Grounding Descriptions in Images informs Zero-Shot Visual Recognition

Shaunak Halbe[1], Junjiao Tian[1], Joseph KJ[2], James Smith[1], Katherine Stevo[1], Vineeth Balasubramanian[3], and Zsolt Kira[1]

[1] Georgia Institute of Technology
[2] Adobe Research
[3] IIT Hyderabad
Corresponding Author: shalbe9@gatech.edu

Abstract. Vision-language models (VLMs) like CLIP have been cherished for their ability to perform zero-shot visual recognition on open-vocabulary concepts. This is achieved by selecting the object category whose textual representation bears the highest similarity with the query image. While successful in some domains, this method struggles with identifying fine-grained entities as well as generalizing to unseen concepts that are not captured by the training distribution. Recent works attempt to mitigate these challenges by integrating category descriptions at test time, albeit yielding only modest improvements. We attribute these limited gains to a fundamental misalignment between image and description representations which is rooted in the pretraining structure of CLIP. In this paper, we propose *GRAIN*, a new pretraining strategy aimed at aligning representations at both fine and coarse levels simultaneously. Our approach learns to jointly ground textual descriptions in image regions along with aligning overarching captions with global image representations. To drive this pre-training, we leverage frozen Multimodal Large Language Models (MLLMs) to derive synthetic annotations that are more fine-grained. We demonstrate the enhanced zero-shot performance of our model compared to current state-of-the art methods across 11 diverse image classification datasets. Additionally, our model is effective in recognizing new, unseen concepts and distinguishing between similar, fine-grained entities. Significant improvements achieved by our model on other downstream tasks such as retrieval further highlights the superior quality of representations learned by our approach.

Keywords: Zero-shot Learning · Open-vocabulary Visual Understanding · Vision-language Pretraining · Fine-grained Visual Recognition

1 Introduction

Traditionally, image classification has operated under the closed-set assumption where models are evaluated on a fixed set of classes that were seen during training. However, in the real and open-world, models need to account for test conditions where the number of classes is unknown during training and can include classes that were not seen. Vision-language models (VLMs) like CLIP [36]

offer a solution in this space, owing to their *open-vocabulary* nature. These models undergo extensive pretraining on large datasets containing paired image-text data and learn to encode images and texts in a shared latent space where semantically similar representations are mapped closed together. For zero-shot classification, CLIP leverages the names of all classes within the test dataset—referred to as the *vocabulary*—as the candidate set, and determines the most probable image-classname pairing by computing the similarity between their latent representations. This vocabulary of classes is unconstrained, enabling the inclusion of any concept, regardless of its presence in the training set. This facilitates classification from an *open-set* of concepts.

Despite this, CLIP's zero-shot capabilities are still limited by a few critical challenges. Firstly, in practice CLIP often struggles to differentiate between fine-grained categories, a limitation highlighted by its under-performance on Fine-Grained Visual Classification (FGVC) datasets [25, 44]. Secondly, while known for its open-vocabulary potential, it can still perform poorly for some domains not well-reprented in the training distribution, especially if the vocabulary used has confounding categories during testing. Using a vocabulary that exceeds the scope of the test dataset significantly diminishes the performance of CLIP even for common datasets like Imagenet [14]. This decline is again largely attributed to CLIP's challenges in differentiating between semantically similar, fine-grained concepts. Additionally, CLIP's inability to recognize novel concepts, such as `Apple Vision Pro` that were not present during its training phase, further restricts its capability to function as a genuinely open-vocabulary model.

Recent works [27, 35] aim to address these challenges by incorporating extra information in the form of class descriptions generated by Large Language Models (LLMs) at test time. These approaches leverage the "visual" knowledge embedded in LLMs to augment the textual representations used in zero-shot classification. As an example, the class `French Bulldog` would be expanded to `A French Bulldog, which has small, pointy ears.` However, in practice, these methods yield only modest improvements over the standard CLIP model across the majority of datasets.

We hypothesize, that the reason for these challenges lies in the poor alignment between image and description embeddings learned by CLIP. As a result, we aim to verify this hypothesis and propose a method to overcome these challenges. Specifically, we posit that the misalignment between images and descriptions stems from CLIP's training structure, which focuses solely on the global objective of matching entire images to their alt-texts (captions), neglecting the rich information that image regions and textual descriptions share with each other. Our observations align with recent research indicating that CLIP tends to overlook fine-grained visual details during pretraining, leading to subpar performance on tasks requiring localization [38], object attributes [51], or physical reasoning [33].

In this work, we propose *GRAIN: Grounding and contrastive alignment of descriptions*, a novel objective for contrastive vision-language pretraining that learns representations more conducive to zero-shot visual recognition. This is achieved through fine-grained correspondence between image region and detailed

text descriptions. As a first step towards our approach, given that pretraining datasets (Conceptual Captions [42], LAION [41], etc.) only contain images with noisy captions but without detailed descriptions, we employ an instruction-tuned Multimodal Large Language Model (MLLM) to generate descriptions or identify salient attributes from the images in these datasets. Following this, we acquire region-level annotations that correspond to these descriptions using an off-the-shelf Open-vocabulary Object Detector (OVD). For our method, we learn to jointly ground text descriptions into specific image regions and the contrastive alignment between images and captions at the global level. This strategy aims to learn representations that encode both coarse-grained (global) and fine-grained (local) information. To achieve this, we introduce a query transformer architecture for encoding images and a text encoder for processing captions and descriptions. The architecture and objectives of our model are specifically crafted to learn object/region-aware image representations that are valuable for zero-shot tasks as we demonstrate in the subsequent sections.

To summarize, our main contributions are as follows:

- We hypothesize and show that CLIP pre-training lacks fine-grained aligned representations, leading to poor zero-shot performance in some domains.
- We propose *GRAIN*, a novel pre-training architecture and losses that simultaneously learns local and global correspondences, obtained via weak supervision from Multimodal LLMs and open-vocabulary detectors.
- We demonstrate significantly improved performance across a range of tasks, including image classification and retrieval, and specifically improve over the state-of-art by up to **9%** in absolute top-1 accuracy for zero-shot classification and up to **25%** across cross-modal retrieval tasks.

2 Related Works

Improving CLIP using Generative Models. Recent works have explored the use of LLMs towards improving the downstream performance of CLIP. Menon *et al*. [27] and CuPL [35] focus on the task of zero-shot classification, and prompt GPT-3 [2] at test-time to generate class descriptions. These descriptions are integrated into the classification prompts to achieve gains in terms of accuracy and interpretability. LaCLIP [12] uses a variety of LLMs to rephrase captions from pretraining datasets and observe noticeable gains on downstream tasks by training on these captions. In this paper, we propose to leverage synthetic annotations generated by a MLLM to drive a pretraining strategy that aligns rich textual information to the local image regions.

Multimodal Large Language Models. MLLMs like LLaVA [22], GPT-4V [32], Mini-GPT4 [54] integrate image tokens into LLMs, leveraging their powerful reasoning capabilities. MLLMs has been found useful in tasks such as scene understanding [39], story-telling [10] etc., where a comprehensive understanding of the images and text is required. We leverage their ability for visual comprehension to generate a set of descriptions for an input image that are used to supervise our fine-grained losses during training.

Zero-shot Learning for Images. Zero-shot learning (ZSL) learning is a challenging problem that requires methods to recognize object categories not seen during training. Various approaches [18,34] have proposed using side information like attributes, hierarchical representations etc. to learn a generalizable mapping. More recent efforts [47, 49] explore the use of generative models to synthesize useful features for unseen categories. Our method aligns more closely with the former, as we learn a fine-grained correspondence conducive for zero-shot classification by leveraging descriptions as side-information.

3 Approach

We propose GRAIN, a novel pretraining approach that simultaneously learns local and global correspondences between image and text representations. Motivated by the observation that CLIP representations lack sufficient fine-grained understanding, we introduce a transformer-based architecture inspired by DETR [4] to infuse the rich context from sub-image regions into learned visual representations. Alongside encoding the image, our model predicts bounding boxes for salient image regions containing discriminative information. These localizations are then aligned semantically with detailed textual descriptions. To supervise this fine-grained objective, we first generate annotations at scale by leveraging Multimodal Large Language Models (MLLMs) and Open-vocabulary Object Detectors (OVDs). In this section, we first elaborate our automated annotation process and then proceed to discussing our architecture and training methodology.

3.1 Weak Supervision from MLLMs and OVDs

We utilize the 3M and 12M versions of the Conceptual Captions [42] (CC3M, CC12M) dataset to train our model. These datasets contain images sourced from the internet, each paired with corresponding alt-texts (or captions). In order to execute our approach, we require fine-grained supervision that is not provided by any existing dataset at scale. Specifically, we find that the captions associated with these images are often noisy, lack detail and may not fully capture the dense visual context. To learn fine-grained correspondence between the two modalities, we propose using rich text descriptions associated with regions within the image as supervision for training our model. For generating these detailed descriptions, we leverage an instruction-tuned Multimodal Large Language Model, LLaVA [22]. LLaVA has been fine-tuned on a visual instruction-tuning dataset containing 150K samples, positioning it as an agent capable of following human instructions. For our annotation purposes, we select the LLaVA v1.6 model which integrates a pretrained Vision Transformer Large (ViT-L) [11] as the visual encoder with the Vicuna-13B LLM [7]. We find that LLaVA is capable of understanding visual scenes effectively, and we rely on LLaVA's ability to provide us with detailed descriptions for these images. We only use LLaVA to describe components of the image at a high level and not pinpoint specific fine-grained categories. A common problem with instruction-tuned models like LLaVA is

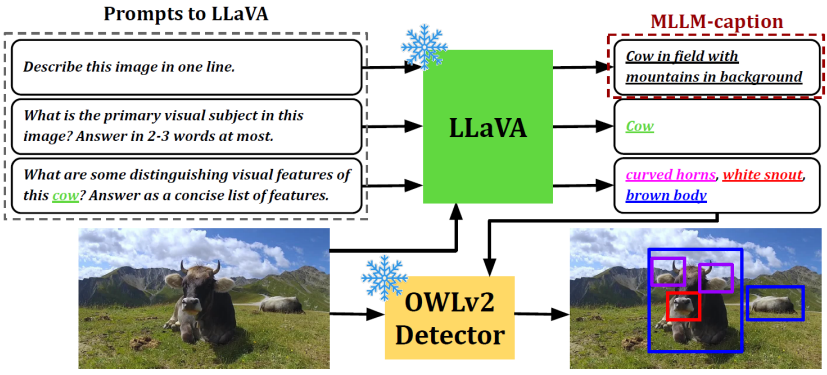


Fig. 1: Overview of our two-stage annotation process: (1) prompting LLaVA for image descriptions and (2) acquiring corresponding region annotations from OWLv2.

their tendency to hallucinate, which causes the model to output sentences that are not well-grounded in the image. To address this, we propose a two-stage approach, as illustrated in Figure 1, to elicit accurate descriptions from LLaVA while minimizing hallucination.

Specifically, the two-stage prompting approach is as follows: in the first stage, we ask LLaVA to identify the primary visual subject in the image using a simple, fixed prompt: "*What is the primary visual subject in this image? Answer in 2-3 words at most.*" By doing this for every image, we collect the main focus of each image. The generations from this prompt typically capture the prominent object, scene, or concept at a high level. Next, we construct specific prompts for each image by asking LLaVA to describe the identified subject: "*What are some distinguishing visual features of this {subject}? Answer as a concise list of features*". We observe that the generations from this two-stage pipeline are more faithful to the visual context and less susceptible to hallucinations. This procedure provides us with a list of descriptions for each image. Additionally, we ask LLaVA to generate a short one-line description-oriented caption for the image by prompting it with "*Describe this image in one line*". This caption gives a high-level overview of the visual context, and it is utilized as text-level data augmentation during training. From this point forward, we refer to this as the MLLM-caption, and the alt-text from the dataset as the original caption.

Next, we are tasked with localizing these generated descriptions within the image to obtain the necessary supervision for training our grounding module. We leverage the OWLv2 Open-vocabulary Detector [28] to localize these descriptions within the image. For each description, we filter out the core attribute being referred to and pass it to the open-world detector for localization. The detector generates several candidate proposals, from which we select detections based on a confidence threshold value. We set this threshold to a relatively high value to ensure high-quality detections. After that, we remove redundant bounding box predictions by performing non-maximum suppression, discarding all boxes

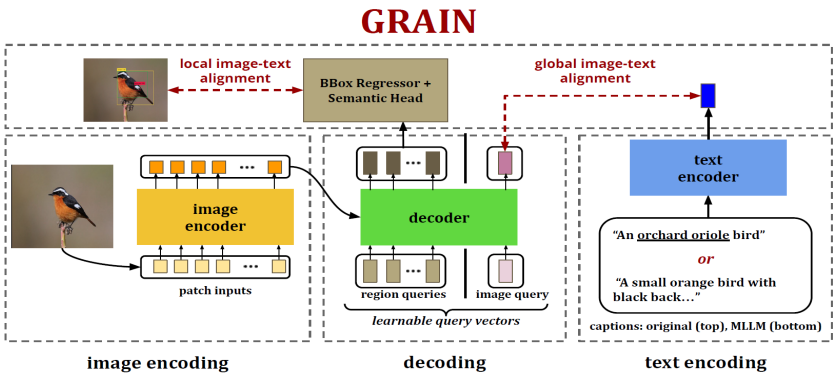


Fig. 2: Architecture overview. Our method, GRAIN, aligns image representations to text captions at a global level while localizing salient image regions and aligning them to text descriptions at the local level.

sharing a high-overlap with confidence scores lower than the maximum for that object.

This procedure enables us to acquire descriptions, bounding boxes, and MLLM captions, which are subsequently utilized to train our model, as detailed in the upcoming section. To our knowledge, we are the first to obtain such fine-grained annotations on a large scale. The overall annotation process took around 600 GPU hours for CC3M and $\sim$ 2200 GPU hours for CC12M using NVIDIA A40s.

3.2 Model Architecture

We adopt a dual-encoding approach similar to CLIP for processing image and text modalities, leveraging contrastive learning to align these representations. For visual representations, we utilize an encoder-decoder network architecture. Notably, all components of our architecture are trained from scratch without any pretrained initialization. In our vision encoder, we adopt a standard vision transformer (ViT) that divides the input image into $\frac{HW}{P^2}$ patches where (H, W) is the input image resolution and P denotes the patch size. The output tokens corresponding to each input patch are fed into our transformer decoder as shown in Figure 2. Both text descriptions and captions are processed by a text transformer which utilizes the same architecture employed in CLIP.

Transformer Decoder. Inspired by DETR [4], we implement a transformer decoder that takes as input a small number of learnable position embeddings called queries and attends to the encoder output. We use two types of queries as input to this model. First we have n_q number of queries that we call region queries, whose corresponding outputs are used to predict bounding boxes. Additionally, we use a single image query to learn the overall image context. The transformer model transforms these input queries through self-attention between region and image queries and cross-attention with the encoder output to form

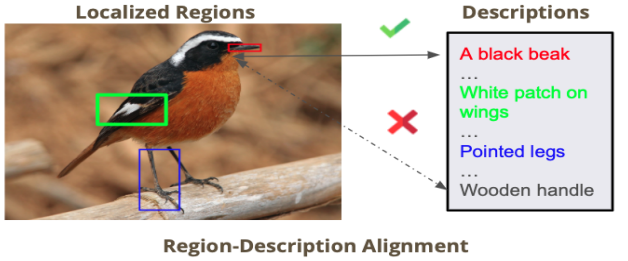


Fig. 3: Region output embeddings are used to predict bounding boxes and align them to textual descriptions via contrastive learning.

output embeddings. The embeddings corresponding to the region queries are utilized for bounding box prediction and serve as semantic representations for local regions, while the embedding corresponding to the image query captures the overall image representation needed for contrastive learning alongside captions. This image query output is passed through a projection layer before contrastive alignment with the text captions. The bounding box prediction module is exclusively used during training to learn region-aware image features and are inactive during evaluation.

Bounding-Box Prediction. The region output embeddings are fed into a multi-layer perceptron for bounding box prediction. The input size of this MLP is equal to the embedding dimension d and the output size is set to 4, corresponding to the four bounding box coordinates. These MLP weights are shared across all queries.

Semantic Representations. Each region output embedding is additionally passed through a projection layer to map it into the shared semantic space. The resulting semantic representations are utilized for contrastive alignment with text descriptions. This region-description alignment procedure is illustrated in Figure 3.

3.3 Training Objectives

Our approach simultaneously optimizes for three objectives: localizing salient regions within the image, contrastively aligning text descriptions to these salient image region representations, and globally aligning images with captions.

Image-Caption Alignment ($\mathcal{L}_{ic}$). We adopt the symmetric cross entropy loss from CLIP to maximize the similarity between correct image-caption pairings while contrasting against incorrect pairings within the batch. As with CLIP, we use the [EOS] token from the last layer of the text transformer and the output embedding corresponding to the image query as feature representations for $\mathcal{L}_{ic}$.

Bounding Box Loss ($\mathcal{L}_{box}$). Our model predicts n_q bounding boxes per image corresponding to the region queries. n_q is set to be greater than or equal

to the maximum number of objects per image in the training set. Given the variable number of objects per image, we employ the Hungarian Matching algorithm to establish a bipartite matching between predicted and ground truth boxes. For the matched boxes, we implement the bounding box loss derived from DETR, which combines the scale-invariant IOU loss and the L1 loss between the bounding box coordinates. Overall, the bounding box $\mathcal{L}_{box}(b_i, \hat{b}_{\sigma(i)})$ is defined as $\mathcal{L}_{iou}(b_i, \hat{b}_{\sigma(i)}) + \|b_i - \hat{b}_{\sigma(i)}\|_1$.

Region-Description Alignment ($\mathcal{L}_{rd}$). We use an InfoNCE loss [31] to learn alignment between output region embeddings and descriptions. Here, the descriptions corresponding to ground bounding boxes serve as supervision. We leverage the matched indices between predicted outputs and ground truth boxes obtained via the Hungarian Matching algorithm in the last step to determine ground-truth descriptions for each predicted region output embedding. These matched ground truths are considered positive pairings, while all other pairings within the batch are treated as negatives for InfoNCE. Optimizing for this loss enables our model to learn fine-grained associations between rich textual descriptions and salient image regions that contain discriminative visual features.

Overall, the final objective function is an equally weighted combination of three components.

$$\mathcal{L}_{total} = \mathcal{L}_{ic} + \mathcal{L}_{box} + \mathcal{L}_{rd} \tag{1}$$

3.4 Inference

At inference time, our model behaves similar to CLIP, conducting zero-shot classification by computing image-text similarities. The image output embedding from our decoder serves as the feature representation for the image. Through self and cross-attention mechanisms, this feature is informed about the fine-grained regions that are characteristic of the given image. The localization modules are inactive during testing; however, they can be used to provide valuable insights for interpreting the model's predictions.

4 Experiments

The goal of our method is to learn fine-grained vision-language representations aimed at improving zero-shot visual recognition tasks. By recognizing and addressing the alignment discrepancy between CLIP's representations of image regions and the rich textual context, our method learns visual representations that are aware of the salient regions in the image and their associations with corresponding textual descriptions. Although the focus of our method is on visual recognition, we observe that our learned representations are of high quality and we conduct experiments on cross-modal retrieval tasks to demonstrate the same. We compare against CLIP as our primary baseline, along with recent works like Menon & Vondrick [27] and CuPL [35], that also leverage complementary information from foundation models to improve upon CLIP. We train all CLIP-based baselines from scratch under the same training conditions.

Table 1: Zero-shot transfer evaluation of different models. We highlight the best performance of each setting in **bold**. We see that GRAIN improves performance under both pretraining datasets, outperforming CLIP by up to **9%** in absolute top-1 accuracy. CLIP* is a version of CLIP with the same number of parameters as our method for fair comparison.

Data	Model	CIFAR-10	CIFAR-100	SUN397	Cars	DTD	Pets	Caltech-101	Flowers	CUB	Places365	Food101	**Average**	**ImageNet**
	LLaVA	89.69	57.72	55.24	15.90	35.37	47.16	75.03	24.69	6.22	29.43	52.80	44.48	35.20
CC3M	CLIP [37]	48.86	18.70	28.44	0.68	9.23	6.94	41.02	8.48	2.51	17.85	8.73	17.40	14.01
	CLIP*	46.99	18.49	29.76	0.52	8.40	6.62	42.56	8.29	3.36	18.70	10.01	17.62	14.04
	Menon&Vondrick [27]	49.35	17.93	29.74	0.60	10.43	7.05	43.89	7.67	2.84	19.12	9.64	18.02	13.86
	CuPL [35]	50.16	18.98	29.66	0.71	9.89	8.22	43.95	8.84	2.91	19.73	10.51	18.51	14.14
	GRAIN (Ours)	**65.86**	**35.20**	**38.07**	**1.34**	**17.24**	**14.15**	**65.20**	**13.24**	**5.47**	**24.96**	**16.18**	**27.00**	**23.34**
CC12M	CLIP [37]	71.24	36.66	48.84	4.57	19.28	42.06	70.09	20.51	7.63	31.84	40.94	35.79	34.66
	CLIP*	70.07	35.63	50.42	4.31	18.35	39.40	74.24	21.04	7.96	32.03	41.36	35.89	33.51
	Menon&Vondrick [27]	72.68	37.08	48.59	5.12	18.45	41.38	72.29	21.15	8.27	31.36	41.20	36.14	34.32
	CuPL [35]	72.85	37.37	49.06	4.88	18.71	41.58	71.17	22.82	7.94	30.28	40.89	36.15	34.65
	GRAIN (Ours)	**81.40**	**46.23**	**55.26**	**8.42**	**25.68**	**48.76**	**81.49**	**26.27**	**10.28**	**36.76**	**45.39**	**42.36**	**41.46**

4.1 Experimental Setup

Model Architectures. For all models, we employ the ViT-B/16 [11] architecture for the vision encoders and the Transformer base model [43] for text encoders as described in CLIP [37]. Our approach, GRAIN, additionally utilizes a query-decoder with 6 transformer decoder layers. We set the number of queries n_q to 10. The outputs from the decoder are processed by projection layers to obtain features in the semantic space, and a 2-layered MLP for predicting bounding boxes. In addition to these comparisons, we evaluate our approach against the substantially larger LLaVA v1.6 model, which includes a ViT-L/14 paired with Vicuna-13 LLM. For this model, we utilize a pretrained checkpoint from huggingface [46].

Datasets. All models are pre-trained on two distinct image-text datasets that vary in scale: Conceptual Captions 3M (CC3M) and Conceptual Captions 12M (CC12M) [42], with the majority of our ablations studies performed on the CC3M dataset. For visual recognition, we evaluate all models on Imagenet [9] and 11 other domain-specific datasets, which include fine-grained datasets. Cross-modal retrieval evaluations are carried out on MS-COCO [21] and Flickr30k [50].

Evaluation Setup. It is important to note that, we evaluate all approaches on all tasks within the framework of a zero-shot evaluation protocol. For classification tasks, we report zero-shot top-k accuracy, and for retrieval tasks, we report zero-shot recall@k for various values of k: 1, 5, 10. For classification tasks, we report performance upon integrating the Menon & Vondrick [27] test-time heuristic—which enhances class names with LLM-generated descriptions—into our approach, observing positive gains. We also include results excluding this

Table 2: Results (Recall@k) on zero-shot image-to-text and text-to-image retrieval tasks on MS-COCO and Flickr30k.

Data	Model	MS-COCO						Flickr30k					
		Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CC3M	CLIP	15.79	38.26	50.70	13.58	33.76	46.04	27.00	53.80	66.30	21.78	44.26	55.10
	GRAIN	**38.26**	**65.96**	**77.03**	**28.81**	**55.86**	**69.00**	**59.90**	**81.80**	**88.40**	**42.82**	**68.21**	**76.54**
	Δ	**+22.47**	**+27.70**	**+26.33**	**+15.23**	**+22.10**	**+22.96**	**+32.90**	**+28.00**	**+22.10**	**+21.04**	**+23.95**	**+21.44**
CC12M	CLIP	41.32	69.40	80.04	30.02	57.32	69.65	59.60	84.70	89.90	43.63	68.75	76.77
	GRAIN	**58.30**	**83.07**	**89.67**	**42.66**	**70.77**	**80.83**	**78.00**	**94.60**	**97.80**	**59.36**	**80.01**	**85.59**
	Δ	**+16.98**	**+13.67**	**+9.63**	**+12.64**	**+13.45**	**+11.18**	**+18.40**	**+9.90**	**+7.90**	**+15.73**	**+11.26**	**+8.82**

heuristic in ablations and note that our approach continues to perform effectively.

Pretraining Setup. All images are resized to a resolution of 224 × 224, and we employ the standard sentencepiece tokenizer for text tokenization. Training for all models is conducted using the AdamW optimizer [23] across 35 epochs, using a cosine learning rate schedule and weight decay regularization. We use a batch size of 1024 for CC3M experiments and 2048 for CC12M. Training GRAIN for CC3M on a 8 NVIDIA H100 DGX machine takes about 16 hours and for CC12M experiments on 2 × 8 H100 machines takes 36 hours. While training GRAIN, we randomly choose between the original caption and the VLM-generated caption as the text supervision:

$$t \sim Uniform([t_{original-caption}, t_{MLLM-caption}]). \tag{2}$$

Baselines. To ensure fair evaluation, all baselines were trained under conditions similar to GRAIN. The introduction of the decoder architecture in our model results in a 22% increase in parameter count compared to CLIP. For a more fair comparison we report numbers for CLIP by leveraging the same architectures as GRAIN but with localization modules turned off. This baseline is reported as CLIP* throughout the paper. Additionally, we report the performance of the LLaVA v1.6 model to benchmark our model's performance against a state-of-the-art MLLM. Despite possesing orders of magnitude more parameters and being trained on billion-scale datasets, our method manages to match and sometimes even surpass LLaVA's performance on the reported datasets.

4.2 Zero-shot image classification

We perform zero-shot classification and evaluate all models on Imagenet and 11 additional datasets encompassing common and fine-grained sets. We measure the top-1 accuracy and report results in Table 1. Our approach, GRAIN, consistently outperforms the current state-of-the-art across all settings and datasets. Specifically, GRAIN improves the zero-shot performance by as much as **9%** in absolute accuracy on Imagenet and achieves similar improvements averaged across all other datasets. Notably, our method surpasses existing benchmarks

Table 3: We report top-1 accuracy (%) for zero-shot attribute-based classification. This is a challenging task as indicated by the results.

Data	Model	CIFAR-10	CIFAR-100	SUN397	Cars	DTD	Pets	Caltech-101	Flowers	CUB	Places365	Food101	**Average**	**ImageNet**
CC3M	CLIP	24.20	7.30	13.65	0.75	6.86	3.43	24.68	1.90	**1.79**	8.93	5.04	8.97	4.53
	GRAIN (Ours)	**46.06**	**18.20**	**20.02**	**0.95**	**14.57**	**4.87**	**45.82**	**2.34**	1.72	**13.06**	**7.63**	**15.93**	**7.87**
	Δ	**+21.86**	**+10.90**	**+6.37**	**+0.20**	**+7.71**	**+1.44**	**+21.14**	**+0.44**	**-0.07**	**+4.13**	**+2.59**	**+6.96**	**+3.34**
CC12M	CLIP	43.71	16.05	23.06	1.67	11.33	7.02	40.61	**4.08**	2.29	14.78	12.74	16.12	9.41
	GRAIN (Ours)	**67.39**	**26.29**	**32.46**	**4.21**	**17.61**	**12.38**	**59.09**	3.66	**2.72**	**20.39**	**18.29**	**24.04**	**14.53**
	Δ	**+23.68**	**+10.24**	**+9.40**	**+2.54**	**+6.28**	**+5.36**	**+18.48**	**-0.42**	**+0.43**	**+5.61**	**+5.55**	**+7.92**	**+5.12**

by significant margins across both fine and coarse-grained datasets, with our most substantial improvement reaching up to **22%** absolute accuracy on the Caltech-101 [13] dataset within the CC3M training setting.

4.3 Cross-modal retrieval

We evaluate the pre-trained models on the task of cross-modal retrieval under the zero-shot setting. Specifically, we focus on the Image-to-Text (I2T) and Text-to-Image (T2I) retrieval tasks using the MSCOCO and Flickr30k datasets in Table 2. Our evaluations are conducted on the standard test sets for both datasets, and we report performance metrics in terms of Recall@k for k values of 1, 5, and 10. Compared to CLIP, our method achieves superior performance with performance gains of up to **33%**. On average, we observe improvements of **23.8%** with CC3M trained models and **12.46%** with models trained on CC12M.

4.4 Zero-shot attribute-based classification

To measure image-description alignment, we design an experiment to classify images by leveraging only descriptions/attributes. This is a challenging task, as image classification is being performed devoid of class names. A model that has learned good alignment between images and descriptions is expected to succeed at this task. Learning this alignment is crucial for succeeding at fine-grained visual recognition and recognizing novel examples by leveraging complementary information in the form of descriptions.

Toward this end, we first prompted GPT-3 using class names from the downstream dataset's vocabulary to obtain descriptions. We gathered around 5-7 descriptions per class. Next, instead of the traditional approach of encoding class names and computing similarities with images, we encoded the description corresponding to the class name and computed similarities with images. As we have multiple descriptions per class, we aggregated them by averaging their encodings. We used this aggregate representation as the text representation for the class and computed similarity with images. The class corresponding to the text

Table 4: Ablation studies on our CC3M trained model reporting top-1 accuracy (%)

Setting	CIFAR-10	CIFAR-100	SUN397	Cars	DTD	Pets	Caltech-101	Flowers	CUB	Places365	Food101	**Average**	**ImageNet**
GRAIN	**65.86**	**35.20**	**38.07**	**1.34**	**17.24**	**14.15**	**65.20**	**13.24**	**5.47**	**24.96**	**16.18**	**27.00**	**23.34**
– Region-description loss	58.21	27.07	35.28	1.01	14.20	9.18	58.86	9.13	3.52	22.31	13.05	22.89	18.73
– Box loss	57.06	26.17	34.38	0.93	14.67	8.87	56.91	8.31	3.20	21.35	13.12	22.27	17.54
– MLLM-caption	47.24	19.92	28.51	0.70	8.78	7.04	43.95	8.20	2.99	20.06	9.01	17.85	14.56
– Menon&Vondrick [27]	46.99	18.49	29.76	0.52	8.40	6.62	42.56	8.29	3.36	18.70	10.01	17.62	14.04

representation that scored the maximum similarity with the test image is considered the prediction for that image. We compute top-1 accuracy as usual and reported for all datasets in Table 3.

From Table 3, we observe that our model is able to achieve strong improvements over CLIP, demonstrating closer image-description alignment. On average, we achieve an improvement of **6-7%** over CLIP, showcasing better alignment.

4.5 Ablations

To assess the importance of the different components in GRAIN, we conduct four ablation experiments. We restrict to models trained on CC3M due to computational constraints. We examine the impact of different training losses—the region-description alignment loss, the bounding box prediction loss—and the effects of incorporating MLLM-generated captions during training and using Menon & Vondrick [27]-style descriptions at test-time. The outcomes of these ablations are reported as top-1 accuracy in Table 4.

Ablating the region-description alignment loss. This component is pivotal to our framework. Omitting this loss results in a significant 5% absolute decline in performance on Imagenet and a similar average drop across other datasets. This considerable decrease underscores the vital role of this loss in establishing fine-grained correspondences between salient image regions and their descriptions.

Ablating the localization loss. Further removing the bounding box prediction losses from our training regime leads to a modest performance drop. This loss is instrumental in identifying and predicting salient regions within the image, and, in conjunction with the alignment loss is crucial to developing fine-grained visual understanding.

Ablating the role of MLLM-caption during training. We employ captions generated by LLaVA as a form of text-level data augmentation during training, alternating between these and the original image captions. The MLLM-generated caption provides a high-level visual summary of the image, proving to be significant for training, as indicated by a 3% decrease in performance upon its removal.

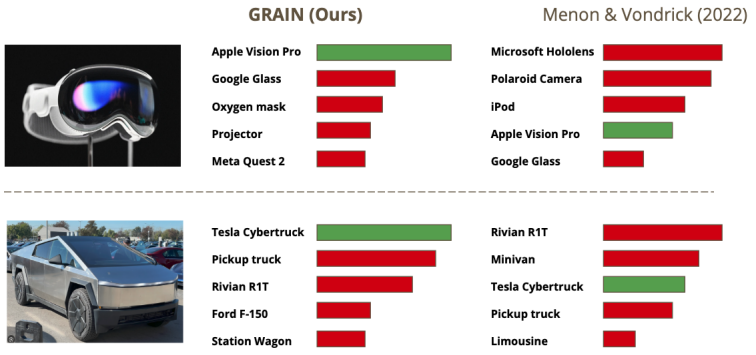


Fig. 4: Visualization of top-5 predictions of our model on novel entities alongside [27]. Our method consistently identifies the ground truth class as the top prediction.

Ablating the role of test-time descriptions. In line with the approach of Menon & Vondrick [27], we utilize descriptions generated by GPT-3 to enrich class names during zero-shot classification. Excluding these augmented descriptions results in a minor performance reduction, suggesting that while beneficial, our model's performance is not reliant on these test-time descriptions.

4.6 Qualitative Analysis

Recognizing Novel Examples As discussed in earlier sections, it is desired for an open-vocabulary model like CLIP to recognize concepts that were poorly represented or absent in its training data. To keep pace with an evolving world, such models must generalize to new and emerging concepts. However, recognizing a concept like `Apple Vision Pro` or `Tesla Cybertruck` that did not exist at the time of training the model, is challenging, as CLIP has not learned the association between images and names of these concepts. Continually training a model on new and emerging concepts is expensive and leads to well known problems like catastrophic forgetting [26]. Instead, Zero-shot learning methods often utilize auxiliary information, such as attributes, for classifying unknown entities. Hence, our approach aims to recognize these concepts by leveraging LLM-generated descriptions. In our experiment, we focus on recognizing newly popular entities, namely the `Apple Vision Pro` and `Tesla Cybertruck`, which emerged after datasets like Conceptual Captions were constructed. First, we add these two classnames to the Imagenet-1K vocabulary. Next, to simulate a real-world open-vocabulary scenario, we also include three related but distinct categories for each novel entity, making this a challenging task. Specifically, for the `Apple Vision Pro`, we add competing Virtual Reality (VR) headsets such as the `Meta Quest 2`, `Microsoft Hololens`, and `Google Glass`. For the `Tesla Cybertruck`, we include other pickup trucks like the `Rivian R1T`, `Ford F-150`, and `Toyota Tundra`. We then utilize GPT-3 (language only) to generate descriptions for the concepts in this extended vocabulary. Following

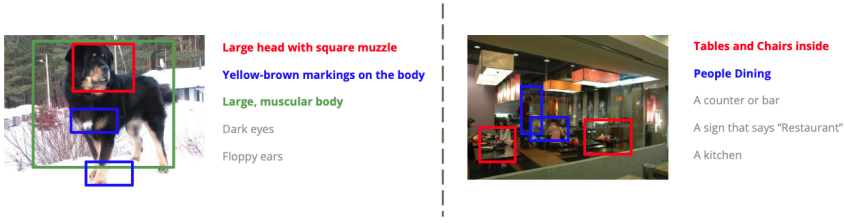


Fig. 5: Localization and region-description matching predictions made by our model on images from ImageNet.

the test regimen introduced in [27], we present the top-5 predictions made by both our model and [27] in Figure 4. Our findings indicate that our model consistently identifies the correct class names with high confidence, whereas the baseline is able to include it in top-5 but fails to rank them as the top choice. This highlights our models ability to recognize novel concepts by leveraging the learned image-description alignment.

Grounding Visualizations. To showcase the efficacy of our grounding module, we present visualizations of its predictions in Figure 5, with images from the Imagenet dataset. We include additional visualizations in the Appendix. These visualizations include LLM-generated descriptions and the corresponding bounding boxes predicted by our model, with each matched pair coded by color. We also include descriptions belonging to this class that are not matched to a bounding box.

5 Conclusion

In this paper, we propose a new pre-training method for contrastive vision-language models. Specifically, we hypothesize that many of the current limitations of CLIP stem from its image-level contrastive pre-training, which neglects fine-grained alignment. As a result, we propose to leverage Multi-Modal Large Language Models (LLaVA) and Open-Vocabulary Object Detectors (OWLv2) to automatically generate weak supervision to drive a more fine-grained pre-training process. Specifically, we propose a two-stage annotation pipeline to create fine-grained descriptions of the images, and use the object detectors to localize them. This automatic annotation pipeline requires no cleaning/curation, and the obtained annotations are used to train a DETR-style encoder-decoder architecture with three proposed losses that encourage fine-grained alignment. We demonstrate superior performance across 11 different classification datasets, including fine-grained ones, as well as additional tasks such as cross-modal retrieval. Our results show significant improvement over the state-of-art, including by up to 9% in absolute top-1 accuracy for zero-shot classification and 25% on retrieval. Our method can even rival the MLLM, which is over 13B parameters (compared to our ∼170M) and was trained on billions of data-points.

References

1. Bica, I., Ilić, A., Bauer, M., Erdogan, G., Bošnjak, M., Kaplanis, C., Gritsenko, A.A., Minderer, M., Blundell, C., Pascanu, R., Mitrović, J.: Improving fine-grained understanding in image-text pre-training (2024) 19, 22
2. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020) 3
3. Burns, K., Witzel, Z., Hamid, J.I., Yu, T., Finn, C., Hausman, K.: What makes pre-trained visual representations successful for robust manipulation? (2023) 20
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers (2020) 4, 6, 20
5. Chen, B., Shvetsova, N., Rouditchenko, A., Kondermann, D., Thomas, S., Chang, S.F., Feris, R., Glass, J., Kuehne, H.: What, when, and where? – self-supervised spatio-temporal grounding in untrimmed multi-action videos from narrated instructions (2023) 20
6. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning (2020) 20
7. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/ 4
8. Conti, A., Fini, E., Mancini, M., Rota, P., Wang, Y., Ricci, E.: Vocabulary-free image classification (2023) 20
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 9
10. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023) 3
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4, 9
12. Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y.: Improving clip training with language rewrites. In: NeurIPS (2023) 3
13. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. pp. 178–178. IEEE (2004) 11
14. Hu, H., Luan, Y., Chen, Y., Khandelwal, U., Joshi, M., Lee, K., Toutanova, K., Chang, M.W.: Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. International Conference on Computer Vision (2023) 2, 20
15. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3922–3931 (2021). https://doi.org/10.1109/ICCV48922.2021.00391 19

16. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision (2021) 19
17. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr – modulated detection for end-to-end multi-modal understanding (2021) 20
18. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 951–958 (2009). https://doi.org/10.1109/CVPR.2009.5206594 4
19. Lei, J., Berg, T.L., Bansal, M.: Qvhighlights: Detecting moments and highlights in videos via natural language queries (2021) 20
20. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J.: Oscar: Object-semantics aligned pre-training for vision-language tasks (2020) 20
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 9
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024) 3, 4
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 10
24. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks (2019) 20
25. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft (2013) 2
26. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989) 13
27. Menon, S., Vondrick, C.: Visual classification via description from large language models. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=jlAjNL8z5cs 2, 3, 8, 9, 12, 13, 14, 19, 21, 24
28. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36** (2024) 5, 22
29. Mirza, M.J., Karlinsky, L., Lin, W., Possegger, H., Feris, R., Bischof, H.: Tap: Targeted prompting for task adaptive generation of textual training instances for visual classification (2023) 19
30. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training (2021) 19
31. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2019) 8
32. OpenAI: Gpt-4v(ision) system card. OpenAI (2023) 3
33. Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., Gatt, A.: VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8253–8280. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.567, https://aclanthology.org/2022.acl-long.567 2

34. Parikh, D., Grauman, K.: Relative attributes. In: 2011 International Conference on Computer Vision. pp. 503–510 (2011). https://doi.org/10.1109/ICCV.2011.6126281 4
35. Pratt, S., Covert, I., Liu, R., Farhadi, A.: What does a platypus look like? generating customized prompts for zero-shot image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15691–15701 (2023) 2, 3, 8, 9, 19, 21
36. 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 1
37. 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: International conference on machine learning. pp. 8748–8763. PMLR (2021) 9, 19, 20, 21
38. Ranasinghe, K., McKinzie, B., Ravi, S., Yang, Y., Toshev, A., Shlens, J.: Perceptual grouping in contrastive vision-language models (2023) 2
39. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. arXiv preprint arXiv:2311.03356 (2023) 3
40. Ren, Z., Su, Y., Liu, X.: Chatgpt-powered hierarchical comparisons for image classification (2023) 19
41. 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. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022) 3, 19
42. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018) 3, 4, 9, 19
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 9
44. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: Cub. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 2
45. Wang, A.J., Ge, Y., Cai, G., Yan, R., Lin, X., Shan, Y., Qie, X., Shou, M.Z.: Object-aware video-language pre-training for retrieval (2022) 20
46. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Huggingface's transformers: State-of-the-art natural language processing (2020) 9, 21
47. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5542–5551 (2018). https://doi.org/10.1109/CVPR.2018.00581 4
48. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training (2021) 19, 22
49. Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T., Kong, L.: Zerogen: Efficient zero-shot learning via dataset generation (2022) 4

50. Young, P., Lai, A., Hodosh, M., 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) 9
51. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? (2023) 2
52. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training (2023) 19
53. Zhang, C., Gupta, A., Zisserman, A.: Helping hands: An object-aware ego-centric video recognition model (2023) 20
54. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 3

Appendix

A Additional Details on Related Work

The scope of our work spans several domains including Vision-Language Representation Learning, Fine-grained Visual Recognition, Visual Grounding and Open-world/Zero-shot learning. Since the related works section in the main paper cannot adequately cover all of these areas, we provide a more comprehensive summary in this supplementary material:

Contrastive Language-Image Pretraining. Methods like CLIP [37] and ALIGN [16] leverage large internet scraped datasets of image-text pairs to learn a joint representation by contrastively aligning the two modalities. The objective of these methods is to pull together image and text representations that are semantically similar and push apart dissimilar pairs. These works employ a dual encoder approach, separately encoding representations for images and text. These learned representations are effective for various downstream vision and language tasks. Follow-up works [30, 52] in this area focus on improving downstream performance by incorporating self-supervision or using other objective functions during pretraining. However, aligning representations at a global (whole image or caption level) is known to only learn coarse-grained features and discard fine-grained visual information. Acknowledging this problem, FILIP [48] introduces a cross-modal late interaction mechanism that utilizes a token-wise maximum similarity between image and text tokens to drive the contrastive objective. In the medical domain, GLORIA [15] proposes an attention-based framework that uses text tokens to attend to sub-image regions and learns local and global representations. Concurrent to our work, SPARC [1] proposes using a sparse similarity metric between image patches and text tokens to learn fine-grained alignment. Our paper shares a motivation to these works in terms of aiming to learn fine-grained representations. However, unlike these methods, we address the fact that image-caption datasets like Conceptual Captions [42] or LAION [41] contain noisy captions that lack descriptive information, thereby limiting the gains that such fine-grained region-token matching objectives can achieve. Secondly, our approach focuses on learning visual representations that would be able to leverage complementary information at test-time (in the form of LLM-generated descriptions as proposed by [27,35]) to recognize fine-grained or novel entities. Finally, in principle, these methods are orthogonal to our contributions and can be coupled with our method.

Zero-shot Learning with CLIP. In image classification, Zero-shot learning methods aim to recognize novel entities that were not seen during training. Relevant to our work, Menon & Vondrick [27] leverage category descriptions generated from a Large Language Model (LLM) as auxiliary information to augment the zero-shot performance of CLIP. On similar lines, CuPL [35] and Ren et. al. [40] use LLMs to generate descriptions in the form of long, cohesive sentences or via nuanced, hierarchy-aware comparisons. TAP [29] learns a text classifier mapping descriptions to categories during training which is used to map

from images to categories at test-time. Different from these works, our method aims to improve alignment between images and descriptions that would further bolster the efficacy of using descriptions at test-time.

Object-aware Vision-Language Pretraining. Encouraging object-oriented representations within a vision-language pretraining objective [3, 5, 6, 20, 24, 45, 53] has been shown to facilitate learning of robust models that can positively impact downstream performance across a variety of tasks in vision-language, video understanding and embodied AI. Many of these approaches follow the DETR line of works [4, 17, 19] that introduce a query-transformer backbone for detection and grounding. We take inspiration from these works to develop our architecture for encoding visual information. However, our approach only uses the grounding task as an auxiliary objective to distill information from local regions to global representations. We leverage our synthetic descriptions to supervise this grounding module, which is then disabled during evaluation as detailed earlier.

Universal Visual Recognition. Recent works [8, 14] introduce the problem of universal visual recognition or vocabulary-free image classification, where the motivation is to free models like CLIP from a constrained vocabulary thereby allowing classification from an unrestricted set of concepts. Corroborating with our claims, these works observe limitations of CLIP toward recognizing novel examples and fine-grained entities. These works formalize this problem and introduce retrieval-based methods as an initial step towards a solution.

B Implementation Details

All baselines reported in the paper (except LLaVA) utilize a ViT-B/16 model as the vision encoder. For encoding text, we utilize a 12-layer transformer network as used with CLIP [37]. The outputs from the vision encoder are 768 dimensional, which are then projected to 512. The outputs embeddings obtained from the decoder are also passed through separate projection layers. The projection layer is shared between all region output embeddings and a separate projection layer is used for the image output embedding. Similarly, the text-encoder output is projected to be of the same 512 dimensional size. Additionally, a two-layer MLP with output size 4 is used to regress on the bounding boxes conditioning on the region output embeddings. The supervision for bounding boxes is obtained through the OWLv2 detector, which is originally for a 960×960 resolution image which is down-scaled to 224×224 following the input resolution of our model. While generating these bounding box annotations from OWLv2, we use a confidence threshold value of 0.3.

C Additional Results and Baseline Analysis

CLIP. We train the CLIP ViT-B/16 variant on the same Conceptual Captions (CC3M and CC12M) datasets as our GRAIN model. For performing zero-

Table A: Zero-shot top-1 accuracy (%) of different methods.

Data	Model	CIFAR-10	CIFAR-100	SUN397	Cars	DTD	Pets	Caltech-101	Flowers	CUB	Places365	Food101	**Average**	**ImageNet**
	LLaVA	89.69	57.72	55.24	15.90	35.37	47.16	75.03	24.69	6.22	29.43	52.80	44.48	35.20
CC3M	CLIP [37]	48.86	18.70	28.44	0.68	9.23	6.94	41.02	8.48	2.51	17.85	8.73	17.40	14.01
	Menon&Vondrick [27]	49.35	17.93	29.74	0.60	10.43	7.05	43.89	7.67	2.84	19.12	9.64	18.02	13.86
	CuPL [35]	50.16	18.98	29.66	0.71	9.89	8.22	43.95	8.84	2.91	19.73	10.51	18.51	14.14
	CLIP*	46.99	18.49	29.76	0.52	8.40	6.62	42.56	8.29	3.36	18.70	10.01	17.62	14.04
	CLIP* + Menon&Vondrick [27]	49.37	17.98	29.94	0.62	10.55	7.14	44.02	8.38	3.51	19.23	10.24	18.27	13.97
	CLIP* + CuPL [35]	50.24	18.86	30.12	0.74	10.14	8.06	43.78	8.95	3.32	19.56	10.77	18.59	14.14
	GRAIN (Ours)	**65.86**	**35.20**	**38.07**	**1.34**	**17.24**	**14.15**	**65.20**	**13.24**	**5.47**	**24.96**	**16.18**	**27.00**	**23.34**
CC12M	CLIP [37]	71.24	36.66	48.84	4.57	19.28	42.06	70.09	20.51	7.63	31.84	40.94	35.79	34.66
	Menon&Vondrick [27]	72.68	37.08	48.59	5.12	18.45	41.38	72.29	21.15	8.27	31.36	41.20	36.14	34.32
	CuPL [35]	72.85	37.37	49.06	4.88	18.71	41.58	71.17	22.82	7.94	30.28	40.89	36.15	34.65
	CLIP*	70.07	35.63	50.42	4.31	18.35	39.40	74.24	21.04	7.96	32.03	41.36	35.89	33.51
	CLIP* + Menon&Vondrick [27]	72.74	37.44	51.20	5.31	18.47	41.74	74.44	21.22	8.32	32.72	41.92	36.87	34.50
	CLIP* + CuPL [35]	72.77	37.85	51.08	5.12	18.98	41.14	74.22	22.68	8.05	32.34	41.65	36.90	34.77
	GRAIN (Ours)	**81.40**	**46.23**	**55.26**	**8.42**	**25.68**	**48.76**	**81.49**	**26.27**	**10.28**	**36.76**	**45.39**	**42.36**	**41.46**

shot testing on all reported datasets, we use the handcrafted prompts specific to each dataset as introduced in the official code-base [37]. These hand-engineered prompts improve the zero-shot performance of CLIP beyond the vanilla, `A photo of {classname}` style prompts.

CLIP*. With the introduction of the decoder and bounding-box modules, our method, GRAIN, uses $\sim 22\%$ more parameters compared to CLIP. For a more fair comparison in terms of number of parameters, we report performance for CLIP by using the same architecture as ours, but with the localization modules turned off. We refer to this baseline as CLIP*

Menon&Vondrick. We leverage the official code-base [27] to report performance for this baseline. In the main paper, we implement this baseline on top of the CLIP method as per the norm. Additionally, in Table A we report performance on combining this method with the CLIP* model. This boosts performance over CLIP* on most datasets as expected while still trailing our method.

CuPL. Similarly, we implement the CuPL baseline leveraging official code [35] and report performance with CLIP and CLIP* in the main paper and in Table A respectively. CuPL shows a similar trend of improving over CLIP baselines but trailing behind our method.

LLaVA. We use a pretrained LLAVA v1.6 checkpoint from huggingface [46] that is composed of a ViT-L/14 vision encoder and a Vicuna-13B LLM. The vision and text encoders of LLAVA have been separately pretrained on billion-scale datasets and conjoined through a projection layer. LLaVA has been trained through multiple stages on a specialized set of $\sim$ 150k instructions. Being a generative model, we ask LLAVA to predict a category for an image by using prompts specific to each dataset as described in Table B. Next, we use a pretrained CLIP

Table B: Prompts to LLaVA for the zero-shot visual recognition task in Table A.

Dataset	Prompt
DTD	Fill in the blank: this is a photo of a {} texture
Pets	What animal is in the image? Be specific about the breed. Fill in the blank: this is a photo of a {}
Places365	What place is this in the image? Fill in the blank: this is a photo of a {}
Food101	What food is in the image? Fill in the blank: this is a photo of a {}
Cars	What type of car is in the image? Be specific about the make and year. Fill in the blank: this is a photo of a {}
Others	Fill in the blank: this is a photo of a {}

text encoder to map the answer generated by LLaVA to the closest category in the vocabulary of the dataset being evaluated on. We use this mapped category as the prediction to compute the top-1 accuracy as usual. Observing Table A, our approach is able to reach and even surpass LLaVA's performance on several datasets despite having orders of magnitude smaller parameters and training datasets.

FILIP. Although FILIP [48] shares a similar motivation to our method, we note that the approach taken for FILIP is orthogonal to ours. FILIP employs a cross-modal late interaction mechanism to learn associations between image patches and caption tokens without using any side information. In contrast, our approach leverages complementary information in the form of descriptions and their corresponding localizations to learn fine-grained alignments. Being orthogonal to our contributions, in principle, the late interaction mechanism from FILIP can be coupled with our approach. Secondly, a concurrent work [1] finds FILIP's results challenging to reproduce due to high training instability, and in practice, observe FILIP to substantially underperform even the zero-shot performance of CLIP on classification tasks. For these reasons, we refrain from comparing our method to FILIP.

D Dataset Details

We train all models on the CC3M and CC12M datasets. As explained earlier, to obtain description and localization annotations, we prompt LLaVA in two stages to extract the primary visual subject of the image and then gather image descriptions by asking LLaVA to focus on the identified visual subject. We obtain bounding boxes corresponding to each description by using OWLv2 [28], an off-the-shelf open-vocabulary object detector. We filter the predicted boxes using a confidence threshold of 0.3 to discard noisy predictions.

Original caption: looking for creative date ideas ? check out this list of date night ideas !

MLLM caption: The image features a night sky filled with stars, set against a backdrop of a rocky landscape.

Primary visual subject: Sky

Descriptions: Dark night sky, Stars visible, Milky way galaxy, Rocky landscape

Original caption: a view from the pond

MLLM caption: A large, white building with a red roof, surrounded by trees and a pond

Primary visual subject: Building

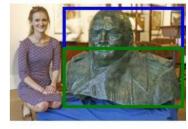
Descriptions: Arched entrance, Red roof, White walls, Multiple Windows, Fountain

Original caption: auction : the bronze bust of wartime prime minister could fetch as much as £ 80,000 in april

MLLM caption: A woman kneeling in front of a large, greenish-brown statue of a man, possibly a bust

Primary visual subject: Statue

Descriptions: Sculpted with realism, Large size, Military uniform, Greenish-brown color



Original caption: uniforms deliver innovation while honoring the past

MLLM caption: Football player in red and white uniform with number 43

Primary visual subject: Football Player

Descriptions: Helmet with mask, Red-white uniform, Black gloves, Number 45 on jersey, white pants

Original caption: astronaut and earth globe in the deep space

MLLM caption: Astronaut floating in space, looking down at Earth.

Primary visual subject: Astronaut

Descriptions: Gloves on hands, No helmet, Earth in background, Floating in space, Arms extended

Original caption: aerial view of road in the middle of forest

MLLM caption: Trees in the image are dense and green, with a clear separation between two distinct areas.

Primary visual subject: Trees

Descriptions: Green leaves, dense foliage, tall trunks, curved branches, layered canopy, uniform height

Fig. A: Sample annotations generated using our two-stage LLaVA prompting scheme followed by OWLv2 localization.

E Sample Annotations

In Figure A, we illustrate the annotations obtained using our two-stage LLaVA prompting followed by bounding box prediction using OWLv2. We randomly select images and captions (original caption) from the CC3M dataset and present the corresponding MLLM caption, primary visual subject, and descriptions generated by our annotation pipeline. The descriptions are color-coded by their associated bounding box. Overall, our annotation pipeline is effective in identifying the primary visual subject, which is the most prominent object or concept in the image, and generating descriptions and corresponding localizations by focusing on this subject. The first five rows show cases where the pipeline successfully localized at least one description, whereas the last row demonstrates a case where no description could be localized due to the vague nature of the image, making the descriptions difficult to localize.

F Two-stage versus Single stage Annotation

In this work, we employ a two-stage annotation pipeline to elicit descriptions from LLaVA. Specifically, in the first stage, we prompt LLaVA to identify the primary visual subject in the image, followed by generating descriptions for this subject. We observe that this approach leads to the generation of descriptions that are more specific and focused on the constituent regions in the image that make up the subject. In Figure B, we compare the descriptions generated by this strategy with a single-stage pipeline that directly prompts LLaVA to generate descriptions without first identifying the subject. We randonly pick samples from the CC12M dataset to illustrate the difference. Contrasting the two setups, we can see that the two-stage approach produces more specific descriptions that are well-grounded in the image compared to the one-stage approach, which either outputs overly generic descriptions or tends to hallucinate (See Rows 1 and 2). This issue is more pronounced for complex scenes involving unusual or fine-grained objects.

G Failure modes in our grounding module

In the main paper, we showcased examples where the grounding module of our approach successfully localized descriptions in images. In this section, we particularly highlight failure cases where the model is unable to correctly localize descriptions within the image. Following the same setup as the main paper, we use images from ImageNet and descriptions generated from an LLM by following Menon & Vondrick's [27] strategy of prompting GPT-3 (language-only) with category names. It is important to note that since these descriptions were generated using only the category name and without access to images, some descriptions might not be visible in every image. We expect our approach to localize descriptions that are present in an image and not localize those that are absent. While

our approach effectively grounds descriptions on average, we illustrate failure cases in Figure C.

Row 1 includes partially successful cases, where the model localizes descriptions but the bounding boxes are either slightly off the mark or does not localize all instances of that description in the image.

Row 2 includes examples where either the model cannot localize a single description in the image or incorrectly associates the description with another region in the image. (the description `typically orange or brown` refers to the *basketball* but was incorrectly assigned to the *jersey of the player* that has a similar color.)

Row 3 includes cases of hallucination, where the model localizes descriptions that are not present in the image.

H Limitations and Broader Impact

Limitations. Our method achieves substantial gains over CLIP and other baselines on zero-shot transfer tasks such as image classification, attribute-based image classification, and cross-modal retrieval. These improvements can be attributed to the fine-grained region-to-description associations learned by our model during the training process. However, learning these correspondences requires annotations in the form of descriptions and bounding box localizations, which are computationally expensive to obtain. As mentioned earlier, our annotation scheme demands significant GPU resources and can take long hours for large datasets. Additionally, since we do not filter or curate these annotations, it might result in some misaligned or inaccurate descriptions or captions, which might not provide the correct signal during the learning process. Future work could explore the use of efficient models to generate annotations as well as a filtering mechanism to ensure all generated text and bounding boxes are correctly aligned with the semantic content of the image.

Broader Impact. We propose a strategy to learn fine-grained image-text correspondences without requiring additional human annotations. Our approach leverages weak supervision from Multimodal Large Language Models (MLLMs) to train a region-aware model that strongly outperforms CLIP across several tasks and datasets. Despite having significantly smaller parameters and training costs, our approach matches and sometimes even outperforms LLaVA, a state-of-the-art MLLM, on zero-shot visual recognition. Although obtaining these annotations is computationally expensive, once acquired, our approach can be viewed as enabling the training of smaller models with small-scale datasets to achieve performance equivalent to a large model trained on extensive data, potentially making Vision and Language Model (VLM) training more accessible.

Image	One stage	Two stage (Ours)
	**Descriptions:** • Colorful salad • Fresh ingredients • Plate on table • Blue background	**Primary Visual Subject:** Salad **Descriptions:** • Colorful vegetables • Fresh tomatoes • Cheese balls • Green herbs • Dressing drizzled over salad
	**Descriptions:** • Bald head • Black beard • Black shirt • Black socks • Hat • Shorts • Wristband	**Primary Visual Subject:** Basketball Players **Descriptions:** • Team jerseys with USA logo • Facial expressions of concern or focus • Sweatbands on wrists • Black hat
	**Descriptions:** • Curtains • Desk • Lamp • Television • Bed	**Primary Visual Subject:** Bedroom **Descriptions:** • White sheets • Brown comforter • Flowerpot • Red curtains
	**Descriptions:** • Man wearing suit • Woman wearing a wedding dress • Dancing in courtyard • Nighttime	**Primary Visual Subject:** Dance **Descriptions:** • Couple dancing • Lights on • People sitting • Tables and chairs • Archway
	**Descriptions:** • Black car • Rear spoiler • Tinted Windows • Sunroof • Custom paint job	**Primary Visual Subject:** Car **Descriptions:** • Black color • Rear spoiler • Tinted windows • Sunroof • Lowered suspension

Fig. B: Qualitative comparison between one-stage (middle) and two-stage (right) LLaVA-based annotation schemes.

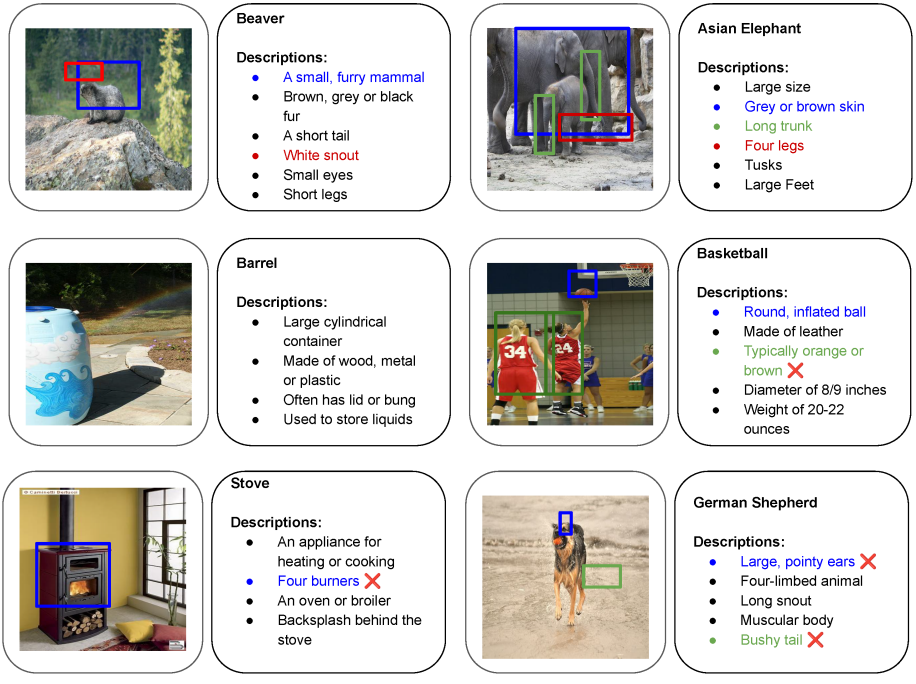


Fig. C: Visualization of failure modes from our grounding module on ImageNet-1K.