Doshisha University at TREC 2025 AVS Task

Dai Morisaki[1], Miho Ohsaki[1] and Kimiaki Shirahama[1]
[1]Doshisha University, 1-3, Tatara Miyakodani, Kyotanabe, 610-0394 Kyoto, Japan
E-mail: kshiraha@mail.doshisha.ac.jp

Abstract – **This paper presents the result obtained by Co-creation Informatics Laboratory (ccilab), Doshisha University on Ad-hoc Video Search (AVS) task. Our initial plan was to test the performance of a latest vision-language model, especially SigLip2 [1]. However, in our preliminary experiments, features extracted by the pre-trained SigLip2 available on [2] did not work at all. Thus, our submitted run *F_M_C_D_ccilab.25.1* was obtained using a basic vision-language model, namely OpenAI CLIP [3], especially the pre-trained one released on [4]. The MAPs of our submitted run are** 0.082 **and** 0.055 **when using the** 2024 **ground truth and the combination of the** 2024 **and** 2025 **ground truths, respectively.**

I. Introduction

Since 2008, we are continuously participating in TREC (or TRECVID until 2023) in order to objectively evaluate the performance of our system through its comparison to systems developed all over the world [5]. Although we participated in medical domain tasks in the past two years [6, 7], this year we return to Ad-hoc Video Search (AVS) task [8] to focus on more fundamental image/video processing modules, especially, feature extraction. Our aim is to examine the performances of latest Vision-Language Models (VLMs) that have been pre-trained on a huge number of image-caption pairs collected on the Web. In particular, we focused on SigLIP2 because its effectiveness as a visual encoder had been proven in [1]. However, in our preliminary experiments on V3C1 dataset [9], features extracted by the SigLIP2 model (siglip2-giant-opt-patch16-384) released on [2] did not work at all. Hence, we decided to submit a result obtained by a basic VLM, OpenAI CLIP [3], especially the pre-trained one available on [4].

II. Method

Because of the background described in the introduction, we only had time to directly use CLIP's text-to-image framework to complete the AVS task on V3C2 dataset for the given 20 topics. Specifically, each topic is encoded into a vector by CLIP's text encoder, and each shot is represented by the NIST-provided keyframe that is then encoded into a vector via CLIP's image encoder. Here, these vectors are normalised to be of unit length. Under this setting, shots in V3C2 dataset are ranked according to their cosine similarities to the topic, and the top-ranked 1000 shots are regarded as our search result.

III. Results

Figs. 1 and 2 show the evaluation results of our submitted run (*F_M_C_D_ccilab.25.1*) using the 2024 ground truth and the combination of the 2024 and 2025 ground truths, respectively. Our run is significantly outperformed by the others due to the simplicity of our search method that just uses the basic OpenAI CLIP. Our future work to improve it is described in the next section.

IV. Conclusion

This paper presented our search method developed for TREC 2025 AVS task and the evaluation results of the run obtained by it (*F_M_C_D_ccilab.25.1*). To improve the current performance, we are now exploring to leverage VLMs (especially, SigLIP2 (ViT-gopt-16-SigLIP2-384) and CoCa (coca_ViT-L-14)) pre-trained in OpenCLIP project [10]. The performances of these VLMs have been already validated through our preliminary experiments on V3C1 dataset. Based on this, we plan to finetune these VLMs to AVS task using V3C1 dataset and past topics in the framework of CLIP-adapter [11].

REFERENCES

[1] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[2] Google. google/siglip2-giant-opt-patch16-384 · hugging face.

[3] 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 *Proc. of the 38th International Conference on Machine Learning (ICML 2021)*, pages 8748–8763, 2021.

[4] OpenAI. https://huggingface.co/openai/clip-vit-large-patch14.

[5] George Awad, Jonathan Fiscus, Afzal Godil, Lukas Diduch, Yvette Graham, and Georges Quénot. Trecvid

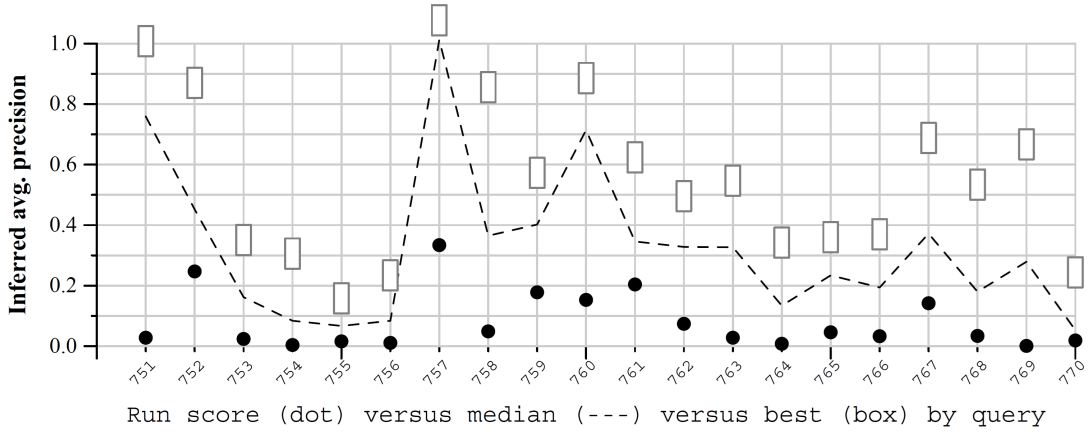

Fig. 1. Inferred average precisions of our submitted run (*F_M_C_D_ccilab.25.1*) for the given 20 topics when using the 2024 ground truth.

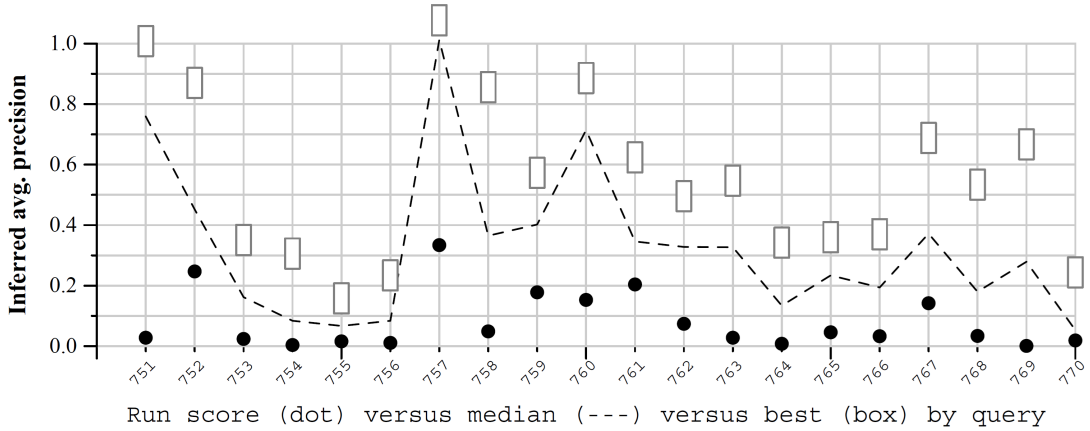

Fig. 2. Inferred average precisions of our submitted run (*F_M_C_D_ccilab.25.1*) for the given 20 topics when using the combination of the 2024 and 2025 ground truths.

2024 - evaluating video search, captioning, and activity recognition. In *Proceedings of TRECVID 2024*. NIST, USA, 2024.

[6] Zihao Chen, Frédéric Li, Marc S. Seibel, Nele S. Brügge, Miho Ohsaki, Heinz Handels, Marcin Grzegorzek, and Kimiaki Shirahama. Doshisha university, universität zu lübeck and german research center for artificial intelligence at trecvid 2023: Miqg task. In *Proc. of the TREC Video Retrieval Evaluation (TRECVID) 2023*, 2023.

[7] Zihao Chen, Falco Lentzsch, Nele S. Brügge, Frédéric Li, Miho Ohsaki, Heinz Handels, Marcin Grzegorzek, and Kimiaki Shirahama. Doshisha university, universität zu lübeck and german research center for artificial intelligence at trecvid 2024: Qfisc task. In *Proc. of the 33rd Text REtrieval Conference (TREC 2024)*, 2024.

[8] Jakub Lokoč, Werner Bailer, Klaus Schoeffmann, Bernd Muenzer, and George Awad. On influential trends in interactive video retrieval: Video browser showdown 2015–2017. *IEEE Transactions on Multimedia*, 20(12):3361–3376, 2018.

[9] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt. V3c–a research video collection. In *International Conference on Multimedia Modeling*, pages 349–360. Springer, 2019.

[10] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2023.