VIREO @ TRECVid 2023 Ad-hoc Video Search

Jiaxin Wu[†⋆], Zhixin Ma[⋆], Sheng-Hua Zhong[§], Chong-Wah Ngo[⋆]

[†]*Department of Computer Science, City University of Hong Kong*

[⋆]*School of Computing and Information Systems, Singapore Management University*

[§]*Department of Computer Science and Software Engineering, Shenzhen University*

jiaxin.wu@my.cityu.edu.hk, zxma.2020@phdcs.smu.edu.sg,
csshzhong@szu.edu.hk, cwngo@smu.edu.sg

Abstract

In this paper, we summarize our submitted runs and results for the Ad-hoc Video Search (AVS) task at TRECVid 2023 [1].

Ad-hoc Video Search (AVS): This year, we applied the generative model pre-trained on large vision-text datasets to understand the AVS query. Specifically, the query is understood by generating images from text. We also perform image captioning on the generated images to obtain textual descriptions as new queries. The generated images and captions are then used for image-to-video and caption-to-video searches. Using different combinations of them, we submitted four automatic runs. For the manual run, we use a large language model (LLM) (i.e., GPT4) to understand the query by rephrasing it several times, and we manually pick one for retrieval. For all the runs, the retrieved results are reranked according to the videos' concept similarity with the textual query. We briefly summarize our runs as follows:

- *F_M_C_D_VIREO.23_1*: This automatic run attains the mean xinfAP= 0.237 on the main task using caption-to-video retrieval.
- *F_M_C_D_VIREO.23_2*: This automatic run attains the mean xinfAP= 0.215 on the main task by image-to-video search.
- *F_M_C_D_VIREO.23_3*: This automatic run obtains the mean xinfAP= 0.256 on the main task. It combines the rank lists of image-to-video and caption-to-video searches with equal weights.
- *F_M_C_D_VIREO.23_4*: This automatic run attains the mean xinfAP= 0.268 on the main task. It ensembles the results of run *F_D_C_D_VIREO.23_3*, BLIP2 [2], CLIP [3] and Imagebind [4].
- *F_M_N_D_VIREO.23_6*: This novelty run is based on the embedding-based search of our interpretable embedding model (ITV) [5] with generated captions. As a result, this run attains mean xinfAP= 0.040 for the main task.
- *M_M_C_D_VIREO.23_1*: This manual run applies the same system with the same settings presented in the run *F_M_C_D_VIREO.23_1*. The difference is that the original queries are rephrased with new terms. The performance drops from 0.237 to 0.222.
- *M_M_C_D_VIREO.23_2*: This manual run is based on the same system with the same settings presented in the run *F_M_C_D_VIREO.23_2* with manual queries. This run obtains 0.002 in the main task.

- *M_ M_ C_ D_ VIREO.23_ 3*: This manual run is based on the same setting presented in the run *F_ M_ C_ D_ VIREO.23_ 3* but with manual queries. It decreases the automatic result from 0.256 to 0.072.
- *M_ M_ C_ D_ VIREO.23_ 4*: This manual run uses the same system with the same settings presented in the run *F_ M_ C_ D_ VIREO.23_ 4* but with manual queries. The performance changes from 0.268 to 0.250 in the main task.
- *M_ M_ N_ D_ VIREO.23_ 5*: This concept run attains the mean xinfAP= 0.235 on the main task produced by the concept-based search of our ITV model with generated captions.

1 Ad-hoc Video Search (AVS)

Having a good understanding of the queries is challenging in the AVS task, especially for those queries that have a few training cases in the retrieval model. Inspired by the ability of generative models (e.g., stable diffusion model [6]) to understand natural language and generate high-quality images, we propose to apply the generative model to understand the query by generating images, and the images are subsequently used for visual similarity search. In addition, we perform image captioning on the generated images, which serves as a rewritten version of the original query. As the generated images may misinterpret the query perhaps due to LLM hallucination and the generated captions may be depicted in a different viewpoint from the query, we also verify the generated images and captions by visual QA and textual QA using LLM before using them for retrieval. Our interpretable embedding model for text-to-video search (ITV) trained with likelihood and unlikelihood losses [5] is used as the core search engine, which enables image-to-video and text-to-video searches and also provides consistent and coherent decoded concepts for video and text embeddings. Different from the original version presented in the paper [5], we use more advanced features (e.g., CLIP [3], BLIP2 [2], imagebind [4]) to encode video and text. We also expand the concept bank with phrases. In the following, we elaborate our approaches for AVS this year and analyze the result.

2 Method

2.1 Understanding the query by generating images and captions

Given an AVS query, it is input to stable diffusion model [6] as a prompt to generate images as a visual understanding of the query. Specifically, we perform the generation with multiple seeds to get diverse images for a query. Fig. 1 shows examples of two AVS queries on the tv23 query set. The first group is for the query-741 *Find shots of a red or blue scarf around someone's neck*, and the second group is for query-746 *Find shots of a man riding a scooter.* As seen, the generated images match the queries with good quality, and they are used as visual queries to search for the target videos.

For each generated image, we also perform the image captioning using BLIP2 model [2], and the image descriptions are used as the textual understanding of the query and input to our ITV model for search. Fig. 1 displays the image captions produced for the generated images. As seen, the image captions detail and simplify the original query to a specific scenario. For example, the original query *a red or blue scarf around someone's neck* is changed to *a man wearing a red scarf.* The new query matches the information need and is also easier to search than the original query which involves logical words. In the second example, the image captions detail the location of riding a scooter on a city street.

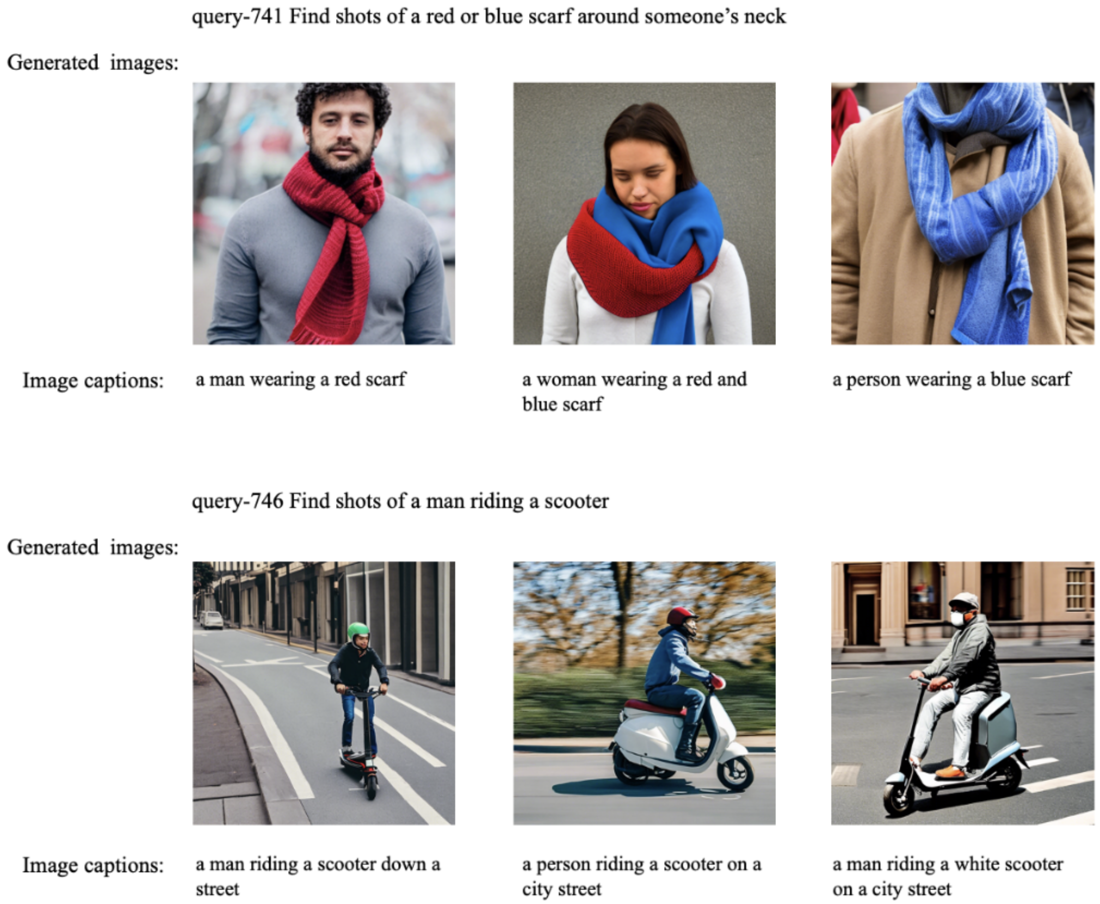
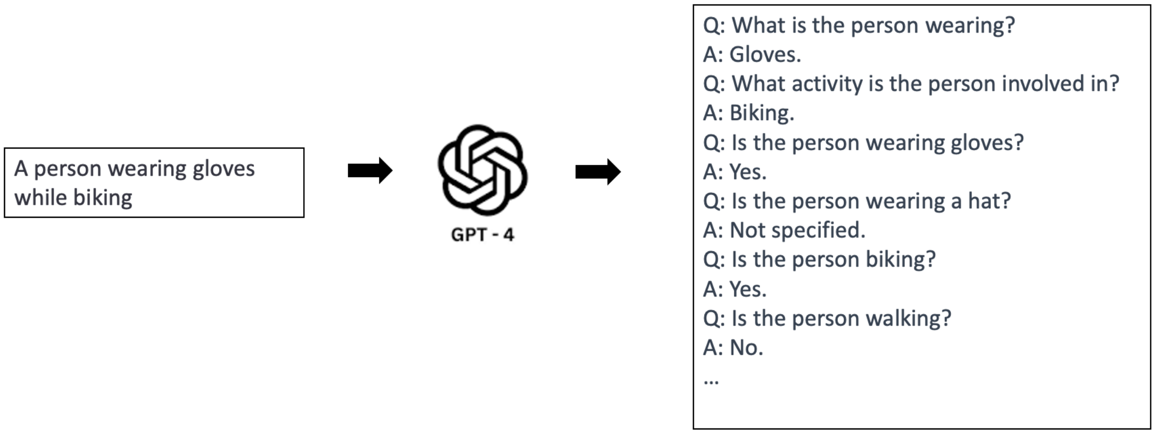

query-741 Find shots of a red or blue scarf around someone's neck
Generated images:
Image captions:
a man wearing a red scarf
a woman wearing a red and blue scarf
a person wearing a blue scarf
query-746 Find shots of a man riding a scooter
Generated images:
Image captions:
a man riding a scooter down a street
a person riding a scooter on a city street
a man riding a white scooter on a city street

Figure 1: The examples of generated images and their corresponding generated image captions.

2.2 Verification of the generations by LLM

As the resulting generated images and captions could be unrelated to the original AVS query, we further perform verification on the generation results using LLM. First, we use GPT4 [7] to generate several kinds of QAs for each AVS query. Specifically, we ask GPT4 to generate two kinds of QA: open-vocabulary QA and YES/NO QA, and all the questions should focus on person, object, action, location, time, color, and quantity. The prompt to the GPT4 is "you act as a question generator. Given a sentence, generate various simple and short QA pairs (including open-vocabulary QA and YES/NO QA) for the given sentence, with extra emphasis on person/being, action, object, location, time, color, and quantity when specified. The number of YES and NO QAs should be balanced." We input both the original AVS query and its narrative to generate QAs. Secondly, we input the QA and the generated image/caption to the LLM (i.e., mPlug-OWL [8])for visual/textual question answering. Assuming the LLM is powerful and can answer the question correctly, those generated images/captions noisy to the AVS query can be removed and would not proceed to the following search. Fig. 2 shows an example of the QA generation result for the query-745 Find shots of a person wearing gloves while biking. The QAs look promising. More importantly, GPT4 performs the QA generation without hallucination. Specifically, when the information that does not appear on the AVS query (e.g., is the person wearing a hat for query-745), the answer is "not specific". In the verification, we remove those QAs with the answer "not



A person wearing gloves while biking

Q: What is the person wearing?
A: Gloves.
Q: What activity is the person involved in?
A: Biking.
Q: Is the person wearing gloves?
A: Yes.
Q: Is the person wearing a hat?
A: Not specified.
Q: Is the person biking?
A: Yes.
Q: Is the person walking?
A: No.
...

Figure 2: An example of QA generation for the query-745 *Find shots of a person wearing gloves while biking.*

specific". The remaining QAs, together with the generated images/captions, are sent to LLM and we keep those images/captions with the best correctness of QAs for the subsequent search.

2.3 Reranking by decoding concepts

After getting the rank list of videos using either image-to-video search or caption-to-video search, we rerank the top 1500 retrieved videos by asking YES/NO QAs and answering by their decoded concepts. For example, for the question "Is the person biking ?", we check whether the concepts *person* and *biking* are in the decoding concept list of the video. Only when both of them appear, the video gets the answer yes to this question. Eventually, the video with the highest number of correctness on QAs ranks first.

2.4 Rewriting query by GPT4 for manual run

For the manual run, we ask GPT4 to rewrite the AVS query several times. We manually check the word frequency in the training set and pick one rewritten sentence with frequently shown words, as it should be better trained for retrieval than others.

3 Results analysis

In this year's AVS benchmarking, the evaluation is conducted on the V3C2 dataset [9] with 20 queries.

3.1 The impact of the generation

Fig. 3 shows the comparison of (i.e., *F_M_C_D_VIREO.23_1*), image-to-search (i.e., *F_M_C_D_VIREO.23_2*) and their fusion (i.e., *F_M_C_D_VIREO.23_3*), and the retrieval performances are driven by the quality of generated image. For most of the queries, when the generated images are related to the AVS query, the performance of either search mode is promising. For example, for the query-741 searching for red or blue scarf around people's neck, our performance doubles the average xinfAP of all other teams due to good-quality images and captions (as shown in Fig. 1). However, when the images

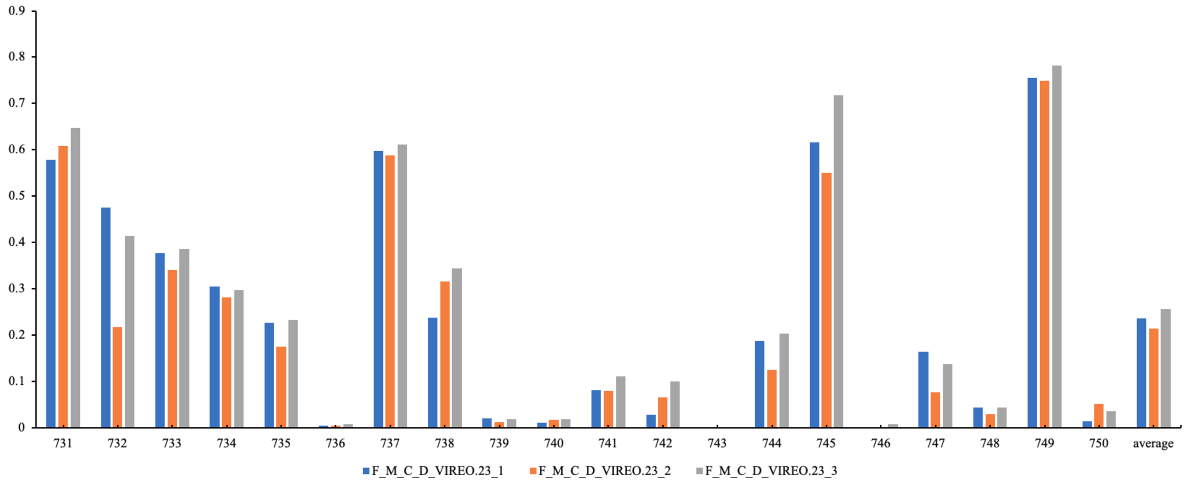


Figure 3: Comparison of our three automatic runs on 20 queries of the tv23 query set. Run 1, 2 and 3 correspond to the caption-to-search search, image-to-search search, and fusion of them, respectively.

cannot represent as much rich information as videos (e.g., multiple motion information), the retrieval performance cannot be guaranteed. For example, for the query-736 *Find shots of a person opens a door and enters a location*, the generated images can only cover one motion. Thus, our model cannot solve this query fully because it involves two actions of the same person. Similarly, the generated captions usually represent partial information, such as "a person opening a door". Also, our system will fail if the stable diffusion model cannot understand the query properly. For example, the query-743 *Find shots of a man talking in a small window located in the lower corner of the screen* is misunderstood, and the model generates images of a man behind a window of a building instead of a screen window, then none of the search mode could retrieve correct videos.

Overall, in the comparison of caption-to-search (i.e., *F_ M_ C_ D_ VIREO.23_ 1*) and image-to-search (i.e., *F_ M_ C_ D_ VIREO.23_ 2*), the textual search have better performance than the visual search on most of the queries, and the fusion of them obtain the best performance.

3.2 Impact of reranking

We compare the performances of before and after reranking on Table 1 using three baselines (i.e., BLIP2, CLIP and Imagebind) with the original AVS queries. The proposed reranking strategy based on decoded concepts brings consistent overall improvements to all three baselines on the tv23 query set. However, for some queries, reranking makes dramatic drops. For example, for the query-746 *Find shots of a man riding a scooter*, all of the three baselines drop by more than 200%. Similarly, for our submission to the TRECVid AVS 2023, we have bad performance on this query, although the generated images and captions are good (as shown in Fig. 1). The reason is that we use imprecise QAs for this query. For example, there is a question asking "Is the person riding a motorbike ?" and the expected answer generated by GPT4 is "No". However, scooter is one of the motorbike. Thus, all the correct videos are pulled down from the top of the rank list (as shown in Fig. 4).

3.3 Impact of the manually modified queries

This year, our manual runs are not able to outperform our automatic runs, and the xinfAP of the image-to-video manual run is problematic due to a coding error. To better evaluate the impact of the

Table 1: Performance comparison before and after reranking on three baselines. The improved performances are in bold.

	731	732	733	734	735	736	737	738	739	740	741	742	743	744	745	746	747	748	749	750	average
BLIP2_ori	0.425	0.206	0.552	0.274	0.238	0.041	0.423	0.216	0.038	0.021	0.031	0.074	0.000	0.077	0.504	0.219	0.129	0.033	0.494	0.057	0.203
BLIP2_rerank	**0.463**	**0.249**	0.523	**0.276**	0.226	0.026	0.400	**0.251**	0.021	**0.031**	**0.064**	**0.080**	0.000	0.068	**0.646**	0.064	0.113	0.026	**0.616**	**0.105**	**0.212**
CLIP-L_14@336px_ori	0.247	0.284	0.285	0.149	0.132	0.030	0.208	0.063	0.020	0.067	0.026	0.086	0.000	0.076	0.411	0.076	0.035	0.034	0.270	0.006	0.125
CLIP-L_14@336px_rerank	**0.285**	**0.342**	**0.382**	**0.264**	0.118	0.017	0.203	**0.137**	0.014	**0.080**	**0.052**	**0.096**	0.000	0.056	**0.486**	0.016	**0.038**	0.031	**0.423**	**0.037**	**0.154**
imagebind_ori	0.348	0.282	0.146	0.162	0.156	0.092	0.388	0.044	0.031	0.047	0.016	0.066	0.000	0.105	0.382	0.122	0.080	0.032	0.276	0.021	0.140
imagebind_rerank	**0.401**	**0.304**	**0.304**	**0.234**	**0.159**	0.051	0.382	**0.115**	0.027	**0.078**	**0.027**	**0.079**	0.000	0.081	**0.472**	0.023	0.070	0.023	**0.449**	**0.071**	**0.167**

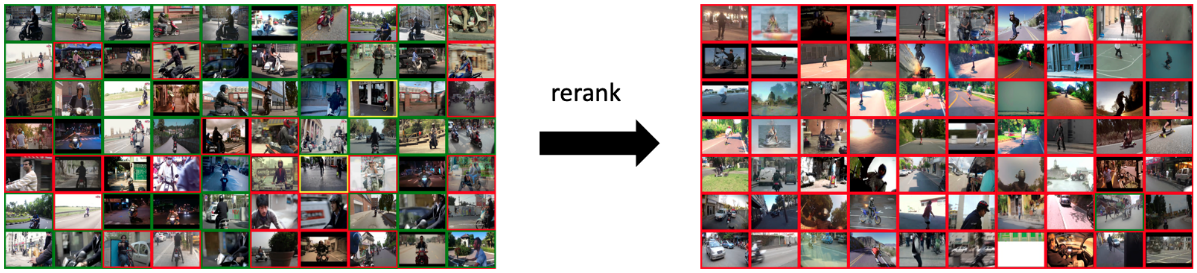


Figure 4: Visualization of the rank lists before and after reranking on the query-746 *Find shots of a man riding a scooter*. Green, red, and yellow borders mean correct, wrong and not judged videos, respectively.

manual query, we re-do the experiments and re-evaluate the results of original and manual queries on caption-to-video and image-to-video searches using the official ground truth. Moreover, to remove the effect of the reranking, in Fig. 5, we report the performances without the reranking process, and the original and modified queries are listed in Table 2. As shown in Fig. 5, the manual queries and the original queries obtain relatively the same xinfAP in terms of caption-to-video search. However, the manual run obtains lower performance than the original on the image-to-video search. The performance increase or decrease trends from automatic to manual runs are consistent on both caption-to-video search and image-to-video search for most queries. For some queries, the rephrased version obtains better results than the original version, such as query-745 and query-750. However, for some queries, there are dramatic drops, e.g., query-735 and query-738. It is hard to conclude what kinds of queries are effective to search, even though we have manually checked the word frequency in the training set. Besides, we also use the manual queries for the novel run but the submission to the TRECVid also problematic because of coding. The result is 0.217 after re-do the experiment.

4 Conclusion

Our study this year aims to explore generative models pre-trained on large datasets in understanding AVS queries. We find that the generative model can understand most of the AVS queries, and using either generated image-to-video search or generated caption-to-video search can have good retrieval performances. Also, the retrieved results of the two modes are complementary, and the fusion of them obtains the best performance. However, the proposed search pipeline will fail if a query is misunderstood or a motion query can not be fully represented by a static image. Besides, we also apply the LLM, e.g., GPT4, to understand the query in the manual run. However, the rewritten query does not bring consistent improvement to all queries, and it is still hard to predict which kind of rephrased query is more effective in video retrieval.

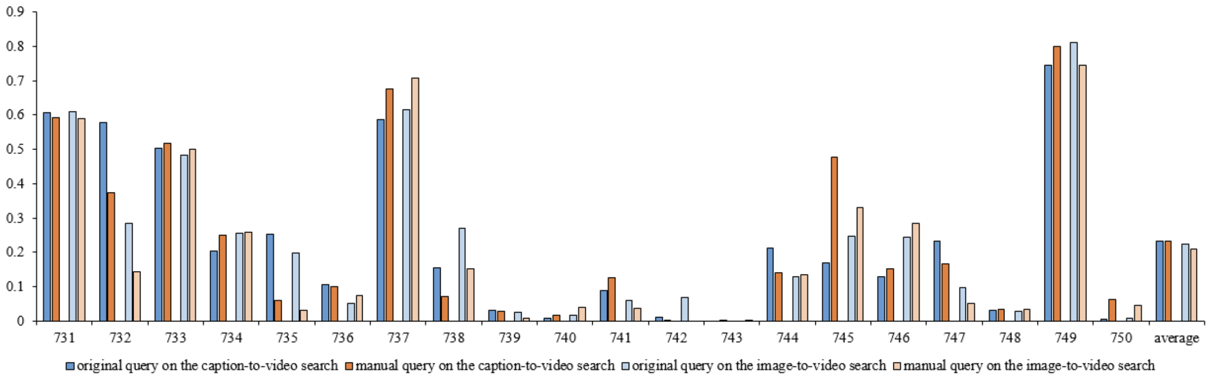


Figure 5: Comparison of the original and manual queries on the tv23 query set.

Table 2: Original and manual AVS queries on tv23 query set.

query id	original query	manual query
731	a man is seen with a baby	father holding a baby
732	a woman with red hair	woman with red hair
733	a golf course	playing golf at golf course
734	a recording studio	a recording studio with a keyboard, monitor, chair and speakers
735	a toy vehicle	toy truck, toy train, toy car
736	a person opens a door and enters a location	a person open a door and walk through door
737	a woman wearing (dark framed) glasses	a woman wearing black frame glasses
738	a police officer wearing a helmet	a police officer wearing a helmet and uniform
739	two or more persons are seen in front of a chain link fence	two persons in front of a chain link fence
740	a heavy man indoors	fat man indoors
741	a red or blue scarf around someone's neck	person wearing red scarf or blue scarf
742	a child climbs an object outdoors	a kid is climbing outdoors
743	a man is talking in a small window located in the lower corner of the screen	a man talking in a slideshow
744	a person taking picture using a cell phone camera	a person taking picture or selfie using a cellphone
745	a person wearing gloves while biking	a close up of a person riding a bike wearing gloves
746	a man riding a scooter	a man riding a moving scooter
747	at least two persons are working on their laptops together in the same room indoors.	two people sitting at a table with laptops indoors
748	a man carrying a bag on one of his shoulders (excluding backbags)	a man is carrying a messenger bag, camera bag, duffel bag on one shoulder
749	a person wearing any kind of face or head mask	a person wearing mask
750	a man with an earring in his left ear	a man with ear piercings

5 Acknowledgments

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

References

[1] G. Awad, K. Curtis, A. A. Butt, J. Fiscus, A. Godil, Y. Lee, A. Delgado, E. Godard, L. Diduch, Y. Graham, , and G. Quénot, "Trecvid 2023 - a series of evaluation tracks in video understanding," in *Proceedings of TRECVID 2023*. NIST, USA, 2023.

[2] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *ArXiv*, vol. abs/2301.12597, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256390509

[3] 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," in *ICML*, 2021.

[4] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, “Imagebind: One embedding space to bind them all,” in *CVPR*, 2023.

[5] J. Wu, C.-W. Ngo, W.-K. Chan, and Z. Hou, “(Un)likelihood training for interpretable embedding,” in *ACM Transactions on Information Systems*, 2023.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2022, pp. 10 674–10 685.

[7] OpenAI, “GPT-4 technical report,” *CoRR*, vol. abs/2303.08774, 2023.

[8] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Jiang, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qi, J. Zhang, and F. Huang, “mplug-owl: Modularization empowers large language models with multimodality,” 2023.

[9] L. Rossetto, H. Schuldt, G. Awad, and A. A. Butt, “V3C–a research video collection,” in *International Conference on Multimedia Modeling.* Springer, 2019, pp. 349–360.