

# Understanding AVS query by generating images and asking questions

**Jiaxin (Nikki) Wu**, Zhixin Ma, Sheng-Hua Zhong, Chong-Wah Ngo

City University of Hong Kong

Singapore Management University

Contact: [jiaxin.wu@my.cityu.edu.hk](mailto:jiaxin.wu@my.cityu.edu.hk)

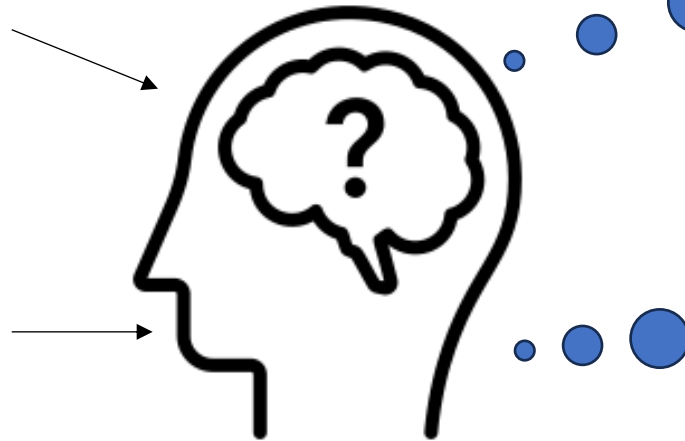


# Key challenge in AVS task

- Understand the query and imagine it visually

A well-trained query:  
a woman is eating something outdoors

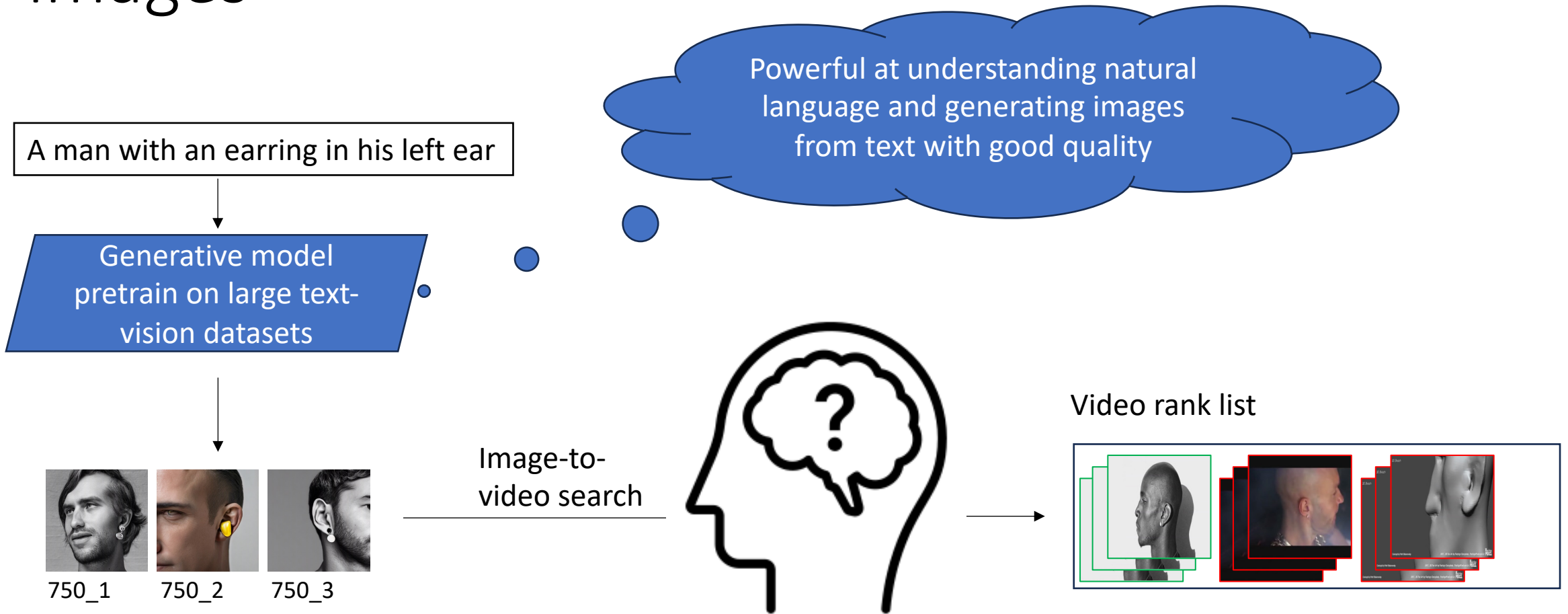
An ill-trained query:  
a person wearing a light t-shirt with  
dark or black writing on it



What is a light t-shirt?  
what is dark? T-shirt or  
writing?

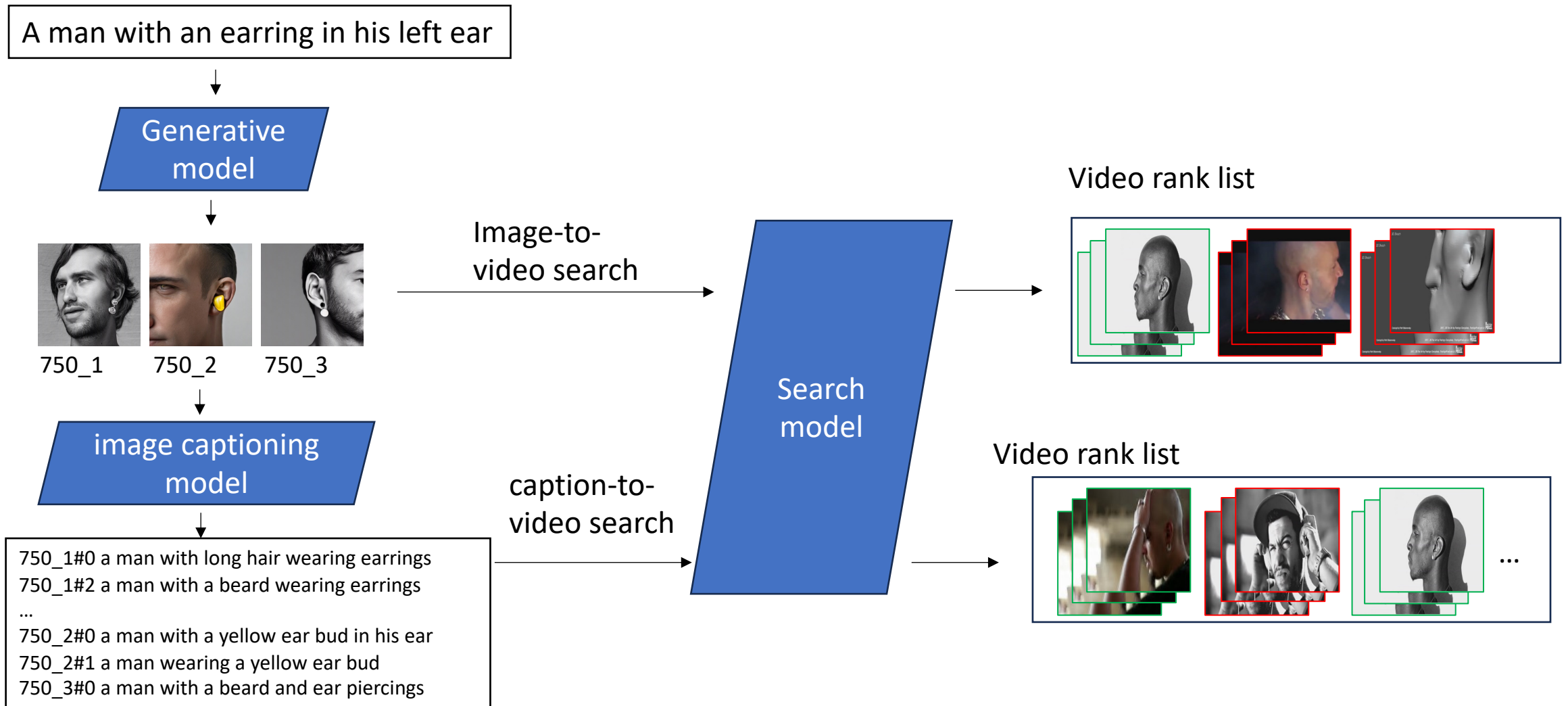
Search model trained on text-video pairs

# Understanding AVS query by generating images

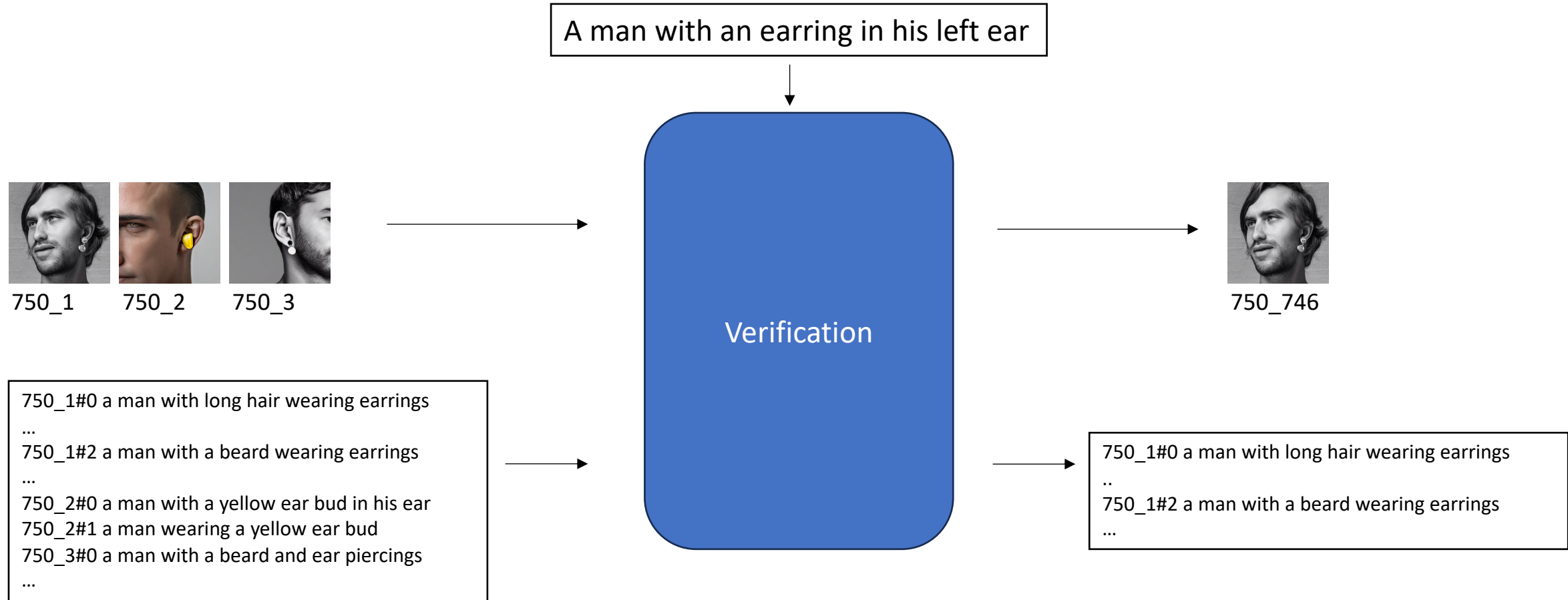


Search model trained on text-video pairs


# Understanding AVS query by generating captions



# Remove the noise in the generation



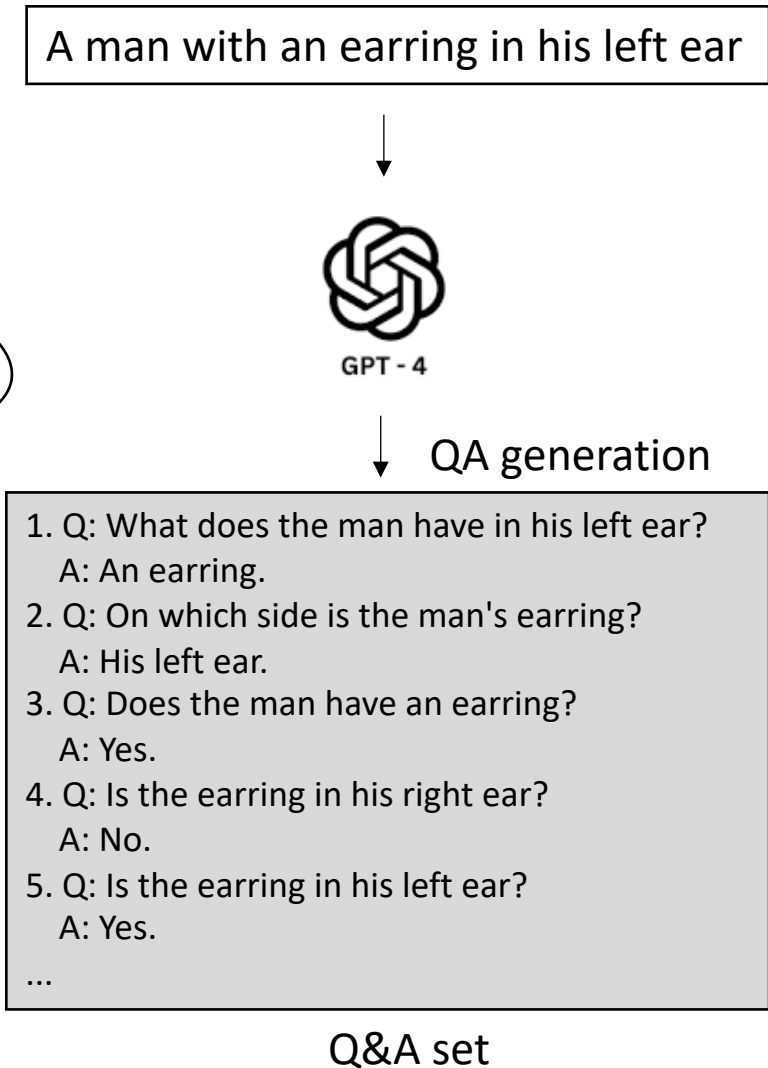
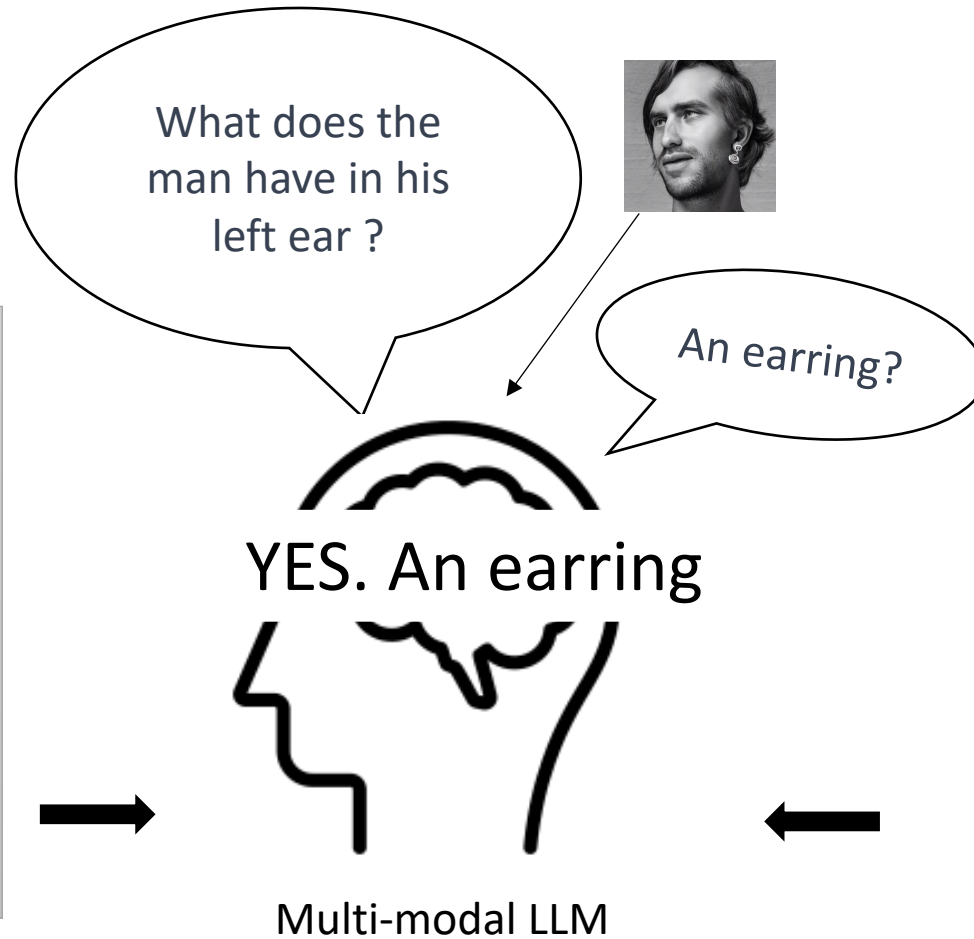
# Verifying the generation by asking questions



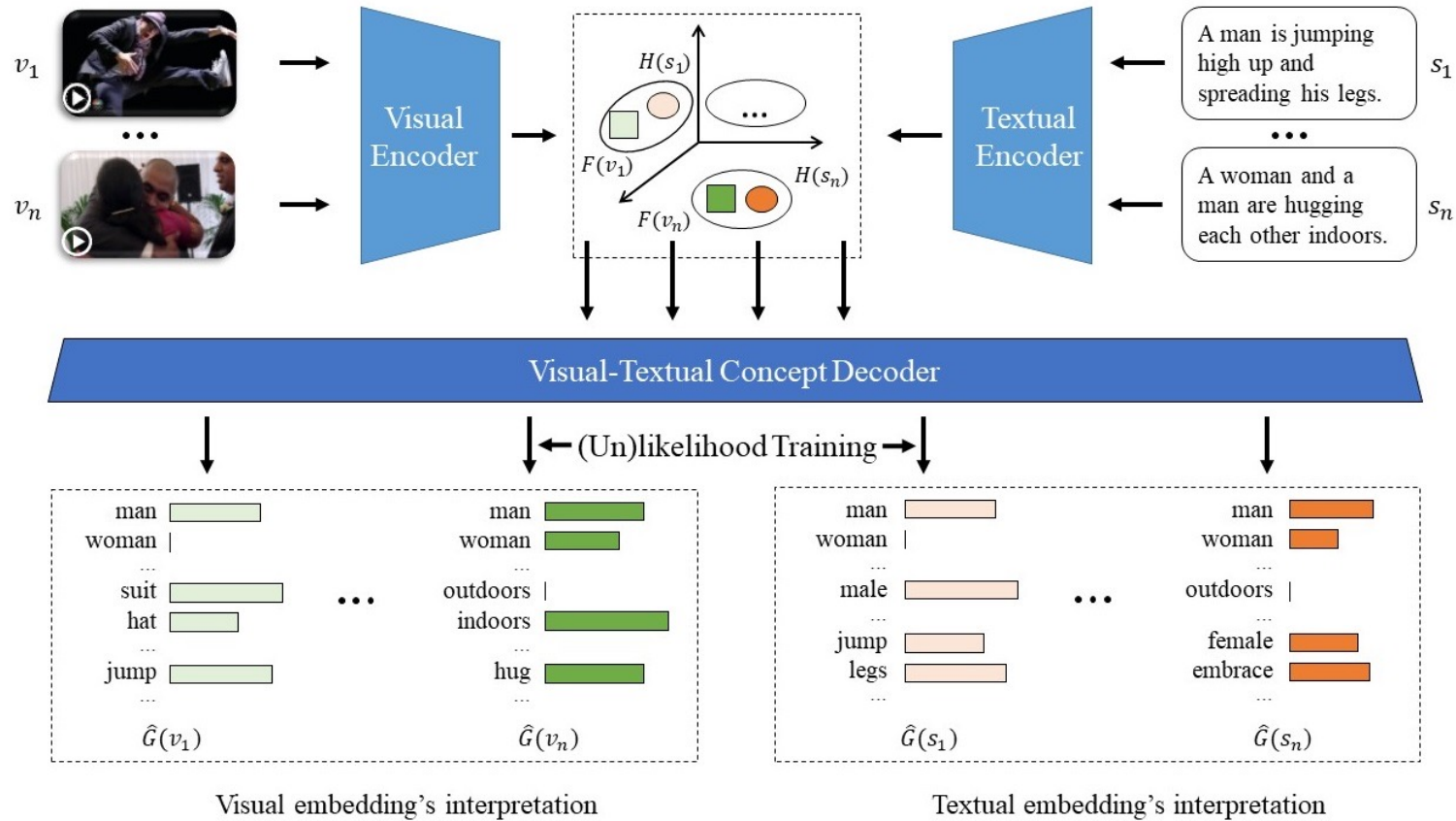
750\_1    750\_2    750\_3

750\_1#0 a man with long hair wearing earrings  
750\_1#2 a man with a beard wearing earrings  
...  
750\_2#0 a man with a yellow ear bud in his ear  
750\_2#1 a man wearing a yellow ear bud  
750\_3#0 a man with a beard and ear piercings

Items waiting for verification

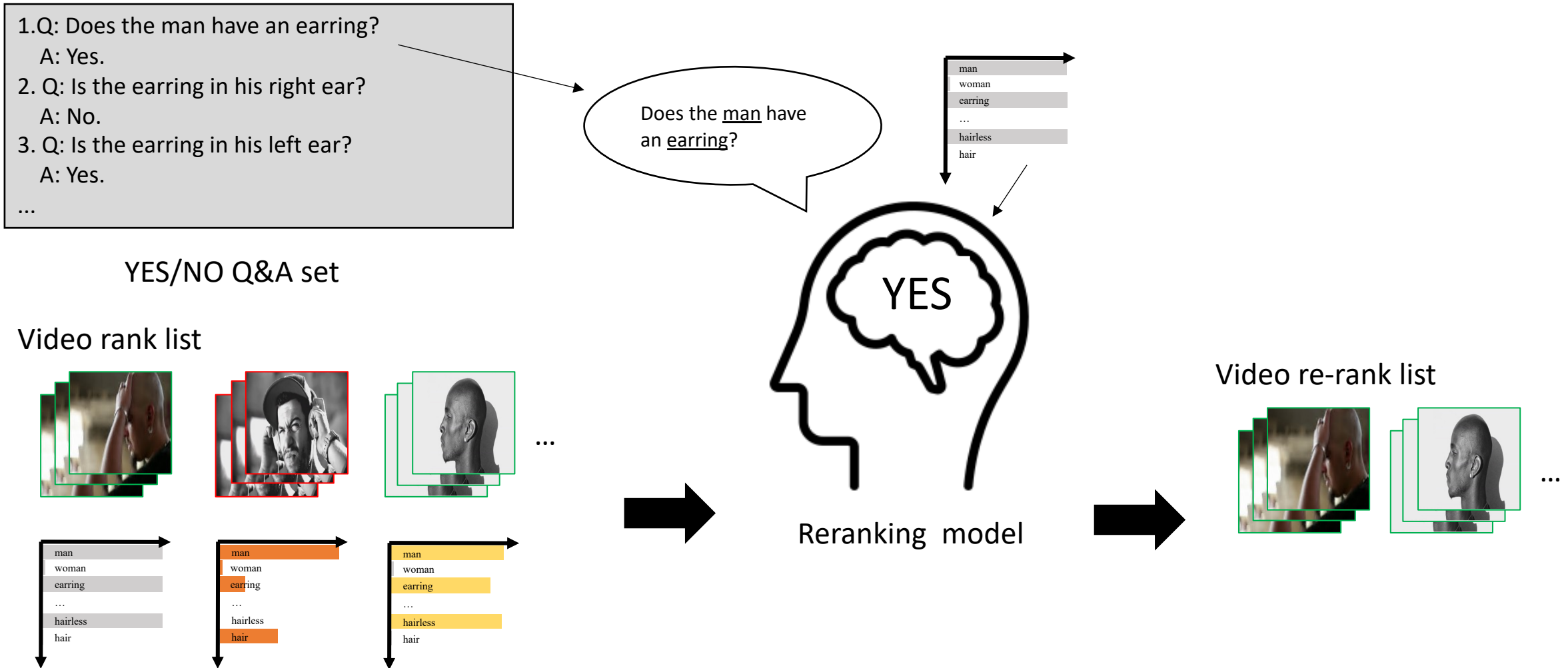


# Core search model: the ITV model



- Highlights:
  - ITV model enables embedding-based search and concept-based search in a unified encoder-decoder framework.
  - Fusion search (concept search + embedding search) obtains the best retrieval performance.
  - Provides consistent and coherent interpretations for video and text embeddings

# Re-ranking video rank list by asking question and answering based on decoded concepts





# Proposed pipeline

A man with an earring in his left ear

Generative model



750\_1    750\_2    750\_3

image captioning model



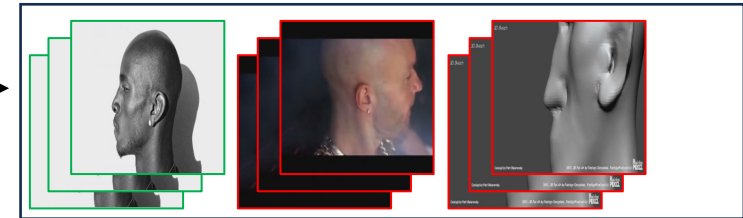
750\_1#0 a man with long hair wearing earrings  
750\_1#2 a man with a beard wearing earrings  
...  
750\_2#0 a man with a yellow ear bud in his ear  
750\_3#0 a man with a beard and ear piercings

Image-to-video search

caption-to-video search

ITV model

Video rank list



Video rank list



Re-rank

# AVS23 submissions

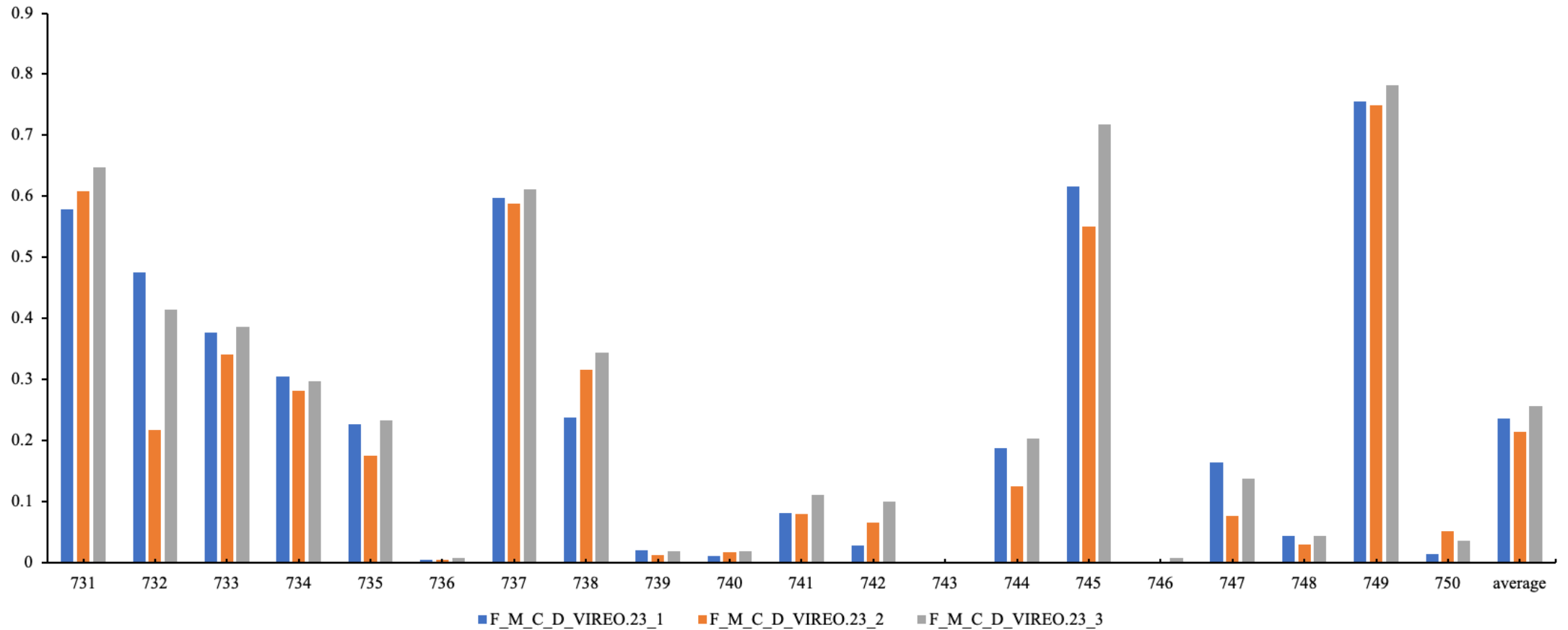
- Run 1: generated captions-to-video search (fusion search of ITV)
- Run 2: generated image-to-video search (visual similarity search)
- Run 3: Run 1 + run 2
- Run 4: Run 3 + BLIP2 + CLIP + Imagebind
- Run 5 (concept run): generated captions-to-video search (concept-based search in ITV)

BLIP2: Li et al., “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models”, arXiv, 2023

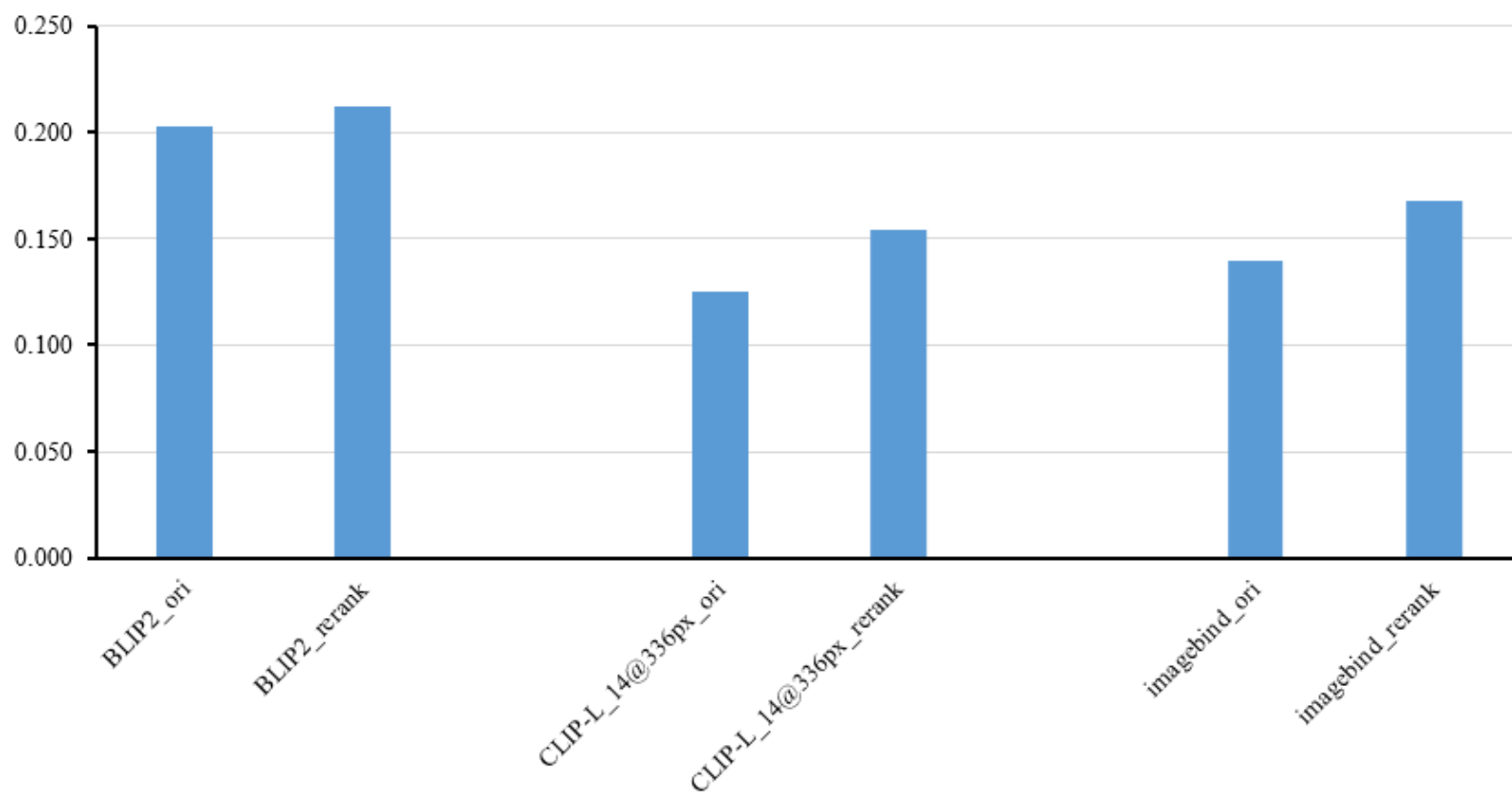
CLIP: Radford et al., “Learning transferable visual models from natural language supervision,” in ICML, 2021.

Imagebind: Girdhar et al., “Imagebind: One embedding space to bind them all,” in CVPR, 2023

# Results: Run 3 > Run 1 > Run 2



# Before and after re-ranking



# Examples of the generated images and captions

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

# Reason for the bad performance—bad QA

746 A man riding a scooter



xinfAP  
=0.200

↓ Rerank

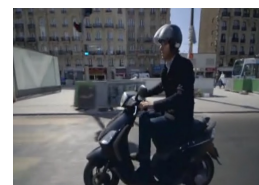


xinfAP  
=0.002

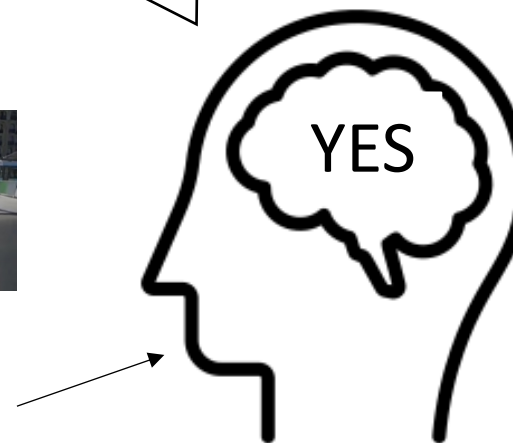
1.Q: Is the man riding a scooter ?  
A: Yes.  
2. Q: Is the man riding a motorbike ?  
A: No.  
...

YES/NO Q&A set

Is the man riding a motorbike ?



- man
- woman
- motorbike
- ...
- scooter
- outdoors



Reranking model

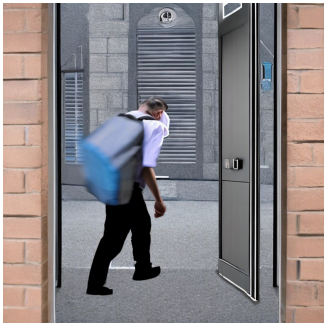
QA generation has an error. Scooter is a subset of motorbikes.

# Limitation of the static image

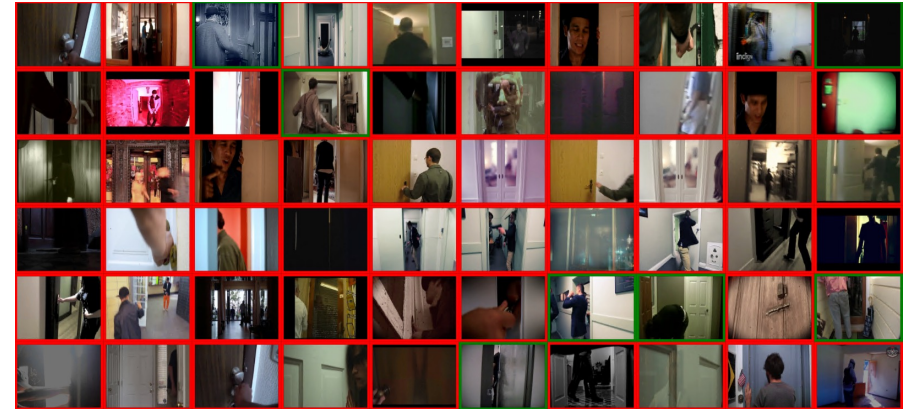
query-736 A person opens a door and enters a location



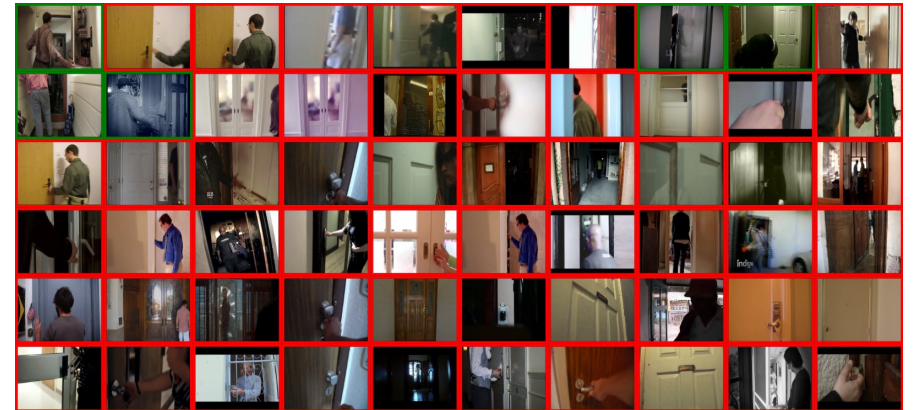
BLIP\_736\_1#0 a man in a red shirt is opening a door



BLIP\_736\_5#0 a man with a backpack walking through an open door



Run 1 xinfAP=0.005



Run 2 xinfAP=0.005

# The generative model misunderstands the search intent

743 A man is talking in a small window located in the lower corner of the screen

- Generated images



- Retrieved video rank list





# Conclusions

- 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 cannot be fully represented by a static image.