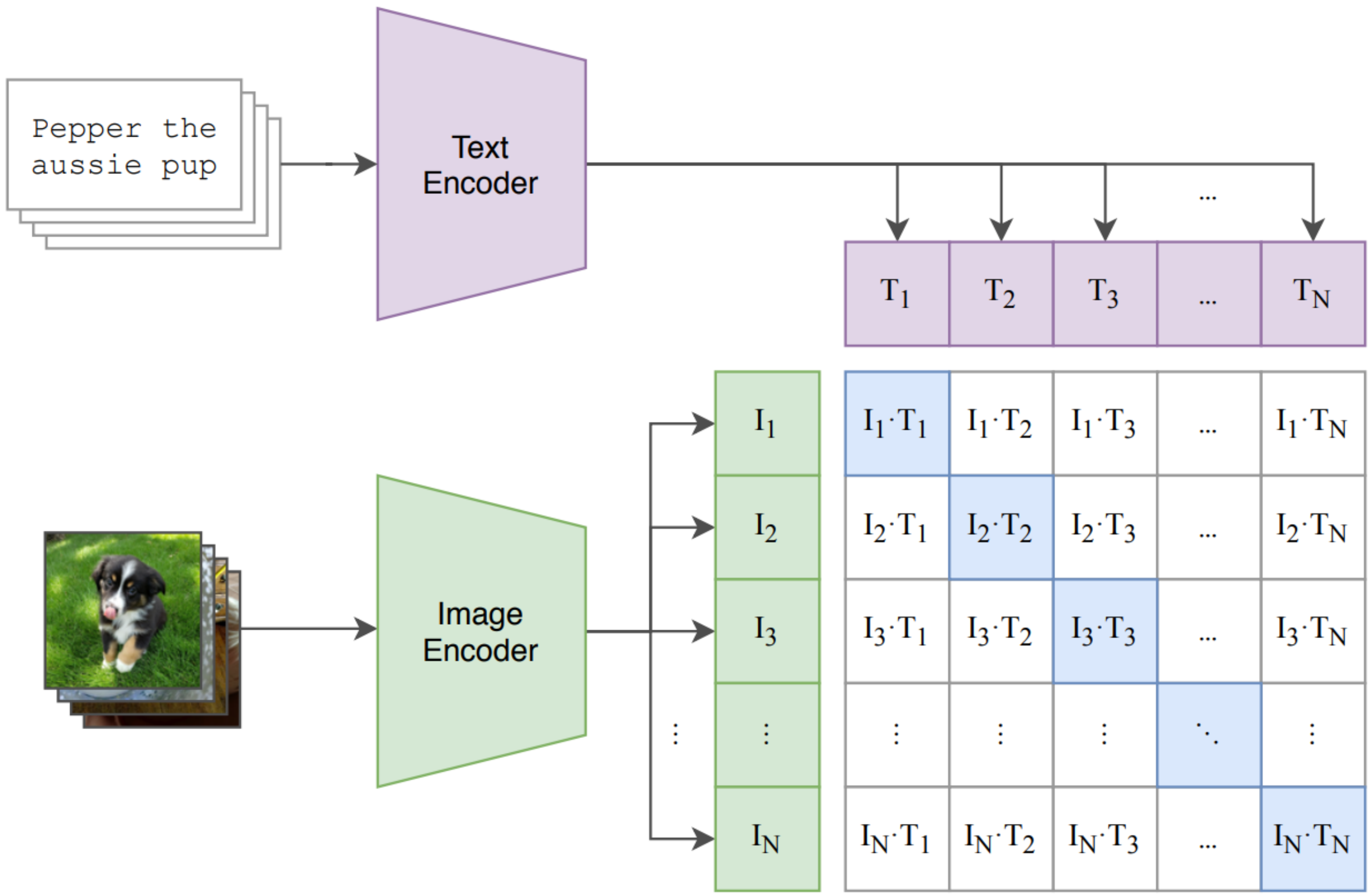
MaskCLIP: Masked Self-Distillation Advances Contrastive Language-Image Pretraining

Xiaoyi Dong*[1], Jianmin Bao*[2], Yinglin Zheng[3], Ting Zhang[2], Dongdong Chen[4], Hao Yang[2], Ming Zeng[3], Weiming Zhang[1], Lu Yuan[4], Dong Chen[2], Fang Wen[2], Nenghai Yu[1]

[1]University of Science and Technology of China [2]Microsoft Research [3]Xiamen University [4]Microsoft Cloud+AI

Contrastive Language-Image Pre-training



- Webley-crawled Image-Text (Annotation free)
- Alignment between Image and text
  - Classical vision/language tasks
  - Zero-shot tasks
  - Text-guided generation

Contrastive Language-Image Pre-training

- Data property difference

Image → Continuous Pixels → Dense Fine-grained semantic → Sparse object semantic

↕ Alignment

Short-Text → Discrete Words → Sparse object semantic → Sparse object semantic

Waste of model capability

Contrastive Language-Image Pre-training

- Data property difference

Image → Continuous Pixels → Dense Fine-grained semantic → Sparse object semantic

↑

Alignment

↓

Short-Text → Discrete Words → Sparse object semantic → Sparse object semantic

Waste of model capability

Image

Text	*"Elementary School Students Studying"*	*"Cosmetic bag or pencil case. Application of fabric and embroidery."*	*"Paramedics wearing hazmat suits responding to an outbreak of Covid-19"*	*"Thumbnail 5 bed property for sale in Ilford"*
Undescribed Object	terrestrial globe, bookshelf	Pen, notebook	Ambulance, cars	Cars

Contrastive Language-Image Pre-training

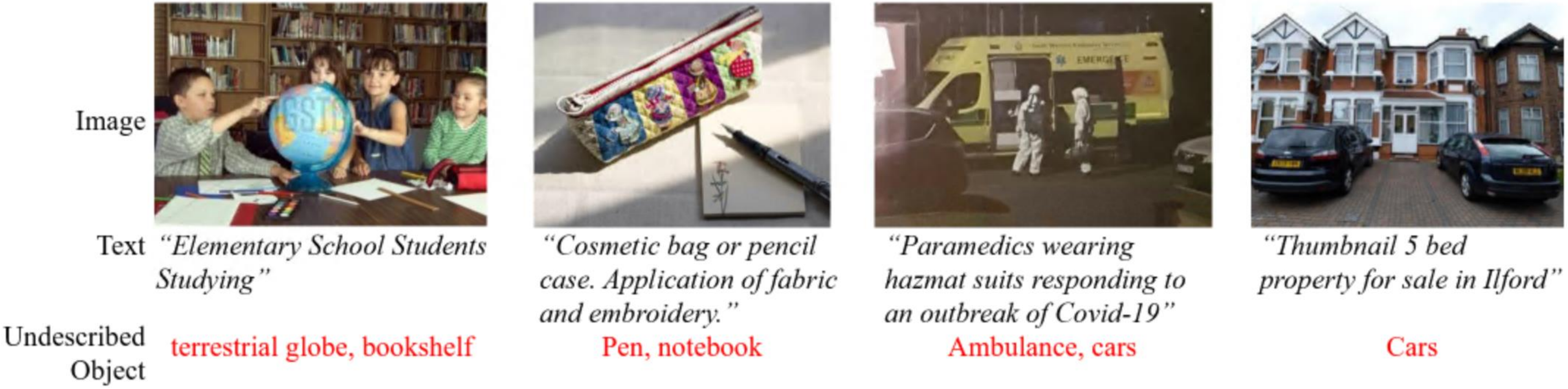
- Data property difference

Image → Continuous Pixels → Dense Fine-grained semantic → Sparse object semantic

Dense Fine-grained semantic → Vision Supervision → Dense Fine-grained semantic

Alignment

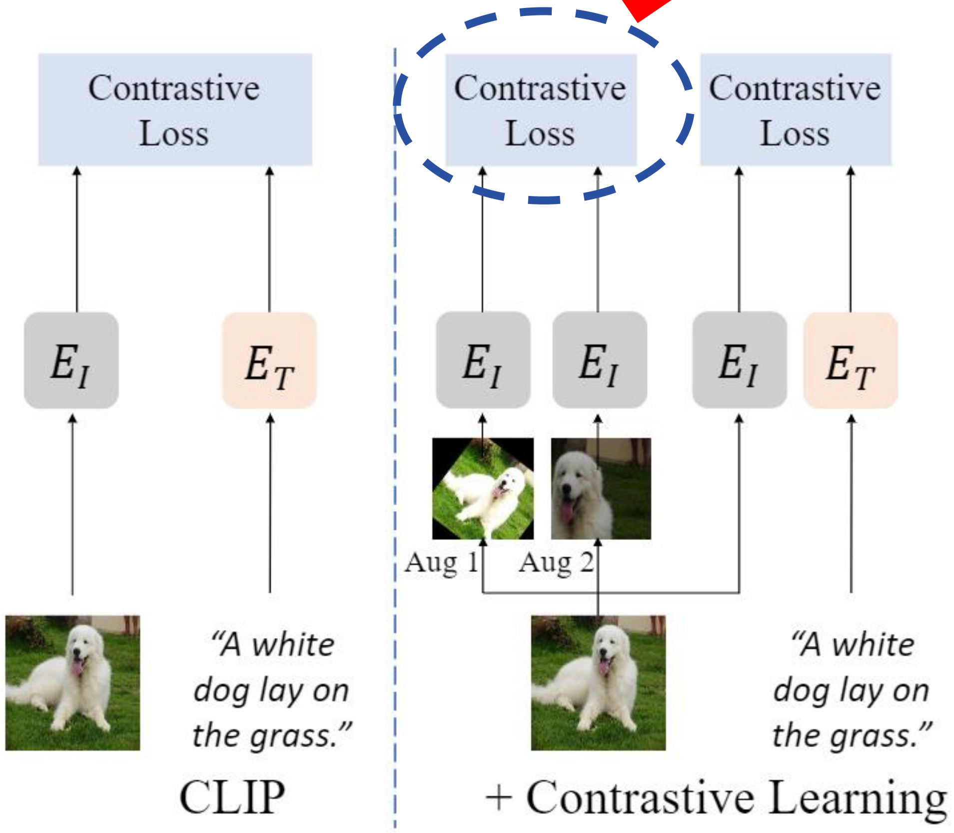
Short-Text → Discrete Words → Sparse object semantic → Sparse object semantic



CLIP + Vision Self-Supervise Learning

- Contrastive Learning

Still Global Supervision



Advantages

· Enhance vision backbone capability

Weaknesses

· Global representation learning

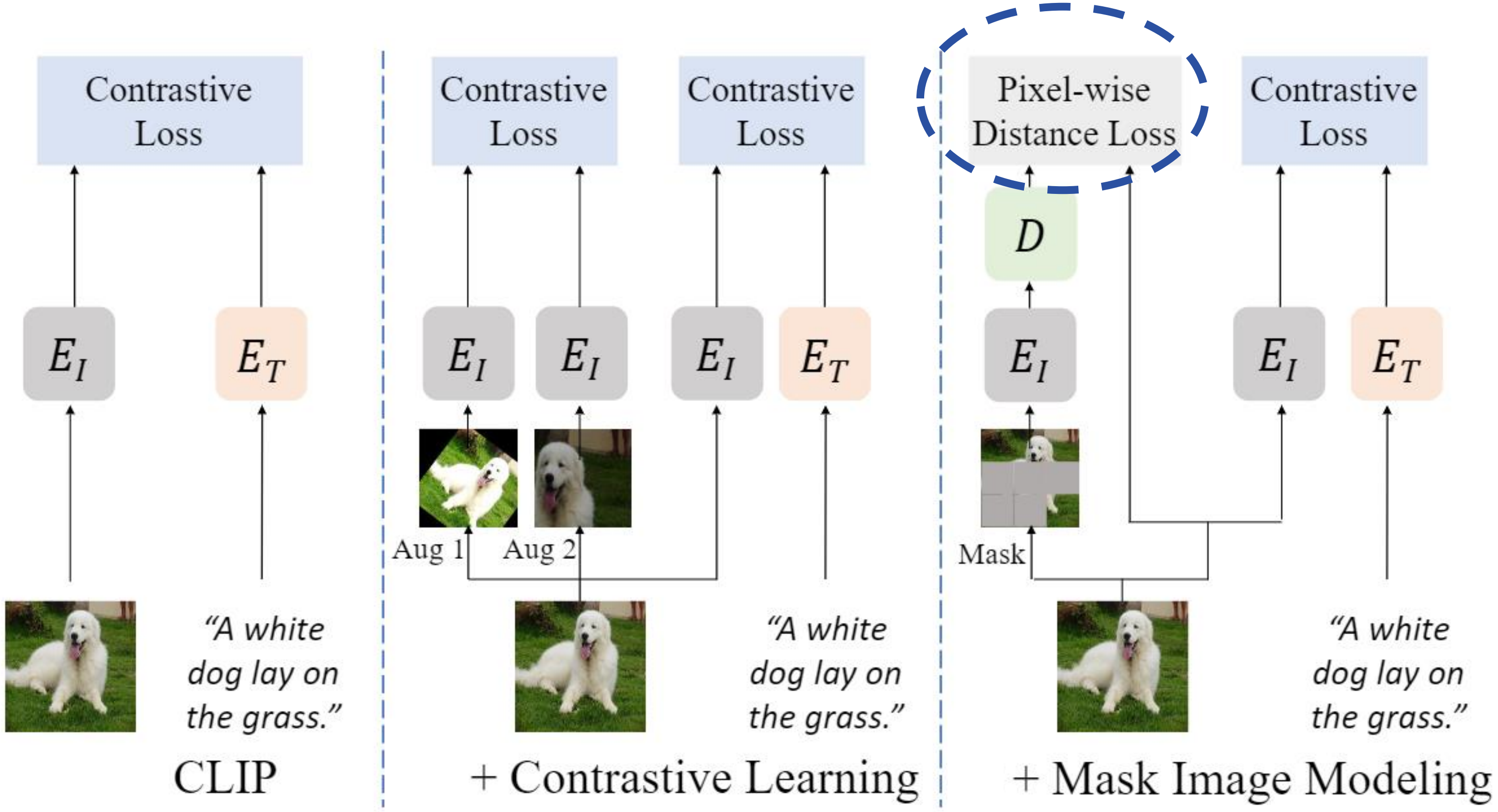
· Huge computation cost

Same as CLIP, still lacks local representation

CLIP + Vision Self-Supervise Learning

- Mask Image Modeling

Low Level Pixel for prediction



Advantages

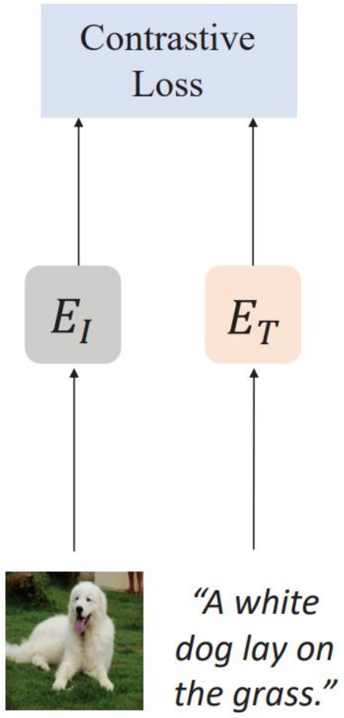
- Local Supervision
- Small computation cost

Weaknesses

- Inefficient pretraining
- Unnecessary target-specific information memorization
- Semantical Conflict with CLIP

CLIP + Vision Self-Supervise Learning

- Two desired properties



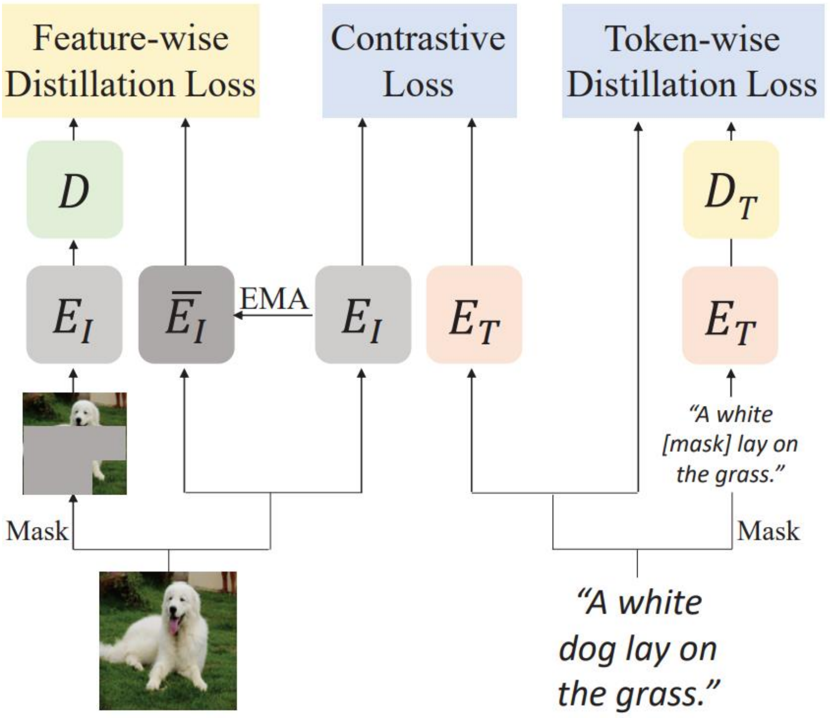
CLIP

Local Supervision

· Fine-grained semantic learning
· Complementary for CLIP global representation

Semantic Output

· Efficient pretraining
· Consist with CLIP output



MaskCLIP

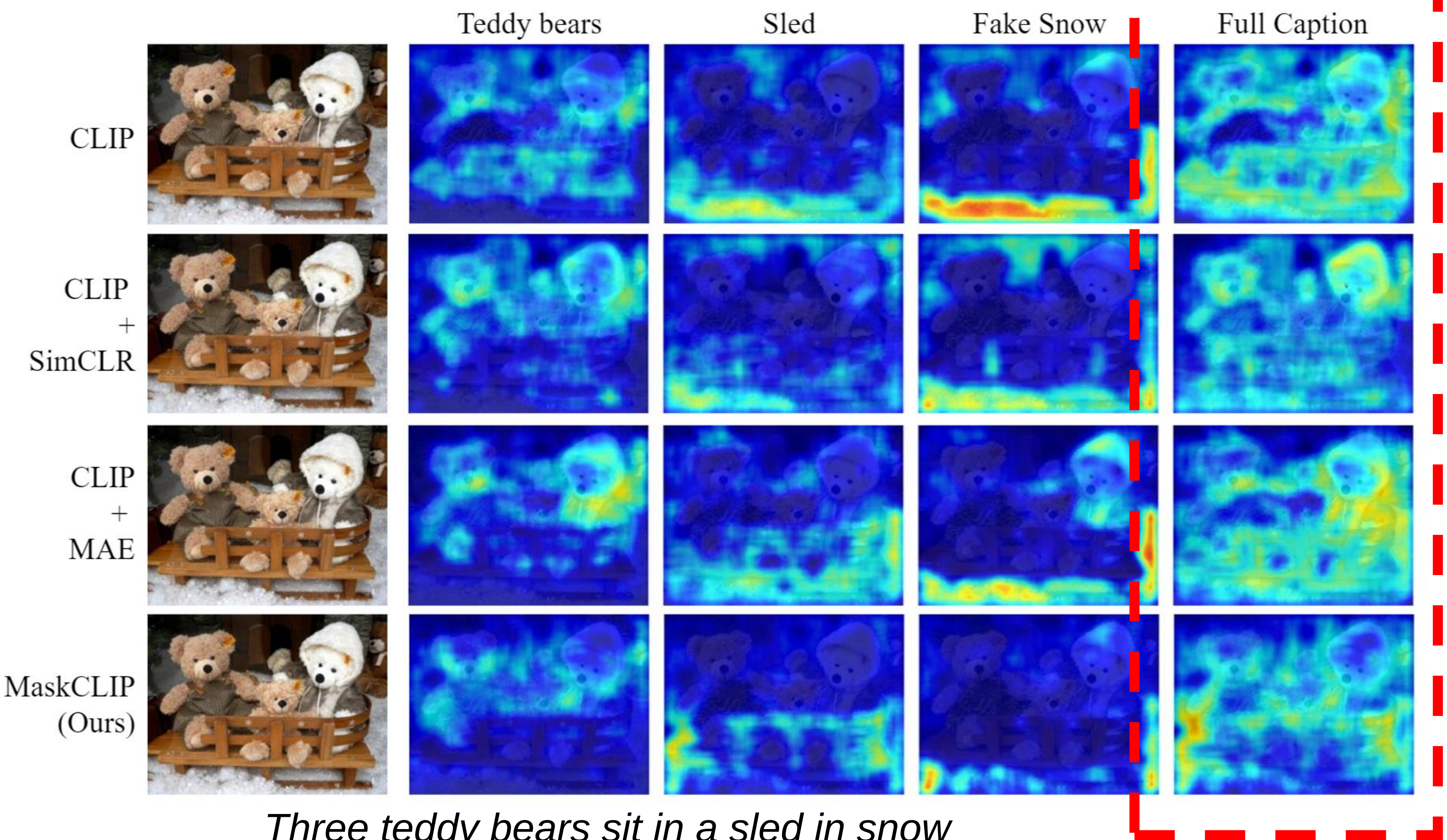
Analysis on MaskCLIP

- Vision self-supervision helps VL contrastive

	Training		IN-1K			Flicker30K	
	Memory	Time	0-shot	Linear	Finetune	I2T	T2I
CLIP	14G	1.00×	37.6	66.5	82.3	52.9	32.8
CLIP+SimCLR	30G	2.67×	42.8	72.1	82.6	58.6	41.3
CLIP+MAE	16G	1.30×	42.1	68.5	83.2	57.3	41.1
MaskCLIP	19G	1.75×	**44.5**	**73.7**	**83.6**	**70.1**	**45.6**

Analysis on MaskCLIP

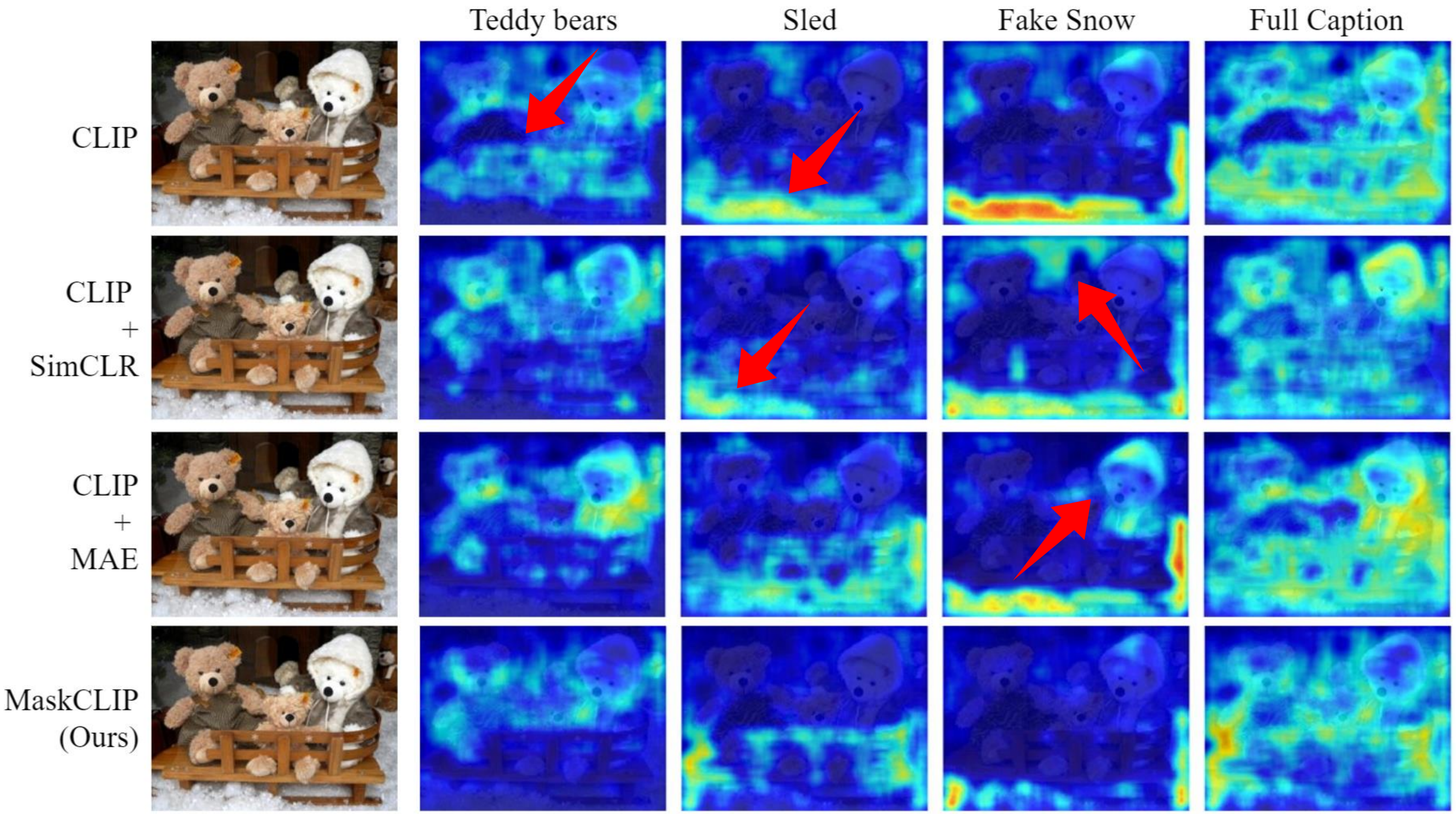
- Masked self-distillation learns semantic representations for local patches.



Three teddy bears sit in a sled in snow

Analysis on MaskCLIP

- Masked self-distillation learns semantic representations for local patches.



Three teddy bears sit in a sled in snow

Experiments

- Vision Tasks

Method	Epoch	IN-1K 0-Shot	IN-1K Lin	IN-1K FT	ADE20K mIoU	MS-COCO AP^b	MS-COCO AP^m
DeiT [59]	300*	–	–	81.8	47.4	44.1	39.8
SimCLR [9]	25	–	64.0	82.5	48.0	44.6	40.2
MAE [26]	25	–	56.2	82.5	46.5	43.2	39.1
CLIP [51]	25	37.6	66.5	82.3	47.8	43.6	39.5
SLIP [49]	25	42.8	72.1	82.6	48.5	44.0	40.3
MaskCLIP	25	**44.5**	**73.7**	**83.6**	**50.5**	**45.4**	**40.9**

+6.9% +1.3% +2.8 mIoU

Experiments

- Zero-shot classification on ICinW challenge

	Average	Caltech-101	CIFAR-10	CIFAR-100	Country211	DTD	EuroSAT	FER-2013	Aircraft	Food-101	GTSRB	Memes	KittiDis	MNIST	Flowers	Pets	PatchCam	SST2	RESISC45	Cars	Voc2007
Pretraining on YFCC-15M																					
CLIP	34.0	58.6	68.5	36.9	10.8	21.4	30.5	16.9	5.1	51.6	6.5	51.1	25.9	5.0	52.7	28.6	51.7	**52.5**	22.4	4.5	79.1
SLIP	37.8	70.9	**82.6**	48.6	11.8	26.6	19.8	18.1	5.6	59.9	**12.6**	51.8	29.4	**9.8**	56.3	31.4	**55.3**	51.5	28.5	5.4	80.5
MaskCLIP	**40.1**	**72.0**	80.2	**57.5**	**12.6**	**27.9**	**44.0**	**20.3**	**6.1**	**64.9**	8.5	**52.0**	**34.3**	4.9	**57.0**	**34.3**	50.1	49.9	**35.7**	**6.7**	**82.1**
Pretraining on ICinW Academic Track Stting: YFCC-15M , GCC3M+12M, ImageNet-21K(ImageNet-1K is removed)																					
1st MaskCLIP	**48.9**	86.4	**95.3**	**78.3**	11.6	**33.0**	**57.7**	18.8	**8.0**	**78.9**	17.3	**52.8**	16.0	7.3	74.2	**74.4**	**52.1**	46.2	**54.3**	**26.5**	**82.3**
2nd KLITE*	45.5	**87.4**	92.7	68.8	8.2	32.2	27.9	17.4	4.3	72.4	11.4	48.4	**31.1**	12.8	**75.6**	65.9	50.6	**52.9**	44.4	10.2	**82.3**
3rd YT-CLIP	44.5	77.8	83.5	58.4	**11.9**	31.9	40.7	27.1	6.9	68.7	**18.8**	52.3	9.1	**18.8**	53.1	69.3	51.5	50.3	52.7	19.7	79.3
4th UniCL†	44.0	84.8	90.2	67.8	6.7	25.4	35.3	**30.8**	3.5	68.3	11.1	51.0	17.9	11.3	71.7	44.9	52.1	49.5	41.4	24.2	81.3
5th Gramer*	43.2	83.9	92.9	69.5	7.3	25.5	24.4	30.4	2.7	71.0	9.0	52.6	12.4	10.1	70.4	52.4	50.6	50.1	44.8	13.8	81.3

- Zero-shot image-text retrieval

	Training Epoch	Flickr30K						MS-COCO					
		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
CLIP [51]	25	52.9	79.6	87.2	32.8	60.8	71.2	27.5	53.5	65.0	17.7	38.8	50.5
SLIP [49]	25	58.6	85.1	91.7	41.3	68.7	78.6	33.4	59.8	70.6	21.5	44.4	56.3
MaskCLIP	25	**70.1**	**90.3**	**95.3**	**45.6**	**73.4**	**82.1**	**41.4**	**67.9**	**77.5**	**25.5**	**49.7**	**61.3**

Thanks