Seeing What Matters: Empowering CLIP with Patch Generation-to-Selection

Gensheng Pei[1], Tao Chen[1], Yujia Wang[2], Xinhao Cai[1], Xiangbo Shu[1], Tianfei Zhou[3], Yazhou Yao[1]*
[1]Nanjing University of Science and Technology, [2]Zhejiang Sci-Tech University, [3]Beijing Institute of Technology
https://github.com/NUST-Machine-Intelligence-Laboratory/CLIP-PGS

Abstract

The CLIP model has demonstrated significant advancements in aligning visual and language modalities through large-scale pre-training on image-text pairs, enabling strong zero-shot classification and retrieval capabilities on various domains. However, CLIP's training remains computationally intensive, with high demands on both data processing and memory. To address these challenges, recent masking strategies have emerged, focusing on the selective removal of image patches to improve training efficiency. Although effective, these methods often compromise key semantic information, resulting in suboptimal alignment between visual features and text descriptions. In this work, we present a concise yet effective approach called Patch Generation-to-Selection (CLIP-PGS) to enhance CLIP's training efficiency while preserving critical semantic content. Our method introduces a gradual masking process in which a small set of candidate patches is first pre-selected as potential mask regions. Then, we apply Sobel edge detection across the entire image to generate an edge mask that prioritizes the retention of the primary object areas. Finally, similarity scores between the candidate mask patches and their neighboring patches are computed, with optimal transport normalization refining the selection process to ensure a balanced similarity matrix. Our approach, CLIP-PGS, sets new state-of-the-art results in zero-shot classification and retrieval tasks, achieving superior performance in robustness evaluation and language compositionality benchmarks.

1. Introduction

The rise of large-scale vision-language models (VLMs) [4, 17, 26, 28, 30, 37, 44, 48, 49, 51, 68] has revolutionized the field of visual representation learning. Pioneering works such as CLIP [45] and ALIGN [23] have demonstrated the potential of contrastive learning to align visual features with natural language descriptions. By leveraging massive datasets [6, 7, 16, 47, 52, 53] composed of image-text pairs collected from the internet, these models learn powerful and transferable visual representations that can be applied to a wide range of downstream tasks [5, 42, 50, 56, 61, 62, 65, 66] without task-specific fine-tuning. For example, CLIP uses a simple yet effective dual-encoder architecture that processes images and text independently and aligns them using contrastive loss. This approach has enabled zero-shot classification, showing competitive performance across various datasets without additional data-specific training.

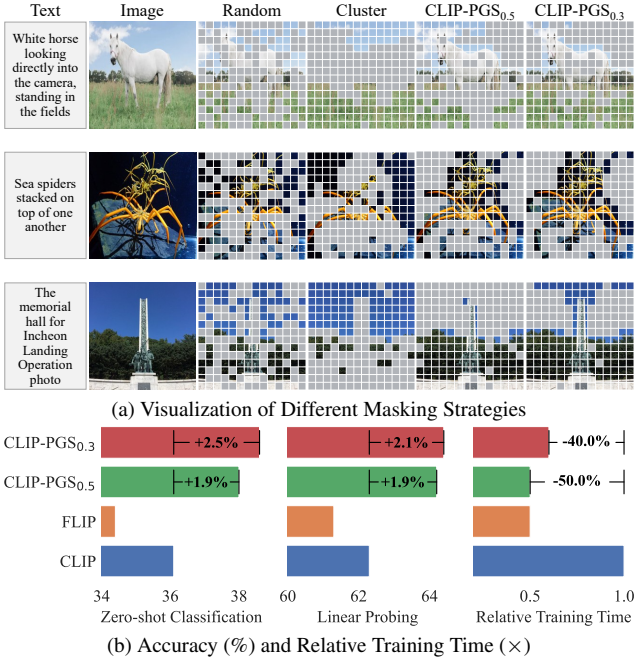


Figure 1. **Advantages of CLIP-PGS**. (a) Visual comparison of masking strategies: random masking (*e.g.*, FLIP [31]), cluster-based masking (*e.g.*, E-CLIP [55]), and our proposed CLIP-PGS. (b) Improvements in zero-shot classification and linear probing tasks, and relative training time reduction achieved by CLIP-PGS.

However, models in vision-language pre-training face the ongoing challenge of high computational costs. Their reliance on vast datasets and complex objectives necessitates extensive GPU resources, making efficient training difficult. Recent research has thus aimed to improve efficiency while maintaining competitive performance, intro-

*Corresponding author.

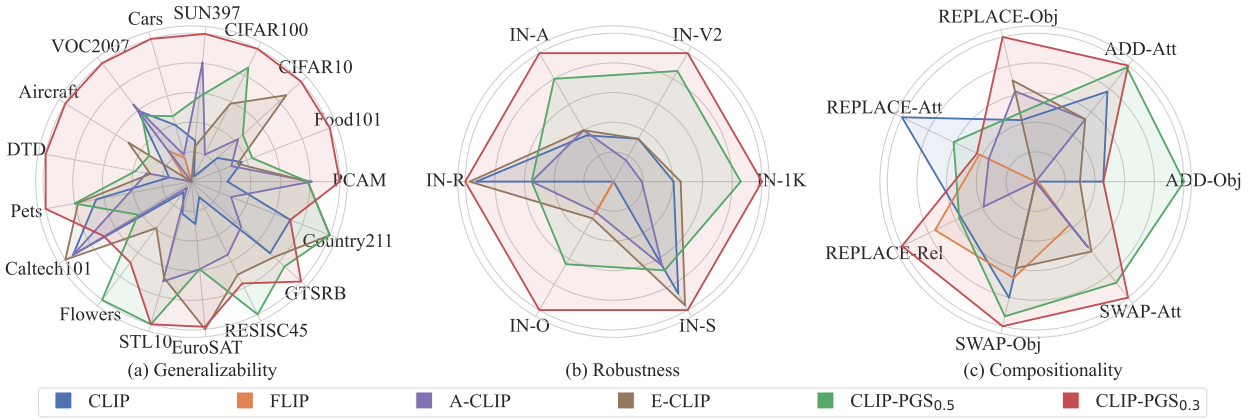


Figure 2. **Performance comparison of vision-language pre-training models**, such as CLIP [45], FLIP [31], A-CLIP [59], E-CLIP [55], and CLIP-PGS, evaluated across three dimensions using normalized scores: (a) generalizability, (b) robustness, and (c) compositionality.

ducing innovative strategies in data augmentation [15, 30], masking [29, 31, 59], and architectural refinement [38, 53].

Within computer vision, masked image modeling (MIM) has emerged as a popular self-supervised learning strategy, exemplified by models such as Masked Autoencoders (MAE) [19] and BEiT [2]. These models mask parts of the image input, guiding the model to reconstruct the obscured content and, in turn, acquire robust feature representations. Inspired by the efficiency gains of MIM, vision-language models have adapted masking techniques to balance computational efficiency and representational quality. Addressing CLIP's high computational demands, recent developments like FLIP [31] introduced random patch masking, allowing the model to process more sample pairs within the same training period by masking 50%∼75% of image patches, enhancing scalability for large-scale datasets. Building on these foundations, MaskCLIP [10] integrates masked image modeling with contrastive language-image training, using self-distillation to align local patch features with semantic text descriptions, which strengthens generalization and transferability. Further advances include A-CLIP [59], which builds on FLIP by introducing an adaptive masking mechanism. A-CLIP retains only the image tokens most semantically related to the paired text, reducing the negative effects of random token removal and ensuring that retained patches maintain meaningful context for visual-text alignment. E-CLIP[55] refines this approach by clustering visually similar patches for masking, preserving or removing coherent visual structures as a group.

Despite the progress made by recent masking strategies in vision-language pre-training, these models still face limitations that may compromise semantic alignment and representation quality. Random masking, as in FLIP [31], can inadvertently remove critical image content, while attention-based methods like A-CLIP [59] often require additional computational modules, adding to the training complexity. Cluster-based masking in E-CLIP [55] helps preserve coherent structures but lacks the granularity needed to selectively retain primary semantic regions, and it may unintentionally obscure regions aligned with textual descriptions.

To address these challenges, we propose CLIP-PGS, a concise masking strategy designed to enhance CLIP's training efficiency while carefully preserving essential semantic content, as shown in Fig. 1a. CLIP-PGS uses a gradual generation-to-selection process: it begins by pre-selecting candidate patches, applies Sobel edge detection to prioritize primary object areas, and then dynamically refines the selection with optimal transport normalization based on similarity scores. Our method minimizes semantic loss while optimizing training efficiency, resulting in superior performance (see Figs. 1b and 2) across zero-shot classification, retrieval tasks, robustness evaluation, and language compositionality benchmarks. CLIP-PGS achieves efficient training without sacrificing the quality of vision-language alignment, demonstrating that a carefully structured masking approach can enhance both efficiency and semantic integrity.

2. Related Work

Vision-Language Pre-training. Recently, vision-language models [4, 26, 28, 48, 51, 67–69] have made significant strides by aligning visual and textual semantics through large-scale pre-training on image-text datasets. Foundational models like CLIP [45] and ALIGN [23] leverage vast internet-sourced image-text pairs, achieving impressive zero-shot classification and retrieval performance across tasks. Despite their success, these models demand substantial computational resources. To enhance efficiency, models like SLIP [39], DeCLIP [30], and MobileCLIP [53] integrate self-supervised learning, reducing dependence on supervised data and boosting robustness. FILIP [60] enhances local similarity matching for finer cross-modal alignment, while more expressive models such as CoCa [64], BLIP [27], and LaCLIP [13] introduce decoders or captioning modules to enhance language generation, though often with added complexity.

Empowering CLIP with Masking. Masked image modeling [1, 2, 14, 19, 34, 43, 58] is a pivotal technique in computer vision that trains models to reconstruct masked image regions, reducing computational load and enhancing data efficiency by guiding models to concentrate on essential information within constrained inputs. Building on MIM's strengths, recent approaches [10, 29, 31, 32, 55, 59] adapt masking strategies to improve CLIP's efficiency while preserving semantic integrity. FLIP [31] introduces random masking of image patches, balancing faster training with accuracy, and establishing an efficient baseline in vision-language pre-training. Recently, more advanced approaches include MaskCLIP [10], which employs masked self-distillation to align local patch features with global semantics, strengthening robustness and transferability across tasks. A-CLIP [59] incorporates attention-based masking to retain tokens aligned with the paired text. Subsequently, E-CLIP [55] uses cluster-based masking to preserve coherent visual structures, optimizing both efficiency and context preservation. However, existing methods face two primary challenges: attention-based masking tends to add computational overhead, while cluster-based masking risks inadvertently masking regions corresponding to textual descriptions within the image. We propose CLIP-PGS, a streamlined approach that enhances pre-training efficiency while selectively preserving critical semantic regions, achieving improved performance with lower computational demands.

3. Method

3.1. Preliminaries

In the CLIP framework [45], visual and textual representations are aligned directly from image-text pairs using a dual-encoder setup. The image encoder $\mathcal{F}_v$ and text encoder $\mathcal{F}_t$ independently process an image $\mathcal{I}$ and text $\mathcal{T}$, projecting them into a shared embedding space with L2 normalization. To optimize alignment, CLIP employs the InfoNCE loss [40], which encourages matched image-text pairs to be similar while pushing mismatched pairs apart. For each batch of pairs, the image encoder's loss is defined as the negative log-likelihood of matched pairs normalized by the total similarity with all text embeddings in the batch, scaled by a temperature parameter. A symmetrical loss is applied to the text encoder, and the final loss $\mathcal{L}_{cl}$ averages both, aligning image and text embeddings effectively.

3.2. CLIP-PGS

We introduce CLIP-PGS, a simple and efficient masking approach designed to meet the computational demands and semantic preservation challenges inherent in VLMs. CLIP-PGS selectively retains semantically meaningful content, enhancing training efficiency without compromising alignment quality between visual and textual representations.

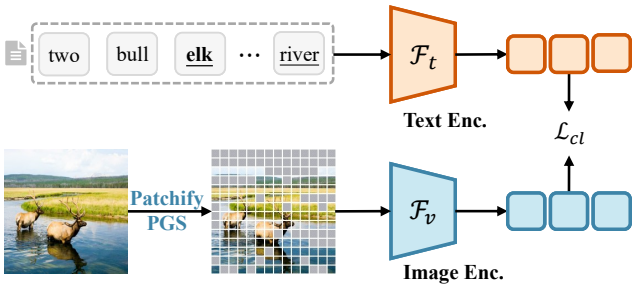


Figure 3. **An illustration of CLIP-PGS**. The text input is processed by the text encoder $\mathcal{F}_t$, while the image undergoes our patch generation-to-selection strategy before entering the image encoder $\mathcal{F}_v$. $\mathcal{L}_{cl}$ subsequently aligns the visual and textual embeddings, strengthening cross-modal representation alignment.

Gradual Process with Dynamic Masking Ratios. Our approach starts by identifying candidate patches for masking. We initialize these patches using the same random masking strategy as FLIP[31], but with a reduced masking ratio of 5% compared to FLIP's 50%. This work introduces two variants of CLIP-PGS: CLIP-PGS$_{0.5}$ and CLIP-PGS$_{0.3}$, where the subscript denotes the lower limit of the masking ratio. Both share an upper limit of 0.5, following FLIP [31]. CLIP-PGS$_{0.5}$ maintains a fixed 0.5 masking ratio as both its lower and upper limits are set to 0.5, while CLIP-PGS$_{0.3}$ dynamically adjusts within $[0.3, 0.5]$.

The dynamic masking process integrates edge detection (ED) and optimal transport normalization (OTN): (*i*) We compute cosine similarities between patches to create a similarity matrix that combines feature-based and image-based affinities. (*ii*) OTN iteratively refines the similarity matrix to satisfy the doubly stochastic constraint, enhancing patch importance allocation. (*iii*) ED preserves critical boundaries, minimizing the masking of semantically significant regions. (*iv*) Patches are ranked by similarity scores to enforce the specified masking bounds, retaining essential patches to meet the lower limit while avoiding excessive masking beyond the upper limit. Next, we will provide a detailed explanation of the proposed method.

Edge Detection. In CLIP-PGS, edge detection (Sobel [24] as used in this work, see §4.3 for ablation details) is applied to the whole image to create an edge map that emphasizes prominent object boundaries and contours. This edge map plays a crucial role in preserving critical semantic information during the masking process. When identifying candidate patches for masking, the edge map is utilized to assign higher importance to patches near strong edges, thereby reducing the likelihood of obscuring key regions such as object outlines and high-contrast details.

Specifically, if a patch is initially marked for removal but exhibits high edge scores, it is retained, while patches with weak edge signals and low candidate heuristics are more likely to be masked. This strategic integration of global edge detection with candidate selection ensures that

Method	R.T.T.	Food101	CIFAR10	CIFAR100	SUN397	Cars	VOC2007	Aircraft	DTD	OxfordPets	Caltech101	Flowers	STL10	EuroSAT	RESISC45	GTSRB	Country211	PCam	Average
CLIP [45]	1.0×	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 [31]	**0.5**×	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 [59]	1.1×	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 [55]	0.6×	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
Ours																			
CLIP-PGS$_{0.5}$	**0.5**×	42.8	62.5	35.5	45.5	17.3	50.0	1.9	17.4	55.7	71.8	**33.2**	**88.2**	20.5	**31.8**	10.1	**4.7**	50.0	37.6
CLIP-PGS$_{0.3}$	0.6×	**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**

Table 1. **Zero-shot classification results**. We evaluate performance on 17 diverse classification datasets, reporting both top-1 accuracy (%) and the overall average. The optimal result is highlighted in **bold**, and the second-best result is underlined. The training time, represented as Relative Training Time (R.T.T.), is benchmarked against CLIP's and set as 1.0×, with other methods shown in relation to this.

semantically meaningful areas are preserved, maintaining the alignment between visual features and their textual descriptions while adhering to the desired masking ratio.

Optimal Transport Normalization. To refine the selection of masked regions and preserve critical semantic information, CLIP-PGS leverages optimal transport normalization (OTN, see Table 6), implemented via the Sinkhorn algorithm [8], to process similarity scores between patches.

The process begins with computing cosine similarity between patches to form a similarity matrix $\boldsymbol{S}$, where each entry $\boldsymbol{S}_{ij}$ denotes the similarity between patch i and patch j. This similarity is calculated using $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^\top$, with $\boldsymbol{X}$ being the normalized embedding matrix ($\boldsymbol{X} = \boldsymbol{X}/\|\boldsymbol{X}\|_2$). The final similarity matrix is computed by integrating both feature and image similarities through a weighted sum: $\boldsymbol{S} = \alpha\boldsymbol{S}_x + (1-\alpha)\boldsymbol{S}_I$, where $\boldsymbol{S}_x$ and $\boldsymbol{S}_I$ are the cosine similarities of features and images, respectively, and α is adjusted based on the training epoch. The Sinkhorn algorithm is then applied to $\boldsymbol{S}$ to iteratively normalize its rows and columns, resulting in a doubly stochastic matrix that ensures a balanced distribution of similarity scores.

The updated similarity matrix $\boldsymbol{S}' = \boldsymbol{S} + \texttt{Sinkhorn}(\boldsymbol{S})$ is crucial for guiding the masking process. OTN uses this balanced similarity matrix to distribute attention across patches, retaining those with high similarity to adjacent regions, thereby preventing the loss of essential visual cues. This balanced selection, facilitated by OTN, not only preserves critical features but also maintains robust alignment between visual and textual modalities during training.

Performance Benefits. CLIP-PGS boosts pre-training efficiency while preserving semantic integrity, yielding high performance in zero-shot classification, retrieval, robustness, and language composition tasks. See §4 for details.

4. Experiments

In this section, we outline the implementation details (§4.1) and present comparison results (§4.2) to verify the proposed method's generalizability on multiple datasets and its robustness to out-of-distribution scenarios. Additionally, we report the training efficiency of existing methods and conduct ablation studies (§4.3) to analyze the proposed components and design choices systematically.

4.1. Implementation

Datasets. Our model is trained on the Conceptual Captions 12M (CC12M) [6] dataset, which comprises approximately 12 million image-text pairs, specifically designed for language-image pre-training. We evaluate all models on 17 diverse classification datasets (*e.g.*, Food101 [3], CIFAR [25] and SUN397 [57]), retrieval datasets (*e.g.*, MS COCO [33] and Flickr [63]), the ImageNet-1K [9] dataset and its out-of-distribution variants, as well as a language compositionality dataset (*i.e.*, SugarCrepe [22]). For additional details on datasets, please refer to the appendix.

Architecture. Following the CLIP [45] and OpenCLIP [7] frameworks, we utilize the ViT-B/16 [11] architecture as the backbone for the image encoder and a 12-layer transformer with 512-dimensional embeddings and 8 attention heads as the text encoder. See the appendix for architecture details.

Pre-training. In our language-image pre-training model, we resize the image to a resolution of 224×224, and tokenize the text into 77 tokens using a 49K token vocabulary, with truncation or padding applied as necessary. The class token is then embedded into a 512 dimensional feature vector through a multi-layer perceptron (MLP). We optimize the model using the AdamW optimizer [36], setting the learning rate to 1e-3, β_1 to 0.9, β_2 to 0.98, and applying a weight decay of 0.2, along with a cosine decay schedule for the learning rate. Our model is trained for 32 epochs with a batch size of 4,096 on 8 NVIDIA TESLA V100 GPUs, using PyTorch's automatic mixed precision library [41].

Downstream Evaluation Tasks. We evaluate our model covering **five** standard benchmark scenarios: zero-shot classification, zero-shot text/image retrieval, linear probing, robustness assessment, and language compositionality, following established evaluation protocols [7, 31, 39, 45].[1]

[1] https://github.com/LAION-AI/CLIP_benchmark

Method	Text Retrieval									Image Retrieval								
	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
CLIP [45]	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
FLIP [31]	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
A-CLIP [59]	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
E-CLIP [55]	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
Ours																		
CLIP-PGS$_{0.5}$	35.2	61.9	72.8	**58.5**	**83.6**	90.6	57.7	82.7	90.4	24.3	48.8	60.5	43.5	70.7	81.0	45.3	72.9	81.2
CLIP-PGS$_{0.3}$	**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**

Table 2. **Zero-shot text/image retrieval results**. We evaluate performance on the MS-COCO [33], Flickr8k [63], and Flickr30k [63] datasets, reporting Recall@1 (%, R@1), Recall@5 (%, R@5), and Recall@10 (%, R@10) for both text and image retrieval tasks.

For the five downstream evaluation tasks outlined above, we adhere strictly to the implementation protocols established in the CLIP baseline [45], ensuring consistency and reliability throughout the evaluation process. This alignment allows for fair comparisons and validates the generalizability and robustness of our methodology.

Baselines. We introduce the four baseline models used for comparison, *i.e.*, CLIP [45], FLIP [31], A-CLIP [59], and E-CLIP [55]. We comprehensively compare different masking strategies in language-image pre-training, reporting baseline results directly from original papers for fairness. All experiments are conducted under consistent settings to ensure reliable conclusions. Performance in various downstream tasks is detailed in the following sections.

4.2. Comparison with SOTA Methods

Zero-Shot Classification. Table 1 presents the zero-shot classification results for our proposed method, CLIP-PGS, with two variants (CLIP-PGS$_{0.5}$ and CLIP-PGS$_{0.3}$) using lower limit masking rates of 0.5 and 0.3. We evaluate CLIP-PGS on 17 diverse classification datasets, comparing it to state-of-the-art models such as CLIP [45], FLIP [31], A-CLIP [59], and E-CLIP [55]. Due to dataset distribution differences, model performance varies. Notably, compact feature spaces (*e.g.*, CIFAR-10) are more sensitive to masking, while greater class diversity (*e.g.*, CIFAR-100) mitigates this effect, as seen in FLIP [31] and E-CLIP [55].

Our CLIP-PGS$_{0.3}$ demonstrates significant improvements on more complex datasets, achieving average accuracy gains of **6.5%** and **3.6%** over FLIP [31] and A-CLIP [59], respectively. In particular, our CLIP-PGS$_{0.5}$, with a 0.5 lower limit masking rate, improves accuracy by 4.6% over FLIP [31] while consuming the same training time. Both variants of CLIP-PGS, with training times reduced to 0.5× and 0.6× of CLIP's duration, either match or exceed baseline accuracy, underscoring the efficiency of our approach. Additionally, CLIP-PGS$_{0.3}$ outperforms E-CLIP [55], showing an average gain of **2.9%**, and achieves the highest top-1 accuracy on **12** out of 17 benchmarking datasets. It performs especially well on challenging benchmarks like Food101 [3], SUN397 [57], and VOC2007 [12], establishing new state-of-the-art results.

Method	CIFAR10	CIFAR100	ImageNet-1K
CLIP [45]	88.0	67.4	62.3
FLIP [31]	85.9	65.5	61.3
A-CLIP [59]	86.4	66.1	62.0
E-CLIP [55]	89.0	69.7	62.7
Ours			
CLIP-PGS$_{0.5}$	89.5 **(+0.5)**	70.3 **(+0.6)**	64.2 **(+1.5)**
CLIP-PGS$_{0.3}$	**90.0 (+1.0)**	**72.3 (+2.6)**	**64.4 (+1.7)**

Table 3. **Linear probing classification results**. We evaluate all models on three common datasets, *i.e.*, CIFAR10 [25], CIFAR100 [25], and ImageNet-1K [9], training each for 10 epochs under a consistent linear training setup. We present top-1 accuracy (%), with gains over the stronger baseline highlighted in **(green)**.

These findings confirm CLIP-PGS's generalizability, efficiency, and adaptability, making it well-suited for real-world applications in various task scenarios.

Zero-Shot Text/Image Retrieval. We evaluate the zero-shot text and image retrieval performance of our proposed method, CLIP-PGS, on MS-COCO [33], Flickr8k [63], and Flickr30k [63], comparing it to state-of-the-art models. Table 2 reports Recall@1, Recall@5, and Recall@10 for both text and image retrieval tasks on these datasets.

The CLIP-PGS$_{0.3}$ variant achieves top-tier performance in most metrics, reaching a Recall@1 of **36.0%** on MS-COCO text retrieval and **25.1%** on MS-COCO image retrieval, outperforming other models in these categories. CLIP-PGS$_{0.5}$ also delivers strong results, securing either the best or second-best scores in multiple retrieval tasks.

In comparison to E-CLIP [55], CLIP-PGS$_{0.3}$ consistently achieves higher retrieval accuracy, particularly on challenging tasks, while CLIP-PGS$_{0.5}$ closely follows and even surpasses E-CLIP [55] on specific metrics. Both variants of CLIP-PGS also show improvements over FLIP [31] and A-CLIP [59], with substantial gains on Flickr8K and Flickr30K. Our method sets new benchmarks over diverse datasets in zero-shot text and image retrieval.

Linear Probing. As shown in Table 3, we assess the linear probing performance on three widely used datasets: CIFAR10 [25], CIFAR100 [25], and ImageNet-1K [9].

Method	ImageNet-1K	ImageNet-V2	ImageNet-A	ImageNet-R	ImageNet-O	ImageNet-Sketch	Average	ID Average	OOD Average
CLIP [45]	36.1	30.7	8.0	47.6	38.4	24.9	31.0	36.1	29.0
FLIP [31]	34.4	29.5	7.1	41.4	39.5	20.1	28.7	34.4	27.5
A-CLIP [59]	35.2	30.1	8.1	45.1	39.4	23.7	30.3	35.2	30.3
E-CLIP [55]	36.3	30.7	8.1	47.9	39.6	25.4	31.3	36.3	30.3
Ours									
CLIP-PGS$_{0.5}$	38.0	32.6	9.1	45.1	41.1	23.9	31.6	38.0	30.4
CLIP-PGS$_{0.3}$	**38.6**	**33.1**	**9.6**	**48.1**	**42.6**	**25.6**	**32.9**	**38.6**	**31.8**

Table 4. **Robustness assessment results**. We evaluate model robustness on ImageNet-1K [9] and five of its variants [20, 21, 46, 54], reporting top-1 accuracy (%) along with overall averages for in-distribution (ID) and out-of-distribution (OOD) performance.

Method	R.T.T.	REPLACE			SWAP		ADD		Average		
		Object	Attribute	Relation	Object	Attribute	Object	Attribute	Object	Attribute	Relation
CLIP [45]	1.0×	85.8	**79.2**	64.5	61.8	58.7	74.2	68.4	73.7	68.8	64.5
FLIP [31]	**0.5**×	84.1	75.9	66.0	60.2	61.6	71.7	63.2	72.0	66.9	66.0
A-CLIP [59]	1.1×	86.6	75.5	63.2	52.4	63.1	71.6	66.8	71.6	68.4	63.2
E-CLIP [55]	0.6×	86.9	73.5	60.2	59.4	63.4	73.3	66.8	73.2	68.4	60.2
Ours											
CLIP-PGS$_{0.5}$	**0.5**×	86.0	77.0	64.6	63.3	65.5	**77.3**	69.8	**75.5**	**70.8**	64.6
CLIP-PGS$_{0.3}$	0.6×	**88.1**	76.0	**67.9**	64.1	**66.5**	74.2	**69.9**	**75.5**	**70.8**	**67.9**

Table 5. **Language compositionality results**. We evaluate the compositionality of vision-language models on the SugarCrepe [22] dataset, which tests models by generating mismatched captions by replacing, swapping, or adding fine-grained atomic concepts (object, attribute, and relation). We report Recall@1 (%) and the overall average for each atomic concept.

Our approach outperforms the baselines, with CLIP-PGS$_{0.3}$ achieving the highest accuracy on all datasets. Notably, CLIP-PGS$_{0.3}$ improves over E-CLIP [55] by **1.0%** on CIFAR10, **2.6%** on CIFAR100, and **1.7%** on ImageNet-1K. CLIP-PGS$_{0.5}$ also achieves competitive results, surpassing E-CLIP [55] by **0.5%** on CIFAR10, **0.6%** on CIFAR100, and **1.5%** on ImageNet-1K. The effectiveness of CLIP-PGS in linear probing highlights its superior capability to capture meaningful representations.

Robustness Assessment. To evaluate robustness, we test CLIP-PGS on ImageNet-1K [9] alongside five of its variants, *i.e.*, ImageNet-V2 [46], ImageNet-A [21], ImageNet-R [20], ImageNet-O [21], and ImageNet-Sketch [54]. We provide the zero-shot top-1 accuracy for each dataset, along with overall in-distribution (ID) and out-of-distribution (OOD) averages, as detailed in Table 4.

CLIP-PGS$_{0.3}$ achieves the highest robustness scores over all datasets, reaching top-1 accuracy of **38.6%** on ImageNet-1K, **9.6%** on ImageNet-A, and **42.6%** on ImageNet-O. Its ID and OOD averages are **32.9%** and **31.8%**, respectively, surpassing all other models. CLIP-PGS$_{0.5}$ shows competitive results, securing the second-best ID and OOD averages at **31.6%** and **30.4%**, respectively.

Compared to E-CLIP [55], which achieves 47.9% on ImageNet-R and 25.4% on ImageNet-Sketch, CLIP-PGS$_{0.3}$ provides improved consistency covering all datasets. This consistency reflects the robustness and adaptability of CLIP-PGS in handling diverse distributional shifts.

Language Compositionality. We evaluate the language compositionality of CLIP-PGS on the SugarCrepe [22] dataset. Table 5 reports Recall@1 for each atomic concept (object, attribute, and relation) in three tasks: replacing, swapping, and adding concepts in captions. Additionally, we provide an overall average for each concept type.

Our method exhibits notable improvements in language compositionality. CLIP-PGS$_{0.3}$ achieving leading scores on most categories. Specifically, CLIP-PGS$_{0.3}$ surpasses A-CLIP [59] and FLIP [31], with **88.1%** in object replacement and **67.9%** in relation handling, marking an average gain over competing models. In addition to outperforming E-CLIP [55], CLIP-PGS$_{0.3}$ achieves higher accuracy in attribute handling compared to other baselines, indicating its robust performance in diverse concept manipulations.

CLIP-PGS$_{0.5}$ likewise shows notable results, leading in object addition with **77.3%** and securing the highest overall averages for object and attribute tasks. Compared to FLIP [31], CLIP-PGS$_{0.5}$ offers an improvement in handling fine-grained attributes and relations. Evaluation results emphasize the effectiveness of CLIP-PGS, especially in tasks requiring nuanced compositionality in vision-language scenarios. Its capabilities make it well-suited for managing complex concept manipulations in real-world applications.

Qualitative results. Fig. 4 illustrates the effectiveness of CLIP-PGS in selectively masking image regions while retaining critical semantic content. By carefully masking non-essential areas, CLIP-PGS maintains the integrity of key elements, allowing for accurate descriptions without compromising important contextual details. Fig. 5 presents the progression of zero-shot top-1 accuracy on ImageNet-1K [9] across training epochs for our models, CLIP-PGS$_{0.5}$ and CLIP-PGS$_{0.3}$, trained on CC12M [6].

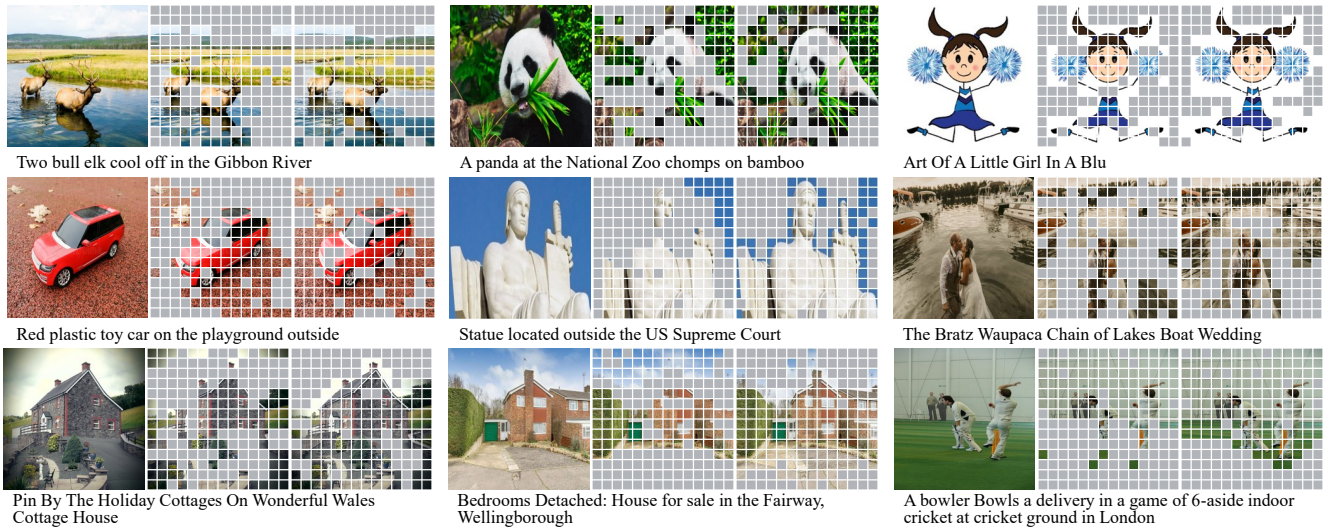


Figure 4. **Visualization of masking regions**. We use ViT-B/16 [11] as the image encoder, displaying each sample with the text description, the original image (**left**), and masking results from CLIP-PGS$_{0.5}$ (**middle**) at a fixed 0.5 masking ratio, and CLIP-PGS$_{0.3}$ (**right**) with a variable masking ratio between 0.3 and 0.5. Our models effectively retain the visual content relevant to the accompanying text context.

Method	Component (extra cost)			ImageNet-1K		MS-COCO	
	MR (<1.0%)	ED (~1%)	OTN (~1%)	ZS	LP	TR	IR
Baseline [45]	-	-	-	36.1	62.3	34.6	23.5
Random Mask [31]	0.5	-	-	34.4	61.3	32.6	22.6
CLIP-PGS$_{0.5}$	0.5	✗	✗	35.2	61.9	33.7	22.8
	0.5	✓	✗	36.2	62.8	34.1	23.4
	0.5	✗	✓	36.3	62.7	33.9	23.2
	0.5	✓	✓	38.0	64.2	35.2	24.3
CLIP-PGS$_{0.3}$	[0.3, 0.5]	✗	✗	35.9	61.7	33.5	23.0
	[0.3, 0.5]	✓	✗	36.8	63.2	34.3	24.0
	[0.3, 0.5]	✗	✓	36.7	63.0	34.5	23.8
	[0.3, 0.5]	✓	✓	38.6	64.4	36.0	25.1

Table 6. **Ablation analysis of key components**. We present the ablation studies of CLIP-PGS's components, covering zero-shot image classification, linear probing, and text/image retrieval tasks.

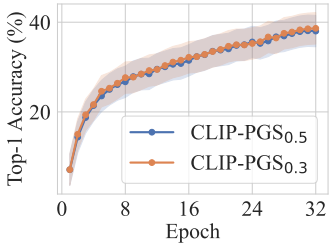


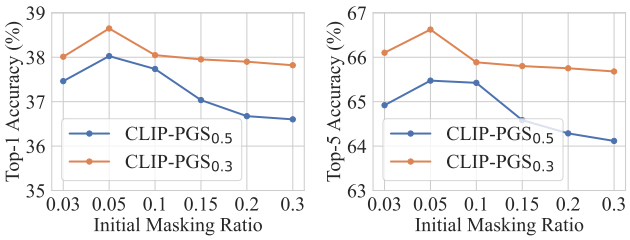
Figure 5. **Zero-shot classification on ImageNet-1K** [9]. We present plots showing the trend of zero-shot accuracy across training epochs for the models trained on CC12M [6] over 32 epochs.

4.3. Ablation Studies

This section provides a detailed ablation analysis of the core design components of CLIP-PGS. By default, we use ViT-B/16 as the image encoder, trained on CC12M [6] for 32 epochs with a batch size of 4,096. Table 6 summarizes results on various downstream tasks, including zero-shot classification (ZS), linear probing (LP), and zero-shot text/image retrieval (TR/IR) on ImageNet-1K [9] and MS-COCO [33]. Please refer to the appendix for more results.

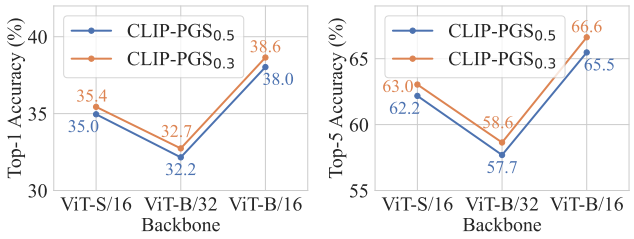
Initial Masking Ratio. We conduct ablations to assess the impact of varying initial masking ratios on CLIP-PGS, as shown in Fig. 6. For top-1 accuracy, CLIP-PGS$_{0.3}$ reaches peak performance at lower initial masking ratios, particularly around 0.05. As the masking ratio increases, accuracy declines for both variants, though CLIP-PGS$_{0.3}$ maintains more stable results than CLIP-PGS$_{0.5}$. A similar trend appears at top-5, as higher masking ratios obscure key semantic information. A lower initial masking ratio provides an optimal balance, preserving essential visual details while minimizing the risk of losing important semantic content.

Lower/Upper Limit Masking Ratio (MR). Table 6 analyzes the influence of lower and upper masking limits on CLIP-PGS performance. Consistent with [31], we set the upper masking ratio to 0.5, with subscripts in CLIP-PGS$_{0.5}$ and CLIP-PGS$_{0.3}$ indicating lower limits of 0.5 and 0.3, respectively. Thus, CLIP-PGS$_{0.5}$ employs a fixed masking ratio, while CLIP-PGS$_{0.3}$ dynamically adjusts between 0.3 and 0.5. In contrast to FLIP's random masking, our approach utilizes progressive masking, starting with a small selection of mask patches and gradually expanding based on similarity scores. This method leads to improvements in zero-shot accuracy for both CLIP-PGS$_{0.5}$ and CLIP-PGS$_{0.3}$, achieving gains of **0.8%** and **1.5%**, respectively, and yielding comparable advancements in linear probing and zero-shot retrieval tasks. In particular, CLIP-PGS$_{0.3}$ consistently outperforms, with zero-shot classification accuracy increasing from 35.2% (for CLIP-PGS$_{0.5}$) to 35.9%



(a) Top-1 *vs.* Initial Masking Ratio (b) Top-5 *vs.* Initial Masking Ratio

Figure 6. **Ablation analysis of initial masking ratio**. We report the zero-shot accuracy results of CLIP-PGS on ImageNet-1K [9] at various initial masking ratios.



(a) Top-1 *vs.* Backbone (b) Top-5 *vs.* Backbone

Figure 7. **Ablation analysis of different backbone and patch sizes**. We report the zero-shot accuracy results of CLIP-PGS on ImageNet-1K [9] at various backbone sizes.

without additional components, and reaching **38.6%** with the addition of ED and OTN. A similar improvement appears in linear probing, where CLIP-PGS$_{0.3}$ attains **64.4%**, surpassing the fixed 0.5 masking setup of CLIP-PGS$_{0.5}$. Expanding the masking ratio range adds flexibility in preserving critical semantic information, thereby reinforcing CLIP-PGS's robustness and overall effectiveness.

Edge Detection (ED). Table 6 presents the ablation results to analyze the effect of ED on CLIP-PGS. Adding ED yields consistent performance gains for both CLIP-PGS$_{0.5}$ and CLIP-PGS$_{0.3}$. For instance, CLIP-PGS$_{0.3}$ increases zero-shot accuracy from 35.9% to **36.8%** and linear probing accuracy from 61.7% to **63.2%**. Text and image retrieval also benefit, underscoring ED's role in preserving key semantics and enhancing feature alignment within masked images. Edge detection is thus essential for improving model performance by emphasizing critical visual regions.

Optimal Transport Normalization (OTN). Table 6 illustrates the impact of incorporating OTN on improving CLIP-PGS performance across all metrics. By establishing a balanced similarity matrix that prioritizes patches with higher similarity to adjacent regions, OTN preserves key semantic details, enhancing feature alignment in masked areas. For example, CLIP-PGS$_{0.3}$ with OTN raises zero-shot accuracy from 35.9% to **36.7%** and linear probing from 61.7% to **63.0%**, with similar gains in text (**34.5%**) and image retrieval (**23.8%**). Combined with ED, OTN yields the highest scores in zero-shot classification and retrieval tasks.

Computational Efficiency. As shown in Table 6, the additional computational cost of CLIP-PGS is minimal. Its key components contribute only a small overhead: MR (random masking and similarity computation) adds less than **1.0%**, while ED (Sobel) and OTN (Sinkhorn) each contribute approximately **1%**, resulting in a total overhead of less than **3%**. Despite these additions, CLIP-PGS maintains training acceleration comparable to FLIP [31], ensuring high efficiency without compromising performance.

Backbone and Patch Sizes. Fig. 7 provides the ablation results of different backbone and patch sizes. We assess three ViT [11] configurations such as ViT-S/16, ViT-B/32 and ViT-B/16. The ViT-B/16 backbone delivers the highest accuracy, demonstrating superior capability in capturing fine-grained features compared to the smaller ViT-S/16 and lower-resolution ViT-B/32. Our model, CLIP-PGS$_{0.3}$ with ViT-B/16, achieves **38.6%** top-1 and **66.6%** top-5 accuracy. The ablation results reinforce the widely accepted view that larger backbones with higher feature resolutions enhance the performance of vision-language pre-training models.

5. Conclusion

In this study, we introduce CLIP-PGS, a simple yet efficient masking framework for enhancing CLIP's efficiency in vision-language alignment through Patch Generation-to-Selection. Our gradual approach begins by pre-selecting candidate patches for masking, followed by Sobel edge detection to create an edge mask that preserves key object regions. We further refine patch selection by computing similarity scores between candidate and neighboring patches, using optimal transport normalization to maintain balanced patch representation. This structured approach ensures effective retention of semantic information while reducing computational demands. CLIP-PGS achieves state-of-the-art performance in zero-shot classification and retrieval tasks. It also demonstrates significant improvements in robustness and language compositionality, showcasing its versatility across diverse downstream applications.

Limitation and Future Work. Due to limited computational resources, our models are pre-trained only on the CC12M dataset, with primary experiments conducted on ViT-B/16. In future work, we aim to extend CLIP-PGS to convolutional network architectures [18, 35], broadening its applicability beyond the current transformer-based models. Additionally, while CLIP-PGS is designed for dual-encoder vision-language models like CLIP [45], our masking strategy is not inherently limited to this framework. We plan to explore its adaptation to other self-supervised learning approaches, such as masked image modeling techniques like MAE [19], to potentially enhance their efficiency.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (No. 62472222, 62222207, 62427808), Natural Science Foundation of Jiangsu Province (No. BK20240080).

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, pages 456–473, 2022. 3

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022. 2, 3

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014. 4, 5

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, pages 1877–1901, 2020. 1, 2

[5] Xinhao Cai, Qiuxia Lai, Yuwei Wang, Wenguan Wang, Zeren Sun, and Yazhou Yao. Poly kernel inception network for remote sensing detection. In *CVPR*, pages 27706–27716, 2024. 1

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568, 2021. 1, 4, 6, 7

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023. 1, 4

[8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 26, 2013. 4

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 4, 5, 6, 7, 8

[10] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *CVPR*, pages 10995–11005, 2023. 2, 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4, 7, 8

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2010. 5

[13] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *NeurIPS*, 36:35544–35575, 2024. 2

[14] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, pages 19358–19369, 2023. 3

[15] Enrico Fini, Pietro Astolfi, Adriana Romero-Soriano, Jakob Verbeek, and Michal Drozdzal. Improved baselines for vision-language pre-training. *Transactions on Machine Learning Research*, 2023. 2

[16] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei W Koh, Olga Saukh, Alexander J Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In *NeurIPS*, pages 27092–27112, 2023. 1

[17] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 8

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2, 3, 8

[20] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021. 6

[21] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 6

[22] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *NeurIPS*, 36:31096–31116, 2023. 4, 6

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 1, 2

[24] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of Solid-State Circuits*, 23(2): 358–367, 1988. 3

[25] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report*, 2009. 4, 5

[26] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learn-

ing with momentum distillation. In *NeurIPS*, pages 9694–9705, 2021. 1, 2

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 2

[28] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1, 2

[29] Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for clip training. In *NeurIPS*, pages 49068–49087, 2023. 2, 3

[30] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022. 1, 2

[31] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, pages 23390–23400, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[32] Mingliang Liang and Martha Larson. Centered masking for language-image pre-training. In *ECML PKDD*, pages 90–106, 2024. 3

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 4, 5, 7

[34] Yuan Liu, Songyang Zhang, Jiacheng Chen, Zhaohui Yu, Kai Chen, and Dahua Lin. Improving pixel-based mim by reducing wasted modeling capability. In *ICCV*, pages 5361–5372, 2023. 3

[35] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 8

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 4

[37] Jiawei Ma, Po-Yao Huang, Saining Xie, Shang-Wen Li, Luke Zettlemoyer, Shih-Fu Chang, Wen-Tau Yih, and Hu Xu. Mode: Clip data experts via clustering. In *CVPR*, pages 26354–26363, 2024. 1

[38] Amin Karimi Monsefi, Kishore Prakash Sailaja, Ali Alilooee, Ser-Nam Lim, and Rajiv Ramnath. Detailclip: Detail-oriented clip for fine-grained tasks. *arXiv preprint arXiv:2409.06809*, 2024. 2

[39] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *ECCV*, pages 529–544, 2022. 2, 4

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32:8024–8035, 2019. 4

[42] Gensheng Pei, Fumin Shen, Yazhou Yao, Guo-Sen Xie, Zhenmin Tang, and Jinhui Tang. Hierarchical feature alignment network for unsupervised video object segmentation. In *ECCV*, pages 596–613, 2022. 1

[43] Gensheng Pei, Tao Chen, Xiruo Jiang, Huafeng Liu, Zeren Sun, and Yazhou Yao. Videomac: Video masked autoencoders meet convnets. In *CVPR*, pages 22733–22743, 2024. 3

[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[46] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400, 2019. 6

[47] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 1

[48] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 1, 2

[49] 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. 1

[50] 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 *CVPR*, pages 13019–13029, 2024. 1

[51] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *ICML*, pages 5100–5111, 2019. 1, 2

[52] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 1

[53] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. In *CVPR*, pages 15963–15974, 2024. 1, 2

[54] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 32, 2019. 6

[55] Zihao Wei, Zixuan Pan, and Andrew Owens. Efficient vision-language pre-training by cluster masking. In *CVPR*, pages 26815–26825, 2024. 1, 2, 3, 4, 5, 6

[56] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *ICLR*, 2024. 1

[57] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene

recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010. 4, 5

[58] 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 *CVPR*, pages 9653–9663, 2022. 3

[59] Yifan Yang, Weiquan Huang, Yixuan Wei, Houwen Peng, Xinyang Jiang, Huiqiang Jiang, Fangyun Wei, Yin Wang, Han Hu, Lili Qiu, et al. Attentive mask clip. In *ICCV*, pages 2771–2781, 2023. 2, 3, 4, 5, 6

[60] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2022. 2

[61] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. In *CVPR*, pages 2623–2632, 2021. 1

[62] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. In *CVPR*, pages 5192–5201, 2021. 1

[63] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 4, 5

[64] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[65] Lu Yu, Haiyang Zhang, and Changsheng Xu. Text-guided attention is all you need for zero-shot robustness in vision-language models. In *NeurIPS*, 2024. 1

[66] Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. In *NeurIPS*, 2024. 1

[67] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE TPAMI*, 46(8):5625–5644, 2024. 2

[68] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1

[69] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2