Waseda_Meisei_SoftBank at TRECVID 2023
Ad-hoc Video Search

Kazuya Ueki (presenter)
Meisei University, Waseda University

Takayuki Hori,
Softbank Corporation, Waseda University

Yuma Suzuki, Hiroki Takushima, Haruki Sato,
Takumi Takada, Hideaki Okamoto, Hayato Tanoue
Aiswariya Manoj Kumar
Softbank Corporation

TRECVID 2023 Workshop   November 13th, 2023

1. Highlights

- Overview

Highlights

- **Submission type**
  - ✓ Fully-automatic
- **Basic approach**
  - ✓ Visual-semantic embedding approach

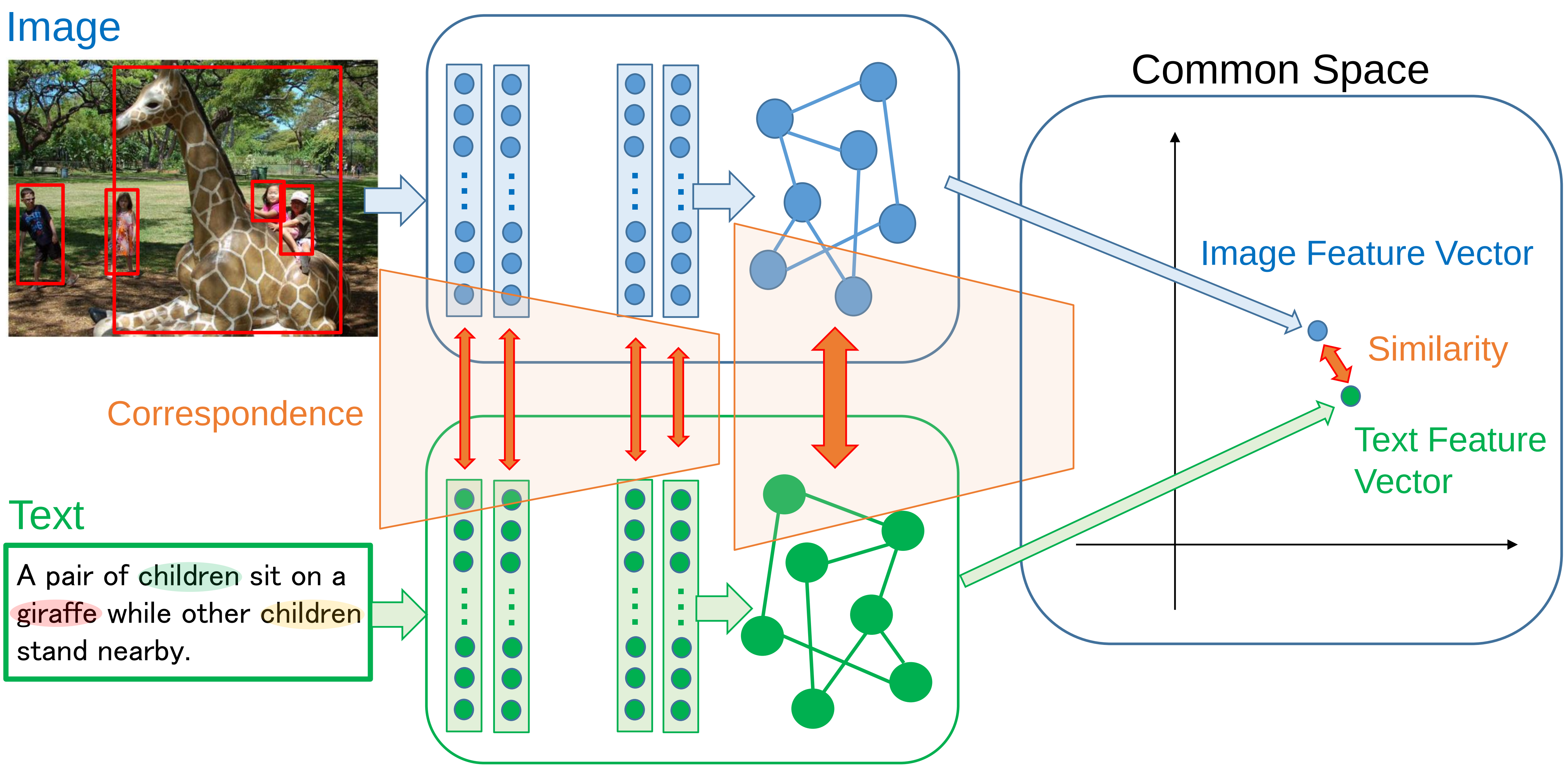
Fusion of multiple embedding models

- VSE++
- GSMN
- CLIP
- SLIP

⋮

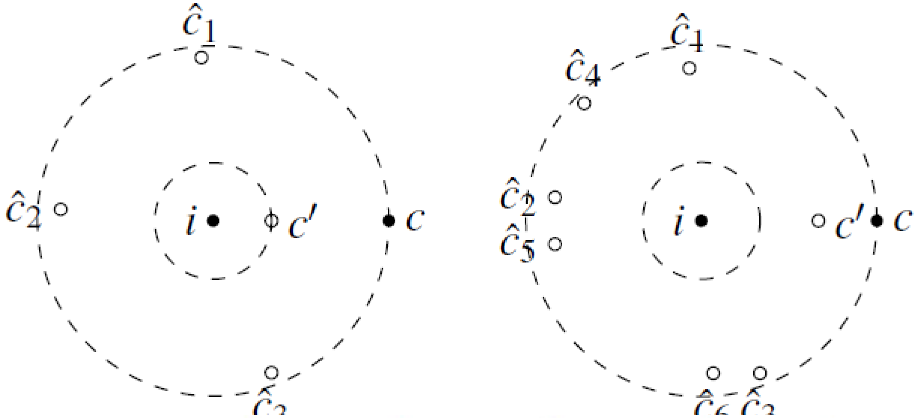
- **This year's update**
  - ✓ Implementation of the latest pre-trained models provided by OpenCLIP
  - ✓ Query expansion by generative language model
- **Results**
  - ✓ 2nd position for main task
  - ✓ Best for progress subtask

Visual-semantic embedding approach



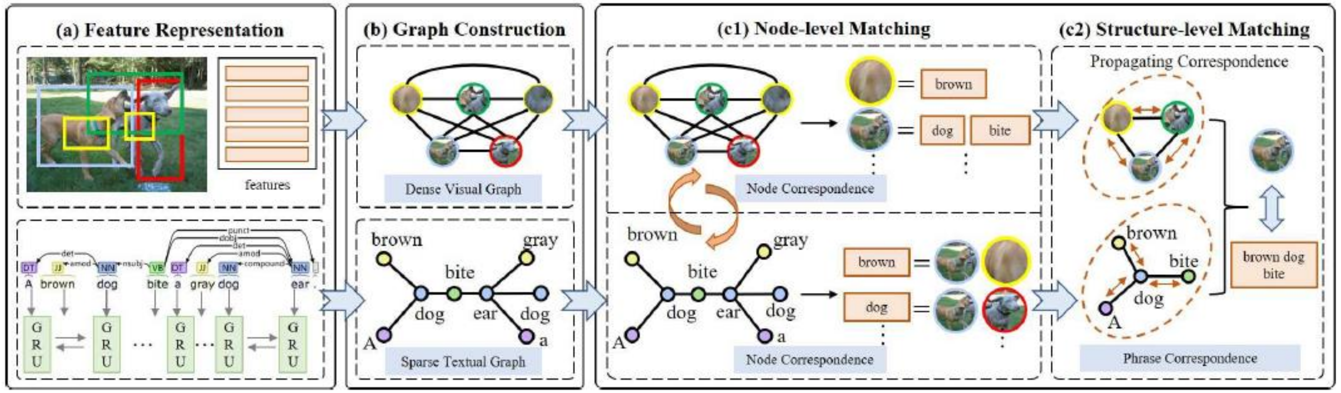
Representative approaches

VSE++ [Faghri+, 2018]

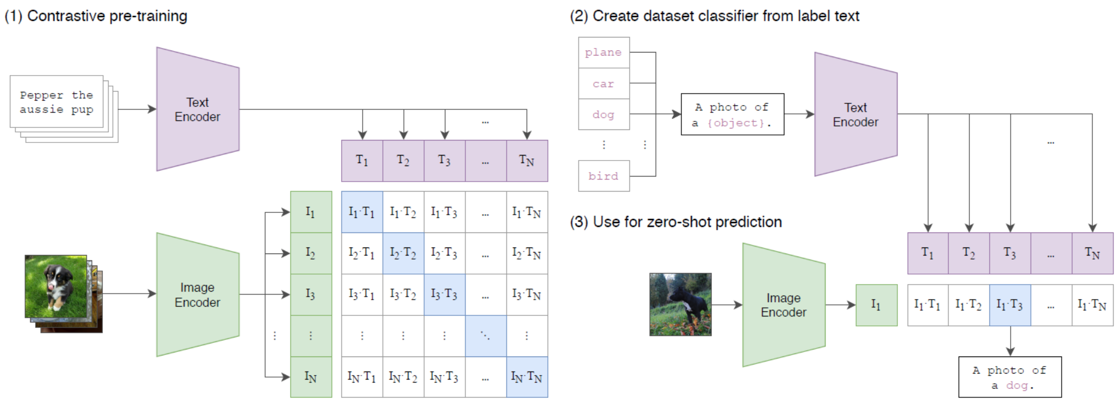


$$\ell_{MH}(i,c) = \max_{c'}\left[\alpha + s(i,c') - s(i,c)\right]_+ + \max_{i'}\left[\alpha + s(i',c) - s(i,c)\right]_+$$

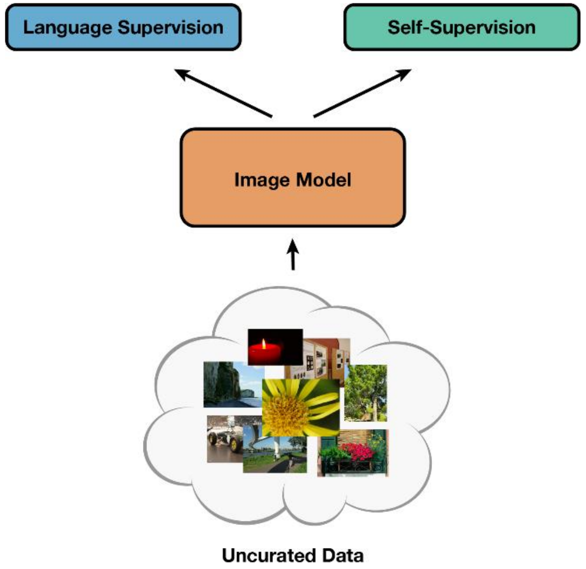
GSMN [Liu+, 2020]



CLIP [Radford+, 2021]



SLIP [Mu+, 2021]



Zero-shot video retrieval techniques

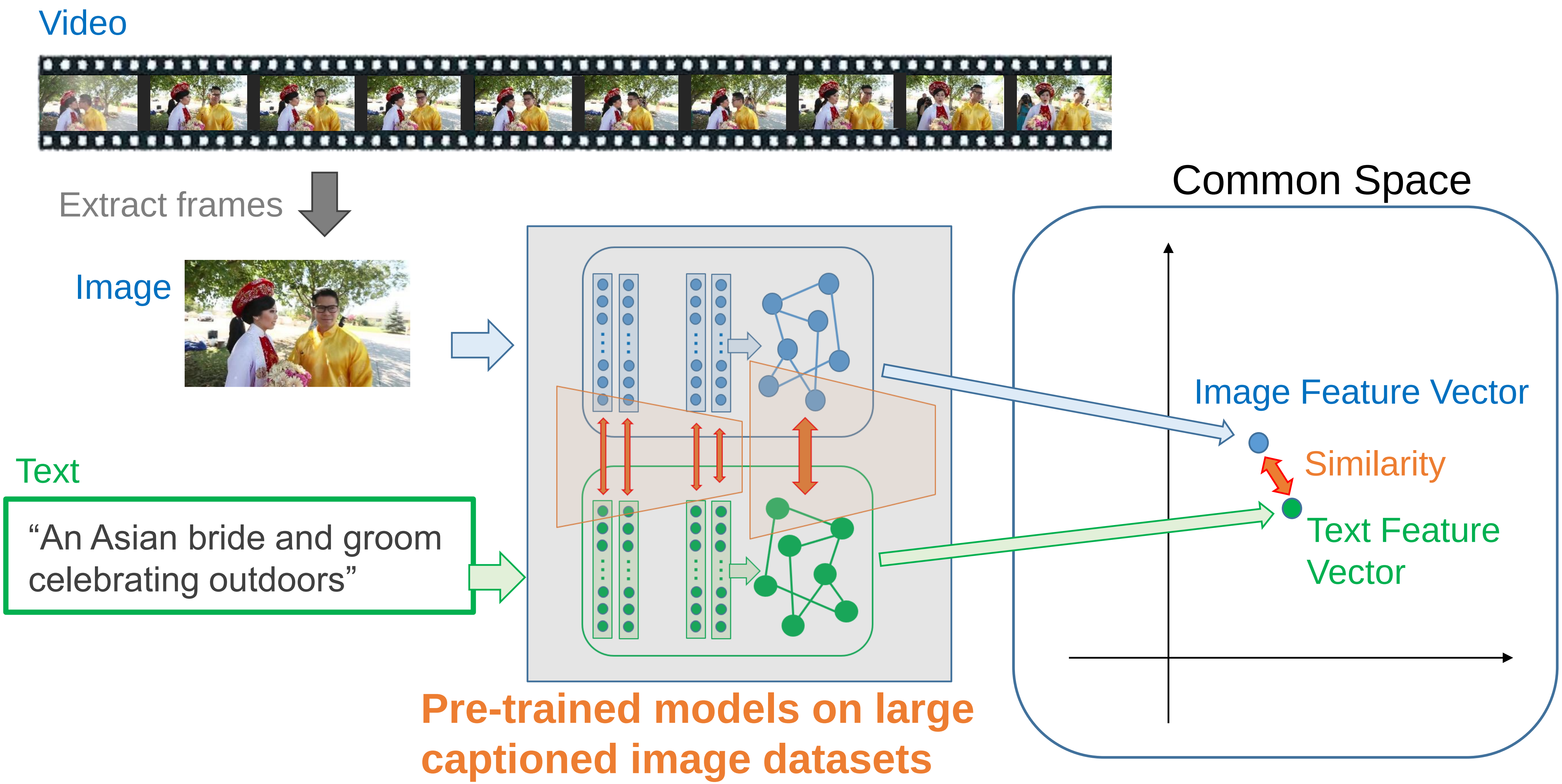
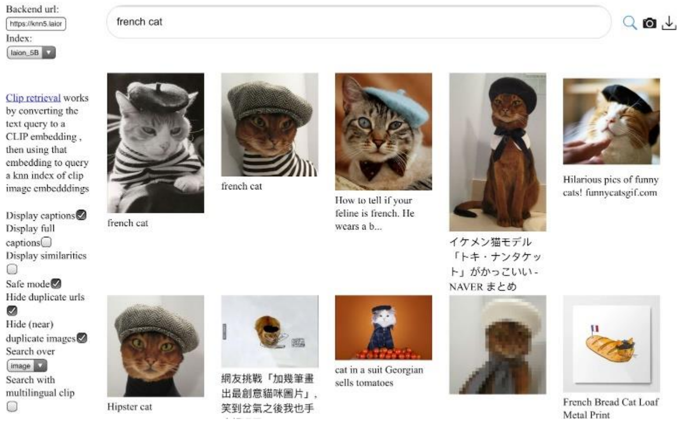


Image caption datasets

Conceptual 12M

LAION-5B



YFCC100M

>>

Video caption datasets

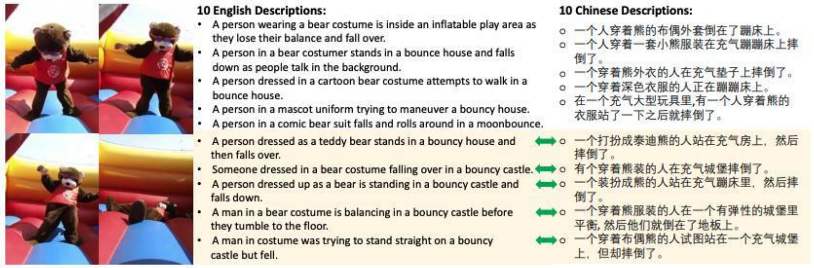
MSR-VTT

10,000 videos
200,000 captions

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

VATEX

41,250 videos
825,000 captions



While there is an abundance of image datasets with captions, the availability of video datasets with accompanying captions is not as extensive.

2. This year's update

- Implementation of the latest pre-trained models provided by OpenCLIP
- Query expansion by generative language model

OpenCLIP

https://github.com/mlfoundations/open_clip/blob/main/docs/openclip_results.csv

open_clip / docs / openclip_results.csv

gabrielilharco and rwightman Fix typo ✓ 91923df · last week Histor

Preview Code Blame 120 lines (120 loc) · 37.3 KB Raw

Search this file

name	pretrained	params (M)	FLOPs (B)	Average perf. on 38 datasets	ImageNet 1k	Caltech-101	CIFAR-10	CIFAR-100	CLEVR Counts	CLEVR Distance
ViT-H-14-378-quickgelu	dfn5b	986.71	1054.05	0.7079	0.8437	0.9517	0.9880	0.9043	0.3596	0.2085
ViT-H-14-quickgelu	dfn5b	986.11	381.68	0.6961	0.8344	0.9552	0.9878	0.9051	0.2967	0.2117
EVA02-E-14-plus	laion2b_s9b_b144k	5044.89	2362.19	0.6930	0.8201	0.9535	0.9934	0.9316	0.2991	0.1998
ViT-SO400M-14-SigLIP-384	webli	877.96	723.48	0.6921	0.8308	0.9599	0.9672	0.8357	0.4071	0.2246
ViT-bigG-14-CLIPA-336	datacomp1b	2517.76	2271.58	0.6842	0.8309	0.9529	0.9904	0.9123	0.1399	0.2161
ViT-bigG-14-CLIPA	datacomp1b	2517.22	1007.93	0.6822	0.8270	0.9513	0.9912	0.9135	0.1357	0.2113
ViT-SO400M-14-SigLIP	webli	877.36	233.54	0.6808	0.8203	0.9600	0.9679	0.8417	0.4210	0.2213
EVA02-E-14	laion2b_s4b_b115k	4704.59	2311.42	0.6690	0.8196	0.9541	0.9925	0.9258	0.1632	0.2499
ViT-L-14-quickgelu	dfn2b	427.62	175.33	0.6687	0.8141	0.9532	0.9836	0.8837	0.3325	0.2481
ViT-L-16-SigLIP-384	webli	652.48	422.91	0.6683	0.8207	0.9611	0.9605	0.8188	0.3275	0.2077
ViT-H-14-CLIPA-336	datacomp1b	968.64	800.88	0.6677	0.8180	0.9467	0.9890	0.8968	0.1326	0.2254
ViT-H-14-quickgelu	metaclip_fullcc	986.11	381.68	0.6671	0.8051	0.9536	0.9804	0.8634	0.2115	0.1881
ViT-bigG-14	laion2b_s39b_b160k	2539.57	1065.36	0.6667	0.8009	0.9484	0.9824	0.8752	0.2989	0.2002

Currently, there are over 100 available models.

OpenCLIP provides a wide range of pretrained embedding models, and the updates come swiftly.

Retieival results on tv22 queries

CLIP (official)

Model name	mAP
RN50	0.0962
RN101	0.1112
RN50x4	0.1075
RN50x16	0.1053
RN50x64	0.0976
ViT-B/32	0.1060
ViT-B/16	0.1209
ViT-L/14	0.1010

OpenCLIP

Model name	Pre-trained	mAP
ViT-B-32	laion400m_e31	**0.1240**
	laion400m_e32	**0.1235**
	laion2b_e16	**0.1337**
	laion2b_s34b_b79k	**0.1425**
ViT-L-14	laion400m_e31	**0.1226**
	laion400m_e32	**0.1228**
	laion2b_s32b_b82k	**0.1460**

The pre-trained models available in OpenCLIP often exhibit higher accuracy compared to official CLIP models, thanks to their training on extensive datasets such as LAION.

Retieival results on tv22 queries

CLIP (official)

Model name	mAP
RN50	0.0962
RN101	0.1112
RN50x4	0.1075
RN50x16	0.1053
RN50x64	0.0976
ViT-B/32	0.1060
ViT-B/16	0.1209
ViT-L/14	0.1010
ViT-L/14@336px	0.0975
Fusion	**0.1941**

simple score fusion

Due to the complementary nature of each model, even a simple score fusion can lead to a significant improvement in accuracy.

This year, we adjusted the fusion weights based on the previous year's ground truth to optimize the mean average precision.

OpenCLIP

open_clip / docs / openclip_results.csv

https://github.com/mlfoundations/open_clip/blob/main/docs/openclip_results.csv

gabrielilharco and rwightman Fix typo ✓ 91923df · last week Histor

Preview | Code | Blame 120 lines (120 loc) · 37.3 KB Raw

Search this file

1	name	pretrained	params (M)	FLOPs (B)	Average perf. on 38 datasets	ImageNet 1k	Caltech-101	CIFAR-10	CIFAR-100	CLEVR Counts	CLEVR Distance
2	ViT-H-14-378-quickgelu	dfn5b	986.71	1054.05	0.7079	0.8437	0.9517	0.9880	0.9043	0.3596	0.2085
3	ViT-H-14-quickgelu	dfn5b	986.11	381.68	0.6961	0.8344	0.9552	0.9878	0.9051	0.2967	0.2117
4	EVA02-E-14-plus	laion2b_s9b_b144k	5044.89	2362.19	0.6930	0.8201	0.9535	0.9934	0.9316	0.2991	0.1998
5	ViT-SO400M-14-SigLIP-384	webli	877.96	723.48	0.6921	0.8308	0.9599	0.9672	0.8357	0.4071	0.2246
6	ViT-bigG-14-CLIPA-336	datacomp1b	2517.76	2271.58	0.6842	0.8309	0.9529	0.9904	0.9123	0.1399	0.2161
7	ViT-bigG-14-CLIPA	datacomp1b	2517.22	1007.93	0.6822	0.8270	0.9513	0.9912	0.9135	0.1357	0.2113
8	ViT-SO400M-14-SigLIP	webli	877.36	233.54	0.6808	[illegible]	0.9600	0.9679	0.8417	0.4210	[illegible]
9	EVA02-E-14	laion2b_s4b_b115k	[illegible]	[illegible]	[illegible]	[illegible]	0.9541	[illegible]	[illegible]	[illegible]	[illegible]
10	ViT-L-14-quickgelu	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
11	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.2989	0.2002

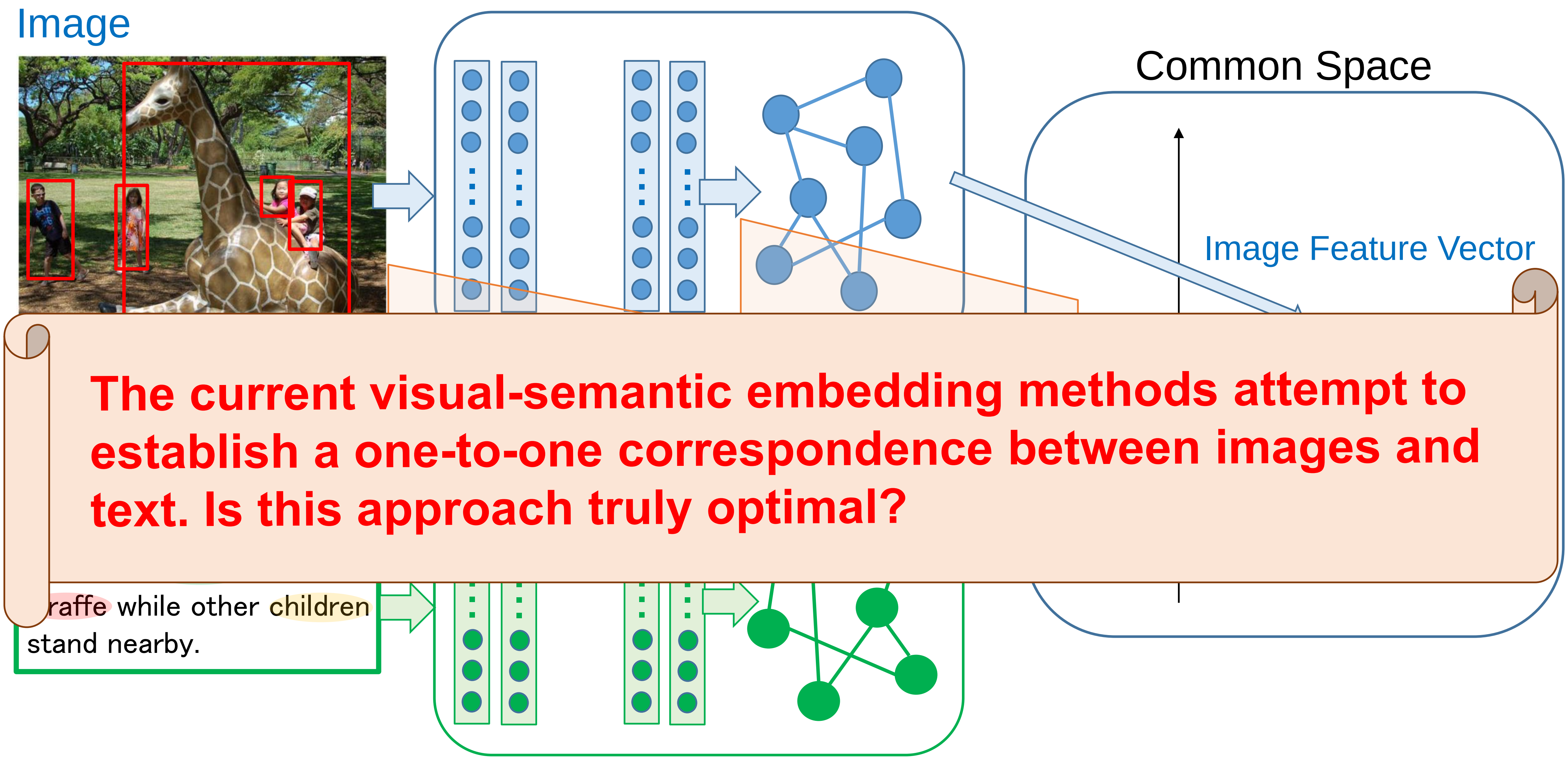
Currently, there are over 100 avail…

Research question

How can we improve the baseline accuracy of video retrieval using pre-trained image/text embedding models?

…CLIP provides a wide range of pretrained embedding models, and the updates come swiftly.

Visual-semantic embedding approach



The current visual-semantic embedding methods attempt to establish a one-to-one correspondence between images and text. Is this approach truly optimal?

One-to-many relationship

There is a dog in the room.

A lovely chestnut-colored miniature dachshund is sitting on the sofa.

There is a gray sofa inside the room.

Marron wants something to eat.

⋮

There are many textual representations for a single image.

⇩

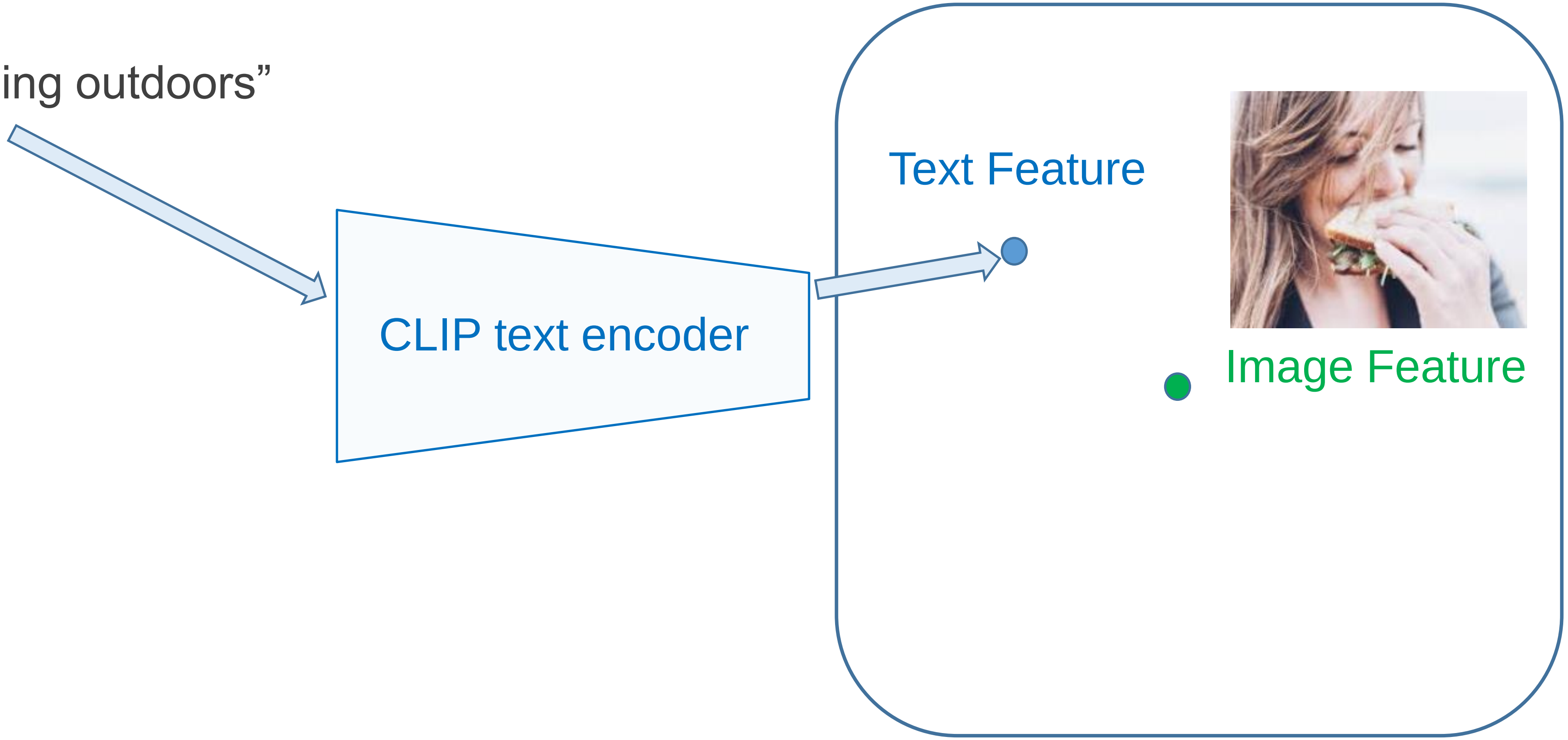
Isn't a rich variety of textual descriptions necessary to search for target videos?

How can we improve the video retrieval performance?

Query expansion!!

Query:

"A woman is eating something outdoors"



How can we improve the video retrieval performance?

Query expansion!!

Query:

"A woman is eating something outdoors"

"A female is having a meal
in the open air."

CLIP text encoder

Text Features

Image Feature

How can we improve the video retrieval performance?

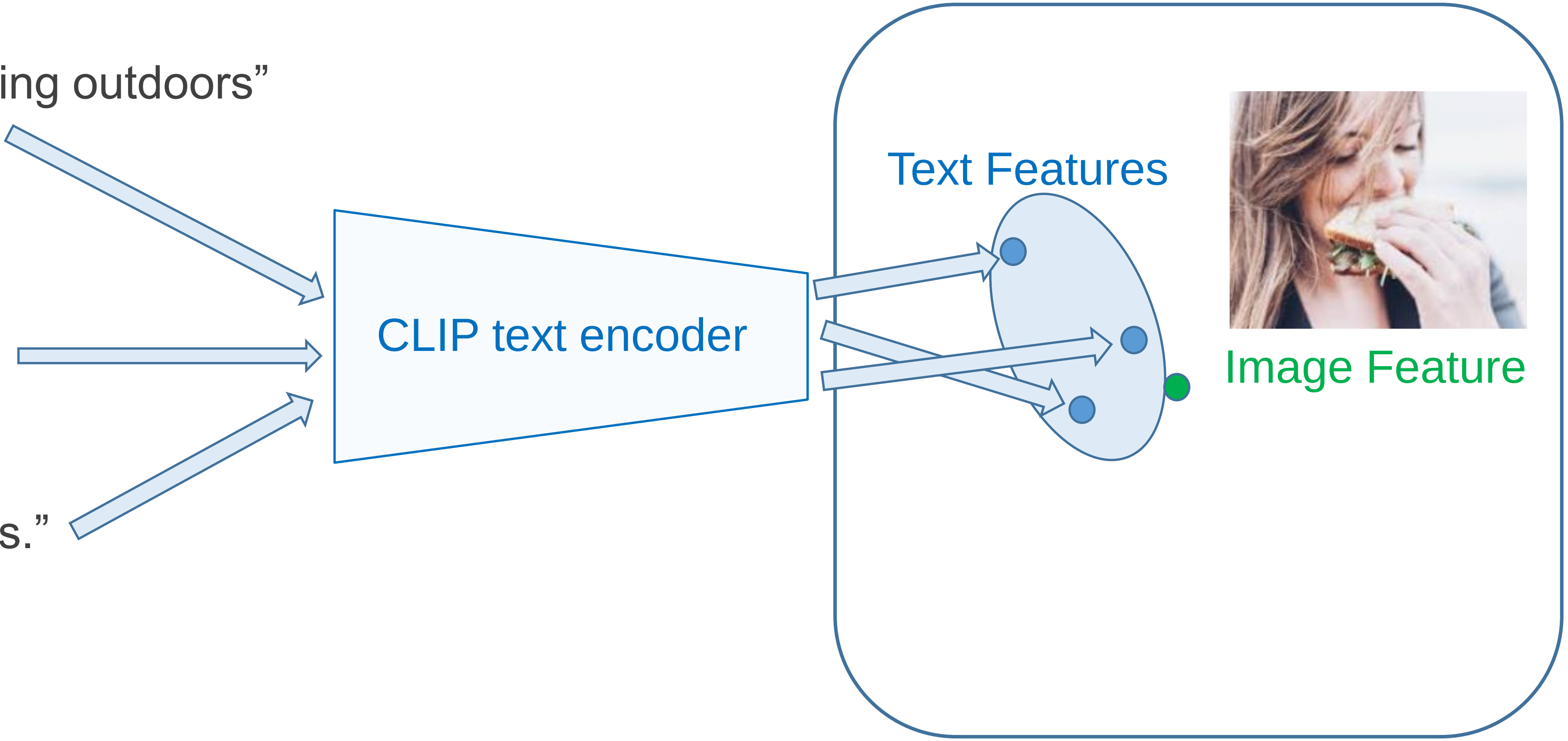
Query expansion!!

Query:

"A woman is eating something outdoors"

"A female is having a meal
in the open air."

"A female is dining outdoors."



How can we improve the video retrieval performance?

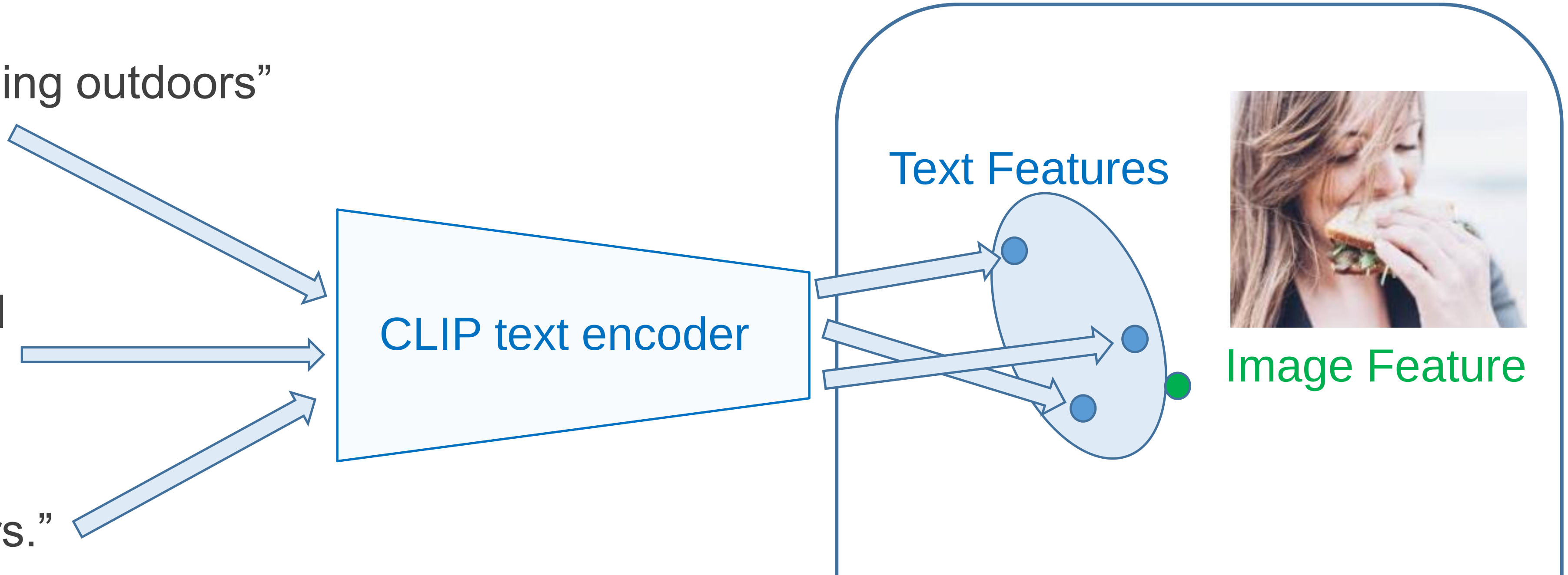
Query expansion!!

Query:

"A woman is eating something outdoors"

"A female is having a meal in the open air."

"A female is dining outdoors."



- The richer the variation of the input query sentence, the more videos can be retrieved.
- Generating sentences that have different expressions but convey the same meaning as the original query sentence is probably very effective.

Query Expansion using ChatGPT

Query

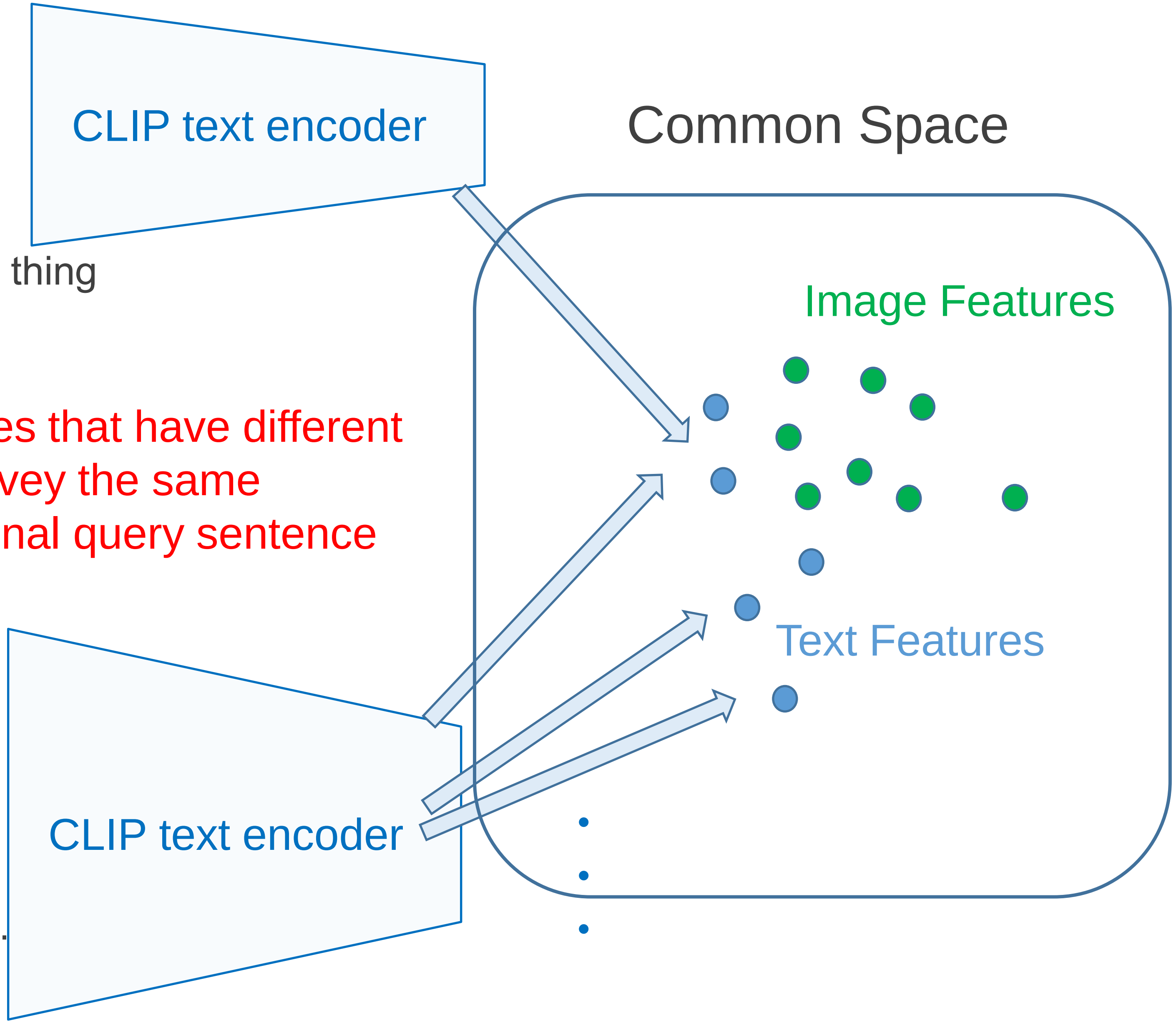
“A woman is eating something outdoors”

“Give me 10 expressions that mean the same thing as ‘query’ with a slight change of sentence”

ChatGPT

Generating sentences that have different expressions but convey the same meaning as the original query sentence

1. A female is consuming something in the open air.
2. A female is partaking of something al fresco.
3. A woman is feasting alfresco.
4. Outdoors, a woman is eating something.
5. Eating something outside, a woman.
6. A female is having a meal in the open air.
7. In the open air, a woman is eating something.
8. Something is being consumed by a woman outdoors.
9. A female is dining outdoors.
10. A woman is taking in something outside.



3. Experiment

- Query expansion using ChatGPT

Query Expansion using ChatGPT

Original query： “A clock on a wall in a room”

We attempted to input the five prompts into ChatGPT twice.

- Give me 10 sentences that mean exactly the same as “A clock on a wall in a room” with slight changes.
- Give me 10 examples of “A clock on a wall in a room” that means exactly the same thing, but with a slight change in the sentence.
- List 10 sentences that mean the same as “A clock on a wall in a room” with slight modifications.
- List 10 examples that mean the same thing as “A clock on a wall in a room,” but with a slight change in the sentence.
- Give me 10 sentences that mean the same as the following sentence with a slight change of wording: “A clock on a wall in a room”

ChatGPT

Generated sentences

1.
2.
3.
⋮
10.

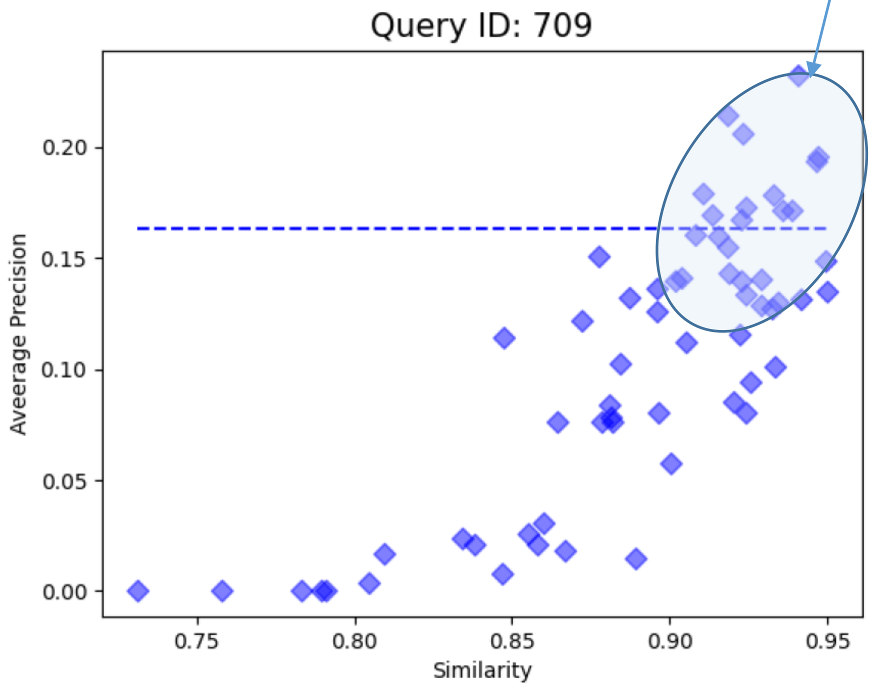
10 x 10 sentences

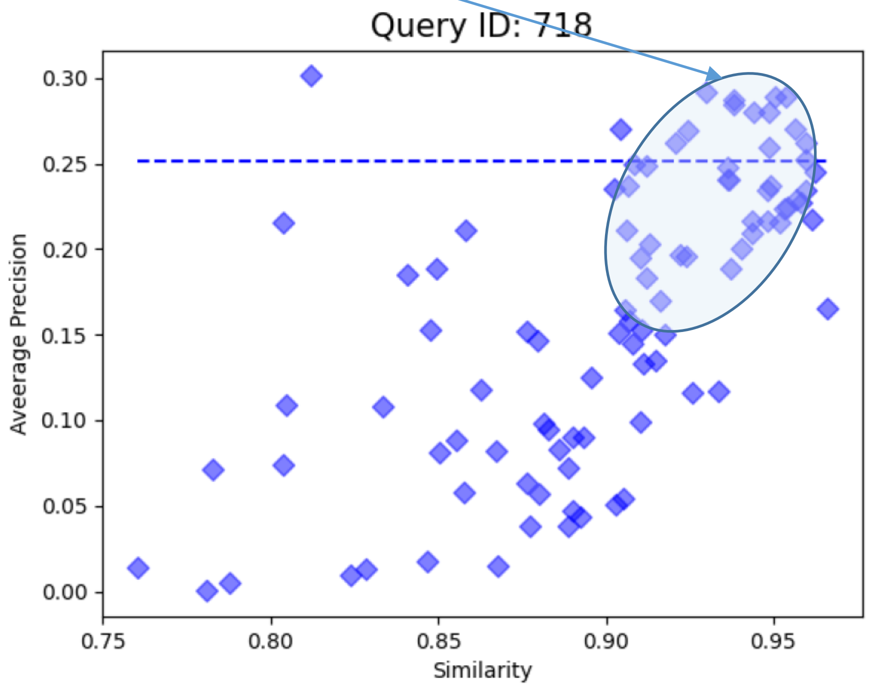
After eliminating the duplicates,

- a minimum of 38 sentences
- a maximum of 100 sentences
- an average of 85.2 sentences

Analysis of the Appropriateness of Generated Sentences

The highly accurate generated sentences exhibited a relatively high level of similarity with the original query sentence!!

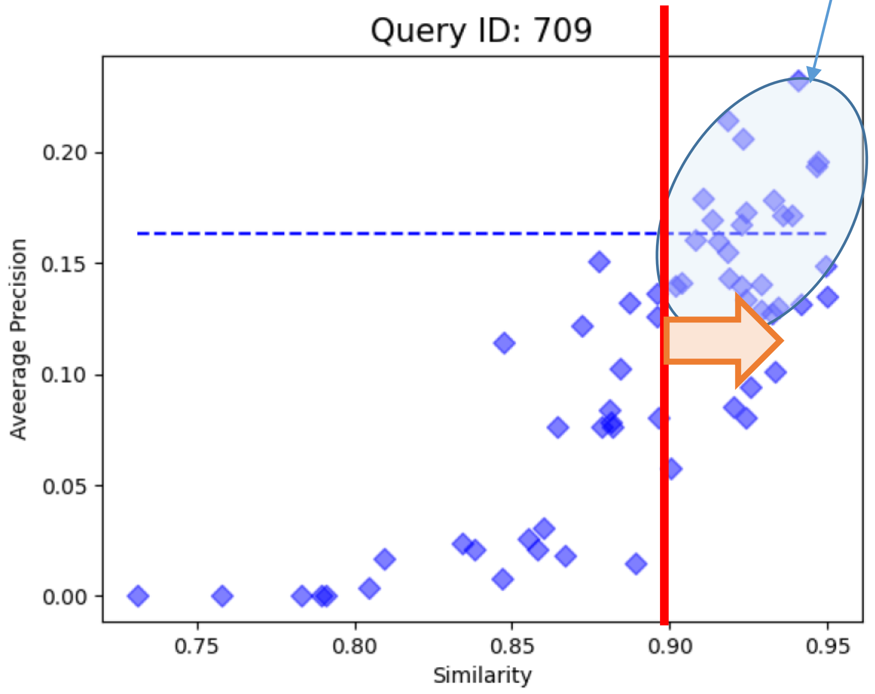


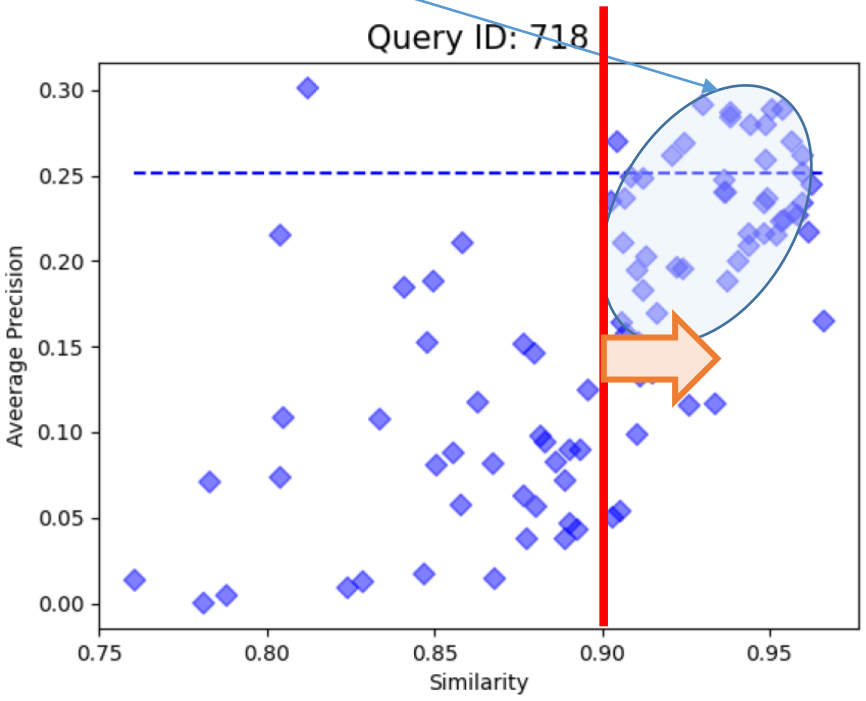


Relationship between “similarity between original and generated sentences” and “retrieval accuracy”

Analysis of the Appropriateness of Generated Sentences

The highly accurate generated sentences exhibited a relatively high level of similarity with the original query sentence!!





Relationship between “similarity between original and generated sentences” and “retrieval accuracy”

Retieival results on tv22 queries

CLIP (official)

Model name	mAP	
	Original query	+ Generated query
RN101	0.1112	**0.1218**
ViT-B/32	0.1060	**0.1123**
ViT-L/14@336px	0.0975	**0.1042**

After conducting evaluations on randomly selected pre-trained models provided by CLIP and OpenCLIP, a slight improvement in the accuracy of video search was observed.

OpenCLIP

Model name	Pre-trained	mAP	
		Original query	+ Generated query
ViT-L-14	datacomp_xl_s13b_b90k	0.1331	**0.1392**
convnext_xxlarge	Laion2b_s34b_k82k_augreg_rewind	0.1563	**0.1618**

4. Submission results

- Results for main task
- Results for progress task

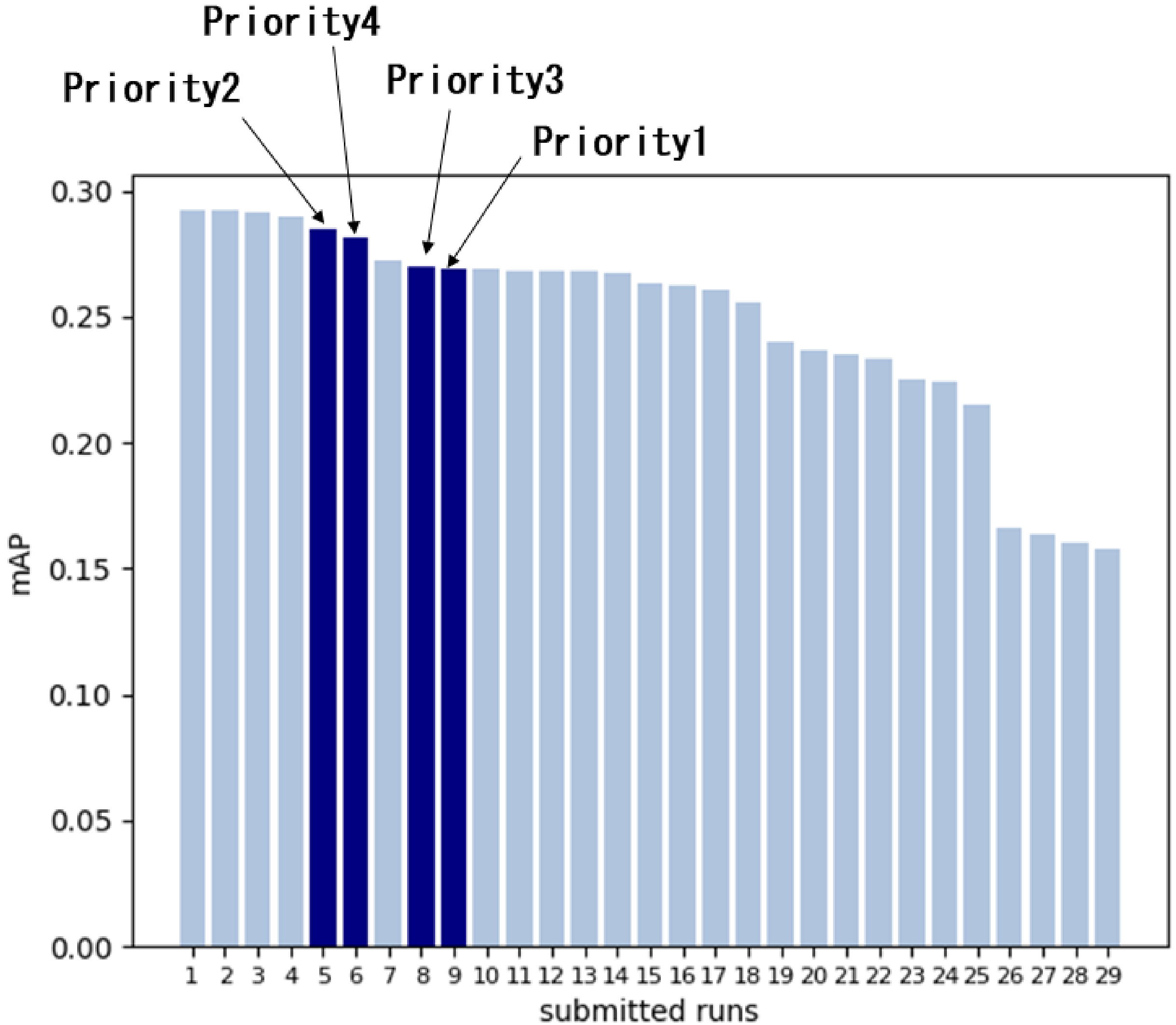
Submissions and results

Run priority	Fusion weight	Query expansion (ChatGPT)	mean average precision	
			Main task	Progress task
1	**Soft**	✓	0.269	0.272
2	**Hard**	✓	**0.285**	**0.286**
3	**Soft**		0.270	0.269
4	**Hard**		0.281	0.283

- This year, we created four different automatic systems and submitted the results.
- The distinctions among these systems lie in the approach used for integrating models, whether by determining hard weights or setting soft weights, as well as the inclusion or exclusion of query expansion using ChatGPT.

Main task (fully-automatic runs)

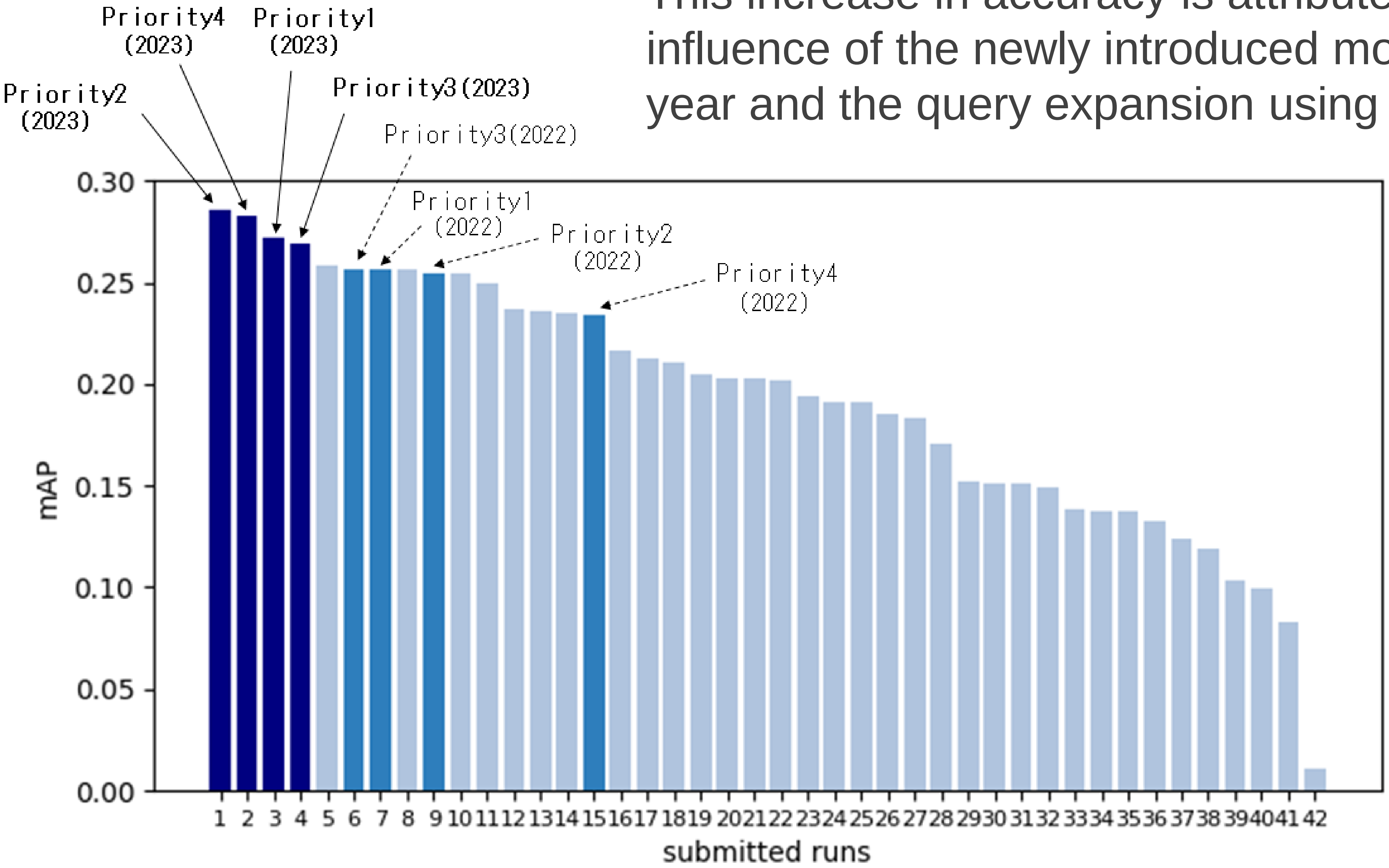
The best-performing system achieved an mAP of 0.285 in the main task, securing the second position among the participating teams.



F	D	C_D_WHU_NERCMS.23	2	0.292
F	D	C_D_WHU_NERCMS.23	1	0.292
F	D	C_D_WHU_NERCMS.23	3	0.291
F	D	C_D_WHU_NERCMS.23	4	0.290
F	D	C_D_WasedaMeiseiSoftbank.23	2	0.285
F	D	C_D_WasedaMeiseiSoftbank.23	4	0.281
F	D	C_D_RUCMM.23	1	0.272
F	D	C_D_WasedaMeiseiSoftbank.23	3	0.270
F	D	C_D_WasedaMeiseiSoftbank.23	1	0.269
F	D	C_D_RUC_AIM3.23	1	0.269
F	D	C_D_VIREO.23	4	0.268
F	D	C_D_RUCMM.23	3	0.268
F	D	C_D_RUCMM.23	2	0.268
F	D	C_D_RUC_AIM3.23	2	0.267
F	D	C_D_RUC_AIM3.23	3	0.263
F	D	C_D_RUC_AIM3.23	4	0.262
F	D	C_D_RUCMM.23	4	0.261
F	D	C_D_VIREO.23	3	0.256
F	D	C_D_ITI_CERTH.23	3	0.240
F	D	C_D_VIREO.23	1	0.237
F	D	C_D_VIREO.23	5	0.235
F	D	C_D_ITI_CERTH.23	4	0.233
F	D	C_D_ITI_CERTH.23	1	0.225
F	D	C_D_ITI_CERTH.23	2	0.224
F	D	C_D_VIREO.23	2	0.215
F	D	C_D_NII_UIT.23	1	0.166
F	D	C_D_NII_UIT.23	3	0.164
F	D	C_D_NII_UIT.23	2	0.160
F	D	C_D_NII_UIT.23	4	0.158

Progress task (fully-automatic runs)

This increase in accuracy is attributed to the influence of the newly introduced models this year and the query expansion using ChatGPT.



5. Summary

- What we have done this year
- Future work

Summary

- In the systems submitted this year, we made efforts to enhance the accuracy of image retrieval by incorporating a multitude of pre-trained models provided by OpenCLIP.

- This year's system update involved not only incorporating newly available high-performance pre-trained models but also experimenting with query expansion using ChatGPT.

Future work

- In the future, we plan to conduct further analysis to understand under which conditions accuracy improves and to refine our approach accordingly.