NEURAL INFORMATION PROCESSING SYSTEMS

🔄 un²CLIP: Improving CLIP's Visual Detail Capturing Ability via Inverting unCLIP

Yinqi Li[1,2], Jiahe Zhao[1,2], Hong Chang[1,2], Ruibing Hou[1], Shiguang Shan[1,2], Xilin Chen[1,2]

[1]Institute of Computing Technology, Chinese Academy of Sciences, China

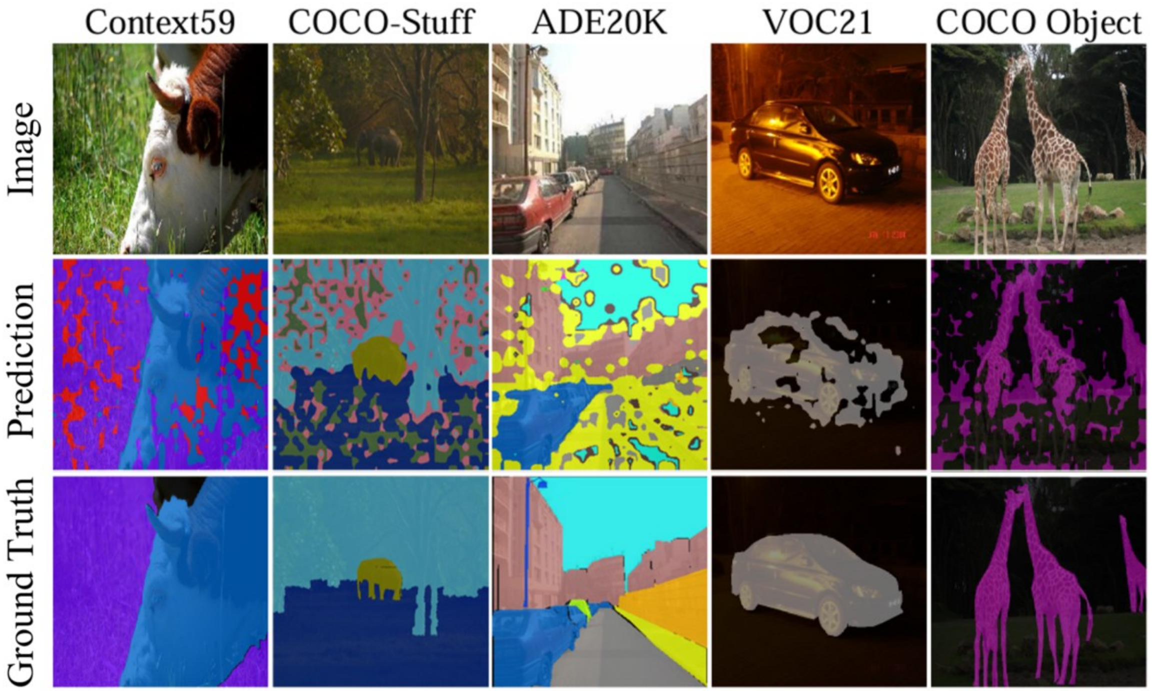
[2]University of Chinese Academy of Sciences, China

Background

- **CLIP falls short in capturing visual details**

(Tong et al., 2024)

Multimodal Understanding



Result of ClearCLIP (Lan et al., 2024)

Open-vocabulary Segmentation

High-Level Idea

- **Refining Existing CLIP Models with Image-Only Data**
  - Challenging to acquire high-quality data (e.g., region-text pairs)
  - Re-training CLIP models is costly

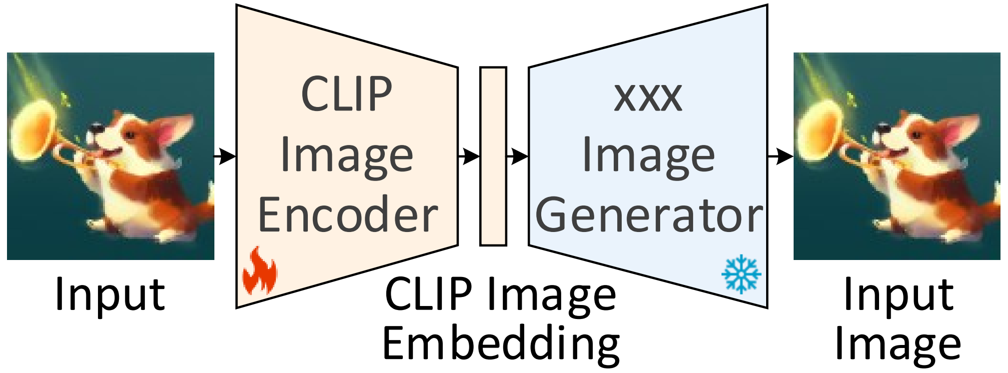
High-Level Idea

- **Refining Existing CLIP Models with Image-Only Data**
  - Challenging to acquire high-quality data (e.g., region-text pairs)
  - Re-training CLIP models is costly
- **Harnessing the Capabilities of Generative Models**
  - Trained to learn the full image data distribution
  - Capture fine-grained visual details better than discriminative models (e.g., CLIP)

High-Level Idea

- Refining Existing CLIP Models with Image-Only Data
- Harnessing the Capabilities of Generative Models

Preliminary Framework

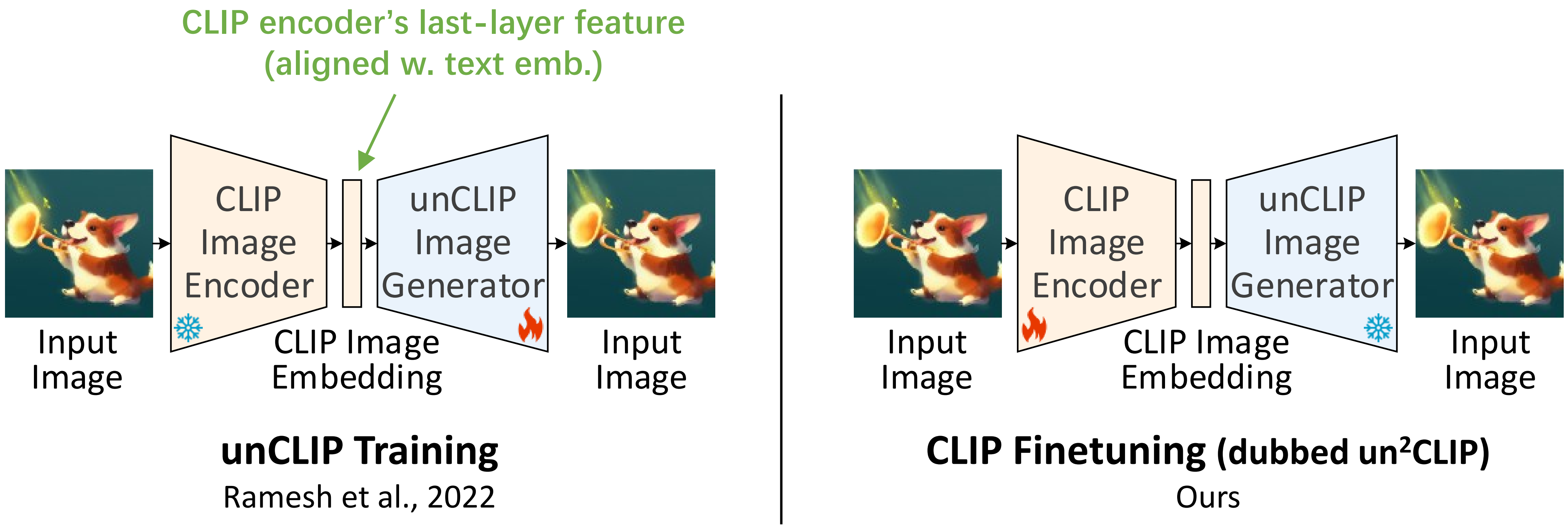


CLIP Finetuning

- General solutions may break the image-text alignment property of CLIP

Method

- Utilize the unCLIP Generator (Ramesh et al., 2022) as the "Decoder" Module
- Freeze the Generator During CLIP Finetuning

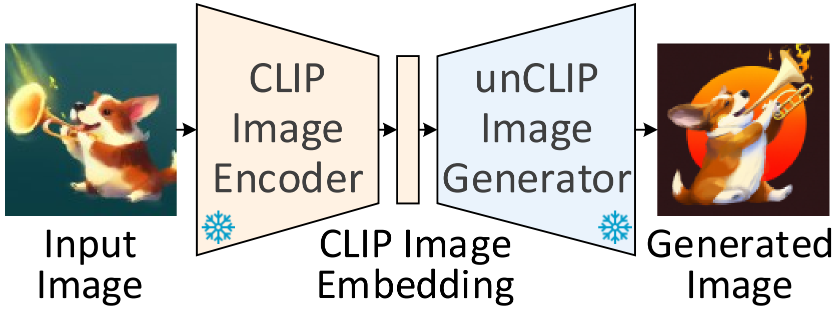


Experiments

- Qualitative Results
- CLIP-Blind Pair (MMVP-VLM) Evaluation
- Dense Vision-Language Inference Evaluation
- Multimodal Large Language Model Evaluation
- Zero-Shot Classification and Retrieval

Experiments

- **Qualitative Results**



unCLIP Sampling

Experiments

- **CLIP-Blind Pair (MMVP-VLM) Evaluation**

Table 1: **MMVP-VLM benchmark evaluation.** The benchmark contains 9 visual patterns that original CLIP models often misinterpret: : Orientation and Direction, : Presence of Specific Features, : State and Condition, : Quantity and Count, : Positional and Relational Context, : Color and Appearance, : Structural and Physical Characteristics, : Texts, : Viewpoint and Perspective. † denotes our reproduced results using official codes correspondingly.

CLIP Model	Resol.	#Params	Method										Avg
OpenAI ViT-L-14	224^2	427.6M	Original	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
			DIVA	13.3	20.0	40.0	6.7	20.0	53.3	46.7	20.0	13.3	25.9
			GenHancer	13.3	33.3	33.3	20.0	6.7	73.3	46.7	20.0	40.0	31.9
			un²CLIP	0.0	33.3	46.7	26.7	13.3	80.0	40.0	20.0	33.3	**32.6**
OpenAI ViT-L-14	336^2	427.9M	Original	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
			DIVA	26.7	20.0	33.3	13.3	13.3	46.7	26.7	6.7	40.0	25.2
			GenHancer	6.7	20.0	33.3	20.0	6.7	73.3	53.3	26.7	26.7	29.6
			un²CLIP	6.7	33.3	46.7	13.3	13.3	80.0	40.0	20.0	20.0	**30.4**
OpenCLIP ViT-H-14	224^2	986.1M	Original	6.7	13.3	53.3	26.7	6.7	73.3	40.0	13.3	26.7	28.9
			DIVA†	13.3	13.3	53.3	26.7	6.7	73.3	46.7	13.3	26.7	30.4
			GenHancer†	13.3	6.7	46.7	20.0	33.3	80.0	26.7	40.0	33.3	33.3
			un²CLIP	26.7	13.3	53.3	20.0	33.3	86.7	46.7	13.3	33.3	**36.3**
SigLIP ViT-SO-14	384^2	878.0M	Original	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
			DIVA	26.7	33.3	53.3	26.7	13.3	80.0	40.0	26.7	46.7	38.5
			GenHancer	26.7	20.0	66.7	33.3	13.3	86.7	40.0	26.7	46.7	40.0
			un²CLIP	20.0	20.0	60.0	46.7	26.7	73.3	40.0	26.7	60.0	**41.5**

Experiments

- **Dense Vision-Language Inference Evaluation**

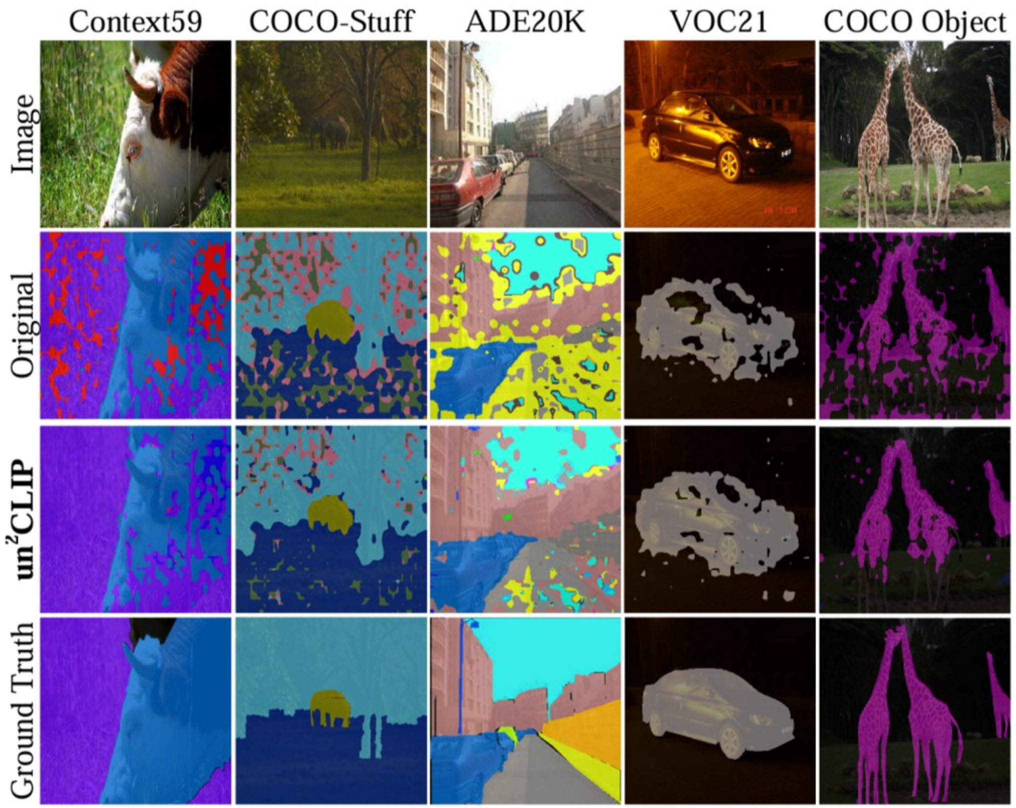


Table 2: **Open-vocabulary semantic segmentation quantitative comparison.** Results of DIVA and GenHancer are obtained using official checkpoints. The CLIP backbone is OpenAI ViT-L-14@336.

Segmentation Method	CLIP-Improve. Method	Without background class					With a background class			Average
		VOC20	Ctx59	Stuff	City	ADE	VOC21	Ctx60	Object	
CLIP	Original	11.7	3.4	1.7	2.5	0.9	7.7	2.9	3.3	4.3
	DIVA	12.0	3.4	1.7	2.5	1.0	7.7	2.9	3.3	4.3
	GenHancer	8.4	2.9	1.3	2.7	0.7	4.6	2.5	1.7	3.1
	un^{2}CLIP	**17.3**	**5.1**	**2.6**	**3.8**	**1.3**	**9.3**	**4.3**	**4.3**	**6.0**
MaskCLIP	Original	24.7	10.1	7.3	10.3	6.1	21.8	9.2	12.1	12.7
	DIVA	25.7	10.4	7.6	10.4	6.3	22.4	9.5	12.6	13.1
	GenHancer	13.5	6.4	3.4	9.2	3.7	12.3	5.9	4.9	7.4
	un^{2}CLIP	**30.0**	**12.9**	**8.9**	**13.1**	**7.5**	**25.2**	**11.6**	**13.5**	**15.3**
SCLIP	Original	37.3	12.7	8.5	10.2	4.6	28.7	11.9	14.9	16.1
	DIVA	37.7	12.8	8.5	10.3	4.6	28.9	11.9	15.0	16.2
	GenHancer	21.0	7.7	3.6	6.8	2.2	15.1	7.0	5.3	8.6
	un^{2}CLIP	**53.8**	**19.5**	**12.0**	**16.1**	**6.9**	**38.6**	**17.9**	**19.3**	**23.0**
ClearCLIP	Original	72.4	26.0	18.1	22.8	14.2	42.6	23.2	27.1	30.8
	DIVA	72.3	25.9	18.1	22.7	14.0	42.6	23.2	27.1	30.7
	GenHancer	52.1	22.9	11.8	17.1	10.3	24.2	20.0	10.2	21.1
	un^{2}CLIP	**76.5**	**30.5**	**20.6**	**26.4**	**16.0**	**47.6**	**27.3**	**29.6**	**34.3**

Experiments

- **Multimodal Large Language Model Evaluation**

Table 3: **MLLM benchmark evaluation.** Best and second best results are highlighted in **bold** and underline. Results on NaturalBench follow the official evaluation protocol [50], which differs from that in GenHancer [43], resulting in some missing entries. Baseline numbers are taken from [43].

LLM	CLIP	Vision-centric Benchmarks								General Benchmarks				
		MMVP [9]	NaturalBench [50]				CV-Bench 2D [12]		CV-Bench 3D [12]	POPE [51]			SciQA-IMG[52]	Hallusion Avg. [53]
			Acc	Q-Acc	I-Acc	G-Acc	ADE20K	COCO		rand	pop	adv		
Vicuna-7B	Original	24.7	67.3	37.7	43.8	12.7	49.6	60.9	58.7	87.3	86.1	84.2	66.8	27.6
	DIVA	**31.3**	-	-	-	-	51.3	63.4	60.2	87.9	87.0	84.6	66.3	**28.6**
	GenHancer	30.7	-	-	-	-	52.9	63.6	**63.2**	**88.1**	86.7	84.6	66.5	28.4
	un²CLIP	**31.3**	**68.7**	**40.0**	**45.9**	**15.1**	**53.9**	**65.1**	61.2	88.0	**87.4**	**85.4**	**68.4**	28.4

Question:

How many clocks are in the image?

(A) 2 (B) 1

(C) 3 (D) 0

Original (A) 2

un²CLIP (B) 1

Question:

Considering the relative positions of the bowl (annotated by the red box) and the broccoli in the image provided, where is the bowl locat-ed with respect to the broccoli?

(A) Left (B) Right

Original (B) Right

un²CLIP (A) Left

Experiments

- **Zero-Shot Classification and Retrieval**
  - Classification tasks generally favor representations that emphasize dominant foreground semantics
  - Contrast with the main objective of our work, which is to enhance CLIP's ability to capture visual details as much as possible

Method	Zero-shot Image Classification							Image-to-Text Retrieval@5		Text-to-Image Retrieval@5	
	IN-1K	C-10	C-100	Cal-101	SUN397	Aircraft	Cars	Flickr30K	COCO	Flickr30K	COCO
Original	**75.5**	**95.6**	75.9	86.7	**67.6**	**31.7**	77.9	**97.3**	79.4	87.3	61.0
DIVA	**75.5**	95.5	**76.3**	**87.1**	67.5	31.6	**78.0**	**97.3**	**79.7**	86.9	61.0
GenHancer	40.2	77.5	44.2	79.3	42.4	7.2	21.0	87.2	61.7	81.6	51.0
un^{2}CLIP	62.4	89.0	65.6	86.8	59.2	22.0	63.3	96.4	77.6	**90.1**	**65.5**

Summary

- **Finding:** unCLIP provides a suitable framework for improving CLIP
- **Proposed method:** un^{2}CLIP - finetunes CLIP image encoder via inverting unCLIP
- **Experiments:** Consistent improvements across CLIP-blind pair, dense vision-language inference, and MLLM evaluations

GitHub HuggingFace OpenReview