PowerCLIP: Powerset Alignment for Contrastive Pre-Training

Supplementary Material

Appendix A. Proof of Theorem 1

In this section, we present a proof of Theorem 1. We first restate the definitions of the T2R aggregation and NLAs.

A.1 Preliminary

Notation. Let $\mathcal{M}_i = \{R_m\}_{m=1}^{M}$ be the set of M region masks for the image I_i, and let $\mathcal{T}_j$ be the parse tree for the text description T_j with $K_j = |\mathcal{T}_j|$ nodes, where i and j index samples in mini-batches. For token masks $P_{m'} \in \mathrm{Leaf}(\mathcal{T}_j)$, we define

$$S^{(0)}_{i,j,m,m'} := \big\langle \phi(I_i|\, R_m),\, \psi(T_j|\, P_{m'}) \big\rangle, \tag{17}$$

where ϕ and ψ are image and text encoders, respectively, satisfying $\|\phi(\cdot)\|_2 = \|\psi(\cdot)\|_2 = 1$ so that $|S^{(0)}_{i,j,m,m'}| \le 1$. For any node $B \in \mathcal{T}_j$, we define aggregated similarities:

$$Q_{i,j,m,B} := \sum_{P_{m'} \in B} S^{(0)}_{i,j,m,m'}, \tag{18}$$

$$Q_{i,j,A,B} := \sum_{R_m \in A} Q_{i,j,m,B}, \tag{19}$$

$$\bar{Q}_{i,j,B} := \sum_{R_m \in \mathcal{M}_i} Q_{i,j,m,B}. \tag{20}$$

Note that, we have $Q_{i,j,A,B} = \langle \boldsymbol{r}^{(i)}_A,\, \boldsymbol{p}^{(j)}_B \rangle$ by bilinearity.

T2R Aggregation. Given similarity $Q_{i,j,A,B}$, we define the T2R similarity as

$$Q^{\leftarrow}_{i,j} := \frac{1}{K_j} \sum_{B \in \mathcal{T}_j} Q^{\leftarrow}_{i,j,B}, \tag{21}$$

where $Q^{\leftarrow}_{i,j,B}$ evaluates the best-matching region subset A for each node B as

$$Q^{\leftarrow}_{i,j,B} = \max_{A \subseteq \mathcal{M}_i} Q_{i,j,A,B}. \tag{22}$$

Non-Linear Aggregators (NLAs). Given $S^{(0)}$, we define the three-layer NLAs with a hyperparameter $\alpha \in [0,1]$ and activation functions $\{\sigma_l\}_{l=1}^{3}$:

$$S^{(1)}_{i,j,m|B} := \sigma_1\left(\sum_{P_{m'} \in B} S^{(0)}_{i,j,m,m'} \right), \tag{23}$$

$$S^{(2)}_{i,j|B} := \sigma_2\left(\sum_{R_m \in \mathcal{M}_i} S^{(1)}_{i,j,m|B} \right), \tag{24}$$

$$S^{(3)}_{i,j} := \sigma_3\left(\frac{1}{K_j^{\,1-\alpha}} \sum_{B \in \mathcal{T}_j} S^{(2)}_{i,j|B} \right). \tag{25}$$

Lemma A.1 (LSE Bound). *Given aggregated similarities $Q_{i,j,A,B}$ and $Q^{\leftarrow}_{i,j,B}$, for any $\tau > 0$, we have*

$$\left| \tau \log \sum_{A \subseteq \mathcal{M}_i} \exp\left(\frac{Q_{i,j,A,B}}{\tau} \right) - Q^{\leftarrow}_{i,j,B} \right| \le \tau M \log 2. \tag{26}$$

Proof. Considering the log-sum-exp (LSE) bound, the Lemma immediately holds; that is, we have

$$\left| \tau \log \sum_{A \subseteq \mathcal{M}_i} \exp\left(\frac{Q_{i,j,A,B}}{\tau} \right) - Q^{\leftarrow}_{i,j,B} \right| \tag{27}$$

$$\le \tau \log \sum_{A \subseteq \mathcal{M}_i} \exp\left(\frac{Q_{i,j,A,B} - Q^{\leftarrow}_{i,j,B}}{\tau} \right) \tag{28}$$

$$\le \tau \log \sum_{A \subseteq \mathcal{M}_i} 1 \tag{29}$$

$$= \tau \log \left| 2^{\mathcal{M}_i} \right| \tag{30}$$

$$= \tau M \log 2. \tag{31}$$

This completes the proof. □

A.2 NLA-T1

We define NLA-T1 and provide a proof for Theorem 1.

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

$$\sigma_1(x) = \tau \cdot \mathrm{Act}\left(\frac{x}{\tau}\right),\ \sigma_2 = \sigma_3 = \mathrm{Id},\ \alpha = 0 \tag{32}$$

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 $|S^{(3)}_{i,j} - Q^{\leftarrow}_{i,j}| < \epsilon$.*

Proof. From Definition 1, the output for the second layer of NLA-T1 is given by

$$S^{(2)}_{i,j|B} = \sum_{R_m \in \mathcal{M}_i} \tau \cdot \mathrm{Act}\left(\sum_{P_{m'} \in B} \frac{S^{(0)}_{i,j,m,m'}}{\tau} \right). \tag{33}$$

When the softplus function is used for the activation func-

tion, we have

$$S^{(2)}_{i,j|B} = \sum_{R_m \in \mathcal{M}_i} \tau \cdot \mathrm{Softplus}\left(\sum_{P_{m'} \in B} \frac{S^{(0)}_{i,j,m,m'}}{\tau} \right) \tag{34}$$

$$= \sum_{R_m \in \mathcal{M}_i} \tau \log \left(1 + \exp \left(\sum_{P_{m'} \in B} \frac{S^{(0)}_{i,j,m,m'}}{\tau} \right) \right) \tag{35}$$

$$= \tau \log \prod_{R_m \in \mathcal{M}_i} \left(1 + \exp \left(\sum_{P_{m'} \in B} \frac{S^{(0)}_{i,j,m,m'}}{\tau} \right) \right) \tag{36}$$

$$= \tau \log \sum_{A \subseteq \mathcal{M}_i} \prod_{R_m \in A} \exp \left(\sum_{P_{m'} \in B} \frac{S^{(0)}_{i,j,m,m'}}{\tau} \right) \tag{37}$$

$$= \tau \log \sum_{A \subseteq \mathcal{M}_i} \exp \left(\sum_{R_m \in A} \sum_{P_{m'} \in B} \frac{S^{(0)}_{i,j,m,m'}}{\tau} \right) \tag{38}$$

$$= \tau \log \sum_{A \subseteq \mathcal{M}_i} \exp \left(\frac{Q_{i,j,A,B}}{\tau} \right) \tag{39}$$

Then, from Lemma A.1, we have

$$\left| S^{(2)}_{i,j|B} - Q^{\leftarrow}_{i,j,B} \right| = \left| \tau \log \sum_{A \subseteq \mathcal{M}_i} \exp \left(\frac{Q_{i,j,A,B}}{\tau} \right) - Q^{\leftarrow}_{i,j,B} \right| \tag{40}$$

$$\leq \tau M \log 2. \tag{41}$$

Hence, for any $\epsilon > 0$, choosing $\tau < \epsilon/(M \log 2)$ ensures $\left| S^{(3)}_{i,j} - Q^{\leftarrow}_{i,j} \right| < \epsilon$. Equivalently, $S^{(3)}_{i,j} \to Q^{\leftarrow}_{i,j}$ holds as $\tau \to 0^+$. This completes the proof. □

Appendix B. Proof of Theorem 2

In this section, we present a proof of Theorem 2. We first restate the definitions of the R2T aggregation and NLA-T2.

B.1 Preliminary

R2T Aggregation We use the notation in Appendix A.1. Given similarity $Q_{i,j,A,B}$, we define the R2T similarity as

$$Q^{\rightarrow}_{i,j} := \frac{1}{2^M} \sum_{A \subseteq \mathcal{M}_i} O_{i,j,A}, \tag{42}$$

where $O_{i,j,A}$ evaluates the best-matching node B for each region subset A as

$$O_{i,j,A} = \max_{B \in \mathcal{T}_j} Q_{i,j,A,B}. \tag{43}$$

Exponential Aggregation. Given similarity $Q_{i,j,A,B}$, we define exponential aggregation $E_{i,j,B}$ as the sum of exponential similarities over the powerset of $\mathcal{M}_i$ with a temperature $\tau > 0$, *i.e.*,

$$E_{i,j,B} = \sum_{A \subseteq \mathcal{M}_i} \exp \left(\frac{Q_{i,j,A,B}}{\tau} \right). \tag{44}$$

Bounding functions. For convenience, we define four auxiliary functions to evaluate upper and lower bounds:

$$\Gamma_B(\alpha) := \frac{1-\alpha}{2} \bar{Q}_{i,j,B} + \alpha \max_A Q_{i,j,A,B} \tag{45}$$

$$\bar{\Gamma}_B(\tau, \alpha) := \frac{1-\alpha}{2} \bar{Q}_{i,j,B} + \alpha \, \tau \log E_{i,j,B} \tag{46}$$

$$\Lambda(\alpha) := \max_{B \in \mathcal{T}_j} \Gamma_B(\alpha) \tag{47}$$

$$\bar{\Lambda}(\tau, \alpha) := \tau \log \left(\sum_{B \in \mathcal{T}_j} \exp \left(\frac{\bar{\Gamma}_B(\tau, \alpha)}{\tau} \right) \right) \tag{48}$$

Lemma B.1 (Summation Over Powerset). *For any similarity $Q_{i,j,A,B}$, we have*

$$E_{i,j,B} = 2^M \exp\left(\frac{\bar{Q}_{i,j,B}}{2\tau} \right) \prod_{R_m \in \mathcal{M}_i} \cosh \left(\frac{Q_{i,j,m,B}}{2\tau} \right). \tag{49}$$

Proof. For any $x_m \in \mathbb{R}$ $(m = 1, 2, \cdots M)$, we have

$$\cosh(x_m) = \frac{1}{2} \left(\exp(x_m) + \exp(-x_m) \right) \tag{50}$$

and thus

$$\prod_{m=1}^{M} \cosh(x_m) = \frac{1}{2^M} \prod_{m=1}^{M} \left(\exp(x_m) + \exp(-x_m) \right) \tag{51}$$

$$= \frac{1}{2^M} \sum_{A \subseteq [M]} \left(\prod_{m \in A} \exp(x_m) \prod_{m \notin A} \exp(-x_m) \right) \tag{52}$$

$$= \frac{1}{2^M} \sum_{A \subseteq [M]} \exp \left(\sum_{m=1}^{M} (2\chi_A(m) - 1) x_m \right) \tag{53}$$

$$= \frac{1}{2^M} \exp(-\bar{x}) \sum_{A \subseteq [M]} \exp \left(2 \sum_{m \in A} x_m \right) \tag{54}$$

where

$$\chi_A(m) = \begin{cases} 1 & (m \in A) \\ 0 & (\text{otherwise}) \end{cases} \tag{55}$$

$$\bar{x} = \sum_{m=1}^{M} x_m \tag{56}$$

By substituting $x_m = Q_{i,j,m,B}/(2\tau)$, we obtain

$$\prod_{R_m \in \mathcal{M}_i} \cosh \left(\frac{Q_{i,j,m,B}}{2\tau} \right) = \frac{1}{2^M} \exp \left(-\frac{\bar{Q}_{i,j,B}}{2\tau} \right) \underbrace{\sum_{A \subseteq \mathcal{M}_i} \exp \left(\frac{Q_{i,j,A,B}}{\tau} \right)}_{E_{i,j,B}}. \tag{57}$$

Therefore, we have

$$E_{i,j,B} = 2^M \exp\left(\frac{\bar{Q}_{i,j,B}}{2\tau}\right) \prod_{R_m \in \mathcal{M}_i} \cosh\left(\frac{Q_{i,j,m,B}}{2\tau}\right) \quad (58)$$

This completes the proof. □

Lemma B.2 (LSE Bound) For any $\tau > 0$ and $\alpha \in [0, 1]$, we have

$$\Lambda(\alpha) \leq \bar{\Lambda}(\tau, \alpha) \leq \Lambda(\alpha) + \tau Z_j \quad (59)$$

where $Z_j = \alpha M \log 2 + \log K_j$.

Proof. By the LSE inequality, for any $\tau > 0$ we have

$$\max_A Q_{i,j,A,B} \leq \tau \log E_{i,j,B} \leq \max_A Q_{i,j,A,B} + \tau \log 2^M. \quad (60)$$

Multiplying by $\alpha \in [0, 1]$ and adding $\frac{1-\alpha}{2} \bar{Q}_{i,j,B}$ to all terms gives

$$\Gamma_B(\alpha) \leq \bar{\Gamma}_B(\tau, \alpha) \leq \Gamma_B(\alpha) + \tau\alpha M \log 2. \quad (61)$$

Next, applying the LSE inequality over B to $\bar{\Lambda}(\tau, \alpha)$, we obtain

$$\max_B \bar{\Gamma}_B(\tau, \alpha) \leq \bar{\Lambda}(\tau, \alpha) \leq \max_B \bar{\Gamma}_B(\tau, \alpha) + \tau \log K_j. \quad (62)$$

Combining this with Eq. (61), we obtain

$$\Lambda(\alpha) \leq \bar{\Lambda}(\tau, \alpha) \leq \Lambda(\alpha) + \tau\alpha M \log 2 + \tau \log K_j. \quad (63)$$

This completes the proof. □

B.2 NLA-T2

We define NLA-T2 and provide a proof for Theorem 2.

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

$$\sigma_1(x) = \zeta_\alpha\left(\frac{x}{2\tau}\right), \sigma_2(x) = \exp(x), \ \sigma_3(x) = \tau \log(x), \quad (64)$$

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* $\left|S^{(3)}_{i,j} - Q^{\rightarrow}_{i,j}\right| < \epsilon$.

Proof. With $\mathrm{Act} = \tanh$, we have

$$\int \tanh(x)\,\mathrm{d}x = \log\cosh(x) \quad (65)$$

and thus, with $\zeta_\alpha(0) = 0$, we obtain

$$\zeta_\alpha(x) = x + \alpha \,\mathrm{logcosh}(x). \quad (66)$$

Then, the output of the first layer is given by

$$S^{(1)}_{i,j,m|B} = \zeta_\alpha\left(\frac{1}{2\tau} \sum_{P_{m'} \in B} S^{(0)}_{i,j,m,m'}\right) \quad (67)$$

$$= \frac{Q_{i,j,m,B}}{2\tau} + \alpha \,\mathrm{logcosh}\left(\frac{Q_{i,j,m,B}}{2\tau}\right). \quad (68)$$

Applying $\sigma_2(x) = \exp(x)$ at the second layer, we have,

$$S^{(2)}_{i,j|B} = \exp\left(\sum_{R_m \in \mathcal{M}_i} S^{(1)}_{i,j,m|B}\right) \quad (69)$$

$$= \exp\left(\frac{\bar{Q}_{i,j,B}}{2\tau}\right) \prod_{R_m \in \mathcal{M}_i} \cosh^\alpha\left(\frac{Q_{i,j,m,B}}{2\tau}\right). \quad (70)$$

From Lemma B.1, we have

$$S^{(2)}_{i,j|B} = 2^{-\alpha M} \exp\left(\frac{1-\alpha}{2\tau} \bar{Q}_{i,j,B}\right) (E_{i,j,B})^\alpha \quad (71)$$

$$= 2^{-\alpha M} \exp\left(\frac{\bar{\Gamma}_B(\tau, \alpha)}{\tau}\right), \quad (72)$$

Applying $\sigma_3(x) = \tau \log(x)$ at the third layer, we have

$$S^{(3)}_{i,j} = \tau \log\left(\frac{1}{|\mathcal{T}_j|^{1-\alpha}} \sum_{B \in \mathcal{T}_j} S^{(2)}_{i,j|B}\right) \quad (73)$$

$$= -\tau Z^\alpha_j + \tau \log\left(\sum_{B \in \mathcal{T}_j} \exp\left(\frac{\bar{\Gamma}_B(\tau, \alpha)}{\tau}\right)\right) \quad (74)$$

$$= -\tau Z^\alpha_j + \bar{\Lambda}(\tau, \alpha), \quad (75)$$

where $Z^\alpha_j = \alpha M \log 2 + (1-\alpha) \log K_j$. From Lemma B.2, we obtain the quantitative bounds

$$\Lambda(\alpha) - \tau Z^\alpha_j \leq S^{(3)}_{i,j} \leq \Lambda(\alpha) + \tau\alpha \log K_j. \quad (76)$$

In particular,

$$\lim_{\tau \to 0^+} S^{(3)}_{i,j} = \Lambda(\alpha). \quad (77)$$

Since $\mathbb{E}_A[Q_{i,j,A,B}] = \frac{1}{2}\bar{Q}_{i,j,B}$ under the uniform distribution over $2^{\mathcal{M}_i}$, Jensen's inequality for the pointwise maximum implies

$$\Lambda(0) = \frac{1}{2} \max_B \bar{Q}_{i,j,B} \leq Q^{\rightarrow}_{i,j} \leq \max_{A,B} Q_{i,j,A,B} = \Lambda(1). \quad (78)$$

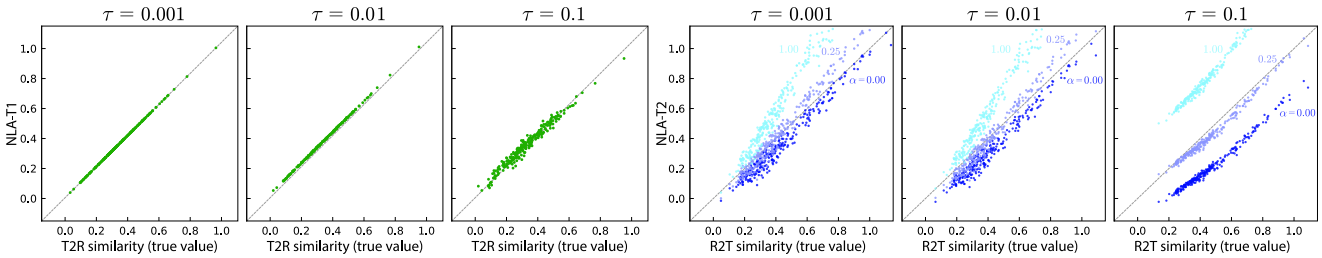

Figure 7. Approximation accuracy evaluation for NLA-T1 and NLA-T2.

Method	Zero-shot classification (Top-1)																
	Food101	CIFAR10	CIFAR100	SUN397	Cars	VOC07	Aircraft	DTD	Pets	Cal101	Flowers	STL10	EuroSAT	RESISC45	GTSRB	Country	PCam
PreviousSOTA	46.5	74.3	37.3	47.5	19.9	**61.1**	**3.1**	19.8	58.1	73.7	30.7	**92.8**	**33.2**	30.4	**10.9**	4.8	50.8
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
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**

Method	Zero-shot retrieval R@1						Robustness (Top-1)						SugarCrepe			Winoground	
	COCO-T	F8K-T	F30K-T	COCO-I	F8K-I	F30K-I	IN-1k	IN-V2	IN-A	IN-R	IN-O	IN-S	Obj	Att	Rel	Text	Image
PreviousSOTA	36.0	58.3	59.9	25.1	44.4	47.1	38.6	33.1	9.6	48.1	**42.6**	25.6	75.5	**70.8**	**67.9**	**25.2**	13.5
PowerCLIP-R	36.7	58.5	61.7	26.3	44.8	46.6	40.3	34.8	11.2	53.2	40.2	28.7	75.6	70.3	**67.9**	22.5	9.5
PowerCLIP-S	**37.3**	**58.6**	**62.4**	**27.0**	**46.3**	**50.4**	**40.8**	**35.1**	**11.9**	**53.5**	40.5	**28.9**	**76.1**	70.4	67.1	24.8	**16.0**

Table 11. Detailed unified comparison of PreviousSOTA (best over CLIP–SPARC), PowerCLIP-R, and PowerCLIP-S. Top: 17 zero-shot classification datasets. Bottom: 6 zero-shot retrieval settings (R@1), 6 ImageNet robustness benchmarks, SugarCrepe compositionality, and Winoground compositionality.

Because $\Lambda(\alpha)$ is continuous in $\alpha \in [0,1]$ (being the maximum of finitely many affine functions of α), there exists $\alpha^\star \in [0,1]$ such that $\Lambda(\alpha^\star) = Q^{\rightarrow}_{i,j}$.

Finally, by the quantitative bounds above, we obtain

$$\left|S^{(3)}_{i,j} - Q^{\rightarrow}_{i,j}\right| = \left|S^{(3)}_{i,j} - \Lambda(\alpha^\star)\right| \tag{79}$$

$$\leq \tau(\alpha^\star M \log 2 + \log K_j)\,. \tag{80}$$

Hence, for any $\epsilon{>}0$, choosing $\tau{<}\epsilon/(\alpha^\star M \log 2 + \log K_j)$ ensures $\left|S^{(3)}_{i,j} - Q^{\rightarrow}_{i,j}\right| < \epsilon$. This shows that NLA-T2 approximates $Q^{\rightarrow}_{i,j}$ with arbitrary precision, completing the proof. □

Appendix C. Approximation Accuracy

To quantitatively evaluate the approximation accuracy of NLAs, we compare the approximated similarity values with the true values. Figure 7 shows the outputs from NLA-T1 and NLA-T2 compared against the true T2R and R2T similarity values computed on synthetic data (randomly generated input vectors) with all parameters initialized randomly. For NLA-T1, approximations closely correlate with the true values. For NLA-T2, although the distribution displays greater variance compared to NLA-T1, the correlation remains strong. Additionally, we see that NLA-T2 with $\alpha = 1.0$ and $\alpha = 0.0$ corresponds to the upper and lower bounds of the R2T similarity, respectively. However, when $\tau = 0.1$, approximations are biased, resulting in larger errors during loss computation. These results are consistent with our theoretical analysis and validate the effectiveness of the proposed NLAs.

Appendix D. Details of Evaluation Task

For a comprehensive comparison with prior methods, Table 11 provides the tabular counterpart of the performance summary shown in Figure 2. The table reports results on 17 zero-shot classification datasets, 6 zero-shot retrieval R@1 settings, 6 ImageNet-based robustness benchmarks, as well as compositional generalization performance on SugarCrepe and Winoground under a unified evaluation protocol. For SugarCrepe, Table 4 in the main paper reports only the average scores for each method, whereas Table 12 presents a more fine-grained breakdown. In particular, our method tends to show better performance than prior

Method	SC-REPLACE			SC-SWAP		SC-ADD	
	Obj	Att	Rel	Obj	Att	Obj	Att
CLIP [49]	85.8	**79.2**	64.5	61.8	58.7	74.2	68.4
FLIP [35]	84.1	75.9	66.0	60.2	61.6	71.7	63.2
A-CLIP [67]	86.6	75.5	63.2	52.4	63.1	71.6	66.8
E-CLIP [60]	86.9	73.5	60.2	59.4	63.4	73.3	66.8
C-PGS [47]	88.1	76.0	**67.9**	**64.1**	66.5	74.2	69.9
FILIP [17]	82.9	61.9	56.8	58.4	58.3	53.4	54.3
SPARC [3]	85.2	75.5	66.9	58.8	**67.4**	76.4	68.6
PowerCLIP-R	**88.3**	76.6	67.8	60.8	64.7	77.8	69.7
PowerCLIP-S	87.5	77.5	67.1	61.6	63.1	**79.1**	**70.7**

Table 12. Detailed compositionality evaluation on SugarCrepe.

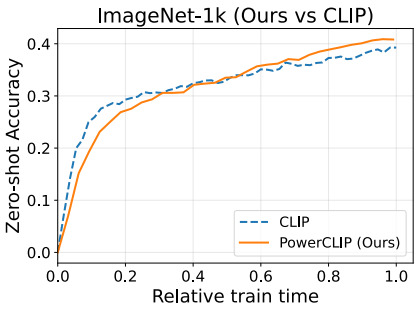

Figure 8. **ImageNet-1k zero-shot accuracy vs relative training time.** CLIP is trained for more epochs so that its total compute matches our method.

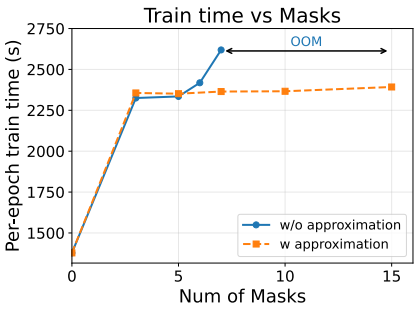

Figure 9. **Per-epoch training time vs number of masks K with and without approximation.** Without approximation, runs with $K>7$ fail due to OOM.

Method	Train time (s)	Rel. to CLIP
CLIP [49]	1378	1.00×
SPARC [3]	1730	1.26×
FILIP [17]	1947	1.41×
PowerCLIP	2366	1.72×

Table 13. **Per-epoch training time of each method under our training setup.** We report wall-clock time in seconds and relative cost normalized by CLIP.

approaches in the replace and add settings, suggesting that it more effectively improves text–image consistency under more challenging compositional transformations.

Appendix E. Computational Cost

Table 13 reports the per-epoch training time compared to prior work, showing that our method incurs about a 1.72× higher cost than CLIP due to the additional computation from region-level features and parse-tree reasoning. To account for this overhead and compare under a matched compute budget, we train CLIP for roughly 1.72× more epochs (from 32 to about 55 epochs) and evaluate ImageNet-1k zero-shot performance. Figure 8 shows the resulting accuracy as a function of relative training time, demonstrating that even under the same total training cost, our method still outperforms CLIP. Figure 9 shows how the per-epoch training time changes with the number of masks, with and without our approximation. Without approximation, the training time already starts to grow noticeably at 6 masks, and increasing the number of masks beyond 7 leads to out-of-memory (OOM) failures. In contrast, with our approximation, we can safely scale the number of masks up to 15 while keeping the per-epoch training time only mildly increased. This demonstrates that the proposed approximation effectively reduces both computation and memory overhead, enabling the use of richer region-level information within a practical training budget.

Appendix F. Ablation Study on λ

We study the sensitivity of PowerCLIP to the mixing coefficient λ under the main training setting in Sec. 3.4. Table 14 reports classification (Cls) and retrieval (Ret) performance for $\lambda \in \{0.1, 0.2, 0.3\}$. While retrieval is relatively stable, classification varies more across λ. We therefore use $\lambda = 0.1$ in all experiments as it yields a balanced trade-off between retrieval and classification.

Value	Cls	Ret
$\lambda=0.1$	42.2	47.0
$\lambda=0.2$	40.7	47.6
$\lambda=0.3$	41.3	47.1

Table 14. Ablation for λ.

Method	AP_{50}^{base}	AP_{50}^{novel}	AP_{50}
CLIP [49]	22.8	1.4	17.2
FLIP [35]	24.1	0.9	18.0
FILIP [17]	21.6	3.2	16.8
SPARC [3]	20.8	7.1	17.2
C-PGS [47]	23.1	2.3	17.7
PowerCLIP	27.6	15.3	24.4

Table 15. Evaluation on OV-COCO.

Appendix G. Open-Vocabulary Evaluation on OV-COCO

To further examine fine-grained recognition beyond closed-set benchmarks, we evaluate PowerCLIP on OV-COCO [71]. As shown in Table 15, PowerCLIP improves over CLIP by 7.2 points in AP_{50} and by 13.9 points in AP_{50}^{novel}. These gains indicate that PowerCLIP captures finer-grained visual–text correspondences, particularly for novel categories.

Appendix H. Qualitative Examples

Figure 10 shows a side-by-side comparison of text–image patch similarity heatmaps with existing models for the same inputs as in Figure 6. Compared to prior methods, our model produces sharper and more localized activations for words, indicating a closer alignment between the textual structure and the corresponding image regions. Figure 11 illustrates the compositional reasoning ability of our model by intentionally altering the order and attributes of objects in the text. When we apply such compositional edits to the caption, the high-similarity regions in the image shift accordingly to the appropriate patches, demonstrating that our model maintains semantically consistent text–image correspondences under such compositional transformations.

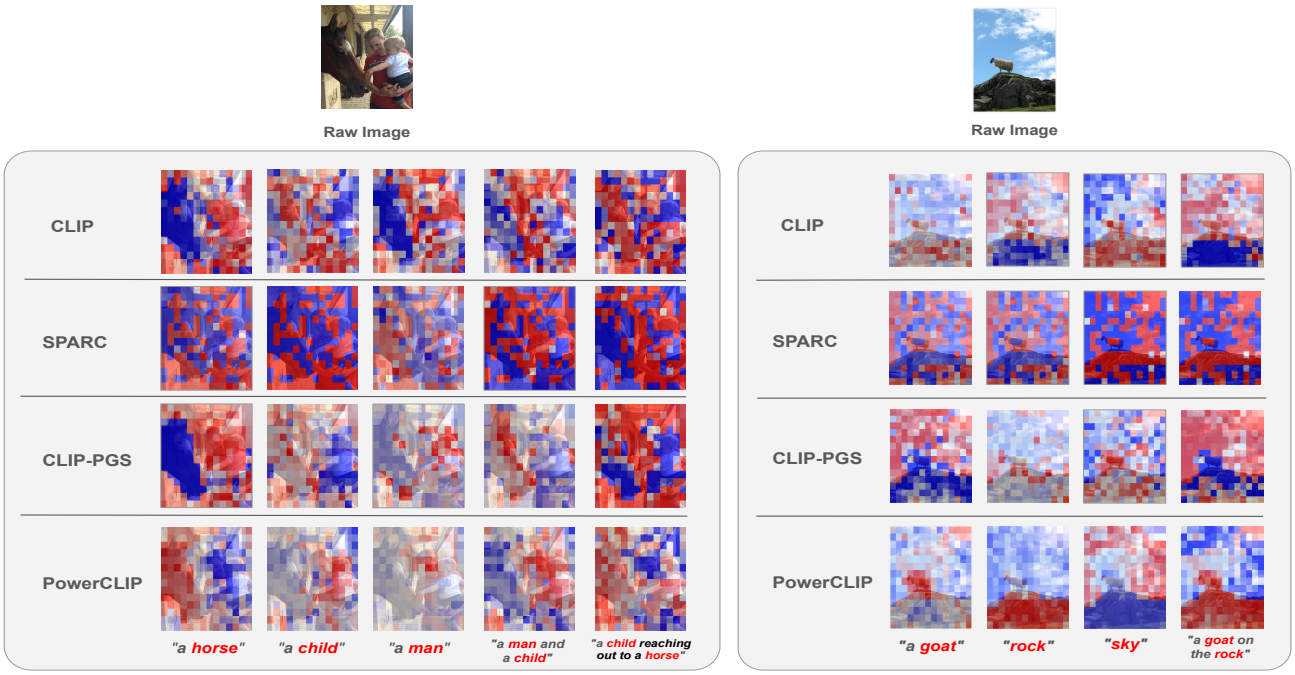

Figure 10. Qualitative comparison of text-to-patch similarity heatmaps across different models.

Figure 11. Qualitative examples of compositional reasoning.