Waseda_Meisei_SoftBank at TRECVID 2021 Ad-hoc Video Search

Kazuya Ueki (presenter) kazuya.ueki@meisei-u.ac.jp
Meisei University, Waseda University

Takayuki Hori
SoftBank Corporation, Waseda University

Yongbeom Kim, Yuma Suzuki
SoftBank Corporation

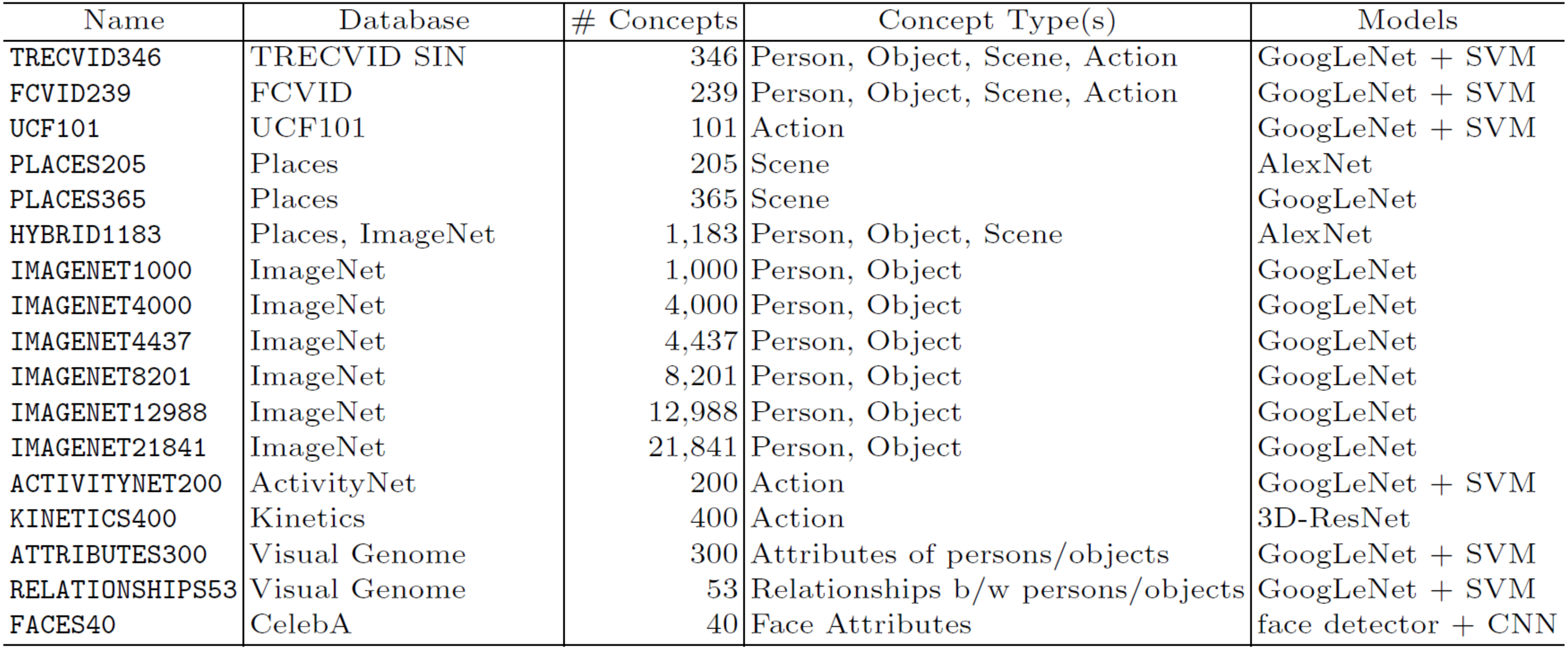
TRECVID 2021 Workshop December 7th, 2021

WASEDA University MEISEI UNIVERSITY SoftBank

Our approach for ad-hoc video search

	Submitted runs	
	Manually assisted	Fully automatic
1. Concept-based	✓	
2. Visual-semantic embedding	✓	✓
rank among all participants	2nd	4th

Concept bank used in our systems in 2020 and 2021



Name	Database	# Concepts	Concept Type(s)	Models
TRECVID346	TRECVID SIN	346	Person, Object, Scene, Action	GoogLeNet + SVM
FCVID239	FCVID	239	Person, Object, Scene, Action	GoogLeNet + SVM
UCF101	UCF101	101	Action	GoogLeNet + SVM
PLACES205	Places	205	Scene	AlexNet
PLACES365	Places	365	Scene	GoogLeNet
HYBRID1183	Places, ImageNet	1,183	Person, Object, Scene	AlexNet
IMAGENET1000	ImageNet	1,000	Person, Object	GoogLeNet
IMAGENET4000	ImageNet	4,000	Person, Object	GoogLeNet
IMAGENET4437	ImageNet	4,437	Person, Object	GoogLeNet
IMAGENET8201	ImageNet	8,201	Person, Object	GoogLeNet
IMAGENET12988	ImageNet	12,988	Person, Object	GoogLeNet
IMAGENET21841	ImageNet	21,841	Person, Object	GoogLeNet
ACTIVITYNET200	ActivityNet	200	Action	GoogLeNet + SVM
KINETICS400	Kinetics	400	Action	3D-ResNet
ATTRIBUTES300	Visual Genome	300	Attributes of persons/objects	GoogLeNet + SVM
RELATIONSHIPS53	Visual Genome	53	Relationships b/w persons/objects	GoogLeNet + SVM
FACES40	CelebA	40	Face Attributes	face detector + CNN

Prepared in advance a large concept classifiers of more than 50,000
to increase the coverage of words in the query sentences.

Video retrieval pipeline of concept-based approach

1. Extract one or more keywords from a query sentence. (manually or automatically)

   ex.) an adult person wearing a backpack and walking on a sidewalk

   "adult" "person" "wearing" "backpack" "walking" "sidewalk"

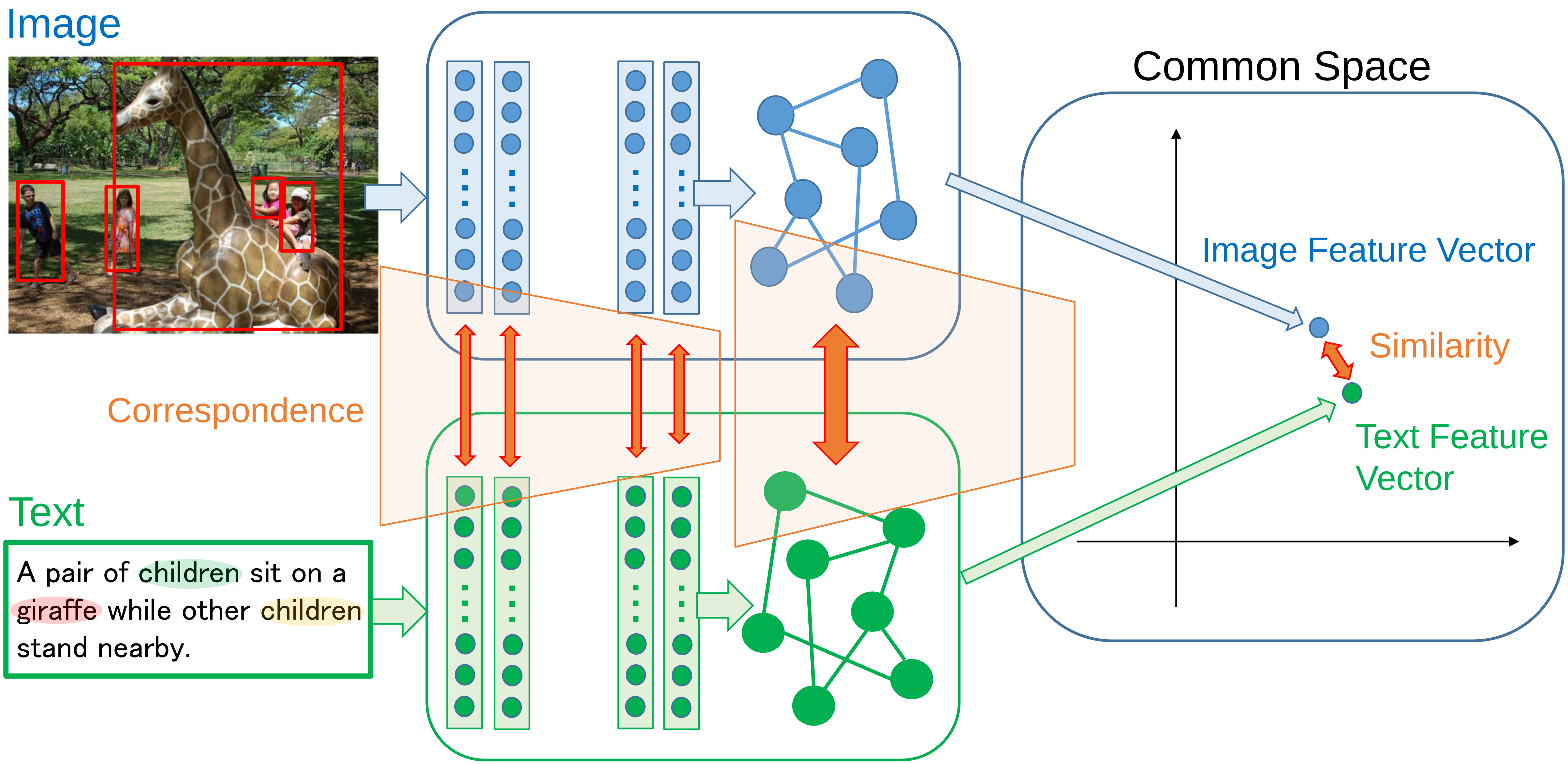
2. Select one or more concept classifiers related to a keyword.
   The corresponding concept may not exist in the concept bank.

   ⇨ Word2vec to obtain more concepts

3. For each video, a score is calculated for the query sentence by integrating the scores from multiple concept classifiers.

   score of "adult" X score of "person" X score of "wearing" X score of "backpack" X score of "walking" X score of "sidewalk"

Visual-semantic embedding approach



Image

Text

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

Correspondence

Common Space

Image Feature Vector

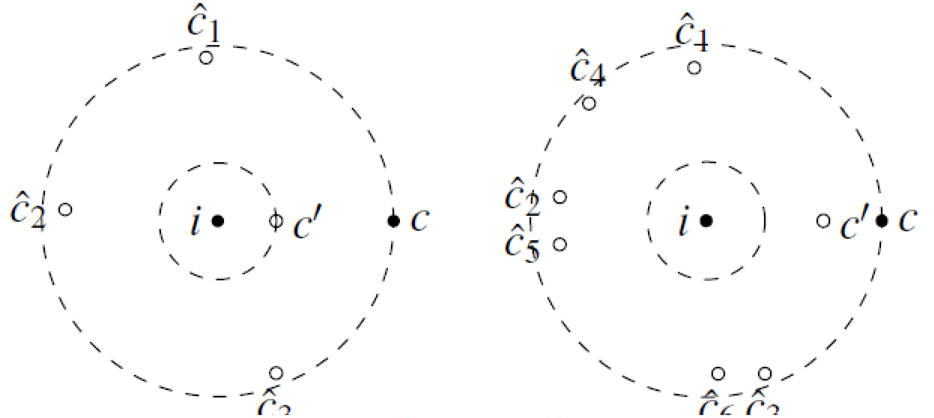
Similarity

Text Feature Vector

Embedding approaches used in our 2021 systems

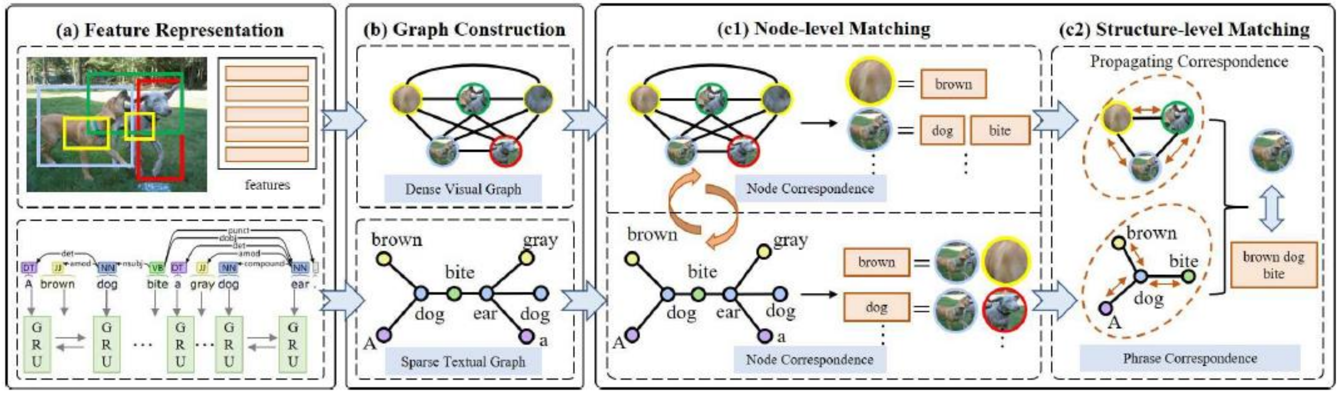
Improved retrieval accuracy by integrating four different embedding methods

VSE++ [Faghri+, 2018]

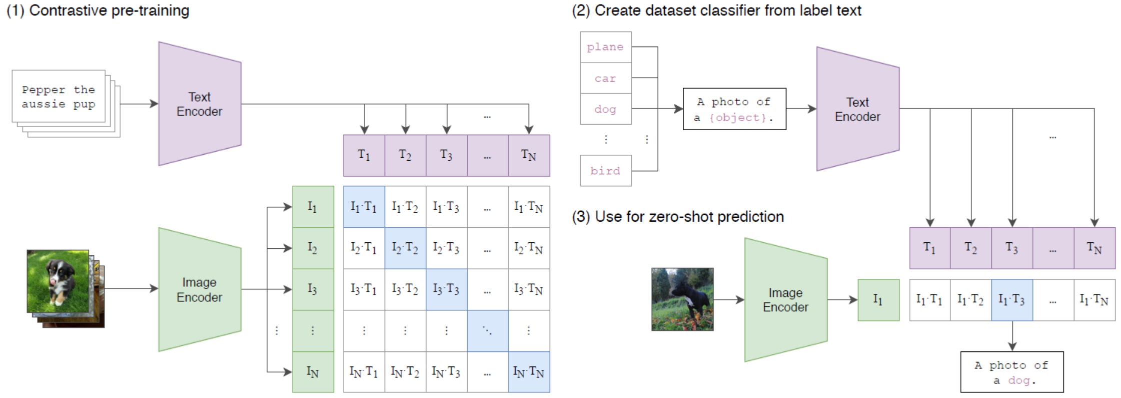


$$\ell_{MH}(i,c) = \max_{c'}\left[\alpha + s(i,c') - s(i,c)\right]_+ + \max_{i'}\left[\alpha + s(i',c) - s(i,c)\right]_+$$

GSMN [Liu+, 2020]

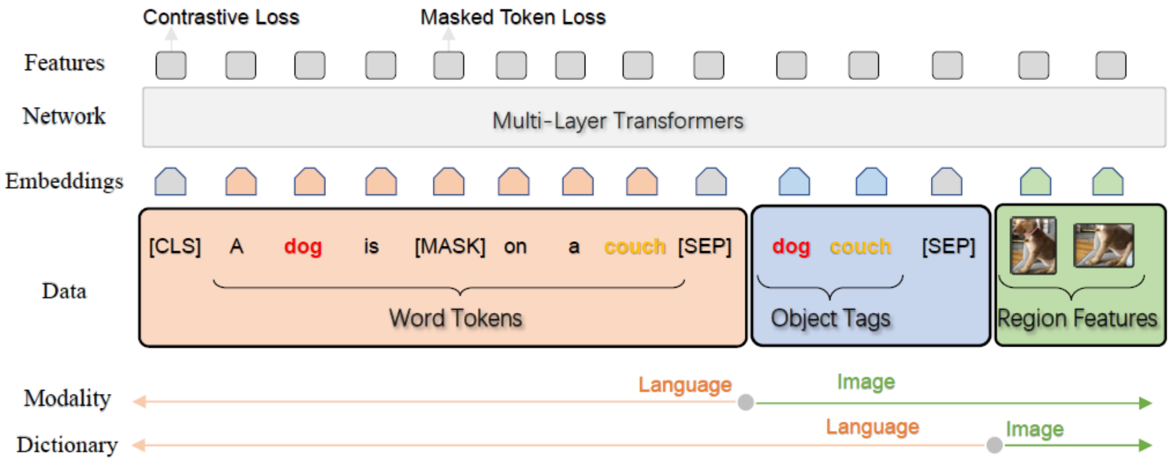


CLIP [Radford+, 2021]



(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
$I_2{\cdot}T_1$ $I_2{\cdot}T_2$ $I_2{\cdot}T_3$ ... $I_2{\cdot}T_N$
$I_3{\cdot}T_1$ $I_3{\cdot}T_2$ $I_3{\cdot}T_3$ ... $I_3{\cdot}T_N$
$I_N{\cdot}T_1$ $I_N{\cdot}T_2$ $I_N{\cdot}T_3$ ... $I_N{\cdot}T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
A photo of a dog.

Oscar [Li+, 2020]



Embedding approaches used in our 2021 systems

The following three types of video-shot frames were used in each approach, depending on when the work was done and how fast the calculations were performed:

$Frame_k$: Use only key frames

$Frame_{10}$: Use the middle 10 frames of the video divided into 11 equal parts

$Frame_{e10}$: Use every 10 frames

	# training data partitions	Model / Features	Type of test data	# score files
VSE++	32	3 (ResNet-50, 101, 152)	2 ($Frame_{10}$, $Frame_{e10}$)	192
GSMN	9	1 (bottom-up attention)	1 ($Frame_{e10}$)	9
CLIP	1	4 (ViT-B/32, RN50, RN101, RN50x4)	2 ($Frame_{10}$, $Frame_{e10}$)	8
Oscar	1	1 (large model)	1 ($Frame_k$)	1

All score files were combined to get the final results

Embedding approaches used in our 2021 systems

The following three types of video-shot frames were used in each approach, depending on when the work was done and how fast the calculations were performed:

$Frame_k$: Use only key frames
$Frame_{10}$: Use the middle 10 frames of the video divided into 11 equal parts
$Frame_{e10}$: Use every 10 frames

	# training data partitions	Model / Features	Type of test data	# score files
VSE++	32	3 (ResNet-50, 101, 152)	2 ($Frame_{10}$, $Frame_{e10}$)	192
Oscar	1	1 (large model)	1 ($Frame_k$)	1

- 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.
- Add 192 scores → min-max normalization (maximum score: 1.0, minimum score: 0.0)

All score files were combined to get the final results

Embedding approaches used in our 2021 systems

The following three types of video-shot frames were used in each approach, depending on when the work was done and how fast the calculations were performed:

$Frame_k$: Use only key frames
$Frame_{10}$: Use the middle 10 frames of the video divided into 11 equal parts
$Frame_{e10}$: Use every 10 frames

	# training data partitions	Model / Features	Type of test data	# score files
VSE++	32	3 (ResNet-50, 101, 152)	2 ($Frame_{10}$, $Frame_{e10}$)	192
GSMN	9	1 (bottom-up attention)	1 ($Frame_{e10}$)	9
		/32, RN50		

- 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.
- Add 9 scores → min-max normalization (maximum score: 1.0, minimum score: 0.0)

All score files were combined to get the final results

Embedding approaches used in our 2021 systems

The following three types of video-shot frames were used in each approach, depending on when the work was done and how fast the calculations were performed:

$Frame_k$: Use only key frames

$Frame_{10}$: Use the middle 10 frames of the video divided into 11 equal parts

$Frame_{e10}$: Use every 10 frames

	# training data partitions	Model / Features	Type of test data	# score files
VSE++	32	3 (ResNet-50, 101, 152)	2 ($Frame_{10}$, $Frame_{e10}$)	192
G				
CLIP	1	4 (ViT-B/32, RN50, RN101, RN50x4)	2 ($Frame_{10}$, $Frame_{e10}$)	8
Oscar	1	1 (large model)	1 ($Frame_k$)	1

- No training → 4 types of pre-trained models
- Add 9 scores → min-max normalization (maximum score: 1.0, minimum score: 0.0)

All score files were combined to get the final results

Embedding approaches used in our 2021 systems

The following three types of video-shot frames were used in each approach, depending on when the work was done and how fast the calculations were performed:

$Frame_k$: Use only key frames
$Frame_{10}$: Use the middle 10 frames of the video divided into 11 equal parts
$Frame_{e10}$: Use every 10 frames

	# training data partitions	Model / Features	Type of test data	# score files
VSE++	32	3 (ResNet-50, 101, 152)	2 ($Frame_{10}$, $Frame_{e10}$)	192
GSMN	[illegible]	1 (bottom-up [illegible]	1 ([illegible])	[illegible]
[illegible]	[illegible]	[illegible] 101, RN50x4)	[illegible]	[illegible]
Oscar	1	1 (large model)	1 ($Frame_k$)	1

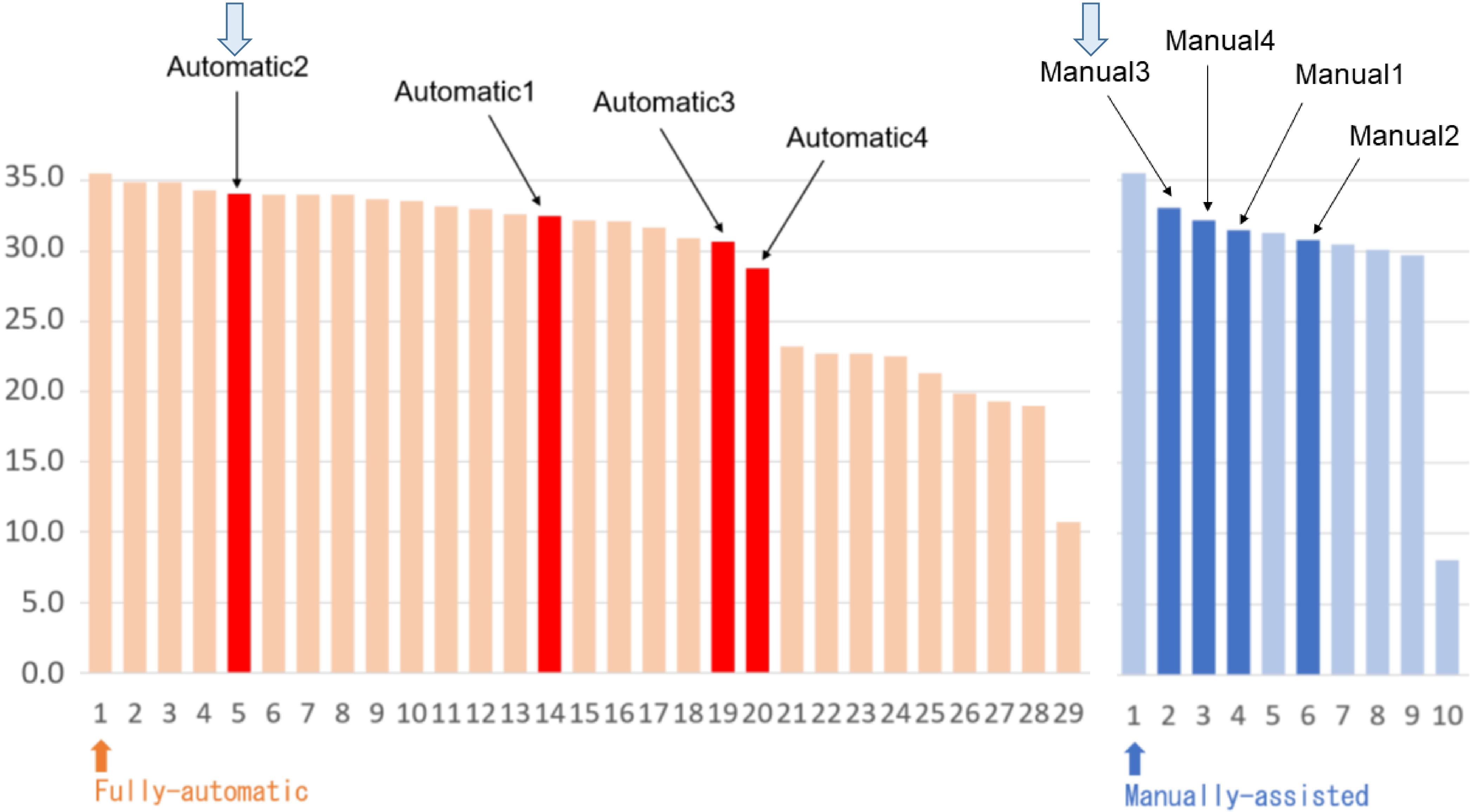
- No training → pre-trained models
- Min-max normalization (maximum score: 1.0, minimum score: 0.0)

All score files were combined to get the final results

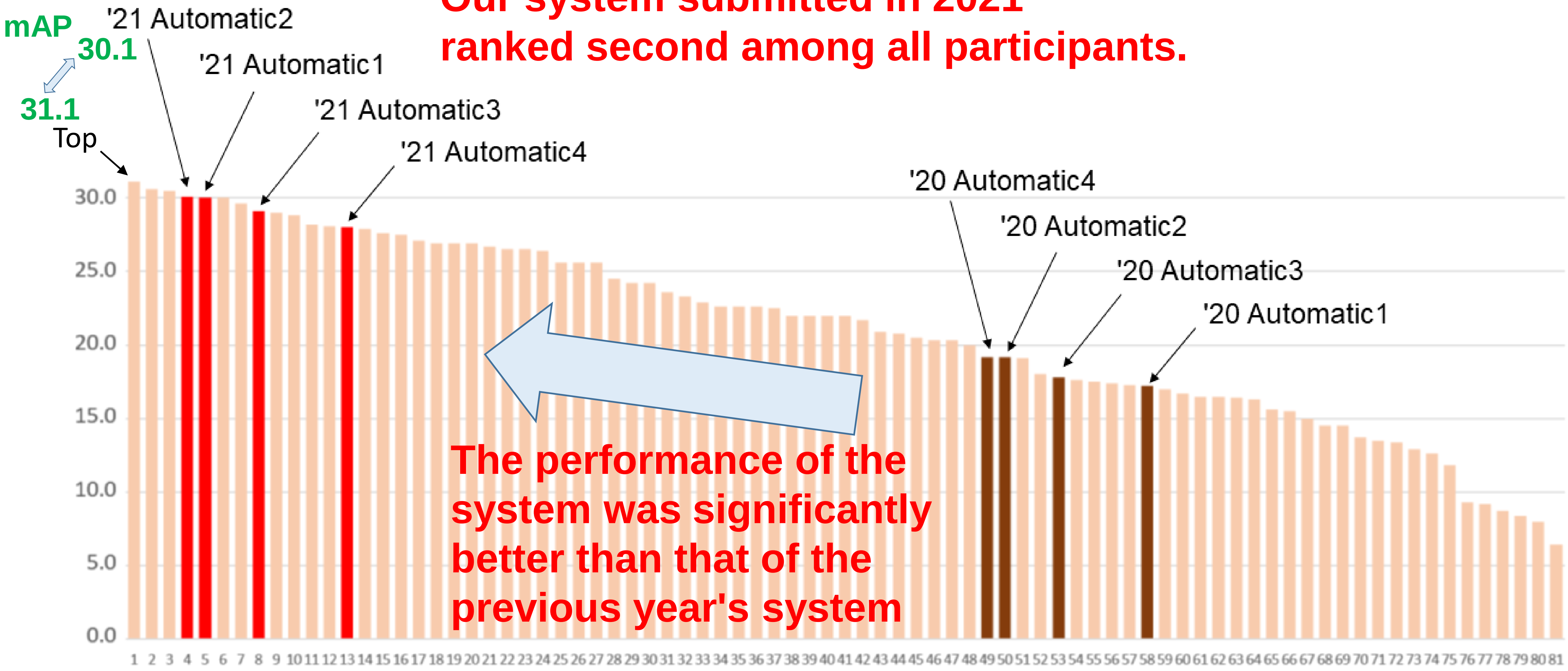
Systems submitted to the main task in 2021

4th among all participants

2nd among all participants



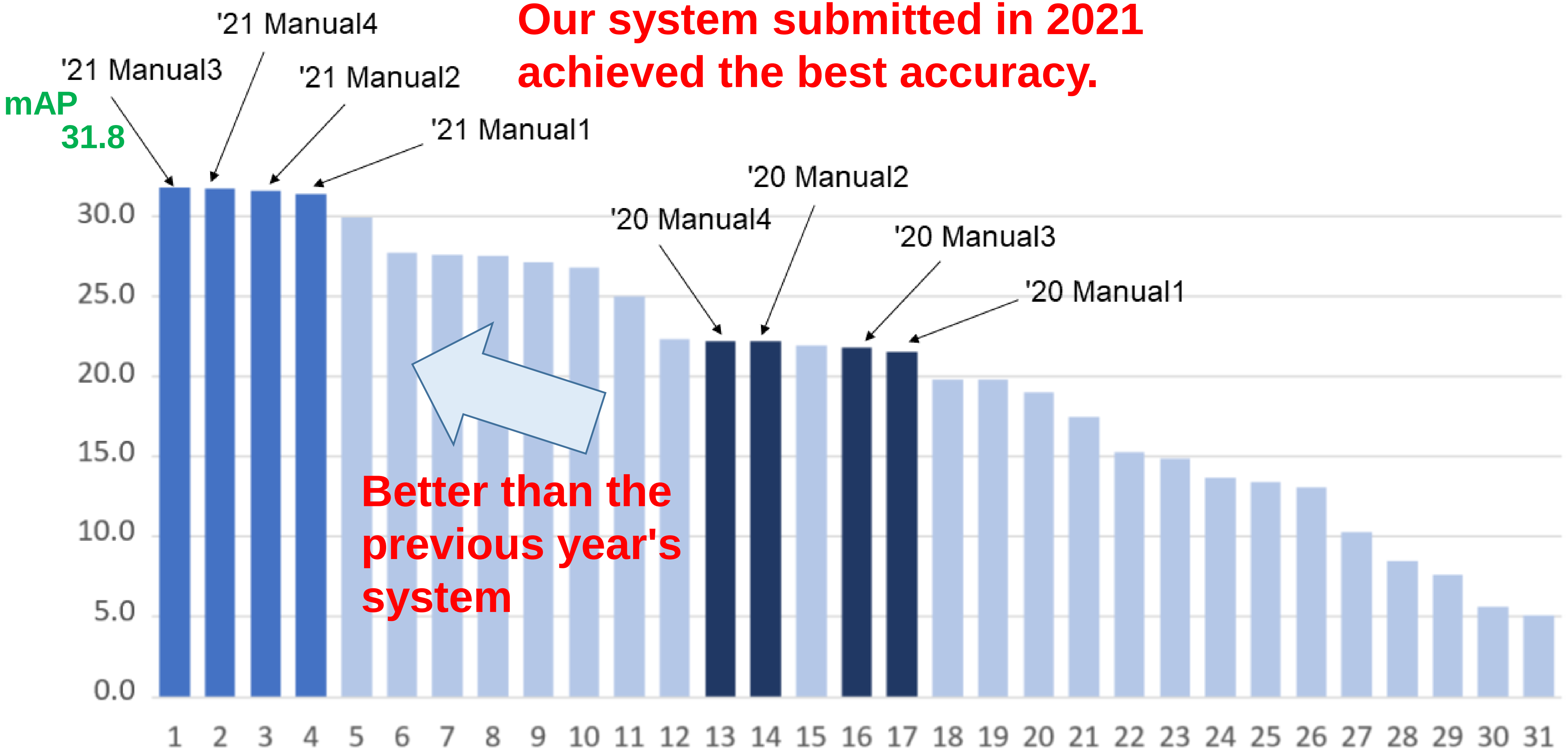
Fully automatic runs for 2019-2021 progress task



Our system submitted in 2021 ranked second among all participants.

The performance of the system was significantly better than that of the previous year's system

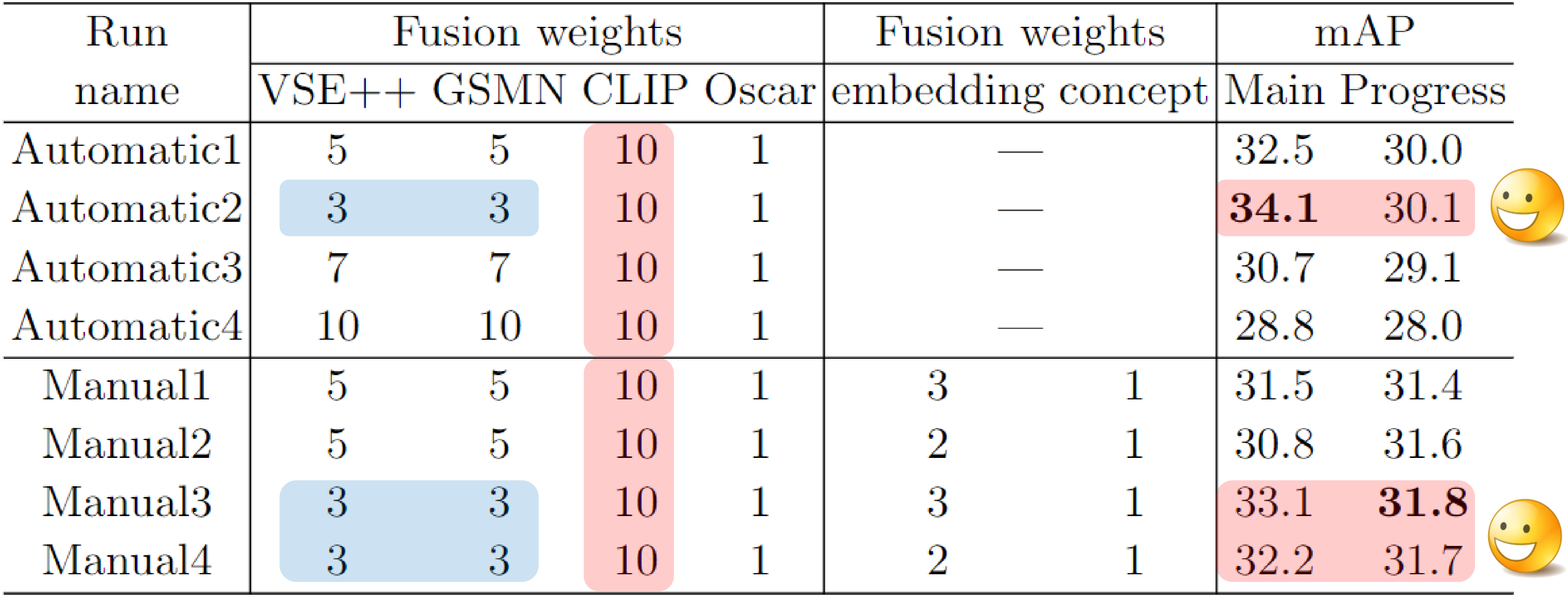
Manually assisted runs for 2019-2021 progress task



Our system submitted in 2021 achieved the best accuracy.

Better than the previous year's system

Our submitted runs for TRECVID 2021 AVS task

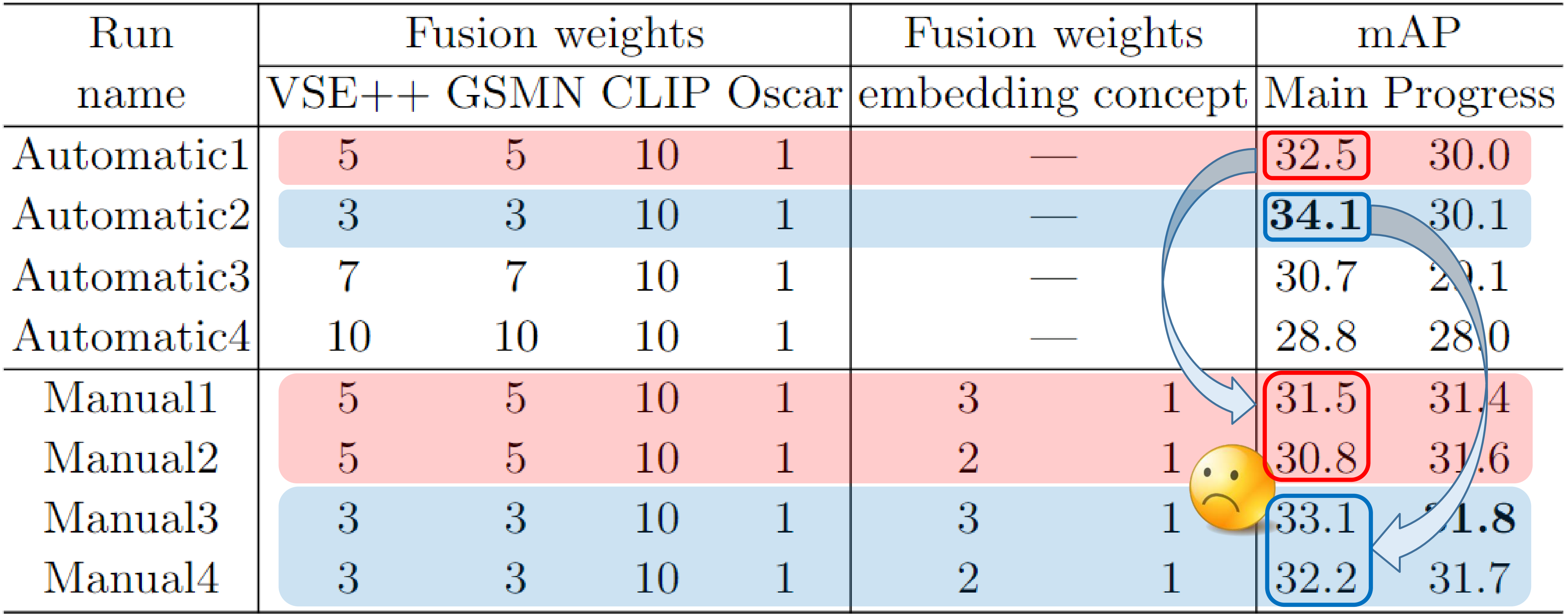


Run	Fusion weights				Fusion weights		mAP	
name	VSE++	GSMN	CLIP	Oscar	embedding	concept	Main	Progress
Automatic1	5	5	10	1	—		32.5	30.0
Automatic2	3	3	10	1	—		**34.1**	30.1
Automatic3	7	7	10	1	—		30.7	29.1
Automatic4	10	10	10	1	—		28.8	28.0
Manual1	5	5	10	1	3	1	31.5	31.4
Manual2	5	5	10	1	2	1	30.8	31.6
Manual3	3	3	10	1	3	1	33.1	**31.8**
Manual4	3	3	10	1	2	1	32.2	31.7

The accuracy is highest when the integration weight of CLIP is large.

CLIP has a different output tendency and higher retrieval accuracy than VSE++ and GSMN.

Our submitted runs for TRECVID 2021 AVS task



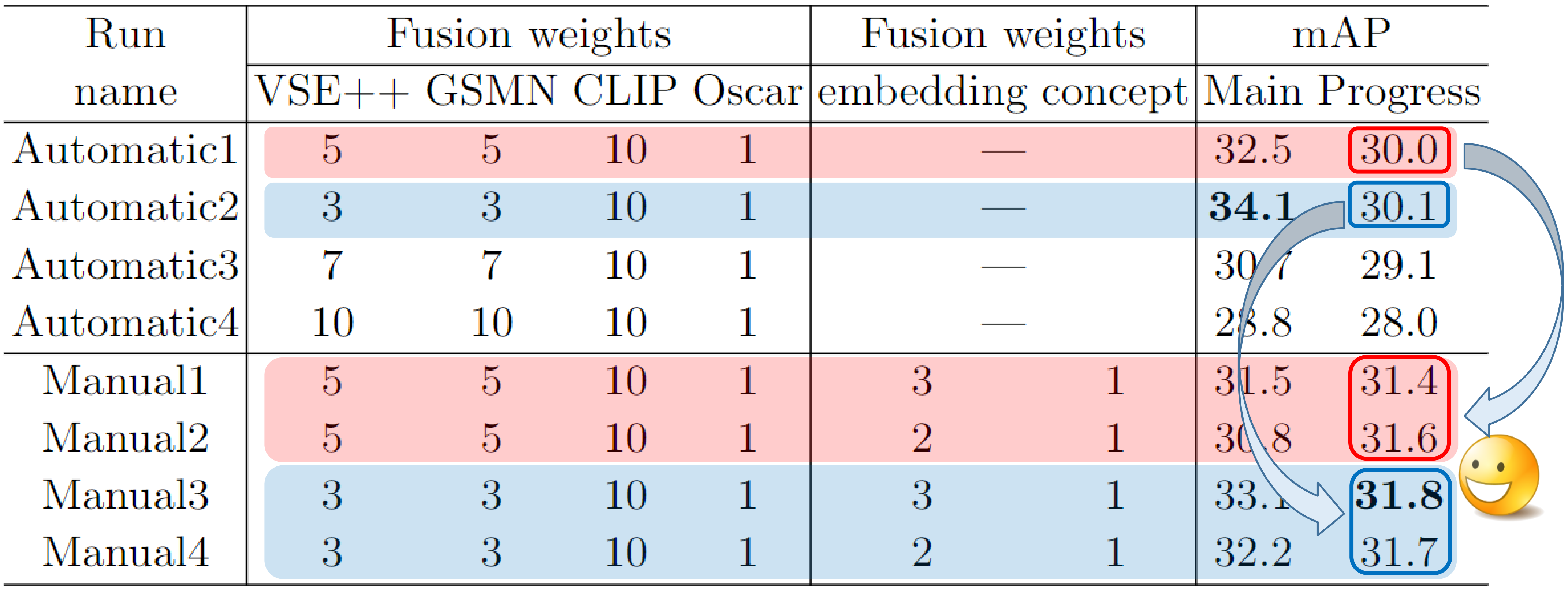
Run	Fusion weights				Fusion weights		mAP	
name	VSE++	GSMN	CLIP	Oscar	embedding	concept	Main	Progress
Automatic1	5	5	10	1	—		32.5	30.0
Automatic2	3	3	10	1	—		**34.1**	30.1
Automatic3	7	7	10	1	—		30.7	29.1
Automatic4	10	10	10	1	—		28.8	28.0
Manual1	5	5	10	1	3	1	31.5	31.4
Manual2	5	5	10	1	2	1	30.8	31.6
Manual3	3	3	10	1	3	1	33.1	**31.8**
Manual4	3	3	10	1	2	1	32.2	31.7

Were the concept-based and embedding methods complementary? ???

⇨ Not so sure.

Main task: Embedding > Embedding + Concept-based

Our submitted runs for TRECVID 2021 AVS task



Run	Fusion weights				Fusion weights		mAP	
name	VSE++	GSMN	CLIP	Oscar	embedding	concept	Main	Progress
Automatic1	5	5	10	1	—		32.5	30.0
Automatic2	3	3	10	1	—		**34.1**	30.1
Automatic3	7	7	10	1	—		30.7	29.1
Automatic4	10	10	10	1	—		28.8	28.0
Manual1	5	5	10	1	3	1	31.5	31.4
Manual2	5	5	10	1	2	1	30.8	31.6
Manual3	3	3	10	1	3	1	33.1	**31.8**
Manual4	3	3	10	1	2	1	32.2	31.7

Were the concept-based and embedding methods complementary? ???

⇨ Not so sure.

Progress task: Embedding < Embedding + Concept-based

Summary

- In the systems submitted this year, we introduced new embedding methods that have been proposed in recent years, such as GSMN, CLIP, and Oscar.

- The evaluation results showed that the accuracy of the system was significantly better than that of the previous year’s system, indicating that the recent pre-training mechanism using large-scale image-text pairs is beneficial.

- All embedding methods we used were image-based and did not take advantage of the characteristics of the video.

⇩

- For future works, it is necessary to consider methods for embedding video features and text features.