SuperCLIP: CLIP with Simple Classification Supervision

Weiheng Zhao[1] **Zilong Huang**[2]* **Jiashi Feng**[2] **Xinggang Wang**[1]

[1] *School of EIC, Huazhong University of Science and Technology*

[2] *ByteDance*

Motivation

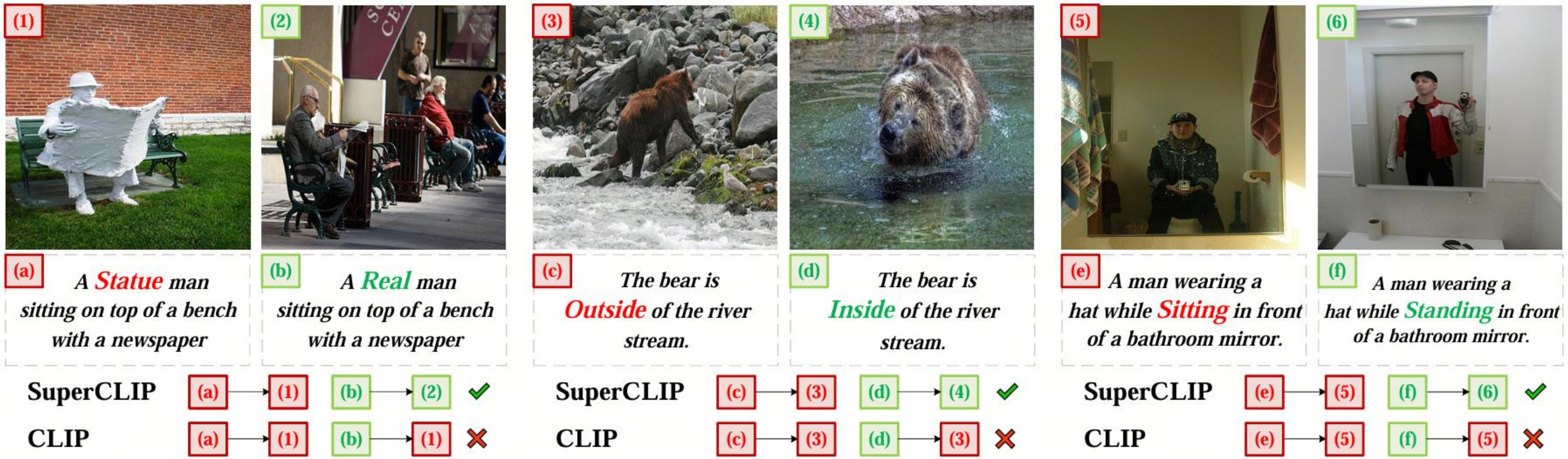
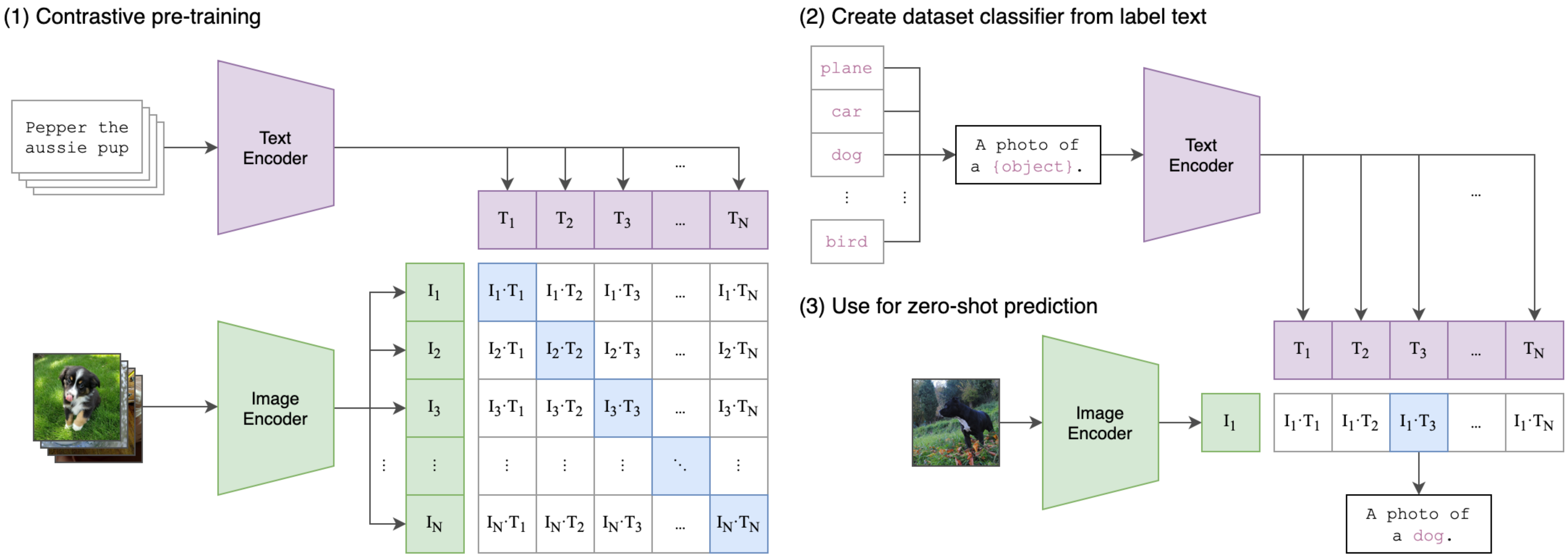


Figure 1: **Evaluating Fine-Grained Alignment in Image-Text Retrieval.** Each row presents pairs of images and captions that are visually and semantically very similar, but differ in fine-grained semantic distinctions such as object status (e.g. **Statue** vs. **Real**), spatial relation (e.g. **Outside** vs. **Inside**), and action (e.g. **Sitting** vs. **Standing**). While both images and texts are close in meaning, SuperCLIP demonstrates a stronger ability than CLIP in correctly distinguishing these fine-grained semantic distinctions. Additional examples are provided in **Appendix A.1.**

Contrastive Language-Image Pretraining



- Enable zero-shot visual understanding
- Exhibit proper scaling characteristics
- Demonstrate robust cross-modal capability

- Limited utilization of fine-grained textual semantics
- Dependence on noisy and weakly aligned web data
- High computational and resource demands

Radford et al. (2021). *Learning Transferable Visual Models from Natural Language Supervision.* (ICML 2021), PMLR.

Limitations of Current CLIP Improvements

- **Limited focus on fine-grained alignment**

  Many improvements primarily concentrate on enhancing model architecture and training efficiency, while paying less attention to capturing fine-grained visual-text correspondences at the word or region level.

- **Dependence on additional labeled data**

  Several methods rely on extra annotated or re-captioned datasets to improve alignment, which limits scalability and diverges from CLIP's original web-scale, weakly supervised paradigm.

- **High computational cost**

  More complex architectures and dense supervision often lead to substantial computational and resource overhead, reducing training efficiency.

- **Limited generalization**

  Some approaches boost performance on specific benchmarks but compromise the generality and simplicity that make CLIP broadly transferable.

Super Simple Classification-based Supervision

Use raw text tokens as classification labels for image encoder

- **IDF Weighting**

  Down-weight frequent tokens using inverse document frequency:

$$w_c = \log\left(\frac{|\mathcal{D}|}{1 + \mathrm{df}(c)}\right)$$

- **Weighted Label**

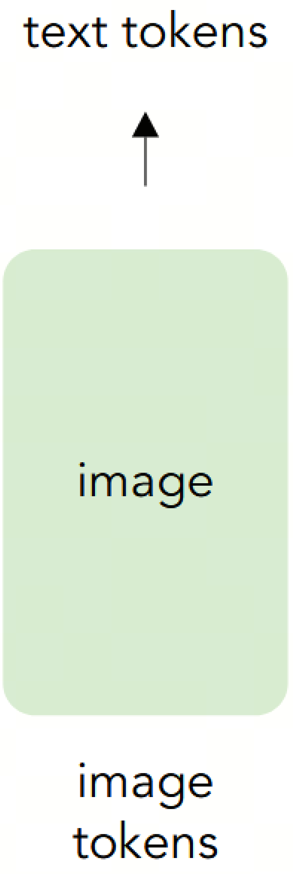
  Normalize weighted token labels:

$$\hat{y}_c = \frac{w_c y_c}{\sum_{c'=1}^{V} w_{c'} y_{c'}}$$

- **Classification Loss**

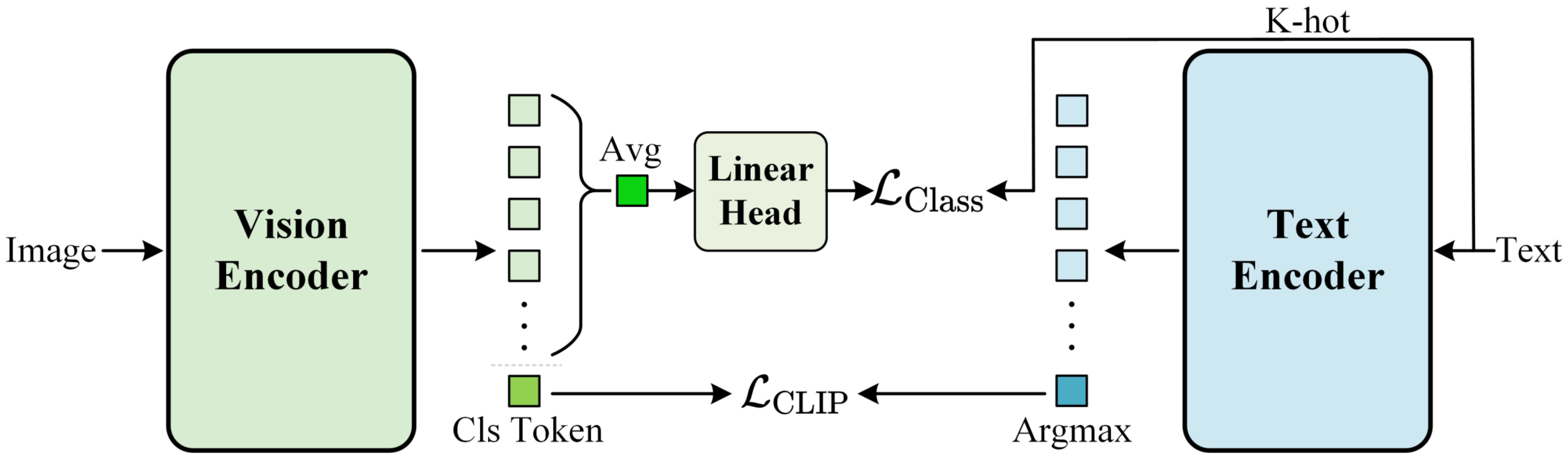
  Align model logits with weighted labels via cross-entropy:

$$\mathcal{L}_{\text{Class}} = -\sum_{c=1}^{V} \hat{y}_c \log\left(\frac{e^{x_c}}{\sum_{c'=1}^{V} e^{x_{c'}}}\right)$$



Huang et al. (2024). *Classification Done Right for Vision-Language Pre-Training*. NeurIPS 2024.

CLIP with Simple Classification Supervision



$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{Class}}$$

Comparison across Different Model Sizes

Model	Pretrain	Image Classification		Image Retrieval		Text Retrieval	
		val	v2	COCO	Flickr	COCO	Flickr
CLIP	B-512M	60.5	53.0	29.0	54.5	46.7	73.3
SuperCLIP	B-512M	63.5 (+3.0)	55.2 (+2.2)	31.3 (+2.3)	56.9 (+2.4)	47.8 (+1.1)	75.6 (+2.3)
CLIP	L-512M	66.1	57.4	32.7	57.0	49.6	76.4
SuperCLIP	L-512M	70.1 (+4.0)	62.5 (+5.1)	35.9 (+3.2)	62.4 (+5.4)	52.2 (+2.6)	79.3 (+2.9)
CLIP	L-12.8B	79.0	72.0	43.9	72.7	62.5	87.0
SuperCLIP	L-12.8B	80.0 (+1.0)	72.8 (+0.8)	45.5 (+1.6)	74.2 (+1.5)	63.1 (+0.6)	88.1 (+1.1)

Table 2: **Comparison with CLIP across Different Model Sizes.** We report **zero-shot** image classification accuracy (%) on ImageNet-1K (val and v2), and **zero-shot** image and text retrieval (Recall@1, %) on COCO and Flickr30K, comparing CLIP and our SuperCLIP under three settings: B-512M, L-512M, and L-12.8B, where models are pretrained on 512M or 12.8B samples from DataComp-1B. Values in parentheses reflect absolute gains or drops for SuperCLIP relative to CLIP.

Brief Analysis of Performance Gain

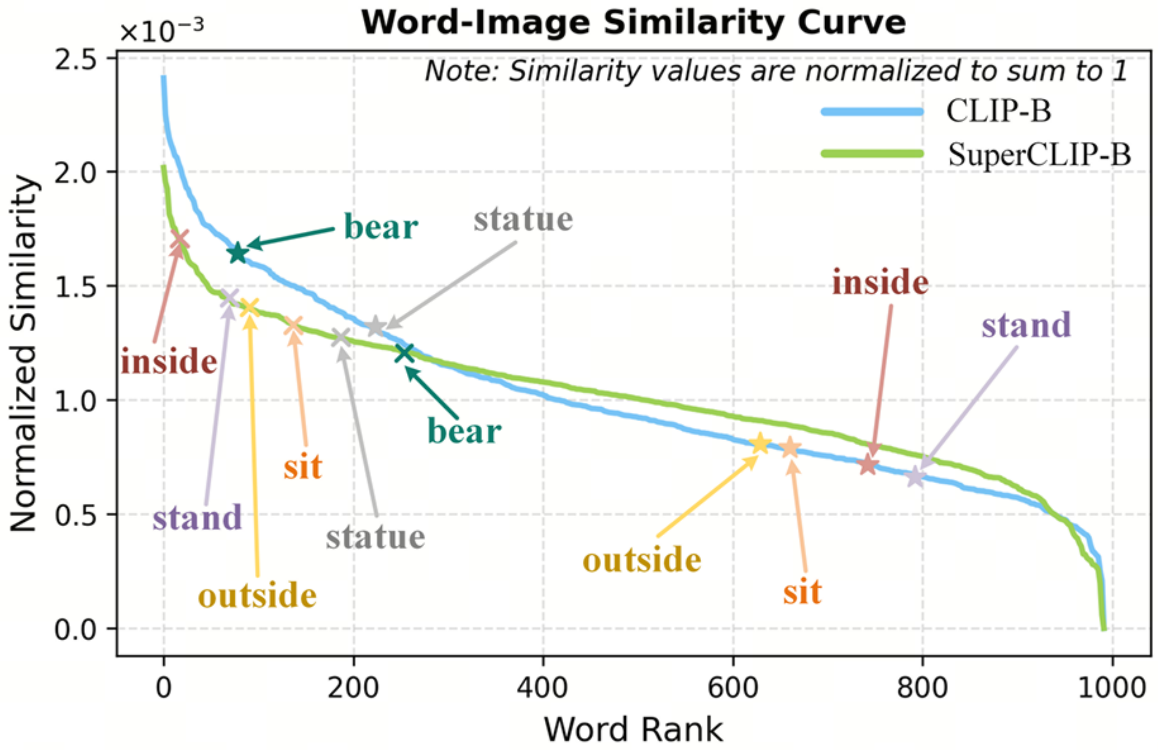


Figure 3: **Visualization of Word-Image Similarity Distribution.** We ranked the similarity scores of 1,000 words that appeared in the captions and highlighted the positions of fine-grained attributes discussed in the above Fig.1.

Component	CLIP	SuperCLIP
Vision Encoder	59.689	59.689
Text Encoder	6.547	6.547
Linear Head	-	0.051

Table 3: **FLOPs Count (GFLOPs).**

Metric	CLIP	SuperCLIP
Total words	992	992
Std Deviation	0.0340	0.0213
Value Range	0.2065	0.1401
Mean Slope	0.000208	0.000141
Top-1→100	0.0702	0.0439

Table 4: **Statistical Summary.** Mean Slope (Δsim): Average drop in similarity between words as the rank goes down. Top-1→100: Difference in similarity between the 1st and 100th word.

Comparison with CLIP using Mixed Caption

Model-Size	Mixed Caption	Image Retrieval		Text Retrieval		Image Classification
	Short / Long	COCO	Flickr	COCO	Flickr	Average. 38
CLIP-B	1.0 / 0.0	29.0	54.4	46.7	73.7	43.4
SuperCLIP-B	1.0 / 0.0	**31.3**	**57.6**	**47.8**	**75.6**	**44.5** (+1.1)
CLIP-B	0.0 / 1.0	23.6	41.8	40.5	66.2	27.8
SuperCLIP-B	0.0 / 1.0	**30.6**	**48.7**	**47.2**	**70.4**	**31.4** (+3.6)
CLIP-B	0.8 / 0.2	32.7	57.5	50.2	76.0	42.8
SuperCLIP-B	Dual	**34.1**	**60.2**	**51.2**	**76.6**	**45.1** (+2.3)
CLIP-L	1.0 / 0.0	32.7	57	49.6	76.4	45.7
SuperCLIP-L	1.0 / 0.0	**35.9**	**62.4**	**52.2**	**79.3**	**48.6** (+2.9)
CLIP-L	0.0 / 1.0	26.2	43.1	42.9	65.9	30.0
SuperCLIP-L	0.0 / 1.0	**34.2**	**55.7**	**52.1**	**75.0**	**33.8** (+3.8)
CLIP-L	0.8 / 0.2	37.0	61.1	53.7	78.8	46.8
SuperCLIP-L	Dual	**37.6**	**65.3**	**54.0**	**82.5**	**49.5** (+2.7)

Table 5: **Comparison with CLIP using Mixed Captions. "Mixed Caption"** refers to the ratio of short (DataComp-1B) and long (Recap-DataComp-1B) captions used during training. The **"0.8/0.2"** mix is the optimal ratio identified in [31] through extensive tuning. **"Dual"** denotes our setup where the contrastive loss uses only short captions and the classification loss uses only long captions. We report average **zero-shot** image classification accuracy (%) across 38 datasets, and **zero-shot** image/text retrieval (Recall@1, %) on COCO and Flickr30K, using **512M** training samples. **Bold** numbers indicate the best results, while values in parentheses show absolute gains or drops of SuperCLIP relative to CLIP.

Generalize to Other CLIP-style Frameworks

Model	Image Classification		Image Retrieval		Text Retrieval	
	val	v2	COCO	Flickr	COCO	Flickr
SigLIP	60.4	52.8	29.8	53.9	45.8	73.2
SuperSigLIP	64.1 (+3.7)	55.9 (+3.1)	32.5 (+2.7)	56.8 (+2.9)	48.6 (+2.8)	75.9 (+2.7)
FLIP	58.1	50.1	27.5	51.8	44.1	66.7
SuperFLIP	61.3 (+3.2)	53.5 (+3.4)	30.1 (+2.6)	54.0 (+2.2)	46.7 (+2.6)	72.0 (+5.3)

Table 6: **Generalization to Other CLIP-Style Frameworks.** We report **zero-shot** performance on image classification accuracy (%) on ImageNet-1K (val and v2), and image/text retrieval (Recall@1, %) on COCO and Flickr30K, comparing SigLIP and FLIP with their SuperCLIP variants (SuperSigLIP and SuperFLIP). All models are pretrained with 512M samples (B-512M). Numbers in parentheses indicate absolute gains over the original models.

Enhance CLIP for Purely Visual Tasks

Model	Pretrian	Class ↑	Segmentation ↑		Depth ↓
		ImageNet-1K	PASCAL	ADE20k	NYUv2
CLIP	B-512M	75.6	57.8	28.0	0.768
SuperCLIP	B-512M	77.1 (+1.5)	65.5 (+7.7)	32.1 (+4.1)	0.746 (-0.022)
CLIP	L-512M	79.7	67.8	34.2	0.740
SuperCLIP	L-512M	81.0 (+1.3)	71.2 (+3.4)	36.3 (+2.1)	0.733 (-0.007)

Table 7: **Enhance CLIP for Purely Visual Tasks.** We report performance on three purely visual tasks: **linear probing** image classification(Class) on ImageNet-1K (Accuracy, %), semantic segmentation(Segmentation) on PASCAL and ADE20K (mIoU), and depth estimation(Depth) on NYUv2 (RMSE). We compare CLIP and SuperCLIP under identical pretraining and evaluation settings to ensure a fair comparison across all purely visual tasks. Numbers in parentheses indicate absolute improvements over the original CLIP models.

Mitigate CLIP's Drop with Limited Batch Sizes

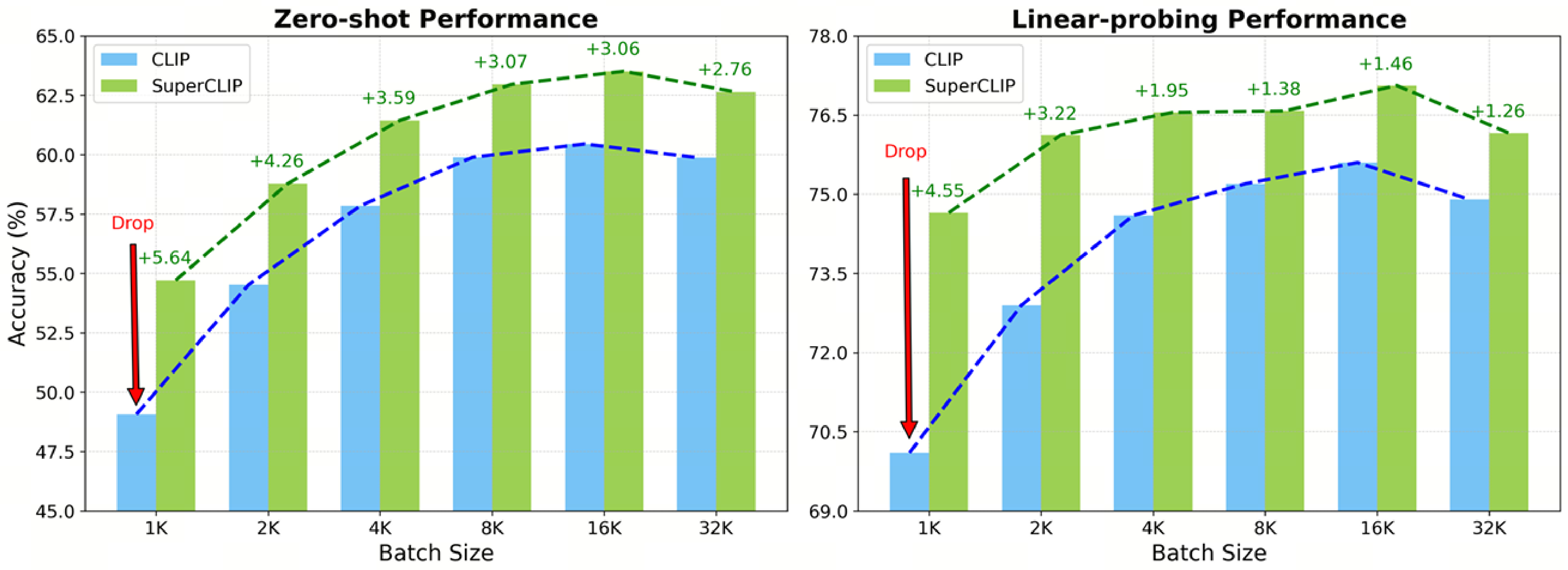


Figure 4: **Mitigate CLIP's Drop with Limited Batch Sizes.** We report zero-shot (Left) and linear-probing (Right) image classification accuracy (%) on ImageNet-1K (val) under varying batch sizes. The green bars represent the performance of SuperCLIP under different batch sizes, while the gray bars indicate the performance of CLIP under the corresponding batch sizes. Green numbers indicate absolute improvements over the original CLIP models at the corresponding batch sizes.

Integrate in Multi-modal LLM

Model	Pretrian	Vision & Language Downstream Tasks							
		VQAv2	GQA	VizWiz	T-VQA	SQA	MMB	MME	POPE
CLIP	B-512M	67.8	55.4	42.1	47.8	**69.3**	49.1	1453	81.7
SuperCLIP	B-512M	**69.6**	**57.5**	**44.4**	**48.4**	69.1	**55.9**	**1562**	**82.0**

Table 8: **Compare with CLIP under Multi-modal LLM Setting.** We report the performance scores on 8 vision & language downstream tasks. **Bold** numbers indicate the best result.

Additional Ablation Studies

Task	Weighting Factor (λ)				
	0.4	**0.6**	**1**	**1.4**	**1.6**
Classification	44.1	45.0	47.1	46.9	47.2
Image Retrieval	41.3	42.1	44.0	43.8	44.2
Text Retrieval	58.3	59.8	61.0	60.9	62.0

Table 9: **Loss Weighting.** We report **zero-shot** classification accuracy (%) on ImageNet-1K (val) and the average retrieval result (Recall@1, %) across COCO and Flickr30K.

Design	Image Retrieval		Text Retrieval		Classification
	COCO	**Flickr**	**COCO**	**Flickr**	**ImageNet-1K**
w/o IDF	31.6	51.7	48.0	71.1	44.8
IDF	**33.2**	**54.7**	**48.9**	**73.1**	**47.1**

Table 10: **IDF Weighting.** We report **zero-shot** classification accuracy (%) on ImageNet-1K (val) and retrieval results (Recall@1, %) on COCO and Flickr30K, respectively.

Thank you !

Code &Models: https://github.com/hustvl/SuperCLIP