Knowledge transfer from CLIP to VGS

BOGON RYU

Table of Contents

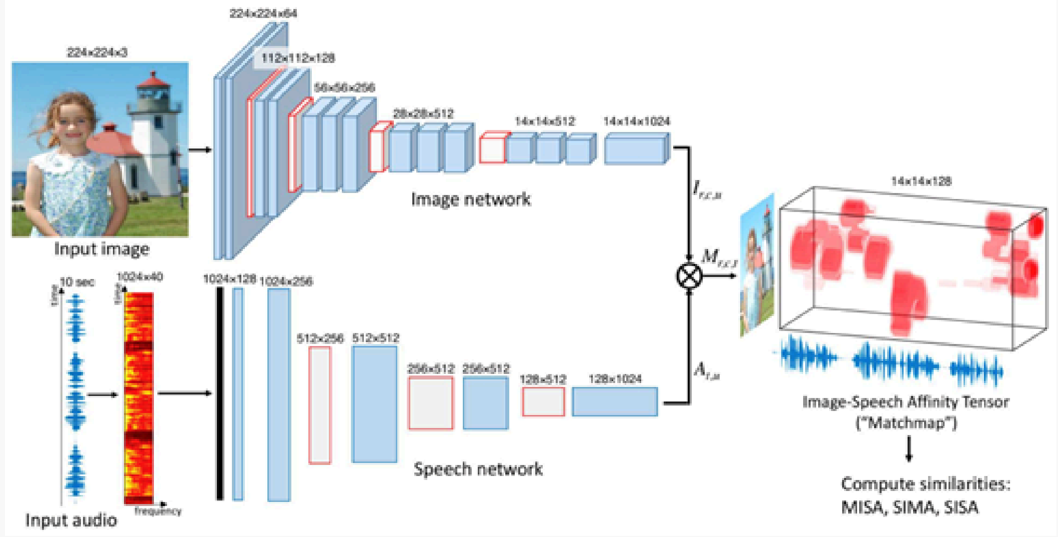
		Page
I	Introduction	3
II	Method	10
III	Experiment	14
IV	Results	15
V	Conclusion	22

What is Audio-Visual Learning?

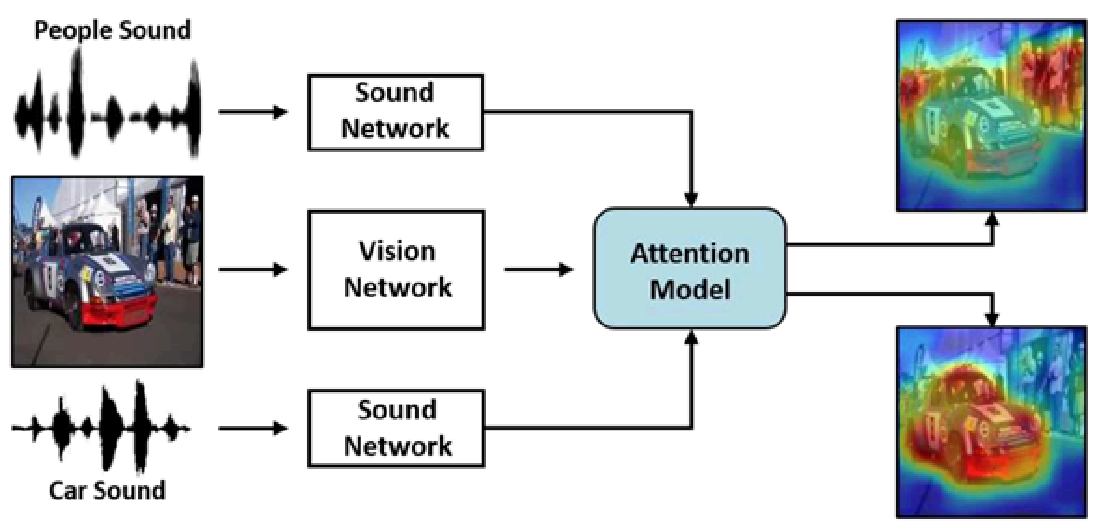
Natural audio paired with the visual signal

Descriptive narration for the visual signal

Speech ? Audio ?



Visually Grounded Speech



Sound Source Localization

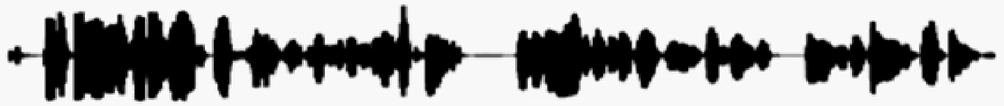
Red and white colored fire truck in front of the station shown during the day

PlacesAudio

VGGSound

Visually grounded speech(VGS)

There's a large open area with very very large rock



Picture of the parking lot of a fire station with three or four firetruck.

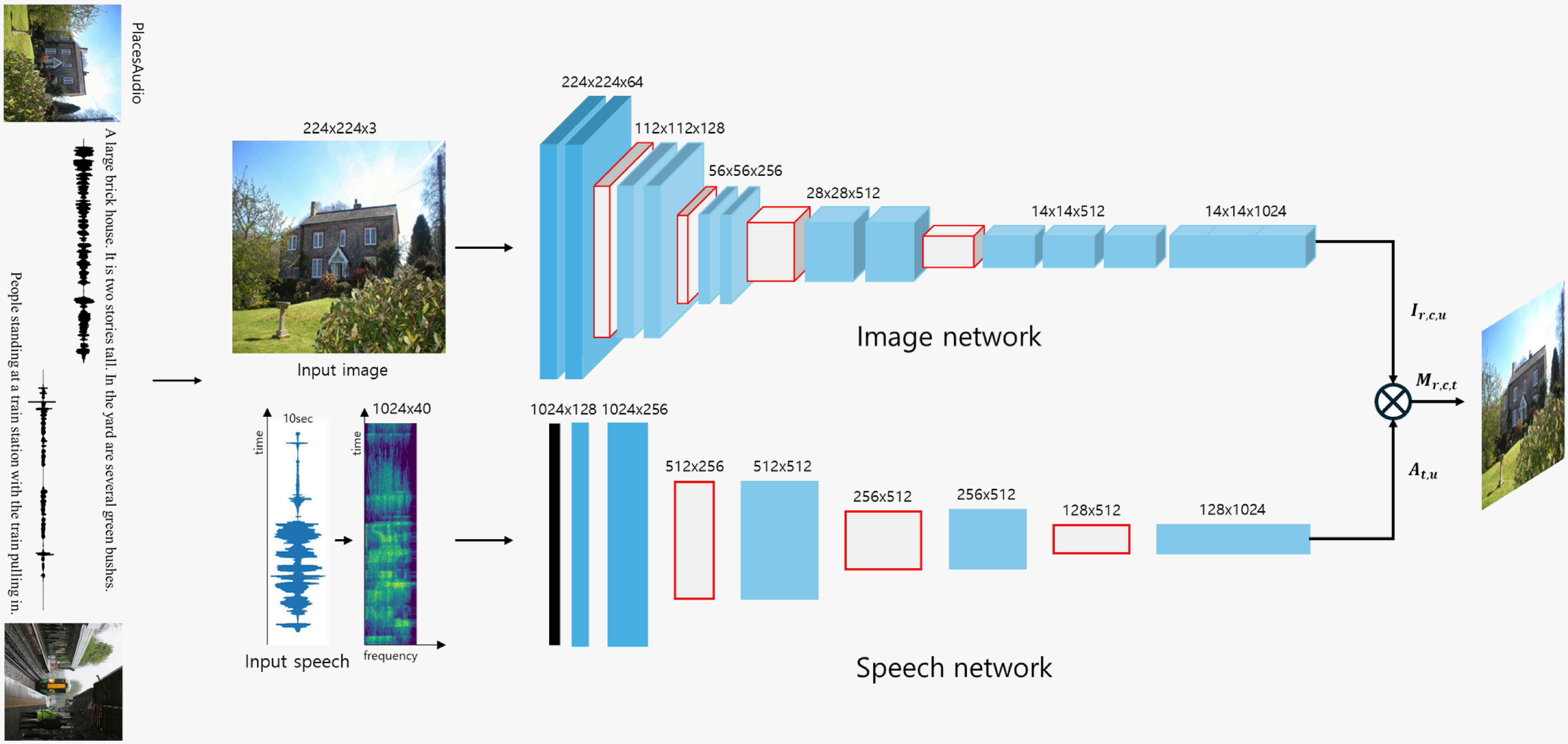
A large brick house. It is two stories tall. In the yard are several green bushes.

People standing at a train station with the train pulling in.

Spoken sentense-Visual Pair is provided.

Retrieve the most proper descriptive narration/image

Visually grounded speech



Difficulties

Collecting high quality spoken sentense-visual pair is difficult.

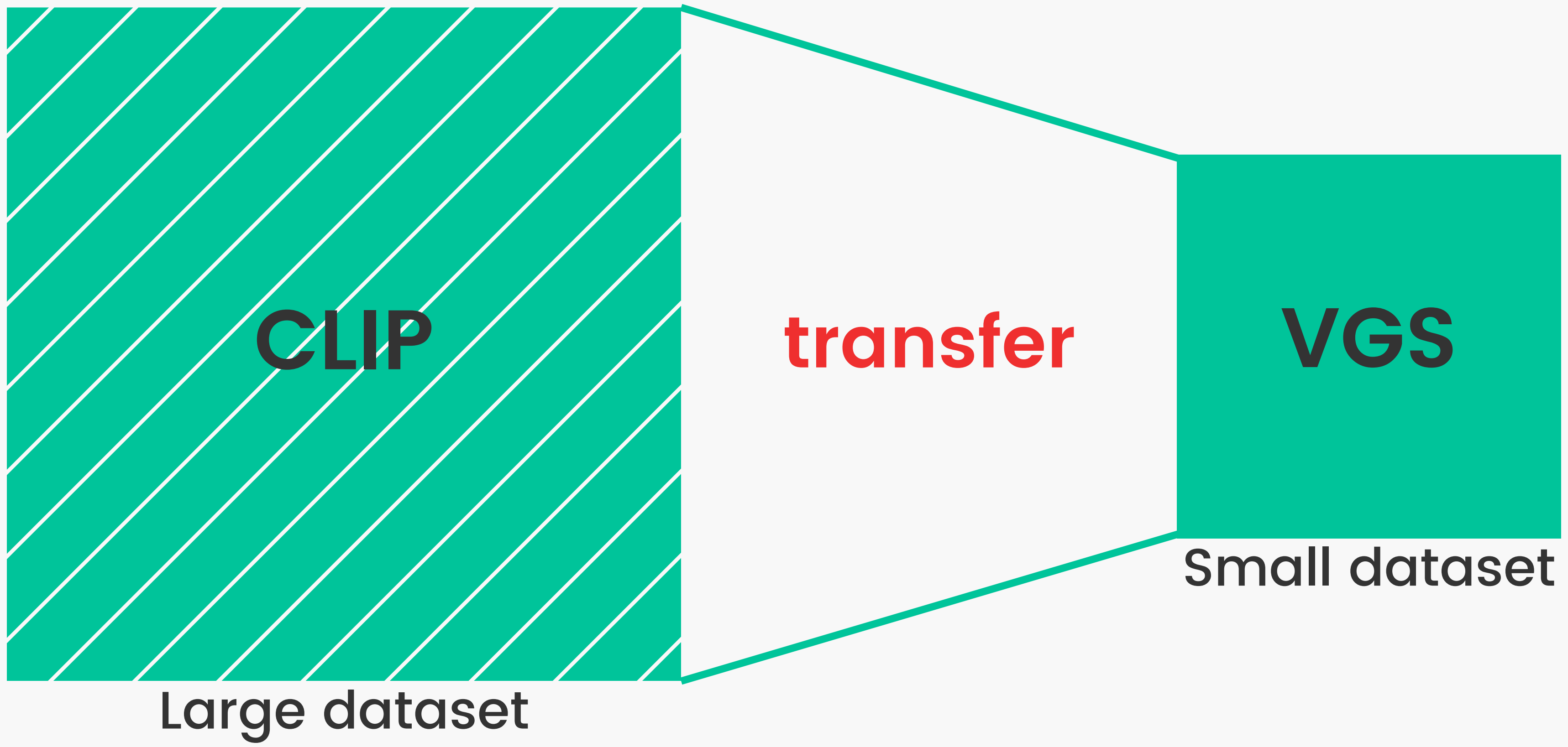
Picture of the parking lot of a fire station with three or four firetruck.

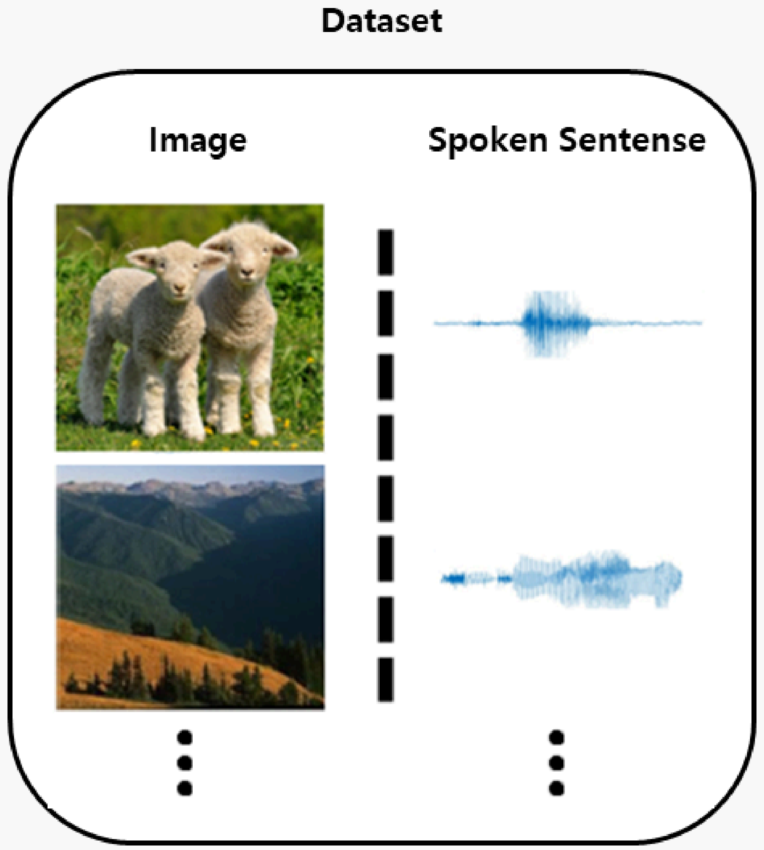
A police car and a yellow car are in a car racing competition.
The cars are making smoke. There is a crowd behind the cars.

Challenges in Creating or Collecting Large-Scale Datasets extremely difficult

Motivation

“Apply knowledge transfer from CLIP to VGS”



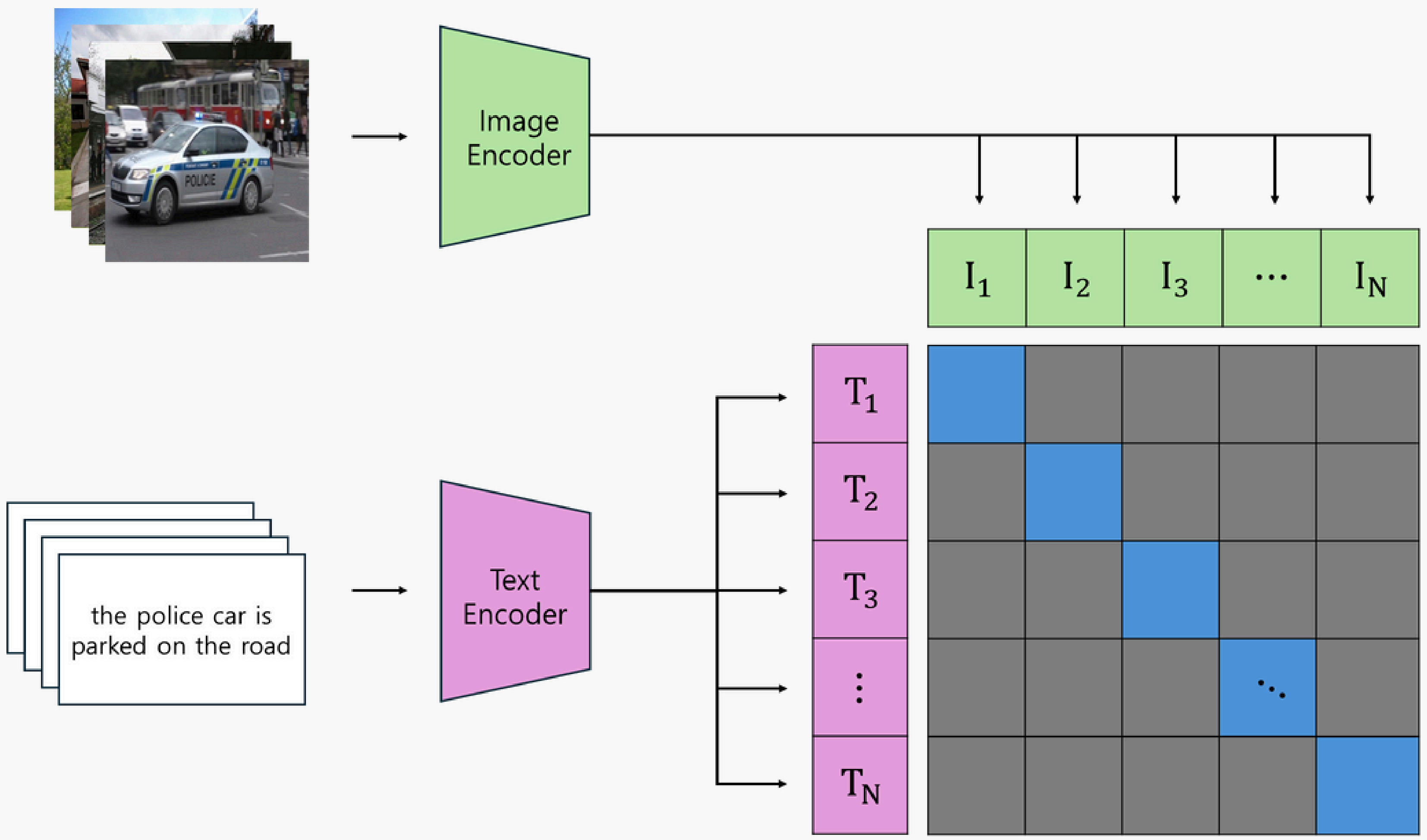


We will use CLIP for similar sample mining.

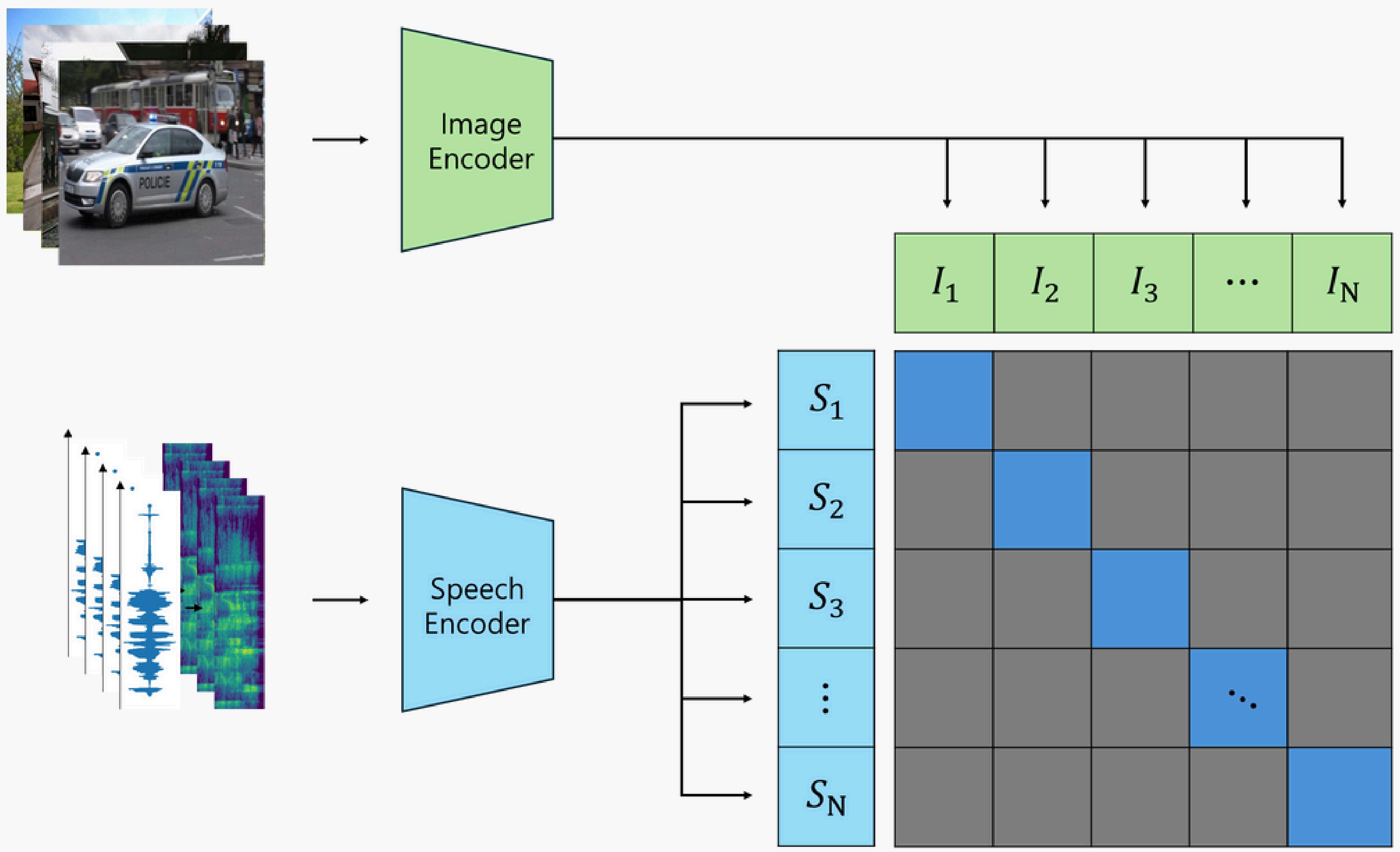
"Our goal is to distill the knowledge from the CLIP model into the Visually Grounded Speech (VGS) system to improve its retrieval score."

Comparison and Contrast: CLIP vs VGS

Why do we use CLIP?

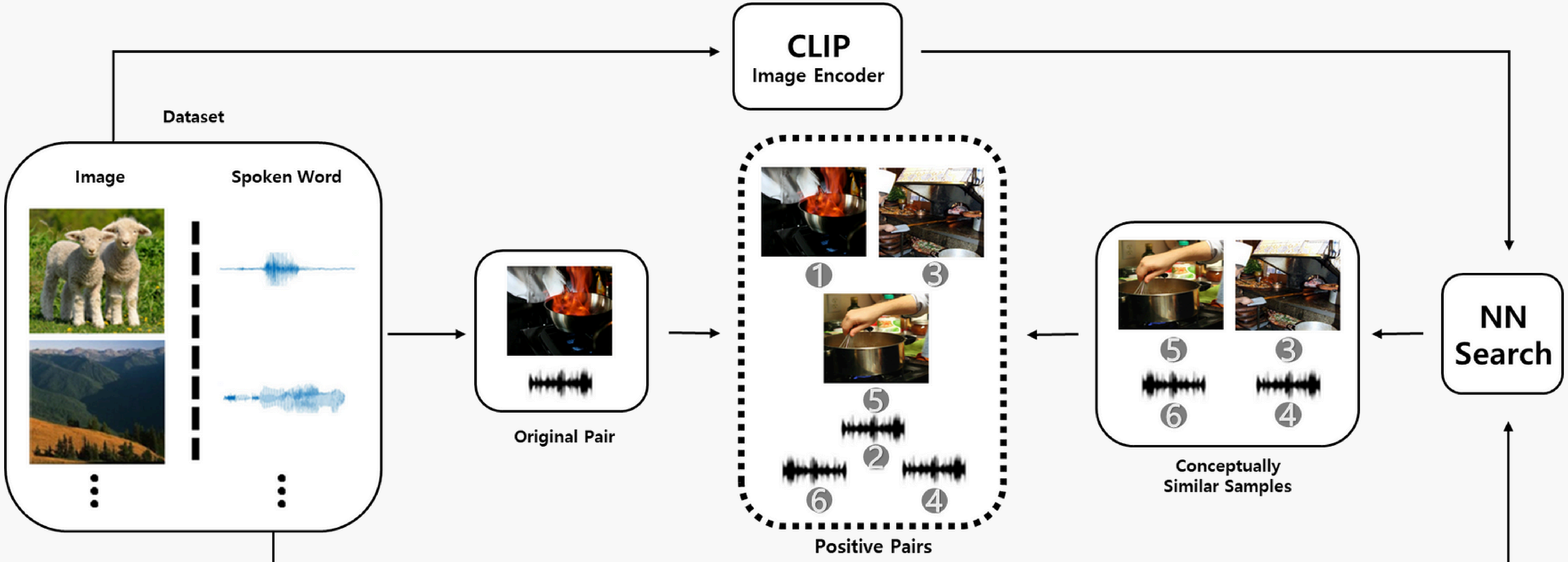


CLIP



VGS

Semantically Similar Samples

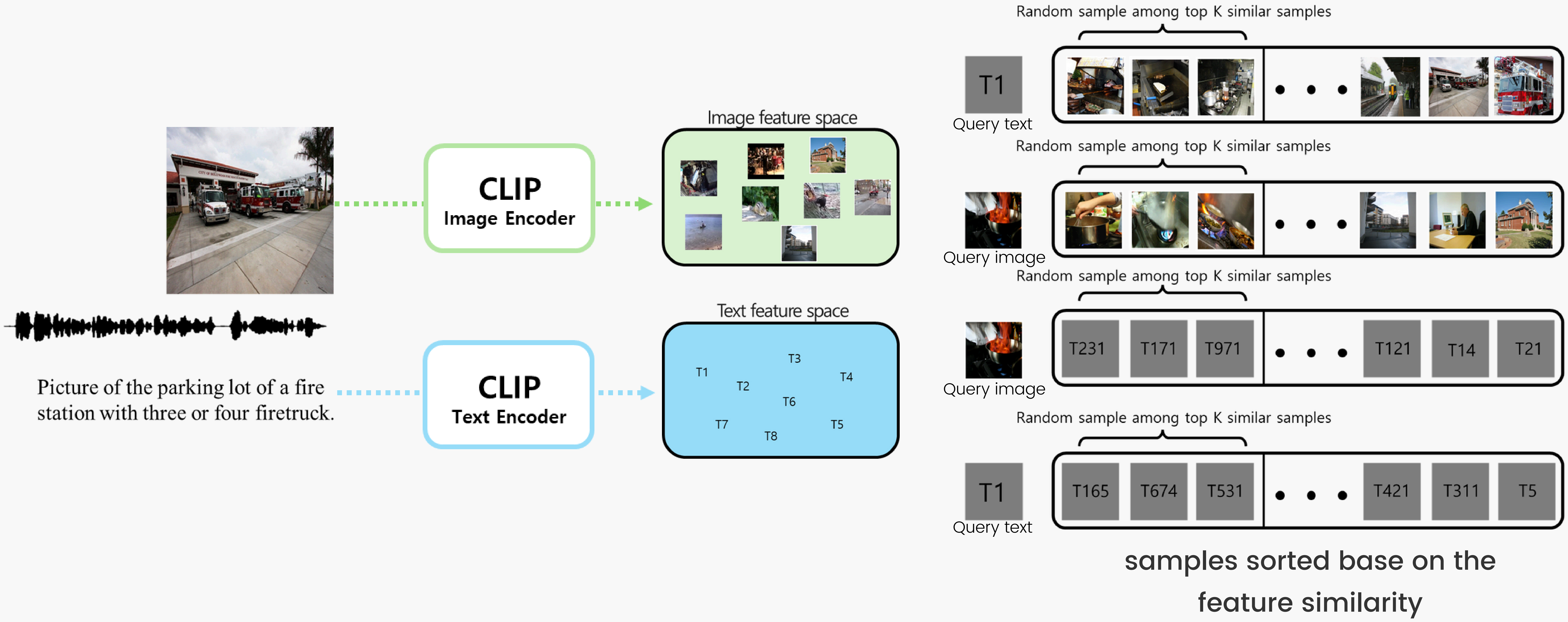


Our Approach

① Original Image	② Original speech
③ Image neighbor based on Text similarity	④ Text neighbor based on Text similarity
⑤ Image neighbor based on Image similarity	⑥ Text neighbor based on Image similarity

How to collect Semantically Similar Samples

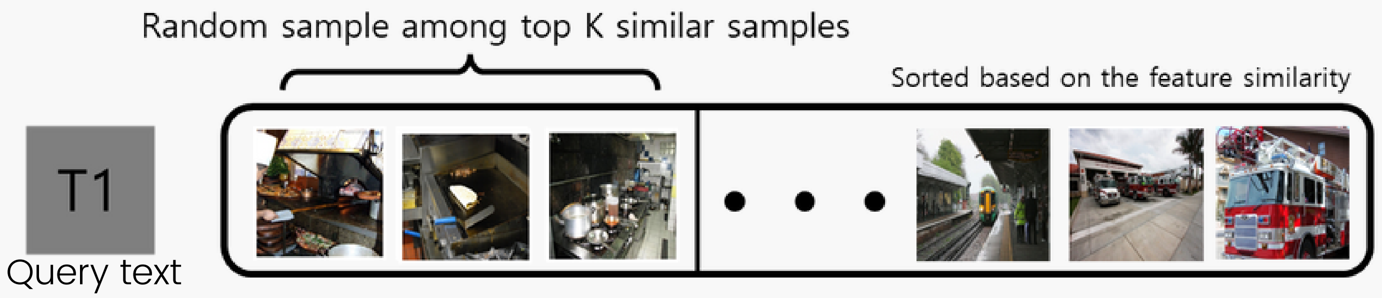
"Apply knowledge transfer from CLIP to VGS"



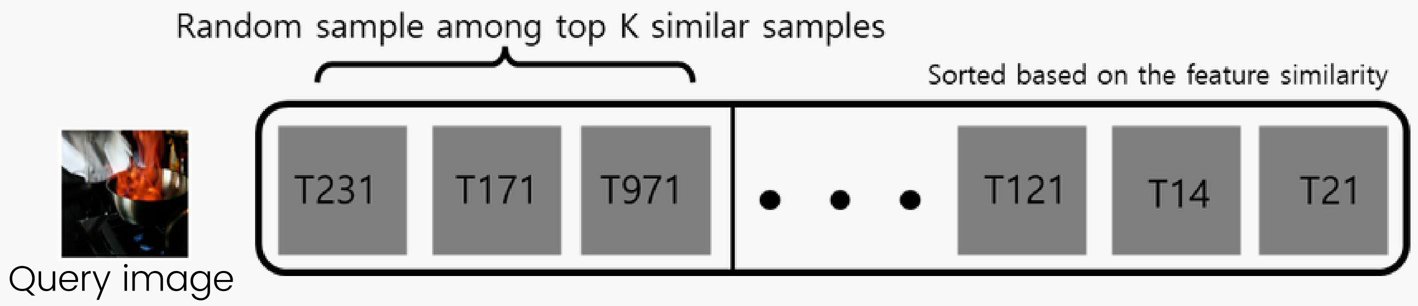
Method

How to collect Semantically Similar Samples

"Apply knowledge transfer from CLIP to VGS"



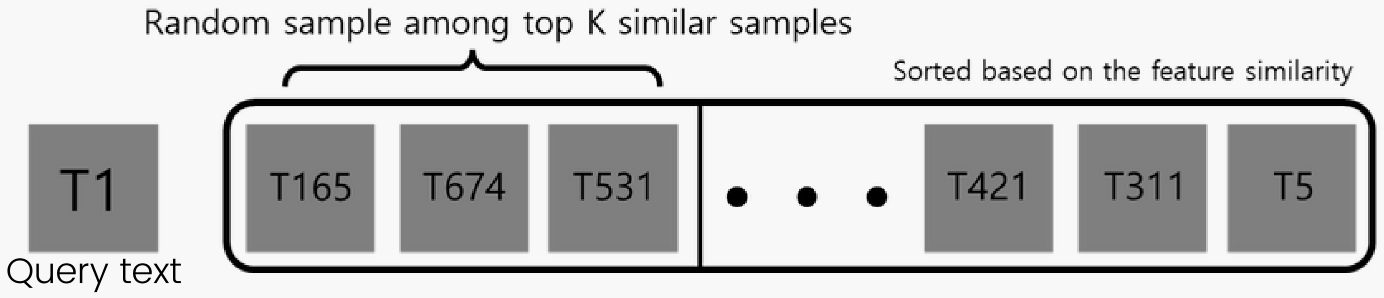
T1 : The food is being cooked on a stove



T231 : Two pots of soup are cooking on a commercial stove

T171 : A man and a chef's hat and white uniform is seen within a bowl of something while a pot

T971 : Close up photo of a chef working on some sort of a dish is pouring sugar or salt on top of it



T1 : The food is being cooked on a stove

T165 : The food is being cooked in skillet there's also some green vegetables

T674 : There are people cooking food in a kitchen

T531 : This picture we also see some delicious food on a plate being prepared for dinner

Method

How to collect Semantically Similar Samples

"Apply knowledge transfer from CLIP to VGS"

T1 : The food is being cooked on a stove

<org,org>

T231 : Two pots of soup are cooking on a commercial stove

<org, i2t>

T674 : There are people cooking food in a kitchen

<org, t2t>

T1 : The food is being cooked on a stove

<i2i, org>

T231 : Two pots of soup are cooking on a commercial stove

<i2i, i2t>

T674 : There are people cooking food in a kitchen

<i2i, t2t>

T1 : The food is being cooked on a stove

<t2i, org>

T231 : Two pots of soup are cooking on a commercial stove

<t2i, i2t>

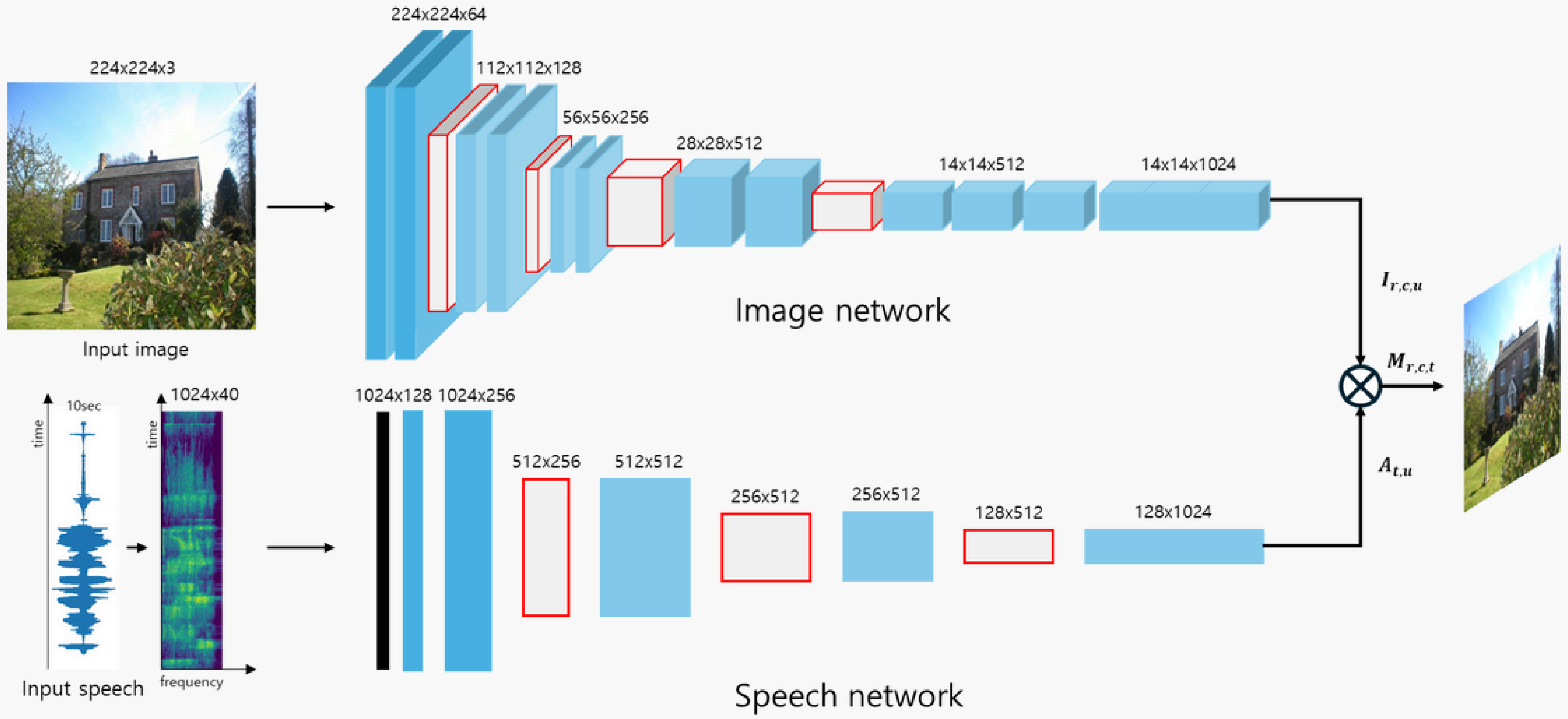
T674 : There are people cooking food in a kitchen

<t2i, t2t>

My Method : Knowledge transfer from CLIP to VGS

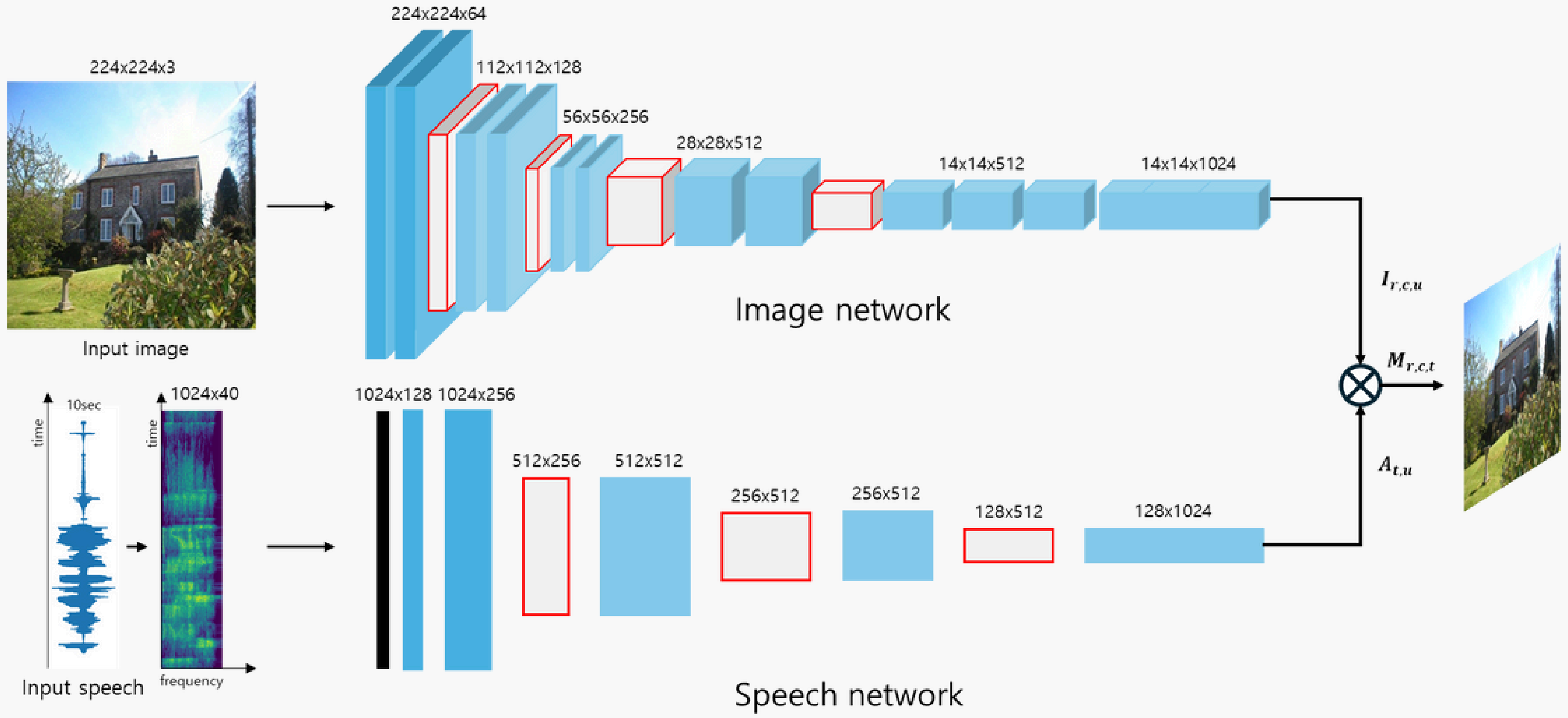
Baseline

Original Pair



Propsed method

Positive Pairs



Quantitative Results : Ablation study

Quantitative results on Places Audio Caption dataset out of 1000 samples

	NN Search					A → I			I → A		
	Original	T2T	T2I	I2T	I2I	R@1	R@5	R@10	R@1	R@5	R@10
(A)	✓	✗	✗	✗	✗	10.9	33.2	46.9	11.3	34.2	45.4
(B)	✓	✓	✗	✗	✗	11.4	36.5	49.8	12.6	**38.5**	49.3
(C)	✓	✗	✓	✗	✗	12.7	38.2	51.9	**15.1**	38.3	**51.5**
(D)	✓	✗	✗	✓	✗	10.6	32.9	49.2	12.5	35.3	48.2
(E)	✓	✗	✗	✗	✓	10	32.7	47.1	10.7	33.1	47.1
(F)	✓	✓	✗	✗	✓	11.5	35.4	50.4	12.1	37.5	50.1
(G)	✓	✓	✗	✓	✗	**12.8**	**38.7**	52.1	13.7	38.1	51
(H)	✓	✓	✓	✓	✗	11.5	36.8	51.5	13.2	38.2	50.9
(I)	✓	✓	✓	✓	✓	10.6	35.6	**52.2**	12.1	36.1	51

Table 1. **Ablation studies on our proposed method to see the impact of each positive pair.**

These experiments were conducted using a single GTX1080ti GPU.

Quantitative Results : Ablation study

Why does using i2i images as similar pairs result in lower performance?

Original pair

The food is being cooked on a stove

Original pair

Top 30 <i2t>

A stove that's heating up a metal pot that has a water boiling in it

Similar pair

A stove that's heating up a metal pot that has a water boiling in it

Original pair

Top 30 <t2t>

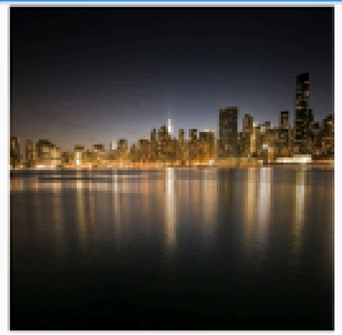
And there is a big pot on the stove

Similar pair

And there is a big pot on the stove

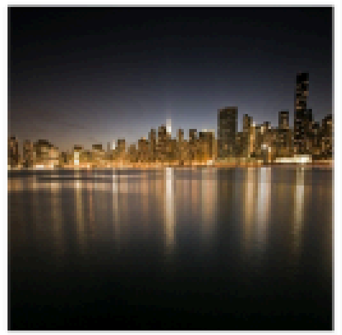
Original pair

Top 30 <i2i>



You can see the inside of a boat made out of wood

Similar pair



The food is being cooked on a stove

Original pair

Top 30 <t2i>

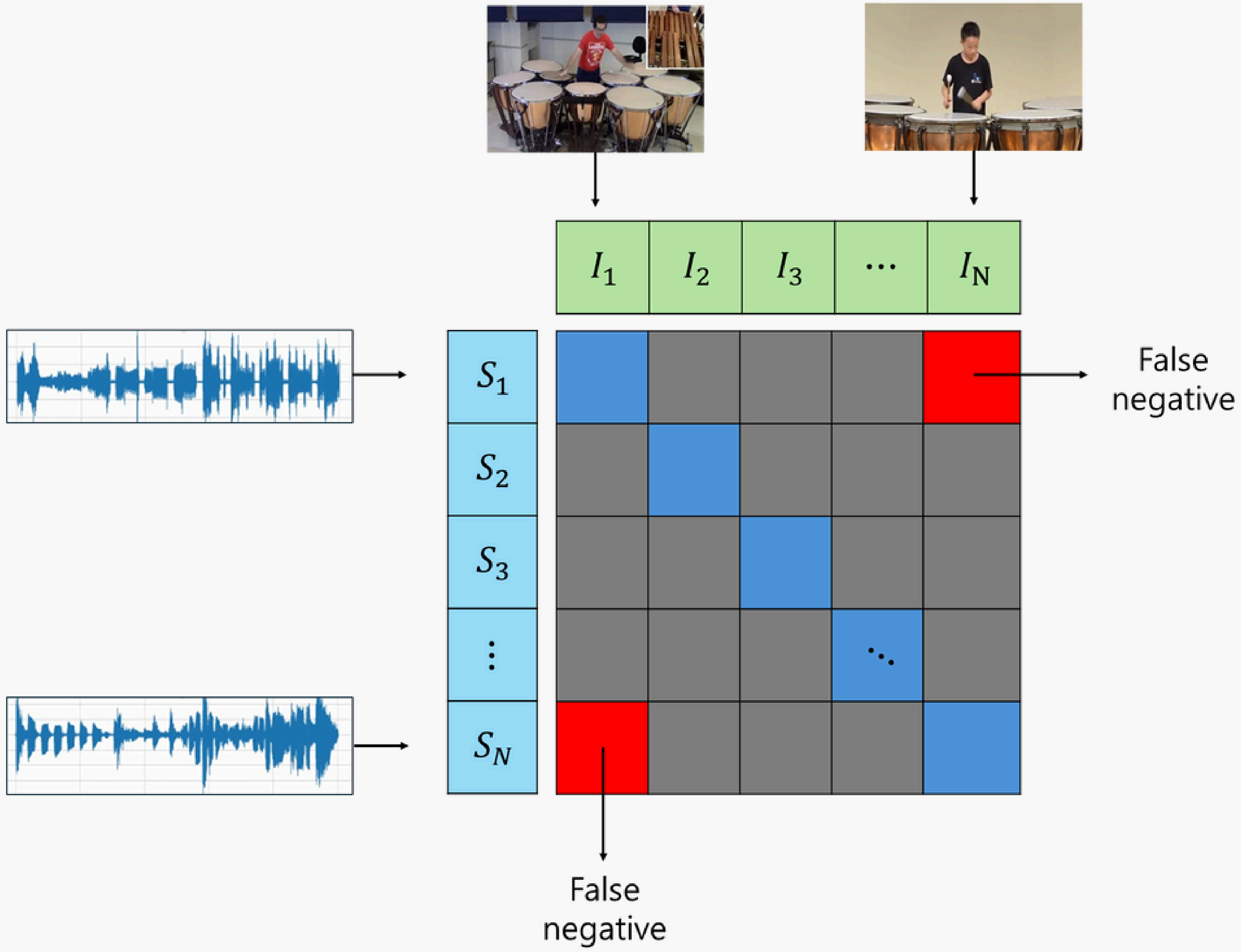
A ceramic bowl that has agreed and said

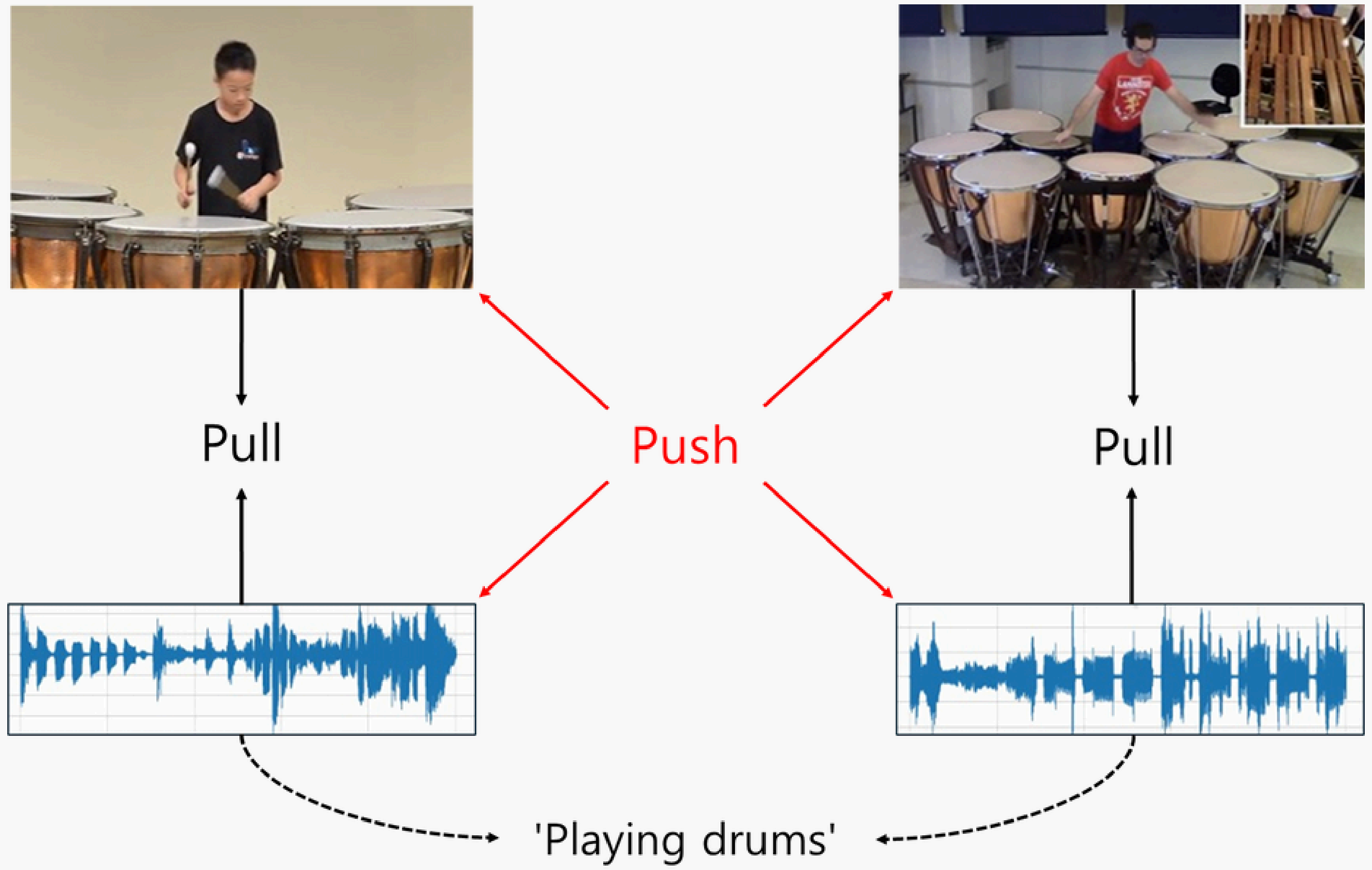
Similar pair

The food is being cooked on a stove

False Negative Aware Contrastive Learning

Why did the Retrieval Score Increase?



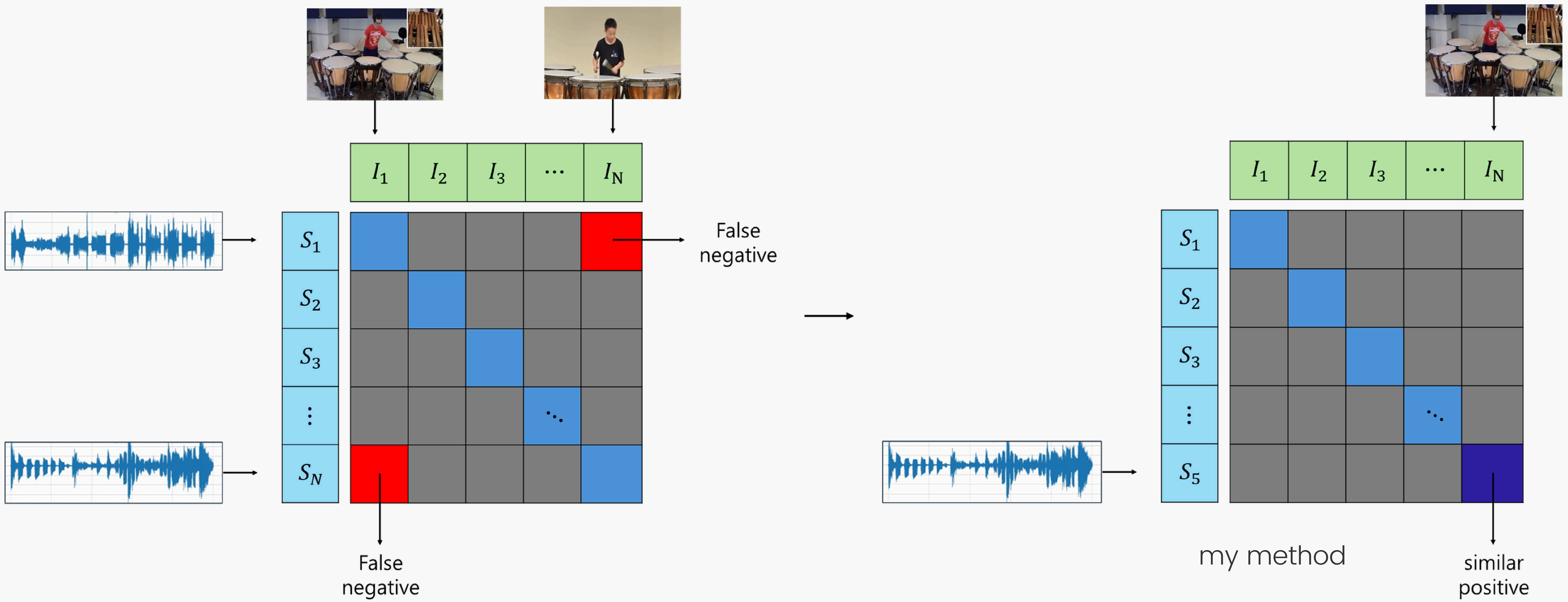


We discover that with a batch size of 128, around 40% of the samples in VGG Sound will encounter at least one false negative sample during training.

Learning Audio-Visual Source Localization via False Negative Aware Contrastive Learning(https://arxiv.org/abs/2303.11302)

False Negative Aware Contrastive Learning

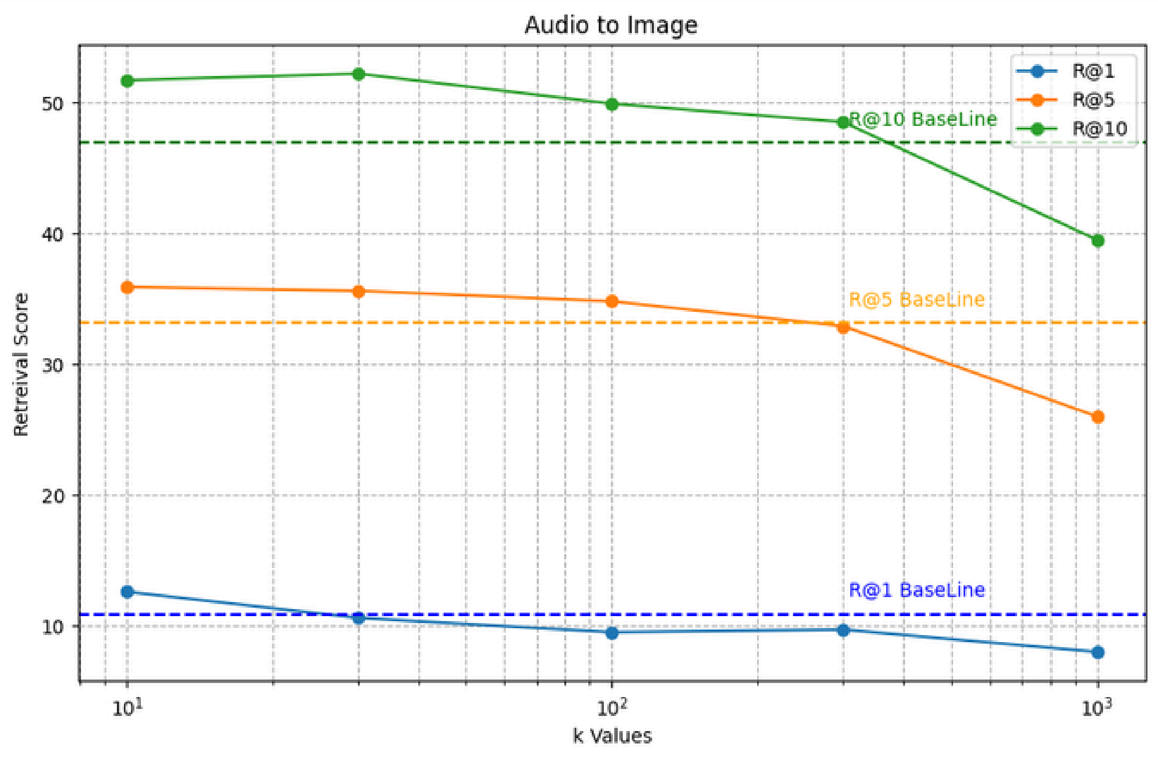
Why did the Retrieval Score Increase?

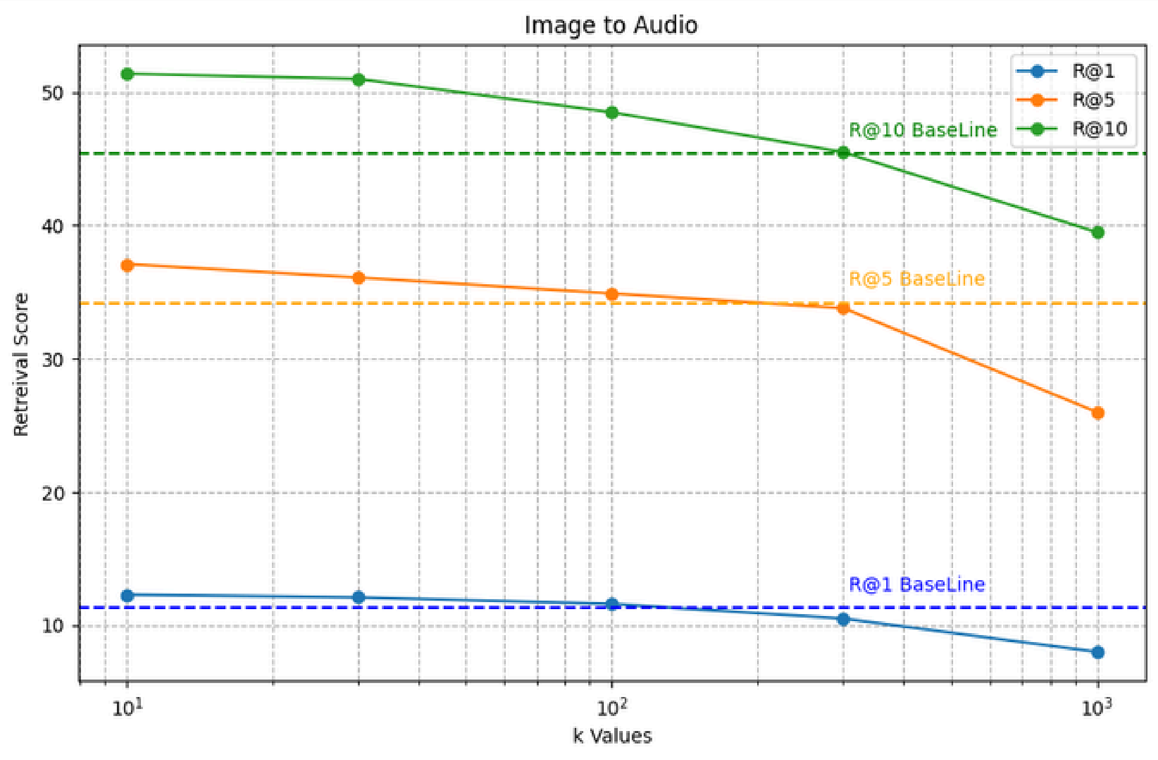


"False negatives lead to push, but they are pulled back due to positive similarities."

Quantitative Results : K - Ablation study

Quantitative results on Places Audio Caption dataset out of 1000 samples





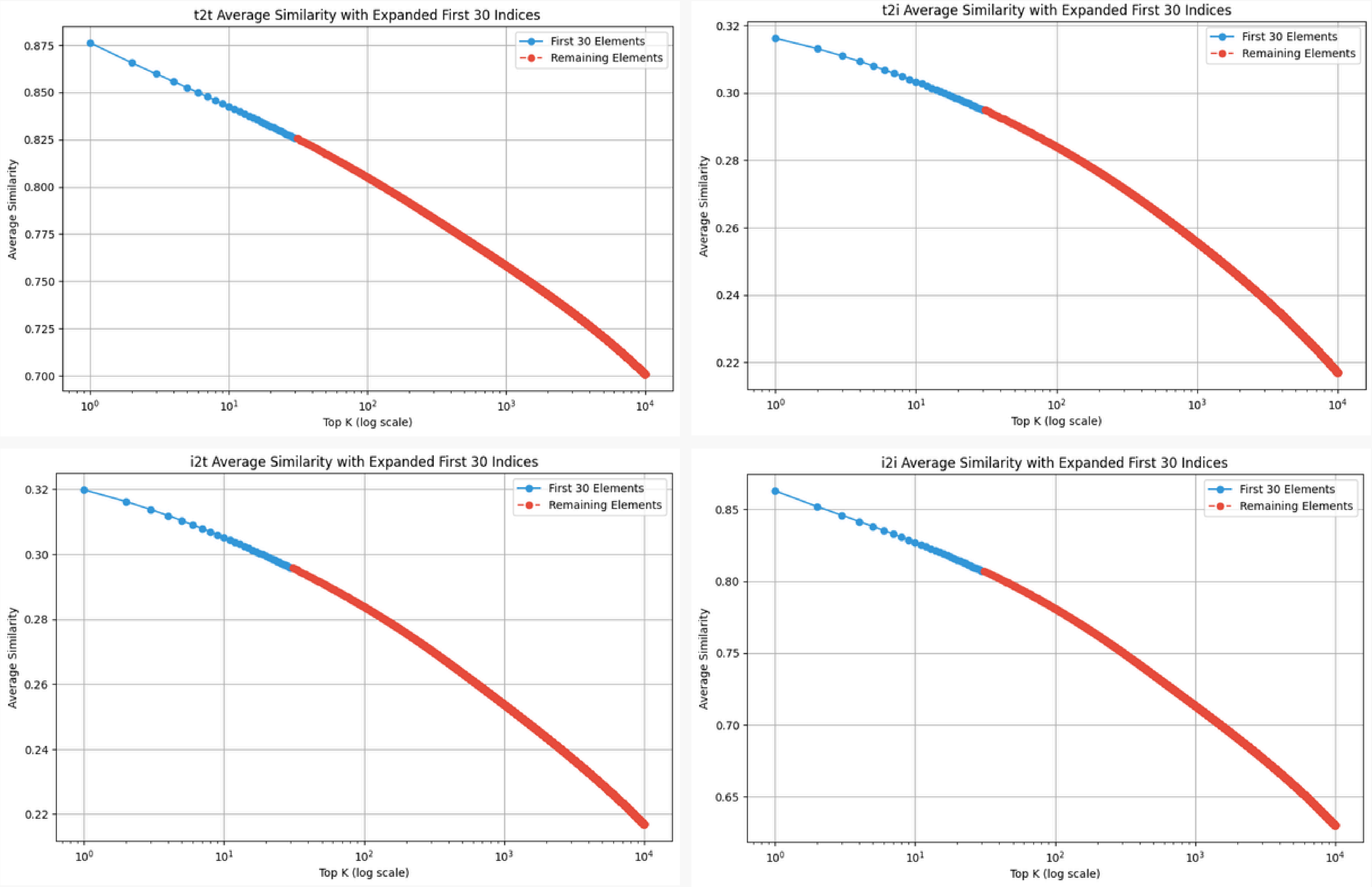
	k in *k*-NN	10	30	100	300	1000
A → I	R@1 ↑	**12.6**	10.6	9.5	9.7	8
	R@5 ↑	**35.9**	35.6	34.8	32.9	26
	R@10 ↑	51.7	**52.2**	49.9	48.5	39.5
I → A	R@1 ↑	**12.3**	12.1	11.6	10.5	8.6
	R@5 ↑	**37.1**	36.1	34.9	33.8	27.9
	R@10 ↑	**51.4**	51	48.5	45.5	40.1

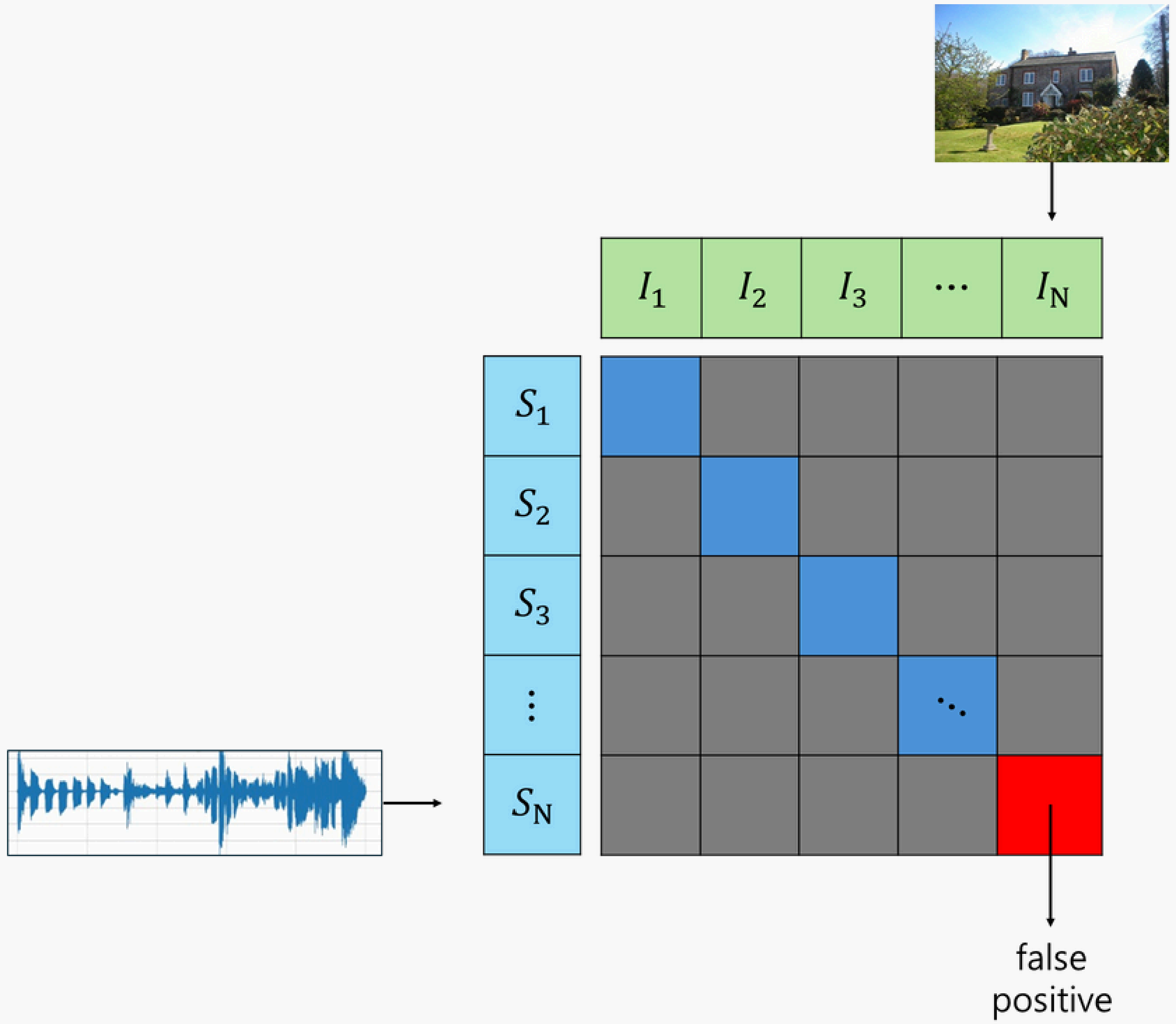
Table 2. **Varying k in conceptually similar sample selection.**

I Results

Quantitative Results : Basis of the research results

Average of the Top K similarity





This graph represents the logarithmic scale of the expression, which is the average of the top K similarities computed for each of the 100,000 features against the other 100,000 features

Quantitative Results: Curve Based on Data Set Size

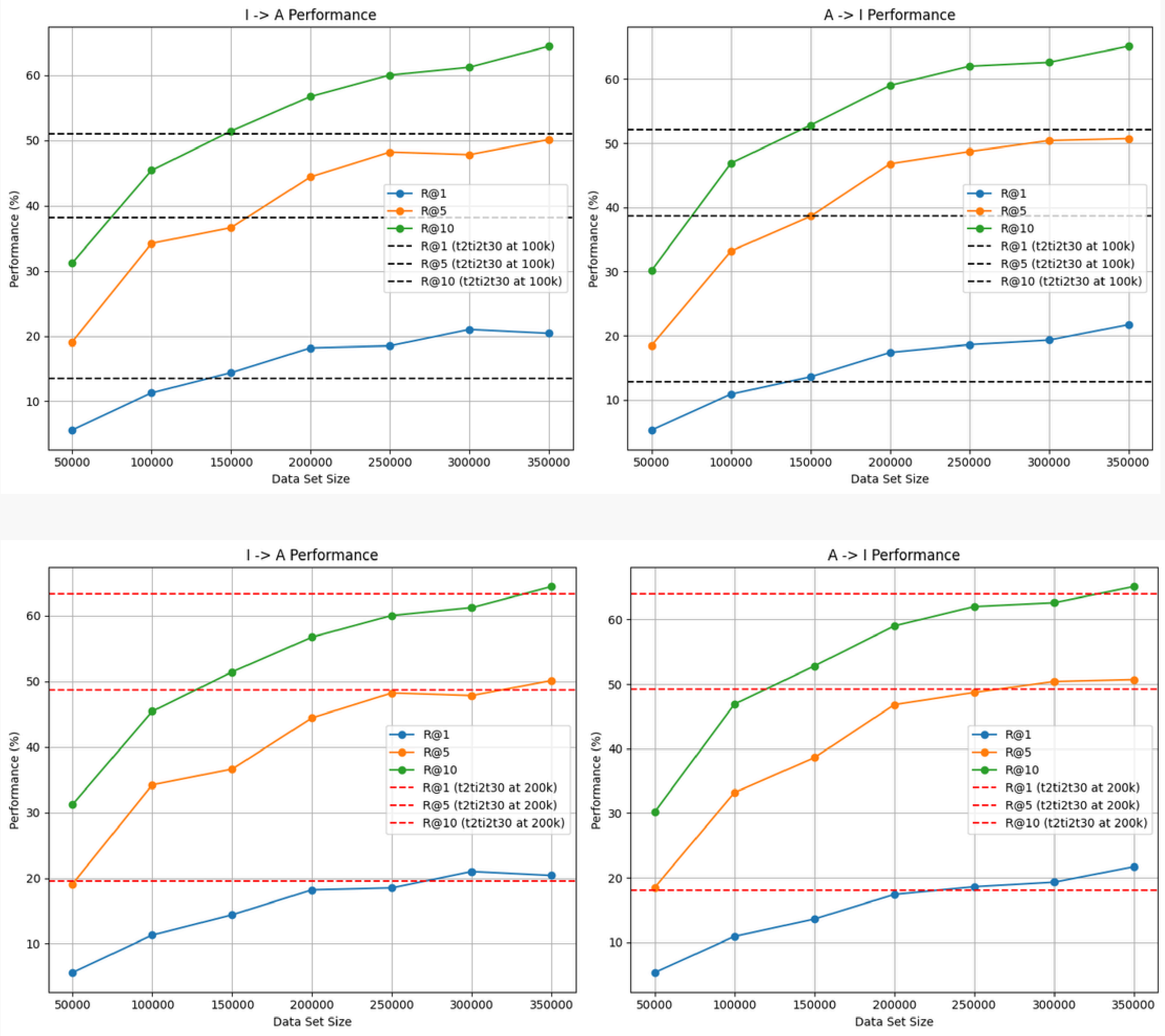
Performance Efficiency Comparison: Baseline vs. Use of Positive Pairs

	Dataset	50K	100K	150K	200K	250K	300K	350K
A → I	R@1 ↑	5.3	10.9	13.6	17.4	18.6	19.3	21.7
	R@5 ↑	18.5	33.2	38.6	46.8	48.7	50.4	50.7
	R@10 ↑	30.2	46.9	52.8	59	62	62.6	65.1
I → A	R@1 ↑	5.6	11.3	14.4	18.2	18.5	21	20.4
	R@5 ↑	19.1	34.2	36.6	44.4	48.2	47.8	50.1
	R@10 ↑	31.2	45.4	51.4	56.7	60	61.2	64.4

Table 3. **Baseline performance varying with dataset size.**

	NN Search			A → I			I → A		
	Original	T2T	I2T	R@1	R@5	R@10	R@1	R@5	R@10
100K	✓	✓	✓	12.8	38.7	52.1	13.7	38.1	51
200K	✓	✓	✓	18	49.2	63.9	19.5	48.7	63.3

Table 4. **Performance of positive pairs varying with dataset size.**



- "We are able to mine similar samples through knowledge transfer from CLIP.

- Using this for training, we observe an improvement in retrieval scores. This allows us to transfer the vast amount of knowledge from CLIP to the Davenet model.

- False negatives can hinder the learning process, but by mining similar samples and using them for training, we achieve improved performance."

Thank you for listening