Laboratory for Analytic Sciences in TREC 2025 Ad-hoc Video Search

Edward Sheriff[1], John Nolan[2], Yue Wang[3], Xi Niu[4],
[1]Department of Defense, USA
[2]North Carolina State University
[3]University of North Carolina at Chapel Hill
[4]University of North Carolina at Charlotte
{esherif, jrnolan2}@ncsu.edu, wangyue@unc.edu, xniu2@charlotte.edu

Abstract

This paper describes the Laboratory for Analytic Sciences (LAS) participation in the 2025 TREC Ad-hoc Video Search (AVS) task on the V3C2 collection. Motivated by deployment settings with constrained bandwidth and compute, our systems use a scalable keyframe-based text-to-video retrieval pipeline with dense 1 fps indexing and cosine-similarity search. We profile contrastive vision–language embedding models to select an efficient visual backbone for large-scale indexing over 4.5M keyframes. At query time, we evaluate training-free enhancements aimed at improving recall and ranking, including LLM-based semantic expansion, modality-aware query decomposition, Vision–Language Model (VLM)–based relevance scoring and reranking, and selectively triggered CLAP audio fusion for topics implying non-speech sounds.

Our results emphasize the role of Vision–Language Models (VLMs), including large multimodal generative models such as `gpt-4.1-mini` and `phi-3.5-vision`, as relevance judges. VLM scores align closely with human judgments on prior topics and provide a useful reranking signal, though some mismatches persist. Disagreements are largely attributable to lexical ambiguity, subjective topic phrasing, and temporal uncertainty from single-keyframe evidence. Across six official runs spanning four workflows, VLM reranking yields the largest gains, with VLM quality emerging as the most influential variable: when operating over identical candidate pools, `gpt-4.1-mini` improves performance on 19 of 20 topics relative to `phi-3.5-vision`. Semantic expansion followed by `gpt-4.1-mini` reranking achieves our best score of 0.399 mean infAP.

1 Introduction

As video content continues to proliferate across platforms and edge devices, scalable, efficient, and accurate video retrieval has become increasingly important. Text-to-video retrieval (T2VR) addresses this need by matching natural-language queries to relevant video segments. Despite recent advances in contrastive vision–language models, bridging the modality gap between text and video remains challenging under practical resource constraints [1, 2]. This is particularly acute in constrained edge networks where bandwidth, storage, and compute limitations make full-resolution processing infeasible, forcing systems to filter and index content early in the pipeline. Keyframe-based retrieval is a common strategy in this regime [3], yet it exposes persistent bottlenecks in embedding throughput and retrieval quality [4], making model selection and system-level speed–accuracy trade-offs central to deployment.

In the context of the TREC Ad-hoc Video Search (AVS) task, we: (1) benchmark contrastive text–image embedding models from the CLIP [5], SigLIP [6], and SigLIP2 [7] families on V3C2 keyframes to characterize hardware-dependent speed–accuracy trade-offs; (2) evaluate lightweight query-time enhancements, including LLM-based semantic expansion and modality-aware query decomposition; (3) assess VLM-based relevance scoring as both a scalable internal proxy for human judgments and a reranking stage; and (4) integrate these components into end-to-end AVS pipelines, submitting six official runs to TREC AVS 2025. Alongside these primary workflows, we also submitted a fine-grained CLIP-derived embedding variant (developed in a parallel study) to enable comparison across visual encoding approaches.

This paper presents our AVS 2025 submissions and analyzes the behavior of key retrieval components within a shared pipeline. We focus on hardware-aware speed–accuracy benchmarking, VLM-based relevance scoring, and selective CLAP audio fusion, highlighting their observed effects and failure modes in end-to-end AVS retrieval based on prior judged topic sets under a common corpus and topic setting.

2 Task, Data, and Evaluation

2.1 TREC 2025 Ad-hoc Video Search Task

This paper describes our participation in the 2025 TREC Ad-hoc Video Search (AVS) task. In AVS, systems receive short natural-language queries and must return, for each query, a ranked list of up to 1,000 video segments from a large corpus. For the 2025 cycle, we submitted six official runs that instantiate different retrieval workflows, exploring combinations of LLM-based semantic expansion, modality-aware query decomposition, VLM-based reranking, and selective CLAP-based audio retrieval. These variations evaluate the impact of lightweight query-time enhancements and multimodal cues under a shared corpus and evaluation setting. Runs follow TREC submission conventions and are scored by NIST on the 2025 topic set using withheld human judgments.

2.2 Dataset and Queries

Our submissions use the V3C (Vimeo Creative Commons Collection) dataset [8], the official corpus for TREC AVS. The full V3C collection contains three subsets: V3C1, V3C2, and V3C3. V3C1 includes 7,475 videos and V3C2 includes 9,760 videos; V3C3 is not used in this work. For AVS 2025, we operate over V3C2, which contains roughly 1.43 million pre-segmented video shots.

Each video in V3C2 has been pre-segmented into shots, and TREC provides one representative keyframe per shot. In addition, we sampled one keyframe per second across V3C2 to improve recall for short or fast-changing events, yielding a dense index of approximately 4.5 million keyframes. While keyframes offer a scalable surrogate for video, they can underrepresent temporal context, particularly for action-centric queries.

TREC has released AVS topics for V3C2 across multiple years. The 2022 and 2023 tasks provided 30 and 20 topics, respectively. The 2024 task released 20 additional topics for V3C2, and these topics are reused in 2025 as a progress set. We use the 50 topics from 2022 and 2023 for system development, while our six official submissions are evaluated on the official 2024/2025 topic set. Human relevance judgments for the 2024/2025 topics were withheld during system development and used by NIST for official evaluation.

2.3 Evaluation and Relevance Judgments

We evaluate retrieval effectiveness using standard TREC AVS measures, including nDCG and inferred Average Precision (infAP) [9], along with precision and recall at multiple cutoffs. Following TREC practice, unjudged items are treated as non-relevant for nDCG and precision-based metrics, while infAP estimates performance under incomplete judgment pools. Human judgments for the 2022 and 2023 topics support internal development analyses, whereas judgments for the official 2024/2025 progress topics are withheld and used only for NIST scoring.

3 System Overview

All submitted systems follow a shared large-scale keyframe retrieval pipeline over V3C2. Each AVS topic is encoded as a text query vector and used to retrieve candidate keyframes by cosine similarity from a vector index. Depending on the workflow, retrieved candidates are optionally enhanced at query time through semantic expansion, modality-aware decomposition, VLM-based relevance reranking, and selective audio fusion. These components are instantiated in four workflows and six official systems, as summarized in Section 4.

3.1 Keyframe Sources and Indexing

We operated over two keyframe sources. For the fine-grained embedding-only configuration, we used the official TREC representative keyframe per pre-segmented scene. For all other workflows, we densely sampled one keyframe per second across V3C2 to improve recall for short or fast-changing events. This yielded approximately 4.5 million keyframes, which were embedded offline and stored in a vector index for efficient cosine-similarity search at query time.

3.2 Visual Embedding Models

We benchmarked nine contrastive text–image embedding models from the CLIP, SigLIP, and SigLIP2 families to characterize system-level speed–accuracy trade-offs (Section 5.1). Benchmarking measured GPU embedding throughput, CPU preprocessing throughput, memory usage, and retrieval effectiveness on previously judged AVS topics. Because these trade-offs depend strongly on hardware and system configuration, embedding selection

must be validated per deployment environment. Based on our measured profile, most submitted systems used `siglip2-base-patch16-naflex` due to its favorable balance of efficiency and retrieval quality.

In addition, a separate fine-grained CLIP-derived embedding model was submitted as an alternative visual configuration to assess the impact of fine-grained visual–text alignment on AVS performance. This model was used only in the embedding-only workflow over official TREC keyframes.

3.3 LLM-Based Semantic Expansion

For expansion-based workflows, we used an LLM to generate 100 semantically similar reformulations of each AVS topic. Each reformulation was embedded using a text encoder, and the resulting vectors were averaged into a pooled query embedding. This pooled vector was then used to retrieve candidates from the dense keyframe index. Semantic expansion is intended to improve recall and robustness to phrasing variation without additional training.

3.4 Modality-Aware Query Decomposition

Query decomposition offers improved video retrieval by leveraging knowledge from LLMs [10]. For decomposition-based workflows, we used an LLM to split each topic into up to five semantic components: `scene setting`, `entities and objects`, `semantic actions`, `spoken content`, and `non-speech audio`. Each component deemed relevant was expanded into 100 phrases. For visual retrieval, we retain only the visually grounded components and fuse their averaged embeddings by unweighted summation into a pooled visual query embedding. In practice, `spoken content` was not activated by the 2025 topic set, while `non-speech audio` occurred infrequently but was used solely to trigger selective CLAP-based audio retrieval (Section 3.6).

3.5 VLM Relevance Scoring and Reranking

VLMs offer the ability to enhance relevance judgements [11]. Some workflows applied a VLM as a relevance judge to rescore top retrieved candidates. Each query–keyframe pair was scored on a discrete 0–3 relevance scale, where 3 indicates strong relevance. Candidates were grouped by score and reranked in descending score order, preserving cosine similarity order within each score tier. We used both `phi-3.5-vision` and `gpt-4.1-mini` as judges in different systems. VLM scoring was also used internally as a scalable proxy for human judgments on the 2022/2023 topics to support embedding model selection and workflow tuning.

3.6 Selective Audio Retrieval with CLAP

For topics where decomposition indicates likely non-speech audio relevance, we applied a complementary retrieval pass over audio. We used CLAP (Contrastive Language–Audio Pretraining) [12] to embed pre-segmented 10-second audio clips and stored these embeddings in a separate vector index. Expanded audio phrases were embedded with CLAP's text encoder and matched to audio clips by cosine similarity. The top audio and visual results were normalized to a [0,1] similarity range and fused by score summation using a simple late-fusion strategy to form a final multimodal ranking. This module was enabled only for a subset of topics to limit unnecessary computation and reduce spurious matches.

3.7 Workflows and Submitted Systems

The components above are assembled into four workflows reflecting different combinations of augmentation, reranking, and fusion. These workflows and the corresponding six official systems are described in Section 4, and their effectiveness is reported in Section 5.

4 Retrieval Workflows and Submitted Runs

We submitted six official systems to the 2025 TREC Ad-hoc Video Search (AVS) task. These systems instantiate four retrieval workflows that vary in their use of query augmentation, VLM-based relevance reranking, and selective audio fusion. All systems operate over a 1 fps visual keyframe index computed with our selected SigLIP2 NAFlex embedding model [7]; as noted in Section 5, such model selection depends on hardware-specific throughput–effectiveness trade-offs. All systems follow TREC submission conventions and are evaluated by NIST on the official 2024/2025 topic set.

4.1 Workflow A: Embedding-Only Retrieval (Fine-Grained CLIP)

We instantiated this workflow using a fine-grained CLIP-derived embedding model over the official TREC representative keyframes. The fine-grained model was developed as part of a parallel study on fine-grained contrastive training for image retrieval; details of its architecture and training are reported separately. We include this system to assess baseline embedding-only performance and to provide context for the improvements achieved through query-time enhancement strategies in Workflows B–D, all of which use SigLIP2 NAFlex as the visual backbone.

4.2 Workflow B: Semantic Expansion with VLM Reranking

In this workflow, each AVS topic is expanded into 100 semantically similar reformulations using an LLM. Each reformulation is embedded with SigLIP2 NAFlex, and the vectors are averaged into a pooled query representation for initial retrieval over the 1 fps keyframe index. The top candidates are then scored by a VLM on a discrete 0–3 relevance scale and reranked by score tiers, preserving cosine-similarity order within tiers. To reduce reranking volatility, one variant applies batched reranking with a sliding window of 250 results with 125 results overlap between adjacent windows, which limits extreme rank shifts while maintaining local similarity structure. We submitted three systems under this workflow: two using `phi-3.5-vision` (full and batched reranking) and one using `gpt-4.1-mini` (full reranking).

4.3 Workflow C: Modality Decomposition with VLM Reranking

This workflow performs LLM-based modality decomposition, splitting each topic into up to five semantic components: `scene setting`, `entities and objects`, `semantic actions`, `spoken content`, and `non-speech audio`. In this workflow, we retain only the visually grounded components (`scene setting`, `entities and objects`, `semantic actions`) and ignore audible components entirely for retrieval. The retained components are each expanded into 100 phrases, embedded, and averaged within modality; the resulting vectors are fused by unweighted summation into a pooled visual query representation for retrieval from the 1 fps keyframe index. Retrieved candidates are then rescored and reranked using the same VLM tiered scoring scheme as in Workflow B. We submitted one system using `phi-3.5-vision` as the judge.

4.4 Workflow D: Modality Decomposition with CLAP Fusion

This workflow reuses modality decomposition and expansion, but activates audio retrieval only when `non-speech audio` is inferred as relevant. Expanded audio phrases are embedded with CLAP's text encoder and compared to precomputed CLAP audio embeddings over 10-second clips. Using a basic late-fusion strategy, the top visual and audio results are each normalized to a [0,1] range and fused by simple score summation to form a final multimodal ranking. This workflow does not apply VLM reranking. We submitted one audio-enhanced system under this workflow.

4.5 Summary of Submitted Systems

Table 1 summarizes the six submitted systems, organized by workflow, judge model (if applicable), and official run IDs.

Wf.	System	Exp.	Decomp.	CLAP	VLM Judge	Run ID
A	Fine-Grained CLIP embedding only	–	–	–	–	`F_M_C_A_ncsu-las.25_3`
B	Semantic expansion + rerank (full)	✓	–	–	phi-3.5-vision	`F_M_C_D_ncsu-las.25_2`
	Semantic expansion + rerank (batched)	✓	–	–	phi-3.5-vision	`F_M_C_D_ncsu-las.25_4`
	Semantic expansion + rerank (full)	✓	–	–	gpt-4.1-mini	`F_M_C_D_ncsu-las.25_1`
C	Visual decomposition + rerank	–	visual-only	–	phi-3.5-vision	`F_M_C_D_ncsu-las.25_3`
D	Decomposition + selective CLAP fusion	–	audio-trigger	✓	–	`F_M_N_D_ncsu-las.25_3`

Table 1: Submitted AVS 2025 systems organized by workflow (Wf.) and enabled components. Workflow C retains only visually grounded decomposition outputs for retrieval and reranking. Workflow D uses `non-speech audio` decomposition solely to trigger CLAP retrieval.

5 Results

5.1 Embedding Model Benchmark: Speed–Accuracy Trade-offs

To inform selection of a practical embedding model for dense keyframe retrieval, we benchmarked nine contrastive text–image models from the CLIP, SigLIP, and SigLIP2 families on V3C2 keyframes. We measured GPU embedding throughput, single-thread CPU preprocessing throughput, peak GPU memory usage, and retrieval effectiveness on development topics using VLM-based relevance judgments.

Across models, larger architectures yielded modest effectiveness gains but at substantially higher computational cost on our server configuration. Smaller CLIP variants achieved high GPU throughput yet were frequently bottlenecked by CPU-side preprocessing, whereas SigLIP2 NAFlex models offered markedly higher preprocessing throughput and a favorable CPU-to-GPU balance, enabling efficient large-scale indexing at 1 fps. Because these trade-offs depend strongly on hardware and system setup, embedding selection must be validated per deployment environment rather than assumed to transfer directly from our configuration. Based on our measured profile, we selected `siglip2-base-patch16-naflex` as the default visual embedding model for all submitted systems except the fine-grained embedding-only configuration.

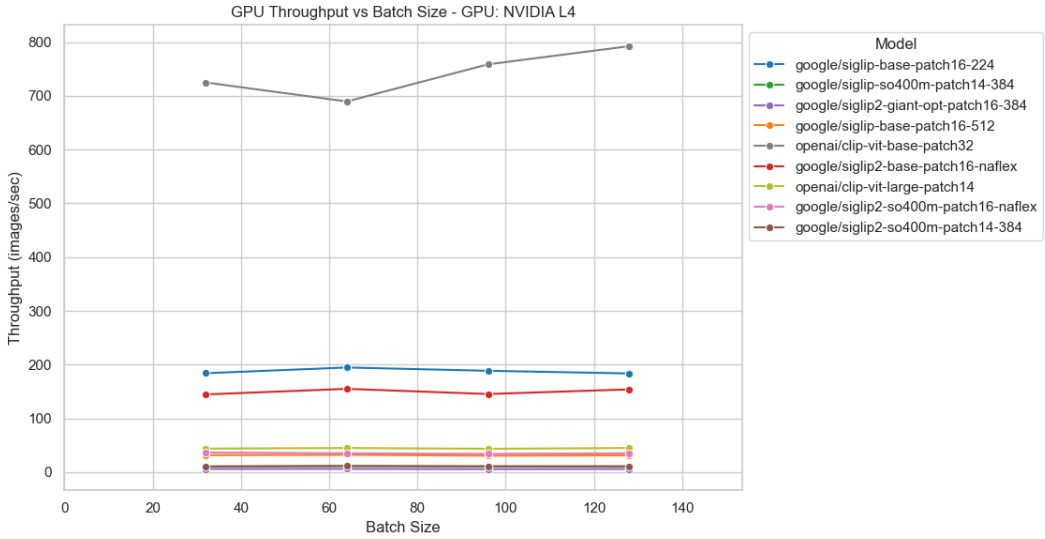

Figure 1: GPU throughput by batch size for all evaluated embedding models.

5.2 VLMs as Relevance Judges

We evaluated vision–language models (VLMs) as automated relevance judges to support scalable internal evaluation and to provide reranking signals in AVS submissions. Using GPT-4.1-mini as the judge, we compared discrete 0–3 relevance scores against human annotations on 1,000 randomly sampled query–keyframe pairs from the judged 2022/2023 AVS topics. Following our reranking configuration, we treated score 3 as relevant and scores 0–2 as non-relevant.

GPT-4.1-mini achieved 88.1% accuracy (881/1000), with recall of 63.2% (91/144) and precision of 60.0% (91/157). As shown in Figure 4, most disagreements were false positives, indicating that the VLM tends to accept semantically plausible matches even when human assessors apply narrower interpretations.

We performed qualitative error analysis to identify recurrent sources of disagreement between the VLM and human annotators. A prominent pattern reflected mismatches in lexical conventions between the model and the evaluation population, particularly for regionally variable terms. For example, for the query *"A man riding a scooter"*, the VLM frequently labeled images of mopeds as relevant, whereas human evaluators did not. This divergence is plausibly attributable to regional differences in the meaning of "scooter": in some locales, mopeds and scooters are used interchangeably, while in others they denote distinct vehicle categories. Similar variability applies to terms such as "bike," which may refer to bicycles or motorcycles depending on regional or contextual usage. Taken together, these

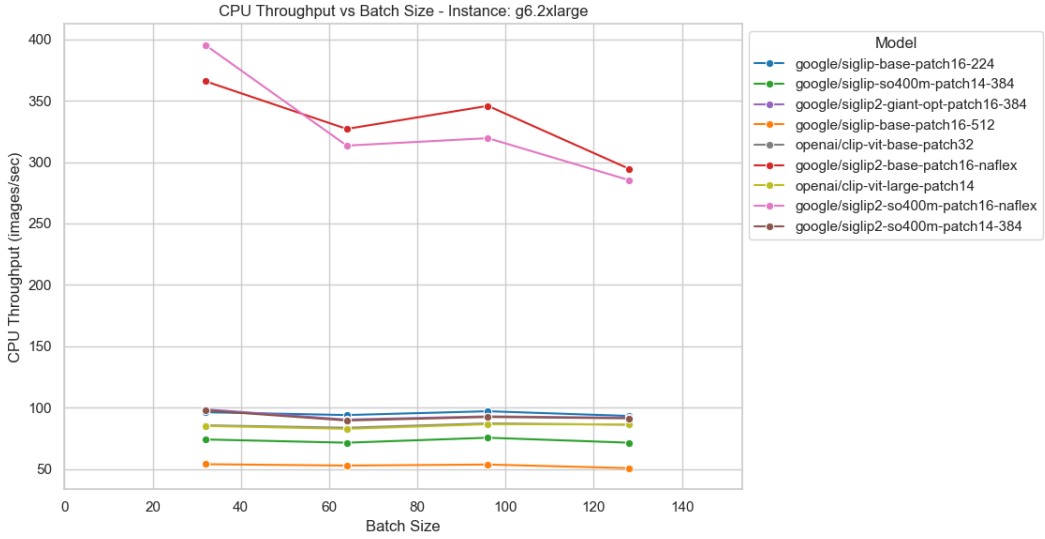

Figure 2: Single-threaded CPU preprocessing throughput for all evaluated embedding models.

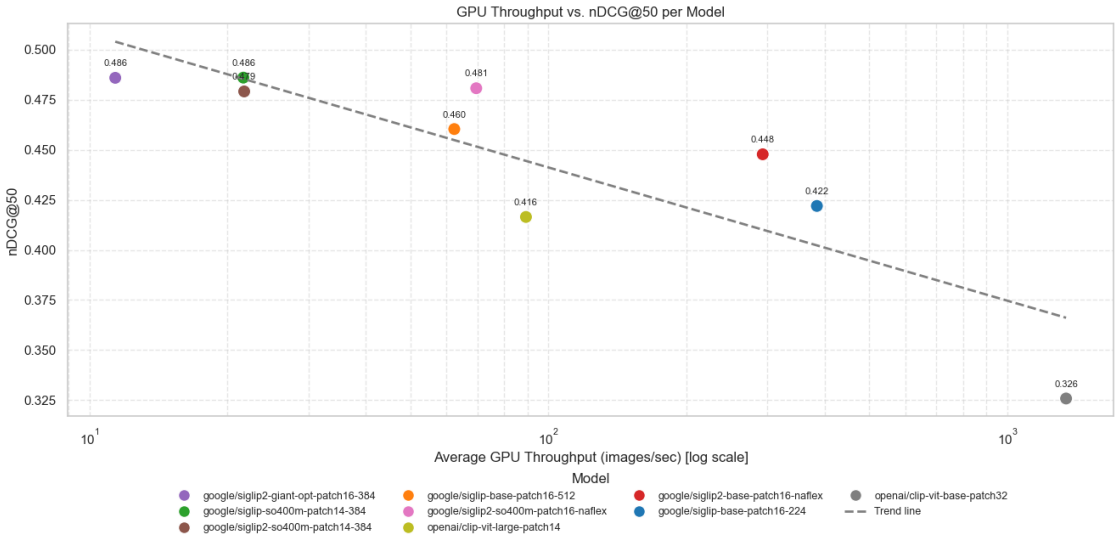

Figure 3: Retrieval effectiveness (nDCG@50) versus GPU throughput all evaluated embedding models.

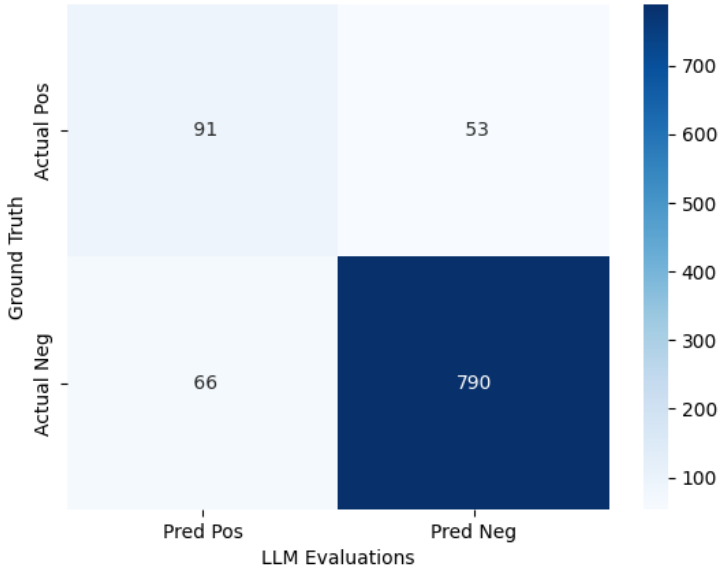

Figure 4: Confusion matrix comparing VLM (GPT-4.1-mini) and human judgments using single keyframes. Scores 3 treated as relevant; 0–2 as non-relevant.

cases suggest that some apparent false positives may reflect differing domain and language assumptions between the VLM and annotators rather than clear model misjudgments.

A second recurring pattern of disagreement concerned temporal misalignment between the action implied by a keyframe and the actual content within the retrieved video segment. For the query *"A person taking a picture using a cell phone camera"*, the VLM frequently judged segments as relevant when the keyframe showed a subject raising or positioning a phone in a manner consistent with photo taking. However, inspection of the full segments revealed that the action was often not completed within the segment's temporal bounds or did not occur at all. This behavior underscores a fundamental limitation of single-frame evaluation: without access to surrounding temporal context, the model cannot reliably determine whether an inferred action is realized within the segment represented by the keyframe.

We further observed disagreements arising from subjective or underspecified query language, which can lead the model and annotators to apply different implicit criteria. For example, for the query *"A heavy man indoors"*, disagreement primarily reflected differing interpretations of what qualifies as "heavy" in visual context. Likewise, for *"A child climbs an object outdoors"*, we noted a case where the VLM treated a child climbing down as relevant while the human judgment did not, illustrating that directionality of motion (ascending versus descending) was not sufficiently constrained and can lead to differing relevance decisions. These patterns emerged from manual inspection of disagreement cases across the sampled query–keyframe pairs and were consistently observed across multiple queries, suggesting systematic rather than idiosyncratic failure modes. While we did not conduct exhaustive error taxonomies within the competition timeline, these recurring themes highlight how lexical ambiguity, temporal uncertainty, and subjective descriptors can yield systematic mismatches in relevance assumptions rather than clear model failures.

The temporal nature of video suggests that single-keyframe evaluation is inherently limited for action-centric and temporally ambiguous queries. We conducted preliminary experiments with five uniformly sampled keyframes per segment but did not observe consistent improvements within our competition timeline, as additional frames introduced false positives that offset recall gains under our simple aggregation approach. However, we believe this negative result reflects limitations in our sampling density and aggregation strategy rather than fundamental constraints of temporal modeling. Incorporating richer temporal context through denser sampling, longer sequences, or models explicitly designed for cross-frame reasoning remains a promising direction for improving relevance estimation on temporally complex topics. Due to these mixed preliminary results and time constraints, all submitted systems relied on single-keyframe judgments.

VLM-based relevance scoring offers a strong, low-overhead signal that tracks human annotations closely, while showing systematic failure modes under lexical ambiguity, subjective topic phrasing, and temporal uncertainty inherent to single-frame evaluation.

5.3 CLAP Retrieval Effectiveness

To evaluate the practical contribution of audio-based retrieval, we analyzed qualitative cases where CLAP retrieved relevant segments missed by visual-only pipelines, focusing on queries where discriminative audio was not evident in the corresponding keyframes.

A representative example is the query *"A rainy day outdoors."* Table 2 shows that keyframes from some of CLAP's top-ranked segments lacked clear visual evidence of rainfall: common cues (umbrellas, wet surfaces, raindrops) were missing, occluded, or out of frame. In contrast, the audio strongly conveyed rainy conditions (e.g., steady rainfall, splashing, muffled ambience), enabling CLAP to retrieve relevant segments despite weak visual confirmation.

Image	Notes
	Close-up of an umbrella edge; the object is difficult to identify visually.
	Rain is present, but only faint streaks are visible against a dark background.
	A distant clear sky is visible, but the ship is situated under active rainfall.

Table 2: Examples of positive results for "A rainy day outdoors" retrieved by CLAP based on audio cues but missed visually.

These cases suggest that CLAP can recover relevant content for topics where characteristic sounds provide salient semantic evidence. Moreover, the benefit is not confined to explicitly sound-centric queries. For instance, *"A man is putting on a jacket or a t-shirt"* may exhibit weak visual cues in isolated frames yet include distinctive fabric-rustling audio. Similarly, *"People inside an airport terminal"* may be supported by ambient intercom announcements or distant aircraft noise.

However, not all topics benefit from audio augmentation. Purely visual queries (e.g., *"A bald man with glasses"*) showed no observable gain, reinforcing the need to trigger audio retrieval selectively when semantic analysis suggests plausible non-speech relevance. Overall, CLAP serves as a complementary modality that can surface content missed by visual-only systems, though careful triggering is necessary to limit unnecessary computation and false positives.

Realizing this potential in practice requires principled multimodal integration. Our fusion approach used unweighted score summation over normalized similarities, treating visual and audio embeddings as directly comparable despite fundamental differences in their training and operational characteristics. SigLIP2 and CLAP were trained independently on different corpora, and their embedding spaces encode relevance at potentially different scales and semantic granularities. Further, the visual index operates at 1 fps keyframe density while audio retrieval uses pre-segmented 10-second clips, such that even when both modalities identify the same segment as relevant, they often retrieve non-overlapping temporal boundaries. Without learned combination weights, modality-specific calibration, or temporal alignment, simple score summation risks amplifying spurious matches rather than reinforcing complementary evidence.

5.4 Official AVS 2025 Evaluation

Table 3 reports the official TRECVID 2025 AVS results for all LAS submissions. All configurations operate over the same corpus and evaluation topics and share a single visual embedding backbone and candidate generation stage. Differences arise only in query interpretation and reranking strategy, enabling controlled substitution of semantic components within a fixed retrieval substrate.

Workflow	System	infAP	infNDCG	iP@10
A	Fine-grained embedding-only	0.143	0.278	0.445
B	Semantic exp. + rerank (phi-3.5, full)	0.294	0.474	**0.780**
B	Semantic exp. + rerank (phi-3.5, batched)	0.271	0.442	0.755
B	Semantic exp. + rerank (gpt-4.1-mini, full)	**0.399**	**0.558**	0.760
C	Visual decomposition + rerank (phi-3.5)	0.287	0.466	0.715
D	Decomposition + selective CLAP fusion	0.061	0.202	0.285

Table 3: Official TRECVID 2025 AVS results for LAS systems. Values are taken directly from official "all" rows and rounded to three decimals. Highest values per metric are shown in bold.

Across 20 evaluation topics, the strongest configuration, semantic expansion followed by full reranking with GPT-4.1-mini, achieves a mean infAP of **0.399**, outperforming all other LAS submissions. Overall, the official results indicate that lightweight query-time enhancements and VLM-based relevance estimation can produce substantial differences in end-to-end AVS effectiveness within a fixed retrieval backbone.

Two findings are particularly salient. First, VLM choice and scale matter for evaluation-oriented reranking. When operating over identical candidate pools, systems reranked with `gpt-4.1-mini` improve infAP on 19 of 20 topics relative to `phi-3.5-vision`, yielding a +0.104 mean per-topic gain and a +35.5% improvement in overall mean infAP. Because the underlying retrieval stage remains unchanged, these gains arise purely from improved relevance discrimination rather than recall expansion or ensemble diversity effects.

Second, while CLAP demonstrates qualitative potential as a complementary signal for topics implying non-speech audio, effective fusion remains challenging. In our current setup, unweighted late fusion does not reliably convert topic-level audio gains into system-level improvements, underscoring that multimodal augmentation requires principled alignment rather than additive scoring.

5.4.1 Workflow-Level Effects

The results illustrate the interaction between candidate generation and VLM-based relevance estimation. Dense visual embeddings determine the feasible recall ceiling, while reranking reshapes ordering within that candidate pool.

The fine-grained embedding-only configuration achieves a mean infAP of **0.143**, indicating that visual similarity alone is insufficient for competitive ranking. Introducing VLM-based reranking with Phi-3.5 increases mean infAP to **0.294**, a +0.151 absolute improvement over embedding-only (+105% relative gain). Replacing Phi-3.5 with GPT-4.1-mini further increases mean infAP to **0.399**, an additional +0.104 absolute gain (+35.5% relative improvement over Phi-3.5).

These comparisons demonstrate that, within our controlled configurations, relevance estimation quality during reranking is the most influential variable affecting end-to-end ranking effectiveness.

Selective CLAP-based audio fusion reduces mean infAP to **0.061**, confirming that unweighted score summation over independently trained embedding spaces can degrade ranking effectiveness despite qualitative evidence of complementary signal (Section 5.3). Because visual and audio embeddings were not jointly calibrated and operate over temporally misaligned retrieval units, naive fusion amplifies noise rather than reinforcing relevance. This outcome underscores that multimodal augmentation requires alignment beyond additive scoring.

5.4.2 Architectural Control and Model Substitution

All LAS configurations share a single visual embedding backbone and candidate generation stage. Variations occur only in query interpretation (e.g., semantic expansion or decomposition) and reranking modules. We do not combine rankings from multiple visual retrieval systems or ensemble independently produced ranking lists.

LLM usage is confined to:

1. Query reformulation prior to retrieval, and/or
2. Shot-level relevance estimation during reranking.

We do not perform text-to-image generation, synthetic visual augmentation, or image captioning as an intermediate retrieval step. The VLM is used strictly as a discriminative relevance judge rather than as a generative model within the retrieval pipeline.

Because the underlying candidate sets remain fixed across controlled comparisons (e.g., GPT vs. Phi), observed performance differences arise from changes in relevance estimation quality rather than recall expansion, generative augmentation, or ensemble diversity effects.

5.4.3 Controlled Comparison: GPT-4.1-mini vs. Phi-3.5

To isolate the effect of VLM capacity, we compare the two full reranking configurations directly. Both operate over identical candidate pools.

Figure 5 shows the per-topic infAP difference:

$$\Delta_{\text{GPT–Phi}} = \text{infAP}_{\text{GPT-4.1-mini}} - \text{infAP}_{\text{Phi-3.5}}$$

GPT-4.1-mini improves performance on **19 of 20 topics**, with:

- Mean per-topic gain: **+0.104 infAP**
- Median per-topic gain: **+0.102 infAP**

Only one topic exhibits marginal regression. Gains are distributed broadly rather than concentrated in a small subset, indicating that increased VLM capacity systematically improves shot–query relevance estimation.

Because candidate sets are identical across these two configurations, the observed improvements arise purely from reranking decisions rather than differences in recall.

5.4.4 Topic-Level Patterns

Per-topic analysis further clarifies the role of VLM-based relevance estimation. Figure 6 visualizes per-topic infAP across all LAS configurations.

Several patterns emerge:

- Topic 1757 exhibits strong performance across all VLM-based systems, indicating clear visual distinctiveness.
- Lower-performing topics (e.g., 1754–1756) show uniformly weak embedding-only scores but substantial gains under GPT-4.1-mini reranking.
- Improvements are not confined to a single semantic class; gains appear across varied query structures.

Notably, larger improvements tend to occur on topics involving lexical ambiguity or action-centric interpretation, aligning with the qualitative disagreement patterns discussed in Section 6.

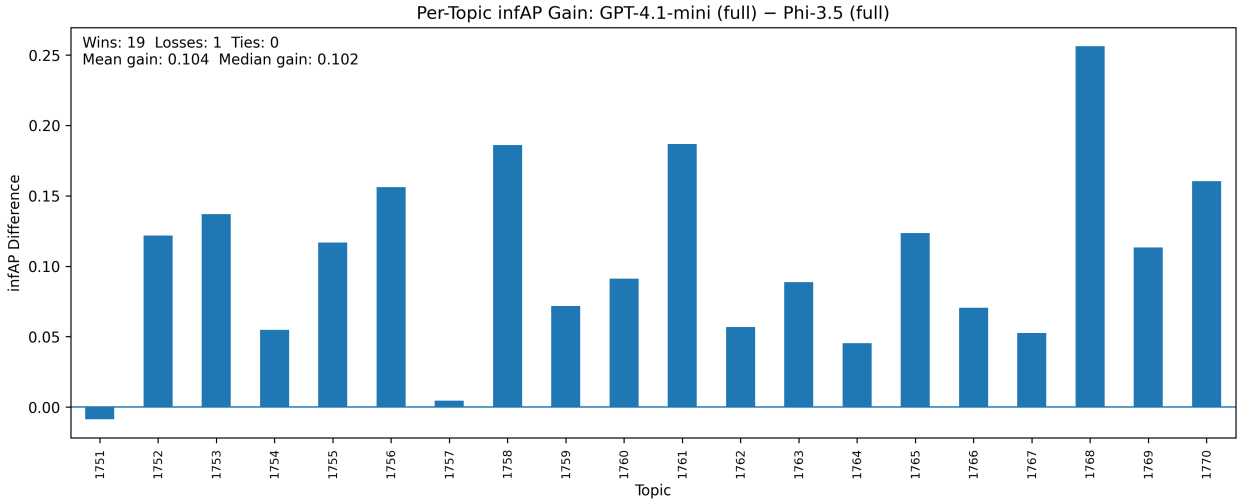

Figure 5: Per-topic infAP difference between GPT-4.1-mini (full) and Phi-3.5 (full) reranking. Positive values indicate improvement under GPT-4.1-mini. GPT-4.1-mini improves performance on 19 of 20 topics, with a mean gain of +0.104 infAP.

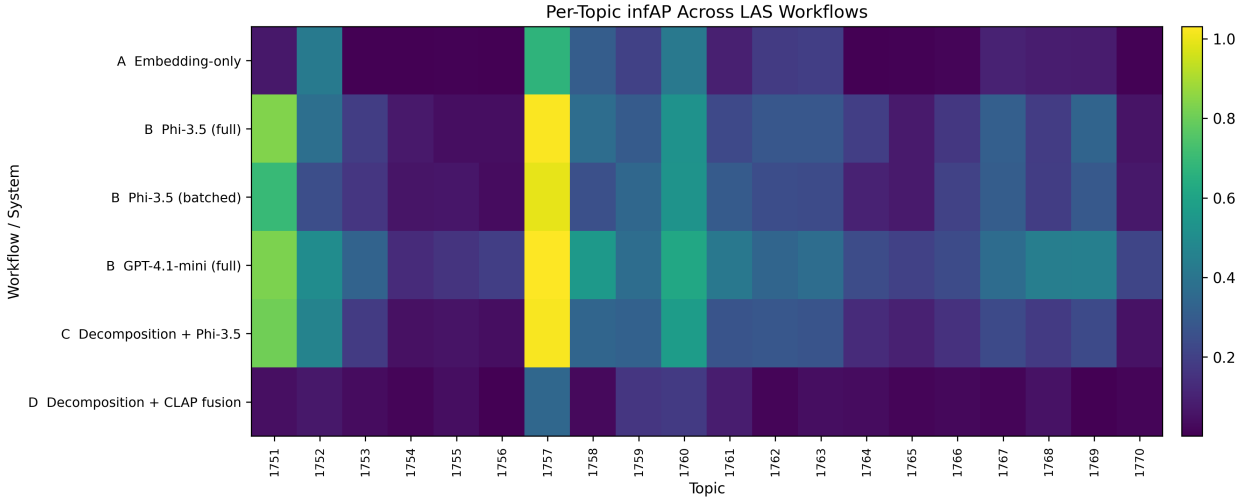

Figure 6: Per-topic infAP across LAS workflows. Rows correspond to runs and columns to topics.

5.4.5 Early Precision and Ranking Stability

To further examine how VLM-based relevance estimation affects ranking behavior, we analyze precision at increasing depth cutoffs.

Figure 7 compares precision-at-k across embedding-only retrieval, Phi-3.5 reranking, and GPT-4.1-mini reranking. GPT-4.1-mini consistently improves precision across early depth thresholds while maintaining stronger overall ranking stability.

Although Phi-3.5 achieves competitive precision at very small cutoffs (e.g., $k = 10$), GPT-4.1-mini produces higher precision across broader depth ranges and yields substantially stronger overall infAP. This pattern indicates that improvements under GPT-4.1-mini are not limited to isolated top-rank adjustments but reflect more consistent ordering of relevant shots throughout the ranked list.

Importantly, because candidate pools are identical across these configurations, differences in early precision arise solely from reranking decisions. The results therefore reinforce that improvements stem from enhanced shot–query relevance estimation rather than changes in recall or candidate diversity.

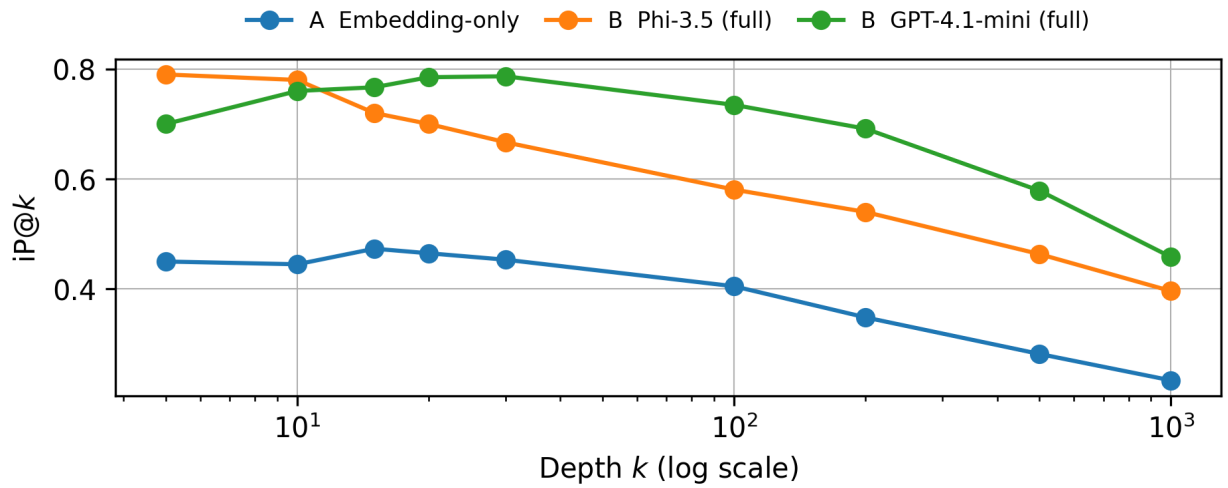

Figure 7: Precision at increasing depth cutoffs for embedding-only retrieval, Phi-3.5 reranking, and GPT-4.1-mini reranking. GPT-4.1-mini improves precision across broader depth ranges, indicating more consistent relevance ordering.

Within a fixed retrieval backbone, increasing VLM capacity improves both early precision and cumulative ranking effectiveness, supporting the interpretation that decision quality during reranking is the primary driver of end-to-end performance.

6 Discussion

This paper reported a system-level study of AVS 2025 retrieval workflows, spanning hardware-aware visual embedding selection, training-free query-time augmentation, VLM-based relevance judging, and selective audio retrieval. While each component influenced performance, our most consistent findings center on the role and behavior of VLM judgments within the pipeline.

First, VLM-based relevance scoring provides a strong, low-overhead signal that tracks historical human annotations closely, making it practically useful for internal evaluation and reranking at scale. GPT-4.1-mini achieved high agreement with 2022/2023 human labels, and the official AVS results show that reranking with a VLM can translate these judgments into substantial end-to-end gains when applied to high-recall candidate sets. Moreover, VLM capacity matters: when operating over identical candidate pools, systems reranked with `gpt-4.1-mini` improved infAP on 19 of 20 topics relative to `phi-3.5-vision`, yielding a mean per-topic gain of +0.104 infAP and a +35.5% overall improvement. Because these comparisons hold the retrieval stage fixed, the gains arise purely from improved relevance discrimination, confirming that VLM quality is the most influential variable affecting end-to-end ranking effectiveness within our workflows. This improvement is not limited to top-rank adjustments but extends across depth cutoffs, with GPT-4.1-mini maintaining higher precision at broader retrieval depths than Phi-3.5, indicating more consistent relevance ordering throughout the ranked list.

Second, the VLM disagreements are often systematic and interpretable rather than erratic. Our qualitative analysis indicates that many "false positives" arise from mismatched lexical or domain conventions between the model and annotators (e.g., regional usage of "scooter" or "bike"), from temporal misalignment when a keyframe implies an action that is not completed within a segment's bounds, and from subjective or underspecified topic language (e.g., thresholds for "heavy" or directionality in "climbs"). In our limited multi-keyframe test, providing five uniformly sampled keyframes increased recall but also increased false positives, producing no consistent net gain for reranking in the current setup. We view this outcome as inconclusive and expect that richer use of temporal context may still help address action-centric ambiguity while preserving precision.

Third, embedding model benchmarking remains foundational because it determines the feasible operating point for dense indexing, which in turn shapes the candidate pool that VLMs rerank. On our hardware profile, SigLIP2 NAFlex models provided the best efficiency–effectiveness trade-off and enabled 1 fps indexing over V3C2. However, the benchmark also underscores that these trade-offs are deployment-specific, and model choice should be revalidated per system configuration.

Finally, CLAP-based audio retrieval shows clear qualitative utility when non-speech audio is semantically implied, recovering segments that visual-only retrieval misses. Yet the official results show that our audio-augmented system achieved a mean infAP of only 0.061, well below even the embedding-only baseline. This outcome reflects the limitations of unweighted late fusion over independently trained embedding spaces that operate at different temporal granularities and encode relevance at incompatible scales. The result highlights that complementary modalities require careful triggering and principled fusion strategies; without calibration or learned combination weights, simple score summation can degrade rather than enhance ranking quality.

In summary, our AVS 2025 submissions demonstrate that practical improvements can be achieved by combining hardware-aware dense retrieval with training-free query augmentation and VLM reranking. Among these, VLM relevance estimation emerges as the most influential lever for end-to-end performance, while its predictable disagreement modes point to targeted calibration opportunities. Audio augmentation via CLAP remains promising but requires stronger fusion schemes to realize consistent gains alongside robust VLM-based ranking.

References

[1] W. Ma, K. Li, Z. Jiang, M. Meshry, Q. Liu, H. Wang, C. Häne, and A. Yuille, "Rethinking video-text understanding: Retrieval from counterfactually augmented data," 2025. [Online]. Available: https://arxiv.org/abs/2407.13094

[2] H.-L. Tran, T.-A. Nguyen-Nhu, H.-P. Phan-Nguyen, T.-H. Nguyen, N.-M. Nguyen-Dich, A. Dao, H.-D. Do, Q. Nguyen, H. M. Le, and Q.-V. Dinh, "Towards efficient and robust moment retrieval system: A unified framework for multi-granularity models and temporal reranking," 2025. [Online]. Available: https://arxiv.org/abs/2504.08384

[3] G. A. (NIST), J. F. (NIST), A. G. (NIST), L. D. (NIST), Y. G. T. C. Dublin), and G. Q. (LIG), "Trecvid 2024 – evaluating video search, captioning, and activity recognition," in *The Thirty-Third Text REtrieval Conference Proceedings (TREC 2024), Gaithersburg, MD, USA, November 15-18, 2024*, ser. NIST Special Publication, vol. 1329. National Institute of Standards and Technology (NIST), 2024. [Online]. Available: https://trec.nist.gov/pubs/trec33/papers/Overview_avs_vtt_actev.pdf

[4] A. Chen, F. Zhou, Z. Wang, and X. Li, "Cliprerank: An extremely simple method for improving ad-hoc video search," 2024. [Online]. Available: https://arxiv.org/abs/2401.08449

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[6] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," 2023. [Online]. Available: https://arxiv.org/abs/2303.15343

[7] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," 2025. [Online]. Available: https://arxiv.org/abs/2502.14786

[8] L. Rossetto, H. Schuldt, G. Awad, and A. A. Butt, "V3c - a research video collection," 2018. [Online]. Available: https://arxiv.org/abs/1810.04401

[9] E. Yilmaz, E. Kanoulas, and J. A. Aslam, "A simple and efficient sampling method for estimating ap and ndcg," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 603–610.

[10] S. R. Dipta and F. Ferraro, "Q2e: Query-to-event decomposition for zero-shot multilingual text-to-video retrieval," 2025. [Online]. Available: https://arxiv.org/abs/2506.10202

[11] J.-H. Yang and J. Lin, "Toward automatic relevance judgment using vision–language models for image–text retrieval evaluation," 2024. [Online]. Available: https://arxiv.org/abs/2408.01363

[12] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," 2024. [Online]. Available: https://arxiv.org/abs/2211.06687