

474 **Appendix Contents**

475	A Additional results	15
476	A.1 Retrieval faithfulness metrics	15
477	A.2 Task performance metrics	15
478	A.3 More visualizations	15
479	A.4 More datasets	15
480	B Method details	17
481	B.1 Concept identification details	17
482	B.2 Algorithm for Surrogate Learning	17
483	B.3 Experimental details	18
484	B.4 User study details	18
485	B.5 Broader impacts	19

A Additional results

A.1 Retrieval faithfulness metrics

Table 2: COCO retrieval faithfulness metrics. For our method, we report the mean over 5 runs.

Method	COCO I→T			COCO T→I		
	R@1	R@5	R@10	R@1	R@5	R@10
SpLiCE [26]	38.12	72.40	86.00	38.26	72.74	85.66
Ours	61.99	89.31	95.51	55.03	86.64	94.46

Table 3: Flickr30k retrieval faithfulness metrics. For our method, we report the mean over 5 runs.

Method	Flickr30k I→T			Flickr30k T→I		
	R@1	R@5	R@10	R@1	R@5	R@10
SpLiCE [26]	40.25	70.41	80.65	40.50	69.56	80.04
Ours	66.37	86.32	91.26	60.31	82.79	89.42

Table 2 and Table 3 shows the detailed R@K metrics for COOC and Flickr30k of our method and SpLiCE, which is supplementary to Table 1 of the main paper.

A.2 Task performance metrics

Table 4: Performance metrics. For our method, we report the mean over 5 runs.

Method	COCO I→T			COCO T→I			SUN397	ImageNet
	R@1	R@5	R@10	R@1	R@5	R@10		
SpLiCE [26]	32.26	62.94	76.44	29.92	59.22	73.00	52.41	42.91
CLIP ViT-B/32	51.90	81.26	90.24	47.48	77.10	87.08	60.71	61.91
Ours	58.03	85.68	93.19	56.30	84.98	93.09	57.98	52.37

Table 4, 5 presents a comparison of performance metrics (evaluated against dataset ground truth) for our method, SpLiCE, and the original CLIP model. The baseline is setup similar to the experiment in Tab. 1. As can be seen in Table 4, our method consistently outperforms the baseline SpLiCE across tasks given the same sparsity.

Furthermore, although designed for post-hoc explanation, our surrogate representation exhibits the ability to sometimes outperform the CLIP model it explains on the zero-shot tasks we tested, despite never having access to the dataset labels and only being trained on the similarities produced by the CLIP model. We hypothesize that performance improvement (when they exists) stems from the increased robustness of concept-based inputs, which may be less susceptible to common image degradations such as occlusion, blurriness, or general noise, compared to raw image inputs.

A.3 More visualizations

In Figure 8, we provide a more comprehensive list of concepts of the images visualized in Fig. 4.

A.4 More datasets

In Table 6, we provide more zero-shot classification results on more datasets (Flowers102, Food101). Our method continues to yield consistent improvements on these fine-grained classification tasks. In Figure 9, we visualize some results from these datasets.

Table 5: Performance metrics (continued). For our method, we report the mean over 5 runs.

Method	Flickr30k I→T			Flickr30k T→I		
	R@1	R@5	R@10	R@1	R@5	R@10
SpLiCE [26]	36.43	65.03	75.36	36.83	63.50	74.20
CLIP ViT-B/32	67.17	88.90	93.80	63.60	86.78	92.26
Ours	74.53	92.58	96.24	73.60	92.67	96.30

Table 6: Faithfulness metrics for more datasets.

Method	Flowers102	Food101
SpLiCE	26.12	51.80
Ours	37.73	63.62



Figure 8: More comprehensive depiction of concept list from our method and the baseline.

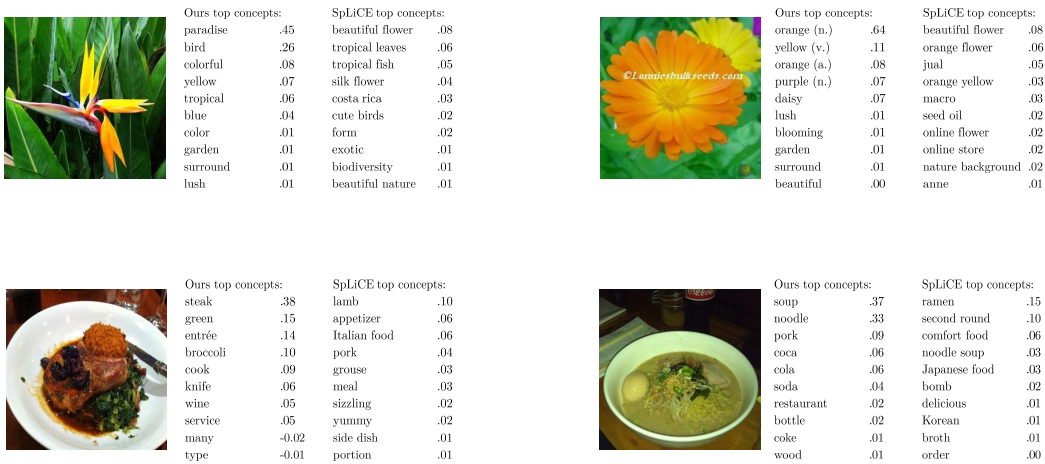


Figure 9: Visualizations for samples from the Flowers102 dataset (upper) and Food101 dataset (lower).

507 B Method details

508 B.1 Concept identification details

509 We redescribe our concept identification process step-by-step with more details:

- 510 1. (If not already present) obtain the captions for each image by using a captioning model
- 511 2. Extract nouns, verbs, and adjectives concepts via part-of-speech tagging with the `nltk`
- 512 library
- 513 3. Filter out words not presenting in WordNet, infrequent and overly frequent concepts to
- 514 obtain the vocabulary C
- 515 4. Estimate concept prevalence score for each image-concept pair to obtain the concept vectors
- 516 \mathbf{c}_i

517 B.2 Algorithm for Surrogate Learning

518 Below is the pseudocode for the algorithm described in Section 3.3.

Algorithm 1: Algorithm for Surrogate Learning

Input: CLIP encoders e_I, e_T , surrogate model f , dataset \mathcal{S} , concepts \mathcal{C} , temperature τ , batch size B , number of iterations T

```

1 for  $t = 0, \dots, T - 1$  do
2   Sample a mini-batch of triplets  $\mathcal{B}_t = \{(\mathbf{x}_i, \mathbf{c}_i, \mathbf{t}_i)\}$  from the dataset
3   Update  $u_{t+1,i}, v_{t+1,i}$  using (5) and (6) for  $i \in \mathcal{B}_t$ 
519 4   Set  $u_{t+1,i} = u_{t,i}, v_{t+1,i} = v_{t,i}$  for  $i \notin \mathcal{B}_t$ 
5   Compute gradient estimator w.r.t.  $f_t$ :
      
$$\frac{1}{|\mathcal{B}_t|^2} \sum_{i \in \mathcal{B}_t} \sum_{j \in \mathcal{B}_t} \frac{\exp\left(\frac{e_I(\mathbf{x}_i)^\top e_T(\mathbf{t}_j)}{\tau}\right)}{u_{t+1,i}} \cdot \frac{\exp\left(\frac{f(\mathbf{c}_i)^\top e_T(\mathbf{t}_j)}{\tau}\right)}{v_{t+1,i}} \cdot \nabla g(f_t, e_T; \mathbf{c}_i, \mathbf{t}_j, \mathcal{B}_t).$$

6   Update  $f_{t+1}$  from  $f_t$  using an optimizer

```

520 We now present the full derivation of Algorithm 1. We will focus only on $D_{KL}(P_i^e \parallel P_i^f)$ since the
521 procedure for optimizing $D_{KL}(Q_j^e \parallel Q_j^f)$ can be derived analogously. From the definition of KL
522 divergence, we can write the first part of (3) as

$$\frac{1}{2n} \sum_{i=1}^n D_{KL}(P_i^e \parallel P_i^f) = -\frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n P_i^e(j) \log P_i^f(j) + \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n P_i^e(j) \log P_i^e(j).$$

523 Note that e is fixed and we want to optimize only f , we will discard the second term on the right
524 hand side since it does not involve f . Plugging (1) and (2) into the above equation, we get

$$\begin{aligned}
& -\frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n P_i^e(j) \log P_i^f(j) \\
&= -\frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n \frac{\exp(e_I(\mathbf{x}_i)^\top e_T(\mathbf{t}_j)/\tau)}{\sum_k \exp(e_I(\mathbf{x}_i)^\top e_T(\mathbf{t}_k)/\tau)} \cdot \log \frac{\exp(f(\mathbf{c}_i)^\top e_T(\mathbf{t}_j)/\tau)}{\sum_k \exp(f(\mathbf{c}_i)^\top e_T(\mathbf{t}_k)/\tau)} \\
&= \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n \left(\sum_{k=1}^n \exp\left(\frac{e_I(\mathbf{x}_i)^\top e_T(\mathbf{t}_k) - e_I(\mathbf{x}_i)^\top e_T(\mathbf{t}_j)}{\tau}\right) \right)^{-1} \\
& \quad \cdot \log \sum_{k=1}^n \exp\left(\frac{f(\mathbf{c}_i)^\top e_T(\mathbf{t}_k) - f(\mathbf{c}_i)^\top e_T(\mathbf{t}_j)}{\tau}\right). \tag{10}
\end{aligned}$$

525 Recall that \mathcal{S} denotes the whole dataset, and

$$\begin{aligned} g(f, e_T; \mathbf{c}_i, \mathbf{t}_j, \mathcal{S}) &:= \frac{1}{n} \sum_{k=1}^n \exp \left(\frac{f(\mathbf{c}_i)^\top e_T(\mathbf{t}_k) - f(\mathbf{c}_i)^\top e_T(\mathbf{t}_j)}{\tau} \right) \\ &= \frac{1}{n} \sum_{k=1}^n \exp \left(\frac{f(\mathbf{c}_i)^\top e_T(\mathbf{t}_k)}{\tau} \right) / \exp \left(\frac{f(\mathbf{c}_i)^\top e_T(\mathbf{t}_j)}{\tau} \right). \end{aligned}$$

526 Then (10) can be equivalently written as

$$\frac{1}{|\mathcal{S}|^2} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} \frac{1}{g(e_I, e_T; \mathbf{x}_i, \mathbf{t}_j, \mathcal{S})} \cdot \log g(f, e_T; \mathbf{c}_i, \mathbf{t}_j, \mathcal{S}). \quad (11)$$

527 The gradient w.r.t. f is given by

$$\frac{1}{|\mathcal{S}|^2} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} \frac{1}{g(e_I, e_T; \mathbf{x}_i, \mathbf{t}_j, \mathcal{S})} \cdot \frac{1}{g(f, e_T; \mathbf{c}_i, \mathbf{t}_j, \mathcal{S})} \cdot \nabla g(f, e_T; \mathbf{c}_i, \mathbf{t}_j, \mathcal{S}).$$

528 Since we only have access to one mini-batch of data \mathcal{B} at each iteration, the obtained mini-batch
529 gradient estimator is

$$\frac{1}{|\mathcal{B}|^2} \sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{B}} \frac{1}{g(e_I, e_T; \mathbf{x}_i, \mathbf{t}_j, \mathcal{B})} \cdot \frac{1}{g(f, e_T; \mathbf{c}_i, \mathbf{t}_j, \mathcal{B})} \cdot \nabla g(f, e_T; \mathbf{c}_i, \mathbf{t}_j, \mathcal{B}).$$

530 However, due to the non-linearity of the reciprocal function, the expectation over \mathcal{B} is not equal to the
531 true gradient. Thus the mini-batch estimator is biased. To address this problem, we use two moving
532 average sequences u and v to approximate $g(e_I, e_T; \mathbf{x}_i, \mathbf{t}_j, \mathcal{S})$ and $g(f, e_T; \mathbf{c}_i, \mathbf{t}_j, \mathcal{S})$ respectively:

$$\begin{aligned} u_{t+1,i} &= (1 - \gamma_1)u_{t,i} + \gamma_1 \frac{1}{|\mathcal{B}_t|} \sum_{k \in \mathcal{B}_t} \exp \left(\frac{e_I(\mathbf{x}_i)^\top e_T(\mathbf{t}_k)}{\tau} \right), \\ v_{t+1,i} &= (1 - \gamma_2)v_{t,i} + \gamma_2 \frac{1}{|\mathcal{B}_t|} \sum_{k \in \mathcal{B}_t} \exp \left(\frac{f(\mathbf{x}_i)^\top e_T(\mathbf{t}_k)}{\tau} \right), \end{aligned}$$

534 where $\gamma_1, \gamma_2 \in (0, 1]$ are two hyperparameters. Then we can approximate $g(e_I, e_T; \mathbf{x}_i, \mathbf{t}_j, \mathcal{S})$ and
535 $g(f, e_T; \mathbf{c}_i, \mathbf{t}_j, \mathcal{S})$ using

$$u_{t+1,i} / \exp \left(\frac{e_I(\mathbf{x}_i)^\top e_T(\mathbf{t}_j)}{\tau} \right), \text{ and } v_{t+1,i} / \exp \left(\frac{f(\mathbf{c}_i)^\top e_T(\mathbf{t}_j)}{\tau} \right).$$

536 B.3 Experimental details

537 **Datasets.** For the Flickr30k dataset, we explain on the full dataset. For the SUN397 dataset, we use
538 the first official testing split. For the COCO 2017 and ImageNet dataset, we use the validation split.
539 Following [26], for computational efficiency, the retrieval metrics are computed in batches of size
540 1000 and averaged over the full dataset.

541 **Training.** We use the Amsgrad variant of the AdamW optimizer with learning rate 10^{-3} and weight
542 decay 10^{-6} . During training, we distill image-text similarities within the batch following Algorithm 1.
543 We perform augmentations on both modalities: selecting a random caption as text augmentation, and
544 random center crop and horizontal flip as image augmentation.

545 B.4 User study details

546 The user study involved 10 volunteers who did not receive monetary compensation. The interface
547 is shown in Figure 10. Notably, we observed that SpLiCE’s weights are often lower than that of us,
548 since they do not sum to one. To reduce the chance of the user differentiating the two methods based
549 on the weights, we scaled SpLiCE’s weights for each individual image so that their sum equals ours.


550 The study was ruled exempt by our institution’s IRB, as no more than minimal risk is posed to the
551 participants. No identifying information was collected.

Criteria	CI	p-value
Relevance	1.61 ± 0.54	9×10^{-8}
Completeness	1.69 ± 0.72	1×10^{-6}
Utility	1.66 ± 0.62	4×10^{-7}

Table 7: The p-values and confidence intervals (CI) for hypothesis testing in our user study. A value of 1 denotes strong preference for EXPLAIN-R, averaged across the 20 shown samples. The hypothesis tested is whether the population mean is less than 3, which denotes neutrality (no preference for SpLiCE or EXPLAIN-R).

552 B.5 Broader impacts

553 Our work addresses the problem of interpreting CLIP image representations in a task independent
554 manner. EXPLAIN-R provides a tool for users and researchers to inspect the semantic content of the
555 representation and provide a simple, intuitive summarization of the learned concepts for each image.
556 For the users, this will enhance transparency and trustworthiness in multimodal representations,
557 which traditionally relies on downstream evaluation. Researchers can use EXPLAIN-R to inspect
558 individual embeddings and model predictions, as well as aggregate them over the dataset to obtain a
559 more holistic view of the concepts learned by the model.



Explanation A:		Explanation B:	
grumpy	.16	bird	.47
cute birds	.14	branch	.27
mid adult	.10	perch	.10
sparrow	.09	beak	.06
banded	.09	closeup	.02
individual	.08	eye	.01
beak	.08	dark	.01
male female	.07	tiny	.01
bird feeder	.07	head	.01
fluffy	.06	line	.00

Which explanation is more relevant to the image? *

1 2 3 4 5

Strongly prefer explanation A ☐ ☐ ☐ ☐ ☐ Strongly prefer explanation B

Which explanation captures concepts that are sufficient to explain the CLIP model's capabilities? *

1 2 3 4 5

Strongly prefer explanation A ☐ ☐ ☐ ☐ ☐ Strongly prefer explanation B

Which explanation is more helpful for understanding what information is encoded in the representation? *

1 2 3 4 5

Strongly prefer explanation A ☐ ☐ ☐ ☐ ☐ Strongly prefer explanation B

Figure 10: The interface for our user study. For each input image, we show the top 10 concepts from both methods, along with the weights. We scale SpLiCE’s weights so they have the same mean as ours.