PowerCLIP: Powerset Alignment for Contrastive Pre-Training

Masaki Kawamura[1,2], Nakamasa Inoue[1,2], Rintaro Yanagi[2], Hirokatsu Kataoka[2,3], Rio Yokota[1,2]
[1]Institute of Science Tokyo, [2]AIST, [3]University of Oxford, VGG

Abstract

*Contrastive vision-language pre-training frameworks such as CLIP have demonstrated impressive zero-shot performance across a range of vision-language tasks. Recent studies have shown that aligning individual text tokens with specific image patches or regions enhances fine-grained compositional understanding. However, it remains challenging to capture compositional semantics that span multiple image regions. To address this limitation, we propose **PowerCLIP**, a novel contrastive pre-training framework enhanced by powerset alignment, which exhaustively optimizes region-to-phrase alignments by minimizing the loss defined between powersets of image regions and textual parse trees. Since the naive powerset construction incurs exponential computational cost due to the combinatorial explosion in the number of region subsets, we introduce efficient non-linear aggregators (NLAs) that reduce complexity from $\mathcal{O}(2^M)$ to $\mathcal{O}(M)$ with respect to the number of regions M, provably approximating the exact loss value with arbitrary precision. Our extensive experiments demonstrate that PowerCLIP outperforms state-of-the-art methods in zero-shot classification and retrieval tasks, underscoring compositionality and robustness of our approach. Our code will be made publicly available.*

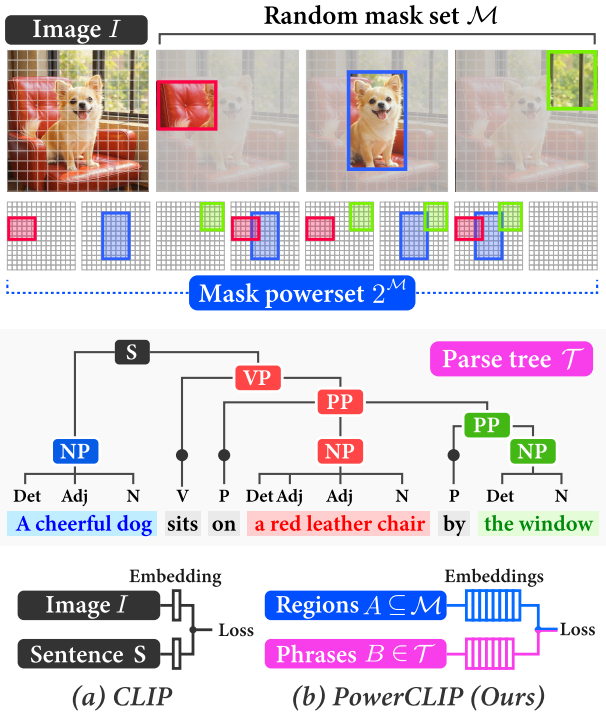

Figure 1. **Overview of PowerCLIP.** (a) CLIP aligns images and sentences globally. (b) PowerCLIP explores all combinations of image regions (*i.e.*, powerset) and aligns them with textual phrases.

1. Introduction

Large-scale contrastive pre-training has established a robust foundation for vision-language understanding. A prominent example is CLIP [49], which aligns visual and textual embeddings within a shared semantic space by minimizing the image-text contrastive loss. To improve robustness and compositionality, recent studies have explored sophisticated local and global alignment techniques [1, 7, 17, 46, 47, 60, 67]. Local alignment approaches, such as SPARC [3] and FineLIP [1], explicitly match textual tokens with corresponding visual patches, facilitating fine-grained correspondences. Global alignment approaches, such as A-CLIP [67] and CLIP-PGS [47], emphasize semantically informative image regions by applying global masks to visual patches. Though effective, both paradigms operate under single-region or masked-region objectives, limiting their ability to capture compositions among multiple visual entities. Motivated by this limitation, we propose **PowerCLIP**, a novel local-to-global alignment framework, which exhaustively aligns image regions with structured textual phrases in a combinatorial manner.

The core idea behind PowerCLIP lies in the *powerset alignment* strategy, which systematically explores all possible subsets of image regions (*i.e.*, the powerset of image regions) and aligns them with phrase structures extracted from textual parse trees. Specifically, since pre-training begins from scratch, we first generate a set of random region masks $\mathcal{M}$ for each image and then define a contrastive objective between the powerset $2^{\mathcal{M}}$ and the textual parse tree $\mathcal{T}$, as illustrated in Figure 1. This approach significantly enhances the compositionality and robustness due to the ex-

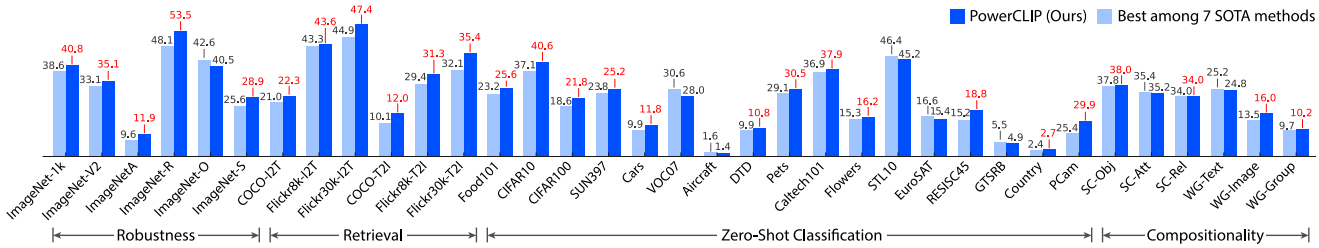

Figure 2. Performance comparison between PowerCLIP and the best-performing method among seven state-of-the-art approaches (CLIP, FLIP, A-CLIP, E-CLIP, C-PGS, FILIP, and SPARC). Performance improvements are highlighted in red.

haustive exploration of local-to-global alignments.

Moreover, since powerset alignment inherently introduces exponential computational complexity, we develop theoretically grounded approximations using **Non-Linear Aggregators (NLAs)** that reduce the complexity to linear in terms of the number of region masks. Through extensive experimentation, we demonstrate that PowerCLIP achieves state-of-the-art performance across 22 out of 28 diverse benchmarks, including classification, image-text retrieval, robustness, and compositionality evaluations, as shown in Figure 2. Our key contributions are summarized as follows:

- We propose PowerCLIP, a novel contrastive pre-training framework leveraging powerset alignment between image regions and textual phrases.
- We develop NLAs that derive computationally tractable approximations for powerset alignment, reducing complexity from exponential to linear. We prove that NLAs approximate the exact loss value with arbitrary precision under mild assumptions (Theorems 1 and 2).
- We demonstrate that PowerCLIP attains state-of-the-art performance across diverse zero-shot benchmarks, improving compositional reasoning and robustness.

2. Related Work

Contrastive Pre-training. Image-text contrastive learning, pioneered by methods such as CLIP [49] and ALIGN [28], has become a cornerstone in large-scale vision-language pre-training [9, 41, 53, 56, 66, 69, 72]. Meanwhile, several studies highlight its limitations, particularly regarding compositionality and robustness, due to inherent difficulties in embedding complex visual and textual structures into a single shared semantic space [16, 27, 31, 55]. Recent efforts to address these limitations have focused on improving alignment from visual, textual, and multimodal perspectives.

Visual Masking Approaches. Inspired by masked image modeling [23, 61, 65], visual masking approaches have significantly enhanced global image-text alignment. For example, FLIP [35] applies random masks for efficient and robust training. MaskCLIP [13] incorporates masking mechanism into self-distillation. Subsequent approaches have focused more on structured and targeted masking. A-CLIP [67] emphasizes informative image regions through attentive masking. E-CLIP [60] employs cluster masking to better capture visual structures. CLIP-PGS [47] proposes gradual masking with the patch generation-to-selection mechanism. In contrast to these approaches, PowerCLIP performs local-to-global alignment by exploring combinations of region masks and aligns them with textual structures, enhancing compositionality and robustness.

Textual Approaches. From the textual perspective, several methods for textual augmentation and refinement have been proposed. For instance, VeCLIP [34] enriches textual descriptions, while LaCLIP [19] rewrites them to better align with visual semantics. NegationCLIP [44] introduces negation terms into textual descriptions to provide richer contrasts. Synthetic approaches have also shown promising effectiveness [46, 62, 70]. TripletCLIP [46] demonstrates that a triplet contrastive loss with hard negative samples improves compositionality. Although we retain the original text during pre-training for a fair comparison with other types of approaches, these studies motivate us to design a triplet margin loss to enhance compositionality.

Multimodal Approaches. Multimodal approaches primarily target fine-grained alignment between textual tokens and visual patches. Prominent examples include FILIP [17], which performs token-level alignment via cross-modal late interaction; SPARC [3], which employs sparse cross-modal alignment; and LAPS [21], which aligns patches and words by identifying redundant visual regions. In fine-tuning scenarios, several methods have addressed alignment with longer textual descriptions via fine-tuning or incremental training, such as LongCLIP [73], FixCLIP [58], GOAL [7], and FineLIP [1]. Additionally, word-to-region correspondences have been explored for downstream tasks via fine-tuning and adaptation for object detection [36, 38, 54, 75, 76] and semantic segmentation [8, 15, 22, 29, 30, 37, 42, 48, 64, 74]. However, capturing compositional semantics across multiple image regions remains challenging. We focus on pre-training scenarios and address compositionality by aligning combinations of image regions with textual parse trees, facilitating effective local-to-global alignment.

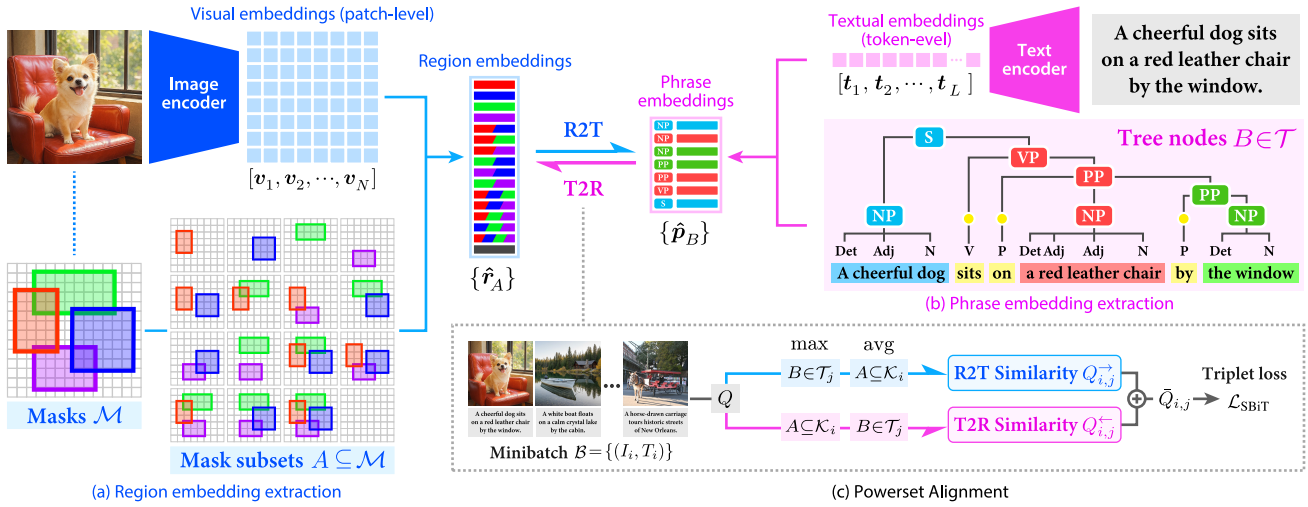

Figure 3. **Overview of the powerset alignment strategy for PowerCLIP.** (a) Region embeddings are extracted for each subset A of region masks in $\mathcal{M}$. (b) Phrase embeddings are extracted for each node B in the parse tree $\mathcal{T}$. (c) Powerset alignment minimizes the triplet loss defined based on the bidirectional similarity: region-set-to-tree (R2T) and vice versa (T2R).

3. Method

This section introduces **PowerCLIP**, a novel contrastive pre-training framework for local-to-global alignment. The core idea behind PowerCLIP is powerset alignment, which exhaustively explores combinatorial correspondences between image regions and textual phrases, improving compositionality and robustness.

3.1. Overview

Problem Setting and Notation. We study image-text contrastive pre-training, where the training dataset consists of images paired with their corresponding textual descriptions. To achieve fine-grained alignment, we adopt Transformer-based encoders for both the image and text modalities to extract patch-level and token-level embeddings, respectively. We denote visual embeddings extracted from an image I by $[\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N] \in \mathbb{R}^{D \times N}$, and textual embeddings from a text description T by $[\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_L] \in \mathbb{R}^{D \times L}$, where N is the number of image patches, L is the length of the token sequence, and D is the shared feature dimension.

Architecture. Figure 3 shows the architectural overview of PowerCLIP, which aligns powersets of image regions and textual parse trees in three primary steps. First, for each training image I, a set of region masks $\mathcal{M}$ is generated on a patch grid either randomly or via a segmentation model. Here, region embeddings corresponding to all subsets of masks $A \subseteq \mathcal{M}$ are extracted as candidates to be matched with textual phrases, as shown in Figure 3(a). Second, phrase embeddings are extracted from each textual description T by identifying phrases using a parse tree as shown in Figure 3(b). Finally, powerset alignment is performed by minimizing the triplet loss defined with similarities in two directions: region-set-to-tree (R2T) and vice versa (T2R). Compared with conventional alignment methods (*e.g.*, SPARC [3]), our approach considers candidate matches more exhaustively over possible combinations of image regions and textual phrases, enhancing compositionality in image-text contrastive learning. Below, we describe details for region embedding extraction (§3.2), phrase embedding extraction (§3.3), and powerset alignment (§3.4).

3.2. Region Embedding Extraction

Region Masks. For each image I, we randomly generate $M \in \mathbb{N}$ bounding boxes on the patch grid by uniformly sampling their centers, heights, and widths. These bounding boxes define the set of region masks $\mathcal{M} = \{R_m\}_{m=1}^{M}$, where each $R_m \in \{0,1\}^N$ is a binary mask over patches. Optionally, this step can utilize segmentation models such as SAM [50] instead of random sampling.

Region Embeddings. To allow comprehensive matching with textual structures, we construct the powerset of the set of region masks: $2^{\mathcal{M}} = \{A \subseteq \mathcal{M}\}$, where each subset A corresponds to a combination of region masks. We then define the region embeddings $\boldsymbol{r}_A$ for each A by

$$\boldsymbol{r}_A = \sum_{R_m \in A} \phi(I \,|\, R_m) \tag{1}$$

where ϕ is a function to encode the image I given an individual region mask R_m. For computational efficiency, we apply masks to visual embeddings obtained from the entire image rather than encoding each image region independently. Specifically, we define ϕ by

$$\phi(I \,|\, R_m) = \frac{\boldsymbol{r}_m}{\|\boldsymbol{r}_m\|_2}, \quad \boldsymbol{r}_m = \sum_{n=1}^{N} R_{mn} \boldsymbol{v}_n, \tag{2}$$

where embeddings are L2 normalized.

3.3. Phrase Embedding Extraction

Parse Trees. For each text description T, we generate a constituency parse tree $\mathcal{T}$ by applying a syntactic parser. Each node $B \in \mathcal{T}$ corresponds to a sentence-level or phrase-level constituent, *e.g.*, Noun Phrase (NP), Verb Phrase (VP), Prepositional Phrase (PP), or Sentence (S).

Token Masks. Analogous to the region masks for the visual modality, we represent each leaf node by a token mask $P_{m'} \in \{0,1\}^L$, where m' indexes the leaf nodes. For example, given the description *"a dog sitting on a red chair,"* the noun phrase *"a dog"* is represented by a mask assigning ones to tokens *a* and *dog* and zeros elsewhere. Consequently, each non-leaf node B is represented by a set of token masks corresponding to its leaf nodes.

Phrase Embeddings. We define the phrase embeddings $\boldsymbol{p}_B$ for each node $B \in \mathcal{T}$ by

$$\boldsymbol{p}_B = \sum_{P_{m'} \in B} \psi(T \,|\, P_{m'}) \tag{3}$$

where ψ is an encoder function that applies the token mask $P_{m'}$ to textual embeddings as follows:

$$\psi(T \,|\, P_{m'}) = \frac{\boldsymbol{p}_{m'}}{\|\boldsymbol{p}_{m'}\|_2}, \quad \boldsymbol{p}_{m'} = \sum_{n=1}^{L} P_{m'n} \boldsymbol{t}_n. \tag{4}$$

These embeddings serve as phrase-level queries to identify corresponding image region subsets.

3.4. Powerset Alignment

Powerset alignment establishes local-to-global alignment by minimizing a triplet margin loss defined based on the bidirectional similarity between region subsets A and tree nodes B. We first define the fine-grained similarity scores, and then aggregate them in two directions, R2T and T2R, as shown in Figure 3(c).

Fine-Grained Similarity. Let $\mathcal{B} = \{(I_i, T_i)\}_{i=1}^{C}$ be a training mini-batch consisting of C image-text pairs. Given an image I_i and a text description T_j (potentially $j \neq i$), we define their fine-grained similarity scores $Q_{i,j,A,B}$ by measuring inner products between the region embeddings $\boldsymbol{r}_A^{(i)}$ and the phrase embeddings $\boldsymbol{p}_B^{(j)}$:

$$Q_{i,j,A,B} = \langle \boldsymbol{r}_A^{(i)}, \boldsymbol{p}_B^{(j)} \rangle \tag{5}$$

where $A \subseteq \mathcal{M}_i$ is a region subset, $B \in \mathcal{T}_j$ is a tree node, and $i, j \in \{1, 2, \cdots, C\}$ index samples in the mini-batch.

R2T Aggregation. This aggregation computes the best-matching phrase for each region subset and then aggregates corresponding scores by averaging. Specifically, we define the R2T similarity matrix $Q^{\rightarrow} \in \mathbb{R}^{C \times C}$ as

$$Q^{\rightarrow}_{i,j} = \frac{1}{2^M} \sum_{A \subseteq \mathcal{M}_i} \max_{B \in \mathcal{T}_j} Q_{i,j,A,B}. \tag{6}$$

This emphasizes region-level coverage, while neglecting less-relevant phrases that do not strongly correspond to any region subset.

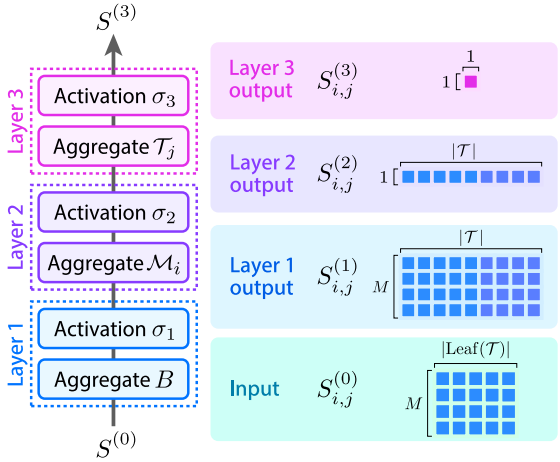

Figure 4. **Non-Linear Aggregator (NLA).** Each layer applies aggregation followed by activation.

T2R Aggregation. Conversely, this aggregation computes the best-matching region subset for each phrase. We define the T2R similarity matrix $Q^{\leftarrow} \in \mathbb{R}^{C \times C}$ as

$$Q^{\leftarrow}_{i,j} = \frac{1}{|\mathcal{T}_j|} \sum_{B \in \mathcal{T}_j} \max_{A \subseteq \mathcal{M}_i} Q_{i,j,A,B}. \tag{7}$$

This emphasizes phrase-level grounding by ensuring each phrase is closely matched to a region subset.

Loss Function. Combining the two similarity matrices, we form the final similarity as $\bar{Q} = Q^{\rightarrow} + Q^{\leftarrow}$. For training, we employ the triplet margin loss [1, 2] due to its effectiveness in encouraging margin-based discrimination between matched and mismatched pairs. Specifically, our triplet loss is defined as

$$\mathcal{L}_{\text{triplet}} = \Phi_\gamma(\bar{Q}) + \Phi_\gamma(\bar{Q}^\top) \tag{8}$$

where $\bar{Q}^\top$ is the transpose of $\bar{Q}$ and $\Phi_\gamma : \mathbb{R}^{C \times C} \to \mathbb{R}$ is the row-wise triplet loss function given by

$$\Phi_\gamma(X) = \frac{1}{C} \sum_{i=1}^{C} \max\left(\max_{j \neq i} X_{i,j} - X_{i,i} + \gamma, 0 \right). \tag{9}$$

The final loss function is a sum of the CLIP contrastive loss and the triplet loss: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CLIP}} + \lambda \mathcal{L}_{\text{triplet}}$, where $\lambda = 0.2$.

Discussion. Compared to token-to-token alignment frameworks [1, 17], PowerCLIP establishes more exhaustive alignment. However, computing the loss function poses a significant challenge, as it involves exponential complexity with respect to the number of region masks. We address this limitation through theoretically grounded approximations.

4. Tractable Approximations

This section introduces **Non-Linear Aggregators (NLAs)**, which provide tractable approximations for the R2T and T2R aggregations. NLAs offer two primary advantages.

First, training stability is improved by employing soft assignment instead of hard assignment computed via max operations in Eqs. (6, 7). Second, computational complexity of aggregation is significantly reduced from $\mathcal{O}(2^M)$ to $\mathcal{O}(M)$ with respect to the number of masks M.

4.1. General Form of NLAs

The NLA comprises three layers, each consisting of an aggregation operation followed by an activation function, as shown in Figure 4. The input is the similarity tensor $S^{(0)}$, obtained by computing inner products between individual region masks and phrases at leaf nodes:[1]

$$S^{(0)}_{i,j,m,m'} = \langle \phi(I_i \,|\, R_m), \psi(T_j \,|\, P_{m'}) \rangle, \tag{10}$$

where $R_m \in \mathcal{M}_i$ is a region mask for the image I_i and $P_{m'} \in \mathrm{Leaf}(\mathcal{T}_j)$ is a token mask at a leaf node for the description T_j. The encoders ϕ, ψ are from Eqs. (2, 4).

At each layer, indexed by $l \in \{1, 2, 3\}$, the similarity scores in $S^{(l-1)}$ are aggregated by summation over a specific dimension followed by an optional activation function $\sigma_l : \mathbb{R} \to \mathbb{R}$. Specifically, the first layer aggregates scores for each node $B \in \mathcal{T}_j$:

$$S^{(1)}_{i,j,m|B} = \sigma_1 \left(\sum_{P_{m'} \in B} S^{(0)}_{i,j,m,m'} \right). \tag{11}$$

The second layer aggregates scores over region masks:

$$S^{(2)}_{i,j|B} = \sigma_2 \left(\sum_{R_m \in \mathcal{M}_i} S^{(1)}_{i,j,m|B} \right). \tag{12}$$

Finally, the third layer aggregates scores over tree nodes:

$$S^{(3)}_{i,j} = \sigma_3 \left(\frac{1}{|\mathcal{T}_j|^{1-\alpha}} \sum_{B \in \mathcal{T}_j} S^{(2)}_{i,j|B} \right), \tag{13}$$

where $\alpha \in [0, 1]$ is a hyperparameter that interpolates between average aggregation ($\alpha = 0$) and summation aggregation ($\alpha = 1$).

These aggregation procedures avoid summation or maximization over powersets, thus reducing computational complexity. Nevertheless, the proposed design enables NLAs to approximate the R2T and T2R similarity matrices through a careful choice of activation functions.

4.2. NLA-T1 for T2R Aggregation

For the T2R aggregation, we introduce a specific class of NLAs, referred to as NLA Type 1 (NLA-T1). Theorem 1 and Corollary 1 prove that NLA-T1 is a soft-assignment variant of the T2R aggregation, computing $Q^{\leftarrow}_{i,j}$ in Eq. (7). We provide a proof in Appendix A.

[1] Since the number of tree nodes depends on each textual description, $S^{(0)}$ is a pseudo tensor comprising $CM \times \sum_{j=1}^{C} |\mathrm{Leaf}(\mathcal{T}_j)|$ scores.

Definition 1 (NLA-T1). *NLA-T1 is a class of NLAs defined by the following activation functions and hyperparameters:*

$$\sigma_1(x) = \tau \cdot \mathrm{Act}\left(\frac{x}{\tau}\right), \ \sigma_2 = \sigma_3 = \mathrm{Id}, \ \alpha = 0 \tag{14}$$

where $\mathrm{Act} : \mathbb{R} \to \mathbb{R}$ *is a nonlinear activation function,* τ *is a temperature hyperparameter, and* Id *is the identity function.*

Theorem 1. *Suppose* $\mathrm{Act} = \mathrm{Softplus}$. *Then, NLA-T1 approximates the T2R similarity* $Q^{\leftarrow}_{i,j}$ *with arbitrary precision. That is, for any* $\epsilon > 0$, *there exists* $\tau > 0$ *such that* $|S^{(3)}_{i,j} - Q^{\leftarrow}_{i,j}| < \epsilon$.

Corollary 1. *Suppose* $\mathrm{Act} = \mathrm{ReLU}$ (*i.e.,* $\tau \to 0$). *Then, NLA-T1 computes the exact T2R similarity* $Q^{\leftarrow}_{i,j}$.

As Corollary 1 is corresponding to the hard assignment in Eq. (7), NLA-T1 using softplus with $\tau > 0$ can be interpreted as a soft assignment variant. In practice, choosing small positive $\tau \simeq 0.001$ leads to improved performance.

4.3. NLA-T2 for R2T Aggregation

Approximating the R2T aggregation is relatively more challenging than approximating the T2R aggregation, as the summation operation is performed over the powerset. Here, we introduce NLA Type 2 (NLA-T2), which evaluates the lower and upper bounds of the R2T similarity and interpolates between these bounds via a hyperparameter $\alpha \in [0, 1]$. Theorem 2 shows that NLA-T2 can approach the true similarity score arbitrarily closely by tuning α.

Definition 2 (NLA-T2). *NLA-T2 is a class of NLAs defined by the following activation functions and hyperparameters:*

$$\sigma_1(x) = \zeta_\alpha\left(\frac{x}{2\tau}\right), \sigma_2(x) = \exp(x), \ \sigma_3(x) = \tau \log(x), \tag{15}$$

where $\zeta_\alpha(x) = x + \alpha \int \mathrm{Act}(x)\mathrm{d}x$ *is a residual antiderivative of a differentiable activation function* Act, *satisfying* $\zeta_\alpha(0) = 0$, *and* τ *is a temperature hyperparameter.*

Theorem 2. *Suppose* $\mathrm{Act} = \tanh$. *Then, NLA-T2 approximates the R2T similarity* $Q^{\rightarrow}_{i,j}$ *with arbitrary precision. That is, for any* $\epsilon > 0$, *there exist* $\tau > 0$ *and* $\alpha \in [0, 1]$ *such that* $\left|S^{(3)}_{i,j} - Q^{\rightarrow}_{i,j}\right| < \epsilon$.

We provide a proof in Appendix B. In practice, the lower bound ($\alpha = 0$) and the upper bound ($\alpha = 1$) are often close to each other when τ is small, making our approach robust to the choice of α (see Figure 5 for analysis).

4.4. Loss Function

Finally, we approximate the triplet loss by replacing $\bar{Q}$ in Eq. (8) with $\bar{S}$ obtained using the two types of NLAs:

$$\bar{S} = \text{NLA-T1}(S^{(0)}) + \text{NLA-T2}(S^{(0)}). \tag{16}$$

Compared with naive computations in Eqs (7, 6), this significantly reduces computational cost while maintaining or even improving performance.

Method	Food101 [4]	CIFAR10 [33]	CIFAR100 [33]	SUN397 [63]	Cars [32]	VOC07 [18]	Aircraft [40]	DTD [10]	OxfordPets [45]	Caltech101 [20]	Flowers [43]	STL10 [11]	EuroSAT [24]	RESISC45 [6]	GTSRB [52]	Country [49]	PCam [57]	**Avg**
CLIP [49]	42.3	57.7	25.0	44.1	17.0	50.5	1.7	16.5	53.9	73.5	26.0	82.0	18.7	26.5	9.4	4.5	48.0	35.1
FLIP [35]	39.9	52.8	24.5	42.8	15.9	46.6	1.4	15.9	46.0	70.4	25.3	80.2	17.0	25.8	5.6	4.0	47.1	33.0
A-CLIP [67]	41.8	61.6	27.1	46.6	16.0	51.1	1.3	17.1	51.2	73.5	25.7	85.8	20.5	29.1	8.0	4.2	50.1	35.9
E-CLIP [60]	42.1	70.7	32.0	43.9	15.1	43.6	2.2	17.0	55.4	73.7	28.4	85.6	22.9	30.0	9.6	4.7	50.0	36.9
C-PGS [47]	46.5	73.5	37.3	47.5	19.9	55.1	**3.1**	19.8	58.1	72.7	30.7	88.2	22.8	30.4	**10.9**	4.5	50.8	39.5
FILIP [17]	33.2	74.3	36.4	44.3	11.0	47.4	1.6	13.9	34.3	64.2	12.2	**92.8**	**33.2**	24.3	8.4	2.8	50.0	34.4
SPARC [3]	42.1	71.9	35.5	45.1	16.0	**61.1**	2.6	19.1	52.4	72.0	27.6	82.9	23.8	24.4	9.8	4.8	50.7	37.8
PowerCLIP-R	50.3	74.7	**43.5**	48.7	22.9	53.2	2.9	**21.5**	58.7	**75.7**	32.4	88.4	30.8	**37.5**	9.8	4.6	50.0	41.5
PowerCLIP-S	**51.2**	**81.3**	40.1	**50.5**	**23.5**	56.0	1.6	21.3	**61.0**	72.9	**32.5**	90.5	29.0	33.9	7.8	**5.4**	**59.7**	**42.2**

Table 1. **Zero-shot classification.** We report Top-1 accuracy (%) for 17 diverse classification datasets. Avg indicates the average accuracy.

Method	Text Retrieval (Image to Text)									Image Retrieval (Text to Image)									Average		
	MS-COCO			Flickr8K			Flickr30K			MS-COCO			Flickr8K			Flickr30K					
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [49]	34.6	62.0	72.7	55.7	81.6	89.9	58.5	83.8	89.1	23.5	47.8	59.7	40.5	68.9	80.2	43.2	70.4	80.4	42.7	69.1	78.7
FLIP [35]	32.6	59.1	70.6	55.0	80.9	88.9	53.8	80.8	88.5	22.6	46.1	58.1	40.3	68.1	78.6	41.5	67.9	77.5	41.0	67.1	77.0
A-CLIP [67]	33.7	60.2	71.0	53.7	80.1	88.0	55.3	81.4	87.6	23.9	48.3	60.0	40.6	68.9	78.9	43.1	70.1	78.8	41.7	68.2	77.4
E-CLIP [60]	34.3	62.0	73.3	57.0	82.7	90.1	55.8	84.2	89.6	23.8	48.2	59.8	42.0	69.4	79.6	43.3	70.9	80.2	42.7	69.6	78.8
C-PGS [47]	36.0	64.4	74.6	58.3	82.9	90.8	59.9	83.5	90.8	25.1	49.5	61.6	44.4	71.7	81.1	47.1	73.5	82.0	45.1	70.9	80.1
FILIP [17]	16.8	38.0	50.8	31.2	55.2	66.8	35.7	61.0	72.5	14.0	33.3	44.8	24.2	50.2	62.3	27.3	55.1	65.8	24.9	48.8	60.5
SPARC [3]	33.7	60.9	72.3	55.2	82.2	90.5	57.1	82.6	89.6	23.8	48.0	59.6	41.0	70.1	79.3	42.7	71.3	80.1	42.3	69.2	78.6
PowerCLIP-R	36.7	64.0	75.0	58.5	**84.8**	91.4	61.7	84.8	91.9	26.3	51.1	62.7	44.8	72.7	82.4	46.6	74.3	82.7	45.8	72.0	81.0
PowerCLIP-S	**37.3**	**64.9**	**75.6**	**58.6**	84.4	**91.5**	**62.4**	**88.5**	**94.2**	**27.0**	**52.9**	**64.0**	**46.3**	**74.1**	**83.2**	**50.4**	**76.6**	**84.6**	**47.0**	**73.6**	**82.2**

Table 2. **Zero-shot image-text retrieval**. R@K indicates recall (%) at top $K = 1, 5$, and 10. Average columns are means across the six settings (MS-COCO, Flickr8K, Flickr30K for both Text Retrieval and Image Retrieval).

5. Experiments

5.1. Experimental Setting

Datasets and Tasks. Following [47, 60], we use the Conceptual Captions 12M (CC12M) dataset [5] for training. For extensive evaluation, models are evaluated across 28 benchmarks. Specifically, we conduct evaluation on (i) 17 diverse datasets for zero-shot classification (listed in Table 1), (ii) 3 datasets for image-text retrieval (COCO [39], Flickr8k [68] and Flickr30k [68]), (iii) 6 datasets for robustness evaluation (ImageNet-1k [12], -V2 [51], -A [26], -R [25], -O [26] and -Sketch [59]), and (iv) 2 datasets for compositionality evaluation (SugarCrepe [27] and Winoground [55]).

Baselines. We compare PowerCLIP with seven baselines: CLIP [49], FLIP [35], A-CLIP [67], E-CLIP [60], C-PGS [47], FILIP [17], and SPARC [3]. These baselines cover different global and local alignment strategies. All models are evaluated under a consistent setting.

Implementation. We adopt the training setting of [47, 60]. Specifically, ViT-B/16 [14] is used as the image encoder, with images resized to a 224×224. The text encoder is a Transformer consisting of 12 layers, 8 attention heads, and embedding dimensions of 512. The models are trained for 32 epochs using the AdamW optimizer with a cosine decay learning rate scheduler, an initial learning rate of 10^{-3}, weight decay of 0.2, and a batch size of 4,096. The number of masks M is set to 10. We use softplus and tanh activations with $\tau = 0.001$ for NLA-T1 and NLA-T2, respectively. For NLA-T2, the function $\zeta_\alpha(x)$ is given by $\log\cosh(x)$ and α is set to 0.75. We implement two variants of our approach: PowerCLIP-R, which uses random masks, and PowerCLIP-S, which uses masks randomly selected from those generated by SAM2 [50].

5.2. Experimental Results

Zero-Shot Classification. Table 1 summarizes zero-shot classification results. We observe that PowerCLIP-R significantly outperforms the CLIP baseline (+6.4%), while PowerCLIP-S further improves the performance, achieving the best average accuracy of 42.2% across 17 datasets. Significant gains are observed on challenging fine-grained datasets such as Cars (+6.5%), Food101 (+8.9%), and RESISC45 (+7.4%). Compared to state-of-the-art methods for global alignment (C-PGS) and local alignment (SPARC), PowerCLIP-S achieves +2.7 and +4.4 points higher average accuracy, respectively, surpassing both on 14 out of 17 datasets. These results demonstrate the superiority of our local-to-global alignment in capturing nuanced semantics.

Method	ImgNet-1k	ImgNet-V2	ImgNet-A	ImgNet-R	ImgNet-O	ImgNet-S	IN	OOD	All
CLIP [49]	36.1	30.7	8.0	47.6	38.4	24.9	36.1	29.0	31.0
FLIP [35]	34.4	29.5	7.1	41.4	39.5	20.1	34.4	27.5	28.7
A-CLIP [67]	35.2	30.1	8.1	45.1	39.4	23.7	35.2	30.3	30.3
E-CLIP [60]	36.3	30.7	8.1	47.9	39.6	25.4	36.3	30.3	31.3
C-PGS [47]	38.6	33.1	9.6	48.1	**42.6**	25.6	38.6	31.8	32.9
FILIP [17]	26.7	22.9	9.3	37.4	25.9	18.2	26.7	22.7	23.4
SPARC [3]	37.2	32.1	9.3	46.8	42.2	24.5	37.2	31.0	32.0
PowerCLIP-R	40.3	34.8	11.2	53.2	40.2	28.7	40.3	33.6	34.7
PowerCLIP-S	**40.8**	**35.1**	**11.9**	**53.5**	40.5	**28.9**	**40.8**	**34.0**	**35.1**

Table 3. **Robustness evaluation.** Top-1 accuracy for six ImageNet (ImgNet) datasets are reported with in-distribution (ID), out-of-distribution (OOD) and overall (All) averages.

Method	Obj	Att	Rel
CLIP [49]	73.9	68.8	64.5
FLIP [35]	72.0	66.9	66.0
A-CLIP [67]	70.2	68.5	63.2
E-CLIP [60]	73.2	67.9	60.2
C-PGS [47]	75.5	**70.8**	**67.9**
FILIP [17]	64.9	58.2	56.8
SPARC [3]	73.5	70.4	66.9
PowerCLIP-R	75.6	70.3	**67.9**
PowerCLIP-S	**76.1**	70.4	67.1

Table 4. **Compositionality evaluation on SugarCrepe.**

Method	Text	Image	Group
CLIP [49]	24.8	8.0	4.3
FLIP [35]	24.8	10.0	5.8
C-PGS [47]	**25.2**	10.5	7.2
FILIP [17]	21.3	13.5	9.7
SPARC [3]	23.3	12.7	9.0
PowerCLIP-R	22.5	9.5	6.5
PowerCLIP-S	24.8	**16.0**	**10.2**

Table 5. **Compositionality evaluation on Winoground.**

Method	Cls	Ret
PowerCLIP-S	42.2	47.0
w/o region sets	41.1	45.7
w/o parse trees	41.1	45.4
w/o R2T agg.	40.8	45.3
w/o T2R agg.	41.8	45.4
w/o Triplet loss	35.1	42.7

Table 6. Ablation study for key components.

Mask	M	Cls	Ret
Random	5	40.5	45.5
Random	10	41.5	45.8
Random	15	40.9	43.9
SAM	5	41.3	45.2
SAM	10	42.2	47.0
SAM	15	41.4	44.9

Table 7. Mask generation. M: Number of masks.

NLA-T1	NLA-T2	Cls	Ret
Softplus	Tanh	42.2	47.0
ReLU	Tanh	40.2	45.0
GELU	Tanh	41.8	44.9
Swish	Tanh	41.1	45.3
Softplus	Tanh	42.2	47.0
Softplus	Sigmoid	40.5	45.0
Softplus	SoftSign	41.0	44.9

Table 8. Activation functions.

Zero-Shot Image-Text Retrieval. Table 2 presents zero-shot retrieval results. PowerCLIP achieves consistent improvements over baseline methods, surpassing CLIP with an average gain of +4.3% for Recall@1 across both retrieval tasks. Notably, PowerCLIP surpasses baselines across all retrieval scenarios. These results demonstrate the effectiveness of compositional alignment between textual phrases and image region combinations in retrieval tasks.

Robustness. Table 3 compares PowerCLIP with baselines across six ImageNet robustness benchmarks. PowerCLIP significantly surpasses the baselines in terms of both in-distribution (ID) and out-of-distribution (OOD) average accuracy. Particularly notable is its performance on ImageNet-R (+5.9%) and ImageNet-Sketch (+4.0%), datasets designed to assess robustness under domain shifts. Overall, the results underscore the generalizability and robustness of PowerCLIP in challenging scenarios.

Compositionality. Tables 4 and 5 evaluate compositional understanding on SugarCrepe (average accuracies for object, attribute and relation subsets) and Winoground (text, image and overall group accuracy), respectively. Consistent with other evaluations, PowerCLIP significantly improves average accuracy over CLIP, confirming stronger compositional grounding of novel elements introduced in images. Performance improvements are particularly pronounced for the object subset of SugarCrepe (+2.2%) and for image retrieval on Winoground (+8.0%). These results demonstrate that explicit phrase-to-region alignment enhances fine-grained compositional understanding, aligning precisely with our motivation.

5.3. Analysis and Discussion

Ablation Study. Table 6 quantifies the contributions of key PowerCLIP components through systematic ablations: 1) replacing region sets with individual regions, 2) replacing parse trees with individual tokens, 3) omitting the R2T aggregation loss, 4) omitting the T2R aggregation loss, and 5) omitting the proposed triplet loss. Results confirm that each component contributes to the overall performance, underscoring their complementary roles.

Mask Generation. Table 7 investigates mask-generation methods by varying the number (M) and type of masks. SAM-generated masks achieve higher performance than random masks overall, with the best results obtained at $M = 10$. Random masks also maintain performance in the same range and do not break down when a sufficient number of masks is used. These results suggest that our method is relatively robust to both the mask generation strategy and the number of masks, while still benefiting from a modest performance gain when using SAM.

Activation Functions. Table 8 compares activation functions for NLAs. For NLA-T1, Softplus consistently outperforms ReLU, GELU and Swish due to its smooth approximation of max operations essential for T2R aggregation. For NLA T2, Tanh performs the best for approximating R2T. Sigmoid and SoftSign still capture the mapping but show clear drops in performance compared with Tanh, similar in scale to the differences seen in NLA T1. Thus, Tanh is the most suitable choice for NLA T2, while other smooth activations remain usable but less accurate.

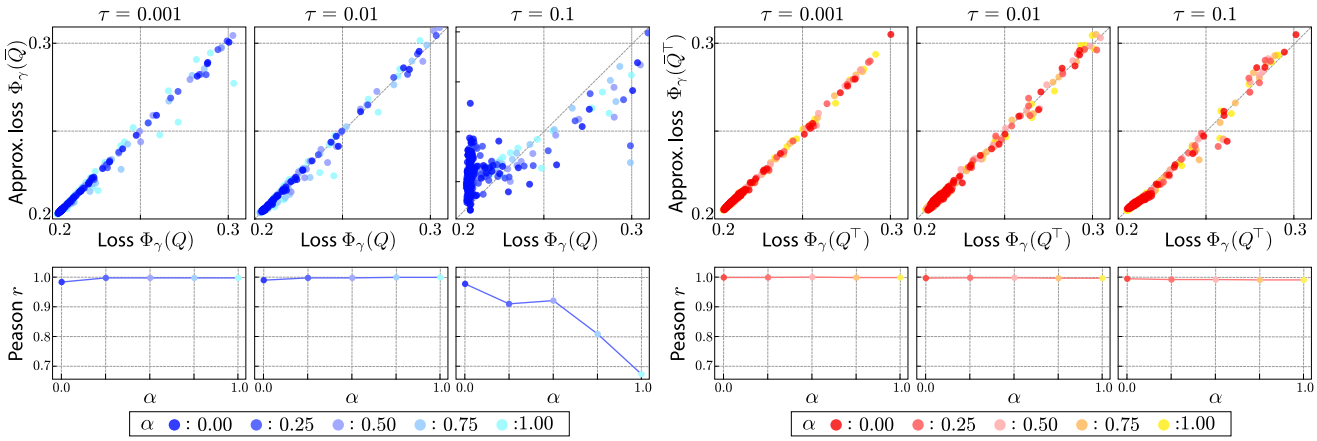

Figure 5. Approximation accuracy evaluation. Top: Comparison between exact and approximated losses for $\tau = \{0.1, 0.01, 0.001\}$ and $\alpha \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$. Bottom: Pearson correlation r between exact and approximated losses.

Figure 6. Visualizations of text-to-patch similarities. For each input text, we compute similarities between the text representation and image patch features and visualize them as heatmaps, showing that high responses are concentrated on the regions referred to in the text.

Method	CF10	CF100	IN1k
CLIP [49]	88.0	67.4	62.3
FLIP [35]	85.9	65.5	61.3
A-CLIP [67]	86.4	66.1	62.0
E-CLIP [60]	89.0	69.7	62.7
FILIP [17]	84.4	56.8	50.4
SPARC [3]	88.7	69.4	62.7
C-PGS [47]	90.0	**72.3**	64.4
PowerCLIP	**91.3**	**72.3**	**65.8**

Table 9. Linear probing.

Value	Cls	Ret
$\alpha = 0.00$	40.7	49.0
$\alpha = 0.25$	41.8	47.9
$\alpha = 0.50$	42.3	49.0
$\alpha = 0.75$	42.2	47.0
$\alpha = 1.00$	42.0	47.4
$\tau = 0.001$	42.2	47.0
$\tau = 0.01$	41.7	46.5
$\tau = 0.1$	40.7	45.6

Table 10. Hyperparameter.

Linear Probing. Table 9 presents linear probing results on CIFAR10 (CF10), CIFAR100 (CF100), and ImageNet-1k (IN1k). Consistent with zero-shot evaluations, PowerCLIP achieves the best performance. This indicates that PowerCLIP learns more discriminative features, enabling improved linear separability for classification tasks.

NLA Accuracy. Figure 5 analyzes the approximation accuracy of NLAs for the two triplet loss terms $\Phi_\gamma(Q)$ and $\Phi_\gamma(Q^\top)$. We observe that loss values approximated by NLAs closely match the exact values when τ is small (0.001 or 0.01), consistently achieving Pearson correlations above 0.98 across all tested α. The highest correlation (0.999) was obtained with $\tau = 0.001$ and $\alpha = 0.75$. Table 10 shows that performance stays high for $\alpha > 0.5$, and peaks at $\alpha = 0.75$. Similarly, τ follows the same tendency predicted by our theoretical analysis, with smaller values leading to better performance.

Quantitative Examples Figure 6 provides qualitative examples. For the illustrated examples, PowerCLIP produces text-to-patch similarity heatmaps whose high responses are concentrated on image regions corresponding to words explicitly mentioned in the text. Across different prompts, the highlighted patches consistently align with the referred objects and actions, indicating that the model attends to the intended visual evidence rather than unrelated areas.

We include computational cost comparisons in Sec. 7.

6. Conclusion

We introduced PowerCLIP, a novel contrastive pre-training framework that leverages powerset alignment. PowerCLIP exhaustively optimizes local-to-global alignments by minimizing bidirectional triplet losses defined over the powersets of image regions and textual parse trees. Extensive experimental results demonstrate that PowerCLIP achieves state-of-the-art performance across diverse benchmarks. For future work, extending PowerCLIP to 3D scene understanding presents a promising avenue for enhancing spatial and semantic alignment in more complex multimodal scenarios.

7. Acknowledgements

This work was supported by Japan Science and Technology Agency (JST) as part of Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE), Grant Number JPMJAP2518. This work was supported by the AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain". We used ABCI 3.0 provided by AIST and AIST Solutions with support from "ABCI 3.0 Development Acceleration Use". We would like to thank Yukito Tajima and Daisuke Nohara for their valuable support with the implementation of this work.

References

[1] Mothilal Asokan, Kebin Wu, and Fatima Albreiki. Finelip: Extending clip's reach via fine-grained alignment with longer text inputs. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 4

[2] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Proc. British Machine Vision Conference (BMVC)*, 2016. 4

[3] Ioana Bica, Anastasija Ili'c, Matthias Bauer, G"oker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A. Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, and Jovana Mitrovi'c. Improving fine-grained understanding in image-text pre-training. In *Proc. International Conference on Machine Learning (ICML)*, pages 3974–3995, 2024. 1, 2, 3, 6, 7, 8, 4, 5

[4] Bossard, Lukas, Guillaumin, Matthieu, Van Gool, and Luc. Food-101–mining discriminative components with random forests. In *Proc. European Conference on Computer Vision (ECCV)*, pages 446–461, 2014. 6

[5] Changpinyo, Soravit, Sharma, Piyush, Ding, Nan, Soricut, and Radu. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568, 2021. 6

[6] Cheng, Gong, Han, Junwei, Lu, and Xiaoqiang. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 6

[7] Choi, Hyungyu, Jang, Young Kyun, Eom, and Chanho. Goal: Global-local object alignment learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4070–4079, 2025. 1, 2

[8] Jiho Choi, Seonho Lee, Minhyun Lee, Seungho Lee, and Hyunjung Shim. Fine-grained image-text correspondence with cost aggregation for open-vocabulary part segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9782–9793, 2025. 2

[9] Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen tau Yih, Shang-Wen Li, and Hu Xu. Meta CLIP 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, pages 1–10, 2025. 2

[10] Cimpoi, Mircea, Maji, Subhransu, Kokkinos, Iasonas, Mohamed, Sammy, Vedaldi, and Andrea. Describing textures in the wild. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014. 6

[11] Coates, Adam, Ng, Andrew, Lee, and Honglak. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pages 215–223, 2011. 6

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 6

[13] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10995–11005, 2023. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *Proc. International Conference on Learning Representations (ICLR)*, 2021. 6

[15] Songsong Duan, Xi Yang, and Nannan Wang. DIH-CLIP: Unleashing the diversity of Multi-Head Self-Attention for Training-Free Open-Vocabulary semantic segmentation. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22794–22803, 2025. 2

[16] Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Shama Sastry, Evangelos E. Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 2

[17] Lewei Yao et al. Filip: Fine-grained interactive language-image pre-training. In *Proc. International Conference on Learning Representations (ICLR)*, 2022. 1, 2, 4, 6, 7, 8, 5

[18] Everingham, Mark, Van Gool, Luc, Williams, Christopher KI, Winn, John, Zisserman, and Andrew. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88:303–338, 2010. 6

[19] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 35544–35575, 2023. 2

[20] Fei-Fei, Li, Fergus, Rob, Perona, and Pietro. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 178–178, 2004. 6

[21] Zheren Fu, Lei Zhang, Hou Xia, and Zhendong Mao. Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26297–26306, 2024. 2

[22] Jiannan Ge, Lingxi Xie, Hongtao Xie, Pandeng Li, Sun-Ao Liu, Xiaopeng Zhang, Qi Tian, and Yongdong Zhang. Clip-adapted region-to-text learning for generative open-vocabulary semantic segmentation. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24034–24044, 2025. 2

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 2

[24] Helber, Patrick, Bischke, Benjamin, Dengel, Andreas, Borth, and Damian. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 6

[25] Hendrycks, Dan, Basart, Steven, Mu, Norman, Kadavath, Saurav, Wang, Frank, Dorundo, Evan, Desai, Rahul, Zhu, Tyler, Parajuli, Samyak, Guo, Mike, Song, Dawn, Steinhardt, Jacob, Gilmer, and Justin. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021. 6

[26] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, 2021. 6

[27] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, page 31096–31116, 2023. 2, 6

[28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. International Conference on Machine Learning (ICML)*, pages 4904–4916, 2021. 2

[29] 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 *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 27896–27918, 2024. 2

[30] Cijo Jose, Th'eo Moutakanni, Dahyun Kang, Federico Baldassarre, Timoth'ee Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Micha"el Ramamonjisoa, Maxime Oquab, Oriane Sim'eoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. DINOv2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24905–24916, 2025. 2

[31] Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is CLIP ideal? no. can we fix it? yes! In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22436–22446, 2025. 2

[32] Krause, Jonathan, Stark, Michael, Deng, Jia, Fei-Fei, and Li. 3d object representations for fine-grained categorization. In *Proc. IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 6

[33] Krizhevsky and Alex. Learning multiple layers of features from tiny images. *Technical Report and University of Tront*, 2009. 6

[34] Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, and Meng Cao. VeCLIP: Improving clip training via visual-enriched captions. In *Proc. European Conference on Computer Vision (ECCV)*, pages 111–127, 2024. 2

[35] Li, Yanghao, Fan, Haoqi, Hu, Ronghang, Feichtenhofer, Christoph, He, and Kaiming. Scaling language-image pre-training via masking. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23390–23400, 2023. 2, 6, 7, 8, 4, 5

[36] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[37] Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. Mask-Adapter: The Devil is in the Masks for Open-Vocabulary Segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14998–15008, 2025. 2

[38] Yunheng Li, Yuxuan Li, Quan-Sheng Zeng, Wenhai Wang, Qibin Hou, and Ming-Ming Cheng. Unbiased Region–Language alignment for Open-Vocabulary dense prediction. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23795–23805, 2025. 2

[39] Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, Zitnick, and C. Lawrence. Microsoft coco: Common objects in context. In *Proc. European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 6

[40] Maji, Subhransu, Rahtu, Esa, Kannala, Juho, Blaschko, Matthew, Vedaldi, and Andrea. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[41] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. SLIP: Self-supervision meets language-image pre-training. In *Proc. European Conference on Computer Vision (ECCV)*, pages 529–544, 2022. 2

[42] Mukhoti, Jishnu, Lin, Tsung-Yu, Poursaeed, Omid, Wang, Rui, Shah, Ashish, Torr, Philip H.S., Lim, and Ser-Nam. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[43] Nilsback, Maria-Elena, Zisserman, and Andrew. Automated flower classification over a large number of classes. In *Proc.*

Indian Conference on Computer Vision and Graphics & Image Processing, pages 722–729, 2008. 6

[44] Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. Know "No" better: A data-driven approach for enhancing negation awareness in CLIP. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2825–2835, 2025. 2

[45] Parkhi, Omkar M, Vedaldi, Andrea, Zisserman, Andrew, Jawahar, and CV. Cats and dogs. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3498–3505, 2012. 6

[46] 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 *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 32731–32760, 2024. 1, 2

[47] Gensheng Pei, Tao Chen, Yujia Wang, Xinhao Cai, Xiangbo Shu, Tianfei Zhou, and Yazhou Yao. Seeing what matters: Empowering CLIP with patch generation-to-selection. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24862–24872, 2025. 1, 2, 6, 7, 8, 4, 5

[48] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Yu Huang, Yaoming Wang, and Wei Shen. Parameter-efficient fine-tuning in hyperspherical space for open-vocabularysemantic segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15009–15020, 2025. 2

[49] Radford, Alec, Kim, Jong Wook, Hallacy, Chris, Ramesh, Aditya, Goh, Gabriel, Agarwal, Sandhini, Sastry, Girish, Askell, Amanda, Mishkin, Pamela, Clark, Jack, et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 1, 2, 6, 7, 8, 4, 5

[50] Ravi, Nikhila, Gabeur, Valentin, Hu, Yuan-Ting, Hu, Ronghang, Ryali, Chaitanya, Ma, Tengyu, Khedr, Haitham, Rädle, Roman, Rolland, Chloe, Gustafson, Laura, Mintun, Eric, Pan, Junting, Alwala, Kalyan Vasudev, Carion, Nicolas, Wu, Chao-Yuan, Girshick, Ross, Dollár, Piotr, Feichtenhofer, and Christoph. Sam 2: Segment anything in images and videos. In *Proc. International Conference on Learning Representations (ICLR)*, 2025. 3, 6

[51] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proc. International Conference on Machine Learning (ICML)*, pages 5389–5400, 2019. 6

[52] Stallkamp, Johannes, Schlipsing, Marc, Salmen, Jan, Igel, and Christian. The german traffic sign recognition benchmark: a multi-class classification competition. In *IJCNN*, pages 1453–1460, 2011. 6

[53] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2

[54] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-CLIP: A CLIP model focusing on wherever you want. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13019–13029, 2024. 2

[55] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 5228–5238, 2022. 2, 6

[56] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier H'enaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding and localization and and dense features. *arXiv prepreint arxiv:2502.14786*, 2025. 2

[57] Veeling, Bastiaan S, Linmans, Jasper, Winkens, Jim, Cohen, Taco, Welling, and Max. Rotation equivariant cnns for digital pathology. In *MICCAI*, pages 210–218, 2018. 6

[58] Bingchao Wang, Zhiwei Ning, Jianyu Ding, Xuanang Gao, Yin Li, Dongsheng Jiang, Jie Yang, and Wei Liu. Fix-clip: Dual-branch hierarchical contrastive learning via synthetic captions for better understanding of long text. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20694–20704, 2025. 2

[59] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 13754–13764, 2019. 6

[60] Wei, Zihao, Pan, Zixuan, Owens, and Andrew. Efficient vision-language pre-training by cluster masking. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26815–26825, 2024. 1, 2, 6, 7, 8, 4

[61] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. MaskFeat: Masked feature prediction for self-supervised visual pre-training. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14668–14678, 2022. 2

[62] Zhixiang Wei, Guangting Wang, Xiaoxiao Ma, Ke Mei, Huaian Chen, Yi Jin, and Fengyun Rao. Hq-clip: Leveraging large Vision-Language models to create high-quality image-text datasets and CLIP models. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22447–22456, 2025. 2

[63] Xiao, Jianxiong, Hays, James, Ehinger, Krista A, Oliva, Aude, Torralba, and Antonio. Sun database: Large-scale scene recognition from abbey to zoo. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, 2010. 6

[64] 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 *Proc. International Conference on Machine Learning (ICML)*, 2025. 2

[65] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple

framework for masked image modeling. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9643–9653, 2022. 2

[66] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *Proc. International Conference on Learning Representations (ICLR)*, 2024. 2

[67] Yang, Yifan, Huang, Weiquan, Wei, Yixuan, Peng, Houwen, Jiang, Xinyang, Jiang, Huiqiang, Wei, Fangyun, Wang, Yin, Hu, Han, Qiu, Lili, et al. Attentive mask clip. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2771–2781, 2023. 1, 2, 6, 7, 8, 4

[68] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[69] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research (TMLR)*, pages 1–20, 2022. 2

[70] Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words and and what to do about it? In *Proc. International Conference on Learning Representations (ICLR)*, 2023. 2

[71] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14393–14402, 2021. 5

[72] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023. 2

[73] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In *Proc. European Conference on Computer Vision (ECCV)*, pages 310–325, 2024. 2

[74] Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing patch correlations in CLIP for open-vocabulary semantic segmentation. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24677–24687, 2025. 2

[75] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based Language-Image pretraining. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16793–16803, 2022. 2

[76] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *Proc. European Conference on Computer Vision (ECCV)*, pages 696–712, 2022. 2