECVid 2025 training split for the Multiple-Choice task. Each row cumulatively adds one proposed module to the InternVL3.5–38B-Instruct baseline.

Configuration	Top-1 Accuracy	MRR
InternVL3.5–38B-Instruct	0.746	0.850
+ Textualized Audio Integration	0.751	0.850
+ Temporal Densification (12 frames)	0.761	0.860
+ Spatial Tiling	0.779	0.866
+ Question-Type–Aware Prompting	**0.795**	**0.876**

Textualized audio integration via Whisper transcripts results in a slight improvement in accuracy to 0.751. This indicates that converting audio content into text allows the LVLM to effectively utilize spoken information without requiring architectural modifications.

Temporal densification achieved by sampling twelve uniformly spaced frames per video, further boosts the MRR to 0.860. This enhancement underscores the model's increased sensitivity to event transitions and repeated actions, which are critical for answering temporally dependent queries.

Spatial tiling, which divides each frame into smaller regions, improves Top-1 Accuracy to 0.779. This result highlights how structured zoom-in representations help the model focus on small, peripheral, or occluded objects, which are essential for fine-grained discrimination in multiple-choice questions.

Question-type-aware thinking is the most significant improvement, as it adapts the reasoning strategy based on the linguistic structure of each query. With this approach, Top-1 Accuracy increases to 0.795 and MRR improves to 0.876. By dynamically adjusting the reasoning depth, encouraging step-wise logic for temporal and counting queries, and concise responses for yes/no or descriptive queries, the model achieves more stable and accurate performance.

In addition to the multiple-choice improvements, we also evaluate the model's performance by category, comparing the accuracy with and without thinking. Table 3 presents the accuracy across various question types. The table demonstrates how the model's performance improves with thinking, particularly in categories such as `summary_description` and `miscellaneous_other`, where accuracy jumps significantly from 0.8235 to 0.9333 and 0.7824 to 0.8908, respectively.

The cumulative effects of these modules are reflected in the final performance, with the Top-1 Accuracy and MRR reaching their highest values in the final submitted configuration.

3.3.2 Answer Generation: A parallel ablation study of the AG pipeline is presented in Table 4. The baseline configuration uses the Aria-8×3.5B, achieving an STS of 0.507, a BERTScore of 0.899, and a METEOR score of 0.187.

Textualized audio integration through Whisper yields a modest but consistent improvement in semantic alignment, with STS rising to 0.508. This improvement reflects better grounding in narrated content and spoken cues, enhancing the model's ability to process audio input in the form of text.

High-resolution visual inputs, where the original image size is used instead of the resized 480px version, further boosts the performance, with STS increasing to 0.521. This result suggests that Aria benefits from finer spatial resolution when generating open-ended descriptions. However, the METEOR score shows a slight decline from 0.186 to 0.181, indicating that higher visual fidelity sometimes leads to more specific lexical choices that do not perfectly match the reference phrasings, despite being semantically accurate.

Output normalization and post-processing lead to improvements across all metrics, with STS rising to 0.537 and BERTScore increasing to 0.901. This enhancement is achieved by eliminating redundant phrasing, regularizing numeric expressions, and aligning the output format with TRECVid's expected standards. This component acts as a lightweight analogue to finetuning, ensuring consistent surface forms in a training-free setting.

Table 3: Accuracy Comparison with and without Thinking by Category

Category	Accuracy		Correct	
	W/ Thinking	W/o Thinking	W/ Thinking	W/o Thinking
miscellaneous_other	**0.8908**	0.8487	**106**	101
event_state_change	**0.8941**	0.8235	**76**	70
counting_quantity	**0.6119**	0.5075	**41**	34
action_activity	0.7258	**0.7581**	45	**47**
reason_why	0.7188	**0.7500**	23	**24**
temporal_time-specific	**0.8235**	0.7647	**14**	13
summary_description	0.9333	0.9333	14	14
yes_no	**0.6667**	0.5833	**8**	7
attribute_property	**0.9091**	0.8182	**10**	9
identity_which	0.5556	**0.6667**	5	**6**
audio_speech	**0.7143**	0.5714	**5**	4
inference_lesson	0.8000	0.8000	4	4
location	1.0000	1.0000	3	3

Table 4: Ablation study on the TRECVid 2025 training split for the Answer Generation task. Modules are added cumulatively to the Aria-8×3.5B baseline.

Configuration	STS	BERTScore	METEOR
Aria-8×3.5B	0.507	0.899	0.187
+ Textualized Audio	0.508	0.899	0.186
+ High-Resolution Visual Input	0.521	0.899	0.181
+ Post-Processing / Normalization	**0.537**	**0.901**	**0.182**

3.4 Submission Results

3.4.1 Answer Generation: For the AG subtask, the NII–UIT team submitted the highest-performing pipeline based on a training-free generation framework. The system produces a primary open-ended answer along with several paraphrastic variants. It is important to note that while stylistic variations, though semantically accurate, slightly reduced the overall score due to the official evaluation process, which averages similarity across all submitted hypotheses.

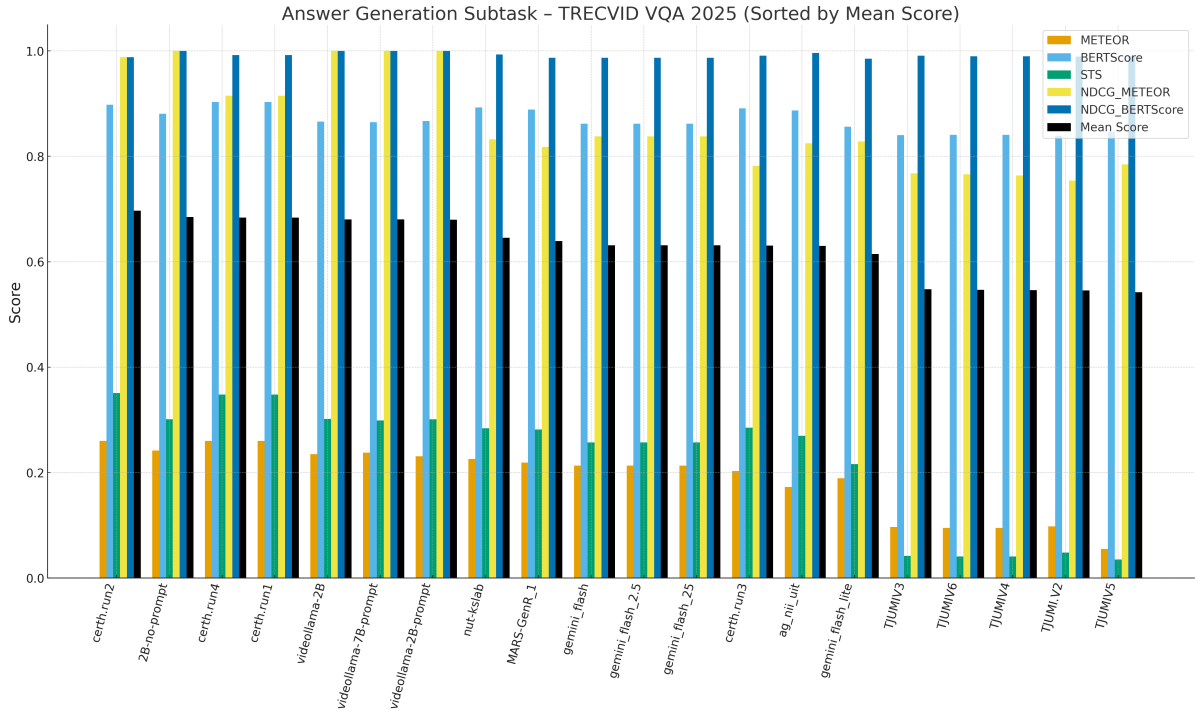

Fig. 3: Performance comparison for the TRECVID 2025 Answer Generation subtask.

Table 5: Multiple Choice VQA Results on the TRECVID 2025 Test Set

Run	Top-1 Accuracy	MRR	Mean Score
HIGHEST_PIPELINE (NII-UIT)	**0.774**	**0.859**	**0.8165**
certh.vqa.mc.run_2	0.761	0.855	0.8080
certh.vqa.mc.run_1	0.742	0.843	0.7925
certh.vqa.mc.run_4	0.559	0.727	0.6430
certh.vqa.mc.run_3	0.530	0.708	0.6190
videollama-7B-vqa	0.522	0.700	0.6110
tv25-kslab-mc	0.499	0.686	0.5925
TjuMI_MC_VQA_V4	0.007	0.009	0.0080
tjumi-v1	0.004	0.007	0.0055

The results of our submission to the TRECVid 2025 evaluation for the AG subtask are as follows:

$$\text{METEOR} = 0.173, \quad \text{BERTScore} = 0.887, \quad \text{STS} = 0.270$$

Additionally, the ranking-based metrics obtained were:

$$\text{NDCG}_{\text{METEOR}} = 0.825, \quad \text{NDCG}_{\text{BERTScore}} = 0.996$$

Post-hoc analysis of the predictions indicates that the top-1 predictions alone exhibit a significantly higher similarity to the reference answers. Thus, the primary limitation appears not in the model's generative capability but in the averaging-based evaluation protocol, which can dilute the quality of multiple hypotheses. Future submissions will focus on a single-answer strategy, which is expected to yield more consistent and competitive scores.

As shown in Figure 3, the performance comparison for the TRECVID 2025 Answer Generation subtask highlights the $METEOR$, $BERTScore$, STS, $NDCG_{METEOR}$, and $NDCG_{BERTScore}$ metrics for each submission. The systems are ranked by their mean score, which is displayed as a separate black bar to highlight overall trends.

3.4.2 Multiple Choice: For the MC subtask, our training-free pipeline achieved the highest performance among all participating teams. The system integrated the InternVL3.5 model with Whisper-based audio transcripts, spatio-temporal input densification, and question-type-aware thinking. The final evaluation results for this subtask were:

$$\text{Top-1 Accuracy} = 0.774, \quad \text{MRR} = 0.859$$

These results demonstrate the power of inference-time strategies, which allow the system to perform robust multimodal reasoning and accurately discriminate fine-grained visual and auditory details, even without task-specific finetuning.

Table 5 illustrates the evaluation results for the TRECVID 2025 Multiple-Choice subtask, where the mean score across Top-1 Accuracy and MRR sorts systems. The NII-UIT run (HIGHEST_PIPELINE) outperforms all other submissions, achieving the highest ranking metrics and accuracy.

4 Conclusion

In this paper, we proposed a training-free VQA framework that enhances large multimodal models without task-specific finetuning. Our approach utilizes Spatio-Temporal Input Densification for improved visual and temporal reasoning, and incorporates lightweight extensions like textualized audio integration, question-type-aware thinking, and answer normalization. We evaluated our method on the TRECVID 2025 VQA benchmark, achieving strong results with a Top-1 Accuracy of 0.774 and an MRR of 0.859 for the Multiple Choice subtask, and placing among the top submissions in Answer Generation with METEOR 0.173, BERTScore 0.887, and $\text{NDCG}_{\text{BERTScore}}$ 0.996. These results demonstrate that our framework offers a scalable, efficient alternative to traditional finetuning approaches, with potential for broader applications in multimodal video understanding.

References

1. Awad, G., Curtis, K., Butt, A.A., Fiscus, J., Godil, A., Lee, Y., Delgado, A., Godard, E., Diduch, L., Graham, Y., et al.: Trecvid 2023-a series of evaluation tracks in video understanding. In: Proceedings of TRECVID. vol. 4 (2023)

2. Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., Niebles, J.C.: Revisiting the" video" in video-language understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2917–2927 (2022)
3. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
4. Fu, T.J., Li, L., Gan, Z., Lin, K., Wang, W.Y., Wang, L., Liu, Z.: Violet: End-to-end video-language transformers with masked visual-token modeling. arXiv preprint arXiv:2111.12681 (2021)
5. Li, D., Liu, Y., Wu, H., Wang, Y., Shen, Z., Qu, B., Niu, X., Zhou, F., Huang, C., Li, Y., et al.: Aria: An open multimodal native mixture-of-experts model. arXiv preprint arXiv:2410.05993 (2024)
6. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
7. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 5971–5984 (2024)
8. P.J., J., Kovoor, B.C.: Video question answering: A survey of the state-of-the-art. Journal of Visual Communication and Image Representation **105**, 104320 (2024). https://doi.org/https://doi.org/10.1016/j.jvcir.2024.104320, `https://www.sciencedirect.com/science/article/pii/S1047320324002761`
9. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International conference on machine learning. pp. 28492–28518. PMLR (2023)
10. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
11. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
12. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
13. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
14. Zhong, Y., Ji, W., Xiao, J., Li, Y., Deng, W., Chua, T.S.: Video question answering: Datasets, algorithms and challenges. In: Proceedings of the 2022 conference on empirical methods in natural language processing. pp. 6439–6455 (2022)