VLM-based Binary Judgment Re-ranking for TREC 2025 Ad-hoc Video Search

Kazuya Ueki[1]

Department of Information Science, Meisei University,
2-1-1 Hodokubo, Hino, Tokyo 191-8506, Japan

kazuya.ueki@meisei-u.ac.jp

Abstract. We participated in the Ad-hoc Video Search (AVS) task at TREC 2025. Building upon our previous system, we aimed to further enhance search performance through a re-ranking approach. Our method employs multiple state-of-the-art Vision-Language Models (VLMs) to verify whether retrieved video shots truly match a given query, enabling more accurate semantic filtering. Among the 29 systems submitted, our four runs achieved the top four ranks, demonstrating the effectiveness of the proposed VLM-based binary judgment strategy. These results confirm the strong potential of recent VLMs to improve large-scale video retrieval.

1 Introduction

Ad-hoc Video Search (AVS) aims to retrieve video shots that are relevant to natural language queries. Recent systems have achieved high performance by employing multiple CLIP-based embedding models [1] to compute the similarity between the text query and video keyframes, enabling robust semantic retrieval in large-scale video collections. However, similarity-based retrieval often suffers from false positives; video shots may obtain high similarity scores despite matching only a partial concept described in the query. This limitation makes it difficult to accurately determine whether a retrieved shot fully satisfies the intent of the query.

Meanwhile, recent advances in Vision-Language Models (VLMs), which are powered by Large Language Models (LLMs) and support conversational interaction, have shown remarkable capabilities across a wide range of image understanding tasks. These models can interpret scenes and respond to queries in a more context-aware manner.

Motivated by these observations, we utilize multiple VLMs to re-verify whether initially retrieved video shots truly match the given query, and re-rank them accordingly. This binary judgment strategy aims to effectively filter out false positives and improve the retrieval accuracy.

The contributions of this paper are summarized as follows:

- We introduce a VLM-based binary judgment method to filter false positives in retrieved shots.
- We integrate multiple VLMs to enhance semantic verification within the re-ranking framework.
- We demonstrate the effectiveness of our approach, achieving the 1st to 4th rank among the 29 systems submitted in the TREC 2025 AVS task.

2 Baseline System from TREC 2024

Our baseline system is based on our previous submission to TREC 2024 [2] and comprises four key components. First, we utilize multiple CLIP-based embedding models to compute the similarity between textual queries and video frames. Second, we expand

each query by generating additional textual and visual descriptions to improve semantic coverage. Third, we fine-tune embedding models using the V3C dataset to better align them with the target domain. Finally, we fuse similarity scores from multiple models to obtain the final retrieval ranking.

2.1 Pre-trained Embedding Models

We adopted multiple state-of-the-art pre-trained embedding models to compute the similarity between textual queries and video frames. Our system relied mainly on CLIP-based models, which are widely used for vision-language representation learning because of their strong semantic alignment performance. We selected several variants from OpenCLIP[1], such as SigLIP [3], EVA-CLIP [4], and MetaCLIP [5], all of which were trained on large-scale image–text datasets including LAION-2B/5B [6]. In addition to CLIP-based models, we further evaluated other advanced multimodal encoders such as BLIP [7], BLIP-2 [8], ALIGN [9], Long-CLIP [10], VeCLIP [11], ViTamin [12], LLaVA [13], and Phi3-Vision [14]. These models provide diverse embedding characteristics, allowing the system to capture a broader range of query semantics and visual concepts.

For each video, keyframes were extracted at fixed intervals, and similarity between each keyframe and the textual query was computed for the initial retrieval ranking. By leveraging multiple embedding models, we obtained complementary similarity scores that contributed to more robust retrieval performance.

2.2 Query Expansion with Textual and Visual Generation

To improve query coverage, we expanded the original query descriptions into multiple semantically similar alternative expressions. A large language model was prompted to generate sentence reformulations conveying the same intent as the original query, and redundant or irrelevant outputs were removed based on semantic similarity evaluation using a multimodal model. This allowed us to retain only a diverse set of text variants that remained faithful to the query meaning.

In addition to the textual expansion, we also generated visual examples corresponding to the original query using image generation models such as Stable Diffusion v2.1 and Stable Diffusion XL [15]. Each generated image was screened by measuring similarity to the original query to filter out visually inconsistent samples.

These multimodal expansions increase the likelihood of retrieving relevant shots that might be missed by the original query alone and improve robustness against visual variations in the target videos.

2.3 Domain Adaptation of Embedding Models

To adapt embedding models to the characteristics of TREC video data, we fine-tuned CLIP models using captions automatically generated from the target dataset. The fine-tuning process follows a lightweight strategy to prevent overfitting while improving domain alignment.

Specifically, inspired by the LiT (Locked-image Tuning) strategy [16], we updated only a small subset of the model parameters: the final projection layers of the image encoder and the entire text encoder, while keeping the remaining weights frozen. This lightweight parameter update limits the risk of overfitting and preserves the robust visual representations originally trained during large-scale pre-training. Training was performed for a single epoch using captioned frame-image pairs extracted from the V3C videos. This limited-domain tuning strengthens the semantic consistency between

[1] `https://github.com/mlfoundations/open_clip`

queries and video frames without sacrificing the zero-shot capabilities of the original models.

Additionally, we applied WiSE-FT to linearly fuse the weights of the fine-tuned model and the original one, enabling a balanced combination of domain adaptation and generalization performance. This adaptation process consistently improved retrieval accuracy in baseline experiments, supporting its effectiveness as a foundation for our re-ranking approach.

2.4 Fusion of Multiple Similarity Scores

Similarity scores produced by different embedding models were normalized into a common range using min–max scaling, enabling balanced score integration. We employed a forward selection strategy in which models were incrementally added based on their contribution to mean Average Precision (mAP). For each newly added model, its fusion weight was adjusted within a predefined range, and the optimal weight was selected. The process continued until no further performance gains were observed. This adaptive fusion allowed us to exploit complementary strengths of multiple embedding models, resulting in more reliable initial retrieval performance before applying VLM-based re-ranking.

3 Proposed Re-ranking Using VLM-based Binary Judgment

The initial similarity-based retrieval often ranks shots highly even if they only partially match the query semantics, leading to false positive results. To address this limitation, we propose a VLM-based re-ranking method which performs strict semantic verification on top of the baseline system. Importantly, the method is entirely zero-shot and training-free, requiring no additional annotated data or model updating.

We adopt the best-performing system from our previous submission as the initial retrieval module. Only the top-k shots of the initial ranking are forwarded to the re-ranking stage, allowing us to focus computation on highly relevant candidates while retaining scalability for large-scale video collections.

For each retrieved shot, a Vision-Language Model (VLM) is prompted in a yes-or-no question answering format to assess whether the visual content truly satisfies the textual query. Specifically, the original query sentence is converted into a binary question such as:

"Does this image contain the following scene?: {query}. Answer YES or NO."

For scoring, we extract the logits for the output tokens "YES" and "NO," apply a softmax function, and use the probability assigned to "YES" as the confidence score of semantic relevance. This allows the VLM to explicitly verify semantic consistency between the query and the retrieved shot.

For each retrieved shot, we uniformly sample multiple keyframes and evaluate them using several different VLMs. Each VLM is prompted in a binary yes-or-no manner to judge whether the content of a given frame is semantically relevant to the query. The highest confidence score among the sampled frames is utilized as the re-ranking score for that VLM. This strict semantic verification effectively removes false positives caused by partial matches or visually ambiguous situations.

Since the VLM inference is performed entirely in a zero-shot, training-free manner, the approach preserves the strong generalization ability of pre-trained VLMs. To further enhance robustness, the final re-ranking score is obtained by combining the scores from multiple VLMs through weighted integration. The fusion weights are determined based on validation performance, allowing complementary strengths from different models to be reflected in the final ranking. By enforcing semantic consistency and integrating multiple VLM judgments, the proposed re-ranking framework significantly improves retrieval precision while keeping computational overhead manageable.

Table 1. Integration ratio optimization settings and evaluation results. The proposed fusion strategy significantly improved retrieval accuracy, achieving mAP scores well above the baseline performance (0.351) obtained with the 2024 ground truth.

Run ID	Validation data	Step	2024 ground truth		2025 ground truth	
			mAP	Rank	mAP	Rank
1	tv22+tv23	0.10	0.522	3	0.540	3
2	tv22+tv23	0.05	0.525	2	0.546	2
3	tv22	0.10	0.515	4	0.538	4
4	tv23	0.10	0.526	1	0.548	1

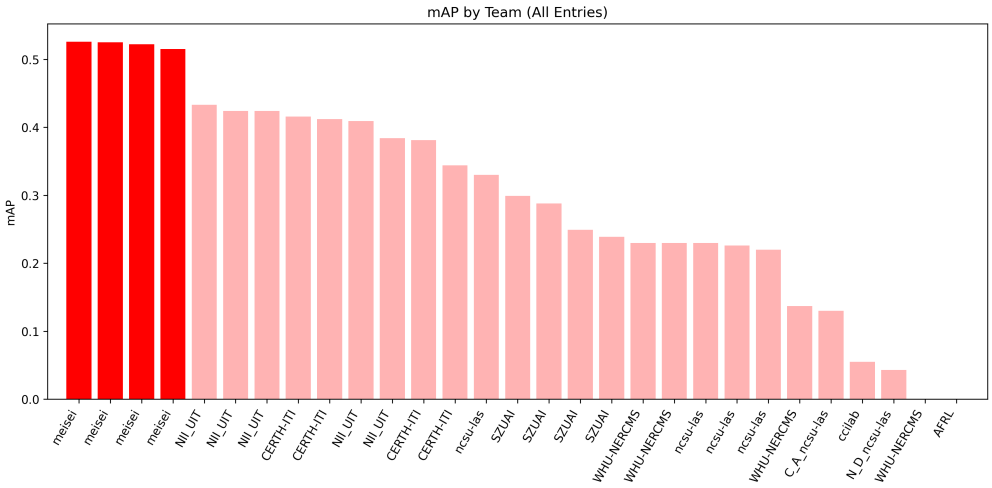

Fig. 1. Ranking results of all 29 submissions in the TREC 2025 AVS task evaluated with the 2024 ground truth. Our four runs (shown in red and labeled as "meisei") achieved the top four mAP scores among all participating teams.

4 Results in TREC 2025 AVS Task

Our submitted systems were evaluated in the TREC 2025 Ad-hoc Video Search (AVS) task. A total of 29 runs were submitted by participating teams. For the official assessment, both the previously available ground truth from the 2024 benchmark and newly expanded annotations added in 2025 were used. The updated annotations extend the coverage of relevant shots, thereby enabling a more comprehensive evaluation of retrieval effectiveness.

The best-performing system from our TREC 2024 submission was used as the baseline, which achieved an mAP of 0.351 under the 2024 evaluation. We applied our VLM-based re-ranking approach to the top 2,000 retrieved shots of the baseline system. Ten VLMs with strong zero-shot visual reasoning capabilities were examined for semantic verification:

- `mistralai/Pixtral-12B-2409`
- `Efficient-Large-Model/Llama-3-VILA1.5-8B-Fix`
- `Efficient-Large-Model/NVILA-8B`
- `microsoft/Phi-4-multimodal-instruct`
- `OpenGVLab/InternVL2_5-4B-MPO`
- `OpenGVLab/InternVL2_5-8B-MPO`
- `OpenGVLab/InternVL3-2B`
- `OpenGVLab/InternVL3-8B`
- `OpenGVLab/InternVL3-9B`
- `Qwen/Qwen2.5-VL-7B-Instruct`

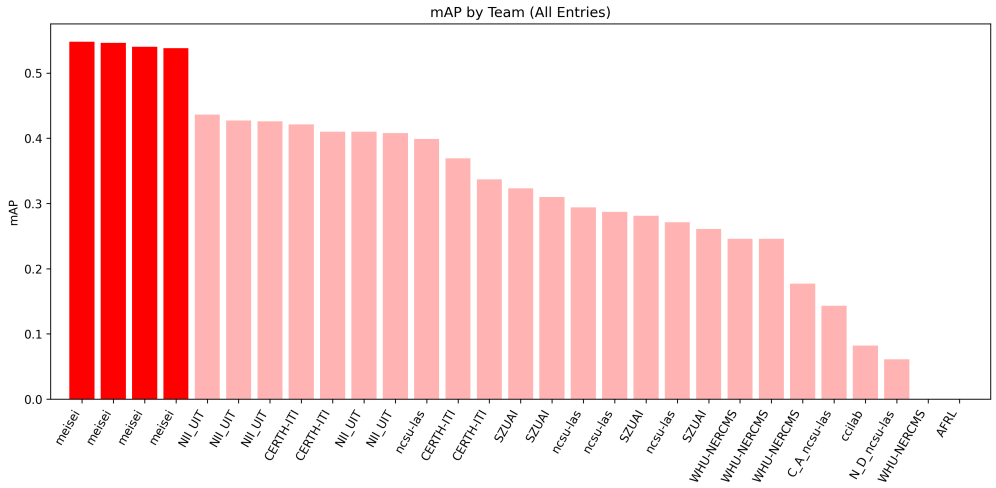

Fig. 2. Ranking results of all 29 submissions in the TREC 2025 AVS task evaluated with the 2025 ground truth. Our four runs (shown in red and labeled as "meisei") achieved the top four mAP scores among all participating teams. This result further confirms the strong effectiveness of incorporating VLM-based semantic verification into the ranking process.

To integrate the baseline similarity scores and the VLM-based binary judgment scores, we applied the same incremental weight optimization strategy as in our previous system, where scores were normalized into a 0.0–1.0 range using min–max scaling and then fused by forward selection to maximize validation performance. The integration ratios were optimized using two step sizes, 0.05 and 0.10. Table 1 summarizes the validation settings and resulting performance of our submitted runs, including different VLM combinations and integration increments.

As shown in Fig. 1, our four runs (highlighted in red and labeled as "meisei") clearly achieved the top four mAP values when evaluated using the 2024 ground truth, demonstrating the substantial advantage gained by introducing VLM-based semantic verification into the ranking pipeline. Similarly, Fig. 2 presents the results using the expanded 2025 ground truth, where our four runs again ranked first through fourth with an even wider performance margin, confirming that the approach remains robust even with the inclusion of additional relevant shots in the updated ground truth.

Further analysis also revealed that leveraging multiple heterogeneous VLMs consistently yielded higher accuracy than relying on any single model, indicating that complementary semantic judgments across models contribute positively to retrieval precision. Overall, these results verify that the proposed VLM-based zero-shot re-ranking framework effectively suppresses false positives while preserving strong generalization capability, and serves as an efficient and highly practical enhancement to existing large-scale AVS systems.

5 Conclusion

We proposed a zero-shot VLM-based re-ranking framework to improve semantic accuracy in Ad-hoc Video Search. By applying strict binary semantic verification and integrating multiple VLMs, our method successfully reduces false positives that remain in similarity-based retrieval. In the TREC 2025 AVS task, our four submitted runs ranked 1st through 4th out of 29 systems, demonstrating the strong effectiveness of the proposed approach.

These results confirm the potential of leveraging recent VLMs to enhance large-scale video retrieval systems without requiring any additional training or annotated

data. In future work, we plan to further improve semantic understanding in complex scenarios and explore more efficient re-ranking strategies that balance accuracy and computational cost.

References

1. 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," arXiv:2103.00020, 2021.
2. K. Ueki, Y. Suzuki, H. Takushima, H. Sato, T. Takada, A. M. Kumar, H. Tanoue, H. Nishihara, Y. Shibata, and Takayuki Hori, "Softbank-Meisei at TREC 2024 Ad-hoc Video Search and Video to Text Tasks," In Proc. of the Text REtrieval Conference (TREC 2024), 2024.
3. X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid Loss for Language Image Pre-Training," In Proc. of IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
4. Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "EVA-CLIP: Improved Training Techniques for CLIP at Scale," arXiv:2303.15389, 2023.
5. H. Xu, S. Xie, X. E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer, "Demystifying CLIP Data," arXiv:2309.16671, 2023.
6. C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5B: An open large-scale dataset for training next generation image-text models," In Proc of the 36th Conference on Neural Information Processing Systems (NeurIPS), 2022.
7. J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," arXiv:2201.12086, 2022.
8. J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," arXiv:2301.12597, 2023.
9. C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," In Proc. of the International Conference on Machine Learning (ICML), 2021.
10. B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang, "Long-CLIP: Unlocking the Long-Text Capability of CLIP," In Proc. of the European Conference on Computer Vision (ECCV), 2024.
11. Z. Lai, H. Zhang, B. Zhang, W. Wu, H. Bai, A. Timofeev, X. Du, Z. Gan, J. Shan, C.-N. Chuah, Y. Yang, and M. Cao, "VeCLIP: Improving CLIP Training via Visual-enriched Captions," arXiv:2310.07699, 2023.
12. J. Chen, Q. Yu, X. Shen, A. Yuille, and L.-C. Chen, "ViTamin: Designing Scalable Vision Models in the Vision-Language Era," In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.
13. H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," In Proc. of the Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS), 2023.
14. Marah Abdin, et al., "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone," arXiv:2404.14219 , 2024.
15. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.10684–10695, 2022.
16. F. Zhai, A. Puig, L. Kirillov, A. Joulin, and I. Misra, "LiT: Locked-image Tuning for Efficient Large-scale Vision-Language Model Training," In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.