FG-CLIP: Fine-Grained Visual and Textual Alignment

Chunyu Xie [1 2 *] **Bin Wang** [2 *] **Fanjing Kong** [2] **Jincheng Li** [2] **Dawei Liang** [2] **Gengshen Zhang** [2]
Dawei Leng [2] **Yuhui Yin** [2]

Code: https://github.com/360CVGroup/FG-CLIP
Model: https://huggingface.co/qihoo360/fg-clip-large
Dataset: https://huggingface.co/datasets/qihoo360/FineHARD

Abstract

Contrastive Language-Image Pre-training (CLIP) excels in multimodal tasks such as image-text retrieval and zero-shot classification but struggles with fine-grained understanding due to its focus on coarse-grained short captions. To address this, we propose Fine-Grained CLIP (FG-CLIP), which enhances fine-grained understanding through three key innovations. First, we leverage large multimodal models to generate 1.6 billion long caption-image pairs for capturing global-level semantic details. Second, a high-quality dataset is constructed with 12 million images and 40 million region-specific bounding boxes aligned with detailed captions to ensure precise, context-rich representations. Third, 10 million hard fine-grained negative samples are incorporated to improve the model's ability to distinguish subtle semantic differences. We construct a comprehensive dataset, termed FineHARD, by integrating high-quality region-specific annotations with hard fine-grained negative samples. Corresponding training methods are meticulously designed for these data. Extensive experiments demonstrate that FG-CLIP outperforms the original CLIP and other state-of-the-art methods across various downstream tasks, including fine-grained understanding, open-vocabulary object detection, image-text retrieval, and general multimodal benchmarks. These results highlight FG-CLIP's effectiveness in capturing fine-grained image details and improving overall model performance. The data, code, and models are available at https://github.com/360CVGroup/FG-CLIP.

[*]Equal contribution [1]Beihang University [2]360 AI Research. Correspondence to: Dawei Leng <lengdawei@360.cn>.

Proceedings of the 42^{st} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

The integration of vision and language (Alayrac et al., 2022; Ramesh et al., 2022; Lin et al., 2023; Gabeff et al., 2024) has been a long-standing goal in artificial intelligence, aiming to develop models that can understand and reason about the world in a visually and linguistically rich manner. Recent advances in multimodal pre-training, such as CLIP (Radford et al., 2021), have made significant strides in this direction by learning joint representations of images and text through contrastive learning. These models have achieved state-of-the-art performance in a variety of downstream tasks, including image-text retrieval (Pan et al., 2023; Sun et al., 2024; Zhang et al., 2024), image captioning (Mokady et al., 2021; Li et al., 2024; Yao et al., 2024), and visual question answering (Li et al., 2023a; Parelli et al., 2023; Team et al., 2024; Wang et al., 2025). However, despite their impressive capabilities, these models often struggle with fine-grained details, particularly in recognizing object attributes and their relationships.

Recent works (Liu et al., 2023a; Wu et al., 2024b; Zhang et al., 2024; Zheng et al., 2024; Jing et al., 2024) point out two primary reasons for the limitations in CLIP's fine-grained learning capability. First, the original CLIP model's text encoder supports only up to 77 tokens, restricting its capacity to process detailed descriptions and hindering its ability to capture nuanced textual information. Second, CLIP aligns entire images with corresponding text descriptions, making it challenging to extract valuable region-specific representations from visual features. Consequently, the model struggles to achieve fine-grained alignment between image regions and their corresponding textual attributes, limiting its effectiveness in complex recognition scenarios.

To address these issues, researchers have proposed extending the positional encoding to support longer token sequences (Wu et al., 2024b; Zhang et al., 2024; Zheng et al., 2024) and integrating object detection datasets into CLIP training (Zhong et al., 2022; Jing et al., 2024). By aligning bounding boxes with category labels, these methods

aim to enhance regional feature extraction. Although these approaches have shown some improvements, they still fall short in fine-grained visual recognition and open-vocabulary object detection. Existing methods (Jing et al., 2024; Zhang et al., 2024) typically introduce relatively few long captions, usually on a million scale, which is inadequate for effective learning of fine-grained details. Additionally, aligning image regions with category labels limits semantic diversity, restricting the model's generalization to open-world scenarios. Furthermore, the lack of hard fine-grained negative samples limits the model's ability to distinguish between objects of the same category but with different attributes. In this work, we introduce Fine-Grained CLIP (FG-CLIP), a novel approach designed to enhance CLIP's fine-grained understanding capabilities through three key innovations.

First, we significantly enhance global-level semantic alignment by generating long captions using state-of-the-art large multimodal models (LMMs) (Hong et al., 2024). This process introduces 1.6 billion long caption-image pairs, providing an unprecedented scale of data that allows FG-CLIP to capture nuanced details at the global-level semantic layer, thereby enhancing its ability to perceive complex and detailed information.

Second, to improve fine-grained alignment between images and text, we develop a high-quality visual grounding dataset. This dataset includes detailed descriptions for 40 million bounding boxes across 12 million images, ensuring that each region is precisely annotated with context-rich captions. By creating such an extensive and richly annotated dataset, we enable the model to learn precise and contextually rich representations, significantly enhancing its performance on tasks that require fine-grained understanding.

Third, to further enhance model robustness and discrimination abilities, we introduce a large-scale corpus of 10 million hard fine-grained negative samples. By incorporating these challenging negative samples into the training process, FG-CLIP learns to distinguish subtle differences in semantically similar but distinct pairs, thereby significantly improving its performance across various downstream tasks. We integrate the high-quality visual grounding data and hard fine-grained negative samples as a whole dataset called FineHARD.

Compared to previous methods, FG-CLIP demonstrates significant improvements across a wide range of benchmark tasks. Our comprehensive enhancements enable the model to achieve superior performance in capturing nuanced visual details, as evidenced by our state-of-the-art results on tasks such as fine-grained understanding, bounding box classification, long caption image-text retrieval, and open-vocabulary object detection. Moreover, when utilized as the backbone for LMMs (Liu et al., 2023b), FG-CLIP also demonstrates performance improvements in tasks involving attribute analysis (Hudson & Manning, 2019), object localization (Kazemzadeh et al., 2014), and reducing output hallucination (Li et al., 2023c). We provide visualization results in Appendix C to demonstrate the improvement in fine-grained understanding. These results highlight FG-CLIP's effectiveness in capturing fine-grained image details and improving overall model performance. To facilitate future research and application, we make the models, datasets, and code publicly available at https://github.com/360CVGroup/FG-CLIP.

2. Related Work

2.1. Contrastive Language-Image Pre-training

Contrastive learning has emerged as a powerful paradigm in multimodal pre-training, significantly advancing the field of image-text alignment. Models like CLIP have revolutionized this area by leveraging large-scale image-text pairs to learn rich representations without explicit supervision. CLIP achieves this through a dual-encoder architecture that maps images and their corresponding text descriptions into a shared embedding space, where semantically similar pairs are pulled closer together while dissimilar pairs are pushed apart. This approach not only simplifies data labeling but also enables zero-shot transfer to downstream tasks, demonstrating impressive performance on various benchmarks such as image classification (Deng et al., 2009; Recht et al., 2019) and image-text retrieval (Young et al., 2014; Lin et al., 2014; Urbanek et al., 2024; Chen et al., 2024a).

2.2. Fine-Grained Understanding

Despite its success, CLIP faces limitations in handling fine-grained visual details. Its text encoder is constrained to 77 tokens, limiting its capacity to process detailed and complex descriptions. Additionally, CLIP aligns entire images with corresponding text, making it challenging to extract valuable region-specific representations. To address these limitations, models like LongCLIP (Zhang et al., 2024) extend the maximum token length of the text encoder, enabling it to handle longer and more detailed textual information. GLIP (Li et al., 2022) and RegionCLIP (Zhong et al., 2022) introduce grounding data, enhancing the model's ability to align specific regions within images with corresponding text, thereby improving performance on downstream detection tasks (Xie et al., 2018; Gupta et al., 2019; Zhou et al., 2022b; Minderer et al., 2024). However, even with these improvements, existing models still struggle to fully capture and align fine-grained features across diverse datasets.

2.3. Image-Text Datasets

Image-text datasets (Gu et al., 2022; Xie et al., 2023; Fu et al., 2024) play a pivotal role in the performance of multimodal models. While existing datasets such as LAION (Schuhmann et al., 2021; 2022), COCO (Lin et al.,

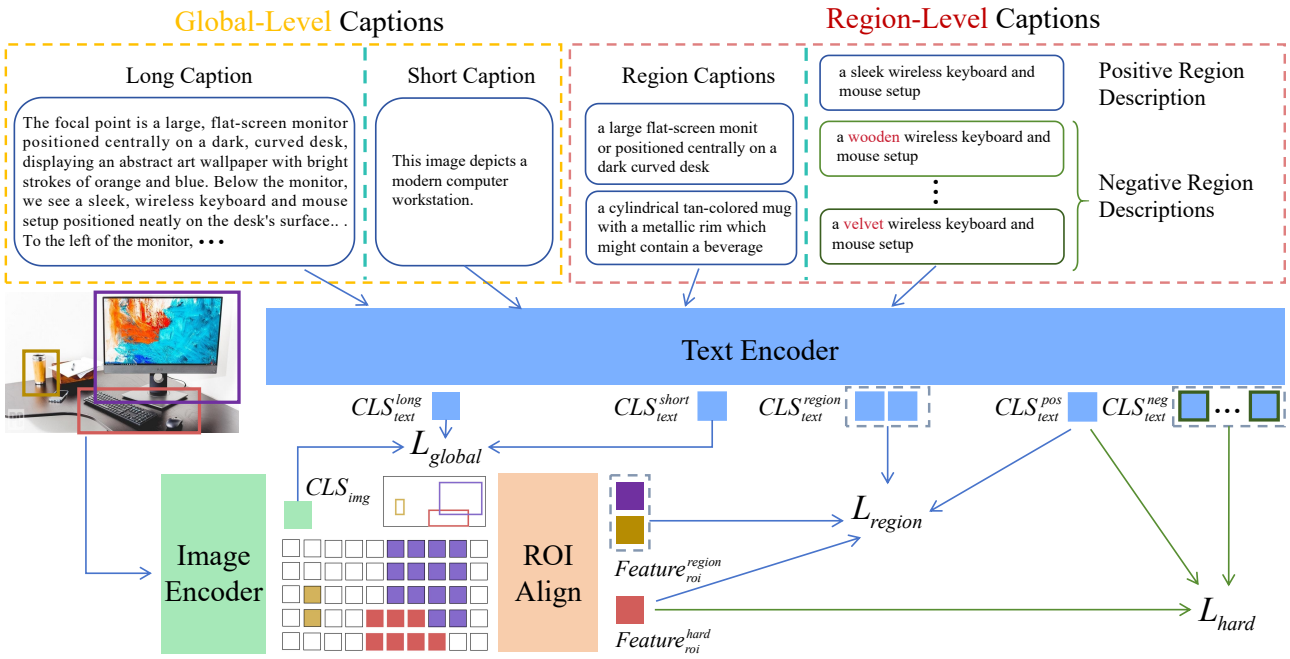


Figure 1. Overview of the FG-CLIP. CLS_{img} denotes the image class features output by the Vision Transformer (ViT), while CLS_{text} represents the class features summarized by the text model for multiple inputs, including long captions, short captions, region captions, and positive&negative descriptions of specific regions within images. FG-CLIP's training proceeds in two stages: the first stage leverages global-level caption-image pairs to achieve initial fine-grained alignment, while the second stage supplements these with additional region-level captions, including detailed region captions and positive/negative region descriptions to further refine the alignment.

2014), Flickr30K (Young et al., 2014), and Conceptual Captions (Sharma et al., 2018; Changpinyo et al., 2021) offer valuable resources, they often emphasize general scene descriptions, neglecting fine-grained details critical for advanced applications. Researchers have adopted several strategies to mitigate these limitations. One approach involves leveraging advanced large multimodal models (Laurençon et al., 2024; Wang et al., 2024; Wu et al., 2024c; Chen et al., 2024b; Team et al., 2024) to refine and enrich text descriptions through recaptioning. For instance, LongCLIP (Zhang et al., 2024) utilizes 1 million long caption-image pairs from ShareGPT4V (Chen et al., 2024a), and FineCLIP (Jing et al., 2024) constructs a dataset of 2.5 million long caption-image pairs. Although these efforts enhance data richness, they remain limited in scale compared to the vast amount of data in the image-text field. Another strategy is to implement pseudo-labeling pipelines using pre-trained object detection models (Li et al., 2023b; Ma et al., 2024; Hou et al., 2024) to automatically generate fine-grained pseudo-labels for region boxes, similar to the GRIT dataset utilized in Kosmos-2 (Peng et al., 2024). These methods help improve region-specific alignment but may introduce noise due to automated labeling.

Another significant challenge is the scarcity of hard fine-grained negative samples. Existing datasets predominantly consist of positive examples that are relatively easy to distinguish, limiting the model's ability to learn subtle variations. The absence of hard negative samples impedes true fine-grained understanding, as models struggle to discern small but meaningful differences in visual and textual features. Addressing this gap is essential for developing models capable of reliably performing fine-grained recognition and alignment tasks, thereby enabling them to handle the nuanced distinctions necessary for advanced applications.

3. Approach

3.1. Fine-Grained CLIP

Figure 1 provides an overview of Fine-Grained CLIP (FG-CLIP). Our proposed FG-CLIP extends the traditional dual-encoder architecture of CLIP to better capture fine-grained details in images and text. We leverage a two-stage training paradigm to achieve this enhancement. In the first stage, FG-CLIP focuses on aligning global representations of images and text using only global contrastive learning. The second stage builds on this foundation by introducing regional contrastive learning and hard fine-grained negative samples learning, leveraging region-text data to further refine the model's understanding of fine-grained details.

Global Contrastive Learning. Global contrastive learning aims to enhance the model's fine-grained understanding

by introducing a method of augmenting long caption alignment utilizing Large Multimodal Models (LMMs). This approach generates additional long captions that provide richer context and finer-grained descriptions. The inclusion of long captions enables the model to perceive and align with global-level semantic details, thereby enhancing fine-grained understanding and context awareness. In addition, we retain the alignment of short caption-image pairs. The long captions complement these short captions, ensuring that the model learns from both detailed, nuanced long captions for complex semantic information and concise, direct short captions for basic concepts. This dual approach improves the model's overall performance in capturing a broader spectrum of visual information.

In our framework, both short and long captions are aligned with images by utilizing the [CLS] token features extracted from the text encoder for the captions and the [CLS] token features from the image encoder for the images. To accommodate longer and more detailed captions while preserving the alignment of short captions, position embeddings of FG-CLIP's text encoder are extended. Specifically, for sequences shorter than or equal to 20 tokens, we use the original position embedding directly. For longer sequences, we apply linear interpolation with a factor of 4 for positions beyond 20, extending the maximum length from 77 to 248 tokens. This modification ensures that the model can effectively handle longer, more descriptive text while maintaining computational efficiency.

During each training step, the model employs both a short caption and a long caption for every image to ensure comprehensive and fine-grained understanding. Given an image-text pair, the outputs of both encoders are embeddings $v \in \mathbb{R}^d$ for images and $t \in \mathbb{R}^d$ for text, where d is the dimensionality of the embedding space. We compute the similarity between each pair using the cosine similarity metric:

$$s(v,t) = \frac{v \cdot t^T}{\|v\|\|t\|}. \tag{1}$$

The objective function for global contrastive learning is based on the InfoNCE loss (He et al., 2020), which maximizes the similarity between matching pairs while minimizing the similarity between mismatched pairs. Specifically, the loss for a batch of N image-text pairs is given by:

$$L_{global} = -\frac{1}{2N}\sum_{i=1}^{N}(\log \frac{\exp(s(v_i,t_i)/\tau)}{\sum_{j=1}^{N}\exp(s(v_i,t_j)/\tau)} + \log \frac{\exp(s(t_i,v_i)/\tau)}{\sum_{j=1}^{N}\exp(s(t_i,v_j)/\tau)}), \tag{2}$$

where τ is a learnable temperature parameter. This global contrastive learning significantly improving its detail perception capabilities in both granular and holistic contexts.

Regional Contrastive Learning. Regional contrastive learning focuses on aligning specific regions within images with corresponding text segments. To achieve this, we employ RoIAlign (He et al., 2017) to extract region-specific features from the image. These extracted features are then processed by applying average pooling over the tokens within each detected region, resulting in a set of region embeddings $\{r_k\}_{k=1}^{K}$, where K denotes the total number of valid bounding boxes across all images within a batch. This approach differs from global contrastive learning, which typically relies on the [CLS] token for deriving image-level features. For text, we segment the full-image caption into phrases or sentences that correspond to individual bounding boxes, obtaining text embeddings l_k. The regional contrastive loss is defined as:

$$L_{regional} = -\frac{1}{2K}\sum_{i=1}^{K}(\log \frac{\exp(s(r_i,l_i)/\tau)}{\sum_{j=1}^{K}\exp(s(r_i,l_j)/\tau)} + \log \frac{\exp(s(l_i,r_i)/\tau)}{\sum_{j=1}^{K}\exp(s(l_i,r_j)/\tau)}). \tag{3}$$

This encourages the model to learn fine-grained alignments between specific regions and textual descriptions.

Hard Fine-Grained Negative Samples Learning. To address the scarcity of challenging fine-grained negative samples, we introduce a hard negative mining strategy. We define hard negative samples as those that are semantically close but not identical to the positive sample. These hard negatives are constructed by rewriting the descriptions of bounding boxes, modifying certain attributes to create subtle differences. The specific process of obtaining hard fine-grained negative samples can be found in Section 3.2.

To incorporate hard negative samples into the learning process, we extend the loss function to include a term for hard negatives. For each region-text pair, we compute the similarity between the regional feature and both the positive description and the corresponding negative sample descriptions. The hard negative loss L_{hard} is defined as:

$$L_{hard} = -\frac{1}{K}\sum_{i=1}^{K}\log \frac{\exp(s(r_i,l_{i,1})/\tau)}{\sum_{j=1}^{M}\exp(s(r_i,l_{i,j})/\tau)}, \tag{4}$$

where M denotes the total number of captions for each region, with $j = 1$ corresponding to the positive sample, and $j > 1$ corresponding to the negative samples.

In the second stage, we integrate all three components: Global Contrastive Learning, Regional Contrastive Learning, and Hard Fine-Grained Negative Samples Learning, to

ensure comprehensive and nuanced alignment tasks. The learning objective in the second stage combines these elements:

$$L = L_{global} + \alpha * L_{regional} + \beta * L_{hard}. \quad (5)$$

Here, the hyperparameters α and β are set to 0.1 and 0.5, respectively, to balance the regional contrastive loss and the hard negative loss, ensuring that each loss operates on similar scales.

This integrated approach ensures that FG-CLIP not only captures global-level semantic details but also distinguishes subtle differences in semantically similar pairs, enhancing its overall performance across various downstream tasks.

3.2. Curated Dataset

In this section, we describe the meticulous process of curating datasets for our FG-CLIP model, emphasizing both scale and quality to address the limitations of existing models in fine-grained understanding.

Enhancing LAION-2B Data with Detailed Recaptioning. In the first stage of training, we utilize an enhanced version of the LAION-2B dataset (Schuhmann et al., 2022), where images are recaptioned with detailed descriptions generated by large multimodal models, i.e., CogVLM2-19B (Hong et al., 2024). This approach generates more detailed and contextually rich captions, crucial for capturing subtle differences in visual content. The original LAION-2B dataset often suffers from overly generic or imprecise captions, leading to suboptimal performance in fine-grained tasks. For instance, an image of a bird might be described as "a bird", without specifying the species or environment. Such generic captions limit the model's ability to recognize fine details.

By leveraging advanced large multimodal models, we generate detailed descriptions that not only identify objects but also provide rich contextual information about their attributes, actions, and relationships within the scene. For instance, rather than a generic description like "a bird", our refined captions read "a red-winged blackbird perched on a tree branch in a park." Utilizing a cluster of 160×910B NPUs, the data processing is completed in 30 days. An ablation study detailed in Section 4.5 evaluates the impact of using these high-quality, detailed captions. The results demonstrate significant improvements in model performance across various tasks, underscoring the critical role of large-scale, high-quality text annotations in enhancing both model accuracy and context understanding.

Fine-Grained Visual Grounding+Reception+Hard Negative Dataset (FineHARD). For the second stage, we develop a high-quality visual grounding dataset named **FineHARD**, featuring precise region-specific captions and hard negative samples. We curate the overall dataset based on GRIT (Peng et al., 2024) images. The process begins with generating detailed image captions using CogVLM2-19B (Hong et al., 2024), ensuring comprehensive and nuanced descriptions that capture the full context of each image. Following (Peng et al., 2024), we then use SpaCy (Honnibal et al., 2020) to parse the captions and extract the referring expressions. Subsequently, the images and referring expressions are fed into the pretrained object detection model, i.e., Yolo-World (Cheng et al., 2024) to obtain the associated bounding boxes. Non-maximum suppression is applied to eliminate overlapping bounding boxes, retaining only those with predicted confidence scores higher than 0.4. This process results in 12 million images and 40 million bounding boxes with fine-grained region captions. We provide examples of the images and their corresponding captions in Appendix A.

Next, to create challenging fine-grained negative samples, we modify attributes of bounding box descriptions while keeping the object names unchanged. For this task, we employ an open-source large language model, Llama-3.1-70B (Dubey et al., 2024), to generate 10 negative samples for each positive sample. To ensure clarity, we remove special symbols such as semicolons, commas, and line breaks from the generated descriptions. A quality check of 3,000 negative samples reveals that 98.9% are qualified, with only 1.1% considered noise—a level within the expected tolerance for unsupervised methods. This process generates subtle variations that better reflect real-world scenarios where objects may appear similar but differ in specific details. We illustrate examples of the fine-grained negative samples in Appendix B.

The resulting dataset includes 12 million images with fine-grained captions, 40 million bounding boxes with detailed region descriptions, and 10 million hard negative samples. The data pipeline utilizes a cluster of 160×910B NPUs and takes 7 days to complete. This comprehensive dataset enhances the model's ability to capture fine-grained details and provides a robust foundation for training FG-CLIP to distinguish subtle differences in visual and textual features.

4. Experiments

4.1. Implementation Details

In the first stage, we train on a dataset of 1.6 billion images, each paired with short and long texts. The model is initialized with weights from the original CLIP (Radford et al., 2021). For both ViT-B and ViT-L (Dosovitskiy, 2021) configurations, the batch size per NPU is set to 384. The learnable temperature parameter τ is initialized to 0.07. We utilize the AdamW optimizer with a learning rate of 1e-4,

Table 1. Results on FG-OVD benchmark. Accuracy is reported.

Method	Backbone	Fine-Grained Understanding			
		hard	medium	easy	trivial
CLIP	ViT-B/16	12.0	23.1	22.2	58.5
EVA-CLIP	ViT-B/16	14.0	30.1	29.4	58.3
Long-CLIP	ViT-B/16	9.2	18.4	16.2	51.8
FineCLIP	ViT-B/16	26.8	49.8	50.4	71.9
FG-CLIP	ViT-B/16	**46.1**	**66.6**	**68.7**	**83.4**
CLIP	ViT-L/14	15.4	25.3	25.7	38.8
EVA-CLIP	ViT-L/14	18.3	38.4	35.2	62.7
Long-CLIP	ViT-L/14	9.6	19.7	16.0	39.8
FineCLIP	ViT-L/14	22.8	46.0	46.0	73.6
FG-CLIP	ViT-L/14	**48.4**	**69.5**	**71.2**	**89.7**

Table 2. Bounding box classification results.

Method	Backbone	BBox Classification		
		COCO	LVIS	Open Images
CLIP	ViT-B/16	44.2	20.9	15.3
EVA-CLIP	ViT-B/16	30.6	14.4	8.8
RegionCLIP	ViT-B/16	40.0	22.2	19.1
CLIPSelf	ViT-B/16	43.7	7.8	11.4
Long-CLIP	ViT-B/16	36.7	18.2	14.9
FineCLIP	ViT-B/16	48.4	23.3	18.1
FG-CLIP	ViT-B/16	**52.3**	**28.6**	**20.6**
CLIP	ViT-L/14	33.8	9.3	8.3
EVA-CLIP	ViT-L/14	32.1	18.3	9.3
Long-CLIP	ViT-L/14	35.6	10.4	8.9
FineCLIP	ViT-L/14	54.5	22.5	19.1
FG-CLIP	ViT-L/14	**63.2**	**38.3**	**23.8**

weight decay of 0.05, β_1 of 0.9, β_2 of 0.98, and warmup steps for the first 200 iterations. The entire training process employs DeepSpeed's Zero-2 optimization technique and Bfloat16 precision to accelerate training, and the model is trained for one epoch.

In the second stage, we train on a dataset of 12 million images. Apart from long and short captions, this dataset includes high-quality visual grounding annotations and hard fine-grained negative samples. The model is initialized with weights obtained from the first stage. The batch size per GPU is set to 512. We employ the AdamW optimizer with a learning rate of 1e-6, weight decay of 0.001, β_1 of 0.9, β_2 of 0.98, and warmup steps for the first 50 iterations. Training acceleration techniques include DeepSpeed's Zero-2 optimization, CUDA's TF32 technology, and Bfloat16 precision, and the model is trained for one epoch.

4.2. Comparisons on Fine-grained Region-level Task

In this section, the primary methods included for comparison are CLIP (Radford et al., 2021), EVA-CLIP (Sun et al., 2023), Long-CLIP (Zhang et al., 2024), and FineCLIP (Jing et al., 2024). Additional methods involved in open-vocabulary detection include OV-RCNN (Zareian et al., 2021), RegionCLIP (Zhong et al., 2022), Detic (Zhou et al., 2022b), VLDet (Lin et al., 2022), RO-ViT (Kim et al., 2023b), CFM-ViT (Kim et al., 2023a), F-ViT(Wu et al., 2024a), and CLIPSelf (Wu et al., 2024a).

Table 3. Performance on open-vocabulary object detection task.

Method	Backbone	OV-COCO		
		AP_{50}^{novel}	AP_{50}^{base}	AP_{50}^{all}
OV-RCNN	RN50	17.5	41.0	34.9
RegionCLIP	RN50	26.8	54.8	47.5
Detic	RN50	27.8	51.1	45.0
VLDet	RN50	32.0	50.6	45.8
RO-ViT	ViT-B/16	30.2	-	41.5
RO-ViT	ViT-L/16	33.0	-	47.7
CFM-ViT	ViT-L/16	34.1	-	46.0
F-ViT	ViT-B/16	17.5	41.0	34.9
F-ViT+CLIPSelf	ViT-B/16	33.6	54.2	48.8
F-ViT+FineCLIP	ViT-B/16	29.8	45.9	41.7
F-ViT+FG-CLIP	ViT-B/16	**35.1**	51.7	47.4
F-ViT	ViT-L/14	24.7	53.6	46.0
F-ViT+CLIPSelf	ViT-L/14	38.4	60.6	54.8
F-ViT+FineCLIP	ViT-L/14	40.0	57.2	52.7
F-ViT+FG-CLIP	ViT-L/14	**41.2**	58.0	53.6

Fine-Grained Understanding. Based on the fine-grained benchmark FG-OVD constructed by (Bianchi et al., 2024), we evaluate open-source image-text alignment models. Unlike previous benchmarks such as MSCOCO (Lin et al., 2014) and Flickr (Young et al., 2014), which rely on global information for matching, this benchmark focuses on identifying specific local regions within images. Each region has one corresponding positive description and ten negative descriptions, with the negative samples derived from the positive text. This benchmark primarily comprises four subsets of varying difficulty levels: hard, medium, easy, and trivial. The increasing difficulty across these subsets is reflected in the degree of distinction between the texts to be matched. In the hard, medium, and easy subsets, one, two, and three attribute words are replaced, respectively. In the trivial subset, the texts are entirely unrelated. For the source collection of specific attribute words, please refer to (Bianchi et al., 2024).

During testing, following FineCLIP, we first extract dense features from the model by removing the last self-attention layer as suggested by (Zhou et al., 2022a). Subsequently, we combine the bounding box information provided by the benchmark with ROIAlign to obtain representative features. These features are used to calculate similarity scores with both positive and negative sample descriptions. Top-1 accuracy is adopted as the evaluation metric.

As shown in Table 1, FG-CLIP achieves significant improvements over existing models, particularly on the challenging hard and medium subsets, thanks to its hard fine-grained negative samples learning strategy. Examples of different

Table 4. Comparisons on image-level tasks, including long/short caption image-text retrieval, and zero-shot image classification.

Method	Backbone	ShareGPT4V		DCI		MSCOCO		Flickr30k		ImageNet-1K	ImageNet-v2
		I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	Top-1	Top-1
CLIP	ViT-B/16	78.2	79.6	45.5	43.0	51.8	32.7	82.2	62.1	68.4	61.9
EVA-CLIP	ViT-B/16	90.5	85.5	41.9	41.2	58.7	41.6	85.7	71.2	**74.7**	**67.0**
Long-CLIP	ViT-B/16	94.7	93.4	51.7	57.3	57.6	40.4	85.9	70.7	66.8	61.2
FineCLIP	ViT-B/16	70.6	73.3	35.5	34.4	54.5	40.2	82.5	67.9	55.7	48.8
FG-CLIP	ViT-B/16	**96.7**	**94.9**	**61.8**	**60.6**	**64.1**	**45.4**	**90.7**	**76.4**	69.0	61.8
CLIP	ViT-L/14	86.5	83.6	37.2	36.4	58.0	37.1	87.4	67.3	76.6	70.9
EVA-CLIP	ViT-L/14	91.5	89.4	47.2	47.8	64.2	47.9	89.2	77.9	**80.4**	**73.8**
Long-CLIP	ViT-L/14	95.8	95.6	44.2	52.5	62.8	46.3	90.0	76.2	73.5	67.9
FineCLIP	ViT-L/14	73.4	82.7	40.1	46.2	-	-	-	-	60.8	53.4
FG-CLIP	ViT-L/14	**97.4**	**96.8**	**66.7**	**66.1**	**68.9**	**50.9**	**93.7**	**81.5**	76.1	69.0

models' performance can be found in Appendix D.1.

Bounding Box Classification. To assess the model's local information recognition capabilities, we conduct zero-shot testing on COCO-val2017 (Lin et al., 2014), LVIS (Gupta et al., 2019), and Open Images (Kuznetsova et al., 2020), following the protocol of (Jing et al., 2024). This evaluation focuses on how well the model can classify objects within bounding boxes using only textual descriptions. Similar to the fine-grained understanding, we integrate known bounding box information from the benchmark with ROIAlign to obtain localized dense representations. Using all categories as textual inputs, we perform matching and recognition for each bounding box, evaluating Top-1 accuracy.

As shown in Table 2, FG-CLIP achieves leading performance in bounding box classification with the help of the regional contrastive learning strategy. Notably, Long-CLIP (Zhang et al., 2024), fine-tuned from CLIP using long texts, shows a significant decline in performance, indicating that long texts affect regional information granularity. Furthermore, FineCLIP uses region alignment data and incorporates a real-time self-distillation scheme, leading to meaningful improvements. While FineCLIP makes significant progress, FG-CLIP excels it by integrating regional and global information. This approach enhances FG-CLIP's ability to accurately recognize and classify regions within images, highlighting the effectiveness of FG-CLIP's training strategy.

Open-Vocabulary Object Detection. To further evaluate the fine-grained localization capability of our method, we employ FG-CLIP as the backbone for downstream open-vocabulary detection tasks. Following prior work (Wu et al., 2024a), we employ a two-stage detection architecture, F-VIT, with a frozen visual encoder. The comparative results are summarized in Table 3. Consistent with previous studies, we report the box AP at IoU 0.5 for base, novel, and all categories (AP_{50}^{novel}, AP_{50}^{base}, and AP_{50}^{all}) on OV-COCO. Notably, AP_{50}^{novel} is the primary focus of interest, as it measures the model's ability to recognize novel objects. Our findings indicate that FG-CLIP achieves leading performance in open-vocabulary detection tasks, highlighting its effectiveness in recognizing and localizing novel objects.

Table 5. Comparisons on General Multimodal Benchmarks.

Method	GQA	POPE	RefCOCO		
			val	testA	testB
LLaVA-v1.5+CLIP	61.9	85.9	76.2	83.4	67.9
	+1.0	*+0.9*	*+5.2*	*+3.1*	*+7.0*
LLaVA-v1.5+FG-CLIP	62.9	86.8	81.4	86.5	74.9

4.3. Comparisons on Image-level Task

Long/short Caption Image-Text Retrieval. To evaluate retrieval performance comprehensively, we conduct experiments on both long caption and short caption image-text retrieval tasks. For long-text retrieval, we follow the protocol of Long-CLIP and use the 1K subset of ShareGPT4V (Chen et al., 2024a) provided by it as the testset. Additionally, we incorporate a more challenging long caption image-text pair dataset from DCI (Urbanek et al., 2024), consisting of 7,805 pairs, into the evaluation. For short-text retrieval, we employ the classic MSCOCO 5K (Lin et al., 2014) and Flickr 1K (Young et al., 2014) evaluation sets, which are widely used benchmarks for assessing image-text alignment models. As shown in Table 4, FG-CLIP achieves significant performance improvements in both long/short caption image-text retrieval tasks. The model's ability to handle diverse caption lengths highlights its versatility and robustness in multimodal alignment.

Zero-shot Image Classification. We evaluate the zero-shot classification performance of our model on ImageNet-1K (Deng et al., 2009) and ImageNet-v2 (Recht et al., 2019). As illustrated in Table 4, despite being marginally behind EVA-CLIP, which is trained on a larger dataset, FG-CLIP demonstrates stable classification performance with enhanced regional and textual understanding capabilities

Table 6. Ablation study results for FG-CLIP. This table compares the performance of different configurations of our FG-CLIP model across multiple evaluation metrics, including long caption image-text retrieval (DCI), short caption image-text retrieval (MSCOCO), bounding box classification (COCO-val2017), and fine-grained understanding (FG-OVD). The results highlight the incremental improvements achieved by incorporating global contrastive learning (L_{global}), regional contrastive learning ($L_{regional}$), and hard fine-grained negative samples learning (L_{hard}).

Method	Long Retrieval		Short Retrieval		BBox Classification		Fine-Grained Understanding		
	I2T	T2I	I2T	T2I	Top-1	Top-5	hard	medium	easy
CLIP	45.5	43.0	51.8	32.7	44.2	72.3	12.0	23.1	22.2
FG-CLIP Stage1	58.3	57.5	64.6	44.9	47.2	74.2	21.8	41.6	36.2
+Stage2 (L_{global})	62.7	61.2	64.4	46.4	46.8	73.6	25.4	46.8	42.9
+Stage2 (L_{global},$L_{regional}$)	62.4	61.1	64.7	45.7	53.7	81.2	24.5	47.1	49.5
+Stage2 (L_{global},$L_{regional}$,L_{hard})	61.8	60.6	64.1	45.4	52.3	79.7	46.1	66.6	68.7

compared to the original baseline, CLIP. Additionally, when compared to Long-CLIP and FineCLIP, both of which aim to enhance fine-grained recognition capabilities, our model exhibits a notable advantage in classification accuracy.

4.4. Comparisons on General Multimodal Benchmarks

We compare FG-CLIP as a visual feature extractor for multimodal large language models with our baseline, CLIP. Specifically, we conduct experiments using LLaVA-v1.5-7B (Liu et al., 2023b), which itself is trained using CLIP. To ensure a fair comparison, all parameter configurations are kept consistent with those in the original LLaVA, and the model is trained using the data provided by LLaVA. Our evaluation focuses on benchmark sets related to attribute analysis, object localization, and output hallucination, which are GQA (Hudson & Manning, 2019), RefCOCO (Kazemzadeh et al., 2014), and POPE (Li et al., 2023c), respectively.

As shown in Table 5, FG-CLIP achieves certain improvements on GQA, which involves attribute-based question answering, and on POPE, which evaluates output hallucination. Additionally, it demonstrates significant gains on RefCOCO, a benchmark set that involves both attribute analysis and object localization. These results indicate the effectiveness of FG-CLIP's training strategy and the data construction, which are specifically designed to enhance fine-grained recognition and regional alignment. We provide more results in Section D.3.

4.5. Ablation Study

To systematically evaluate the contributions of different components in our FG-CLIP model, we conduct an ablation study with results summarized in Table 6.

Global Contrastive Learning and Detailed Recaptioning Data. We start by comparing the original CLIP model with FG-CLIP Stage 1 and Stage 2 incorporating global contrastive learning L_{global}. The results demonstrate that generating detailed captions significantly enhances performance across various tasks. Specifically, FG-CLIP Stage 1 outperforms CLIP in all metrics, highlighting the importance of fine-grained training data. Further improvements are observed when adding L_{global} in Stage 2, particularly in long caption image-text retrieval (DCI (Urbanek et al., 2024)) and fine-grained understanding (FG-OVD (Bianchi et al., 2024)). This underscores the effectiveness of detailed caption data combined with global contrastive learning in improving model performance.

Regional Contrastive Learning. We introduce regional contrastive learning $L_{regional}$ to evaluate its impact on capturing detailed image regions. Compared to configurations using only L_{global}, adding $L_{regional}$ leads to substantial improvements in bounding box classification accuracy from 46.8% to 53.7%, and FG-OVD easy dataset accuracy from 42.9% to 49.5%. These gains highlight the effectiveness of $L_{regional}$ in refining the model's ability to understand fine-grained details within specific image regions. Moreover, this component maintains strong performance in both retrieval and classification tasks, demonstrating its versatility.

Hard Fine-Grained Negative Samples Learning. We incorporate hard fine-grained negative samples learning L_{hard} to distinguish subtle differences in semantically similar but distinct region-text pairs. By comparing configurations with and without L_{hard}, we observe significant improvements in FG-OVD performance. Accuracy on the hard dataset increases from 24.5% to 46.1%, while on the medium dataset it rises from 47.1% to 66.6% and on the easy dataset it jumps from 49.5% to 68.7%. These results underscore the importance of L_{hard} in distinguishing subtle semantic differences. Hard fine-grained negative samples learning effectively addresses challenge cases, thereby enhancing the model's stability and discriminative power.

5. Conclusion

In this work, we introduce Fine-Grained CLIP (FG-CLIP), a novel approach that significantly advances fine-grained understanding. By integrating advanced alignment techniques

with large-scale, high-quality datasets and hard negative samples, FG-CLIP captures global-level and region-level semantic details and distinguishes subtle differences more effectively. Extensive experiments across diverse downstream tasks validate the model's superior performance. In addition, we propose FineHARD as a unified dataset that combines high-quality region-specific annotations with challenging fine-grained negative samples, offering a valuable resource for advancing multimodal research. Looking ahead, exploring the integration of more sophisticated multimodal models and expanding dataset diversity will be crucial for pushing the boundaries of fine-grained understanding.

Impact Statement

This paper aims to advance the field of Machine Learning, which has broad implications for society. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, pp. 23716–23736, 2022.

Bianchi, L., Carrara, F., Messina, N., Gennaro, C., and Falchi, F. The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding. In *CVPR*, pp. 22520–22529, 2024.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pp. 3558–3568, 2021.

Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., and Lin, D. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, pp. 370–387, 2024a.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pp. 24185–24198, 2024b.

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., and Shan, Y. Yolo-world: Real-time open-vocabulary object detection. In *CVPR*, pp. 16901–16911, 2024.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Fu, L., Datta, G., Huang, H., Panitch, W. C.-H., Drake, J., Ortiz, J., Mukadam, M., Lambeta, M., Calandra, R., and Goldberg, K. A touch, vision, and language dataset for multimodal alignment. In *ICML*, pp. 14080–14101, 2024.

Gabeff, V., Rußwurm, M., Tuia, D., and Mathis, A. Wildclip: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. *IJCV*, pp. 1–17, 2024.

Gu, J., Meng, X., Lu, G., Hou, L., Minzhe, N., Liang, X., Yao, L., Huang, R., Zhang, W., Jiang, X., et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In *NeurIPS*, volume 35, pp. 26418–26431, 2022.

Gupta, A., Dollar, P., and Girshick, R. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pp. 5356–5364, 2019.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *ICCV*, pp. 2961–2969, 2017.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pp. 9729–9738, 2020.

Hong, W., Wang, W., Ding, M., Yu, W., Lv, Q., Wang, Y., Cheng, Y., Huang, S., Ji, J., Xue, Z., et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., et al. spacy: Industrial-strength natural language processing in python. 2020.

Hou, X., Liu, M., Zhang, S., Wei, P., and Chen, B. Salience detr: Enhancing detection transformer with hierarchical salience filtering refinement. In *CVPR*, pp. 17574–17583, 2024.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pp. 6700–6709, 2019.

Jing, D., He, X., Luo, Y., Fei, N., Yang, G., Wei, W., Zhao, H., and Lu, Z. Fineclip: Self-distilled region-based clip for better fine-grained understanding. In *NeurIPS*, 2024.

Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pp. 787–798, 2014.

Kim, D., Angelova, A., and Kuo, W. Contrastive feature masking open-vocabulary vision transformer. In *ICCV*, pp. 15602–15612, 2023a.

Kim, D., Angelova, A., and Kuo, W. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *CVPR*, pp. 11144–11154, 2023b.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981, 2020.

Laurençon, H., Marafioti, A., Sanh, V., and Tronchon, L. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pp. 19730–19742, 2023a.

Li, J., Xie, C., Wu, X., Wang, B., and Leng, D. What makes good open-vocabulary detector: A disassembling perspective. *arXiv preprint arXiv:2309.00227*, 2023b.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al. Grounded language-image pre-training. In *CVPR*, pp. 10965–10975, 2022.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023c. URL https://openreview.net/forum?id=xozJw0kZXF.

Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., and Bai, X. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*, 2024.

Lin, C., Sun, P., Jiang, Y., Luo, P., Qu, L., Haffari, G., Yuan, Z., and Cai, J. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., and Xie, W. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *MICCAI*, pp. 525–536, 2023.

Liu, C., Zhang, Y., Wang, H., Chen, W., Wang, F., Huang, Y., Shen, Y.-D., and Wang, L. Efficient token-guided image-text retrieval with consistent multimodal contrastive training. *IEEE Transactions on Image Processing*, 32:3622–3633, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, volume 36, pp. 34892–34916, 2023b.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, pp. 216–233, 2024.

Ma, C., Jiang, Y., Wu, J., Yuan, Z., and Qi, X. Groma: Localized visual tokenization for grounding multimodal large language models. In *ECCV*, pp. 417–435, 2024.

Minderer, M., Gritsenko, A., and Houlsby, N. Scaling open-vocabulary object detection. In *NeurIPS*, volume 36, 2024.

Mokady, R., Hertz, A., and Bermano, A. H. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Pan, J., Ma, Q., and Bai, C. A prior instruction representation framework for remote sensing image-text retrieval. In *ACM MM*, pp. 611–620, 2023.

Parelli, M., Delitzas, A., Hars, N., Vlassis, G., Anagnostidis, S., Bachmann, G., and Hofmann, T. Clip-guided vision-language pre-training for question answering in 3d scenes. In *CVPR*, pp. 5607–5612, 2023.

Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Ye, Q., and Wei, F. Kosmos-2: Grounding multimodal large language models to the world. In *ICLR*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *ICML*, pp. 5389–5400, 2019.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, volume 35, pp. 25278–25294, 2022.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pp. 2556–2565, 2018.

Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., and Wang, J. Alpha-clip: A clip model focusing on wherever you want. In *CVPR*, pp. 13019–13029, 2024.

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V., and Romero-Soriano, A. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *CVPR*, pp. 26700–26709, 2024.

Wang, B., Xie, C., Leng, D., and Yin, Y. Iaa: Inner-adaptor architecture empowers frozen large language model with multimodal capabilities. In *AAAI*, volume 39, pp. 21035–21043, 2025.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wu, S., Zhang, W., Xu, L., Jin, S., Li, X., Liu, W., and Loy, C. C. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In *ICLR*, 2024a. URL https://openreview.net/forum?id=DjzvJCRsVf.

Wu, W., Zheng, K., Ma, S., Lu, F., Guo, Y., Zhang, Y., Chen, W., Guo, Q., Shen, Y., and Zha, Z.-J. Lotlip: Improving language-image pre-training for long text understanding. *arXiv preprint arXiv:2410.05249*, 2024b.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024c.

Xie, C., Li, C., Zhang, B., Han, J., Zhen, X., and Chen, J. Memory attention networks for skeleton-based action recognition. In *IJCAI*, pp. 1639–1645, 2018.

Xie, C., Cai, H., Li, J., Kong, F., Wu, X., Song, J., Morimitsu, H., Yao, L., Wang, D., Zhang, X., et al. Ccmb: A large-scale chinese cross-modal benchmark. In *ACM MM*, pp. 4219–4227, 2023.

Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. 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.

Zareian, A., Rosa, K. D., Hu, D. H., and Chang, S.-F. Open-vocabulary object detection using captions. In *CVPR*, pp. 14393–14402, 2021.

Zhang, B., Zhang, P., Dong, X., Zang, Y., and Wang, J. Long-clip: Unlocking the long-text capability of clip. In *ECCV*, pp. 310–325, 2024.

Zheng, K., Zhang, Y., Wu, W., Lu, F., Ma, S., Jin, X., Chen, W., and Shen, Y. Dreamlip: Language-image pre-training with long captions. In *ECCV*, pp. 73–90, 2024.

Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pp. 16793–16803, 2022.

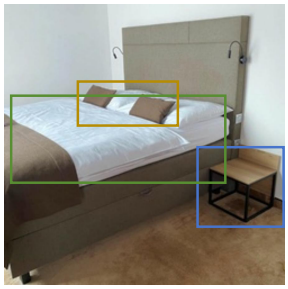
Zhou, C., Loy, C. C., and Dai, B. Extract free dense labels from clip. In *ECCV*, pp. 696–712, 2022a.

Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., and Misra, I. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, pp. 350–368, 2022b.

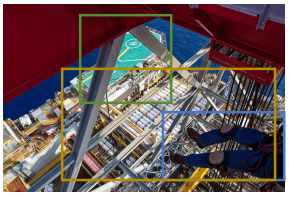
A. Examples of Curated Visual Grounding Data

Figure 2 shows visual grounding examples utilized in our experiments. Each example comprises an image, accompanied by its corresponding long and short captions, as well as multiple region-specific annotations, each with a detailed description.

---------long caption---------
The image shows a sewing project in progress. There is a sewing machine with its needle and presser foot engaged in the process. A piece of fabric with a pattern of pink shapes on an orange background is being sewn; the edges of the fabric are neatly folded over one another, and it appears to be pinned in place to hold its shape during the sewing. The sewing machine is positioned on a pink cutting mat that has a grid pattern, which is typically used to ensure accurate cutting when working with fabric.. . In the background, there is a white ceramic pot partially visible with a green plant that adds a touch of color to the setting. The lighting seems to be indoor, likely coming from overhead, as there are no shadows visible on the objects, and the overall lighting appears even and consistent.. . The style of the image is a close-up photograph, capturing the details of the sewing process with high clarity. There are no people or characters in the image, so there are no emotions to convey. The image is factual and does not contain any subjective interpretation of emotions or style.
---------short caption---------
sewing machine with a semi-finished pink baby hat
---------region caption---------
a pink cutting mat that has a grid pattern which is typically used to ensure accurate cutting when working with fabric
---------region caption---------
a white ceramic pot partially visible with a green plant that adds a touch of color to the setting
---------region caption---------
a baby hat with white spots to be finished



---------long caption---------
The image presents a bedroom scene centered on a large bed with a tall, upholstered headboard. The headboard is light brown and has a smooth texture without any visible patterns. Two reading lights are mounted on the wall on each side of the headboard, extending outward with a flexible arm and a small light at the end. These lights are positioned at a height that seems suitable for reading. The bed frame is also light brown and has a contemporary design.. . On the bed, there is a white duvet and a couple of pillows, with one of them being a smaller pillow that matches the color of the reading lights, and the other being a larger pillow that appears to have a different color, possibly a neutral tone. At the foot of the bed, there is a small wooden side table with a black metal frame, which has a minimalistic design with a flat surface and a single visible drawer. The table's color palette includes brown and black, complementing the overall room tones.. . The lighting within the image is not discernible due to the lack of shadows or highlights that could indicate a specific light source. Instead, the room appears to be lit by ambient light that creates a soft glow, giving the space a cozy atmosphere.. . The style of the image is a photograph. The edges and shadows suggest a natural light source and the clarity of the objects points towards a high-resolution image captured by a camera.
---------short caption---------
Modern bedside table in a simple style in the hotel room – metal furniture
---------region caption---------
a small wooden side table with a black metal frame which has a minimalistic design with a flat surface and a single visible drawer
---------region caption---------
a white duvet and a couple of pillows with one of them
---------region caption---------
a smaller pillow that matches the color of the reading lights and a larger that appears to have a different color

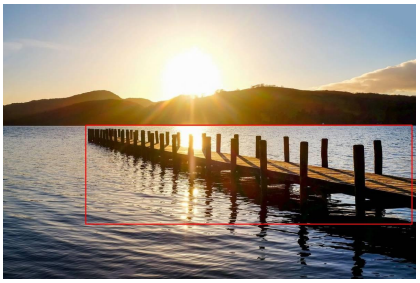


---------long caption---------
This image depicts an aerial view of an oil rig's platform. In the foreground, there is a worker engaged in an activity on the edge of the platform, wearing safety gear such as jeans, a yellow hard hat, knee pads, and brown work boots. The worker appears to be suspended or working at height, secured by a safety harness and other equipment. The rig's infrastructure is comprised of metal beams, pipes, and various work platforms.. . The background reveals a large body of water, which could be the ocean. On the water's surface, there is a green helipad with a circular symbol, suggesting that this is a designated area for landing helicopters. The helipad is connected to the rig by a walkway. The weather appears to be clear, with bright sunlight casting shadows on the structure of the rig. The image is taken from a high vantage point, looking downwards towards the worker and the rig's platform, highlighting the scale and complexity of the offshore oil operation.. . The style of the image is a high-resolution photograph, capturing the intricate details of the industrial setting. There are no people or characters other than the worker, so there are no emotions to convey. The description is factual and does not include any subjective interpretations or personal opinions about the image.
---------short caption---------
An overhead view of the legs and harness of an industrial painter suspended from the underside of an offshore rig derrick, high above the deck below.
---------region caption---------
a worker engaged in an activity on the edge of the platform wearing safety gear such as jeans a yellow hard hat knee pads and brown work boots
---------region caption---------
a circular symbol suggesting that this is a designated area for landing helicopters
---------region caption---------
a high vantage point looking downwards towards the worker and the rig's platform highlighting the scale and complexity of the offshore oil operation

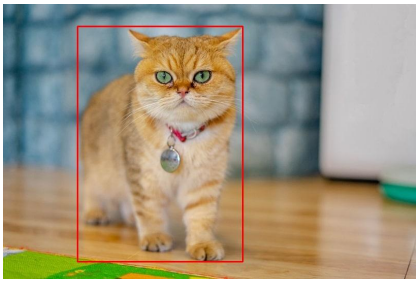
Figure 2. Examples of curated visual grounding data.

B. Positive and Negative Descriptions Related to Image Regions

To generate hard fine-grained negative samples, we modify the attributes of bounding box descriptions while keeping object names unchanged. Figure 3 illustrates examples of positive and corresponding negative descriptions for image regions.

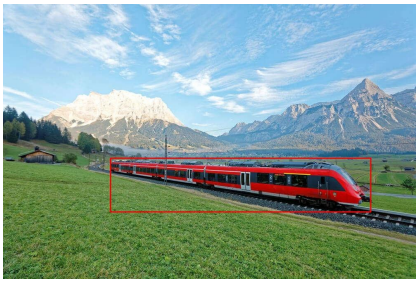


---------positive---------
very long wooden jetty, jutting out into a calm blue wooden lake with mountains sunsetting in background
---------negatives---------
'0': 'very long iron jetty, jutting out into a calm pink wooden lake with mountains sunsetting in background'
'1': 'very long silver jetty, jutting out into a calm turquoise wooden lake with mountains sunsetting in background'
'2': 'very long brass jetty, jutting out into a calm lavender wooden lake with mountains sunsetting in background'
'3': 'very long golden jetty, jutting out into a calm peach wooden lake with mountains sunsetting in background'
'4': 'very long bronze jetty, jutting out into a calm mint wooden lake with mountains sunsetting in background'
'5': 'very long copper jetty, jutting out into a calm coral wooden lake with mountains sunsetting in background'
'6': 'very long steel jetty, jutting out into a calm magenta wooden lake with mountains sunsetting in background'
'7': 'very long chrome jetty, jutting out into a calm charcoal wooden lake with mountains sunsetting in background'
'8': 'very long titanium jetty, jutting out into a calm plum wooden lake with mountains sunsetting in background'
'9': 'very long aluminum jetty, jutting out into a calm amber wooden lake with mountains sunsetting in background'

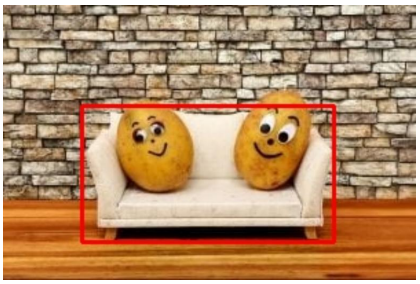


---------positive---------
british shorthair with a red collar
---------negatives---------
'0': 'british shorthair with a green collar'
'1': 'british shorthair with a copper collar'
'2': 'british shorthair with a grass green collar'
'3': 'british shorthair with a deep blue collar'
'4': 'british shorthair with a white collar'
'5': 'british shorthair with a black collar'
'6': 'british shorthair with a light blue collar'
'7': 'british shorthair with a silver collar'
'8': 'british shorthair with a dark blue collar'
'9': 'british shorthair with a yellow collar'

---------positive---------
A two-box fastback configuration combined with cleverly concealed air intakes and lightweight roof assembly
---------negatives---------
'0': 'A brass riveted configuration combined with cleverly concealed air intakes and copper roof assembly'
'1': 'A copper bolted configuration combined with cleverly concealed air intakes and copper roof assembly'
'2': 'A metallic braided configuration combined with cleverly concealed air intakes and metallic roof assembly'
'3': 'A woven mesh configuration combined with cleverly concealed air intakes and woven roof assembly'
'4': 'A translucent grid configuration combined with cleverly concealed air intakes and translucent roof assembly',
'5': 'A diamond-plate configuration combined with cleverly concealed air intakes and diamond-plate roof assembly'
'6': 'A chrome-plated configuration combined with cleverly concealed air intakes and chrome-plated roof assembly'
'7': 'A striped metal configuration combined with cleverly concealed air intakes and striped roof assembly',
'8': 'A glossy ceramic configuration combined with cleverly concealed air intakes and glossy roof assembly'
'9': 'A beaded mesh configuration combined with cleverly concealed air intakes and beaded roof assembly'



---------positive---------
a red train in Swiss alps
---------negatives---------
'0': 'a light blue train in Swiss alps'
'1': 'a grass green train in Swiss alps'
'2': 'a copper train in Swiss alps'
'3': 'a deep blue train in Swiss alps'
'4': 'a white train in Swiss alps'
'5': 'a black train in Swiss alps'
'6': 'a turquoise train in Swiss alps'
'7': 'a brown train in Swiss alps'
'8': 'a silver train in Swiss alps'
'9': 'a grey train in Swiss alps'



---------positive---------
a beige couch with a brick background
---------negatives---------
'0': 'A turquoise couch with a silver background'
'1': 'A copper couch with a silver background'
'2': 'A turquoise couch with a copper background'
'3': 'A copper couch with a turquoise background'
'4': 'A bronze couch with a silver background'
'5': 'A bronze couch with a copper background'
'6': 'A copper couch with a bronze background'
'7': 'A turquoise couch with a bronze background'
'8': 'A silver couch with a turquoise background'
'9': 'A silver couch with a copper background'

Figure 3. Examples of positive and negative descriptions related to image regions.

C. Visualization Comparison

As illustrated in Figure 4, we present a comparison of similarity matrix visualizations for different methods using challenging sample images. We utilize the dense image feature extraction strategy introduced by (Zhou et al., 2022a). In the figure, warmer colors (e.g., yellow) denote higher similarity, whereas cooler colors (e.g., blue) indicate lower relevance. Our goal is for the model to precisely comprehend and interpret the fine-grained details within the images.

In the first scenario, the image contains three dogs of different colors, and we compute the similarity matrix using only the phrase "Black dog" with each image token. It can be observed that CLIP and EVA-CLIP fail to accurately identify the target dog, FineCLIP captures some relevant tokens but not all, whereas FG-CLIP identifies a larger number of relevant tokens, demonstrating superior performance.

In the second scenario, the image contains multiple black entities, and we compute the similarity matrix using only the phrase "Black nose", which occupies a very small area within the image. CLIP fails to identify the target, while EVA-CLIP and FineCLIP locate the target but also respond to many other black regions. In contrast, FG-CLIP accurately identifies the target, showcasing its precision in fine-grained localization.

In the third scenario, the image features gemstones of three different colors, and we compute the similarity matrix using only the phrase "Red gemstone". Both CLIP and EVA-CLIP fail to locate the target at the bottom and exhibit high responses to gemstones of other colors. FineCLIP shows slightly lower localization accuracy compared to FG-CLIP, which precisely distinguishes between the colors of different gemstones and achieves more accurate localization.

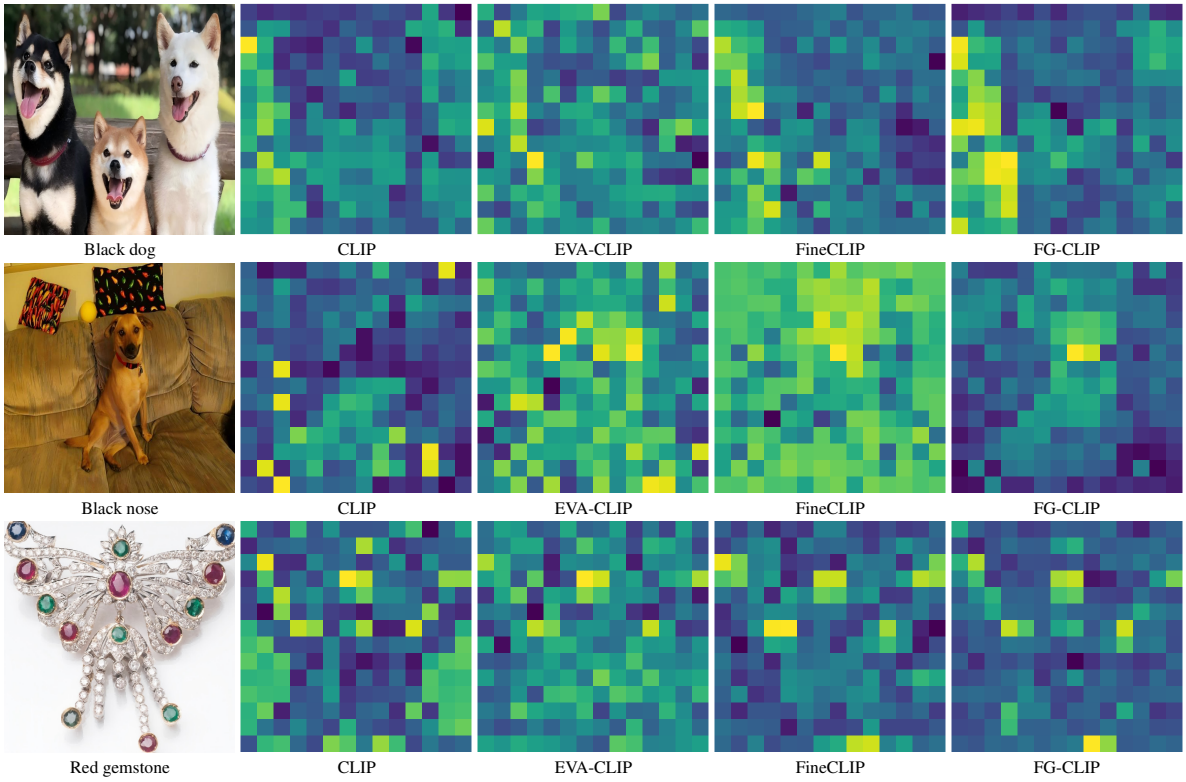


Figure 4. Feature visual comparisons of different methods.

Additionally, we utilize FG-CLIP to conduct a correlation analysis between different input texts and the same image. The results in Figure 5 indicate that FG-CLIP provides precise positional understanding of different targets within the image. This demonstrates the model's stable visual localization capabilities and its fine-grained understanding of image content.

To evaluate the impact of hard fine-grained negative samples learning, we further provide the qualitative results in Figure 6. After performing hard negative sampling, our FG-CLIP can capture the regions more accurately. For example, the highlighted region of "Man in red clothes" with hard negative loss in 1st row shows significantly better than that without hard negative loss.

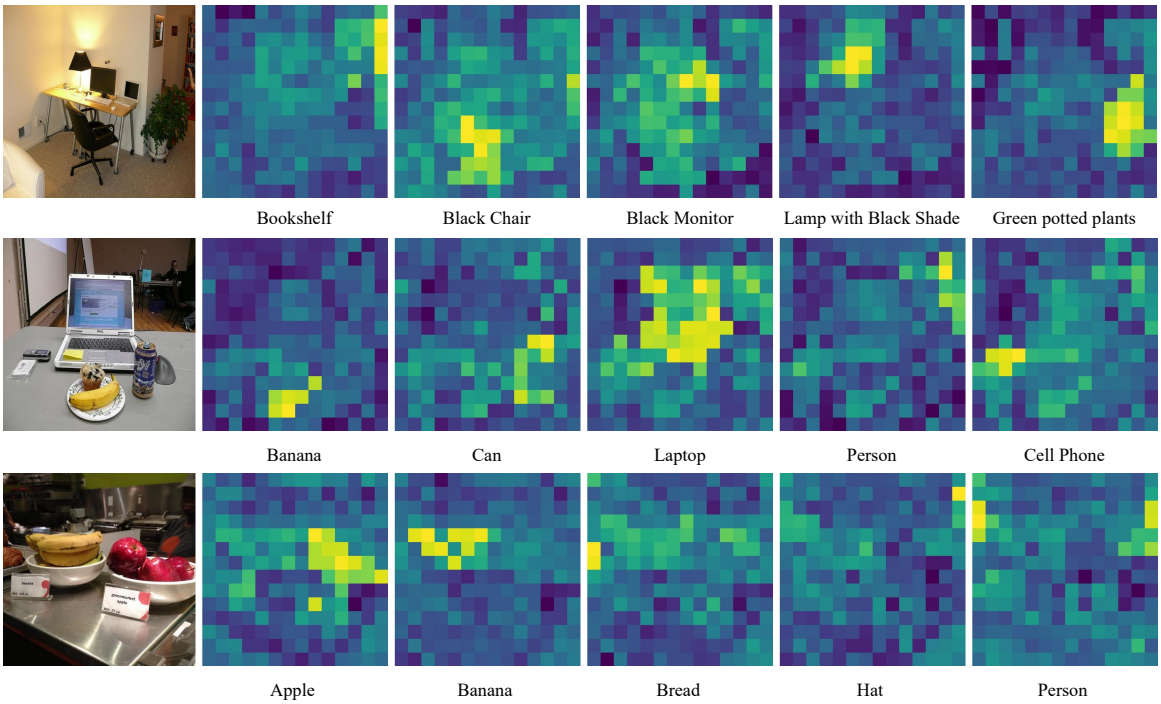


Figure 5. Feature visual comparisons of different input texts.

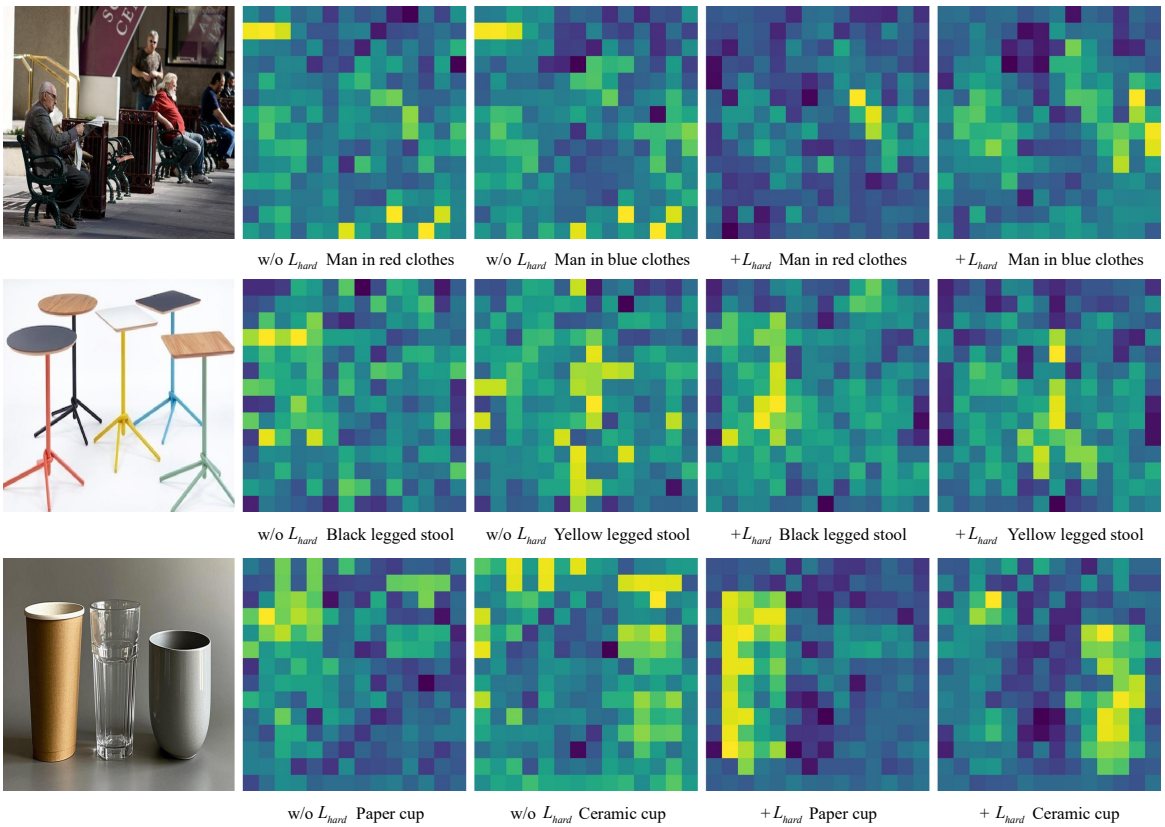
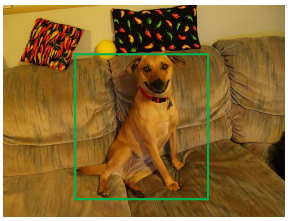
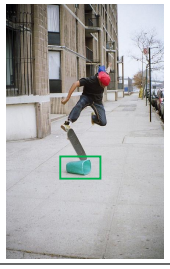


Figure 6. Performance of hard fine-grained negative samples learning.

D. Further Experiments

Table 7. Comparisons of different methods on fine-grained benchmark.

Image with region	Positive and Negative Region Descriptions	CLIP	EVA-CLIP	FineCLIP	FG-CLIP
	Origin: A table made of dark brown wood.	0.73	0.79	0.62	1.0
	1: A table made of pink wood.	0.0	0.59	0.58	0.0
	2: A table made of dark brown paper.	0.48	0.34	0.48	0.48
	3: A table made of light grey wood.	0.27	0.0	0.39	0.55
	4: A table made of dark brown wool.	0.27	0.11	0.03	0.21
	5: A table made of yellow wood.	0.38	0.44	0.38	0.14
	6: A table made of dark brown velvet.	1.0	0.63	0.0	0.01
	7: A table made of dark brown text.	0.94	0.72	0.38	0.55
	8: A table made of grey wood.	0.53	0.47	1.0	0.61
	9: A table made of dark brown plastic.	0.45	1.0	0.86	0.26
	10: A table made of green wood.	0.47	0.16	0.64	0.46
	Origin: A brown leather handbag.	0.62	0.79	0.78	1.0
	1: A brown metal handbag.	0.45	0.88	0.56	0.37
	2: A brown text handbag.	0.53	1.0	0.90	0.50
	3: A brown wool handbag.	0.0	0.0	0.63	0.13
	4: A orange leather handbag.	0.32	0.74	0.34	0.53
	5: A brown paper handbag.	0.41	0.87	1.0	0.60
	6: A light orange leather handbag.	0.08	0.72	0.08	0.56
	7: A brown glass handbag.	1.0	0.59	0.75	0.0
	8: A dark red leather handbag.	0.43	0.59	0.15	0.93
	9: A dark yellow leather handbag.	0.31	0.82	0.11	0.73
	10: A purple leather handbag.	0.21	0.91	0.0	0.27
	Origin: A brown dog with black nose.	0.75	0.0	0.76	1.0
	1: A brown dog with light red nose.	0.37	0.45	0.64	0.80
	2: A brown dog with dark yellow nose.	0.10	0.39	0.93	0.90
	3: A light blue dog with black nose.	0.40	0.39	0.0	0.0
	4: A brown dog with yellow nose.	0.38	0.40	1.0	0.86
	5: A red dog with black nose.	1.0	1.0	0.59	0.32
	6: A brown dog with light orange nose.	0.16	0.67	0.71	0.85
	7: A brown dog with light yellow nose.	0.0	0.58	0.83	0.83
	8: A brown dog with light blue nose.	0.26	0.44	0.48	0.52
	9: A dark green dog with black nose.	0.88	0.32	0.41	0.13
	10: A brown dog with dark purple nose.	0.41	0.12	0.48	0.73
	Origin: A light blue plastic trash can.	0.89	0.95	0.52	1.0
	1: A light blue stone trash can.	0.90	1.0	1.0	0.77
	2: A dark purple plastic trash can.	0.60	0.70	0.03	0.64
	3: A light blue wool trash can.	0.68	0.57	0.37	0.29
	4: A dark green plastic trash can.	0.92	0.74	0.99	0.65
	5: A dark orange plastic trash can.	0.58	0.35	0.0	0.87
	6: A black plastic trash can.	0.66	0.80	0.65	0.77
	7: A purple plastic trash can.	0.68	0.90	0.36	0.75
	8: A light blue crochet trash can.	0.0	0.0	0.10	0.0
	9: A light blue glass trash can.	1.0	0.72	0.52	0.64
	10: A light orange plastic trash can.	0.68	0.77	0.02	0.88
	Origin: A red plastic bucket.	0.97	0.53	0.70	1.0
	1: A green plastic bucket.	0.94	0.77	0.46	0.68
	2: A red metal bucket.	0.83	0.29	0.54	0.61
	3: A red crochet bucket.	0.0	0.10	0.10	0.39
	4: A red ceramic bucket.	0.58	0.06	0.15	0.58
	5: A red fabric bucket.	0.55	0.05	0.0	0.43
	6: A red stone bucket.	1.0	0.10	0.84	0.48
	7: A red rattan bucket.	0.47	0.51	0.58	0.0
	8: A red wool bucket.	0.39	0.0	0.14	0.31
	9: A yellow plastic bucket.	0.77	1.0	1.0	0.67
	10: A light green plastic bucket.	0.70	0.87	0.45	0.28

D.1. Comparison of Different Methods on Fine-Grained Benchmark

As shown in Table 7, we select several samples from the test set of FG-OVD (Bianchi et al., 2024) and visualize the comparison results of different methods. We employ the testing strategy detailed in Section 4.2 and match the text with localized dense feature. The similarity scores computed between regions and texts are normalized, where the sentence with the lowest similarity is assigned a score of 0.0, and the sentence with the highest similarity is assigned a score of 1.0.

FG-CLIP demonstrates strong capability in identifying these extremely difficult samples, whereas other methods struggle to achieve comparable performance.

D.2. Performance Comparison on Identical Datasets

Table 8. Comparisons of different methods on the same dataset.

Method	Data Source	COCO-Box-Top-1	COCO-Retrieval-I2T	COCO-Retrieval-T2I
FineCLIP	FineCLIP (CC2.5M)	50.7	54.4	40.2
FineCLIP	FG-CLIP (12M)	53.5	59.6	46.2
FG-CLIP (Ours)	FG-CLIP (12M)	56.1	65.9	47.1

To evaluate the effectiveness of our proposed FG-CLIP method, we conduct experiments on the same dataset to ensure a fair comparison. Specifically, we compare FineCLIP and FG-CLIP using the 12M dataset due to time constraints, instead of the larger 1.6B+12M setup. From Table 8, the substantial improvements (Row 1 -> Row 2 and Row 2 -> Row 3) highlight that both our proposed dataset and model architecture are significant for FG-CLIP.

D.3. Performance on General Multimodal Benchmarks

Table 9. Comparisons on General Multimodal Benchmarks.

Method	GQA	POPE	RefCOCO			MMBench-EN		MMBench-CN	
			val	testA	testB	dev	test	dev	test
LLaVA-v1.5+CLIP	61.9	85.9	76.2	83.4	67.9	65.1	66.5	58.2	58.4
	+1.0	*+0.9*	*+5.2*	*+3.1*	*+7.0*	*+1.5*	*+0.2*	*+0.6*	*+0.9*
LLaVA-v1.5+FG-CLIP	62.9	86.8	81.4	86.5	74.9	66.6	66.7	58.8	59.3

In addition to GQA, POPE, and RefCOCO, we conduct experiments on other general multimodal benchmarks (Liu et al., 2024). The experimental results in Table 9 show that LLaVA with FG-CLIP achieves better performance.