WHU-NERCMS AT TRECVID2025: AD-HOC VEDIO SEARCH(AVS) AND VIDEO QUESTION ANSWER(VQA) TASK

Fangyun Duan[*] **Haixiang Ni**[*] **Xiusong Wang, Chao Liang**[†]
Hubei Key Laboratory of Multimedia and Network Communication Engineering
National Engineering Research Center for Multimedia Software
School of Computer Science, Wuhan University
cliang@whu.edu.cn

ABSTRACT

The WHU-NERCMS team participated in the Ad-hoc Vedio Search (AVS) and Video Question Answer(VQA) tasks at TRECVID 2025. For AVS task, we continued to use multiple visual semantic embedding methods, combined with ranking aggregation techniques to integrate different models and their outputs to generate the final ranked video shot list. For VQA task, we propose to use the VLM model to generate answer that serve as baseline answer. The answer is then embedded in the same vector space with the four options, and then compute the similarity of these vectors to sort the results.

1 AVS TASK

1.1 Introduction

The AVS task aims to return a list of similarity ranking results as accurately as possible from large video datasets based on given text ad-hoc queries. This year's dataset is still V3C2[1], a dataset that contains 9,760 videos with a total of 1,300 hours.Following last year's approach, we used multiple models to generate the base ranking and then reorder the results to get better results. This year we have adopted some new models for getting better basic search retrieval.

1.2 Method

We used the following Language-Image pre-trained models to construct our system. Overall, we generally continue to use the Language-Image pre-trained models that performed well last year.

[*] Both authors contributed equally to this research.

[†] Corresponding author.

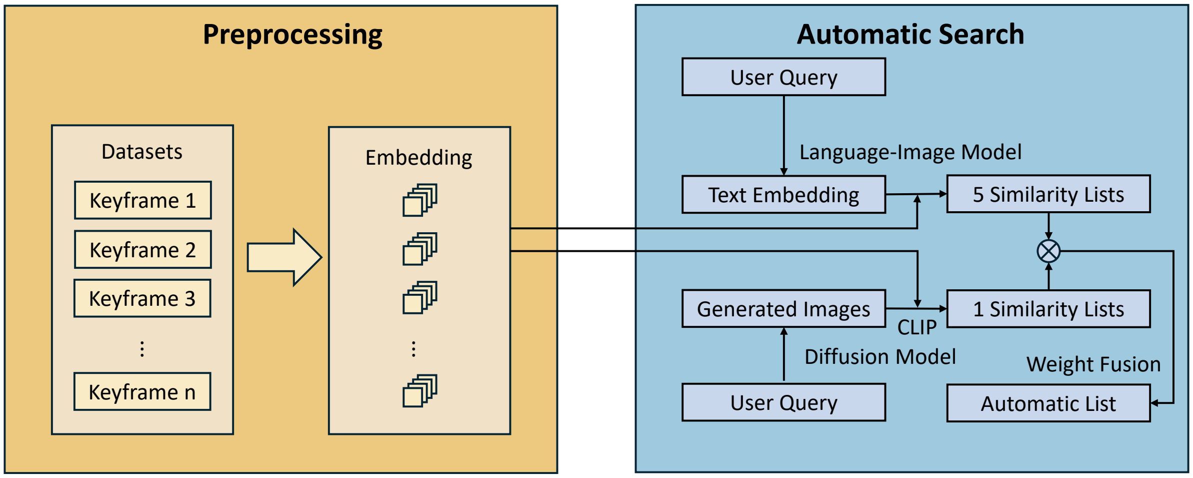

Figure 1: AVS framework

1.2.1 Embedding Models

This year, we employed the following model to generate the base ranking. The BEIT3 and InternVL models are newly adopted this year.

CLIP[2]: RN50x4, RN50x16, RN50x64, RN50, RN101, ViT-B/16, ViT-B/32, ViT-L/14.

SLIP[3]: ViT-Small, ViT-Base, ViT-Large, ViT-Base(CC3M), ViT-Base(CC12M).

BLIP[4]: ViT-B(COCO), ViT-B(Flickr30k), ViT-L(COCO), ViT-L(Flickr30k).

BLIP-2[5]: BLIP2(COCO).

LaCLIP[6]: LaCLIP(CC12M).

BEiT3[7]: BEiT3-base.

InternVL[8]: InternVL-14B-Flickr30K-FT-364px.

For all the models mentioned, we extract feature vectors for each keyframe in the dataset and store them in a database for retrieval. When a user submits a text query, we extract the relevant text vectors and compute their similarity with the stored model features.

For different pretrained models of the same type, we sum the similarity scores for the same keyframes, ensuring they are between 0 and 1 with Min-Max Normalization. This can be expressed by the following formula:

$$s_i^{id} = \mathrm{norm}\left(\sum_{p \in M} s_p^{id}\right) \tag{1}$$

Here, "id" represents the shot id corresponding to the keyframe, "i" denotes the i th model type , and "p" refers tovarious pre-trained models of the same type. Following this calculation, we generate five lists of similarity scores.

Additionally, we employed the Stable Diffusion v2-1 to construct the "mean image query." It was used to generate 1,000 images based on the given text queries. Subsequently, we extracted the visual features of each image using CLIP's VIT-B/32. The mean similarity score was calculated by averaging the similarity between the features of these 1,000 images and the visual features in the dataset.

1.2.2 Ranking Aggregation

Through different cross-modal models, we obtain multiple score lists s_i, with w_i as an appropriate weight. The final score s^{id} for a query is calculated as:

$$s^{id} = \sum w_i s_i^{id} \tag{2}$$

For the Ranking Aggregation, we have adopted the following approaches. In addition to the traditional parameterization, we also used the unsupervised RA method, the HPA[9] method.

1.3 Conclusion

The results are shown in the table below.

Table 1: AVS Results

Run_ID	**infAP**		**Weight(B3:B:B2:C:I:L:S:D)**
	2024	**2024+2025**	
F_1	–	–	–
F_2	0.246	0.230	1:1:1:1:1:1:1:1
F_3	0.177	0.137	HPA[9]
F_4	0.246	0.230	6:16:4:10:5:3:3:3

Table 2: Model of the abbreviation in our method

Abbreviation	**Description**	**Abbreviation**	**Description**
B3	BEiT3	B	BLIP
B2	BLIP2	C	CLIP
I	InternVL	L	LaCLIP
S	SLIP	D	Diffusion

2 VQA

2.1 Introduction

Video Question Answering (Video QA) stands as a pivotal task at the intersection of computer vision and natural language processing, aiming to enable machines to comprehend video content and answer natural language questions about it. This task presents significant challenges, primarily due to the necessity for temporal modeling to understand actions and causal relationships across frames, effective multimodal fusion to jointly reason over visual and textual information, and managing the high computational complexity of processing lengthy videos. Furthermore, answering often requires commonsense reasoning beyond the pixels and involves mitigating dataset biases that might lead models to rely on spurious patterns rather than genuine understanding.

2.2 Method

Our framework is shown in the figure 2 .

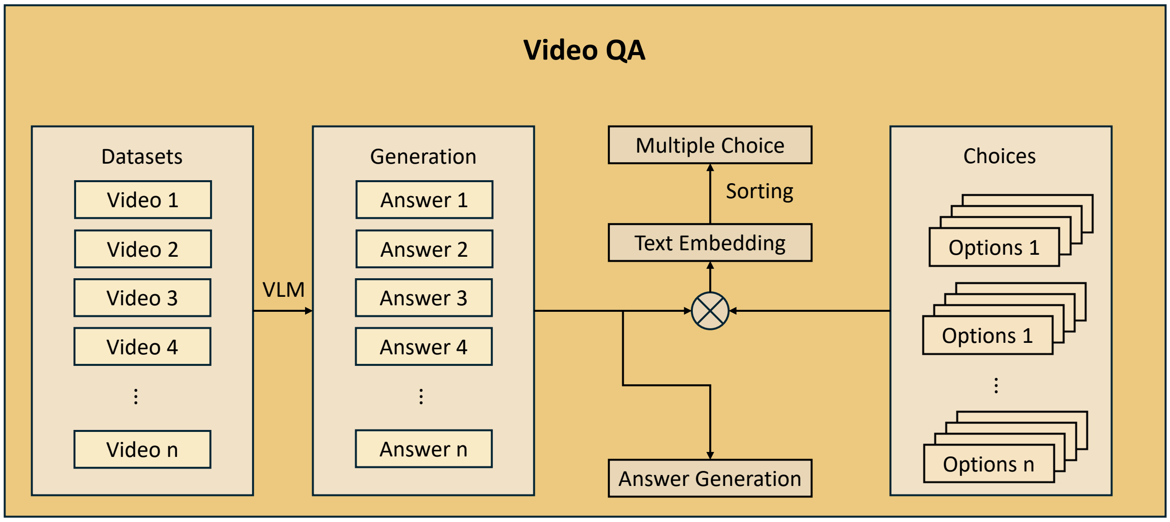

Figure 2: VQA framework

The model VideoLLaMA[10] based on QWEN2.5 is used to generate the base answer. Then, in the Answer Generation task, we used two models for generating results.In the Multiple Choice task, we found that the results generated using the 7B model contained more information and were more conducive to word embedding.

In the word embedding session, we embed the generated answer and the four options into the same vector space, then perform cosine similarity computation on the word vectors of the answer and the four options, and sort the results to get the final results.

2.3 Conclusion

The results of the AG are shown in Table 3, and the results of the MC are shown in Table 4.

Table 3: VQA AG Results

Run_ID	METEOR	BERTScore	STSscore
1	0.234	0.865	0.302
2	0.231	0.866	0.300
3	0.238	0.863	0.298

Table 4: VQA MC Results

Run_ID	Top1Correct	MRRScore
1	0.527	0.706

3 Acknowledgement

This work is supported by the National Natural Science Foundation of China (No.62372339), Hubei Provincial Science and Technology Plan Project (No.2025CSA057, 2025BAB020), Key Science and Technology Research Project of Xinjiang Production and Construction Corps in 2025, the Fundamental Research Funds for the Central Universities (No.2042023kf1033) and the Ministry of Education Industry-University Cooperative Education Project (No.240700006245501). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A. Butt. V3c – a research video collection. In Ioannis Kompatsiaris, Benoit Huet, Vasileios Mezaris, Cathal Gurrin, Wen-Huang Cheng, and Stefanos Vrochidis, editors, *MultiMedia Modeling*, pages 349–360, Cham, 2019. Springer International Publishing.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[3] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 529–544, Cham, 2022. Springer Nature Switzerland.

[4] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba

Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 17–23 Jul 2022.

[5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 23–29 Jul 2023.

[6] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 35544–35575. Curran Associates, Inc., 2023.

[7] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 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.

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[9] Soichiro Fujita, Hayato Kobayashi, and Manabu Okumura. Unsupervised ensemble of ranking models for news comments using pseudo answers. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, pages 133–140, Cham, 2020. Springer International Publishing.

[10] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, and Lidong Bing andDeli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.