Waseda_Meisei_SoftBank at TRECVID 2022 Ad-hoc Video Search

Kazuya Ueki (1),(2) (presenter)
kazuya.ueki@meisei-u.ac.jp

Yuma Suzuki (3)

Hiroki Takushima (3)

Hideaki Okamoto (3)

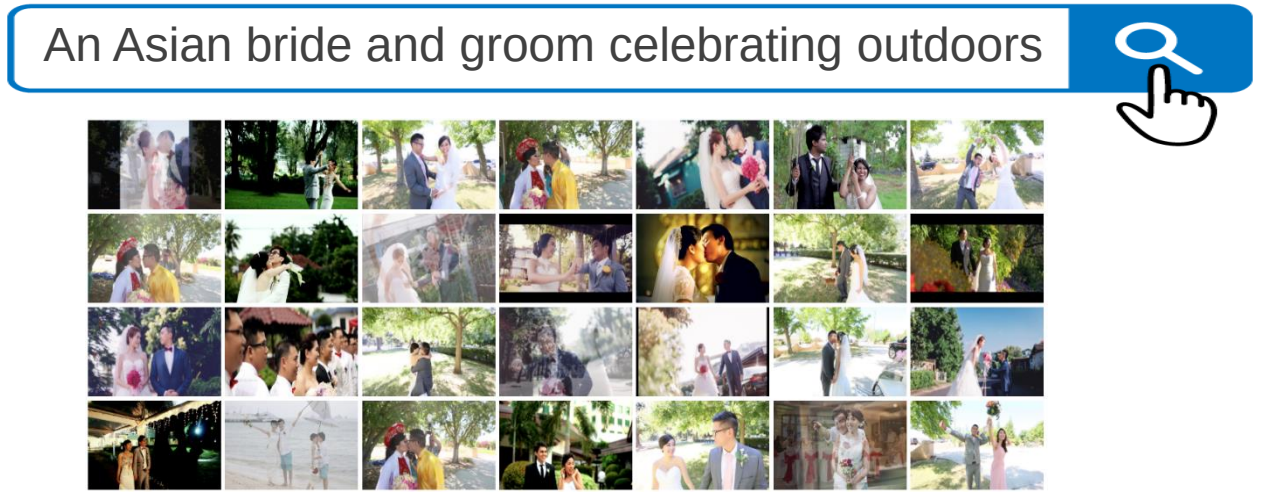
Hayato Tanoue (3)

Takayuki Hori (2),(3)

(1) Meisei University

(2) Waseda University

(3) SoftBank Corporation



TRECVID 2022 Workshop December 6th, 2022

Highlights

- **Submission type**

  ✓ Fully-automatic — Visual-semantic embedding

  ✓ ~~Manually-assisted~~ — Concept-based approach

- **This year's approach**

  Fusion of multiple visual-semantic embedding models.

  - **VSE++**
  - **GSMN**
  - **CLIP**
  - **SLIP**

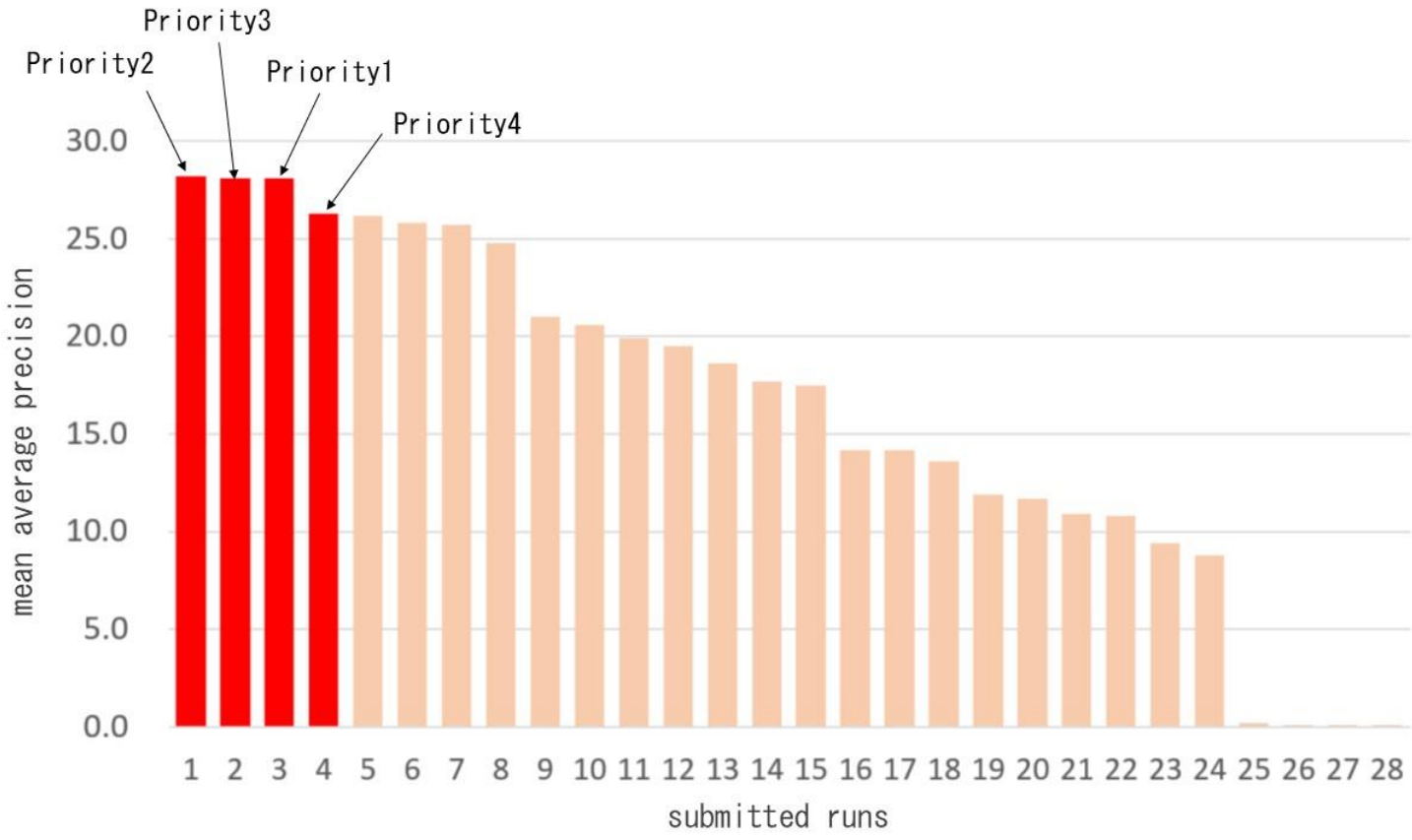
- **Results**

  Highest average precision among all submitted systems. 

Highlights

- **Submissior**
  - ✓ Fully-a
  - ✓ ~~Manua~~
- **This year's**

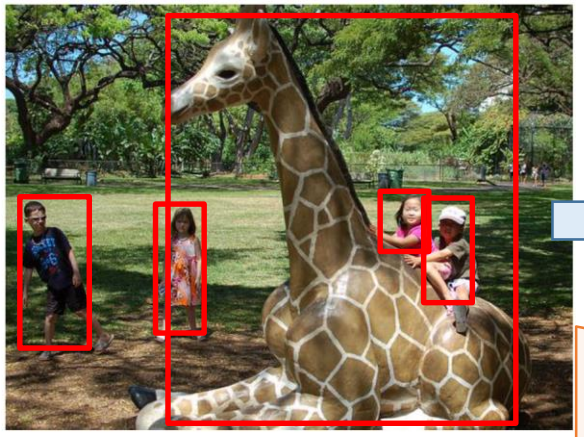
  Fusion of
- **Results**



Highest average precision among all submitted systems.

Visual-semantic embedding approach

Image



Correspondence

Text

A pair of children sit on a giraffe while other children stand nearby.

Common Space

Image Feature Vector

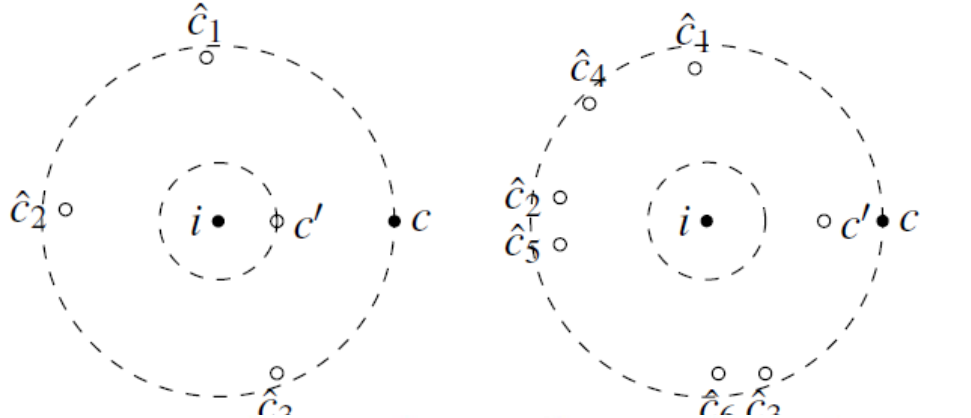
Similarity

Text Feature Vector

Embedding approaches used in our 2022 systems

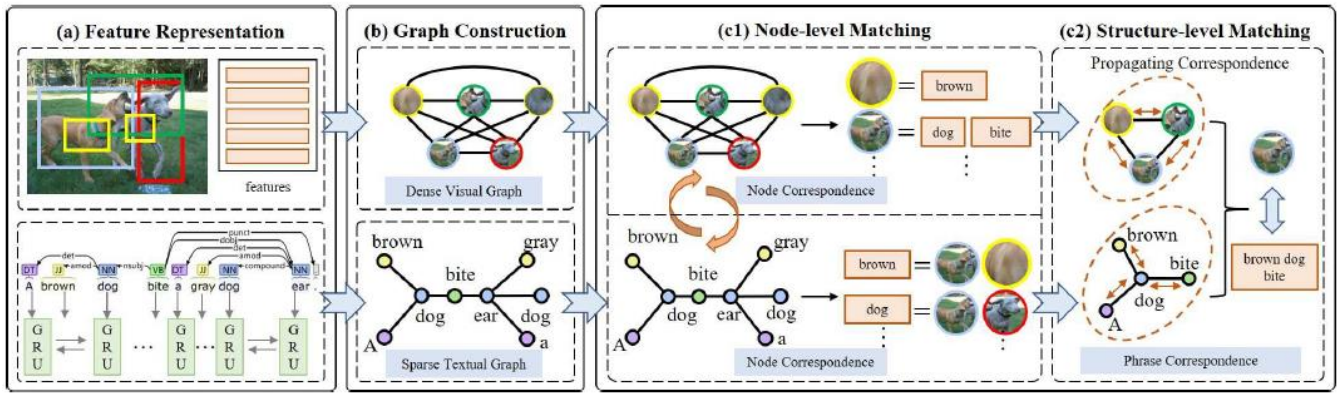
Improved retrieval accuracy by integrating four different embedding methods

VSE++ [Faghri+, 2018]

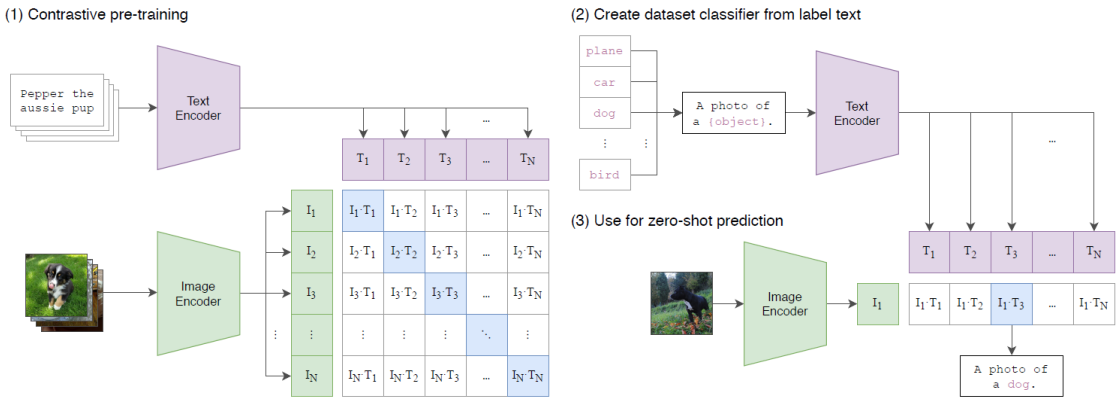


$$\ell_{MH}(i,c) = \max_{c'}\left[\alpha + s(i,c') - s(i,c)\right]_+ + \max_{i'}\left[\alpha + s(i',c) - s(i,c)\right]_+$$

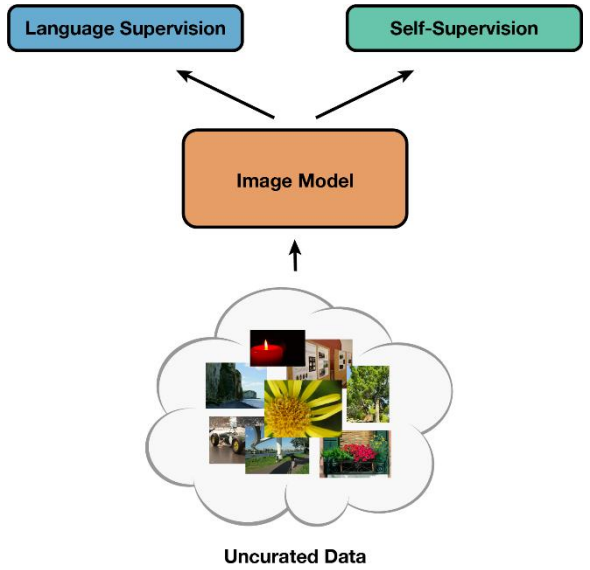
GSMN [Liu+, 2020]



CLIP [Radford+, 2021]



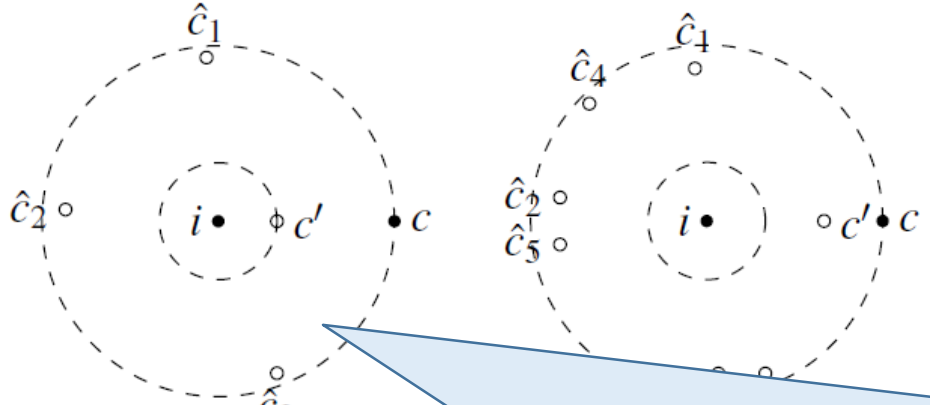
SLIP [Mu+, 2021]



Embedding approaches used in our 2022 systems

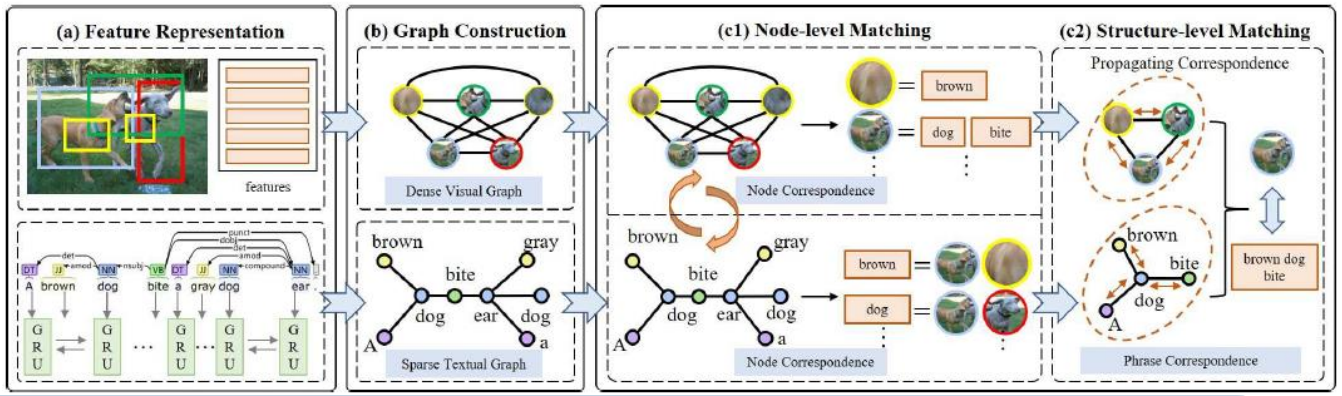
Improved retrieval accuracy by integrating four different embedding methods

VSE++ [Faghri+, 2018]



$\ell_{MH}(i,$

GSMN [Liu+, 2020]



CL

- Datasets for training: Flickr8k, Flickr30k, MS-COCO, Conceptual Captions
- # image captions: 3,428,009
- 500,000 training data and 50,000 validation data were randomly selected to train models.
- Repeated this data-selection process 32 times for each of the three types of ResNet model (ResNet-50, ResNet-101, and ResNet-152, respectively), and totally trained 96 embedding models.

Uncurated Data

Embedding approaches used in our 2022 systems

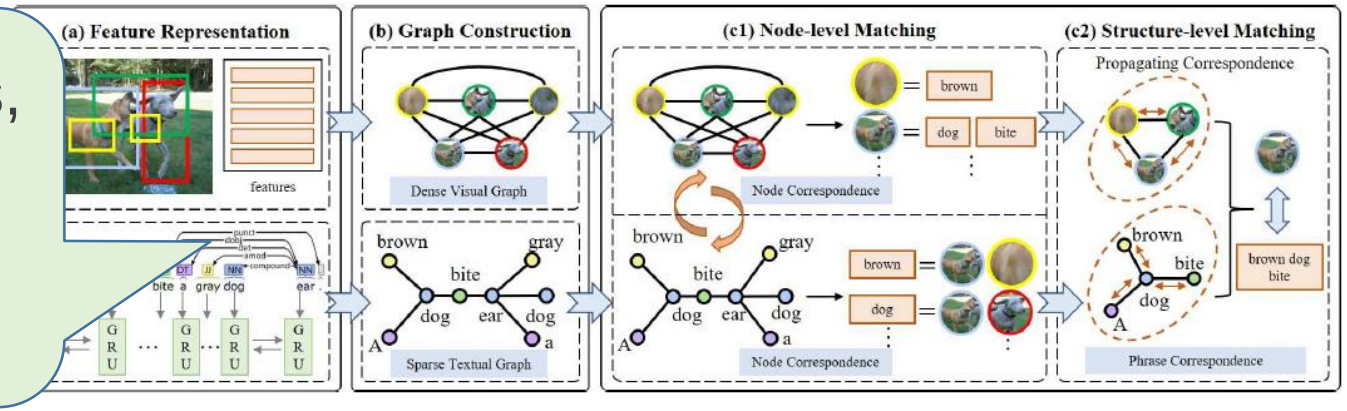
Improved retrieval accuracy by integrating four different embedding methods

VSE++ [Faghri+, 2018]

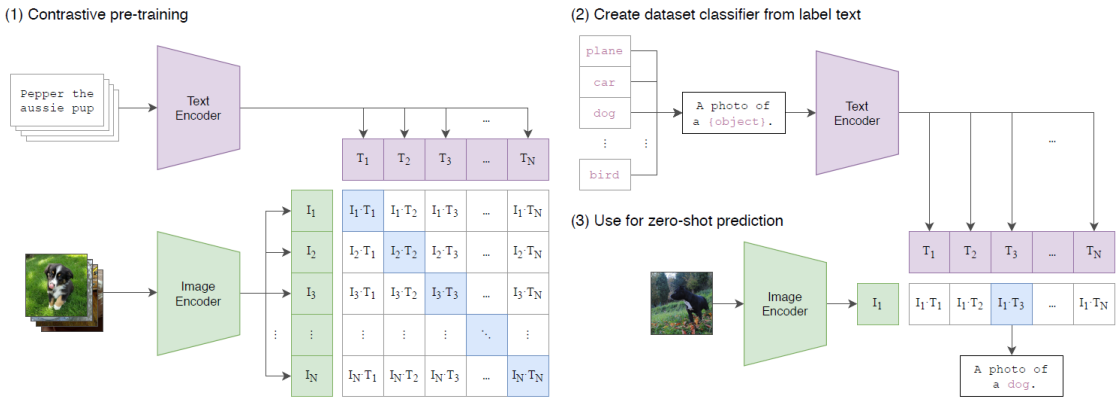
- GSMN models objects, relationships, and attributes as structured phrases through node- and structure-level correspondences.

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

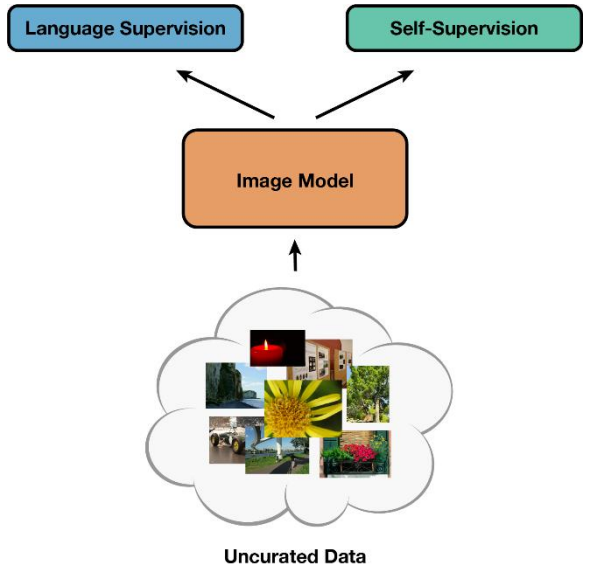
GSMN [Liu+, 2020]



CLIP [Radford+, 2021]



SLIP [Mu+, 2021]



Embedding approaches used in our 2022 systems

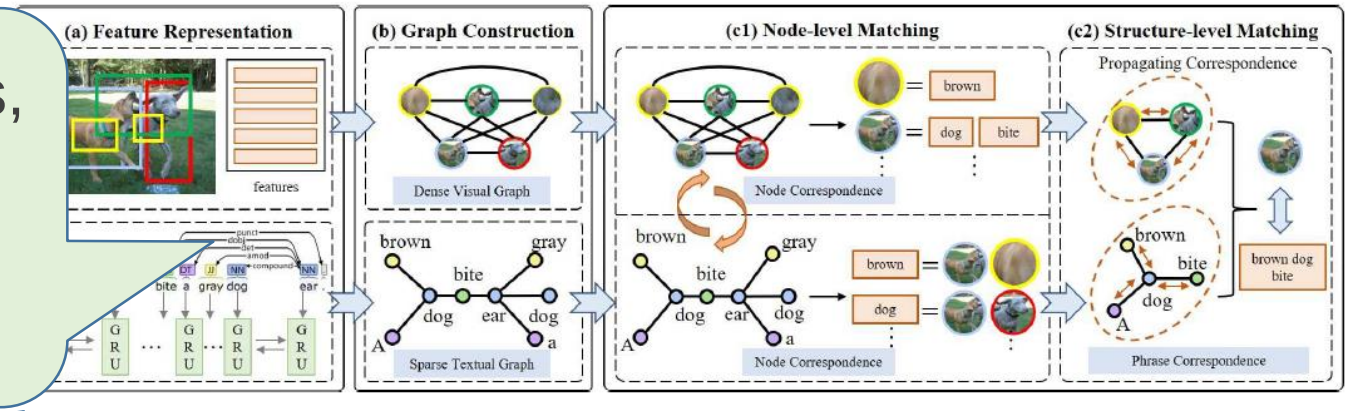
Improved retrieval accuracy by integrating four different embedding methods

VSE++ [Faghri+, 2018]

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

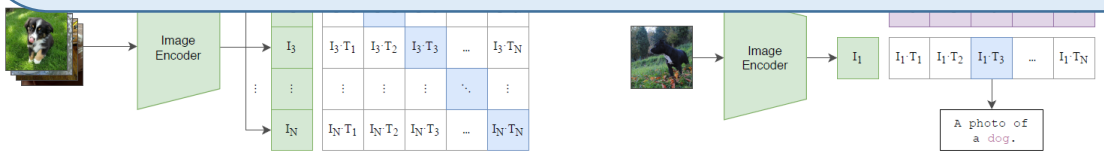
- GSMN models objects, relationships, and attributes as structured phrases through node- and structure-level correspondences.

GSMN [Liu+, 2020]



CL

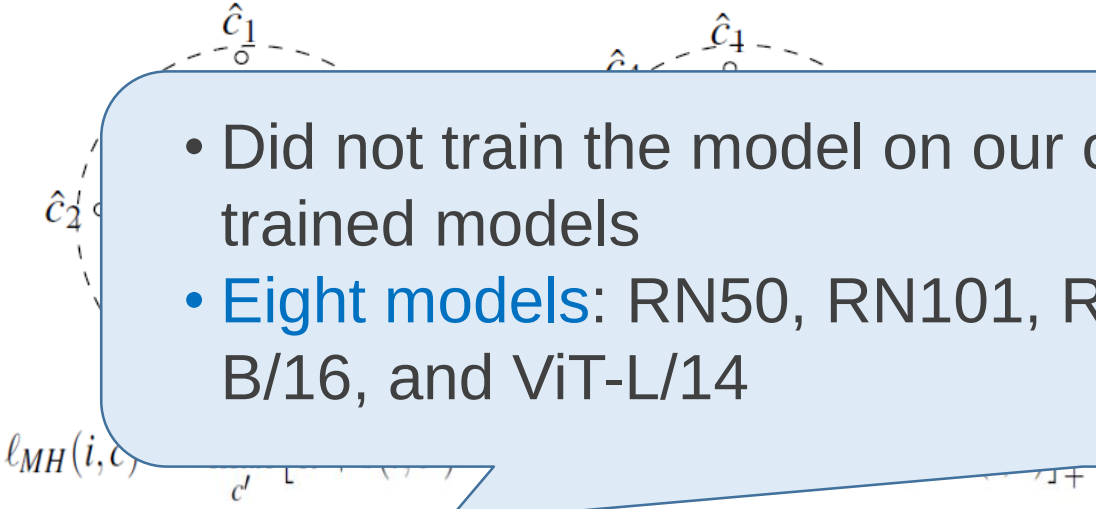
- Datasets for training: Flickr8k, Flickr30k, MS-COCO, Conceptual Captions, MSR-VTT
- # image captions: 3,755,503
- We divided the training data and created nine models.



Embedding approaches used in our 2022 systems

Improved retrieval accuracy by integrating four different embedding methods

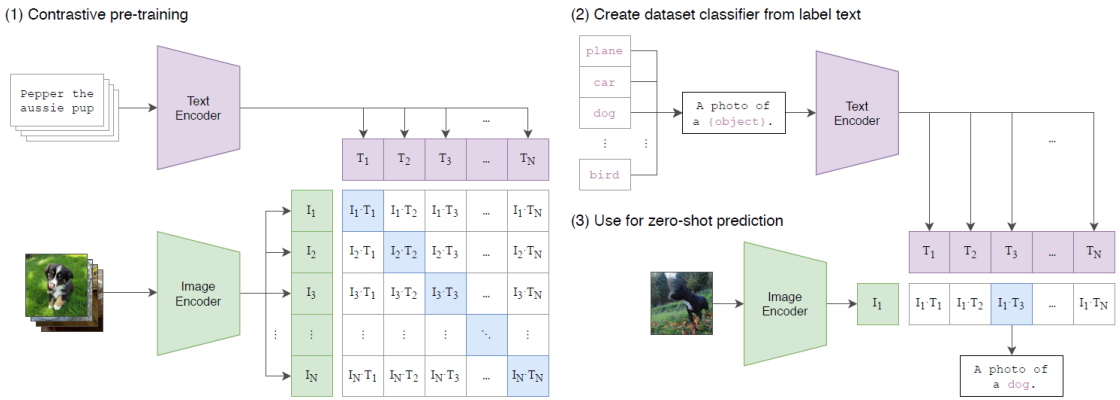
VSE++ [Faghri+, 2018]



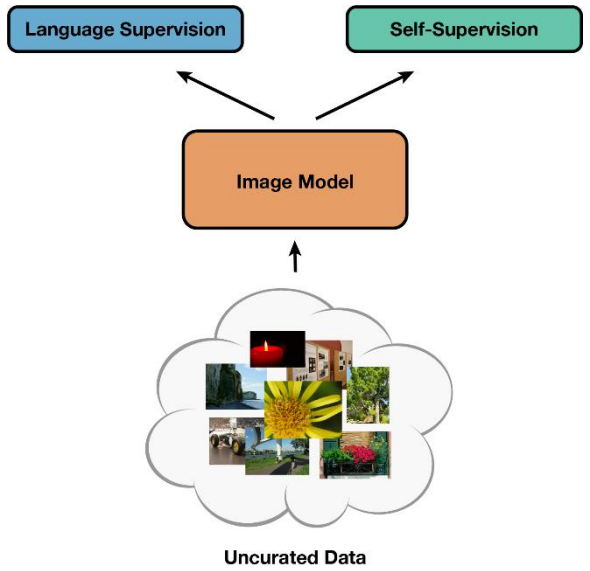
GSMN [Liu+, 2020]

- Did not train the model on our own, but instead used publicly available pre-trained models
- Eight models: RN50, RN101, RN50x4, RN50x16, RN50x64, ViT-B/32, ViT-B/16, and ViT-L/14

CLIP [Radford+, 2021]



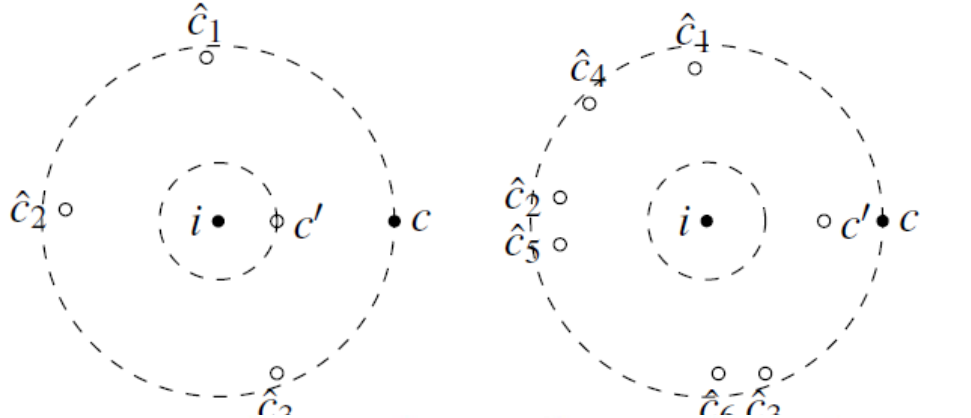
SLIP [Mu+, 2021]



Embedding approaches used in our 2022 systems

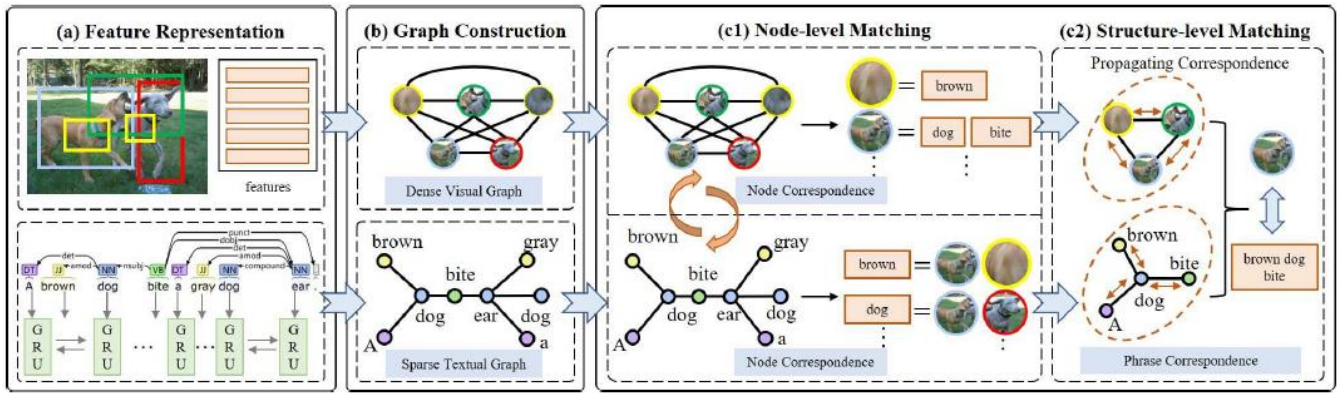
Improved retrieval accuracy by integrating four different embedding methods

VSE++ [Faghri+, 2018]



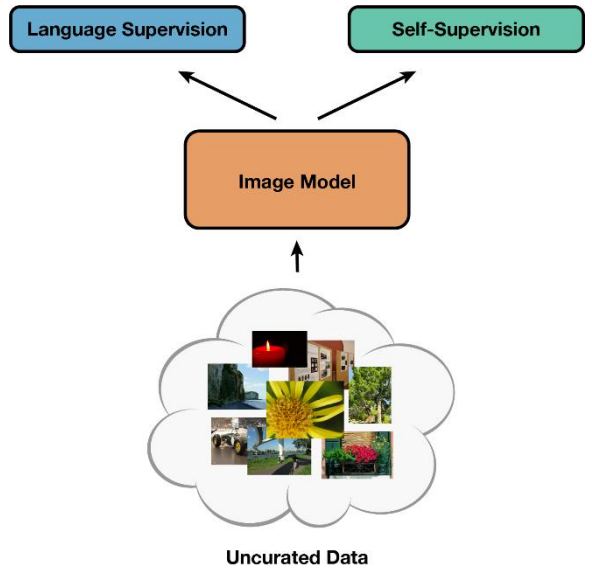
$$\ell_{MH}(i,c) = \max_{c'}\left[\alpha + s(i,c') - s(i,c)\right]_+ + \max_{i'}\left[\alpha + s(i',c) - s(i,c)\right]_+$$

GSMN [Liu+, 2020]



CLIP [Radford+, 2021]

SLIP [Mu+, 2021]



- Multi-task learning framework for combining self-supervised learning and CLIP pre-training.
- Across ImageNet and additional datasets, SLIP improves accuracy by a large margin.

Embedding approaches used in our 2022 systems

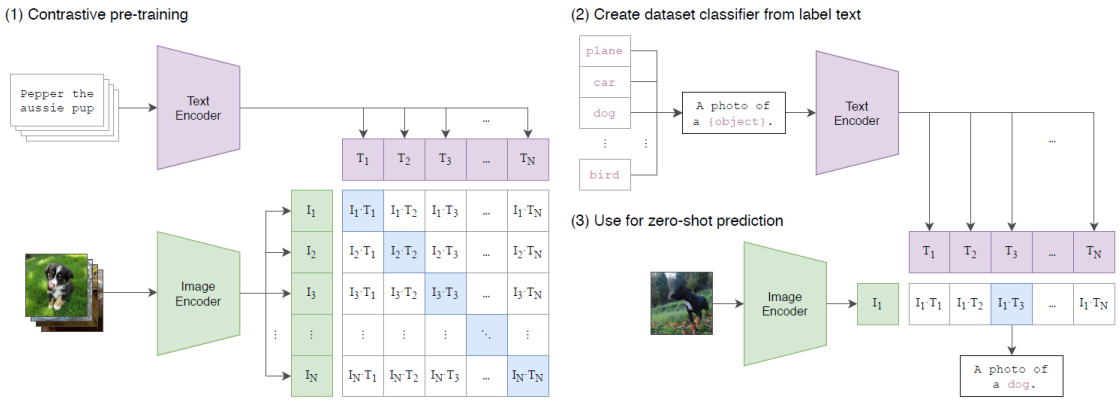
Improved retrieval accuracy by integrating four different embedding methods

VSE++ [Faghri+, 2018]

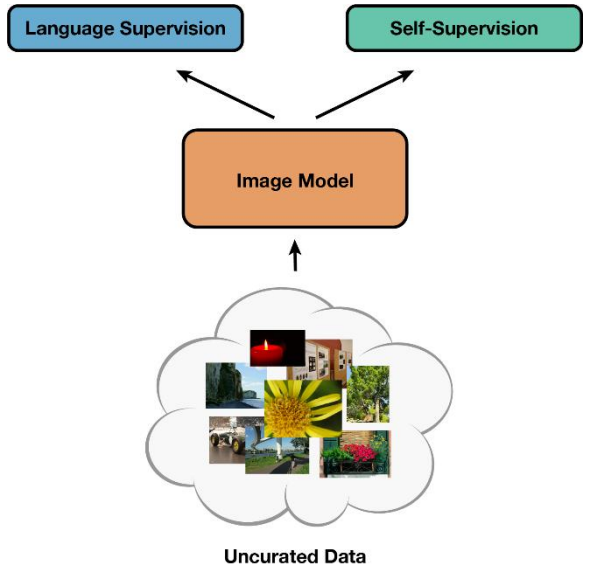
GSMN [Liu+, 2020]

- Did not train the model on our own as well, but instead used publicly available pre-trained models
- Four models: ViT-Small(yfcc15m), ViT-Base(yfcc15m), ViT-Base(cc3m), and ViT-Base(cc15m)
- Would like to use ViT-Large, but could not complete the calculations…

CLIP [Radford+, 2021]



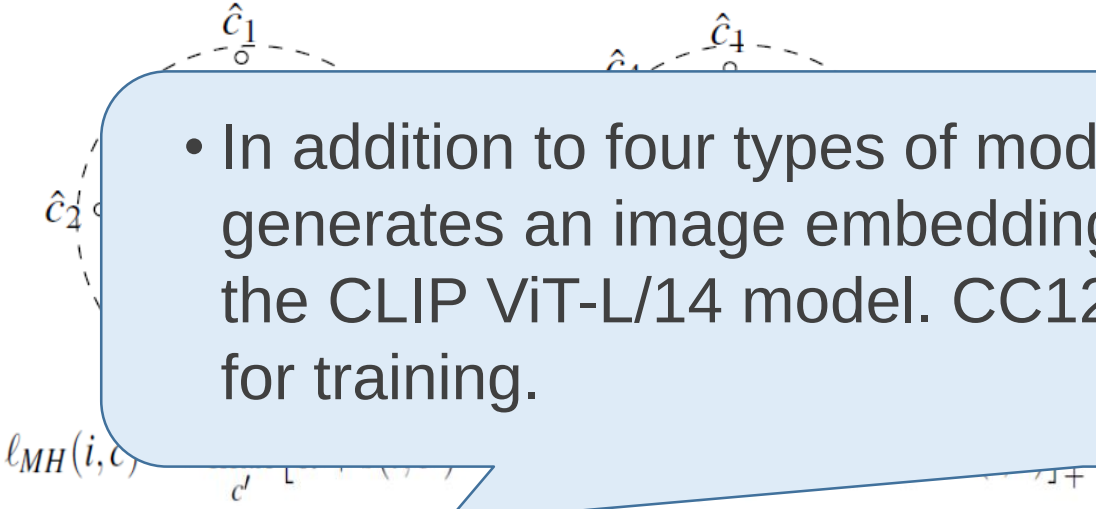
SLIP [Mu+, 2021]



Embedding approaches used in our 2022 systems

Improved retrieval accuracy by integrating four different embedding methods

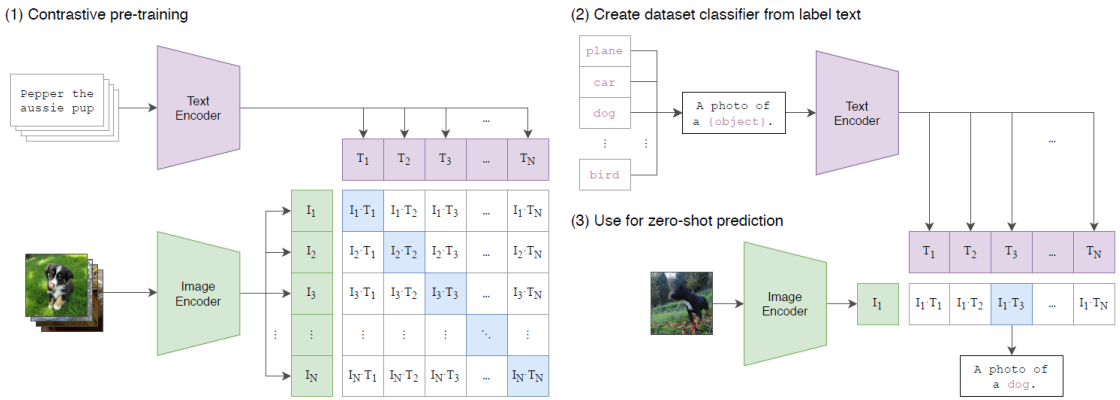
VSE++ [Faghri+, 2018]



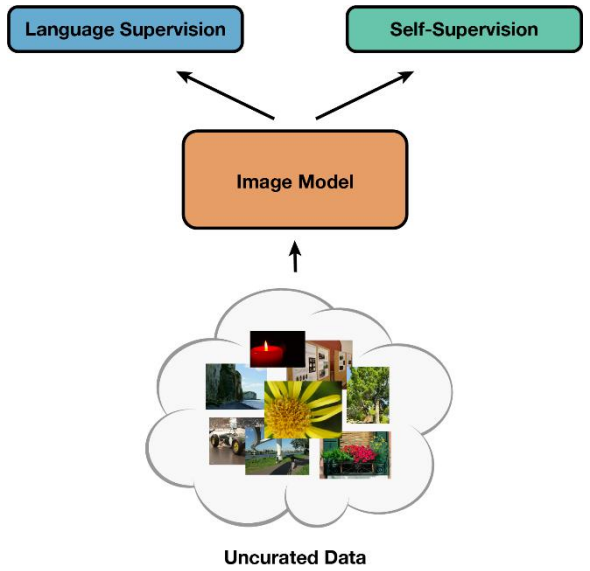
GSMN [Liu+, 2020]

- In addition to four types of model, we also trained a diffusion model that generates an image embedding of CLIP conditioned on a text embedding of the CLIP ViT-L/14 model. CC12M and a portion of LAION400M were used for training.

CLIP [Radford+, 2021]



SLIP [Mu+, 2021]



Score calculations

- Using the trained models, we calculated the scores for V3C2 based on whether it matched the query sentence.
- We used the cosine similarity between the frame images extracted from the videos and the query sentence to search for videos that match the query sentence.
- Images were extracted from the video every 10 frames, the similarity of the images to the query sentence was calculated, and the maximum value was used as the score for that video.
- For the diffusion model, we generated 1,000 image embeddings for one query and calculated the similarity between the generated images and the video frames.

Score calculations

- After calculating all scores for the test dataset, to obtain the final score, a min-max normalization was conducted for each model,

  ⇨ The maximum score: 1.0, and the minimum score: 0.0.

- For each embedding method, all scores from multiple models were added and normalized again using the min-max normalization.

- The final search result was determined using the score computed using the weighted sum of each embedding method.

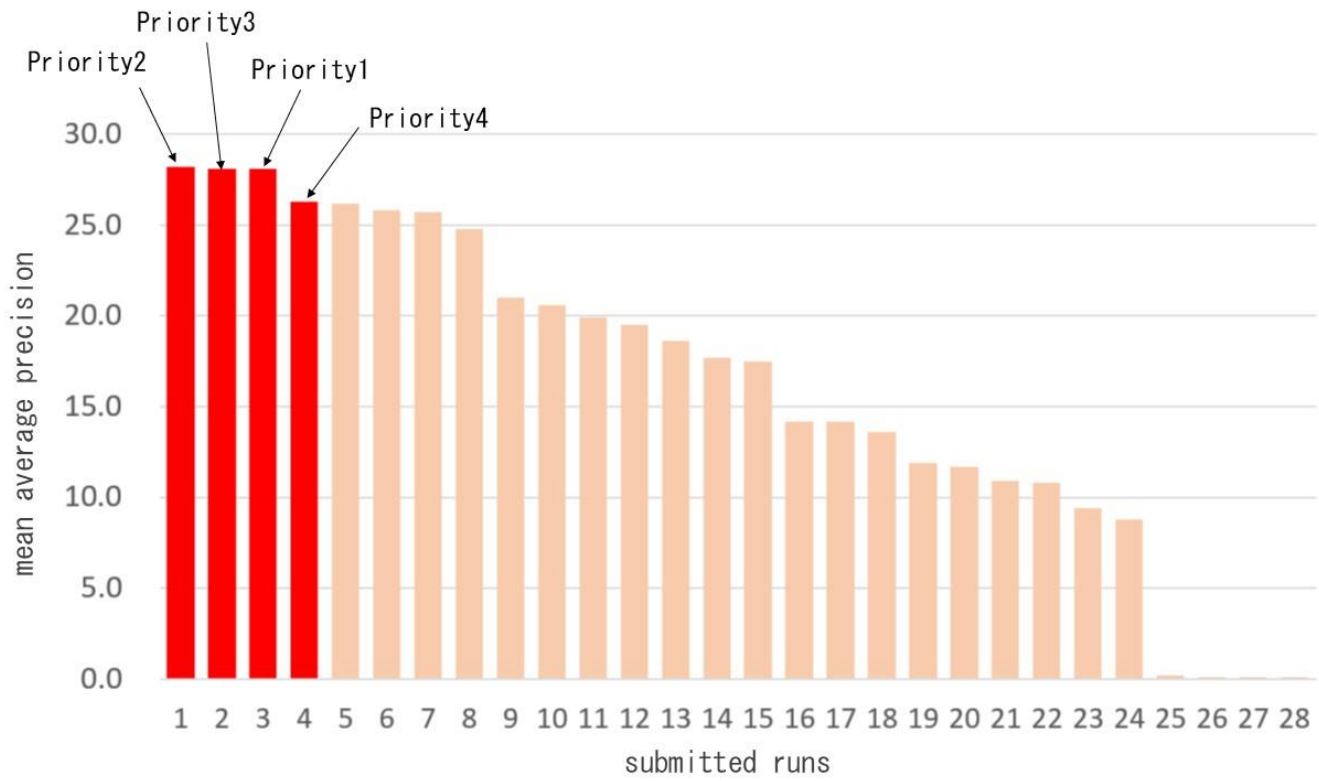
  ⇨ The fusion weights were manually determined by evaluating the AVS tasks of 2019, 2020, and 2021 TRECVID.

Submissions and results

Run priority	Fusion weights					mAP
	VSE++	GSMN	CLIP	SLIP	Diffusion	
1	3	3	15	3	3	**28.1**
2	3	3	10	3	3	**28.2**
3	3	3	15	5	3	**28.1**
4	3	3	15	3	0	**26.3**

Submissions and results

Run priority	Fusion weights					mAP
	VSE++	GSMN	CLIP	SLIP	Diffusion	
1						28.1
2						28.2
3						28.1
4						26.3



Highest average precision among all submitted systems.

Submissions and results

Run priority	Fusion weights					mAP
	VSE++	GSMN	CLIP	SLIP	Diffusion	
1	3	3	15	3	3	**28.1**
2	3	3	10	3	3	**28.2**
3	3	3	15	5	3	**28.1**
4	3	3	15	3	0	**26.3**

As the reason for the highest fusion weights for the CLIP models, they had the highest precision and largest contribution over VSE++ and GSMN on the benchmark from last year.

Submissions and results

Run priority	Fusion weights					mAP
	VSE++	GSMN	CLIP	SLIP	Diffusion	
1	3	3	15	3	3	**28.1**
2	3	3	10	3	3	**28.2**
3	3	3	15	5	3	**28.1**
4	3	3	15	3	0	**26.3**

SLIP might be better than CLIP; however, we set the integration weights lower than CLIP because we did not finish all feature extraction calculations and were unable to evaluate it sufficiently.

Submissions and results

Run priority	Fusion weights					mAP
	VSE++	GSMN	CLIP	SLIP	Diffusion	
1	3	3	15	3	3	**28.1**
2	3	3	10	3	3	**28.2**
3	3	3	15	5	3	**28.1**
4	3	3	15	3	0	**26.3**

In addition, the diffusion model introduced this year was given a lower fusion weight, partly because we had not yet obtained sufficient validation results, and only some of the models could be trained.

Submissions and results

Run priority	Fusion weights					mAP
	VSE++	GSMN	CLIP	SLIP	Diffusion	
1	3	3	15	3	3	**28.1**
2	3	3	10	3	3	**28.2**
3	3	3	15	5	3	**28.1**
4	3	3	15	3	0	**26.3**

- However, the results of this year's benchmark show that priorities 1, 2, and 3 had a higher mean average precision and contributed more than priority 4.
- Because sufficient validation experiments could not be conducted, a detailed analysis and validation will be conducted in the future to confirm the effectiveness of this approach.

Retrieved videos (Good results)

mAP = 70.6 703 A construction site

mAP = 69.8 723 A person is biking through a path in a forest

mAP = 68.3 709 A person is in the act of swinging

mAP = 50.6 723 Building with columns during daytime

Retrieved videos (Bad results) ☹

mAP = 4.2 726 Two teams playing a game where one team have their players wearing white t-shirts

mAP = 4.4 713 A kneeling man outdoors

mAP = 5.1 710 A person wearing a light t-shirt with dark or black writing on it

mAP = 6.1 702 A room with blue wall

Summary

- In the systems submitted this year, we use four types of embedding approaches; VSE++, GSMN, CLIP, and SLIP to improve the retrieval performance.

- The complementarity of the results from each model allowed for the best accuracy among the submitted systems.

Future work

- Since we were not able to conduct sufficient validation experiments on the use of the diffusion model, we plan to conduct detailed analysis and validation in the future to confirm its effectiveness.