

# **FILIP: Fine-Grained Interactive Language-Image Pre-Training**

**ICLR 2022 Poster (336 Citations)**

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang,  
Zhenguo Li, Xin Jiang, Chunjing Xu

Presented by:

Anthony Bilic, Kunyang Li, David Shatwell, Zain Ulabedeen Farhat, Kevin Zhai

# Outline

1. Background/Motivation
2. Method
3. Results
4. Conclusion
5. Limitations

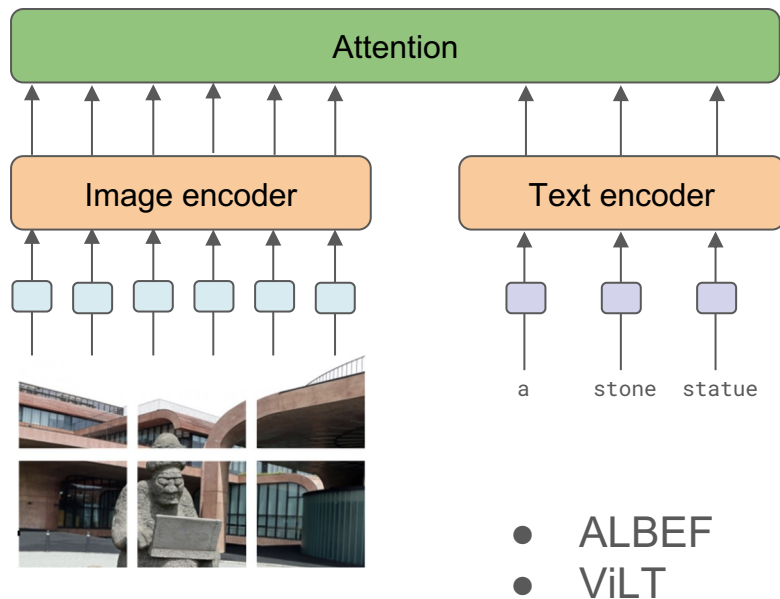
# Background / Motivation

# Problem

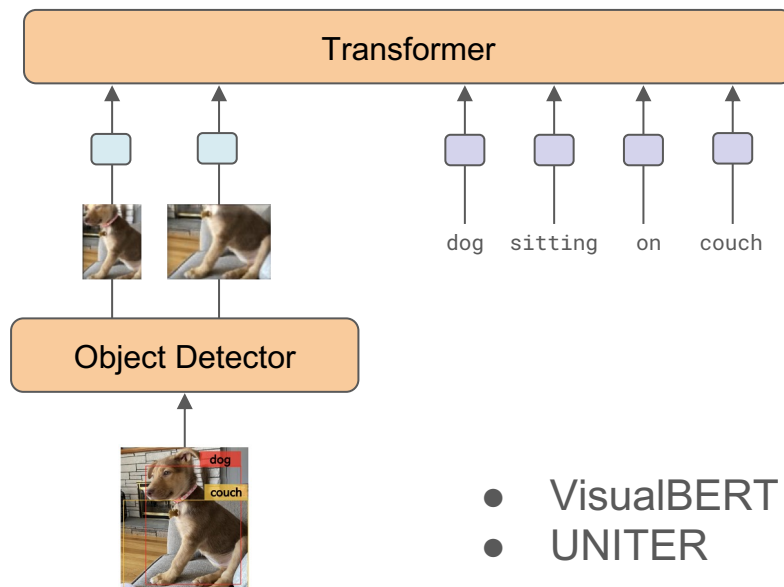
- CLIP is not able to capture fine-grained interactions
  - Uses global features (entire images and sentences)
  - Cannot capture relationship between image patches and textual words
  
- “... CLIP also **struggles** compared to task specific models **on very fine-grained classification**, such as telling the difference between car models, variants of aircraft, or flower species.” - OpenAI

# Previous Works in Learning Fine-Grained Interactions

## Attention-based



## “Region-of-Interest”-based

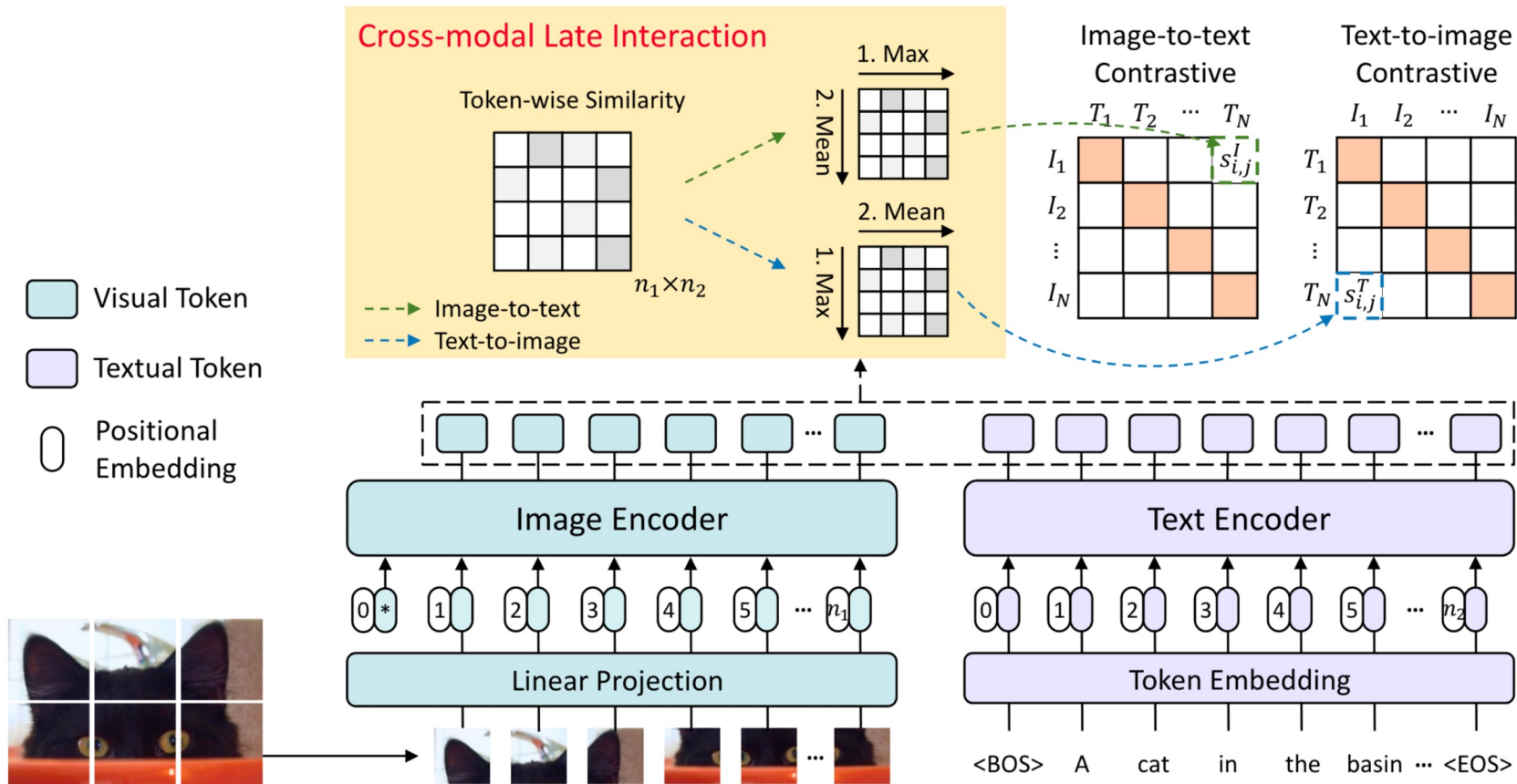


# FILIP Main Contributions

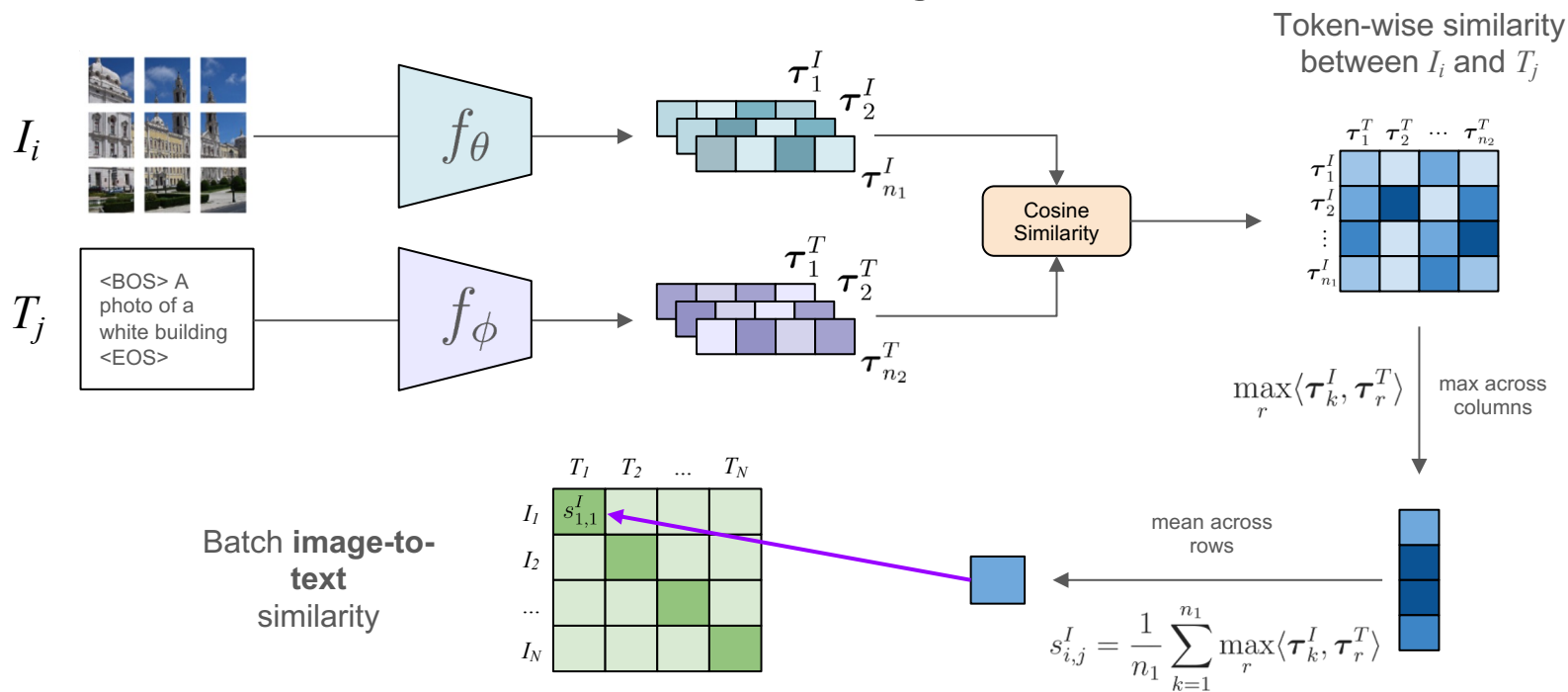
- Overcomes previous issues using token-wise maximum similarity
- Prompt templates for downstream tasks
- Introduces several optimizations to reduce the training time
- Demonstrates improved zero-shot classification and I2T retrieval over CLIP

# Method

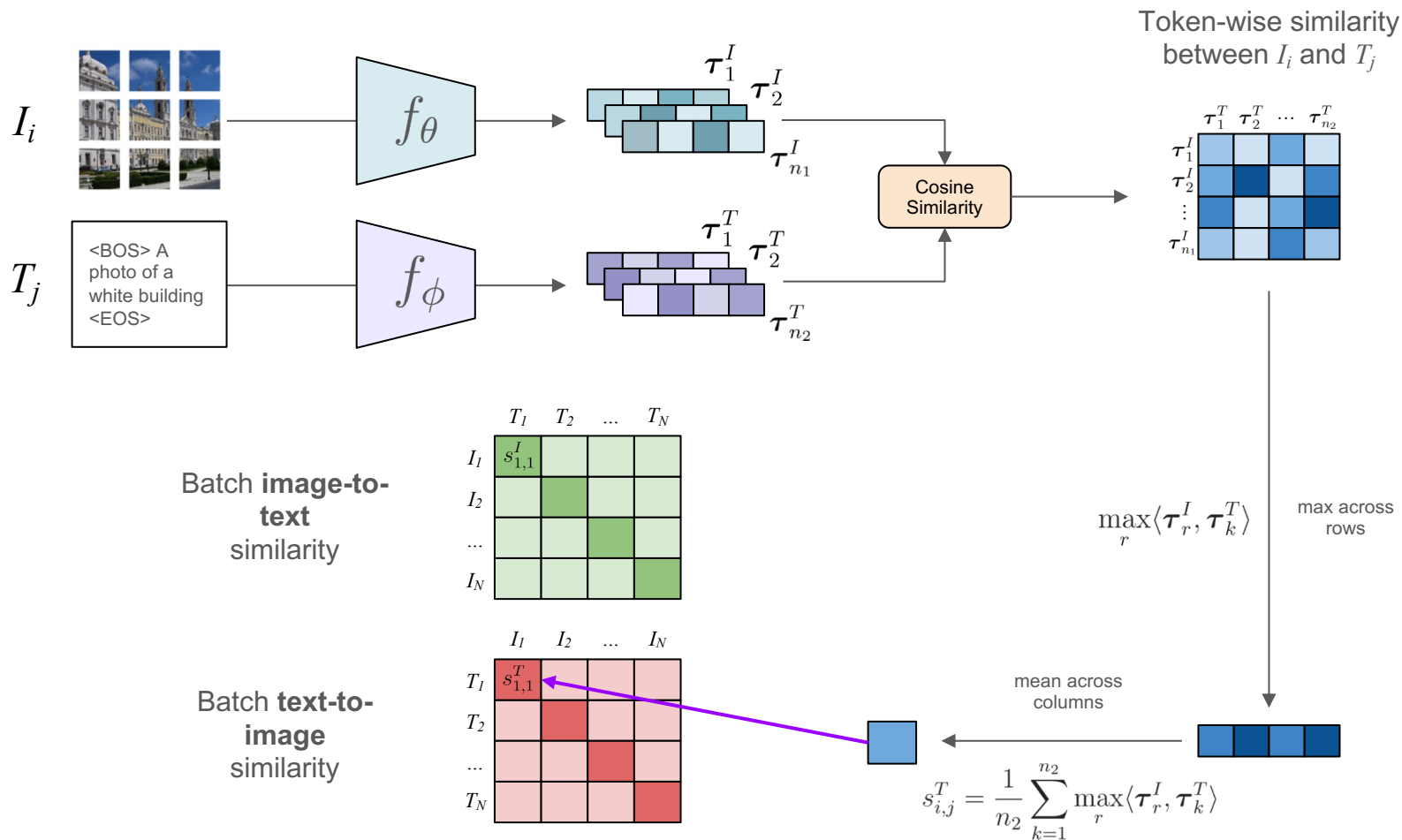
# Overall architecture



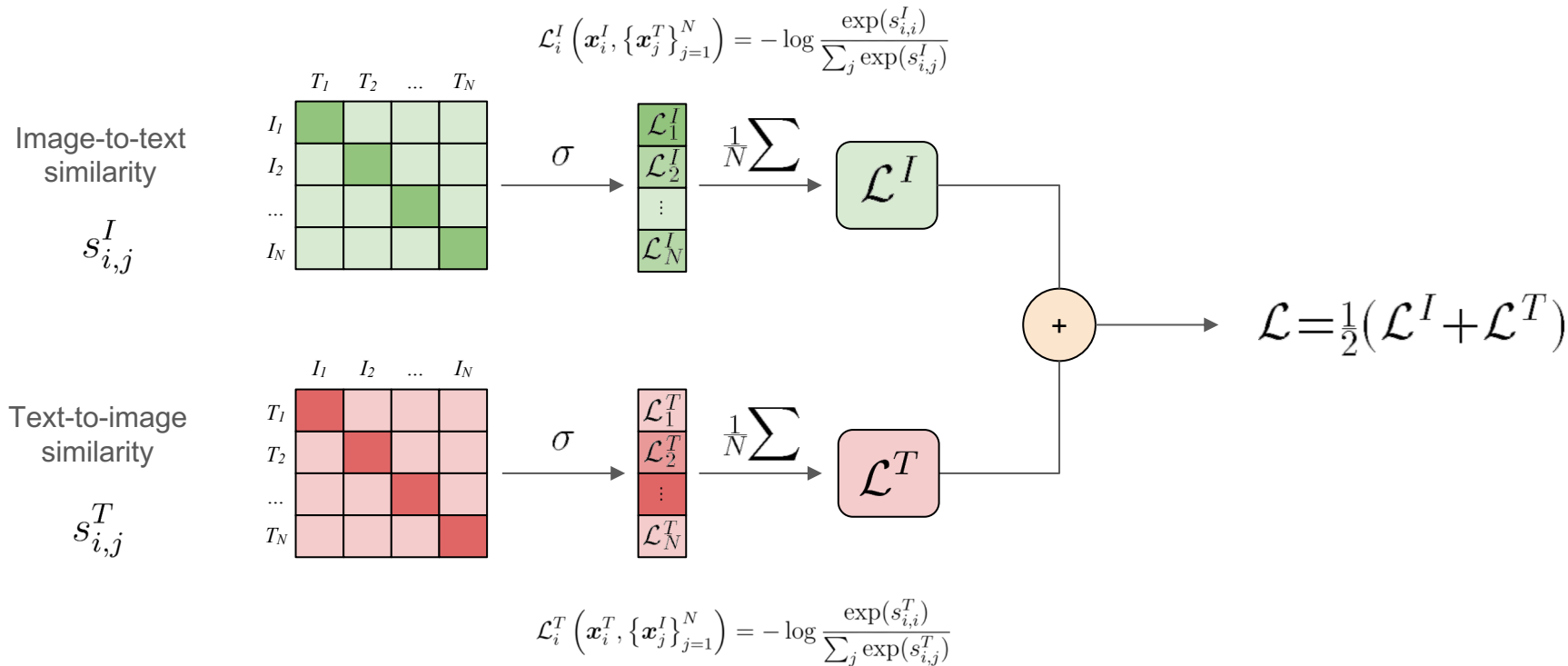
# Cross-modal late interaction: image-to-text



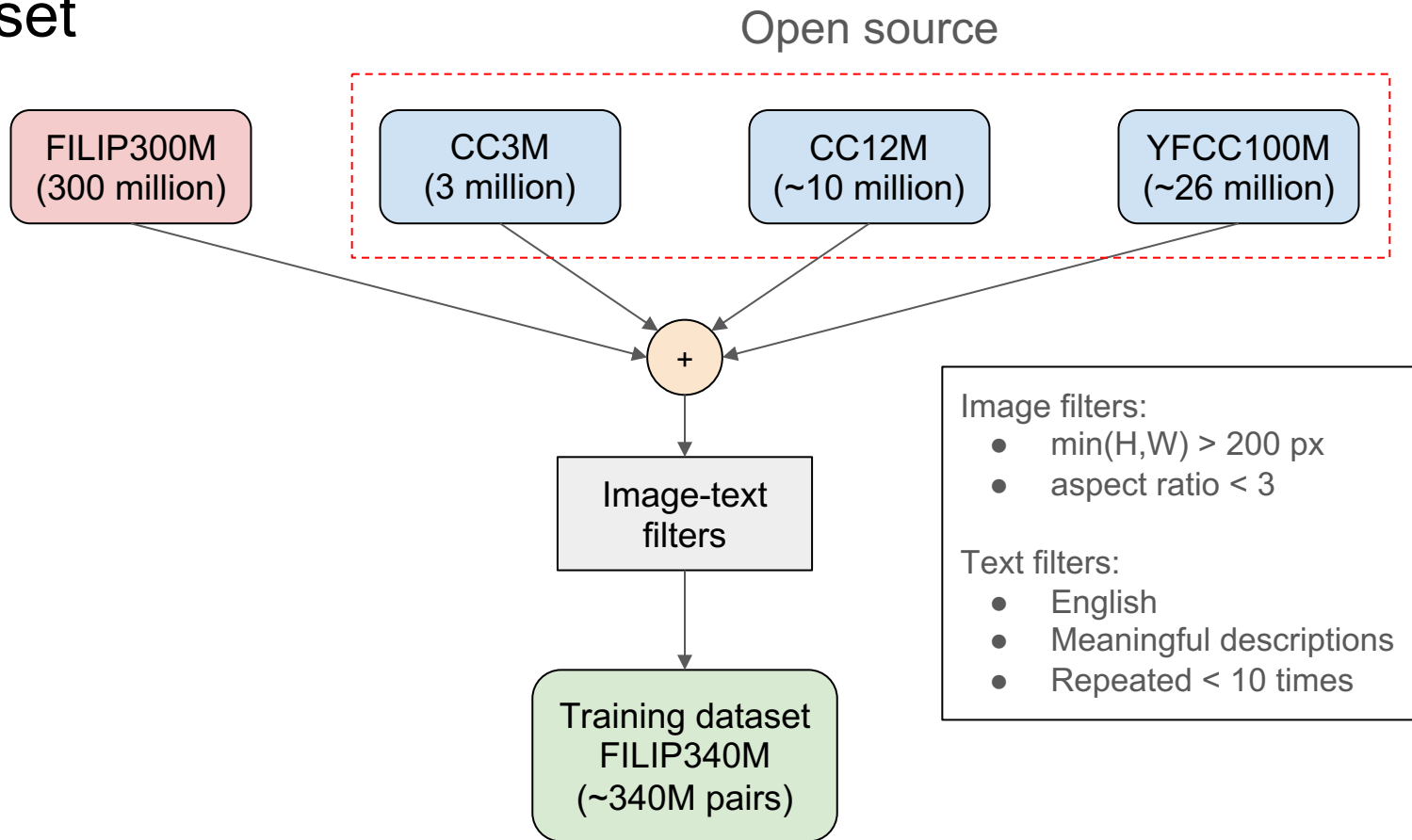
# Cross-modal late interaction: text-to-image



# Loss function: contrastive loss

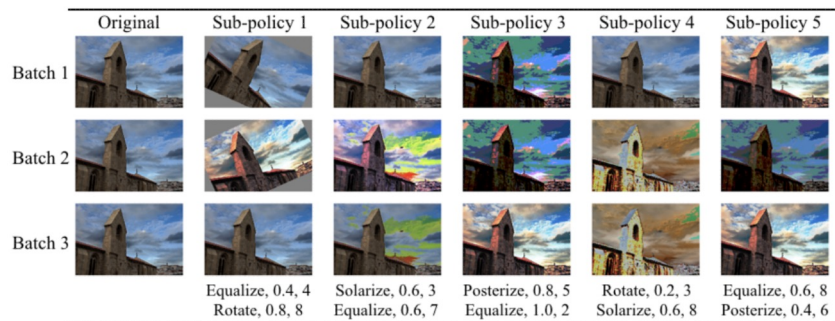


# Dataset



# Data Augmentation Methods for Pre-Training

## Image Augmentation



## Text Augmentation

Original [EN]: This is a photo of a **nice** house.

Translation [RU]: Это фото красивого дома.

Back-translation [EN]: This is a photo of a **beautiful** house.

# Prompt Ensemble

[prefix] {label}, [category description]. [suffix].

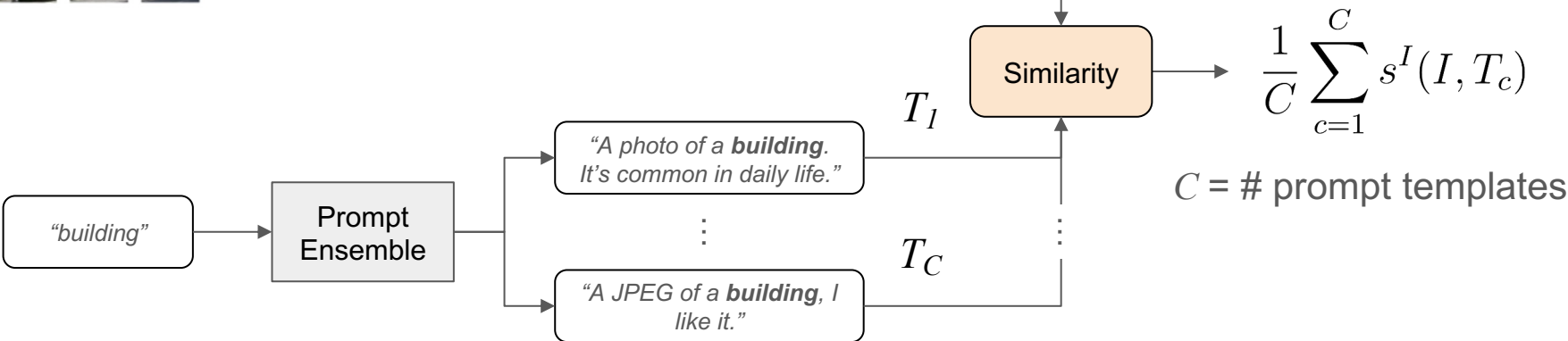
Dataset	Prefix	Category	Suffix
Stanford-Car	“a photo of a”, “a close-up photo of a”, “a good photo of a”, “a bad photo of a”	“a type of car”, “a type of automobile”	“I like it”, “It belongs to my friend”, “It’s brand new”, “It’s popular recently”, “It’s important to me”, “I take it today”

- “A photo of an {Audi 100 Sedan 1994}, a type of car. It’s important to me.”
- Also for CIFAR10, etc.

# FILIP Inference



$I$



# Pre-Training Experimental Setup

Model	Embedding dimension	Input resolution	Image Encoder			Text Encoder		
			#layers	width	#heads	#layers	width	#heads
FILIP <sub>base</sub>	256	224 × 224	12	768	12	12	512	8
FILIP <sub>large</sub>	256	224 × 224	24	1024	16	12	768	12

- FILIP<sub>base</sub> uses ViT-B/32 → 128 Nvidia V100s, 9 days
- FILIP<sub>large</sub> uses ViT-L/14 → 192 Nvidia V100s, 24 days

# Pre-Training Experimental Setup

- Maximum # Text Tokens: 77
- Vocabulary Size: ~49k
- LAMB Optimizer
- Cosine Learning Rate Schedule + Linear Warmup
- Weight Decay → Training Stability

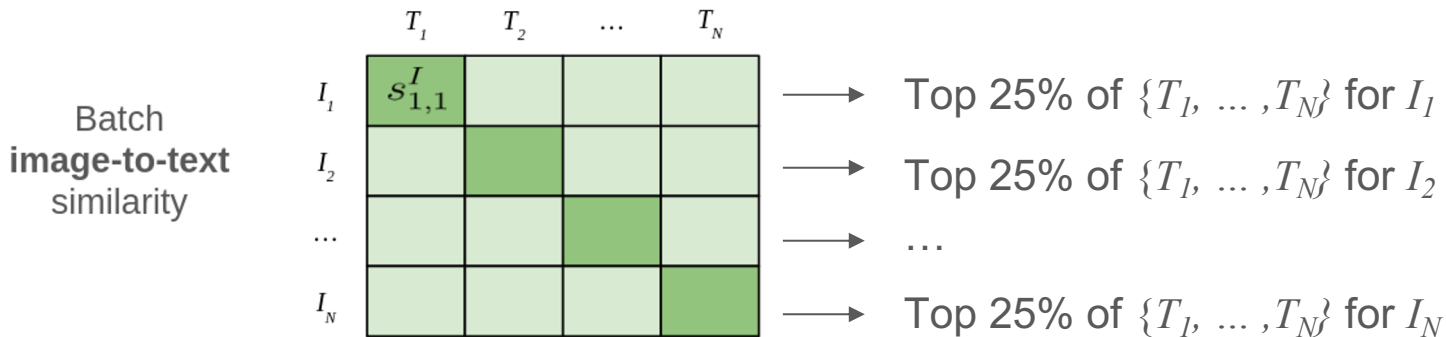
	Model	Dataset	Batch size	Base LR	Weight decay
For ablations	FILIP <sub>base</sub>	YFCC100M	$1024 \times 8$	$6 \times 10^{-3}$	3e-2
	FILIP <sub>base</sub>	FILIP340M	$320 \times 128$	$2 \times 10^{-3}$	3e-3
	FILIP <sub>large</sub>	FILIP340M	$160 \times 192$	$8 \times 10^{-4}$	3e-3

# Pre-Training Efficiency

- FILIP determines similarity between **tokens**
- Reduced embedding size: 512 → 256
- Reduced precision in computing I2T and T2I similarities: fp32 → fp16
  - **fp32**:  $\sim 1.18e-38$  ...  $\sim 3.40e38$  with 6–9 significant decimal digits precision
  - **fp16**:  $\sim 5.96e-8$  ... 65504 with 4 significant decimal digits precision

# Pre-Training Efficiency

- Intuition: each sample can be represented by a few tokens
- Select 25% of tokens with the highest token-wise maximum similarity score
  - For both I2T and T2I similarities



# Effects of FILIP Model Optimizations

Loss	Embed dim	Embed precision	Token %	Training time (sec/iter)	Memory (MB)	ImageNet ZS Top1
orig (baseline)	512	fp32	-	1.31	14300	30.4
late	512	fp32	100%	2.85	26000	34.6
late	512	fp16	100%	2.67	23468	34.5
late	256	fp16	100%	2.31	22382	<b>35.2</b>
late	256	fp16	50%	1.61	16336	34.5
late*	256	fp16	25%	1.39	16100	34.3

- \* denotes final configuration (embed dim/precision, token %)
- 1.39 sec/iter vs 2.31 sec/iter

# I2T Retrieval Fine-Tuning Experimental Setup

Hyperparameter	Value
Image size	$392 \times 392$
Training epochs	3
Optimizer	LAMB
Batch size	5120
Base LR	$2 \times 10^{-4}$
Weight decay	$3 \times 10^{-4}$

- For Flickr30K (30K pairs), MSCOCO (113K pairs)

# Results

# FILIP's Evaluation Setting

- Evaluated with Zero-Shot Image Classification and Image-Text Retrieval
- Across many natural image datasets

Bald eagle  
(5,6)



Bullock cart  
(5,6)



“A photo of a {label}.”

“[BOS] A photo of a bald eagle [EOS]”

0 1 2 3 4 5 6 7

“[BOS] A photo of a bullock cart [EOS]”

0 1 2 3 4 5 6 7

# FILIP vs CLIP Zero-Shot Classification

- Evaluated on 12 downstream classification (augmented) datasets
- FILIP outperforms CLIP in average top-1 accuracy over 12 datasets

	CIFAR10	CIFAR100	Caltech101	StanfordCars	Flowers102	Food101	SUN397	DTD	Aircrafts	OxfordPets	EuroSAT	ImageNet	Average
CLIP-ViT-B/32	91.3	65.1	87.9	59.4	66.7	84.4	63.2	44.5	21.2	87.0	49.4	63.2	65.3
FILIP <sub>base</sub> -ViT-B/32	86.9	65.5	91.9	55.4	85.3	82.8	69.1	49.3	57.2	88.1	49.9	68.8	<b>70.9<sup>+5.6</sup></b>
CLIP-ViT-L/14	96.2	77.9	92.6	77.3	78.7	92.9	67.7	55.3	36.1	93.5	59.9	75.3	75.3
FILIP <sub>large</sub> -ViT-L/14	95.7	75.3	93.0	70.8	90.1	92.2	73.1	60.7	60.2	92	59.2	77.1	<b>78.3<sup>+3.0</sup></b>

# Domain-Specific Dataset Performance

- 30% increase in performance on the FGVC Aircraft dataset



# Image-Text Retrieval Results

- Tested on two retrieval benchmark datasets: Flickr30K and MSCOCO
- FILIP is 2.7% higher than ALIGN, which is trained on a 6x larger dataset

	Flickr30K						MSCOCO					
	image-to-text			text-to-image			image-to-text			text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Unicoder-VL	64.3	85.8	92.3	48.4	76.0	85.2	—	—	—	—	—	—
ImageBERT	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
UNITER	83.6	95.7	97.7	68.7	89.2	93.9	—	—	—	—	—	—
CLIP	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
ALIGN	88.6	98.7	99.7	<b>75.7</b>	<b>93.8</b>	<b>96.8</b>	58.6	83.0	89.7	45.6	69.8	78.6
<b>FILIP</b>	<b>89.8</b>	<b>99.2</b>	<b>99.8</b>	75.0	93.4	96.3	<b>61.3</b>	<b>84.3</b>	<b>90.4</b>	<b>45.9</b>	<b>70.6</b>	<b>79.3</b>
ALBEF*	94.1	99.5	99.7	82.8	96.3	98.1	—	—	—	—	—	—
<b>FILIP*</b>	<b>95.4</b>	<b>99.8</b>	<b>100.0</b>	<b>84.7</b>	<b>97.0</b>	<b>98.7</b>	—	—	—	—	—	—

\* Denotes zero-shot results on Flickr30K after fine-tuning on MSCOCO

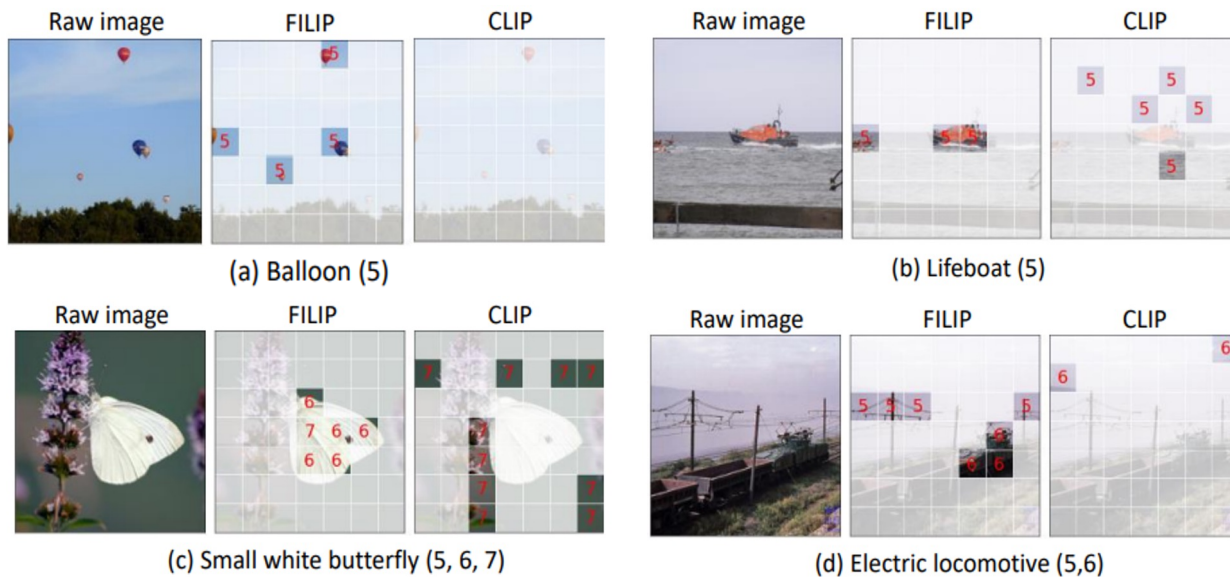
# Image-Text Retrieval Ablations

- R@1 improvement of 5.5 % over vanilla CLIP ViT-B/32
- Effective in both Zero-Shot and fine-tuned Image-Text Retrieval tasks

Model	MSCOCO				ImageNet
	I2T R@1	I2T R@5	T2I R@1	T2I R@5	ZS Top1
Baseline (ViT-B/32)	25.0	49.5	14.7	34.7	30.4
w/ image augmentation	26.1	51.8	16.5	37.5	32.5
w/ back translation	29.2	55.0	17.9	39.8	33.9
w/ cross-modal late interaction	<u>30.5</u>	<u>55.3</u>	<u>18.5</u>	<u>40.0</u>	<u>34.3</u>
Our FILIP <sub>base</sub>	<b>33.4</b>	<b>60.1</b>	<b>23.0</b>	<b>46.2</b>	<b>37.8</b>

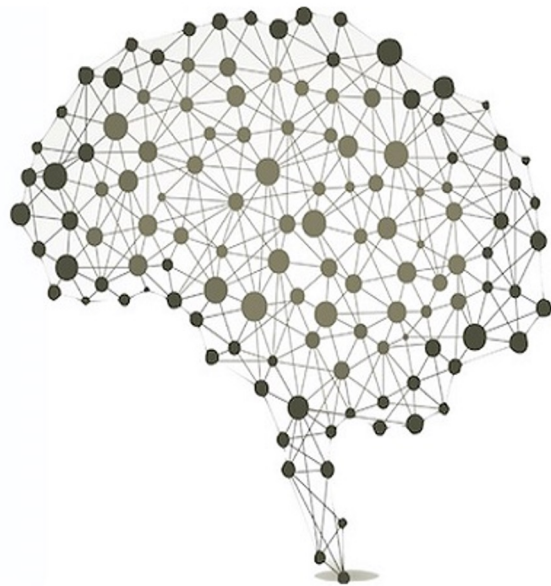
# Word-Patch Alignment Visualizations

- Match images patches with captioned text tokens that have the highest similarity



# Conclusion

- FILIP  $\Rightarrow$  Fine-Grained Vision-Language Pre-Training Model
- Uses token-wise maximum similarity
- SoTA downstream tasks
- Later papers improve on the performance
  - E.g. BLIP-2



# References

- [1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.
- [2] Yao, Lewei, et al. "Filip: Fine-grained interactive language-image pre-training." *arXiv preprint arXiv:2111.07783* (2021).