SuperCLIP: CLIP with Simple Classification Supervision

Weiheng Zhao[1] **Zilong Huang**[2]* **Jiashi Feng**[2] **Xinggang Wang**[1]
[1]*School of EIC, Huazhong University of Science and Technology*
[2]*ByteDance*
Code & Models: hustvl/SuperCLIP

Abstract

Contrastive Language-Image Pretraining (CLIP) achieves strong generalization in vision-language tasks by aligning images and texts in a shared embedding space. However, recent findings show that CLIP-like models still underutilize fine-grained semantic signals in text, and this issue becomes even more pronounced when dealing with long and detailed captions. This stems from CLIP's training objective, which optimizes only global image-text similarity and overlooks token-level supervision—limiting its ability to achieve fine-grained visual-text alignment. To address this, we propose SuperCLIP, a simple yet effective framework that augments contrastive learning with classification-based supervision. By adding only a lightweight linear layer to the vision encoder, SuperCLIP leverages token-level cues to enhance visual-textual alignment — with just a 0.077% increase in total FLOPs, and no need for additional annotated data. Experiments show that SuperCLIP consistently improves zero-shot classification, image-text retrieval, and purely visual tasks. These gains hold regardless of whether the model is trained on original web data or rich re-captioned data, demonstrating SuperCLIP's ability to recover textual supervision in both cases. Furthermore, SuperCLIP alleviates CLIP's small-batch performance drop through classification-based supervision that avoids reliance on large batch sizes.

1 Introduction

CLIP [45] has become a cornerstone in vision-language learning, excelling in tasks like zero-shot classification, retrieval, and text-to-image generation [72, 29, 26, 46]. By aligning images and text in a shared embedding space and leveraging large-scale noisy web data [11, 6, 47], it learns rich, transferable representations. However, despite its strong performance, CLIP's representations still have room for improvement, and further enhancing them remains crucial for advancing multimodal applications [51, 65, 52, 57].

Recent works have proposed various improvements to CLIP in three main dimensions: training strategies [33, 67, 42, 32, 22, 64, 12, 60], architectural modifications [5, 56, 63, 70, 16, 58, 50], and data collection techniques [23, 17, 27, 38, 1, 11, 59]. These approaches have significantly enhanced the performance of the CLIP model in zero-shot and other downstream tasks [55].

However, despite these advances, an interesting phenomenon emerges: Contrastive models like CLIP still struggle to fully exploit the rich supervision in captions, especially when those captions are long and detailed re-captioned texts [31, 10, 27]. This counterintuitive phenomenon highlights a fundamental issue: contrastive learning fails to make full use of fine-grained semantic signals in text, even when they are explicitly available.

*Corresponding author (zilong.huang2020@gmail.com).

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

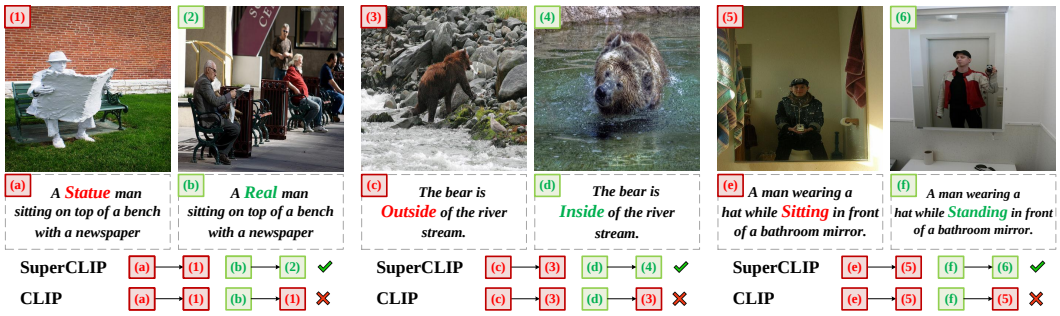

Figure 1: **Evaluating Fine-Grained Alignment in Image-Text Retrieval.** Each row presents pairs of images and captions that are visually and semantically very similar, but differ in fine-grained semantic distinctions such as object status (e.g. **Statue** vs. **Real**), spatial relation (e.g. **Outside** vs. **Inside**), and action (e.g. **Sitting** vs. **Standing**). While both images and texts are close in meaning, SuperCLIP demonstrates a stronger ability than CLIP in correctly distinguishing these fine-grained semantic distinctions. Additional examples are provided in **Appendix A.1.**

This challenge stems from how CLIP is trained by optimizing only for global image-text similarity, while overlooking the dense semantic cues encoded in individual words or phrases [70, 68, 61, 36, 24, 62]. The problem is further compounded by the characteristics of typical web data, which tend to be short, ambiguous, and only loosely aligned with the visual content [11]. As a result, CLIP-like models often miss subtle but important distinctions in object attributes, spatial relationships, and actions. As illustrated in Fig.1, CLIP may confuse a statue with a real person or fail to distinguish whether a bear is inside or outside the river. This lack of fine-grained alignment limits their effectiveness in downstream multimodal tasks that require precise visual-textual understanding [54]. Existing works have attempted to address this issue, but they either rely on additional annotated datasets beyond the web-scale data typically used for CLIP training, or introduce substantial computational overhead [31, 70, 68, 30, 64, 24, 62, 54, 55].

Thus, in this work, we propose SuperCLIP, a super simple yet effective approach that introduces a classification-based supervision method [20] into the contrastive learning paradigm of image–text pretraining. With only a lightweight linear layer added to the vision encoder, SuperCLIP directly leverages raw text tokens to guide the vision encoder to attend to semantic entities mentioned in the text and their visual manifestations in the image. In this way, SuperCLIP fully leverages the rich textual supervision from all words in the text, thereby enhancing the model's ability to achieve fine-grained visual-text alignment — with just a 0.077% increase in total FLOPs, and without requiring additional annotated data.

Extensive experiments demonstrate that our method effectively helps CLIP models recover rich textual supervision from all words in the text—whether trained on original web data or rich re-captioned data—leading to consistent improvements in zero-shot performance on classification and retrieval tasks, while also enhancing the vision encoder's features for purely visual tasks. Furthermore, SuperCLIP is simple and easy to implement, making it readily applicable to other CLIP-style training frameworks such as SigLIP [67] and FLIP [33], where it also brings consistent performance gains. Finally, thanks to its classification-based supervision and independence from large batch sizes, SuperCLIP alleviates the performance degradation typically observed in CLIP under small-batch training settings.

Our main contributions can be summarized as follows:

1. We propose SuperCLIP, a simple yet effective vision–language pretraining framework that seamlessly integrates classification-based supervision into contrastive learning, enabling CLIP models to effectively recover rich textual supervision from all words in the text.

2. Without introducing heavy computational cost or requiring additional annotated data, SuperCLIP enhances CLIP's ability to achieve fine-grained visual-text alignment and mitigates its performance degradation under small batch sizes.

3. Empirical results demonstrate that SuperCLIP achieves improved performance on zero-shot classification and retrieval tasks, as well as on purely visual downstream tasks, thereby confirming its broad effectiveness.

2 Related Work

In this section, we first provide a brief overview of representative efforts to improve CLIP. Then, we discuss existing approaches that specifically aim to address the underlying limitations of CLIP highlighted in the introduction.

2.1 Contrastive Vision-Language Pretraining

Contrastive learning has become the dominant approach for vision-language pretraining, with CLIP [45] demonstrating strong zero-shot transfer by aligning images and texts in a shared embedding space using large-scale noisy web data. Subsequent efforts have improved CLIP along three major dimensions. **Training-centric** [33, 67, 42, 32, 22, 64, 12, 60] strategies improve learning efficiency and robustness by modifying optimization objectives and training dynamics, such as using a sigmoid-based contrastive loss in SigLIP [67], applying masked image modeling to accelerate training in FLIP [33], and leveraging nearest-neighbor supervision to enhance data efficiency in DeCLIP [32]. **Model-centric** [5, 56, 63, 70, 16, 58, 50] improvements include designing stronger vision encoders such as Vitamin [5], rethinking input representations as in CLIPPO [56], and introducing more robust attention mechanisms like DiffCLIP [16]. **Data-centric** [23, 17, 27, 38, 1, 11, 59]. approaches focus on scaling dataset size and diversity to enhance model generalization, exemplified by ALIGN [23], LAION-5B [47], and DataComp [11]. In summary, these methods have effectively boosted the CLIP model's performance by refining the data, model architecture, and training techniques.

2.2 Improve CLIP with Additional Supervision

A number of recent works have considered the underlying problem that CLIP struggles with fine-grained visual-text alignment due to its reliance on global image-text similarity and weak, ambiguous supervision from web-sourced captions [31, 70, 68, 30, 64, 24, 62, 54, 55]. To address this issue, these works introduce additional forms of supervision to enhance fine-grained visual-text alignment. Recap-DataComp-1B [31] recaptions the original web data using LLaMA-3 to produce more informative captions for improving CLIP, but their findings show that CLIP is not fully effective at leveraging such rich textual supervision. While RegionCLIP [70] introduces region-level supervision without manual labels, it inherits CLIP's semantic limitations, overlooks inter-region relationships, and incurs additional computation due to region proposal processing. Long-CLIP [68] extends CLIP to long-text understanding via positional embedding stretching and component matching, but it compromises zero-shot image classification performance by disrupting the alignment between short-text prompts and visual features. UniCL [64] enhances CLIP by unifying contrastive learning across image-text and image-label pairs, but it relies on additional human-annotated category labels, which limits its scalability compared to purely web-supervised approaches. Eyes Wide Shut [54] and SigLIP 2 [55] both improve visual grounding and understanding through dense feature integration, but their methods introduce substantial computational and data overhead. In summary, these methods either fall short of fully resolving the issue, rely on additional annotated datasets beyond the web-scale data typically used for CLIP training, or introduce substantial computational overhead.

3 Motivation and Method

In this section, we first revisit the inherent limitations of the CLIP contrastive learning paradigm to motivate our approach. Then, we present a super simple classification-based method to recover rich textual supervision and improve CLIP's fine-grained visual-text alignment. The overall framework of our proposed SuperCLIP is illustrated in Fig.2.

3.1 Limitations of the Contrastive Learning Paradigm

Overall Review of CLIP Training CLIP [45] learns joint image-text embeddings using a large collection of paired examples $\{(I_k, T_k)\}_{k=1}^{M}$. The model consists of two encoders—f_θ for images

Keyword Group	**Man + Newspaper**		**Bear + River**		**Man + Mirror**	
Condition	Basic Pair	+ Real/Stat	Basic Pair	+ In/Out	Basic Pair	+ Sit/Stand
Matching Captions	333	6	219	2	1216	19
Percentage (%)	0.00333	0.0006	0.00219	0.00002	0.01216	0.00019

Table 1: **Keyword Co-occurrence Statistics in Datacomp-1B [11] (10M captions).** "In/Out" = Inside/Outside; "Real/Stat" = Real/Statue; "Sit/Stand" = Sitting/Standing. Each column shows how many captions match specific keyword combinations. Percentages refer to frequency in 10M captions. More keyword combination results are provided in **Appendix A.2.**

and g_ϕ for text—and normalizes their outputs to unit length. Specifically, for each image I_i and text T_i, their embeddings are computed as:

$$u_i = \frac{f_\theta(I_i)}{\|f_\theta(I_i)\|_2}, v_i = \frac{g_\phi(T_i)}{\|g_\phi(T_i)\|_2}. \tag{1}$$

For a batch of N pairs, CLIP computes the similarity matrix:

$$S_{ij} = \frac{u_i^\top v_j}{\tau}, \tag{2}$$

where τ is the temperature parameter. The objective function $\mathcal{L}_{\text{CLIP}}$ is designed to maximize the similarity between matching image-text pairs while minimizing the similarity between non-matching pairs in the shared embedding space. It is defined as:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{\exp(S_{ii})}{\sum_{k=1}^{N}\exp(S_{ik})} + \log\frac{\exp(S_{ii})}{\sum_{k=1}^{N}\exp(S_{ki})}\right), \tag{3}$$

where $\log\frac{\exp(S_{ii})}{\sum_{k=1}^{N}\exp(S_{ik})}$ corresponds to the image-to-text part, and $\log\frac{\exp(S_{ii})}{\sum_{k=1}^{N}\exp(S_{ki})}$ corresponds to the text-to-image part. By aligning matching image-text pairs and separating non-matching ones, CLIP learns robust visual and textual features, which are applicable to various downstream tasks.

Impact of Batch Size and Web Data Sparsity A key assumption of CLIP is that each batch must contain enough positive and negative pairs for effective learning [45]. When batch size is small, performance degrades rapidly [67], which is why CLIP training typically relies on very large batches—often 16k or more—demanding significant computational resources [33, 49]. Large batches help CLIP learn diverse object categories from web data [23, 47, 11], contributing to its strong zero-shot classification performance. Despite CLIP's strong performance in object recognition, it struggles with fine-grained attributes like actions, spatial relations, and object states. This is largely due to the nature of web data, where captions are often short, ambiguous, and poorly structured [11]. As a result, semantic combinations needed to learn fine-grained distinctions are rare and inconsistent. For example, "man + newspaper" appears 333 times in 10M captions, but "man + newspaper + real/statue" appears only 6 times, and some combinations—like "bear + river + in/out"—are nearly absent (see Table 1). These low-frequency cases rarely form effective contrastive pairs [2, 3, 54], and even when they do exist, they are unlikely to co-occur in the same batch—making contrastive learning of such concepts nearly impossible without extremely large, computationally expensive batch sizes.

Limitation in Using Rich Textual Supervision CLIP is trained with a contrastive objective that aligns image and text embeddings by pulling matched pairs closer and pushing mismatched pairs apart within each batch. While effective at capturing global semantic alignment, this objective tends to overlook fine-grained or detailed semantic information present in the captions [70, 68, 61, 36, 24, 62], limiting the model's ability to fully leverage rich textual supervision. An interesting exploration in [31] re-captioned web data using powerful MLLMs like LLaMA-3 [14], enriching the captions with more semantic information. In theory, this should provide stronger supervision and improve performance. However, contrary to expectations, CLIP models trained under the contrastive learning paradigm actually showed a drop in performance when the original data was entirely replaced with re-captioned data. Our experiments in section 4.3 further confirm this phenomenon, showing that simply enriching captions does not necessarily lead to better performance under the contrastive learning paradigm. This suggests that CLIP's contrastive learning paradigm struggles to take advantage of rich textual supervision—in fact, the added complexity introduced by richer captions can hinder learning and lead to a drop in performance.

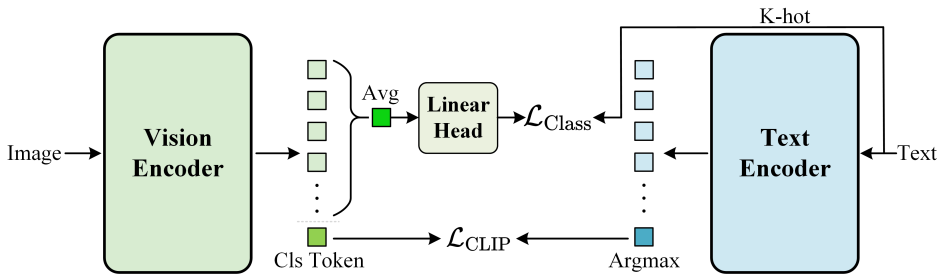

Figure 2: **Overall Framework of Our Proposed SuperCLIP.** Introducing simple classification-based supervision into the CLIP framework is straightforward. It only involves adding a lightweight linear layer to the image encoder to map the averaged image features to text classification targets, without requiring any changes to the original contrastive learning paradigm.

3.2 Super Simple Classification-based Supervision

Using text as supervision for visual backbones is well studied [41, 53, 25, 28]. However, existing methods often rely on manual filtering or heuristics to construct classification vocabularies [19, 40], which limits scalability to noisy web data. To overcome this, we follow [20] and use raw text tokens—prior to CLIP's text encoder—as direct classification labels for the vision encoder. Specifically, consider an image-text dataset $\mathcal{D} = \{(I_i, T_i) \mid i \in [1, N]\}$ used for contrastive training following the CLIP framework. Each caption T_i is tokenized using CLIP's subword-level tokenizer with a vocabulary size of V, resulting in a set of token IDs $\mathcal{C}$. These tokens are then represented as a V-dimensional K-hot vector $\mathbf{y} \in \mathbb{R}^V$, where $y_c = 1$ if $c \in \mathcal{C}$, and $y_c = 0$ otherwise. While the original label $\mathbf{y}$ treats all subwords equally, some frequent stopwords or generic terms carry less discriminative information. To address this imbalance, an Inverse Document Frequency (IDF) based weighting is applied:

$$w_c = \log\left(\frac{|\mathcal{D}|}{1 + \mathrm{df}(c)}\right), \tag{4}$$

where $|\mathcal{D}|$ denotes the total number of image-text pairs in the dataset, and $\mathrm{df}(c)$ is the document frequency of subword c, i.e., the number of captions in which c appears. Using these weights, a normalized weighted label distribution $\hat{y}_c$ is computed as:

$$\hat{y}_c = \frac{w_c y_c}{\sum_{c'=1}^{V} w_{c'} y_{c'}}. \tag{5}$$

Given the normalized label distribution $\hat{y}_c$, the goal is to train the model such that its output distribution aligns closely with this weighted supervision signal. Let x_c denote the logit output of the model for class $c \in \{1, \ldots, V\}$, obtained by applying a linear classification head on top of the image features extracted by the CLIP vision encoder. The final classification loss is defined as the cross-entropy between the weighted label distribution $\hat{y}_c$ and the softmax-normalized model predictions:

$$\mathcal{L}_{\text{Class}} = -\sum_{c=1}^{V} \hat{y}_c \log\left(\frac{e^{x_c}}{\sum_{c'=1}^{V} e^{x_{c'}}}\right). \tag{6}$$

This classification loss encourages alignment between the model predictions and all subword tokens extracted from the text, ensuring that the full textual supervision signal is utilized. Finally, since both the training data and the vision encoder are taken directly from the existing CLIP training pipline (except for a simple linear layer that maps image features to classification targets), this loss can be easily added to the CLIP optimization objective:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{Class}}. \tag{7}$$

In this way, our method extends CLIP to effectively recover rich textual supervision from all words in the text, naturally guiding the model to attend to fine-grained visual-text alignment that is often overlooked by standard CLIP. What's more, since the classification loss does not rely on batch size, it can alleviates the performance degradation typically observed in CLIP under small-batch training settings.

Model	Pretrain	Image Classification		Image Retrieval		Text Retrieval	
		val	v2	COCO	Flickr	COCO	Flickr
CLIP	B-512M	60.5	53.0	29.0	54.5	46.7	73.3
SuperCLIP	B-512M	63.5 (+3.0)	55.2 (+2.2)	31.3 (+2.3)	56.9 (+2.4)	47.8 (+1.1)	75.6 (+2.3)
CLIP	L-512M	66.1	57.4	32.7	57.0	49.6	76.4
SuperCLIP	L-512M	70.1 (+4.0)	62.5 (+5.1)	35.9 (+3.2)	62.4 (+5.4)	52.2 (+2.6)	79.3 (+2.9)
CLIP	L-12.8B	79.0	72.0	43.9	72.7	62.5	87.0
SuperCLIP	L-12.8B	80.0 (+1.0)	72.8 (+0.8)	45.5 (+1.6)	74.2 (+1.5)	63.1 (+0.6)	88.1 (+1.1)

Table 2: **Comparison with CLIP across Different Model Sizes.** We report **zero-shot** image classification accuracy (%) on ImageNet-1K (val and v2), and **zero-shot** image and text retrieval (Recall@1, %) on COCO and Flickr30K, comparing CLIP and our SuperCLIP under three settings: B-512M, L-512M, and L-12.8B, where models are pretrained on 512M or 12.8B samples from DataComp-1B. Values in parentheses reflect absolute gains or drops for SuperCLIP relative to CLIP.

4 Empirical Results

4.1 Experimental Setup

Pretraining Setup. We pretrain our proposed SuperCLIP and CLIP on a standard subset of the Datacomp dataset [11], which contains about 1.3B image-text pairs. All images are resized to a fixed resolution of 224×224, and the text is minimally processed with only basic tokenization. Note that, except for the experiment (**Comparison with CLIP with Mixed Caption**) in Section 4.3 which consider useing the Recap-DataComp[31] data, all other experiments are conducted on the original Datacomp dataset. All experiments are conducted with a batch size of 16k, except for those under varying batch sizes analyzing the impact on CLIP. For fair comparison, all models adopt AdamW with a cosine schedule, using the same learning rate and weight decay as CLIP.

Evaluation Protocol. For zero-shot evaluation, we use the open-source LAION CLIP Benchmark framework [47] to assess all models on zero-shot classification and image-text retrieval. For linear probing image classification experiments, we follow the training protocol introduced in MAE [18]. For semantic segmentation and depth estimation, we follow a protocol similar to DINOv2 [44].

4.2 Main Results

Comparison across Different Model Sizes. We demonstrate that our method consistently benefits CLIP across different model sizes, through **zero-shot** image classification on ImageNet-1K [8] (val and v2) and image-text retrieval on COCO [35] and Flickr30K [66]. Detailed results are presented in Table 2. By training both B- and L-sized models with varying amounts of pretraining data, we compare CLIP and our proposed SuperCLIP under three settings: B-512M, L-512M, and L-12.8B (ViT-B/L pretrianed with seen 512M/12.8B samples). Under the B-512M setting, SuperCLIP improves classification and retrieval performance across all tasks, including gains of over +3% in classification accuracy and up to +2.4% in retrieval. With the L-512M setting, the improvements are more substantial, reaching up to +5.4% in image retrieval and over +5.1% in classification. At the largest scale (L-12.8B), SuperCLIP still delivers consistent improvements across all benchmarks. For the L-size model, we estimate the additional computation introduced by the linear head added for classification-based supervision (see Table 3) using a single image-text pair, which accounts for only 0.077%. These results demonstrate that our method not only scales well with model and data size, but also consistently enhances CLIP's performance by better leveraging classification supervision—without introducing significant computational overhead. More FLOPs statistics of the models are provided in **Appendix A.3.**

Analysis of Performance Gain. We analyze word-image similarity statistics to demonstrate that, compared to CLIP, our SuperCLIP more effectively captures fine-grained visual attributes beyond global semantics. Visualization and statistical results are presented in Fig.3 and Table 4. We compute the similarity between each image and the words in its captions on the COCO validation set, measuring how much attention the model gives to different words. For each word, we then average its similarity across the dataset by dividing the total similarity by its frequency. After filtering

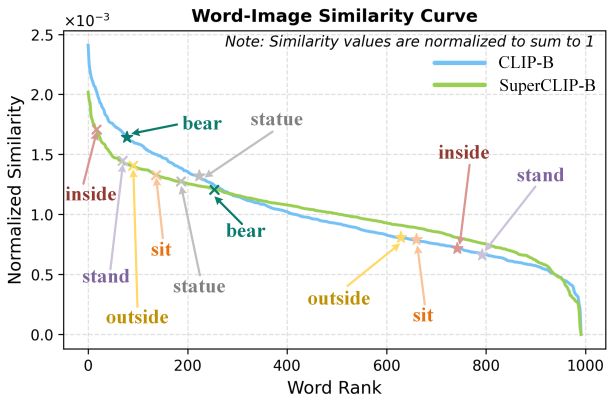

Figure 3: **Visualization of Word-Image Similarity Distribution.** We ranked the similarity scores of 1,000 words that appeared in the captions and highlighted the positions of fine-grained attributes discussed in the above Fig.1.

Component	CLIP	SuperCLIP
Vision Encoder	59.689	59.689
Text Encoder	6.547	6.547
Linear Head	-	0.051

Table 3: **FLOPs Count (GFLOPs).**

Metric	CLIP	SuperCLIP
Total words	992	992
Std Deviation	0.0340	0.0213
Value Range	0.2065	0.1401
Mean Slope	0.000208	0.000141
Top-1→100	0.0702	0.0439

Table 4: **Statistical Summary.** Mean Slope (Δsim): Average drop in similarity between words as the rank goes down. Top-1→100: Difference in similarity between the 1st and 100th word.

out low-frequency words (fewer than 20 occurrences), we obtain a set of about 1,000 words and rank them by average similarity. The results are both interesting and expected. CLIP tends to rank object category words (e.g., zebras, kites, elephants, often in the top 20) highest, as they are easily captured by global visual features. In contrast, SuperCLIP, with added classification supervision, raises the ranks of words describing object status, spatial relations, and actions (see Fig. 3). This is because classification supervision encourages the model to focus more on fine-grained visual details overlooked by CLIP. As shown in Table 4, SuperCLIP also produces more stable similarity rankings with lower variance across words, reducing the long-tail effect seen in CLIP. Overall, these results show that SuperCLIP better captures fine-grained attributes that CLIP often overlooks, leading to improved performance across multiple tasks. More statistical analyses on word-image similarity are provided in **Appendix A.4.**

4.3 Recover Textual Supervision

Comparison with CLIP using Mixed Caption. We demonstrate that our method helps CLIP recover overlooked textual supervision through extensive experiments, including **zero-shot** classification on 38 datasets [47] and image-text retrieval on COCO and Flickr30k (Table 5). Following [31], we train our models on a mix of DataComp-1B and Recap-DataComp-1B, with the latter containing longer, semantically richer captions. While such captions provide more detailed and fine-grained supervision, increasing their proportion in training tends to degrade CLIP's performance—suggesting that contrastive learning alone may not fully benefit from rich textual signals. In contrast, the classification supervision introduced in our method is better equipped to utilize this additional semantic information, thereby mitigating the limitations of contrastive learning. Under the **DualCaption** setting, contrastive learning captures coarse-grained semantics from short captions, while our classifier extracts fine-grained details from long ones—achieving strong performance without the need for the carefully tuned **"0.8/0.2"** mixing ratio identified through extensive search in [31]. Complete evaluation results across 38 datasets are provided in **Appendix A.5.**

4.4 Generalization Analysis

Generalize to Other CLIP-style Frameworks. We test the generalizability of our method on two CLIP-style models, SigLIP and FLIP, using **zero-shot** classification on ImageNet-1K and image-text retrieval on COCO and Flickr30K. Detailed results are presented in Table 6. Our SuperCLIP variants (SuperSigLIP and SuperFLIP) consistently outperform their baselines under the same pretraining setup. SuperSigLIP achieves gains of up to +3.7% in image classification and +2.9% in image/text retrieval. Similarly, SuperFLIP improves by +3.4% in classification, +2.6% in image retrieval, and up to +5.3% in text retrieval. This demonstrates that our method is not limited to CLIP, but is a generally effective enhancement to vision-language pretraining.

Model-Size	Mixed Caption	Image Retrieval		Text Retrieval		Image Classification
	Short / Long	COCO	Flickr	COCO	Flickr	Average. 38
CLIP-B	1.0 / 0.0	29.0	54.4	46.7	73.7	43.4
SuperCLIP-B	1.0 / 0.0	**31.3**	**57.6**	**47.8**	**75.6**	**44.5** (+1.1)
CLIP-B	0.0 / 1.0	23.6	41.8	40.5	66.2	27.8
SuperCLIP-B	0.0 / 1.0	**30.6**	**48.7**	**47.2**	**70.4**	**31.4** (+3.6)
CLIP-B	0.8 / 0.2	32.7	57.5	50.2	76.0	42.8
SuperCLIP-B	Dual	**34.1**	**60.2**	**51.2**	**76.6**	**45.1** (+2.3)
CLIP-L	1.0 / 0.0	32.7	57	49.6	76.4	45.7
SuperCLIP-L	1.0 / 0.0	**35.9**	**62.4**	**52.2**	**79.3**	**48.6** (+2.9)
CLIP-L	0.0 / 1.0	26.2	43.1	42.9	65.9	30.0
SuperCLIP-L	0.0 / 1.0	**34.2**	**55.7**	**52.1**	**75.0**	**33.8** (+3.8)
CLIP-L	0.8 / 0.2	37.0	61.1	53.7	78.8	46.8
SuperCLIP-L	Dual	**37.6**	**65.3**	**54.0**	**82.5**	**49.5** (+2.7)

Table 5: **Comparison with CLIP using Mixed Captions. "Mixed Caption"** refers to the ratio of short (DataComp-1B) and long (Recap-DataComp-1B) captions used during training. The **"0.8/0.2"** mix is the optimal ratio identified in [31] through extensive tuning. **"Dual"** denotes our setup where the contrastive loss uses only short captions and the classification loss uses only long captions. We report average **zero-shot** image classification accuracy (%) across 38 datasets, and **zero-shot** image/text retrieval (Recall@1, %) on COCO and Flickr30K, using **512M** training samples. **Bold** numbers indicate the best results, while values in parentheses show absolute gains or drops of SuperCLIP relative to CLIP.

Model	Image Classification		Image Retrieval		Text Retrieval	
	val	v2	COCO	Flickr	COCO	Flickr
SigLIP	60.4	52.8	29.8	53.9	45.8	73.2
SuperSigLIP	64.1 (+3.7)	55.9 (+3.1)	32.5 (+2.7)	56.8 (+2.9)	48.6 (+2.8)	75.9 (+2.7)
FLIP	58.1	50.1	27.5	51.8	44.1	66.7
SuperFLIP	61.3 (+3.2)	53.5 (+3.4)	30.1 (+2.6)	54.0 (+2.2)	46.7 (+2.6)	72.0 (+5.3)

Table 6: **Generalization to Other CLIP-Style Frameworks.** We report **zero-shot** performance on image classification accuracy (%) on ImageNet-1K (val and v2), and image/text retrieval (Recall@1, %) on COCO and Flickr30K, comparing SigLIP and FLIP with their SuperCLIP variants (SuperSigLIP and SuperFLIP). All models are pretrained with 512M samples (B-512M). Numbers in parentheses indicate absolute gains over the original models.

Enhance CLIP for Purely Visual Tasks. We demonstrate how our method enhances CLIP for purely visual tasks, through **linear probing** image classification on ImageNet, semantic segmentation on Pascal [9] and ADE20K [71], and depth estimation on NYUv2 [43]. Detailed results are presented in Table 7. For linear probing image classification experiments, we freeze the backbone and train a linear classification head. For the semantic segmentation and depth estimation tasks, we similarly attach a linear head to the backbone, but fine-tune the entire model. SuperCLIP consistently improves performance across all tasks, indicating that the vision encoder trained with our method learns more effective and discriminative visual representations.

4.5 Impact of Batch Size

Mitigate CLIP's Drop with Limited Batch Sizes. We examine the extent to which our method mitigates CLIP's performance degradation under small batch sizes, through **zero-shot** and **linear probing** classification on ImageNet across batch sizes ranging from 1K to 32K. Detailed results are presented in Fig.4. For zero-shot classification (Left), SuperCLIP shows clear advantages under small-batch training, where CLIP suffers significant degradation. For linear probing (Right), SuperCLIP

Model	Pretrian	Class ↑	Segmentation ↑		Depth ↓
		ImageNet-1K	PASCAL	ADE20k	NYUv2
CLIP	B-512M	75.6	57.8	28.0	0.768
SuperCLIP	B-512M	77.1 (+1.5)	65.5 (+7.7)	32.1 (+4.1)	0.746 (-0.022)
CLIP	L-512M	79.7	67.8	34.2	0.740
SuperCLIP	L-512M	81.0 (+1.3)	71.2 (+3.4)	36.3 (+2.1)	0.733 (-0.007)

Table 7: **Enhance CLIP for Purely Visual Tasks.** We report performance on three purely visual tasks: **linear probing** image classification(Class) on ImageNet-1K (Accuracy, %), semantic segmentation(Segmentation) on PASCAL and ADE20K (mIoU), and depth estimation(Depth) on NYUv2 (RMSE). We compare CLIP and SuperCLIP under identical pretraining and evaluation settings to ensure a fair comparison across all purely visual tasks. Numbers in parentheses indicate absolute improvements over the original CLIP models.

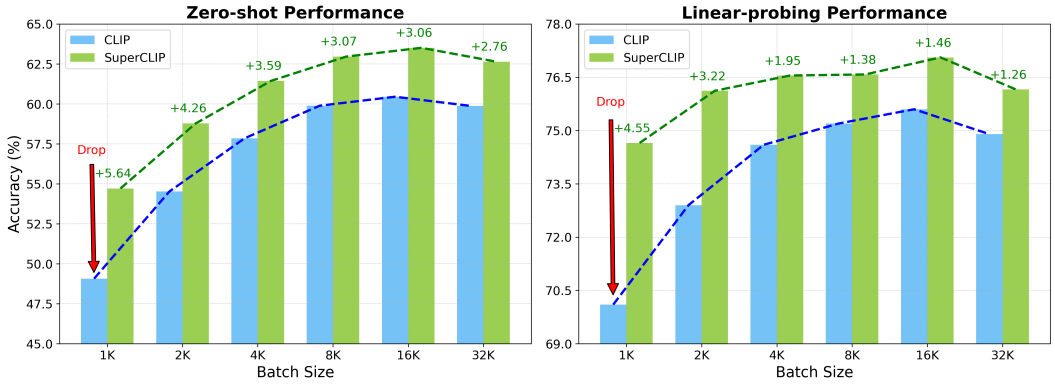

Figure 4: **Mitigate CLIP's Drop with Limited Batch Sizes.** We report zero-shot (Left) and linear-probing (Right) image classification accuracy (%) on ImageNet-1K (val) under varying batch sizes. The green bars represent the performance of SuperCLIP under different batch sizes, while the gray bars indicate the performance of CLIP under the corresponding batch sizes. Green numbers indicate absolute improvements over the original CLIP models at the corresponding batch sizes.

maintains stable performance across all batch sizes, preserving high-quality visual representations even in low-resource settings where CLIP's performance drops noticeably. The above results validate that our method effectively mitigates the performance degradation of CLIP under limited batch size conditions. This improvement is attributed to the introduction of classification supervision, which is inherently insensitive to batch size.

4.6 Integrate in Multi-modal LLM

Compare with CLIP under Multi-modal LLM Setting. We evaluate SuperCLIP beyond CLIP-style contrastive pretraining by integrating both CLIP and SuperCLIP (ViT-B/16, 512M samples) into the LLaVA-1.5 framework [37], combined with the Vicuna-7B language model [7], enabling a fair comparison within the multi-modal LLM setting. This setup supports effective multi-modal reasoning and instruction following across a broad range of downstream benchmarks, including VQAv2 [13], GQA [21], VizWiz [15], T-VQA [48], SQA [69], MMB (MMBench) [39], MME [4] and POPE [34]. Detailed results are presented in Table 8. These experiments confirm SuperCLIP's superior performance over CLIP encoders across multiple benchmarks, particularly on VQAv2 and MMBench, which focus on general visual question answering and fine-grained recognition, respectively. The strong transfer performance demonstrates that SuperCLIP is not only effective in contrastive pretraining but also exhibits excellent cross-modal generalization when integrated into large-scale multi-modal frameworks.

4.7 Additional Ablation Studies

Ablation on Loss Weighting and IDF Weighting We study the effect of weighting the classification loss by multiplying the $\mathcal{L}_{\text{Class}}$ term in Eq. 7 by a factor λ. As shown in Table 9, performance improves

Model	Pretrian	Vision & Language Downstream Tasks							
		VQAv2	GQA	VizWiz	T-VQA	SQA	MMB	MME	POPE
CLIP	B-512M	67.8	55.4	42.1	47.8	**69.3**	49.1	1453	81.7
SuperCLIP	B-512M	**69.6**	**57.5**	**44.4**	**48.4**	69.1	**55.9**	**1562**	**82.0**

Table 8: **Compare with CLIP under Multi-modal LLM Setting.** We report the performance scores on 8 vision & language downstream tasks. **Bold** numbers indicate the best result.

Task	Weighting Factor (λ)				
	0.4	**0.6**	**1**	**1.4**	**1.6**
Classification	44.1	45.0	47.1	46.9	47.2
Image Retrieval	41.3	42.1	44.0	43.8	44.2
Text Retrieval	58.3	59.8	61.0	60.9	62.0

Table 9: **Loss Weighting.** We report **zero-shot** classification accuracy (%) on ImageNet-1K (val) and the average retrieval result (Recall@1, %) across COCO and Flickr30K.

Design	Image Retrieval		Text Retrieval		Classification
	COCO	**Flickr**	**COCO**	**Flickr**	**ImageNet-1K**
w/o IDF	31.6	51.7	48.0	71.1	44.8
IDF	**33.2**	**54.7**	**48.9**	**73.1**	**47.1**

Table 10: **IDF Weighting.** We report **zero-shot** classification accuracy (%) on ImageNet-1K (val) and retrieval results (Recall@1, %) on COCO and Flickr30K, respectively.

across all tasks as λ increases from 0.4 to 1.0. Beyond 1.0, text retrieval continues to improve, whereas image retrieval and classification saturate. This confirms the effectiveness of the classification loss, and we recommend $\lambda \geqslant 1.0$ in practice. Then, we assess the role of IDF weighting by comparing IDF-weighted and unweighted K-hot labels. As shown in Table 10, IDF consistently improves performance across all benchmarks, confirming its benefit. All additional ablation studies use a ViT-B/16 model trained on 256M samples with all other settings unchanged.

5 Conclusion and Future Work

We introduce SuperCLIP, a simple yet effective framework that adds classification supervision to CLIP-style vision–language pretraining. By treating raw text tokens as classification labels, SuperCLIP recovers rich semantic signals often missed by contrastive learning, enabling better use of full textual content beyond coarse image-text alignment. SuperCLIP consistently improves performance across a wide range of tasks, including zero-shot classification, image-text retrieval, linear probing, and purely visual benchmarks. It enhances CLIP's ability to achieve fine-grained visual-text alignment, while requiring no additional annotations or significant computational cost. Its batch-size-independent classification loss also mitigates CLIP's performance drop under small-batch settings, making it more practical for real-world applications. We hope these findings encourage further research into combining classification and contrastive learning in large-scale multimodal models. For the future work, SuperCLIP focuses on enhancing the supervision from text to the vision encoder. A promising direction is to explore whether a similar approach can improve the supervision from images to the text encoder.

Acknowledgement: This work was partially supported by the National Natural Science Foundation of China (No. 62276108).

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication, 2023. *URL https://arxiv.org/abs/2303.09540*, 2021.

[2] Mothilal Asokan, Kebin Wu, and Fatima Albreiki. Finelip: Extending clip's reach via fine-grained alignment with longer text inputs. *arXiv preprint arXiv:2504.01916*, 2025.

[3] Lorenzo Bianchi, Fabio Carrara, Nicola Messina, and Fabrizio Falchi. Is clip the main roadblock for fine-grained open-world perception? In *2024 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–8. IEEE, 2024.

[4] Fu Chaoyou, Chen Peixian, Shen Yunhang, Qin Yulei, Zhang Mengdan, Lin Xu, Yang Jinrui, Zheng Xiawu, Li Ke, Sun Xing, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 3, 2023.

[5] Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Vitamin: Designing scalable vision models in the vision-language era. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12954–12966, 2024.

[6] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[7] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Mark Everingham, Luc Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[10] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023.

[11] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

[12] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022.

[13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[14] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[15] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[16] Hasan Abed Al Kader Hammoud and Bernard Ghanem. Diffclip: Differential attention meets clip. *arXiv preprint arXiv:2503.06626*, 2025.

[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.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[19] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023.

[20] Zilong Huang, Qinghao Ye, Bingyi Kang, Jiashi Feng, and Haoqi Fan. Classification done right for vision-language pre-training. *Advances in Neural Information Processing Systems*, 37:96483–96504, 2024.

[21] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[22] Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Retrieval-enhanced contrastive vision-text models. *arXiv preprint arXiv:2306.07196*, 2023.

[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 *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[24] Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Wei Wei, Huiwen Zhao, Zhiwu Lu, et al. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *Advances in Neural Information Processing Systems*, 37:27896–27918, 2024.

[25] Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 67–84. Springer, 2016.

[26] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021.

[27] Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, et al. Veclip: Improving clip training via visual-enriched captions. In *European Conference on Computer Vision*, pages 111–127. Springer, 2024.

[28] Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192, 2017.

[29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[30] 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 *CVPR*, 2022.

[31] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024.

[32] 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. *arXiv preprint arXiv:2110.05208*, 2021.

[33] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23390–23400, 2023.

[34] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[35] 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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[36] Chong Liu, Yuqi Zhang, Hongsong Wang, Weihua Chen, Fan Wang, Yan Huang, Yi-Dong Shen, and Liang Wang. Efficient token-guided image-text retrieval with consistent multimodal contrastive training. *IEEE Transactions on Image Processing*, 32:3622–3633, 2023.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[38] Yanqing Liu, Xianhang Li, Zeyu Wang, Bingchen Zhao, and Cihang Xie. Clips: An enhanced clip framework for learning with synthetic captions. *arXiv preprint arXiv:2411.16828*, 2024.

[39] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[40] Sachin Mehta, Maxwell Horton, Fartash Faghri, Mohammad Hossein Sekhavat, Mahyar Najibi, Mehrdad Farajtabar, Oncel Tuzel, and Mohammad Rastegari. Catlip: Clip-level visual recognition accuracy with 2.7 x faster pre-training on web-scale image-text data. *arXiv preprint arXiv:2404.15653*, 2024.

[41] Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First international workshop on multimedia intelligent storage and retrieval management*, volume 18. Citeseer, 1999.

[42] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pages 529–544. Springer, 2022.

[43] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[48] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[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.

[50] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024.

[51] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13019–13029, 2024.

[52] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[53] 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.

[54] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.

[55] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[56] Michael Tschannen, Basil Mustafa, and Neil Houlsby. Clippo: Image-and-language understanding from pixels only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11006–11017, 2023.

[57] Bin Wang, Chunyu Xie, Dawei Leng, and Yuhui Yin. Iaa: Inner-adaptor architecture empowers frozen large language model with multimodal capabilities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21035–21043, 2025.

[58] Xinze Wang, Chen Chen, Yinfei Yang, Hong-You Chen, Bowen Zhang, Aditya Pal, Xiangxin Zhu, and Xianzhi Du. Clip-up: A simple and efficient mixture-of-experts clip training recipe with sparse upcycling. *arXiv preprint arXiv:2502.00965*, 2025.

[59] Zihao Wei, Zixuan Pan, and Andrew Owens. Efficient vision-language pre-training by cluster masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26815–26825, 2024.

[60] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21970–21980, 2023.

[61] Wei Wu, Kecheng Zheng, Shuailei Ma, Fan Lu, Yuxin Guo, Yifei Zhang, Wei Chen, Qingpei Guo, Yujun Shen, and Zheng-Jun Zha. Lotlip: Improving language-image pre-training for long text understanding. *arXiv preprint arXiv:2410.05249*, 2024.

[62] 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. *arXiv preprint arXiv:2505.05071*, 2025.

[63] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18134–18144, 2022.

[64] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19163–19173, 2022.

[65] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[66] 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.

[67] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[68] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer, 2024.

[69] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[70] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022.

[71] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

[72] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper is empirical in nature and does not include theoretical results, assumptions, or formal proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Section 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and models will be made publicly available upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include error bars, confidence intervals, or statistical significance tests. It primarily reports deterministic results from large-scale pretraining and evaluation. Due to the scale and cost of such pretraining (e.g., hundreds of millions of samples), repeated runs or variance analysis are impractical.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix A.6.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research fully complies with the NeurIPS Code of Ethics. It uses only publicly available datasets and does not involve human subjects, private data, or sensitive content.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The work does not have any foreseeable societal impact. It focuses on improving vision-language pretraining methods without involving sensitive data, human subjects, or deployment-oriented applications. Therefore, we believe there are no direct or indirect risks associated with the proposed approach.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not involve the release of models or datasets that pose a high risk of misuse. It builds on publicly available datasets and does not include pretrained models or components with dual-use concerns.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses publicly available datasets that are properly cited with their original sources and licenses. We also rely on the official implementation of CLIP from OpenCLIP, which is licensed under the MIT License. All assets used are cited with appropriate references and used in accordance with their licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We will release the SuperCLIP model weights and training code upon acceptance. The release will include documentation covering model usage, training details, and licensing information to ensure reproducibility and responsible use.

Guidelines:

- The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects, and thus IRB or equivalent approval is not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve the use of LLMs as any important, original, or non-standard component.

    Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.