No Hard Negatives Required: Concept Centric Learning Leads to Compositionality without Degrading Zero-shot Capabilities of Contrastive Models
- *Supplementary Materials* -

Hai X. Pham David T. Hoffmann Ricardo Guerrero Brais Martinez
Samsung AI Center Cambridge, UK

A. Additional experimental results

We provide additional results of our method, C^{2}LIP, and the baseline contrastive models employing the same ViT-B backbone listed in Tab. A1. Please refer to the main paper for details of the benchmarks and evaluation protocol.

This section is organized as follows. First, we show that our proposed loss function is not sensitive to scaling hyperparameters in Sec. A.1. Next, the evaluation results on compositionality benchmarks including SugarCrepe and SugarCrepe++ are described in Sec. A.2, and the evaluation of VLM tuned with our frozen C^{2}LIP in Sec. A.3. Furthermore, we discuss zero-shot retrieval performance in Sec. A.4. Finally, the zero-shot classification results are given in Sec. A.5.

Table A1. **Baseline methods.** Summary of baseline methods and their corresponding training datasets. The last column indicates whether the provided checkpoints were trained from scratch (✓).

Model	Training data	Train from scratch
Composition-aware		
CE-CLIP [15]	MSCOCO	
NegCLIP [14]	MSCOCO	
CLIC [10]	LAION-1.5B	
DAC [3]	CC3M	✓
SLVC [2]		
CoN-CLIP [11]		
TripletCLIP [9]		
Codebook-based		
Codebook-CLIP [1]		✓
IL-CLIP [16]		✓
Fine-grained training		
DreamLIP-3m [17]		✓
FLAIR-3m [12]		✓
FG-CLIP [13]	LAION-2B + FineHARD	✓
FineCLIP [5]	MSCOCO	
LLIP [6]	Common Crawl 12.8B	✓

A.1. Sensitivity to hyperparameters in objective function

Tab A2 shows the avarage accuracies on SugarCrepe of models trained with different values of λ_{hnc} & λ_{xac}. The variation in performance scores is marginal, showing that our proposed loss function is insensitive to the values of these hyperparameters. We selected the combination of $(1, 0.01)$ in all our experiments.

Table A2. **Ablation of trade-off hyperparameters in the objective function.** We experimented with different values of λ_{hnc} & λ_{xac} showing our proposed method incurs little sensitivity to these hyperparameters.

λ_{hnc}	λ_{xac}	SugarCrepe Average Accuracy
0.5	0.5	84.6
0.5	0.1	85.2
0.5	0.01	85.2
1	0.5	85.1
1	0.01	**85.6**

A.2. Compositionality evaluation

In addition to the task-specific average scores shown in Tab. 2 of the main paper, we include all accuracy scores of all competing methods on all sub-tasks of SugarCrepe and SugarCrepe++ benchmarks in Tab. A3 and Tab. A4, respectively.

SugarCrepe benchmark. As can be seen in Tab. A3, our concept centric contrastive learning method improves the performance of the original base model, SigLIP, by 7.67% on average, surpassing most other composition-aware methods as the second best performing model, only behind DAC-LLM [3] by 0.8 percentage points. Notably, these methods rely on training with hard-negatives to induce the compositionality representations. Instead our method emphasizes better exploiting regular data, with auxiliary concept cen-

tric objectives, to improve compositional representations. In particular, C^{2}LIP excels at recognizing incorrect objects in the image, evidenced by the highest scores on "Replace Object" and "Add Object" sub-tasks. Moreover, C^{2}LIP exhibits strong attribute-binding capabilities, with the highest score on "Swap Attribute" and second best on "Replace Attribute". Our method, however, lags behind in "Replace Relation" and "Swap Object" sub-tasks. On the other hand, the substantial improvements on these two sub-tasks compared to the original SigLIP model demonstrate the effectiveness of our method. The results on SugarCrepe should be interpreted carefully, as the benchmark is insufficient to evaluate lexical sensitivity and semantic understanding [4].

SugarCrepe++ benchmark. SugarCrepe++ [4] resolves this problem by extending the protocol of SugarCrepe by using two positive captions and requiring both of them to be higher ranked than the negative, to be considered correct. By that, it aims to address the limitation of SugarCrepe, where the caption patterns can be, to some extent, imitated to create custom training data. However, the second positive captions in SugarCrepe++ aim to evaluate the generalization capabilities of contrastive models. Intuitively, the concepts in the second caption remain the same as in the first, described differently, thus a model trained to fit the pattern in the first caption may no longer align to the second. As shown in Tab. A4, the methods relying on custom hard-negative training data incur a substantial performance drop across SugarCrepe++ tasks. In contrast, our method performs consistently well across all tasks, showcasing both effective compositional representation as well as strong generalization capabilities. On average C^{2}LIP outperforms the baselines by a large margin, made possible by our proposed concept centric learning framework.

A.3. C^{2}LIP improves visual instruction tuning

In our compositionality experiments, C^{2}LIP demonstrates consistent superior accuracies in comparison to similar-sized models. However, these experiments are limited within the cross-modal retrieval scope, and do not provide insights on whether these capabilities could influence other downstream applications, such as when combining with an LLM in the vision-language models (VLMs). Thus, we conduct experiments based on the LLaVa [8] instruction tuning framework, where the frozen image encoder is combined with an LLM for image-to-text generation. In particular, we replace the CLIP ViT image encoder of LLaVa with SigLIP and C^{2}LIP vision encoders, and train the adapter MLP and LLM following the 2-stage recipe of LLaVa on the same data, resulting in two VLMs, LLaVA-SigLIP and LLaVA-C^{2}LIP, respectively. The resulting models are evaluated on SugarCrepe and SugarCrepe++, where the cosine similarity score is substituted with the VQA image-text matching score (VQAScore) [7]. The performance metrics are summarized in Tab. A5. LLaVA-C^{2}LIP outperforms LLaVA-SigLIP on both SugarCrepe and SugarCrepe++ by 0.4% and 1.6%, respectively, showing that the enhanced compositionality comprehension capabilities of C^{2}LIP also transfers to the VLM that utilizes its vision encoder.

A.4. Zero-shot retrieval evaluation

We evaluate our proposed method and the baselines on two regular retrieval benchmarks: MSCOCO and Flickr30k, and two fine-grained retrieval benchmarks: DOCCI and Image-in-words (IIW). We report Recall@5 scores of image-to-text and text-to-image retrieval tasks, as summarized in Tab. A6. As shown in this table, the composition-aware methods incur degraded retrieval performance compared to the base CLIP model. In contrast, C^{2}LIP performs consistently well, it is the best performing model on most tasks, and on par with the top models for the remaining tasks. Our model enjoys 1.9 percentage point improvement over the original SigLIP on average. These results prove that our training method not only effectively maintains, but can also improve the generalization capability of the original model.

A.5. Zero-shot classification evaluation

Pretrained contrastive V&L models are increasingly employed in a variety of downstream tasks in computer vision. One of them is zero-shot classification, which made this class of models popular in the first place. We evaluate our method and all baselines on 11 classification benchmarks, their accuracies are summarized in Tab. A7. Among the composition-aware baselines, all methods significantly drop their performance on zero-shot classification. Only CLIC almost reaches the accuracy of the original CLIP model (68.1 and 69, respectively), thanks to the generated training data that helps improve generalization in addition to the hard-negatives. Our model incurs a slight 2.9% performance drop compared to the original SigLIP, which is comparatively small in comparison to the other fine-tuned methods. The drop in performance is not surprising, as our model was fine-tuned with scene centric objectives and data, whereas the classification tasks are object centric. Despite the distributional shift, C^{2}LIP can still retain most zero-shot performance.

References

[1] Yuxiao Chen, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N Metaxas, and Hongxia Yang. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1

[2] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja

Table A3. **SugarCrepe compositionality benchmark – all subtasks**. The results in this table are summarized in Tab. 2 of the main paper. We compare the accuracy of C^{2}LIP with baseline methods on different tasks. On average, C^{2}LIP is the second best method, only 0.8 percentage points below DAC-LLM, while being trained on much less data, without custom compositional captions.

Models	Replace				Swap			Add			**Average**
	Obj	Attr	Rel	Avg	Obj	Attr	Avg	Obj	Attr	Avg	
SigLIP ViT-B/16	95.3	86.7	70.3	84.1	60.0	71.5	65.8	89.1	83.8	86.5	79.5
CLIP (OpenAI) ViT-B/32	90.9	80.0	69.2	80.0	61.2	64.1	62.7	77.2	68.8	73.0	73.1
CLIP (OpenAI) ViT-B/16	93.5	81.1	66.7	80.4	60.0	65.0	62.5	78.5	66.9	72.7	73.1
FG-CLIP	95.9	87.1	72.2	85.1	66.5	73.3	69.9	87.6	81.8	84.7	80.6
FineCLIP	95.6	85.2	75.0	85.3	63.3	70.3	66.8	90.0	80.8	85.4	80.0
DreamLIP-3m	87.2	77.3	68.1	77.5	56.3	72.1	64.2	74.3	71.5	72.9	72.4
FLAIR-3m	91.4	82.3	70.5	81.4	63.2	78.5	70.9	84.5	76.6	80.6	78.1
LLIP	89.6	79.4	67.6	78.9	55.9	59.6	57.8	79.1	63.7	71.4	70.7
Codebook-CLIP	53.5	51.4	58.7	54.5	45.7	49.2	47.5	57.3	43.8	50.6	51.4
IL-CLIP	53.1	54.1	51.1	52.8	57.6	52.7	55.2	54.9	48.6	51.8	53.2
CE-CLIP	93.1	88.8	79.0	87.0	72.8	77.0	74.9	92.4	93.4	92.9	85.2
NegCLIP	92.7	85.9	76.5	85.0	**75.2**	75.4	**75.3**	88.8	82.8	85.8	82.5
CLIC	95.6	86.6	75.3	85.8	71.0	74.0	72.5	88.4	91.2	89.8	83.1
DAC-SAM	91.2	85.9	83.9	87.0	71.8	75.1	73.5	87.5	95.7	91.6	84.4
DAC-LLM	94.5	**89.5**	**84.4**	**89.5**	75.1	74.2	74.6	89.7	**97.7**	93.7	**86.4**
SLVC-R	91.3	81.4	64.1	78.9	68.6	69.1	68.8	79.5	91.3	85.4	77.9
SLVC-RL	88.1	76.8	62.7	75.9	64.5	66.5	65.5	75.8	81.2	78.5	73.7
CoN-CLIP	92.5	79.7	60.1	77.4	58.8	66.1	62.4	86.7	78.2	82.4	74.6
TripletCLIP	94.4	85.7	80.9	87.0	70.2	69.7	69.9	90.4	86.1	88.3	82.5
SigLIP ViT-B/16 (ft. CC3m)	95.8	87.4	73.5	85.6	65.3	74.2	69.7	88.9	87.0	87.9	81.7
C^{2}LIP	**96.7**	89.2	78.9	88.3	67.3	**78.8**	73.1	**93.5**	94.9	**94.2**	85.6

Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured Vision&Language concepts to vision&language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

[3] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Dense and aligned captions (DAC) promote compositional reasoning in VL models. In *Neural Information Processing Systems*, 2023. 1

[4] Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. SUGARCREPE++ dataset: vision-language model sensitivity to semantic and lexical alterations. In *Neural Information Processing Systems - Datasets and Benchmarks Track*, 2024. 2

[5] Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. FineCLIP: Self-distilled region-based CLIP for better fine-grained understanding. In *Neural Information Processing Systems*, 2024. 1

[6] Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mido Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. Modeling caption diversity in contrastive vision-language pretraining. In *International Conference on Machine Learning*, 2024. 1

[7] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 2024. 2

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2, 4

[9] Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. TripletCLIP: Improving compositional reasoning of CLIP via synthetic vision-language negatives. In *Neural Information Processing Systems*, 2024. 1

[10] Amit Peleg, Naman Deep Singh, and Matthias Hein. Advancing compositional awareness in CLIP with efficient fine-tuning. In *Neural Information Processing Systems*, 2025. 1

[11] Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn "no" to say "yes" better: Improving vision-language models via negations. In *Winter Conference on Applications of Computer Vision*, 2025. 1

[12] Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu,

Table A4. **SugarCrepe++ compositionality benchmark – all subtasks**. The results in this table are summarized in Tab. 2 of the main paper. We compare the accuracy of C^{2}LIP with baseline methods on different compositionality tasks. SugarCrepe++ addresses the limitation of the SugarCrepe benchmark, where the positive and negative captions are "hackable" by using custom training data created using similar rules. Here we observe significant performance drops from the baseline methods, while our model, which maintains generalization capabilities, achieves the best result overall.

Models	Replace - I2T				Replace - TOT				Swap - I2T			Swap - TOT			**Average**
	Obj	Attr	Rel	Avg	Obj	Attr	Rel	Avg	Obj	Attr	Avg	Obj	Attr	Avg	
SigLIP ViT-B/16	91.2	75.5	54.8	73.8	79.2	64.2	45.0	62.8	39.6	56.3	48.0	22.9	46.4	34.7	57.5
CLIP (OpenAI) ViT-B/32	86.7	65.6	56.3	69.5	83.7	59.3	38.6	60.5	46.1	45.2	45.7	19.1	35.6	27.4	53.6
CLIP (OpenAI) ViT-B/16	89.6	67.6	53.2	70.1	84.4	57.2	39.0	60.2	39.2	48.4	43.8	16.3	31.4	23.9	52.6
FG-CLIP	92.6	76.8	58.0	75.8	90.6	67.8	44.2	67.5	47.8	55.3	51.5	**29.4**	47.0	38.2	60.9
FineCLIP	90.8	70.7	57.0	72.8	91.3	67.9	47.0	68.7	39.6	48.2	43.9	20.8	32.7	26.8	56.6
DreamLIP-3m	75.8	60.8	46.9	61.2	71.4	50.8	32.4	51.5	34.3	54.4	44.4	20.0	40.1	30.1	48.7
FLAIR-3m	84.3	64.2	50.3	66.3	77.3	58.4	36.8	57.5	40.8	58.3	49.5	24.9	45.5	35.2	54.1
LLIP	84.1	66.0	51.9	67.3	71.2	54.8	44.5	56.8	36.7	44.9	40.8	24.5	30.2	27.3	50.9
Codebook-CLIP	32.3	32.6	37.8	34.3	18.2	9.8	13.5	13.8	28.2	28.5	28.3	8.6	11.6	10.1	22.1
IL-CLIP	38.0	32.9	32.4	34.4	55.8	18.5	21.3	31.8	39.2	34.7	36.9	9.4	20.0	14.7	30.2
CE-CLIP	71.9	52.0	45.5	56.5	86.3	64.2	50.5	67.0	36.3	32.3	34.3	28.6	36.3	32.5	50.4
NegCLIP	87.0	67.1	53.1	69.1	**93.3**	70.7	48.6	70.9	51.8	55.0	53.4	27.8	50.3	39.1	60.5
CLIC	91.6	75.9	62.3	76.6	84.7	52.4	37.3	58.1	**55.9**	62.2	**59.0**	22.9	31.2	27.0	57.6
DAC-SAM	64.3	44.3	48.7	52.4	75.9	56.0	48.7	60.2	27.8	33.5	30.6	11.4	25.4	18.4	43.6
DAC-LLM	65.7	47.7	47.6	53.7	76.8	59.5	42.3	59.6	31.4	32.9	32.2	11.4	24.8	18.1	44.0
SLVC-R	82.9	61.6	47.7	64.0	89.5	67.4	47.7	68.2	49.4	53.2	51.3	20.4	36.3	28.4	55.6
SLVC-RL	81.0	57.1	47.5	61.9	91.6	67.0	51.3	70.0	43.3	48.8	46.0	18.4	34.3	26.3	54.0
CoN_CLIP	87.9	68.1	48.2	68.1	91.5	64.7	53.9	70.1	40.0	50.3	45.2	18.8	37.2	28.0	56.1
TripletCLIP	84.9	66.0	58.7	69.9	89.0	71.1	48.1	69.4	38.4	4.4	21.4	18.8	38.1	28.5	51.7
SigLIP ViT-B/16 (ft. CC3m)	91.7	74.2	54.6	73.5	85.4	69.3	48.9	67.9	42.5	57.2	49.8	23.7	49.6	36.6	59.7
C^{2}LIP	**93.9**	**79.8**	**65.4**	**79.7**	91.1	**80.2**	**54.7**	**75.3**	44.1	**66.2**	55.2	28.6	**59.8**	**44.2**	**66.4**

Table A5. **Performance of VLM using C^{2}LIP vision encoder.** We follow the LLaVa [8] recipe to tune VLM with frozen SigLIP and C^{2}LIP image encoders, resulting in LLaVA-SigLIP and LLaVA-C^{2}LIP, respectively. LLaVA-C^{2}LIP shows improvements on compositionality, demonstrating that the vision encoder trained with our proposed concept-centric approach also benefits the VLM that utilizes it.

Model	SugarCrepe				SugarCrepe++		
	Replace	Swap	Add	**Average**	Replace	Swap	**Average**
SigLIP-ViT-B-16	84.1	65.8	86.5	79.5	73.8	48.0	63.5
C^{2}LIP	**89.1**	75.2	**93.3**	**86.3**	**79.7**	55.2	**69.9**
LLaVA-SigLIP	86.9	77.0	85.0	83.5	76.5	55.4	68.1
LLaVA-C^{2}LIP	86.8	**78.5**	85.0	83.9	77.6	**57.8**	69.7

Zeynep Akata, and Stephan Alaniz. FLAIR: Vlm with fine-grained language-informed image representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025. 1

[13] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. FG-CLIP: Fine-grained visual and textual alignment. In *International Conference on Machine Learning*, 2025. 1

[14] Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. 1

[15] Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 1

[16] Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Iterated learning improves compositionality in large vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 1

[17] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, 2024. 1

Table A6. **Zero-shot retrieval benchmarks – all subtasks**. The results in this table are summarized in Tab. 2 of the main paper. We record Recall@5 metrics of all models on text-to=image (t2i) and image-to-text (i2t) tasks, across two standard retrieval benchmarks: MSCOCO and Flickr30K, and two fine-grained retrieval benchmarks: DOCCI and Image-in-words (IIW). It can be observed that C^2LIP can maintain or improve the performance of the original base model across all tasks, with the best average score.

Models	MSCOCO		Flickr30k		DOCCI		IIW		**Average**
	t2i	i2t	t2i	i2t	t2i	i2t	t2i	i2t	
SigLIP ViT-B/16	72.4	85.4	92.3	98.0	35.8	82.0	53.0	97.5	77.1
CLIP (OpenAI) ViT-B/32	56.0	74.9	83.4	94.6	27.1	67.1	44.7	94.3	67.8
CLIP (OpenAI) ViT-B/16	58.4	76.7	85.6	96.2	29.3	71.5	46.7	95.9	70.0
FG-CLIP	71.4	85.5	93.0	98.6	33.9	79.4	52.5	**98.7**	76.6
FineCLIP	73.7	84.2	90.2	96.7	27.9	61.9	44.6	89.5	71.1
DreamLIP-3m	55.2	67.2	76.6	89.6	39.3	78.2	**58.5**	97.4	70.2
FLAIR-3m	65.6	77.3	86.5	94.3	30.9	63.4	53.0	92.7	70.5
LLIP	62.3	72.8	87.3	93.1	28.6	65.2	48.3	93.1	68.8
Codebook-CLIP	0.1	0.1	0.5	0.6	0.1	0.1	0.8	0.5	0.3
IL-CLIP	0.1	0.1	0.3	0.5	0.1	0.1	0.7	0.7	0.3
CE-CLIP	69.5	74.3	86.4	88.4	19.1	42.4	31.4	68.8	60.0
NegCLIP	68.4	79.3	89.5	95.2	26.4	64.0	43.5	89.4	69.5
CLIC	62.9	71.9	88.2	94.0	33.1	69.4	52.2	94.6	70.8
DAC-SAM	59.7	57.9	85.5	82.5	26.9	35.5	45.5	55.7	56.2
DAC-LLM	63.5	54.5	87.8	79.6	24.8	29.4	40.9	46.9	53.4
SLVC-R	62.0	71.7	87.2	93.1	29.9	54.3	48.6	83.7	66.3
SLVC-RL	62.3	71.8	87.4	92.5	29.7	53.9	48.0	86.6	66.5
CoN-CLIP	54.6	67.9	84.2	87.9	27.2	64.9	42.6	91.5	65.1
TripletCLIP	53.3	55.6	80.9	82.6	25.3	54.2	43.0	82.0	59.6
SigLIP ViT-B/16 (ft. CC3m)	73.7	87.0	92.8	98.4	36.2	**83.3**	53.0	97.9	77.8
C^2LIP	**77.9**	**87.5**	**95.2**	**98.8**	**38.0**	81.9	56.1	96.7	**79.0**

Table A7. **Zero-shot classification benchmarks**. We evaluate classification accuracies of C^2LIP and the baselines on 11 standard zero-shot classification benchmarks. We observe significant performance gaps between our model and the composition-aware baselines in general, particularly on the challenging *Cars* and *Aircraft* benchmarks where the classes are actually sub-classes of the same object category.

Model	ImageNet1K	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	**Average**
SigLIP ViT-B/16	**76.1**	**91.6**	92.3	72.2	69.9	**90.9**	43.8	64.7	94.1	**88.0**	86.0	79.0
CLIP (OpenAI) ViT-B/32	63.3	83.9	89.8	64.3	63.2	59.7	19.6	44.0	87.5	83.8	66.5	66.0
CLIP (OpenAI) ViT-B/16	68.4	88.8	90.8	67.0	65.5	64.7	24.5	45.2	89.2	84.0	71.5	69.0
FG-CLIP	69.0	85.2	93.9	**76.4**	**71.2**	84.2	24.4	57.2	90.5	86.2	70.1	73.5
FineCLIP	55.8	60.1	94.3	69.0	55.9	6.3	10.7	41.6	57.9	85.4	41.4	52.6
DreamLIP-3m	31.6	23.6	75.7	43.7	41.3	3.5	1.6	18.8	28.9	70.3	18.5	32.5
FLAIR-3m	33.7	24.8	81.9	51.6	47.0	4.0	2.0	24.3	35.8	72.4	20.4	36.2
LLIP	60.8	80.5	**94.4**	69.8	57.7	77.1	22.3	53.3	78.8	85.3	42.9	65.7
Codebook-CLIP	0.1	1.0	11.0	1.1	0.4	0.5	1.1	2.0	2.2	0.9	1.2	1.9
IL-CLIP	0.1	1.0	10.0	1.0	0.1	0.5	1.0	1.6	2.7	0.9	1.1	1.8
CE-CLIP	40.4	59.8	81.2	55.0	44.0	26.1	9.2	28.5	60.9	76.0	37.3	47.1
NegCLIP	55.7	74.1	85.9	60.9	55.9	46.0	11.8	39.1	82.3	82.5	58.0	59.3
CLIC	66.6	88.9	91.1	68.3	64.0	60.8	23.7	46.7	88.0	84.0	67.5	68.1
DAC-SAM	52.3	72.3	89.9	63.7	51.4	39.8	9.0	40.2	77.0	78.0	54.2	57.1
DAC-LLM	51.1	74.5	90.4	63.9	52.1	39.5	11.3	38.5	74.9	79.8	54.9	57.3
SLVC-R	58.5	81.3	92.3	66.0	62.6	49.6	14.9	39.4	85.8	81.9	59.7	62.9
SLVC-RL	59.8	81.7	92.0	66.7	63.7	50.6	14.5	39.8	85.0	82.7	61.3	63.4
CoN-CLIP	63.7	84.5	88.7	63.0	64.0	55.5	19.0	40.4	85.2	83.3	63.3	64.6
TripletCLIP	45.9	58.7	86.9	56.6	53.6	11.7	7.9	33.1	55.2	78.3	45.2	48.5
SigLIP ViT-B/16 (ft. CC3m)	75.9	**91.6**	92.4	72.7	70.3	90.8	**44.4**	**65.1**	**94.4**	**88.0**	**86.1**	**79.2**
C^2LIP	73.5	88.7	92.7	72.6	68.1	87.4	32.8	**65.1**	92.3	87.2	82.9	76.7