Improving CLIP Training with Language Rewrites

Lijie Fan[1,2,*] Dilip Krishnan[1] Phillip Isola[2] Dina Katabi[2] Yonglong Tian[1,*]

[1]Google Research, [2]MIT CSAIL, [*]equal contribution

NeurIPS 2023

HUMANE Lab 박현빈

25.05.09

Background

- In CLIP Training, data augmentations are applied to image inputs to prevent overfitting
- However, text inputs remain unchanged
- Such asymmetry presents two issues
  - the language aspect provides less guidance to the image encoders
  - the text encoders repeatedly encounter the exact same texts in each epoch, which increases the risk of text overfitting
- Previous approaches focus on word-level treatments like replacement or masking, but these have less impact compared to image augmentations

Preliminary

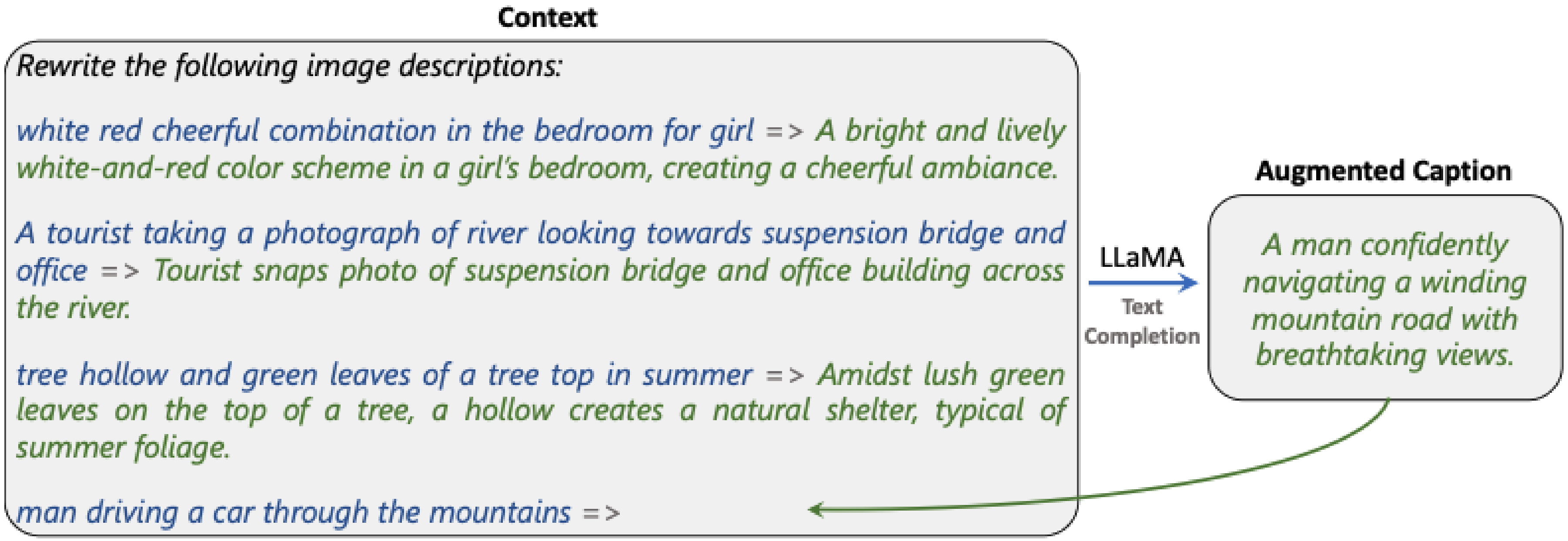
$$L_I = -\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(f_I(\text{aug}_I(x_I^i)), f_T(x_T^i))/\tau\right)}{\sum_{k=1}^{N} \exp\left(\text{sim}(f_I(\text{aug}_I(x_I^i)), f_T(x_T^k))/\tau\right)}$$

- f: normalization functions
- L_I: image-to-text loss
- L_T: text-to-image loss
- $L = (L_T + L_I)/2$

$$L_I = -\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(f_I(\text{aug}_I(x_I^i)), f_T(\text{aug}_T(x_T^i)))/\tau\right)}{\sum_{k=1}^{N} \exp\left(\text{sim}(f_I(\text{aug}_I(x_I^i)), f_T(\text{aug}_T(x_T^k)))/\tau\right)}$$

Meta-Input-Output Text Pair Generation

- To harness ICL for text rewriting, generate several rewriting examples to be included in the prompt



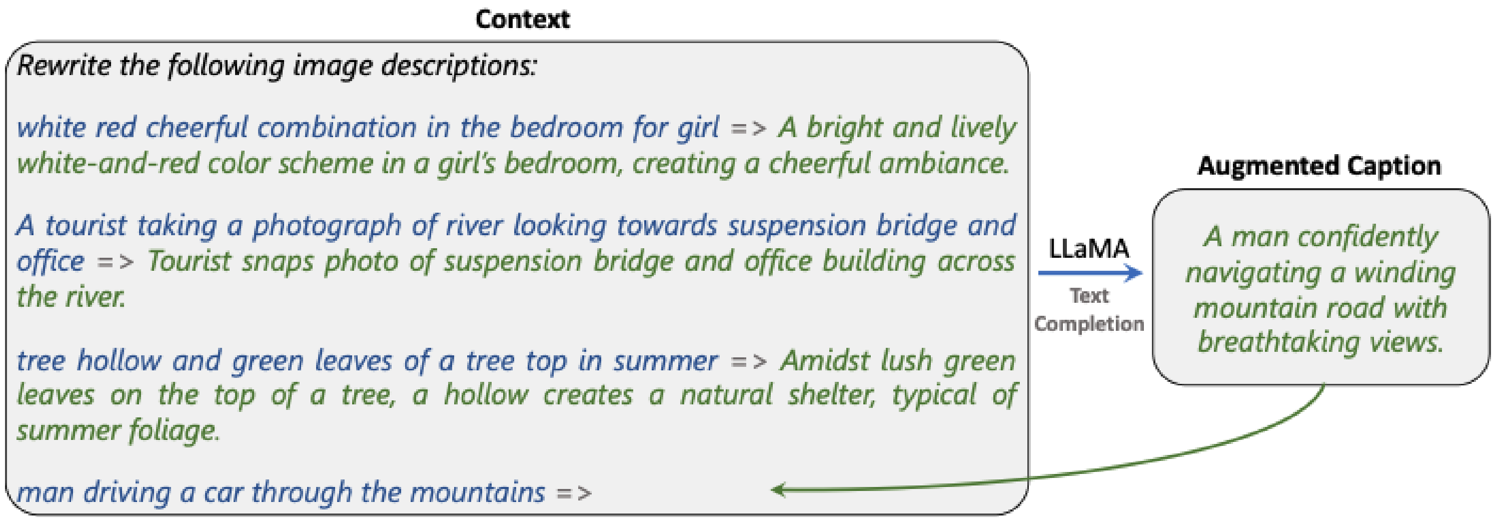
Meta-Input-Output Text Pair Generation

- Rewriting with Chatbots
  - randomly sample texts from image-text datasets, and prompt ChatGPT and Bard web portals
  - prompt : "Rewrite this caption of an image vividly, and keep it less than thirty words: "

- MSCOCO Sampling
  - within this dataset, each image is associated with five distinct text descriptions
  - randomly select a subset of images
  - choose one description as the meta-input text and another as the meta-output text

- Human Rewriting

Language Rewriting



- A sentence that informs the LLM about the task
- Three examples sampled from the meta-input-output pairs
- The text sample that requires rewriting
- LLaMA-7B generates four distinct rewrites (ChatGPT, Bard, COCO, Human)

Language Rewriting

Original: Traffic jam on the road, a lot of cars which go towards each other and to the different directions

LLM ChatGPT: Traffic jam on the road, with a lot of cars moving in different directions, as well as towards each other.
LLM Bard: A lot of cars line the street, all heading toward the same intersection.

LLM MSCOCO: Traffic jam on the road with cars going in different directions, some cars are on the same lane, others are on different lanes.
LLM Human: A traffic jam on the road. There are a lot of cars and many of them are going in different directions.

LaCLIP (Language augmented CLIP)

$$\text{aug}_T(x_T) \sim \text{Uniform}([x_{T0}, x_{T1} \ldots, x_{TM}])$$

- Randomly select a text sample from either original text or one of the generated rewrites

$$L_I = -\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(f_I(\text{aug}_I(x_I^i)), f_T(\text{aug}_T(x_T^i)))/\tau\right)}{\sum_{k=1}^{N} \exp\left(\text{sim}(f_I(\text{aug}_I(x_I^i)), f_T(\text{aug}_T(x_T^k)))/\tau\right)}$$

- The additional text augmentation does not bring additional computation or parameter overheads

Experiments Setup

- Datasets
  - Training set: CC3M, CC12M, RedCaps, LAION-400M
  - Evaluation set: 15 common downstream datasets

- Training Parameters
  - CC3M, CC12M, RedCaps
    - ViT-B/16
    - 8,192 batch size
  - LAION-400M
    - ViT-B/32, ViT-B/16
    - 32,768 batch size

Zero-Shot(ZS) Evaluation

- The class text embeddings are used to compute distance with the image feature, and images are classified to class with the shortest distance

Data	Model	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	GTSRB	Country211	**Average**	**ImageNet**
Model Architecture: ViT-B/32																		
LAION-400M	CLIP	**79.9**	91.8	72.0	64.6	77.0	15.8	49.9	84.8	89.3	64.4	95.3	43.2	60.6	36.9	14.5	62.7	62.0
	LaCLIP	79.7	**92.4**	**73.0**	**64.9**	**81.9**	**20.8**	**55.4**	**87.2**	**91.8**	**70.3**	**97.3**	**50.6**	**61.5**	**49.4**	**16.0**	**66.1**	**64.4**
Model Architecture: ViT-B/16																		
CC3M	CLIP	10.3	54.9	21.8	25.0	0.8	1.4	10.5	12.8	43.3	10.2	77.6	14.1	19.1	**6.9**	0.6	20.6	15.8
	LaCLIP	**14.2**	**57.1**	**27.5**	**35.1**	**1.6**	**1.6**	**16.6**	**15.6**	**52.7**	**14.7**	**86.2**	**15.0**	**24.3**	6.4	**1.0**	**24.6**	**21.5**
CC12M	CLIP	50.8	64.9	38.5	44.7	24.1	2.4	19.4	64.1	77.4	33.2	91.0	20.1	38.9	7.3	5.1	38.8	40.2
	LaCLIP	**60.7**	**75.1**	**43.9**	**57.0**	**36.3**	**5.6**	**31.0**	**72.4**	**83.3**	**39.9**	**95.1**	**27.3**	**44.3**	**12.7**	**8.9**	**46.2**	**48.4**
	SLIP	52.5	80.7	46.3	48.8	24.9	2.3	25.1	58.6	77.6	29.2	89.1	**25.8**	36.6	6.0	5.7	40.6	42.1
	LaSLIP	**62.9**	**82.0**	**50.2**	**59.6**	**32.2**	**4.4**	**30.1**	**70.6**	**82.4**	**37.4**	**95.0**	20.4	**45.6**	**10.1**	**9.2**	**46.1**	**49.7**
RedCaps	CLIP	81.5	70.4	39.9	33.2	19.2	1.9	19.7	**82.7**	72.8	53.9	**92.8**	23.3	33.6	**8.3**	6.2	42.6	42.9
	LaCLIP	**85.0**	**74.8**	**40.7**	**40.3**	**21.3**	**2.2**	**23.9**	78.2	**76.4**	**59.0**	91.4	**27.1**	**41.3**	5.6	**7.6**	**45.0**	**46.2**
LAION-400M	CLIP	85.5	93.0	71.7	66.8	83.5	16.7	52.8	90.1	91.2	63.9	97.3	42.4	63.3	**46.2**	17.8	65.5	67.0
	LaCLIP	**86.5**	**93.5**	**73.9**	**67.9**	**87.1**	**24.2**	**58.9**	**90.9**	**92.4**	**73.1**	**98.4**	**48.3**	**65.8**	46.1	**19.6**	**68.4**	**69.3**

- LaCLIP outperform CLIP across various datasets
- LaCLIP is compatible with other techniques intended to enhance CLIP

Few-Shot(FS) Evaluation

- Perform 5-way 5-shot classification with Prototypical Network as classifier on top of the frozen features

Data	Model	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	GTSRB	Country211	**Average**
Model Architecture: ViT-B/32																	
LAION-400M	CLIP	92.5	87.2	89.0	98.0	98.5	78.9	87.4	94.5	99.2	99.0	96.1	**82.8**	**94.3**	79.8	49.7	88.5
	LaCLIP	**93.5**	**91.0**	**90.7**	**98.2**	**99.1**	**82.2**	**87.5**	**95.7**	**99.4**	**99.2**	**97.2**	80.1	94.2	**80.4**	**52.2**	**89.4**
Model Architecture: ViT-B/16																	
CC3M	CLIP	67.6	64.2	73.6	94.1	54.4	46.1	74.4	76.7	93.3	94.3	84.6	**81.4**	**87.1**	66.9	37.3	73.1
	LaCLIP	**70.0**	**69.1**	**76.8**	**95.2**	**57.6**	**49.2**	**75.8**	**77.4**	**95.2**	**95.0**	**89.5**	81.1	85.5	**71.0**	37.3	**75.0**
CC12M	CLIP	87.0	77.5	82.1	97.2	90.9	62.0	83.3	91.1	98.2	97.6	92.6	**83.4**	91.2	70.6	44.3	83.3
	LaCLIP	**89.9**	**81.3**	**85.0**	**98.0**	**95.3**	**68.1**	**84.9**	**93.4**	**98.9**	**98.4**	**95.9**	83.0	**92.4**	**76.4**	**46.7**	**85.8**
	SLIP	87.6	79.2	83.0	97.5	85.6	56.4	85.8	88.1	97.7	97.1	92.5	**84.9**	91.0	62.4	43.0	82.1
	LaSLIP	**90.5**	**84.9**	**86.6**	**98.1**	**91.6**	**61.0**	**86.7**	**89.8**	**98.7**	**97.8**	**94.2**	84.0	**92.8**	**65.8**	**45.4**	**84.5**
RedCaps	CLIP	94.4	80.6	85.3	95.9	88.5	54.5	**82.6**	**94.5**	97.8	99.0	94.8	84.9	91.3	75.3	40.6	84.0
	LaCLIP	**95.8**	**81.4**	**85.4**	**96.2**	**90.9**	**58.8**	82.4	94.1	**98.0**	**99.2**	**95.6**	**86.2**	**92.1**	**76.5**	**42.6**	**85.0**
LAION-400M	CLIP	95.0	90.1	90.7	98.2	99.2	80.8	88.7	96.2	99.5	99.4	97.1	**84.5**	95.0	77.7	55.1	89.8
	LaCLIP	**95.8**	**92.7**	**91.9**	**98.4**	**99.5**	**86.1**	**89.0**	**97.1**	**99.6**	**99.5**	**98.1**	82.9	95.0	**80.9**	**57.9**	**91.0**

Linear-Probing(LP) Evaluation

- Train a linear classifier on top of the frozen features

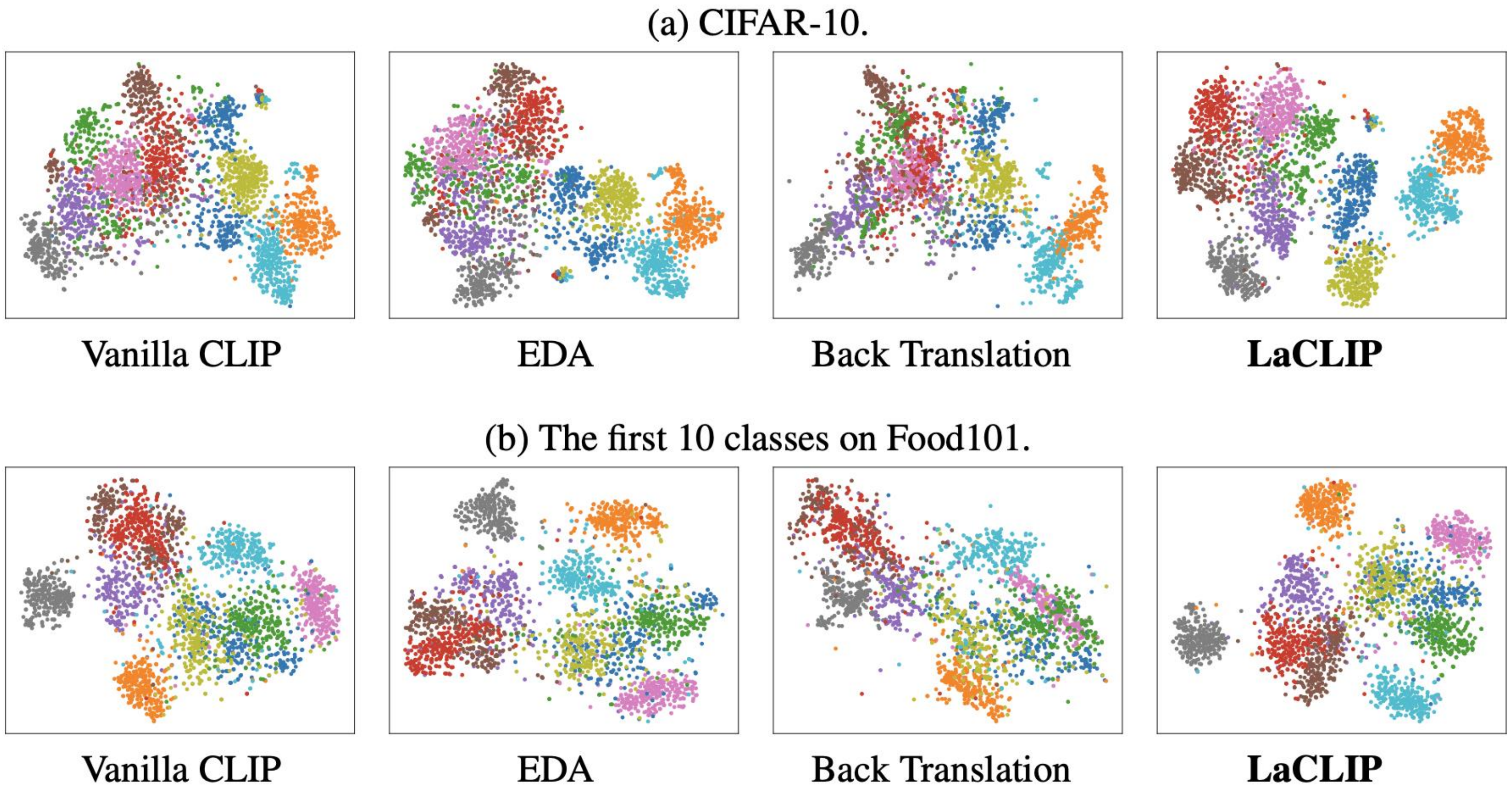
Data	Model	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	GTSRB	Country211	**Average**	**ImageNet**
Model Architecture: ViT-B/32																		
LAION-400M	CLIP	**85.8**	95.8	83.6	75.1	89.2	54.3	**79.7**	86.9	94.5	96.8	97.9	**96.3**	93.5	88.6	**23.1**	82.7	74.6
	LaCLIP	85.1	**96.2**	**84.2**	**75.6**	**90.1**	**56.1**	79.6	**89.1**	**94.8**	**97.7**	**98.4**	95.8	**93.6**	88.6	22.9	**83.2**	**75.3**
Model Architecture: ViT-B/16																		
CC3M	CLIP	62.6	86.8	68.1	58.5	**32.8**	40.9	63.4	69.6	82.0	89.4	91.7	**95.9**	89.0	71.9	**13.3**	67.7	54.5
	LaCLIP	**63.8**	**87.7**	**69.5**	**60.2**	32.4	**42.7**	**64.0**	**71.1**	**83.3**	**90.2**	**93.4**	95.8	**89.7**	**74.6**	13.2	**68.8**	**56.5**
CC12M	CLIP	81.6	93.8	79.3	72.0	75.1	52.6	75.6	86.2	92.2	95.3	97.3	**96.7**	**93.1**	80.6	19.7	79.4	70.3
	LaCLIP	**82.9**	**94.7**	**79.7**	**73.8**	**79.9**	**54.5**	**75.7**	**87.7**	**93.0**	**96.4**	**98.0**	96.4	93.0	**81.9**	19.7	**80.5**	**72.3**
	SLIP	84.4	94.2	79.1	73.5	74.2	**54.6**	76.5	**86.1**	92.7	95.7	97.6	96.8	**93.7**	74.0	20.6	79.6	73.2
	LaSLIP	**85.2**	**94.6**	**80.8**	**75.1**	**77.0**	53.8	**78.5**	85.6	**93.7**	**96.5**	**97.9**	96.8	93.5	**76.1**	**21.1**	**80.4**	**74.4**
RedCaps	CLIP	89.1	94.1	**78.8**	65.6	74.0	52.5	73.2	**91.5**	91.4	97.7	**98.0**	96.3	**93.5**	80.8	17.0	79.6	71.8
	LaCLIP	**90.1**	**94.3**	78.5	**66.6**	**77.6**	**53.6**	**73.9**	90.8	**91.5**	**97.9**	97.6	**96.6**	92.7	80.8	**17.2**	**80.0**	**71.9**
LAION-400M	CLIP	90.5	**96.9**	85.0	78.1	92.1	57.2	80.0	90.9	95.7	98.0	98.7	**96.7**	**94.7**	**90.3**	27.0	84.8	78.6
	LaCLIP	**90.7**	96.7	**85.5**	**78.7**	**92.8**	**63.1**	**81.3**	**92.8**	**96.2**	**98.8**	**99.1**	96.4	94.6	89.5	**27.5**	**85.6**	**79.9**

Ablation Study

- Varying augmentation strategies
  - EDA: word-level synonym replacement
  - back translation: translate text to another language and then back to the original language

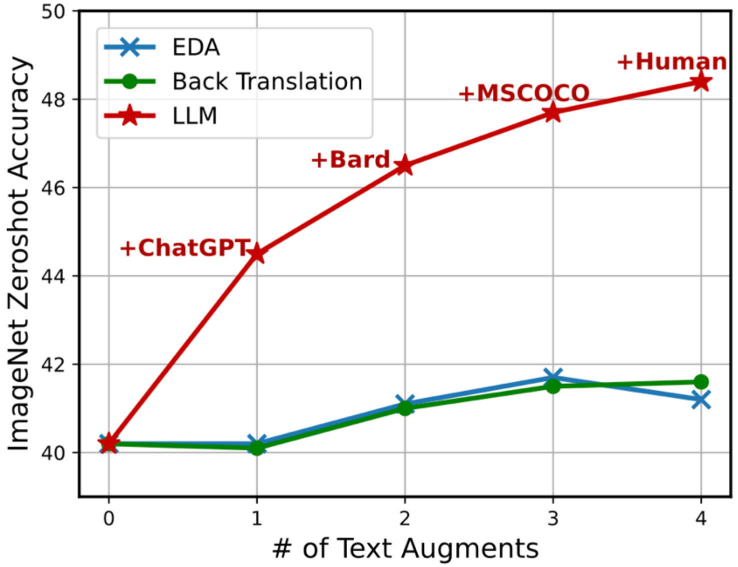
Augment	ZS		FS	LP	
	DS	**IN**		**DS**	**IN**
N/A (CLIP)	38.8	40.2	83.3	79.4	70.3
EDA [58]	40.6	41.2	83.4	79.4	70.5
Back Trans [50]	40.4	41.6	83.9	79.8	70.7
LLM (Ours)	**46.2**	**48.4**	**85.8**	**80.5**	**72.3**

Ablation Study



Ablation Study

- Scaling with number of augmentations



  - simpler augmentation strategies exhibit poor scalability because of there limited diversity
  - conversely, LLM-based text augmentation consistently improves performance as more augmentations are added

Ablation Study

- Different meta-input-output pair for ICL

Source	ZS		FS	LP	
	DS	**IN**		**DS**	**IN**
ChatGPT	42.3	44.5	84.8	79.8	71.2
Bard	41.7	44.8	85.0	79.6	71.2
MSCOCO	42.1	44.6	84.8	79.8	71.3
Human	43.0	45.1	84.8	79.9	71.3

  - with the model trained with augmentations using the Human pair slightly outperforming the others
  - humans have the advantage of viewing the corresponding image, which allows them to generate more accurate and diverse rewrites

Ablation Study

- Comparison with Pre-trained Text Encoder

Method	Text Encoder		ZS		FS	LP	
	Pre-train	**Freeze**	**DS**	**IN**		**DS**	**IN**
CLIP	✘	✘	38.8	40.2	83.3	79.4	70.3
	✔	✘	42.1	42.9	83.6	79.5	70.4
	✔	✔	24.5	23.2	80.3	74.9	66.0
LaCLIP	✘	✘	**46.2**	**48.4**	**85.8**	**80.5**	**72.3**

  - fine-tuning based on the pre-trained text encoder exhibits some improvements
  - LaCLIP outperforms all configurations, underscoring the benefit and necessity for text augmentation strategies

Ablation Study

- LLaMA model size

(a) Zero-shot and Linear-probing Experiment Results

Model Size	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	GTSRB	Country211	**Average**	**ImageNet**
Zero-shot																	
N/A (CLIP)	50.8	64.9	38.5	44.7	24.1	2.4	19.4	64.1	77.4	33.2	91.0	20.1	38.9	7.3	5.1	38.8	40.2
7B	57.0	71.1	38.9	51.2	**31.6**	3.9	25.5	63.0	80.8	**36.9**	92.9	24.5	39.6	10.1	6.9	42.3	44.5
13B	55.4	71.5	**39.3**	51.3	29.6	4.0	**26.4**	**65.7**	80.7	36.0	93.8	17.0	38.7	9.0	**7.6**	41.7	**44.8**
33B	56.7	**76.0**	37.7	**52.0**	31.2	**4.5**	24.3	60.7	**80.9**	35.4	**94.4**	26.7	**40.4**	11.6	7.0	42.6	44.4
65B	**57.5**	69.2	38.9	51.6	31.1	4.1	25.3	65.2	79.0	36.8	93.1	**31.7**	40.2	**15.0**	7.4	**43.1**	44.4
Linear-Probing																	
N/A (CLIP)	81.6	93.8	79.3	72.0	75.1	52.6	75.6	86.2	92.2	95.3	97.3	96.7	93.1	80.6	19.7	79.4	70.3
7B	81.8	**94.3**	**79.7**	73.3	77.5	55.0	75.4	87.4	92.5	**96.3**	**97.6**	**96.9**	92.6	81.3	**20.2**	80.1	71.2
13B	82.1	93.7	78.2	73.0	77.6	**55.6**	74.6	87.4	**92.7**	96.0	97.4	96.3	**93.2**	82.5	20.0	80.0	71.2
33B	81.8	94.1	79.4	73.3	78.6	54.1	75.0	86.4	92.4	96.1	97.3	96.6	93.1	81.5	19.8	80.0	**71.4**
65B	**82.2**	94.2	79.3	73.0	**78.7**	54.0	75.4	87.3	91.9	95.4	97.5	96.7	92.7	**82.5**	20.0	80.1	71.3

Ablation Study

- LLaMA model size

(b) Few-shot Experiment Results

Model Size	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	GTSRB	Country211
N/A (CLIP)	87.0±0.5	77.5±0.6	82.1±0.7	97.2±0.2	90.9±0.5	62.0±1.0	83.3±0.6	91.1±0.5	98.2±0.2	97.6±0.2	92.6±0.4	83.4±0.5	91.2±0.4	70.6±0.8	44.3±0.7
7B	88.8±0.5	78.4±0.6	83.3±0.6	97.7±0.2	93.4±0.4	66.5±1.0	84.4±0.6	92.5±0.4	98.6±0.2	98.0±0.2	94.3±0.3	84.0±0.5	**92.3±0.4**	**73.7±0.8**	45.6±0.7
13B	**89.1±0.5**	79.2±0.6	82.8±0.7	**97.9±0.2**	94.0±0.4	66.3±1.0	84.1±0.6	92.9±0.4	98.5±0.2	**98.2±0.2**	94.4±0.3	83.2±0.5	91.6±0.4	73.6±0.8	45.7±0.7
33B	88.6±0.5	**80.3±0.6**	**83.6±0.6**	97.8±0.2	**94.3±0.4**	65.4±1.0	**84.7±0.6**	92.8±0.4	98.6±0.2	**98.2±0.2**	**94.5±0.3**	84.2±0.5	92.1±0.4	72.0±0.8	**45.8±0.7**
65B	88.8±0.5	79.2±0.6	82.9±0.6	97.8±0.2	94.1±0.4	**66.6±1.0**	84.3±0.6	**93.1±0.4**	98.6±0.2	98.1±0.2	**94.5±0.3**	**85.6±0.5**	91.9±0.4	72.5±0.8	45.6±0.7

Multi-Text Training Loss with LaCLIP

- LaCLIP-MT, which incorporates a multi-positive contrastive training loss

$$L_{I*} = -\frac{1}{M}\sum_{i=1}^{N}\sum_{j=0}^{M}\log\frac{\exp\left(\text{sim}(f_I(\text{aug}_I(x_I^i)), f_T(x_{Tj}^i))/\tau\right)}{\sum_{k=1}^{N}\exp\left(\text{sim}(f_I(\text{aug}_I(x_I^i)), f_T(x_{Tj}^k))/\tau\right)}$$

- Final training loss: $L = (L_{I*} + L_T)/2$

Dataset	Method	ZS		FS	LP	
		DS	IN		DS	IN
CC12M	CLIP	38.8	40.2	83.3	79.4	70.3
	LaCLIP	**46.2**	48.4	85.8	80.5	72.3
	LaCLIP-MT	45.2	**49.0**	85.8	**80.6**	**72.4**
RedCaps	CLIP	42.6	42.9	84.0	79.6	71.8
	LaCLIP	45.0	46.2	85.0	80.0	71.9
	LaCLIP-MT	**46.1**	**48.1**	**85.3**	**80.3**	**72.4**

Conclusion

- Training CLIP with rewritten texts generated using the ICL capability of LLM can improve performance

My Review

- **Enhancing Encoder Quality:** Effectively improved the quality of both CLIP's text and image encoders by generating augmented texts via LLMs.
- **Effective Use of ICL:** create diverse meta-input-output pairs from sources like ChatGPT, Bard, MSCOCO, and Human annotators to leverage the ICL capabilities of LLMs.
- **Missing Comparison with SimCSE:** A comparison with the SimCSE was absent and would have been an insightful addition.

Open Question

- M or M+1

$$L_{I*} = -\frac{1}{M}\sum_{i=1}^{N}\sum_{j=0}^{M}\log\frac{\exp\left(\text{sim}(f_I(\text{aug}_I(x_I^i)), f_T(x_{Tj}^i))/\tau\right)}{\sum_{k=1}^{N}\exp\left(\text{sim}(f_I(\text{aug}_I(x_I^i)), f_T(x_{Tj}^k))/\tau\right)}$$