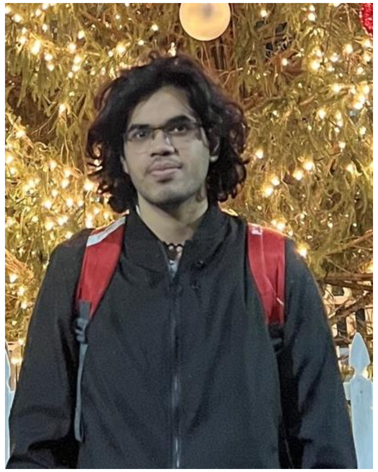
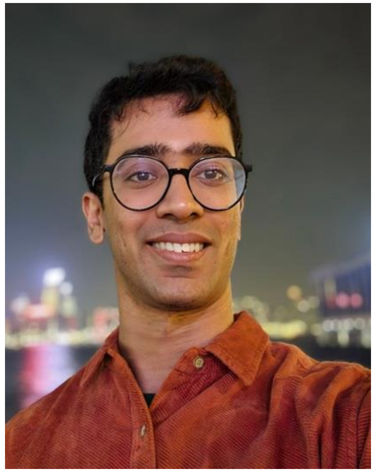
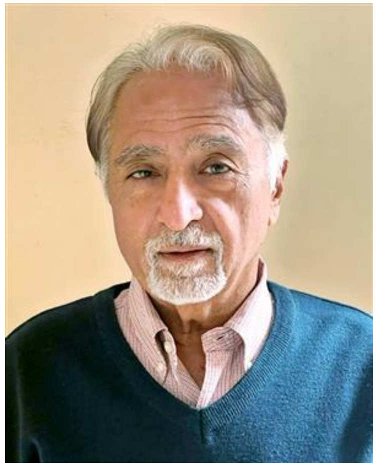
PositiveCoOp: Rethinking Prompting Strategies for Multi-Label Recognition with Partial Annotations

Samyak Rawlekar, Shubhang Bhatnagar, Narendra Ahuja

WACV 2025

Problem: Multi-Label Recognition (MLR) with Partial Labels

Categories	Complete Labels	Partial Labels
Computer	✓	✓
Speakers	✓	?
Oven	✗	?
Desk	✓	✓
Books	✓	?
Lamp	✓	✓
Traffic Lights	✗	✗

Multi-Label Recognition (MLR): The task involves identifying all the objects present in an image

MLR with Partial Labels:

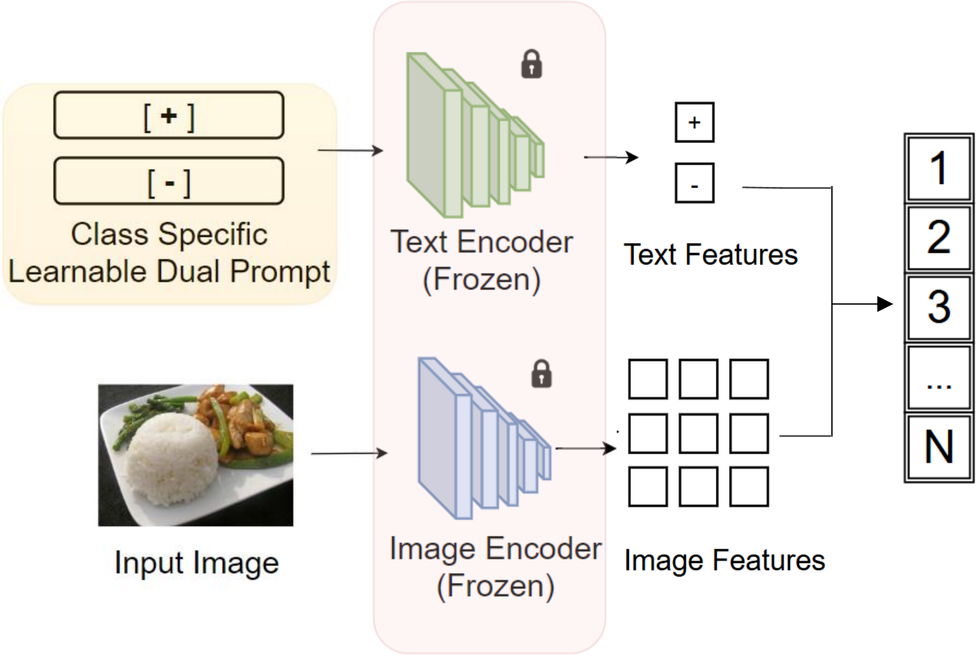
- **Training:** In real-world MLR datasets, not all objects in an image are annotated
- **Inference:** Our goal is to correctly identify all the classes present in the image

Recent Work in MLR with Partial Annotations

Vision-Language Models for MLR

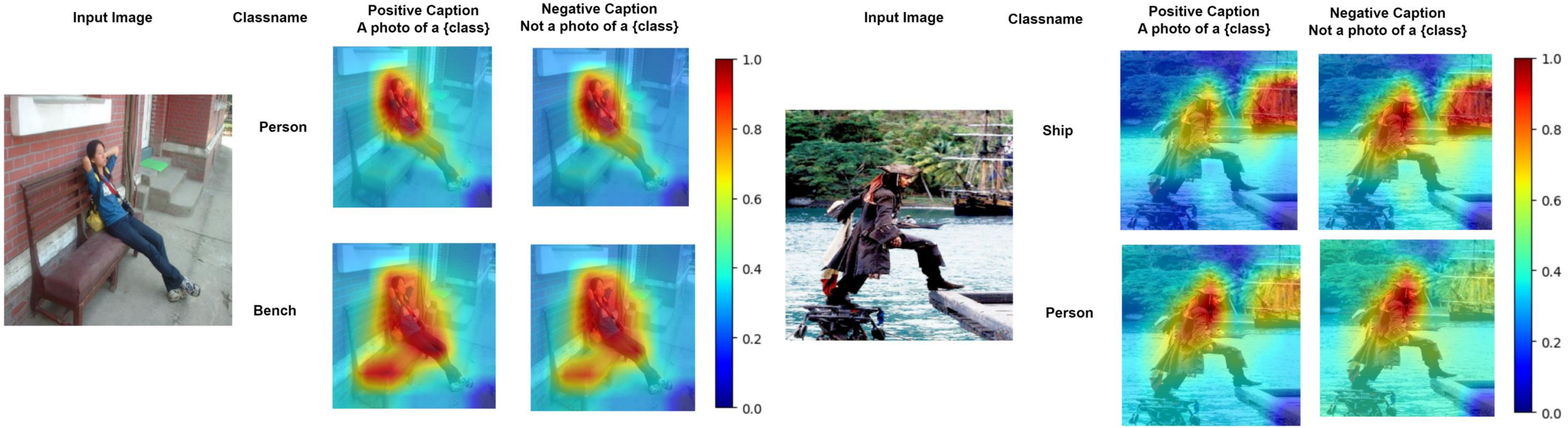
Recent work addresses challenges in MLR by:

- Adapt information from pretrained vision language models (e.g. CLIP [1])
- To preserve the feature extraction priors, these models are kept frozen
- Learnable positive and negative text prompts are then used as classifiers on the image features
- The positive prompt detect the presence and negative prompt detect class absence [2]



[1] Radford et al " Learning transferable visual models from natural language supervision." *ICML* (2021)
[2] Sun et al "Dualcoop: Fast adaptation to multi-label recognition with limited annotations." *NIPS* (2022)

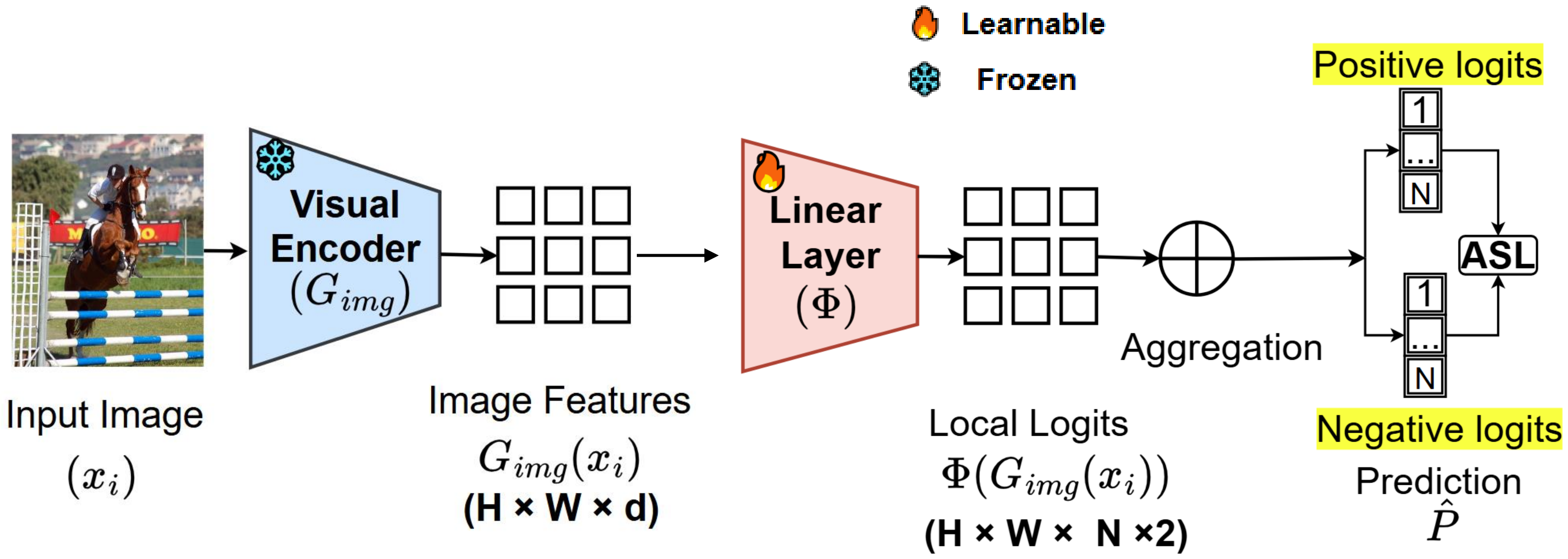
Prompting with VLMs



- **Similarity Map Visualization**: We analyze the similarity maps of CLIP image features with both positive and negative prompt for a given class
- **Activated Regions**: Both captions activate regions that corresponds to the presence of object

Are Negative Prompts Truly Analyzing Features Related to Class Absence ?

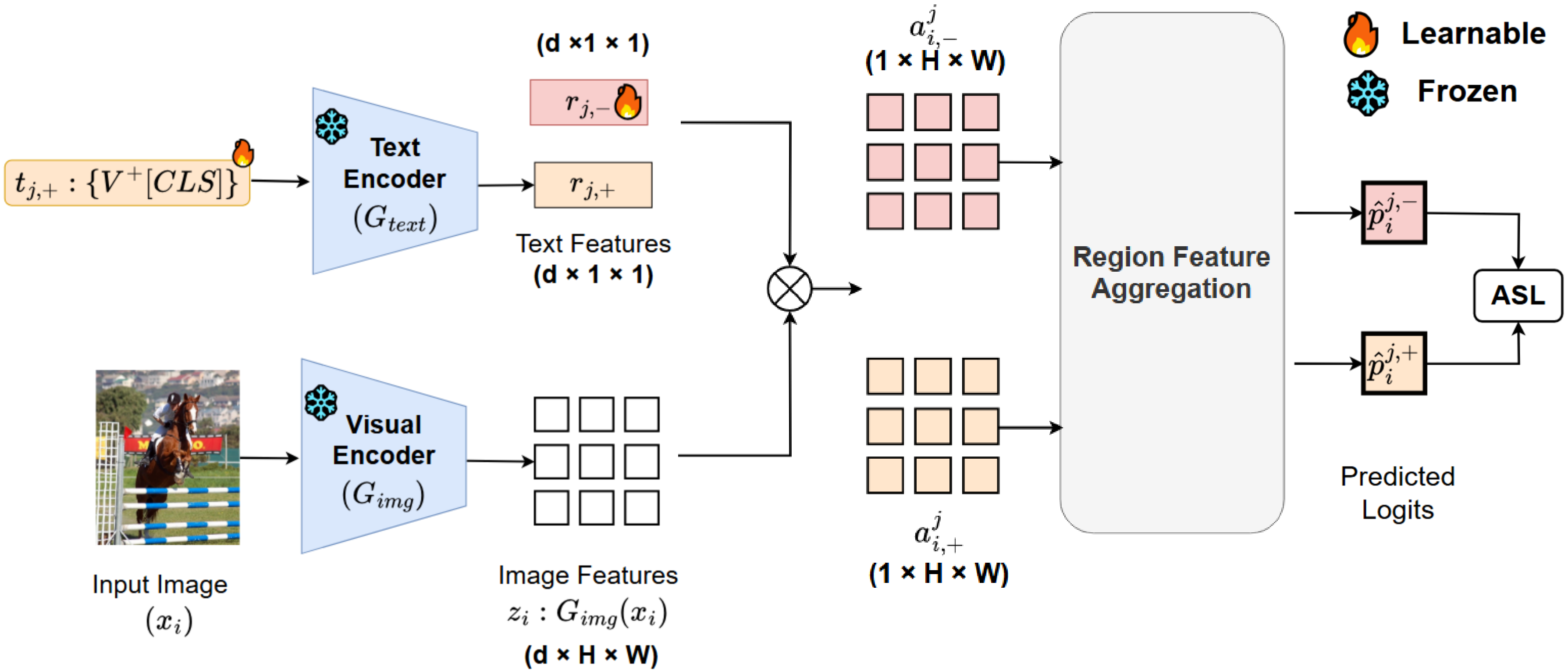
Baseline



Prior VLM based MLR works do not compare with such a vision only baseline

Baseline relies solely on CLIP visual features and helps estimate the impact of different prompting strategies

PositiveCoOp (NegativeCoOp)



PositiveCoOp

Class presence features: Learn positive prompt

Class absence features: Learn negative embeddings in feature space

NegativeCoOp

Class presence features: Learn positive embeddings in feature space

Class absence features: Learn negative prompt

Performance Evaluation – COCO

Methods	#Params	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg.
SSGRL	64.7M	62.5	70.5	73.2	74.5	76.3	76.5	77.1	77.9	78.4	74.1
GCN-ML	44.9M	63.8	70.9	72.8	74.0	76.7	77.1	77.3	78.3	78.6	74.4
KGGR	$\geq$ 25M	66.6	71.4	73.8	76.7	77.5	77.9	78.4	78.7	79.1	75.6
CL	$\geq$ 38M	26.7	31.8	51.5	65.4	70.0	71.9	74.0	77.4	78.0	60.7
Partial BCE	$\geq$ 38M	61.6	70.5	74.1	76.3	77.2	77.7	78.2	78.4	78.5	74.7
SST	33.5M	68.1	73.5	75.9	77.3	78.1	78.9	79.2	79.6	79.9	76.7
SARB	29.6M	71.2	75.0	77.1	78.3	78.9	79.6	79.8	80.5	80.5	77.9
SST*	33.5M	69.1	78.5	79.3	79.9	80.1	80.5	81.1	80.7	80.7	78.9
SARB*	29.6M	75.5	78.5	79.0	79.5	80.4	80.2	80.8	80.6	80.8	79.4
DualCoOp	1.3M	78.7	80.9	81.7	82.0	82.5	82.7	82.8	83.0	83.1	81.9
SCPNet	3.4M	**80.3**	**82.2**	82.8	83.4	**83.8**	83.9	84.0	84.1	84.2	**83.2**
Baseline	80k	78.9	80.6	81.3	81.9	82.7	82.8	82.9	83.2	83.5	82.0
NegativeCoOp	730k	77.8	80.3	81.0	81.9	82.2	82.4	82.7	82.8	82.9	81.6
PositiveCoOp	730k	79.8	82.1	**83.0**	**83.5**	83.7	**83.9**	**84.0**	**84.2**	**84.4**	**83.2**

Comparison of Baseline, PositiveCoOp, and NegativeCoOp with SOTA methods on COCO

Performance Evaluation – VOC2007

Methods	#Params	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg.
SSGRL	66.6M	77.7	87.6	89.9	90.7	91.4	91.8	91.9	92.2	92.2	89.5
GCN-ML	44.9M	74.5	87.4	89.7	90.7	91.0	91.3	91.5	91.8	92.0	88.9
KGGR	$\geq$ 25M	81.3	88.1	89.9	90.4	91.2	91.3	91.5	91.6	91.8	89.7
CL	$\geq$ 38M	44.7	76.8	88.6	90.2	90.7	91.1	91.6	91.7	91.9	84.1
Partial BCE	$\geq$ 38M	80.7	88.4	89.9	90.7	91.2	91.8	92.3	92.4	92.5	90.0
SST	32.4M	81.5	89.0	90.3	91.0	91.6	92.0	92.5	92.6	92.7	90.4
SARB	29.6M	83.5	88.6	90.7	91.4	91.9	92.2	92.6	92.8	92.9	90.7
DualCoOp	0.3M	90.3	92.2	92.8	93.3	93.6	93.9	94.0	94.1	94.2	93.2
SCPNet	-	91.1	92.8	**93.5**	93.6	93.8	94.0	94.1	94.2	94.3	93.5
Baseline	20k	90.5	92.2	92.8	93.0	93.3	93.8	93.9	94.0	94.2	93.1
Negative CoOp	170k	88.9	89.3	89.6	89.9	90.7	91.2	91.8	92.1	92.4	90.8
Positive CoOp	170k	**91.4**	**92.8**	93.4	**93.6**	**93.8**	**94.0**	**94.2**	**94.2**	**94.3**	**93.6**

Across the 10%-90% partial available labels, the performance order is :
PositiveCoOp > DualCoOp ≈ Baseline > NegativeCoOp.

Negative Prompting Hurts MLR !

Computation Comparison

Dataset	Method	#Params	GPU Hours
VOC	DualCoOp	0.3M	3.55
	SCPNet	-	3
	Baseline	20k	1.5
	NegativeCoOp	0.17M	3
	PositiveCoOp	0.17M	3
COCO	DualCoOp	1.3M	16
	SCPNet	3.4M	26
	Baseline	80k	7.97
	NegativeCoOp	0.73M	16
	PositiveCoOp	0.73M	16

Comparison of training parameters and GPU hours of the three setups with SOTA.

Baseline uses fewer parameters and GPU hours than all others, while PositiveCoOp and NegativeCoOp require about half the parameters of DualCoOp

Why Negative Prompt Learning is Ineffective?

The LAION dataset contains about 2 million captions (0.47% of 400 million) that include a negative word.

- CLIP may fail to distinguish between positive and negative prompts
- Too few negative captions for it to learn this!

Cosine Similarity (80 cls-1 prompt)	P1:'photo of a{}' N1:'Not a photo of a {}'	P1:'photo of a{}' P2:'picture of a {}'
Mean $\pm$ Std (Min,Max)	0.58 ± 0.06 (0.37, 0.69)	0.53 ± 0.04 (0.51, 0.67)

To test empirically, we calculate cosine similarity between:

- Positive-positive feature pairs
- Positive-negative feature pairs

Cosine Similarity (80cls-85prompt)	P1-N1 Pairs	P1-P2 Pairs
Mean $\pm$ Std (Min, Max)	0.56 ± 0.06 (0.37, 0.67)	0.61 ± 0.01 (0.55, 0.63)

Results indicate that CLIP projects positive and negative prompts very closely in the feature space

Conclusion

- We investigated the impact of positive and negative prompts in VLM-based multi-label recognition with partial annotations
- Our ablations (PositiveCoOp and NegativeCoOp) show that learning only positive prompts while using learned negative embeddings outperforms dual prompt learning approaches
- Our analysis of LAION-400M suggests that the absence of negative prompts in large-scale pretraining data contributes to the poor performance of negative prompting
- In settings with fewer missing labels, a vision-features-only baseline performs strongly while being significantly more computationally efficient

Visit Our Project Page