TREC 2025 Ad-hoc Video Search (AVS) Track Overview

George Awad
Information Access Division
Information Technology Laboratory
National Institute of Standards and Technology (NIST)
geo.awad@nist.gov

Abstract

The Ad-hoc Video Search (AVS) task at TREC continues to serve as a long-running benchmark for measuring progress in open-vocabulary video retrieval. The 2025 cycle builds on more than a decade of work and reflects the rapidly evolving landscape of multimodal and vision-language models. This overview describes the task design, dataset characteristics, evaluation protocol, participating teams, and the general retrieval trends observed during this assessment cycle.

1 Introduction

Searching video using free-form natural language is a deceptively simple objective: people want to express a thought, an observation, or a visual moment, and immediately retrieve the segments that contain it. The TRECVID Ad-hoc Video Search (AVS) task was created to explore that challenge in a rigorous, large-scale evaluation setting. Instead of relying on predefined labels or closed vocabularies, the task embraces the open and unpredictable nature of human queries. Participants must sift through millions of video shots and identify those that genuinely capture the semantics expressed in a query, however broad or fine-grained it may be.

Over the years, AVS has tracked the evolution of retrieval models—from early concept banks, to deep convolutional embeddings, and now to multimodal foundation models capable of surprisingly nuanced image–text alignment. The 2025 cycle follows this trajectory and highlights the impact of two-stage pipelines, token-level alignment strategies, and large vision-language models reintroduced into the retrieval loop.

What follows is a detailed walkthrough of the task dataset, judging procedures, participating teams, and the thematic patterns that shaped this year's results.

2 Task Definition

The AVS task requires participating systems to return a ranked list of up to 1000 shots from a pool of 1,425,454 possible candidates. Each run is evaluated against a set of textual queries that describe people, objects, actions, locations, or various combinations of these elements. Unlike classification tasks, there are no predefined categories; systems must rely entirely on their ability to interpret the semantics expressed in text and match them to visual content in the videos.

The queries for the 2025 cycle consisted of 20 topics, identical to those used in 2024. Some queries are visually straightforward—a specific article of clothing or a single object—while others weave together multiple facets and require systems to detect subtle relationships between people, their actions, and their environments.

3 Dataset Overview

3.1 Evaluation Collection: V3C2

The evaluation set is drawn from the V3C2 portion of the Vimeo Creative Commons (V3C)[2] collection. It contains 9760 publicly licensed videos totaling approximately 1300 hours. Each video averages around eight minutes in duration and covers a broad range of subjects, camera styles, and visual conditions. This diversity is intentional: retrieval systems must generalize to the unpredictable nature of real-world video. The dataset attempts to capture the variability of ordinary web video uploads— from hand-held clips and amateur content to semi-professional short productions.

3.2 Development Data

Teams were free to leverage several earlier TRECVID datasets for development such as the IACC.1–3 collections (about 2000 hours combined), used extensively in AVS tasks between 2010 and 2018, and V3C1 (about 1000 hours), which supported the AVS cycles from 2019 to 2021. These datasets include concept annotations and prior ad-hoc query assessments, which continue to serve as a valuable foundation for model training and system tuning.

4 Queries

The 20 queries span the spectrum of complexity that characterizes the AVS task: from simple entity mentions to multi-facet combinations involving who, what, where, and the manner in which an action is performed. Table 1 lists all 20 queries.

Table 1: Ad-hoc Video Search Queries

Category	**Query Text**
A Person, Location, or Object	A bald man with glasses
	Rainy day outdoors
	Pink necktie
	A white sweater
Person + Action	Person is wiping himself or an object using their bare hands or other object
	A man is putting on a jacket or t-shirt
Person + Location	People inside an airport terminal
	A man inside a workshop
Person + Object	Man wearing a checked shirt
	A woman wearing a floral top or dress
Person + Object + Location	Two or more persons indoors with coffee cups or mugs seen with them
	Two women together wearing hats, excluding caps, outdoors
Object + Location	A traffic light seen at an intersection of a road or street
	A map seen on a wall indoors
Person + Action + Object	An adult is wrapped in a blanket
	A person holding a pen
Person + Action + Location	At least two persons in a hallway are seen walking
	An adult is sitting in a glass walled building
Person + Action + Object + Location	A seated person reading from a paper or book outdoors during daytime
	A woman wearing a silver necklace around her neck

5 Participating Teams

In total, 8 teams submitted 28 automatic runs and 1 manually-assisted run (where the query can be modified manually before submitting to the system for official results submission). These groups, representing a mixture of academic and industrial research labs, are listed in Table 2.

Table 2: Teams and Organizations

Team Identifier	Organization
ncsu-las	North Carolina State University, Laboratory for Analytic Sciences
NII_UIT	National Institute of Informatics; HCM-UIT
CERTH-ITI	Centre for Research and Technology Hellas
AFRL	Air Force Research Laboratory
ccilab	Co-creation Informatics Laboratory, Doshisha University
Meisei	Meisei University
WHU_NERCMS	Wuhan University
SZUAI	Shenzhen University Artificial Intelligence

These teams contributed multiple runs each, often testing variations in system design, model combinations, and scoring pipelines.

6 Evaluation Methodology

6.1 Pooling and Judgment

The evaluation followed the standard AVS protocol. For each query, all submitted ranked lists were pooled, and the top 300 results in all runs received a full human evaluation. Ranks 301–1000 were sampled at a rate of 25%.

Across all runs and topics:

- at least 22.6% of a system's retrieved shots were judged,
- in some cases, 100% were judged,
- the average judgment rate was 86.71%.

The judgments were binary: a shot either contained the described event or did not. Assessors watched the entire shot, including audio when present, to avoid misinterpreting ambiguous situations.

6.2 Quality Control

Judgments were re-evaluated in two circumstances: when a shot appeared in at least ten systems' top results yet was labeled as a false positive, and when neighboring shots (within a window of five) received inconsistent judgments despite appearing visually similar. These checks help stabilize the labeling process by reducing annotation noise.

6.3 Scoring

Systems were evaluated using extended inferred average precision (xinfAP)[1], computed through the official `sample_eval` tool[1]. Two sets of scores were produced—one based solely on 2024 judgments and another that combined 2024 and the new 2025 judgments.

7 Human Assessment

The evaluation relied on five human assessors who collectively contributed over 109,000 new judgments during the 2025 cycle. The total effort across 2024 and 2025 amounted to near 1000 hours of annotation work. In total, the combined ground truth of both years consisted of 164,843 judgments with 25,513 relevant video shots.

8 Results and Observations

8.1 Overall Performance

When the systems' final scores were sorted, the median xinfAP was 0.317. Scores spanned a wide range, reflecting both the variability in system design and the inherent difficulty of some queries. Figure 1 shows the overall run scores when evaluated using the 2024 ground truth as well as the combined ground truth. Majority of the runs scored higher based on the 2025 combined ground truth, except two teams (NII_UIT and CERTH_ITI) who had higher scores based on the 2024 ground truth with the exception of the manually-assisted run by NII_UIT (highlighted by red).

Figure 2 shows how the 2025 systems compare to 2024 systems when evaluated on the 2024 ground truth only. Results show that 12 systems in 2025 performed higher than the top 2024 systems signaling possible improvements in model learning and approaches. Roughly 36% of all true hits in the pooled results were unique to a single system.

8.2 Query-Level Behavior

Figure 3 shows the performance of the top 10 highest scores for each query. Some queries showed strong agreement across models when the visual cues were clear, or the described moment was distinctive enough that embeddings captured it reliably. Other queries revealed large gaps between systems or universal difficulty, often

[1] https://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/sample_eval/

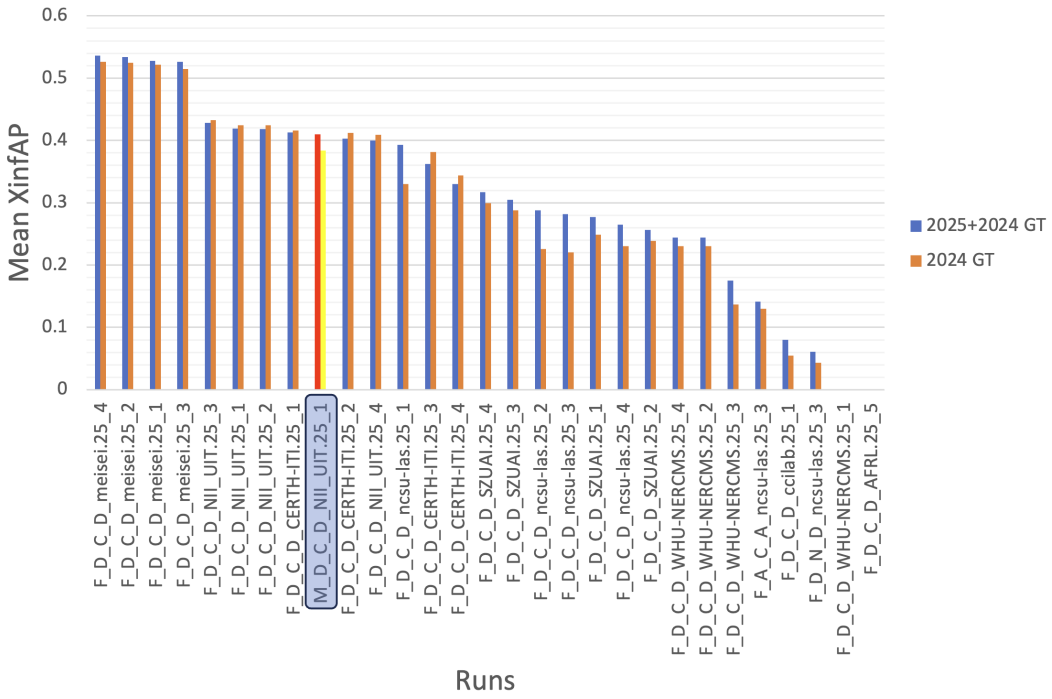

Figure 1: Overall system performance distribution.

tied to subtle actions or rare visual combinations. Multi-facet queries that require detection of both action and spatial context remain challenging.

8.3 Task Difficulty and Topic Characteristics

A simple yet informative method for analyzing difficulty is to count how many systems achieve a score above or below 0.5 for each query. Queries rich in visual cues—such as clothing attributes or recognizable indoor environments— tend to be easier. Queries involving interactions, partial occlusion, or less common actions remain substantially more difficult. The 5 easiest queries based on the 0.5 threshold were ”A man wearing a checked shirt”, ”A man inside a workshop”, ”A bald man with glasses”, ”A rainy day outdoors”, and ”A woman wearing a floral top or dress”. On the other hand, the 5 most difficult queries were ”A person is wiping himself or an object using their bare hands or other object”, ”Two women together wearing hats, excluding caps, outdoors”, ”A white sweater”, ”A man is putting on a jacket or a t-shirt”, and ”An adult is sitting in a glass walled building”.

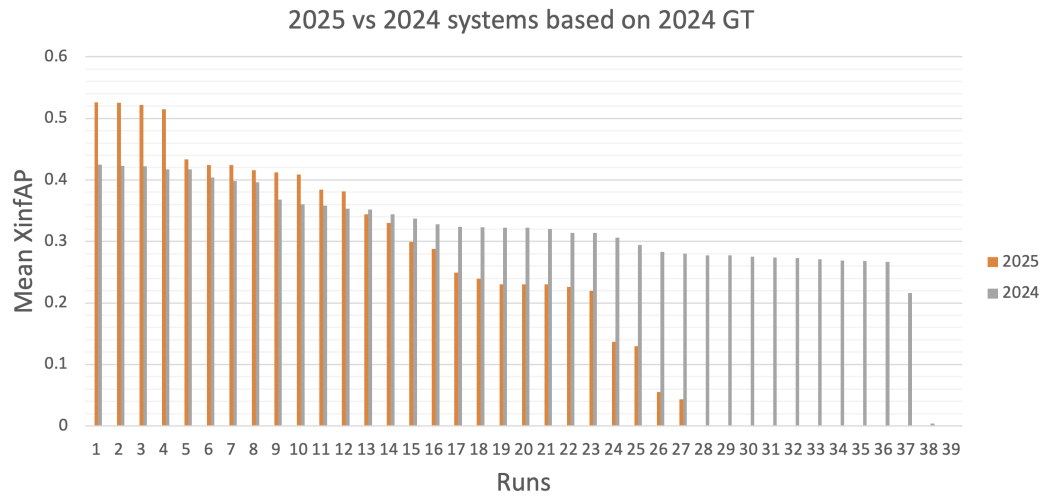

Figure 2: Comparing 2025 systems vs 2024 systems

9 Novelty Retrieval

The optional novelty task revisits the question: *Can a system retrieve relevant shots that other systems overlook?*

Each relevant shot receives a weight based on how frequently it appears across systems:

$$w = 1 - \frac{N}{M},$$

where N is the number of systems retrieving the shot and M is the total number of systems.

The novelty score for a run is the average of these weighted sums across all 20 queries.

We received 1 novelty run from north Carolina state university (solid red). For each team who didn't submit a novelty run, we select the best performed run for comparison purposes. Figure 4 shows that ncsu team best run scored higher than the rest and their own novelty run. These results indicates that the only novelty run submitted did not achieve the objective fully of retrieving rare relevant shots.

10 System Approaches in 2025

We summarize here the main trends of the system approaches that participated this year in the AVS track.

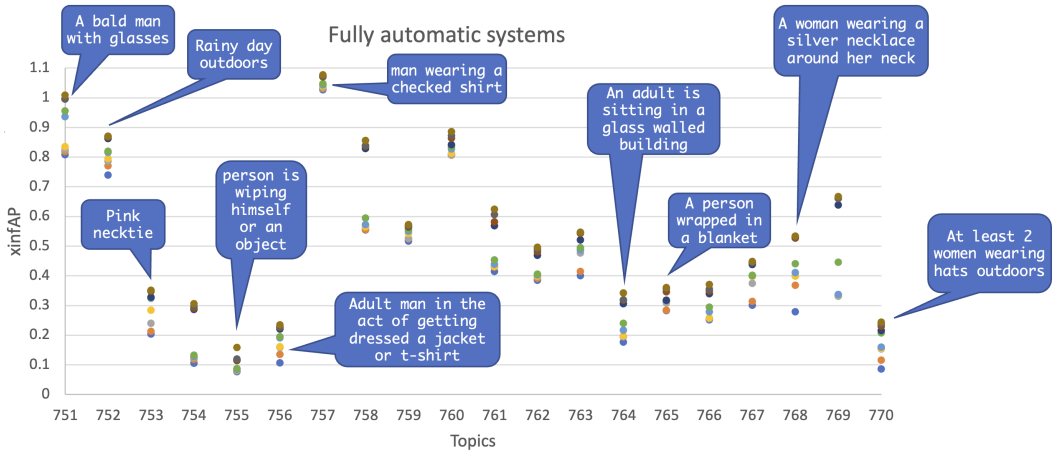

Figure 3: Top 10 scores per-query performance

10.1 Single-Stage Embedding Approaches

Some teams relied almost exclusively on modern text–image encoders such as CLIP. These models remain efficient and surprisingly strong for simple or highly visual queries.

10.2 Two-Stage Retrieval Pipelines

The dominant pattern involved a two-stage structure: an initial high-recall retrieval pass using CLIP, BLIP2, or SIGLIP2 models, followed by a more precise re-ranking or verification stage using large vision-language models such as GPT-4.1 or Phi-3 Vision. The second stage often helps filter out false positives involving ambiguous interactions or subtle semantic cues.

10.3 Model Fusion and Ensembles

Teams also experimented with model ensembles incorporating BEIT3, BLIP, BLIP2, CLIP variants, InternVL, LaCLIP, SLIP, and even diffusion models for semantic refinement. These systems attempt to combine multiple representation spaces to yield more resilient retrieval scores.

10.4 Fine-Grained Alignment

Approaches such as FG-CLIP highlight an increased interest in token-level alignment. These methods target nuanced details such as object–action interactions or

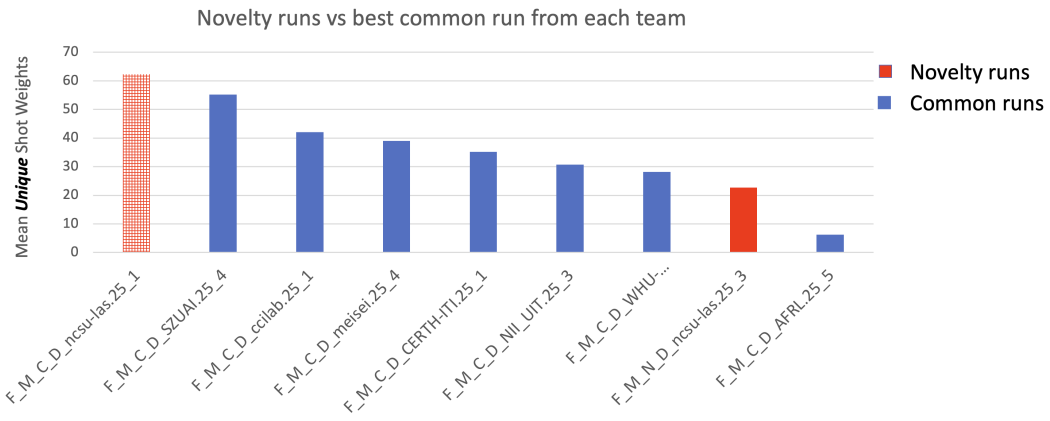

Figure 4: novelty scores

specific relational attributes that broader embeddings often gloss over.

10.5 Query Expansion

Large language models were frequently enlisted to paraphrase or decompose queries. These textual variations help capture alternative phrasing and latent attributes, giving systems a broader semantic footprint.

10.6 External Training Data

Use of external datasets was generally limited, with occasional references to MSR-VTT, but most systems leaned heavily on pretrained backbones and internal fusion strategies.

11 Conclusions

The 2025 AVS cycle reflects a field in active transition. Where systems once relied on engineered concept banks and heavy supervision, they now integrate rich, pretrained multimodal models capable of attending to a wide range of visual phenomena. The two-stage retrieval pattern of fast embedding-based recall followed by VLM-driven verification has emerged as a particularly effective strategy.

Despite these advances, challenges remain: queries involving uncommon combinations of facets or subtle interactions continue to highlight the limitations of current representations. Nonetheless, the diversity of retrieved hits and the steady participation of research teams underscore the vitality of the track.

AVS has produced 270 evaluated queries across its history, with 92 teams participating over the last decade. The V3C1 and V3C2 collections have supported this stable progress, and the unexplored V3C3 subset offers opportunities for future track development.

Disclaimer

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

References

[1] J. A. Aslam, V. Pavlu, and E. Yilmaz. Statistical Method for System Evaluation Using Incomplete Judgments. In *Proceedings of SIGIR*, 2006.

[2] L. Rossetto, H. Schuldt, G. Awad, and A. Butt. V3C: A Research Video Collection. In *Proceedings of the 25th International Conference on Multimedia Modeling*, 2019.