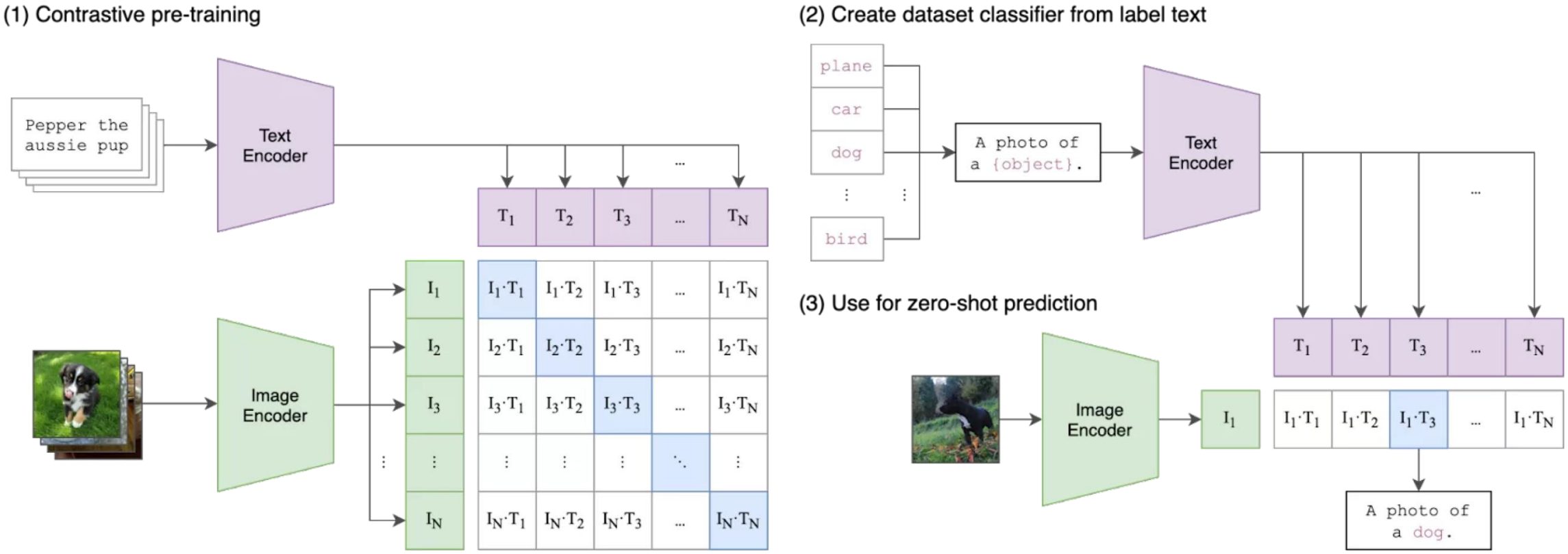
TripletCLIP: Improving Compositional Reasoning of CLIP via Synthetic Vision-Language Negatives

Maitreya Patel, Abhiram Kusumba*, Sheng Cheng*, Changhoon Kim,
Tejas Gokhale, Chitta Baral, Yezhou Yang

1 Introduction to CLIP and Compositionality

CLIP Overview



$$\mathcal{L}_{\mathcal{X}\to\mathcal{Y}}^{CL} = \frac{-1}{N}\sum_{i=1}^{N}\log\frac{\exp(\langle F_{\mathcal{X}}(x_i), F_{\mathcal{Y}}(y_i)\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle F_{\mathcal{X}}(x_i), F_{\mathcal{Y}}(y_k)\rangle/\tau)},$$

Radford et al. "Learning transferable visual models from natural language supervision." ICML, 2021.

Compositionality Challenges

- CLIP models are the cornerstone of recent advancements in Generative AI.
  - From Text-to-Image/Video /3D synthesis to Vision-LLMs.
- However, the core building block of our recent advancements itself is fundamentally flawed.
- CLIP models despite being trained billions of image-text pairs they lack simple compositionality understanding.

Thrush et. al. "Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality," CVPR 2022.

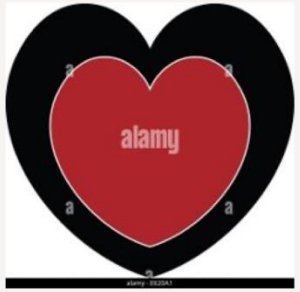
What kind of synthetic data is needed? And how can we utilize it more effectively?

2 TripletData

Synthetic Hard Negatives at Scale!

Synthetic Dataset Pipeline

- We propose to use the synthetic image-caption negatives.
- Step 1: Large Language Models (LLMs) to generate hard negatives of the input captions.
- Step 2: Stable Diffusion Turbo to synthesize the image corresponding to the negative captions.



"Heart trimmed in white on top of black frame."

"Heart trimmed in red on bottom of silver frame."

"A cartoon illustration of a stressed man taking cover from gunfire."

"A cartoon illustration of a peaceful man hiding from rain."

"man breaks free from a tackle during the football game."

"man gets tackled during the basketball game."

"A patterned background with a vintage color scheme."

"A solid-colored background with a futuristic color scheme."

"small peony in the garden."

"large peony in the greenhouse."

"A photo of a map with pins or a stack of pins."

"A photo of a map with marbles or a stack of marbles."

Data Quality Checks

- We evaluate the SOTA models on the subset of TripletData like Winoground style.
- We find that despite being synthetic TripletData is very challenging for the SOTA CLIP models.
- Additionally, from Table 2, we can observe that TripletData maintains the unique # of concept synsets.
- Hence, it does not focus on adding more concept diversity but instead focuses on the diversity of compositionality.

Table 1: Winoground-style evaluation of pretrained CLIP models on `TripletData`.

	Img Score	Text Score	Grp Score
ViT-B/32	40.29	68.17	36.53
ViT-L/14	44.84	69.21	40.91
ViT-bigG	42.94	77.61	40.98
Siglip-so400m	44.24	71.27	26.10
Humans (on Winoground)	**88.50**	**89.50**	**85.50**

Table 2: Wordnet synset analysis of captions from CC3M and `TripletData`.

	CC3M	TripletData (Negative Only)	TripletData	Intersection
# unique	59094	59616	62741	**55969**
# total synsets	231M	215M	446M	-

A quick review:

1. We propose a straightforward yet effective hard negative data generation pipeline at scale.

2. We release 13M synthetic negatives data (TripletData) to complement CC3M/12M for the community to build upon.

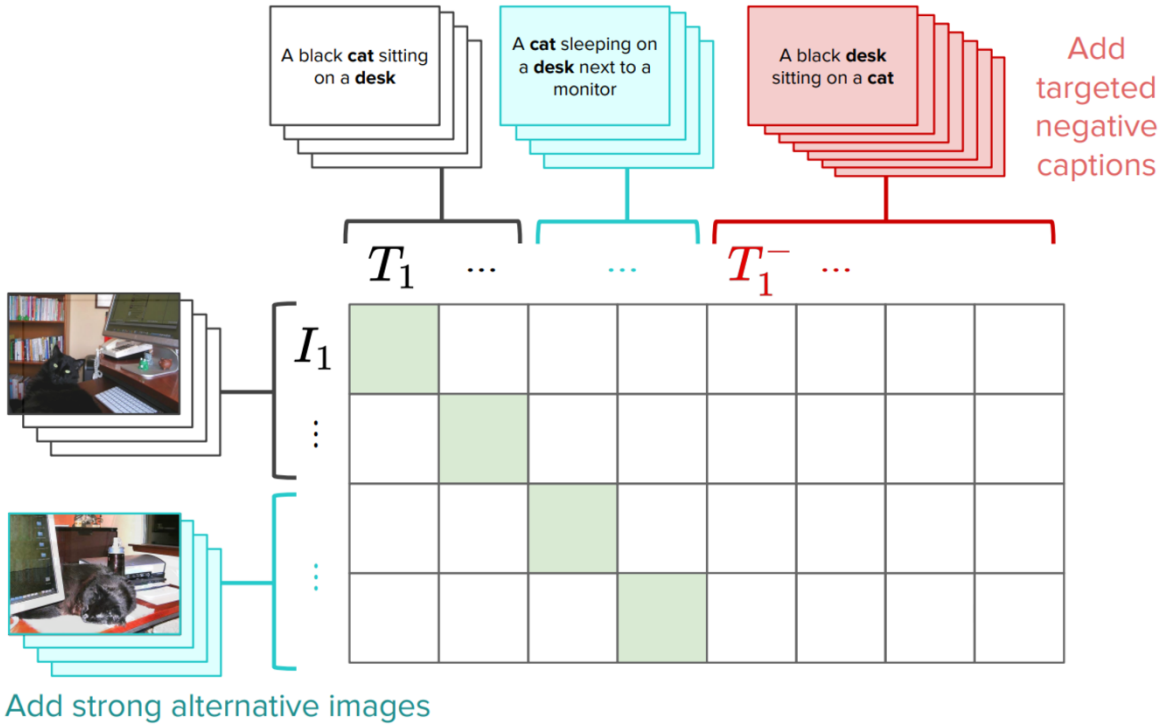
But how to utilize this data effectively?

3 TripletCLIP

NegCLIP

Empirical evidence shows that having hard negative images for loss (similar to NegCLIP) only leads to suboptimal performance.

Hard negative can be extracted either by mining or synthetic augmentations.



$$\mathcal{L}^{NegCL}_{\mathcal{X}\to\mathcal{Y};\mathcal{Y}'} = \frac{-1}{N}\sum_{i=1}^{N}\log\frac{\exp(\langle F_{\mathcal{X}}(x_i), F_{\mathcal{Y}}(y_i)\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle F_{\mathcal{X}}(x_i), F_{\mathcal{Y}}(y_k)\rangle/\tau) + \sum_{m=1}^{N}\exp(\langle F_{\mathcal{X}}(x_i), F_{\mathcal{Y}}(y'_m)\rangle/\tau)}.$$

$$\mathcal{L}_{NegCLIP}(\mathcal{X},\mathcal{Y},\mathcal{Y}') = \mathcal{L}^{CL}_{\mathcal{Y}\to\mathcal{X}} + \mathcal{L}^{NegCL}_{\mathcal{X}\to\mathcal{Y};\mathcal{Y}'}.$$

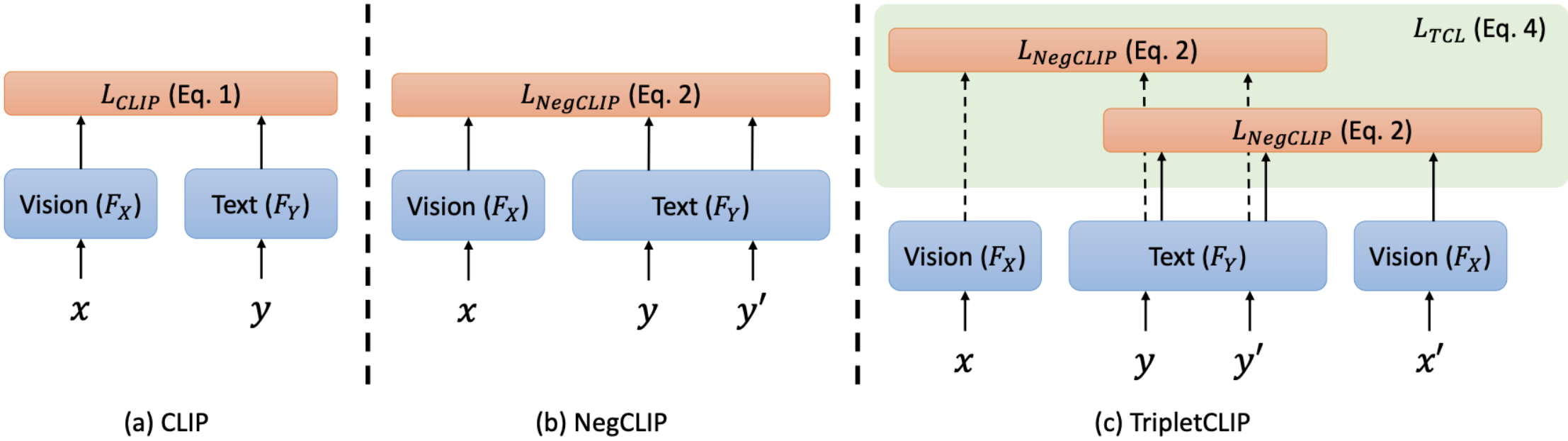
Importance of hard negative contrastive loss

Table 3: **Importance of image-text hard negatives.** We measure the importance of various modality-specific hard negatives on SugarCrepe, image-text retrieval, and ImageNet1k. We find that `Triplet-CLIP` results into the most optimal solution. **Bold** number indicates the best performance.

Models	Negative Captions	Negative Images	SugarCrepe	Retrieval	ImageNet1k
LaCLIP	×	×	54.09	8.19	3.79
NegImage	×	✓	56.28	9.20	4.48
NegCLIP++	✓	×	61.69	8.36	3.84
TripletCLIP	✓	✓	**63.49**	**16.42**	**7.31**

Empirical evidence shows that having hard negative images for loss (similar to NegCLIP) only leads to suboptimal performance. Therefore, we propose a better strategy to utilize the negative images.

Proposed Approach: TripletLoss



(a) CLIP (b) NegCLIP (c) TripletCLIP

$$\mathcal{L}_{TCL} = \mathcal{L}_{NegCLIP}(\mathcal{X}, \mathcal{Y}, \mathcal{Y}') + \mathcal{L}_{NegCLIP}(\mathcal{X}', \mathcal{Y}', \mathcal{Y}).$$

4 Results

Pretraining Setup

Table 10: Detailed pre-training hyper-parameters for CLIP training across various experiments and ablations.

Hyperparameters	CC3M	CC12M	LiT	Concept Coverage Ablations
Batch size	1024	1024	1024	1024
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}
Weight decay	0.5	0.5	0.5	0.5
Adam β	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Adam ϵ	1×10^{-8}	1×10^{-8}	1×10^{-8}	1×10^{-8}
Total steps	90,000	230,000	90,000	200,000
Learning rate schedule	cosine decay	cosine decay	cosine decay	cosine decay

We perform pretraining and finetuning experiments from scratch (incl. baselines) for comprehensive evaluations.

Compositionality Evals: SugarCrepe

Table 4: **Composition evaluations of the methods on SugarCrepe benchmark. Bold** number indicates the best performance and underlined number denotes the second-best performance. † represents the results taken from SugarCrepe benchmark.

	Methods	Replace			Swap		Add		Overall
		Object	Attribute	Relation	Object	Attribute	Object	Attribute	Avg.
CC3M	**LaCLIP**	59.44	53.17	51.42	54.69	49.25	55.29	55.35	54.09
CC3M	**LaCLIP + HN**	63.44	55.96	50.71	50.60	48.57	56.98	51.16	53.92
CC3M	**NegCLIP**	62.71	58.12	54.48	**56.33**	51.20	56.26	61.13	57.18
CC3M	**NegCLIP++** *(ours)*	64.77	66.12	**65.93**	55.51	55.41	59.65	**64.45**	61.69
CC3M	**TripletCLIP** *(ours)*	**69.92**	**69.03**	64.72	**56.33**	**57.96**	**62.61**	63.87	**63.49**
	Performance Gain *w.r.t.* LaCLIP	**10.48%**	**18.56%**	**13.30%**	**1.64%**	**8.71%**	**7.32%**	**8.52%**	**9.40%**
CC12M	**LaCLIP**	75.06	65.48	58.68	53.47	57.66	67.65	66.76	63.54
CC12M	**NegCLIP**	77.84	69.29	63.23	**66.53**	62.31	67.17	69.65	68.00
CC12M	**NegCLIP++** *(ours)*	82.99	78.68	75.75	61.63	**65.47**	70.08	**76.01**	72.94
CC12M	**TripletCLIP** *(ours)*	**83.66**	**81.22**	**79.02**	64.49	63.66	**73.67**	75.43	**74.45**
	Performance Gain *w.r.t.* LaCLIP	**8.60%**	**15.75%**	**20.34%**	**11.02%**	**6.00%**	**8.67%**	**7.35%**	**10.91%**
DataComp	`small`: **ViT-B/32**† (13M)	56.90	56.85	51.99	50.81	50.00	53.93	60.55	54.43
DataComp	`medium`: **ViT-B/32**† (128M)	77.00	69.54	57.68	57.72	57.06	66.73	64.88	64.37
DataComp	`large`: **ViT-B/16**† (1B)	**92.68**	79.82	63.94	56.10	57.66	**84.34**	**78.61**	73.31
DataComp	`xlarge`: **ViT-L/14**† (13B)	**95.52**	**84.52**	69.99	**65.04**	**66.82**	**91.03**	**84.97**	**79.70**

Traditional Evals

Table 5: **Zero-shot image-text retrieval and classification results. Bold** number indicates the best performance and underlined number denotes the second-best performance.

	Methods	Retrieval (R@5)				Zero-shot Classification			
		Image-to-Text		Text-to-Image		VTAB		ImageNet1k	
		MSCOCO	Flickr30k	MSCOCO	Flickr30k	top-1	top-5	top-1	top-5
CC3M	**LaCLIP**	5.06	10.90	5.97	10.84	11.56	34.72	3.79	10.49
CC3M	**LaCLIP + HN**	8.08	16.10	8.64	16.64	**12.31**	37.14	5.75	15.22
CC3M	**NegCLIP**	6.32	13.80	6.61	12.96	12.25	36.38	4.67	12.69
CC3M	**NegCLIP++** *(ours)*	5.8	11.20	6.19	10.24	11.65	35.47	3.84	10.52
CC3M	**TripletCLIP** *(ours)*	**10.38**	**22.00**	**11.28**	**22.00**	**12.31**	**41.45**	**7.32**	**18.34**
CC3M	**Performance Gain**	**5.32%**	**11.1%**	**5.31%**	**11.16%**	**0.75%**	**6.73%**	**3.53%**	**7.85%**
CC12M	**LaCLIP**	25.86	42.70	19.78	36.30	19.08	49.06	19.72	41.39
CC12M	**NegCLIP**	30.16	46.60	23.11	41.70	19.12	50.56	20.22	42.63
CC12M	**NegCLIP++** *(ours)*	26.96	43.90	22.69	42.86	18.48	50.38	19.06	40.91
CC12M	**TripletCLIP** *(ours)*	**33.00**	**55.90**	**28.50**	**52.38**	**20.81**	**53.40**	**23.31**	**47.33**
CC12M	**Performance Gain**	**7.14%**	**13.2%**	**8.72%**	**16.08%**	**1.73%**	**4.34%**	**3.59%**	**5.94%**

Finetuning Evals

Table 7: **Finetuning-based composition evaluations of the methods on SugarCrepe benchmark. Bold** number indicates the best performance and underlined number denotes the second-best performance.

Methods	Replace			Swap		Add		Overall
	Object	Attribute	Relation	Object	Attribute	Object	Attribute	Avg.
CLIP	90.92	80.08	69.13	61.22	64.26	77.16	68.64	73.06
CLIP (finetuned)	90.92	79.69	64.01	60.82	64.26	84.67	78.76	74.73
NegCLIP	91.53	83.25	73.97	72.24	67.72	86.95	88.44	80.59
Baseline [44]	93.22	84.39	67.35	62.04	70.12	88.31	79.48	77.84
CoN-CLIP [50]	93.58	80.96	63.3	**87.29**	**79.62**	59.18	65.16	75.58
TSVLC (RB) [12]	91.34	81.34	64.15	68.16	69.07	79.49	91.33	77.84
TSVLC (LLM+RB) [12]	88.13	76.78	62.73	64.08	66.67	75.80	81.07	73.61
DAC [11]	**94.43**	**89.48**	**84.35**	75.10	74.17	89.67	**97.69**	**86.41**
TripletCLIP *(ours)*	**94.43**	85.53	80.94	69.80	69.82	**90.40**	86.27	82.46

Filtering high-quality TripletData

Table 6: **Ablation on filtering high-quality image-text pairs from `TripletData`.** We evaluate the `TripletCLIP` after applying the filters to ensure the quality similar to DataComp and compare the baselines on three benchmarks. We find that `TripletCLIP` results in the most optimal solution. **Bold** number indicates the best performance. † represents that results are borrowed from DataComp.

Models	Filtering Strategy	Data Size	Augmentations	SugarCrepe	Retrieval	ImageNet1k
CLIP†	No filtering	12.8	-	55.61	6.49	2.7
	CLIP Score	3.8	-	57.31	9.08	5.1
	Image-based ∩ CLIP Score	1.4	-	54.75	5.63	3.9
LaCLIP	No filtering (CC3M)	2.6	-	54.09	8.19	3.79
TripletCLIP	No filtering (CC3M)	2.6	2.6	63.49	16.42	7.31
TripletCLIP++	CLIP Score (from CC12M)	1.4	1.4	**66.09**	**19.85**	**8.85**

What’s holding back CLIP models?

Vision Encoder!

Table 8: **Frozen encoder ablation.** LiT style fine-tuning ablations on SugarCrepe, image-text retrieval, and ImageNet1k. **Bold** number indicates the best performance.

Models	Train Text	Train Vision	SugarCrepe	Retrieval	ImageNet1k
LaCLIP	✓	×	**0.6373**	0.5345	31.21%
TripletCLIP *(ours)*	✓	×	0.6227	**0.6817**	**34.25%**
LaCLIP	×	✓	0.5886	0.1134	5.51%
TripletCLIP *(ours)*	×	✓	**0.6923**	**0.2626**	**12.51%**

- Vision encoder is also important for the compositionality.
- So, it is important for future works to consider this modality!

Limitations & Future Works

In academic setting, we perform small-scale yet comprehensive evaluations. However, future work should scale this to wide varity of models and tasks.

LLMs and T2I models will limit the diversity of the of synthetic datasets. Hence, alternative solutions are needed.

Synthetic data creation is resource-intensive task.

Future works should focus on finding the representation level solutions.

We release our code, data, and weights for the open-source community!

Thank you!

maitreyapatel.com

 patelmaitreya