SPARC: Score Prompting and Adaptive Fusion for Zero-Shot Multi-Label Recognition in Vision-Language Models

Kevin Miller
Boston University
nivek@bu.edu

Aditya Gangrade
Boston University
gangrade@bu.edu

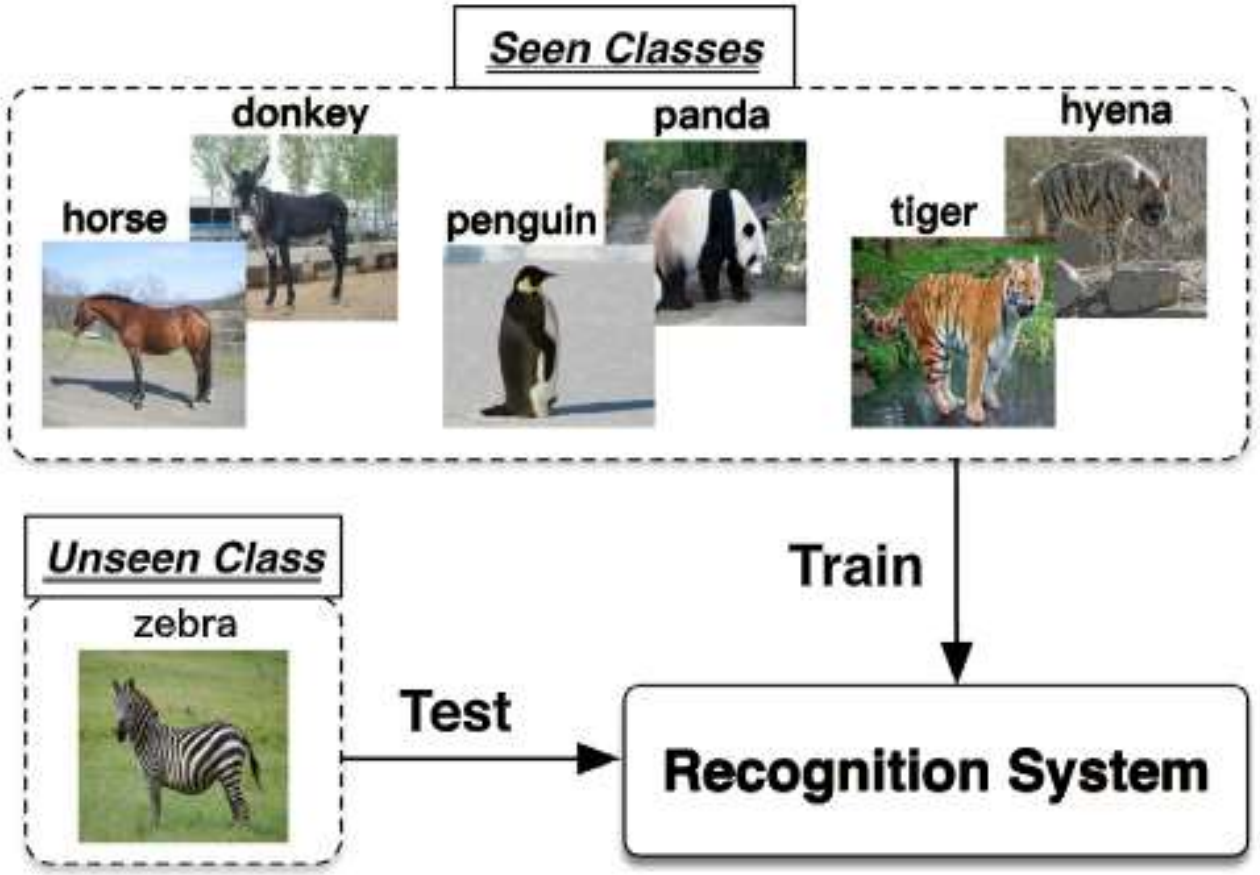
Samarth Mishra
Boston University
samarthm@bu.edu

Kate Saenko
Boston University and Meta AI (FAIR)
saenko@bu.edu

Venkatesh Saligrama
Boston University
srv@bu.edu

CVPR 2025

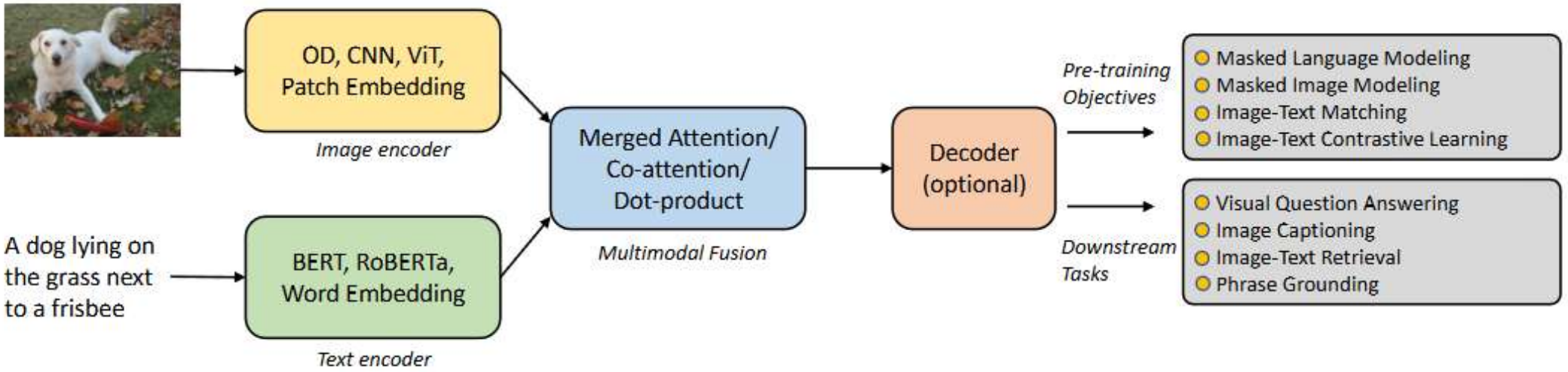
Zero-shot Learning



Multi-Label Recognition

	Multi-Class	Multi-Label
Image		
Labels	cat ✓ dog hamster rabbit fish	✓ dog long-haired ✓ glasses ears up ✓ collar

Vision-Language Models



Motivation

VLMs like CLIP Struggle in Multi-Label Scenario

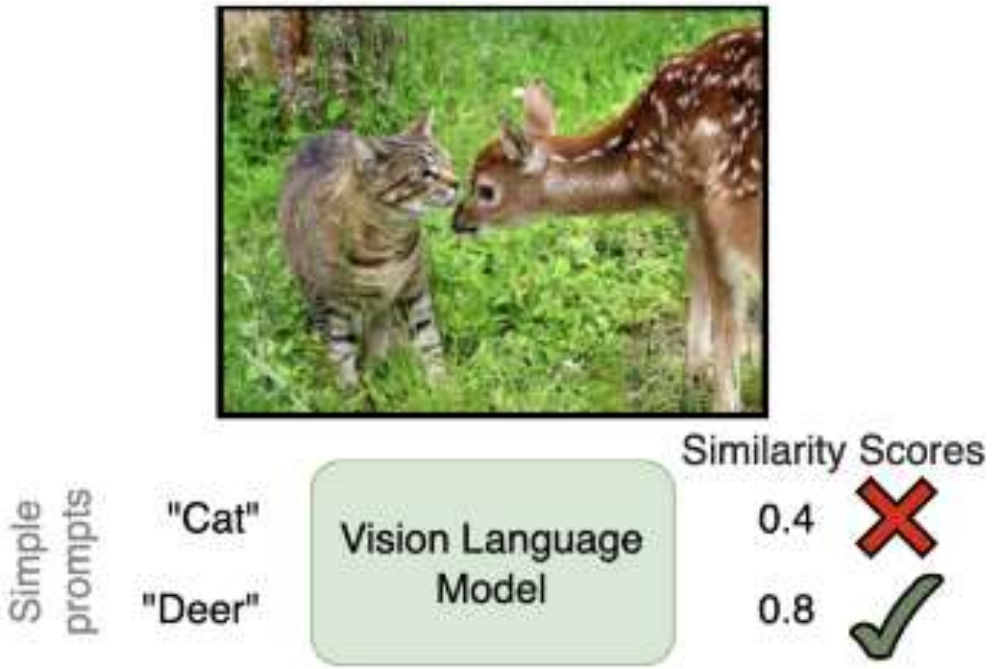


Image- and Prompt-specific Biases

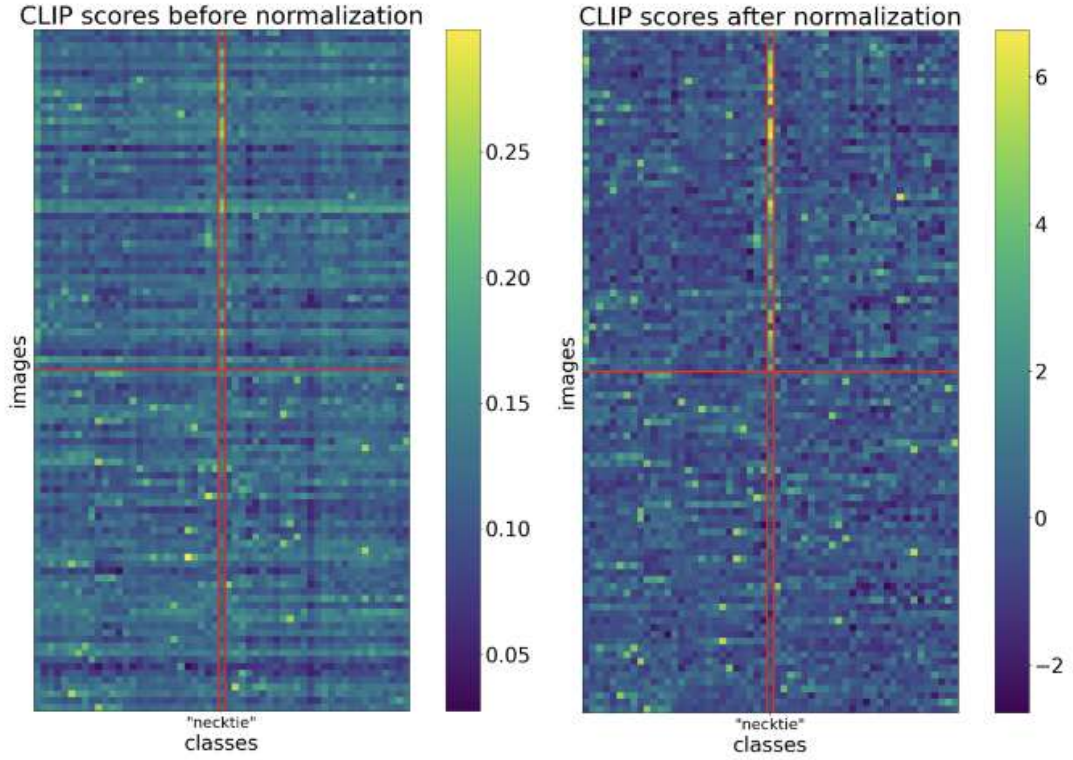


Figure 2. A motivating example for the Normalization module, with CLIP scores on a slice of the COCO dataset, before and after normalization. The "necktie" class is present for examples in the top half of the plot and absent in the bottom half. Image- and prompt-level biases show up as horizontal and vertical striations; normalization removes these and creates better separation.

Suboptimal maximum compound prompt

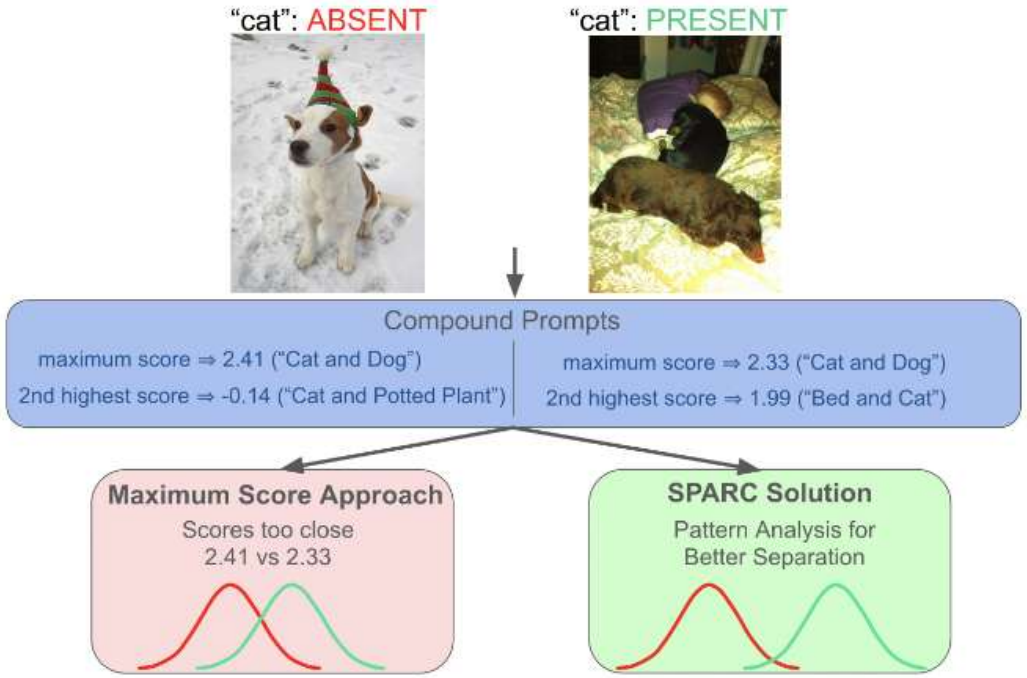
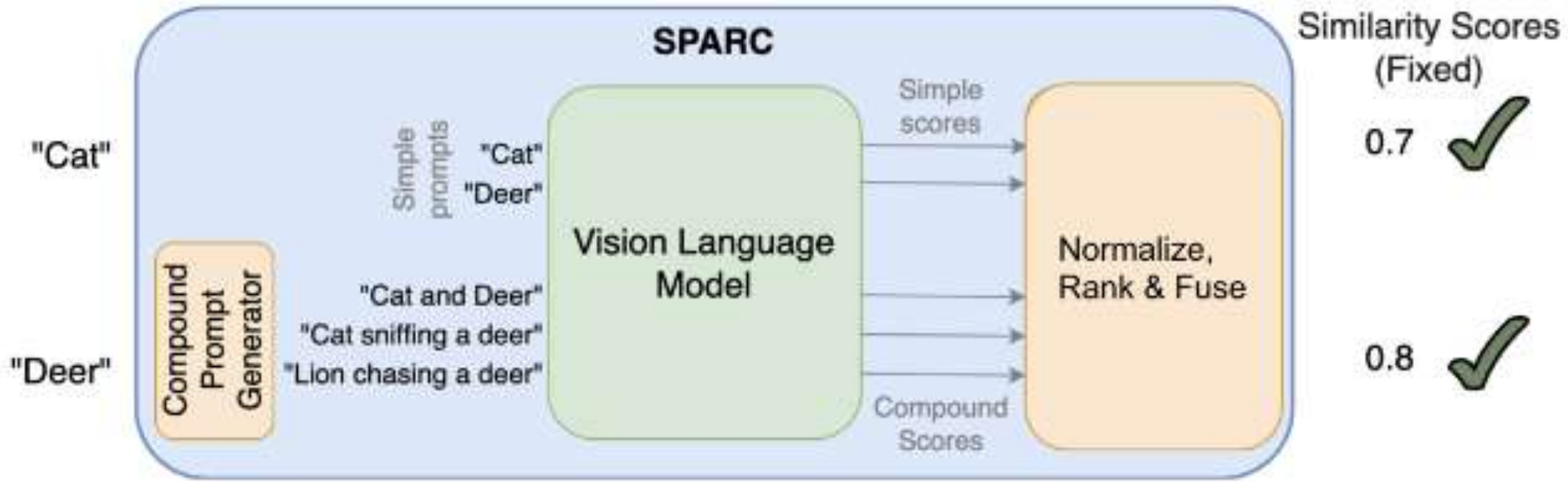


Figure 3. A motivating example for the Rank Fusion module, with an image where class "cat" is absent (left) and one where it is present (right). The highest compound prompt score is an unhelpful signal because it gives a high score to both negatives and positives, while the second-highest is more discriminative. Our method adaptively fuses the most informative order statistics, resulting in a strong signal.

Method

Overview



Compound Prompt Generation

The process takes as input the classnames $c_1, ..., c_N$, ground-truth cooccurrence statistics $\mathbb{P}$, thresholds τ_2 and τ_3 (fixed to 0.05 and 0.025 for all datasets), and off-the-shelf LLM Φ.

We use probability thresholding to select pairs and triplets of classes that could plausibly coocur in realistic visual scenes. These pairs and triplets are used to make formulaic compound prompts of the form “A and B” and “A, B, and C”. We feed these formulaic prompts to LLM Φ, and ask it to generate natural sentences from them.

Our total set of compound prompts comprises the formulaic pair, formulaic triplet, and natural sentence prompts, which add up to on average 20 compound prompts per class in all of our datasets.

Normalization

Query the VLM to obtain singleton scores $s_1^t, \ldots, s_N^t$, and compound scores $\{s_p^t : p \in P\}$ for each image t.

Address image-level bias:

$$\tilde{s}_i^t := \frac{s_i^t - \hat{\mu}(s_\cdot^t)}{\hat{\sigma}(s_\cdot^t)} \quad \textit{and} \quad \tilde{s}_p^t := \frac{s_p^t - \hat{\mu}(\check{s}_\cdot^t)}{\hat{\sigma}(\check{s}_\cdot^t)}, \qquad (1)$$

$\hat{\mu}(s_\cdot^t),\ \hat{\mu}(\check{s}_\cdot^t),\ \hat{\sigma}(s_\cdot^t),\ \hat{\sigma}(\check{s}_\cdot^t)$ Sample means and standard deviations across the prompt dimension for a single image.

$\check{s}_1^t, \ldots, \check{s}_N^t$ Scores of “auxiliary” prompts that mention classnames in isolation

Address prompt-level bias:

$$\overline{s}_i^t := \frac{\tilde{s}_i^t - \hat{\mu}(\tilde{s}_i^\cdot)}{\hat{\sigma}(\tilde{s}_i^\cdot)} \quad \textit{and} \quad \overline{s}_p^t := \frac{\tilde{s}_p^t - \hat{\mu}(\tilde{s}_p^\cdot)}{\hat{\sigma}(\tilde{s}_p^\cdot)}, \qquad (2)$$

$\hat{\mu}(\tilde{s}_i^\cdot),\ \hat{\mu}(\tilde{s}_p^\cdot),\ \hat{\sigma}(\tilde{s}_i^\cdot),$ and $\hat{\sigma}(\tilde{s}_p^\cdot)$ Sample means and SDs across the image.

Rank Fusion

Do fusion by taking a weighted sum where the weighting vector is the direction of highest variance and compute the fused compound score $\tilde{\zeta}_i^t$ as

$$w^{i*} := \arg\max_{w^i} \mathrm{Var}_t(w_0^i \bar{s}_i^t + \sum_k w_k^i r_{i,k}^t) \qquad (3)$$

$$\tilde{\zeta}_i^t := w_0^{i*} \bar{s}_i^t + \sum_k w_k^{i*} r_{i,k}^t. \qquad (4)$$

$r_{i,k}^t$ the score of the k-th highest scoring compound prompt that mentions class i

Add this fused score into the original singleton score to get our final score

$$\zeta_i^t := s_i^t + \tilde{\zeta}_i^t \qquad (5)$$

Method

Pipeline

Algorithm 1 SPARC Pipeline

Input: Images $\mathcal{I}$, Class Names $\mathcal{C}$
Output: Final scores ζ_i^t for each image t and class i
$P \leftarrow \text{GenerateCompoundPrompts}(\mathcal{C})$ *(Details in Supp.)*
for $t \in \mathcal{I}$ **do**
  $s_i^t \leftarrow \text{GetVLMScores}(t, i)$ *(Singleton scores)*
  $\forall p \in P,\ s_p^t \leftarrow \text{GetVLMScores}(t, p)$ *(Compound scs.)*
  // (Normalize scores (Eqs. 1,2))
  $\tilde{s}_i^t \leftarrow \text{ImageNorm}(s_i^t),\ \tilde{s}_p^t \leftarrow \text{ImageNorm}(s_p^t)$
  $\bar{s}_i^t \leftarrow \text{PromptNorm}(\tilde{s}_i^t),\ \bar{s}_p^t \leftarrow \text{PromptNorm}(\tilde{s}_p^t)$
  for $i \in \mathcal{C}$ **do**
    $r_{i,k}^t \leftarrow \text{GetOrderStatistics}(\{\bar{s}_p^t : i \in p\})$
    $w^{i*} \leftarrow \text{MaxVarDirection}([\bar{s}_i^t, r_{i,k}^t])$ *(Eq. 3)*
    $\tilde{\zeta}_i^t \leftarrow w_0^{i*} \bar{s}_i^t + \sum_k w_k^{i*} r_{i,k}^t$ *(Eq. 4)*
    $\zeta_i^t \leftarrow \bar{s}_i^t + \tilde{\zeta}_i^t$ *(Eq. 5)*
return $\{\zeta_i^t\}$

Comparison with ZSCLIP

Dataset		Architectures									
		ViT-L/14 336px	ViT-L/14	ViT-B/16	ViT-B/32	RN50 x64	RN50 x16	RN50 x4	RN101	RN50	Avg (archs)
COCO	ZSCLIP	59.1	58.1	55.5	50.9	58.6	58.1	55.7	52.7	53.0	55.7
	Ours	**70.5**	**69.5**	**67.4**	**64.0**	**70.6**	**70.1**	**69.5**	**67.7**	**65.7**	**68.3**
VOC	ZSCLIP	81.3	79.9	79.9	77.5	83.1	81.6	80.6	80.3	79.7	80.4
	Ours	**88.9**	**88.3**	**88.7**	**87.5**	**90.5**	**90.3**	**90.0**	**89.7**	**89.2**	**89.2**
NUSWIDE	ZSCLIP	41.4	41.0	40.9	38.7	39.8	38.5	38.8	35.7	38.5	39.3
	Ours	**47.5**	**47.1**	**47.3**	**46.9**	**47.4**	**47.9**	**48.3**	**46.5**	**45.8**	**47.2**

Table 2. Results over three datasets and nine CLIP backbones. Our proposed method consistently outperforms the ZSCLIP baseline across all datasets and architectures exhibiting its effectivness on multi-label recognition tasks.

Complementarity with other methods

		RN50 x64	RN50 x16	RN50 x4	RN 101	RN 50
COCO	CLIP-DPT	70.5	64.2	64.2	59.7	57.1
	Ours-DPT	**77.8**	**73.9**	**72.7**	**69.8**	**68.4**
VOC	CLIP-DPT	86.6	82.4	86.1	83.5	82.6
	Ours-DPT	**91.1**	**89.8**	**88.5**	**88.5**	**89.9**
NUS	CLIP-DPT	39.8	37.8	39.3	38.4	38.1
	Ours-DPT	**43.5**	**45.5**	**42.7**	**42.6**	**44.5**

Table 3. Comparing SPARC with local-feature method CLIP-DPT [4]. Our method still improves upon local features, showing that its strength is complementary.

	COCO ViT-B/16	VOC ViT-B/16	Avg
TagCLIP	70.9	91.7	81.3
+Ours	**73.8**	**92.0**	**82.9**

Table 4. Shows complementarity with architectural approaches, improving TagCLIP by 1.6%

Ablation study

Normalize	Pair prompts	Triplets + Descriptive	Rank Fusion Strategy	COCO	VOC	NUS	Avg
✓	-		-	65.9	87.7	45.1	66.2
✓	all pairs		ours	67.5	88.5	46.4	67.5
✓	all pairs		mean	67.5	88.5	46.4	67.5
✓	cooccurrence-filtered		ours	68.1	89.0	47.0	68.0
✓	cooccurrence-filtered		mean	67.9	88.5	46.8	67.7
✓	cooccurrence-filtered	✓	ours	**68.3**	**89.2**	**47.2**	**68.3**

Table 6. Ablations on the makeup of our compound prompts. We find that we can still enjoy some benefit from pairwise prompts without any cooccurrence filtering.

Compound	Normalize	COCO	VOC	NUS	Avg
		55.7	80.4	39.3	58.5
	✓	65.9	87.7	45.1	66.2
✓		58.4	80.5	40.0	59.7
✓	✓	**68.3**	**89.2**	**47.2**	**68.3**

Table 7. Ablations on Normalization module. Quantifies impact of normalization both with and without compound prompts.

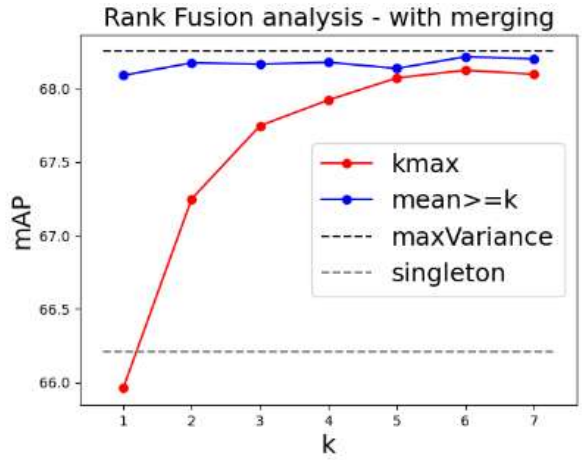


Figure 4. Average mAP for different Rank Fusion strategies demonstrates superiority of adaptive fusion over fixed strategies.

Thanks