# Exploiting Temporal and Semantic Diversity: A Multi-Stage Retrieval–Reranking Pipeline for AVS 2025

Thuyen Tran Doan[1,2], Bao Tran[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.** With the explosive growth in video content and volume, efficient video retrieval systems have become increasingly essential. However, our system still underperforms on queries involving temporal or action-related information. This limitation stems from the reliance on Text-to-Image (T2I) retrieval models, such as BEiT and BLIP, whose architectures are inherently image-based. In contrast, Text-to-Video (T2V) retrieval models, such as CLIP4Clip and TS2Net, are built upon pretrained backbones like CLIP and incorporate simple yet effective temporal modeling mechanisms, which enhance the system's ability to understand temporal aspects in textual queries. For our participation in the TRECVID 2025 Ad-hoc Video Search (AVS) task, we have integrated several T2V models into both the initial retrieval stage and the fusion step, in addition to the existing T2I models. This integration not only boosts the overall average precision (AP) score but also improves system diversity and recall. To further leverage the increased recall, we employ a reranking step using several Large Vision-Language Models (LVLMs). These models, equipped with advanced reasoning capabilities, can better interpret complex or ambiguous query elements, such as exclusion terms, that are often challenging for smaller T2I/T2V models to handle effectively. Evaluated on the AVS 24 and 25 main tasks, our system achieves xinfAP scores of **0.4334** and **0.4361**, respectively, demonstrating the effectiveness of combining diverse T2V models with multi-VLLM reranking.

## 1 Introduction

The TRECVID Ad-hoc Video Search (AVS) workshop series evaluates the effectiveness of various techniques applied to video retrieval tasks. Building upon our previous system, the NII_UIT team actively participated in this year's AVS task, submitting entries for both the fully automatic and manually assisted retrieval tracks in the main task.*

At its core, the AVS task is a text-to-video retrieval challenge. As such, many participating systems (e.g., [25,10,21]) typically adopt Vision-Language Models (VLMs) such as Contrastive Language–Image Pre-training (CLIP), along with its variants like BLIP-2 [9] and BEiT [20], which are originally trained for Text-to-Image (T2I) retrieval (and its inverse). These T2I models are generally trained using a contrastive learning scheme, wherein paired image–text data are projected into a shared embedding space.

However, as highlighted in [2], these models tend to underperform when handling queries that include temporal or action-related cues. This limitation arises from their inherently image-based design: they process individual frames independently, lacking the ability to capture temporal dynamics across video sequences. To address this issue, Text-to-Video (T2V) retrieval models, such as CLIP4Clip [15], TS2Net, and TeachCLIP, have adapted CLIP for video–text retrieval tasks on datasets such as MSVD [4], MSR-VTT [23], ..., by integrating temporal modelling mechanisms. These temporal mechanisms often involve neural architectures such as Transformer encoders [1], enabling the models to capture sequence-level temporal dependencies critical for effective video understanding.

Therefore, in our participation in the TRECVID 2025 AVS task, we integrated T2V models into both the initial retrieval stage and the fusion step, in addition to the existing T2I models used in previous systems. This integration led to an increase in the average precision ($AP$) score at the initial retrieval stage.

Furthermore, we hypothesise that in AVS 24, the retrieved shots produced by T2I-based models were largely duplicated (see Tab. 1), resulting in a limited number of unique relevant shots, only **1861** in total. While this may not initially seem detrimental, the high degree of duplication, evidenced by a total of **84,937** retrieved shots with only **16,701** relevant ones, indicates a lack of diversity in the retrieval results. This redundancy significantly reduces

*The AVS 25 main task uses the same queries as AVS 24. No progress task is included this year, as the three-split evaluation (2022–2024) has already been completed.

Table 1: Number of unique and overlapping retrieved shots (along with their corresponding relevant counts) among top-performing T2I-only methods on AVS 2023 queries. The final row, *Fusion*, includes more VLMs beyond BEiT-3 and InternVL, but for brevity, we report results using these representative models only. This fusion setup achieved an AP of **0.2850** on AVS 2023.

| Query ID | Models | #Ret. shots that unique | #Ret. shots that unique & rel. | #Ret. shots that overlap | #Ret. shots that overlap & rel. |
|---|---|---|---|---|---|
| 1731 | BEiT-3 | 1587 | 56 | 1913 | 703 |
| | InternVL | 1412 | 57 | 2088 | 717 |
| 1732 | BEiT-3 | 961 | 51 | 2539 | 751 |
| | InternVL | 836 | 59 | 2664 | 786 |
| 1733 | BEiT-3 | 1038 | 2 | 2462 | 240 |
| | InternVL | 1183 | 27 | 2317 | 258 |
| 1734 | BEiT-3 | 1085 | 52 | 2415 | 667 |
| | InternVL | 860 | 52 | 2640 | 696 |
| 1735 | BEiT-3 | 1833 | 23 | 1667 | 378 |
| | InternVL | 1409 | 71 | 2091 | 439 |
| 1736 | BEiT-3 | 1505 | 7 | 1995 | 213 |
| | InternVL | 1195 | 19 | 2305 | 246 |
| 1737 | BEiT-3 | 715 | 78 | 2785 | 1243 |
| | InternVL | 631 | 60 | 2869 | 1205 |
| 1738 | BEiT-3 | 1376 | 14 | 2124 | 210 |
| | InternVL | 1175 | 14 | 2325 | 218 |
| 1739 | BEiT-3 | 1646 | 8 | 1854 | 166 |
| | InternVL | 1349 | 10 | 2151 | 190 |
| 1740 | BEiT-3 | 1430 | 41 | 2070 | 595 |
| | InternVL | 1452 | 129 | 2048 | 649 |
| 1741 | BEiT-3 | 2014 | 10 | 1486 | 126 |
| | InternVL | 1658 | 87 | 1842 | 199 |
| 1742 | BEiT-3 | 1589 | 1 | 1911 | 77 |
| | InternVL | 1338 | 7 | 2162 | 83 |
| 1743 | BEiT-3 | 2621 | 2 | 879 | 2 |
| | InternVL | 1646 | 6 | 1854 | 6 |
| 1744 | BEiT-3 | 1807 | 17 | 1693 | 156 |
| | InternVL | 1630 | 51 | 1870 | 192 |
| 1745 | BEiT-3 | 1329 | 344 | 2171 | 997 |
| | InternVL | 1189 | 74 | 2311 | 1056 |
| 1746 | BEiT-3 | 1639 | 6 | 1861 | 182 |
| | InternVL | 1324 | 6 | 2176 | 199 |
| 1747 | BEiT-3 | 1670 | 0 | 1830 | 127 |
| | InternVL | 1118 | 53 | 2382 | 175 |
| 1748 | BEiT-3 | 1314 | 17 | 2186 | 194 |
| | InternVL | 1074 | 42 | 2426 | 247 |
| 1749 | BEiT-3 | 1545 | 145 | 1955 | 875 |
| | InternVL | 1341 | 112 | 2159 | 912 |
| 1750 | BEiT-3 | 1274 | 11 | 2226 | 150 |
| | InternVL | 1265 | 40 | 2235 | 176 |
| Total | Fusion | 55063 | 1716 | 27514 | 7235 |

the diversity of the candidate pool, thereby limiting the number of distinct shots that can be judged or re-ranked in later stages. As a consequence, the overall effectiveness of the system may be constrained.

By incorporating T2V models, we were able to retrieve a more diverse set of shots, thereby improving recall diversity. This, in turn, benefits the reranking stage, which can then operate on a richer and more representative set of candidate shots, potentially enhancing overall system performance (see Tab. 2) Specifically, the number of *#unique and relevant* shots increases from **1716** to **1821**, while the *#overlapping and relevant* shots decreases from **7235** to **4888**. This reduction in redundancy provides more room for the LLM-based reranker, which can only process a limited subset of candidates, to more effectively filter out irrelevant shots and promote relevant ones, ultimately improving overall performance.

In the following sections, Section 2 will detail the approach used for each module within our pipeline, while Section 3 will cover the experimental setup and findings. Finally, Section 4 will provide a summary of our approach and outline potential directions for future work.

## 2 Methodology

### 2.1 Embedding-based Search with Query expansion:

**Embedding-based Search using VLMs** Recently, embedding-based search models have significantly transformed image retrieval tasks, dramatically improving how visual information is accessed and processed. By converting images and text into high-dimensional embeddings, these models effectively capture complex relationships and semantic nuances across modalities. To enhance performance and precision, our retrieval system integrates several cutting-edge text-image models, harnessing their unique strengths to deliver more accurate and relevant search results.

- **InternVL-G** [5]: We utilize the InternVL-14B-224px model, a VLLM renowned for its strong performance in image retrieval tasks. This model generates 768-dimensional embeddings, contributing to the system's robust capabilities;
- **BEiT-3** [20]: Selected for its superior ability to capture complex multimodal relationships, BEiT-3 significantly enhances retrieval precision by producing detailed 1024-dimensional embeddings;

Table 2: Number of unique and overlapping retrieved shots (along with their corresponding relevant counts) among top-performing T2I methods (BEiT-3 and InternVL) and T2V methods (CLIP4Clip and TeachCLIP) on AVS 2023 queries. The final row, *Fusion*, includes more VLMs beyond BEiT-3, InternVL CLIP4Clip and TeachCLIP, but for brevity, we report results using these representative models only. This fusion setup achieved an AP of **0.3218** on AVS 2023.

| Query ID | Models | #Ret. shots that unique | #Ret. shots that unique & rel. | #Ret. shots that overlap | #Ret. shots that overlap & rel. |
|---|---|---|---|---|---|
| 1731 | BEiT-3 | 1504 | 16 | 1996 | 743 |
| | InternVL | 1350 | 12 | 2150 | 762 |
| | TeachCLIP | 1540 | 84 | 1960 | 787 |
| | CLIP4Clip | 1733 | 48 | 1767 | 693 |
| 1732 | BEiT-3 | 1028 | 10 | 2472 | 792 |
| | InternVL | 850 | 23 | 2650 | 822 |
| | TeachCLIP | 1296 | 42 | 2204 | 807 |
| | CLIP4Clip | 1516 | 43 | 1984 | 744 |
| 1733 | BEiT-3 | 1115 | 3 | 2385 | 239 |
| | InternVL | 1215 | 14 | 2285 | 271 |
| | TeachCLIP | 1595 | 2 | 1905 | 239 |
| | CLIP4Clip | 1961 | 2 | 1539 | 226 |
| 1734 | BEiT-3 | 1284 | 62 | 2216 | 657 |
| | InternVL | 854 | 35 | 2646 | 713 |
| | TeachCLIP | 1318 | 49 | 2182 | 619 |
| | CLIP4Clip | 1282 | 62 | 2218 | 657 |
| 1735 | BEiT-3 | 1786 | 8 | 1714 | 393 |
| | InternVL | 1377 | 26 | 2123 | 484 |
| | TeachCLIP | 2194 | 24 | 1306 | 371 |
| | CLIP4Clip | 2270 | 16 | 1230 | 317 |
| 1736 | BEiT-3 | 1725 | 6 | 1775 | 214 |
| | InternVL | 1301 | 8 | 2199 | 257 |
| | TeachCLIP | 1568 | 2 | 1932 | 192 |
| | CLIP4Clip | 1565 | 5 | 1935 | 206 |
| 1737 | BEiT-3 | 905 | 73 | 2595 | 1248 |
| | InternVL | 742 | 34 | 2758 | 1231 |
| | TeachCLIP | 1268 | 125 | 2232 | 1125 |
| | CLIP4Clip | 1498 | 112 | 2002 | 1010 |
| 1738 | BEiT-3 | 1447 | 4 | 2053 | 220 |
| | InternVL | 1220 | 3 | 2280 | 229 |
| | TeachCLIP | 1616 | 21 | 1884 | 255 |
| | CLIP4Clip | 1737 | 18 | 1763 | 228 |
| 1739 | BEiT-3 | 1683 | 8 | 1817 | 166 |
| | InternVL | 1308 | 10 | 2192 | 190 |
| | TeachCLIP | 1993 | 3 | 1507 | 122 |
| | CLIP4Clip | 1958 | 2 | 1542 | 129 |
| 1740 | BEiT-3 | 1881 | 69 | 1619 | 567 |
| | InternVL | 1442 | 78 | 2058 | 700 |
| | TeachCLIP | 2135 | 61 | 1365 | 447 |
| | CLIP4Clip | 2288 | 31 | 1212 | 374 |
| 1741 | BEiT-3 | 2197 | 8 | 1303 | 128 |
| | InternVL | 1545 | 46 | 1955 | 240 |
| | TeachCLIP | 2172 | 20 | 1328 | 199 |
| | CLIP4Clip | 2250 | 17 | 1250 | 166 |
| 1742 | BEiT-3 | 1772 | 1 | 1728 | 77 |
| | InternVL | 1423 | 2 | 2077 | 88 |
| | TeachCLIP | 1815 | 1 | 1685 | 76 |
| | CLIP4Clip | 2006 | 0 | 1494 | 68 |
| 1743 | BEiT-3 | 2947 | 0 | 553 | 4 |
| | InternVL | 1828 | 4 | 1672 | 8 |
| | TeachCLIP | 2427 | 2 | 1073 | 4 |
| | CLIP4Clip | 2589 | 3 | 911 | 3 |
| 1744 | BEiT-3 | 1921 | 4 | 1579 | 169 |
| | InternVL | 1388 | 5 | 2112 | 238 |
| | TeachCLIP | 2004 | 43 | 1496 | 212 |
| | CLIP4Clip | 2210 | 35 | 1290 | 192 |
| 1745 | BEiT-3 | 1556 | 339 | 1944 | 1002 |
| | InternVL | 1560 | 130 | 1940 | 1000 |
| | TeachCLIP | 1302 | 62 | 2198 | 913 |
| | CLIP4Clip | 1388 | 142 | 2112 | 974 |
| 1746 | BEiT-3 | 1859 | 0 | 1641 | 188 |
| | InternVL | 1450 | 0 | 2050 | 205 |
| | TeachCLIP | 1867 | 3 | 1633 | 200 |
| | CLIP4Clip | 2096 | 1 | 1404 | 177 |
| 1747 | BEiT-3 | 1984 | 0 | 1516 | 127 |
| | InternVL | 1004 | 33 | 2496 | 195 |
| | TeachCLIP | 1745 | 2 | 1755 | 136 |
| | CLIP4Clip | 1911 | 4 | 1589 | 129 |
| 1748 | BEiT-3 | 1647 | 18 | 1853 | 193 |
| | InternVL | 1303 | 26 | 2197 | 263 |
| | TeachCLIP | 1658 | 15 | 1842 | 186 |
| | CLIP4Clip | 1827 | 12 | 1673 | 147 |
| 1749 | BEiT-3 | 1550 | 101 | 1950 | 919 |
| | InternVL | 1200 | 67 | 2300 | 957 |
| | TeachCLIP | 1553 | 54 | 1947 | 873 |
| | CLIP4Clip | 1907 | 41 | 1593 | 681 |
| 1750 | BEiT-3 | 1475 | 12 | 2025 | 149 |
| | InternVL | 1184 | 12 | 2316 | 204 |
| | TeachCLIP | 1626 | 6 | 1874 | 156 |
| | CLIP4Clip | 1943 | 8 | 1557 | 147 |
| Total | Fusion | 131437 | 1821 | 12004 | 4888 |

– **BLIP-2** [9]: For image retrieval tasks, we employ BLIP-2, which has been fine-tuned on the COCO [12], further enriching our model selection;
– **CLIP** [17]: Our approach includes two variants of the CLIP-H/14 model, trained on the DataComp [7] and LAION-2B [18] datasets, as well as a CLIP-L/14 model from DataComp. Additionally, we incorporate CLIP-RN101, all sourced from OpenCLIP, to provide a diverse range of embedding outputs.

For T2V methods, we explored the following methods:

– **CLIP4Clip** [15]: This approach builds upon the CLIP pretraining framework, incorporating a temporal transfer module to enhance inference for T2V retrieval tasks. As the authors did not release pretrained weights, we trained

Table 3: Main queries of AVS 2023.

| Query ID | Original Query |
|---|---|
| 1731 | A man is seen with a baby |
| 1732 | A woman with red hair |
| 1733 | A golf course |
| 1734 | A recording studio |
| 1735 | A toy vehicle |
| 1736 | A person opens a door and enters a location |
| 1737 | A woman wearing (dark framed) glasses |
| 1738 | A police officer wearing a helmet |
| 1739 | Two or more persons are seen in front of a chain link fence |
| 1740 | A heavy man indoors |
| 1741 | A red or blue scarf around someone's neck |
| 1742 | A child climbs an object outdoors |
| 1743 | A man is talking in a small window located in the lower corner of the screen |
| 1744 | A person taking a picture using a cell phone camera |
| 1745 | A person wearing gloves while biking |
| 1746 | A man riding a scooter |
| 1747 | At least two persons are working on their laptops together in the same room indoors. |
| 1748 | A man carrying a bag on one of his shoulders (excluding backpacks) |
| 1749 | A person wearing any kind of face or head mask |
| 750 | A man with an earring in his left ear |

Table 4: Main queries of AVS 24 and AVS 25.

| Query ID | Original Query |
|---|---|
| 1751 | A bald man with glasses |
| 1752 | A rainy day outdoors |
| 1753 | A pink necktie |
| 1754 | A white sweater |
| 1755 | A person is wiping themselves or an object using their bare hands or other object. |
| 1756 | A man is putting on a jacket or a t-shirt |
| 1757 | A man wearing a checked shirt |
| 1758 | A woman wearing a floral top or dress |
| 1759 | People inside an airport terminal |
| 1760 | A man inside a workshop |
| 1761 | A traffic light seen at an intersection of a road or street |
| 1762 | A map seen on a wall indoors |
| 1763 | At least two persons in a hallway are seen walking |
| 1764 | An adult is sitting in a glass walled building |
| 1765 | An adult is wrapped in a blanket |
| 1766 | A person holding a pen |
| 1767 | A seated person reading from a paper or book outdoors during daytime |
| 1768 | A woman wearing a silver necklace around her neck |
| 1769 | Two or more persons indoors with coffee cups or mugs seen with them. |
| 1770 | Two women together wearing hats, excluding caps, outdoors |

the model on the MSR-VTT dataset using the ViT-32 variant (see CLIP). The resulting model produces 512-dimensional embeddings;

- **TS2Net** [13]: This method extends CLIP4Clip by incorporating a "shifting module" that shifts feature representations along the temporal axis. We employ the ViT-32 variant provided by the TeachCLIP authors, with the resulting model producing 512-dimensional embeddings;
- **XCLIP** [16]: To enhance video understanding, we integrate XCLIP, pre-trained on Kinetics-600 [3], which extracts 768-dimensional embeddings and bolsters our retrieval system's performance;
- **TeachCLIP** [19]: This method extends the student–teacher training paradigm by employing pretrained teachers such as TS2Net, XPool, .... However, as the authors only released the pretrained teacher models, we adopted

the same training configuration as used for CLIP4Clip, employing XCLIP and XPool as teachers. The resulting model produces 512-dimensional embeddings;
– **Side4Video** [22]: This method attaches a lightweight spatial–temporal side network to a frozen image backbone instead of inserting adapters (which have been introduced NLP tasks) inside the model. This side branch uses multi-level features plus lightweight 3D convolutions and token-shift attention to capture temporal information efficiently, enabling video-aware embeddings with minimal memory cost. Side4Video produces 512-dimensional embeddings.

**Query Expansion** Query expansion is an essential component of the Information Retrieval (IR) module, designed to improve the clarity and expressiveness of user queries. AVS queries are often underspecified, frequently consisting of only a few nouns (e.g., Query #1751, 1753, 1754; see Tab. 4), which limits retrieval effectiveness. To mitigate this issue, we incorporate a query expansion strategy that enriches the semantic content of these queries.

Following our 2024 approach, we use GPT-4 to generate $N$ paraphrased variants for each ambiguous query, producing multiple ranked lists that reflect different interpretations. These lists are then averaged to capture the collective semantics of the expanded queries. To retain the influence of the original query, we fuse the aggregated expanded results with the initial ranked list using a weighted combination, assigning a weight of $0.1/N$ to each paraphrase and 0.9 to the original query. This weighting scheme provides a stable balance between the expanded semantic context and the user's original intent, resulting in more accurate and robust retrieval performance.

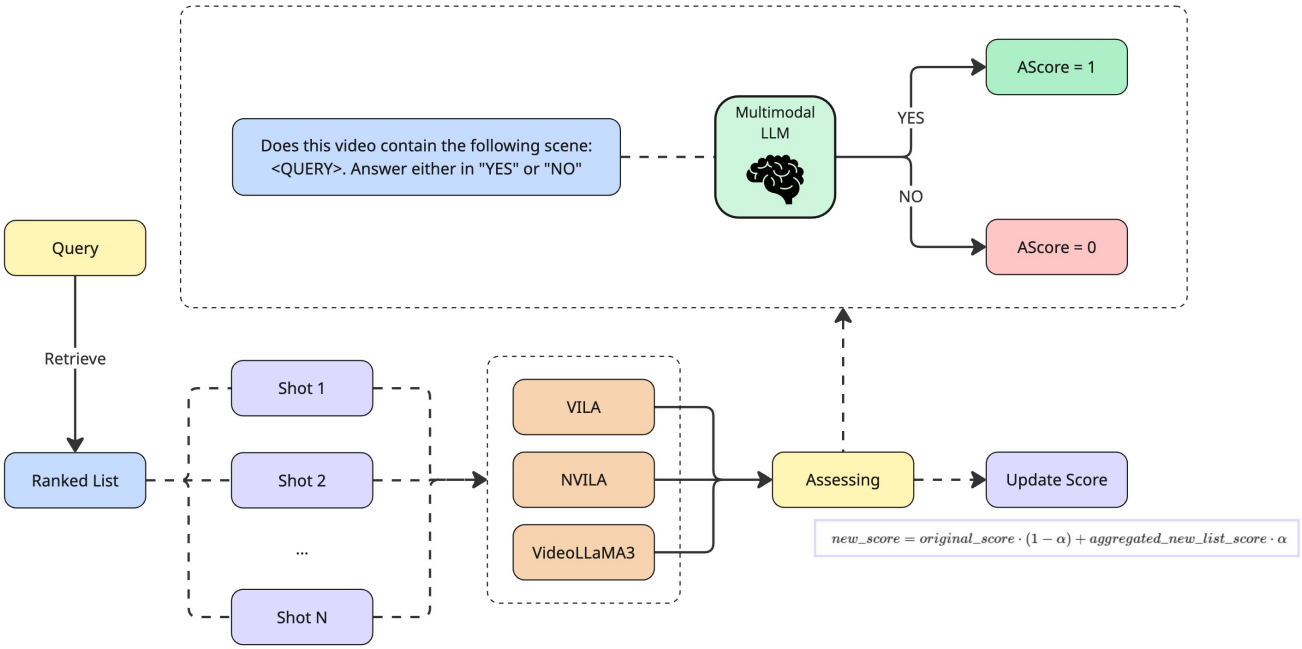


Fig. 1: Overview of our 2025 reranking pipeline, which integrates multiple VLLMs to improve answer robustness. By aggregating predictions across models, the system establishes a stronger consensus when handling complex queries, such as Query #1743 ("A man is talking in a small window located in the lower corner of the screen"), leading to more reliable and accurate reranking performance.

**Fusion** combines the outputs of multiple IR models to exploit their complementary strengths. In our system, fusion is applied both across models and within each model by merging results from the original and expanded queries. We first generate two ranked lists, one from the original query and one from its expanded variants, then integrate them using the CombMNZ [6] algorithm, which rewards agreement across lists while preserving score differences. This approach, consistent with our 2024 pipeline, yields a richer and more robust representation by capturing diverse semantic cues from different models and query formulations.

### 2.2 Reranking

A reranking module serves as a second-pass reordering mechanism within a ranked list. One promising approach is to use Video Large Language Models (VLLMs), which have shown strong performance on various video understanding tasks. In our 2024 pipeline, we employed VILA [11] as a single reranker. However, as shown in Fig. 2, the fitted regression line reveals a mild negative correlation, indicating that greater model diversity is associated

with larger ensemble gains. Specifically, queries with lower IoU scores (i.e., higher disagreement among models) tend to achieve larger improvements in xinfAP over VILA. This finding supports the hypothesis that diversity among VLLMs strengthens ensemble effectiveness. Motivated by this observation, our 2025 AVS pipeline integrates multiple complementary VLLMs in the reranking stage to further enhance overall accuracy. The VLLMs used are as follows:

- VILA-40B [11]: The largest VLLM used in our pipeline (40B parameters). VILA leverages *in-context learning*, enabling the model to infer new tasks or behaviors directly from examples provided in the prompt. This capability helps it adapt effectively to complex constraints in AVS queries;
- NVILA-15B [14]: Its core innovation is a "*scale-then-compress*" design: the model first increases spatial and temporal resolution to preserve fine-grained visual details, and then compresses visual tokens to reduce computation. This enables NVILA to process high-resolution images and long videos efficiently, reducing training cost and inference latency by up to 2–5 times.
- VideoLLaMA3-7B [24]: Architecturally, VideoLLaMA3 has *adaptive vision token design* i.e variable number of tokens, similarity-based token reduction for videos;
- Aria-8x3.5B [8]: ARIA is an open multimodal Mixture-of-Experts (MoE) model capable of processing text, images, and video within a unified framework. It also supports *long-context inputs of up to 64K tokens*, making it well suited for handling extended long video content.

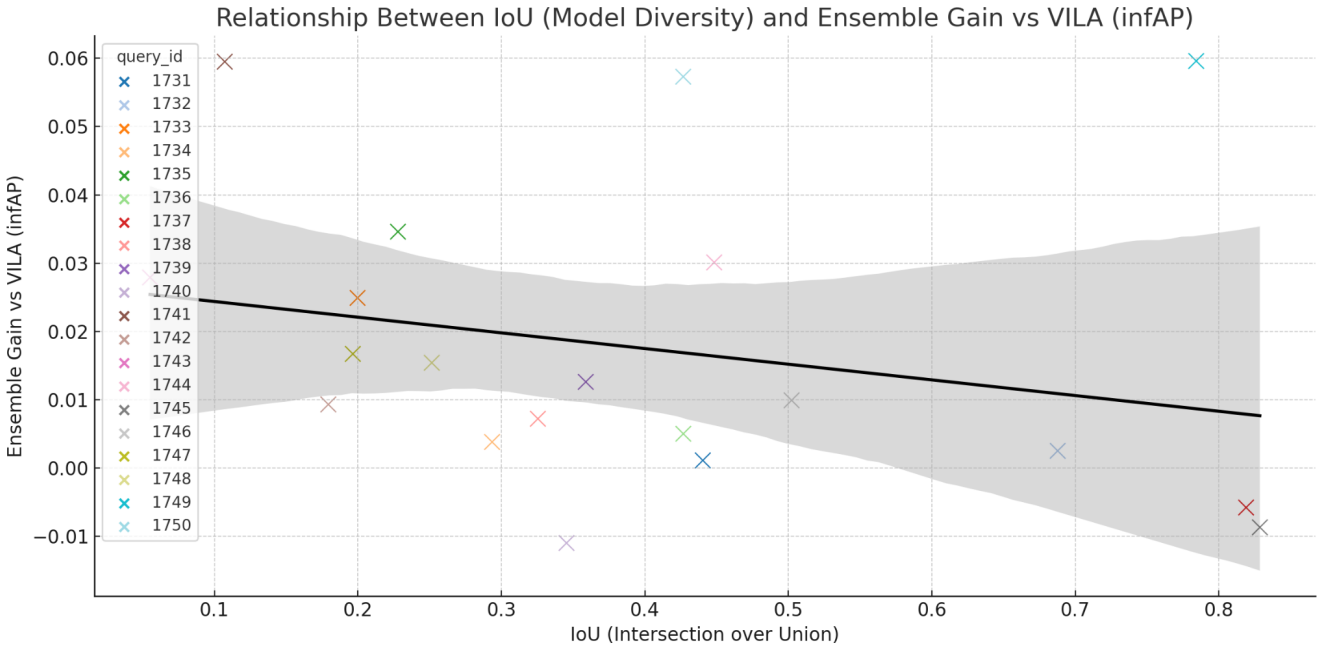


Fig. 2: Relationship between model diversity and ensemble gain over VILA on AVS 2023 queries. Each point corresponds to a query, with the x-axis showing the IoU (Intersection over Union) of relevant shots between VILA and the other rerankers (lower IoU indicates higher diversity), and the y-axis showing the xinfAP gain of the ensemble compared to VILA alone.

The final score for each shot is calculated as follows:

$$\text{Final Score} = \text{Initial Score} \times (1 - \alpha) + \text{Relevant} \times \alpha$$

where:

- **Final Score** represents the adjusted score after reranking
- **Initial Score** is the original score assigned to each shot during fusion
- **Relevant** is a binary value (1 if the shot is relevant to the query, 0 otherwise)
- $\boldsymbol{\alpha}$ is a hyperparameter that controls the influence of the reranking module on the final score

This weighting approach allows us to balance the original fused score with the relevance assessment, ensuring that shots closely aligned with the query are prioritized in the final ranked results.

Table 5: Model performance (T2I/T2V): Comparison of model performance on original vs. expanded queries for TRECVid AVS 2023 and 2024 queries. Abbreviations: DC = DataComp, L2B = Laion2B, DF = DFN5b. Fusion applies a weight of 4 to InternVL–G, BEiT–3, and BLIP–2, while all other models are assigned a weight of 1. Results are reported separately for T2I† models and T2V* models.

| **Models** | **xinfAP** | | |
|---|---|---|---|
| | **AVS 23** | **AVS 23 QE** | **AVS 25** |
| InternVL-G† | 0.2400 | 0.2531 | 0.2567 |
| BEiT-3† | 0.2160 | 0.2211 | 0.2370 |
| BLIP-2† | 0.2057 | 0.2064 | 0.1896 |
| CLIP-L/14 (DC)† | 0.1019 | 0.1225 | 0.1447 |
| CLIP-H/14 (L2B)† | 0.1534 | 0.1696 | 0.1882 |
| CLIP-H/14 (DF)† | 0.1556 | 0.1674 | 0.2140 |
| CLIP-RN101† | 0.0915 | 0.1113 | 0.0976 |
| XCLIP* | 0.0523 | 0.0579 | 0.0356 |
| Fusion (2024 pipeline) | - | 0.3043 | - |
| CLIP4Clip* | 0.1215 | 0.1254 | 0.1206 |
| TeachCLIP* | 0.1795 | 0.1810 | 0.1713 |
| TS2Net* | 0.1332 | - | 0.1404 |
| Side4Video* | 0.1826 | - | 0.1896 |
| Fusion (2025 pipeline) | - | 0.3215 | 0.2916 |

Table 6: Additional T2I/T2V: Results on AVS 23 are reported for other T2I† models and T2V* models.

| **Models** | **xinfAP** |
|---|---|
| | **AVS 23** |
| EVA-L336† | 0.1688 |
| EVA-BigE† | 0.1909 |
| MugStan* | 0.1299 |

## 3 Experiments results

In this section, we report the evaluation results of each module on AVS 23 (see Tab. 3), and AVS 24/25 (see Tab. 4); note that TRECVid AVS 24 and 25 share the same query set. Note that Tab. 5 presents results for AVS 25 but not for AVS 24, whereas Tab. 9 includes both AVS 24 and AVS 25. This is because the organizers evaluate the same submission runs for both years, as shown in Tab. 9, but do not release the AVS 24 ground-truth annotations. As a result, we are able to evaluate the retrieval stage only on AVS 25 (Tab. 5).

### 3.1 Embedding-based Search with Query expansion

We evaluated our retrieval system using the AVS 23 queries and the AVS 24/25 queries. The corresponding results of this module are presented in Tab. 5, Tab. 6 and Tab. 7.

Particularly, in Tab. 5, each VLM of the T2I approach and T2V approach performs very well on both query sets, except for XCLIP, which showed the lowest scores of 0.0523 and 0.0356. Furthermore, with the addition of the Query Expansion module, each method's performance improved significantly, particularly for lower-scoring models.

Table 7: Ablation study: step-wise addition of T2I and T2V models on AVS 2023 queries, evaluated at different TOP-K settings. Abbreviations: TE = TeachCLIP, C4 = CLIP4Clip, IT = InternVideo2, EB = EVA-BigE, E3 = EVA-L336, MS = MugSTAN, TS = TS2Net, S4 = Side4Video. The last row shows the configuration used for the Fusion (2025 pipeline) reported in Tab. 5.

| **Methods** | **TOPK = 1000** | | **TOPK = 3500** | |
|---|---|---|---|---|
| | AP | #Rel. shot retrieved | AP | #Rel. shot retrieved |
| Fusion (2024 pipeline) | 0.3043 | 7045 | 0.2786 | 13425 |
| + C4 | 0.3118 | 7272 | 0.2868 | 13752 |
| + TE | 0.3215 | 7386 | 0.2938 | 13857 |
| + TE + C4 | 0.3209 | 7376 | 0.2972 | 14103 |
| + TE + C4 + EB | 0.3231 | 7387 | 0.2987 | 14055 |
| + TE + C4 + E3 | 0.3227 | 7389 | 0.2993 | 14083 |
| + TE + C4 + MS | 0.3175 | 7321 | 0.2949 | 14166 |
| + TE + C4 + TS | 0.3218 | 7404 | 0.3001 | 14230 |
| + TE + C4 + TS + S4 | 0.3215 | 7434 | 0.3004 | 14314 |

Notably, models that initially performed well on the original query, such as InternVL-G and BEiT-3, seemed to reach a performance plateau on these test sets, as indicated by their limited improvement when compared to the results from the Query Expansion module. This increase demonstrates the power of Query Expansion in enhancing retrieval precision, particularly for ambiguous or sparse queries.

Additional T2I and T2V model performances are reported in Tab. 6, but these methods are not included in our 2025 pipeline despite showing strong results on the AVS 2023 queries. As shown in Tab. 7, integrating these models into the 2024 pipeline (+ TE + C4 + EB/E3/MS) leads to lower AP at $TOP-K = 3500$ and fewer *#Rel. shot retrieved* compared to the configuration using + TE + C4 + EB + TS. This reduction in relevant shots is harmful for the reranking stage.

Finally, following our experiments on AVS 2023, we assign a weight of 1 to most models to maintain balanced contributions, while giving higher weights of 4 to the top-performing models, InternVL-G, BEiT-3, and BLIP-2, to better exploit their superior retrieval accuracy (see Tab. 7). This weighting scheme allows the system to capture fine-grained cues from a diverse set of models while emphasizing the high-precision outputs of the strongest ones, resulting in a more robust and effective fused representation. With this module, we achieve an xinfAP of **0.3215** on AVS'23. From the AP and #Rel. shot retrieved results at $TOP\text{-}K = 3500$, it is evident that the gains from adding more models at the retrieval stage have largely plateaued (0.0003+ in AP and 83+ in relevant shots). With the candidate pool nearing saturation, additional retrieval combinations offer minimal benefit. Therefore, we decided to halt further retrieval-stage experiments and focus our resources on enhancing the reranking step, where improvements are more impactful.

### 3.2 Reranking

After fusing the ranked shot lists from individual models on both the original and expanded queries, we applied reranking exclusively within the top 4000 shots to further refine the output (see Tab. 8). This step prioritizes highly relevant shots by exploiting subtle differences in relevance that may only become apparent within this top-ranked subset.

However, a single VLLM may occasionally fail to correctly interpret certain queries. To mitigate this issue, we integrate multiple VLLM-based rerankers and fuse their outputs during the reranking stage. This ensemble approach yields consistently better overall performance (see Tab. 9). This observation is also consistent with the empirical relationship between AP and IoU shown in Fig. 2 and Fig. 3.

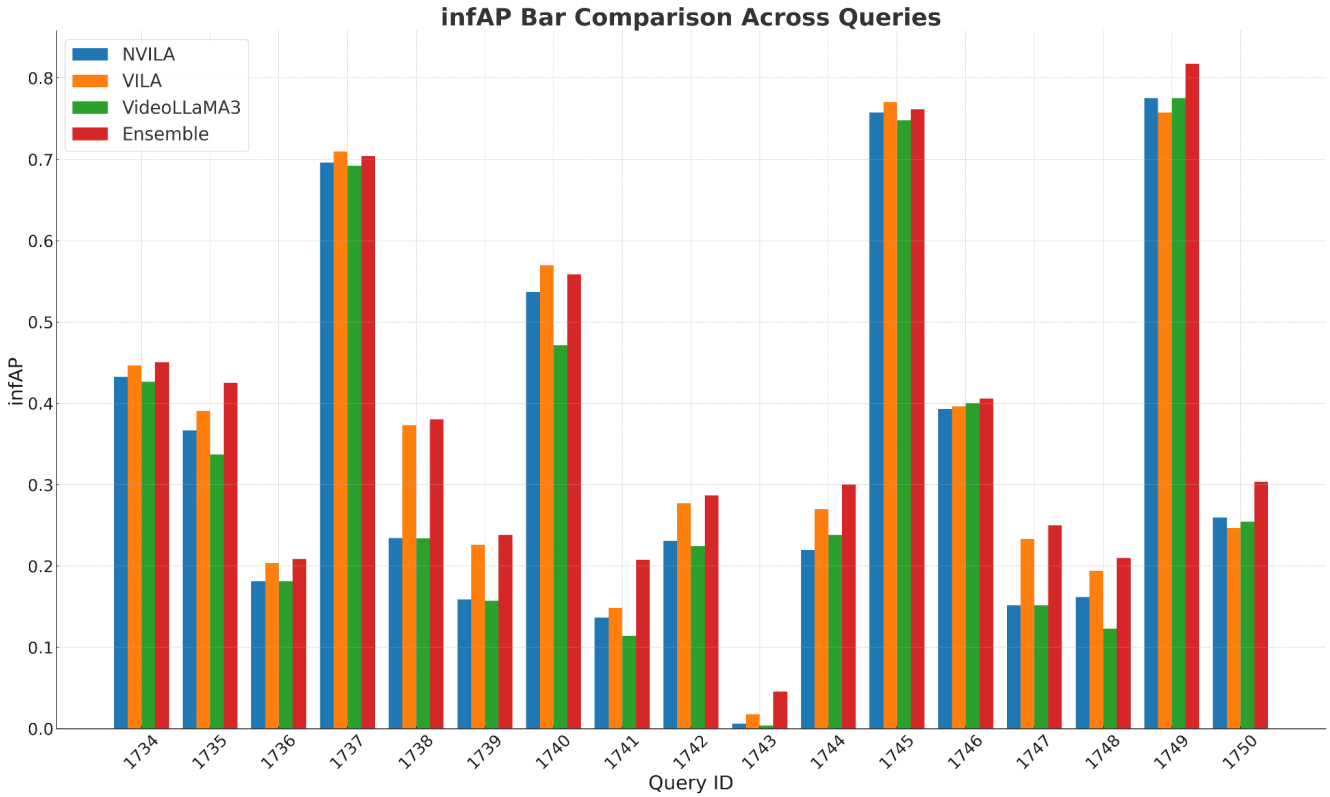


Fig. 3: infAP scores when using an ensemble of multiple VLLMs across queries. Notably, Query #1743 shows a substantial improvement compared to individual models, further demonstrating the benefit of integrating multiple VLLMs for reranking.

Table 8: Reranking (single VLLM): Evaluation results on AVS 23 queries for the reranking stage with a separate VLLM model.

| **Methods** | **AVS 23** |
|---|---|
| Fusion (2024 pipeline) | 0.3043 |
| + VILA-40B | 0.3840 |
| Fusion (2025 pipeline) | 0.3215 |
| + VILA-40B | 0.4035 |
| + NVILA-15B | 0.3671 |
| + VideoLLaMA3 | 0.3544 |

### 3.3 Submission Results

This year, we conducted four fully automatic runs to evaluate our system in the main task (see Fig. 4 and Fig. 5 for a comparison with other teams). Note that our team was the only participant to submit a run for the manually assisted task in 2025.

- F_M_C_D_NII_UIT_4 (Run #4): This run applied VILA-40B as the sole reranker on top of the Fusion (2025 pipeline), achieving xinfAP scores of **0.4050** on AVS 23, **0.4095** on AVS 24, and **0.4076** on AVS 25.
- F_M_C_D_NII_UIT_3 (Run #3): This run combined four VLLMs, VILA-40B, NVILA-15B, VideoLLaMA3, and NVILA (each with weight 1). This run achieved the highest performance among all submissions, scoring **0.4334** on AVS 24 and **0.4361** on AVS 25.
- F_M_C_D_NII_UIT_2 (Run #2): This run combined three VLLMs, VILA-40B, NVILA-15B, and VideoLLaMA3 (each with weight 1). It obtained **0.4240** on AVS 24 and **0.4258** on AVS 25.
- F_M_C_D_NII_UIT_1 (Run #1): This configuration increased the weight of VILA-40B to 4 while keeping NVILA-15B and VideoLLaMA3 with a weight of 1. It obtained **0.4245** on AVS 24 and **0.4266** on AVS 25.

Table 9: Reranking + Submission: Evaluation results on AVS 23, AVS 24, and AVS 25 for the reranking stage and the final submission runs. Abbreviations: V = VILA-40B, NV15 = NVILA-15B, VL = VideoLLaMa3, AR = Aria. The symbol $w$ denotes the fusion weights assigned to each reranker.

| Run | Methods | AVS 23 | AVS 24 | AVS 25 |
|---|---|---|---|---|
| | Fusion (2025 pipeline) | 0.3215 | - | 0.2916 |
| #4: | + V | 0.4050 | 0.4095 | 0.4076 |
| #3: | + V (w = 1) + NV15 (w = 1) + VL (w = 1) + AR (w = 1) | - | 0.4334 | 0.4361 |
| #2: | + V (w = 1) + NV15 (w = 1) + VL (w = 1) | 0.4211 | 0.4240 | 0.4258 |
| #1: | + V (w = 4) + NV15 (w = 1) + VL (w = 1) | 0.4215 | 0.4245 | 0.4266 |

For the manually-assisted approach, we conducted only one experimental run:

– M_M_C_D_NII_UIT_1: This run used Fusion + Query Expansion with T2V models, enhanced with multiple VLLM rerankers (VILA-40B, NVILA-15B, VideoLLaMA3) at TOP-K = 4000, alongside manually refined paraphrased queries. It achieved an xinfAP of **0.384** on AVS 24 and **0.410** on AVS 25.

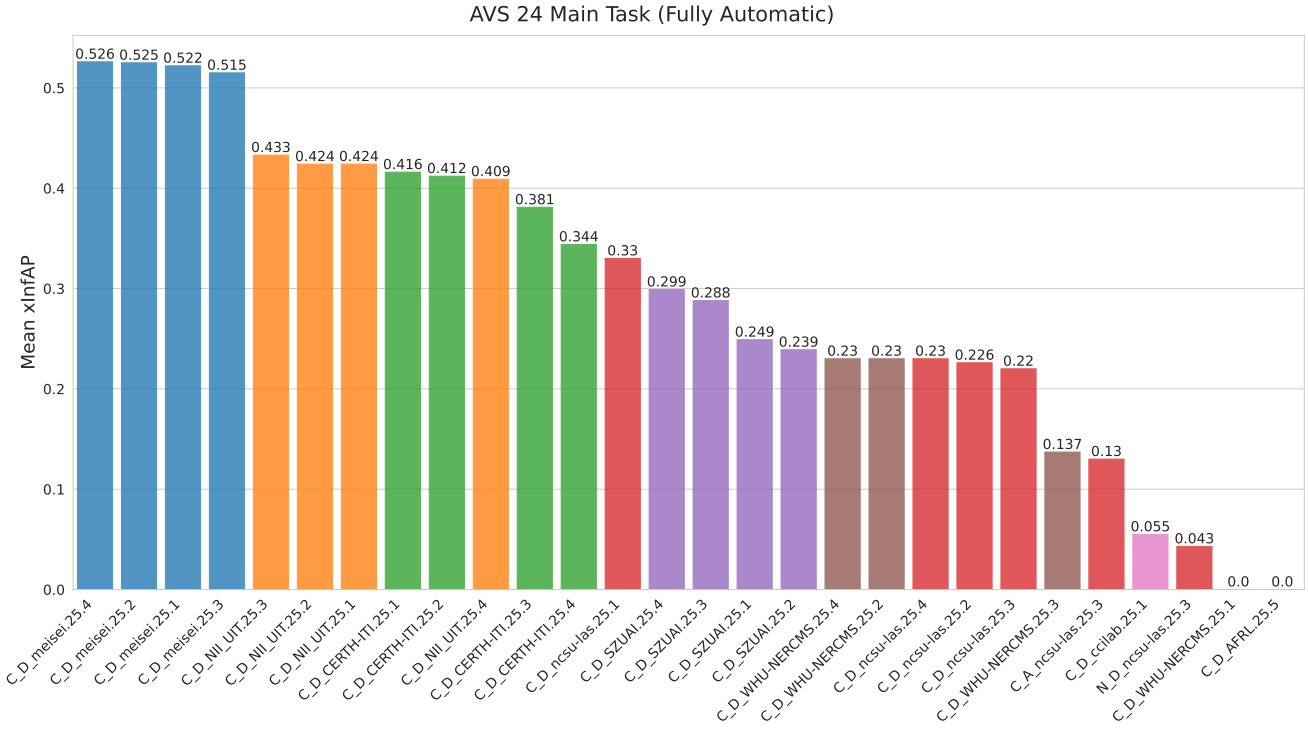


Fig. 4: Evaluation scores of Fully-Automatic runs on the Main task in AVS 24.

## 4 Conclusion

In the TRECVid Ad-hoc Video Search 2025 task, we further advance our retrieval system by emphasizing model diversity across both the retrieval and reranking stages (see the summary in Tab. 10). Building upon last year's framework, we integrate a broader set of Text-to-Video (T2V) models, such as CLIP4Clip, TS2Net, TeachCLIP, and Side4Video, to complement existing T2I backbones and enhance temporal understanding. This richer collection of retrieval models increases the semantic and temporal diversity of the retrieved shot pool, providing a stronger foundation for downstream reranking.

Motivated by our empirical observation that greater model diversity correlates with higher ensemble gains (as indicated by the relationship between AP and IoU), we also introduce multiple VLLM-based rerankers, such as VILA-40B, NVILA-15B, and VideoLLaMA3, to replace the single-model reranking strategy used in 2024. These

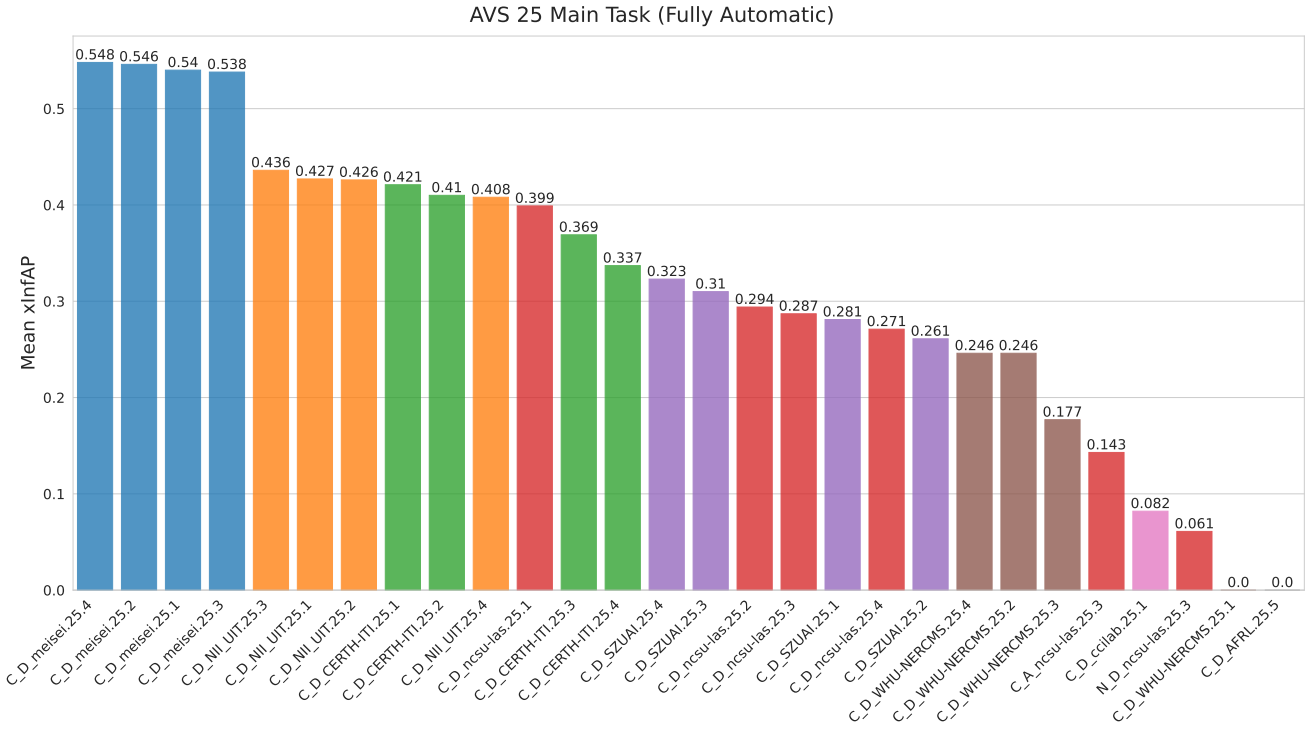


Fig. 5: Evaluation scores of Fully-Automatic runs on the Main task in AVS 25.

Table 10: Summary of incremental improvements from the 2024 pipeline to the enhanced 2025 pipeline. Step-wise additions include T2V models for increased recall diversity and multiple VLLM rerankers for improved reasoning and ranking accuracy.

| **System Variant** | **AVS 23** | **AVS 24** | **AVS 25** |
|---|---|---|---|
| Baseline (2024 best: InternVL-G) | 0.2400 | - | - |
| + Query Expansion (QE) | 0.2531 | - | - |
| + Fusion (T2I CombMNZ) | 0.3043 | - | - |
| + Single VLLM: VILA-40B | 0.3840 | 0.4250 | - |
| **2025 Additions** | | | |
| + Text-to-Video (T2V) Models (C4, TE, TS, S4) | 0.3215 | - | 0.2916 |
| + Single VLLM: VILA-40B (Run#4) | 0.4050 | 0.4095 | 0.4076 |
| + Multiple VLLMs: V, NV15, VL, AR (Run#3) | **-** | **0.4334** | **0.4361** |

complementary VLLMs bring different reasoning strengths and visual-language alignment capabilities, yielding more consistent query interpretation and improved final ranking quality.

Together with GPT-4-assisted query expansion, our 2025 system forms a unified, diversity-driven framework. By leveraging heterogeneous models at every stage of the pipeline, we achieve substantial improvements in retrieval robustness and ranking accuracy, demonstrating the effectiveness of model diversity as a guiding principle for large-scale ad-hoc video search.

## References

1. Ashish, V.: Attention is all you need. Advances in neural information processing systems **30**, I (2017)
2. 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. 2023 (2023)
3. Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A.: A short note about kinetics-600. CoRR **abs/1808.01340** (2018), `http://arxiv.org/abs/1808.01340`
4. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 190–200 (2011)

5. 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)
6. Fox, E.A., Shaw, J.A.: Combination of multiple searches. In: TREC. NIST Special Publication, vol. 500-215, pp. 243–252. National Institute of Standards and Technology (NIST) (1993)
7. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems **36** (2024)
8. 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)
9. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
10. Li, X., Chen, A., Wang, Z., Hu, F., Tian, K., Chen, X., Dong, C.: Renmin university of china at trecvid 2022: Improving video search by feature fusion and negation understanding. arXiv preprint arXiv:2211.15039 (2022)
11. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26689–26699 (2024)
12. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (2014)
13. Liu, Y., Xiong, P., Xu, L., Cao, S., Jin, Q.: Ts2-net: Token shift and selection transformer for text-video retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
14. Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al.: Nvila: Efficient frontier visual language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4122–4134 (2025)
15. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
16. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. In: European Conference on Computer Vision (ECCV) (2022)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
18. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
19. Tian, K., Zhao, R., Xin, Z., Lan, B., Li, X.: Holistic features are almost sufficient for text-to-video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17138–17147 (2024)
20. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., Wei, F.: Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
21. Wu, J., Ngo, C.W., Wei, X.Y., Li, Q.: Polysmart and vireo@ trecvid 2024 ad-hoc video search. arXiv preprint arXiv:2412.15494 (2024)
22. Yao, H., Wu, W., Li, Z.: Side4video: Spatial-temporal side network for memory-efficient image-to-video transfer learning. arXiv preprint arXiv:2311.15769 (2023)
23. Yu, Y., Kim, J., Kim, G.: A joint sequence fusion model for video question answering and retrieval. In: Proceedings of the European conference on computer vision (ECCV). pp. 471–487 (2018)
24. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
25. Zhao, Y., Song, Y., Chen, S., Jin, Q.: Ruc_aim3 at trecvid 2020: Ad-hoc video search & video to text description. In: TRECVID. vol. 3, p. 6 (2020)