Grounding Descriptions in Images informs Zero-Shot Visual Recognition

Shaunak Halbe*[1] Junjiao Tian[1] K J Joseph[2] James Seale Smith[1]
Katherine Stevo[1] Vineeth N Balasubramanian[3] Zsolt Kira[1]
[1]Georgia Institute of Technology [2]Adobe Research [3]Indian Institute of Technology, Hyderabad

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 modest improvements. We attribute these limited gains to a misalignment between image regions and textual descriptions, which stems from CLIP's global alignment objective. 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 large-scale synthetic annotations. 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, we introduce Products-2023, a newly curated, manually labeled dataset featuring novel concepts, and showcase our model's ability to recognize these concepts by benchmarking on it. Significant improvements achieved by our model on other downstream tasks like retrieval further highlight the superior quality of representations learned by our approach. Code available at https://github.com/shaunak27/grain-clip.

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 [21] 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 [15, 28]. Secondly, while known for its open-vocabulary potential, it can still perform poorly for some domains not well-represented 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 [9]. 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 [16, 20] 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 repre-

*Correspondence to shalbe9@gatech.edu

sentations used in zero-shot classification. As an example, the class `French Bulldog` would be expanded to `A French Bulldog, which has small and pointy ears.` These methods provide some improvements over standard CLIP models, though they leave room for further advancements.

We attribute the limited gains from injecting descriptions to the training schema of CLIP, which optimizes for global representation alignment between image and text modalities that might not be suitable for fine-grained tasks. 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 overarching captions, neglecting the rich information that image regions and textual descriptions share with each other. Our observations align with recent research [5, 30, 35] indicating that CLIP tends to overlook fine-grained visual details during pretraining, leading to subpar performance on tasks requiring localization [23], object attributes [32], or physical reasoning [19].

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 regions and detailed text descriptions. As a first step towards our approach, given that pretraining datasets (Conceptual Captions [25], LAION [24], etc.) only contain images with noisy captions but without detailed descriptions, we employ an instruction-tuned Multimodal Large Language Model (MLLM) to generate descriptions and 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). We then propose a method that learns to jointly ground text descriptions into specific image regions along with aligning image and caption representations at a 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. Finally, to evaluate our model's ability to recognize novel concepts, we curate and manually label a new image classification dataset, *Products-2023*, and benchmark upon it.

To summarize, our main contributions are as follows:

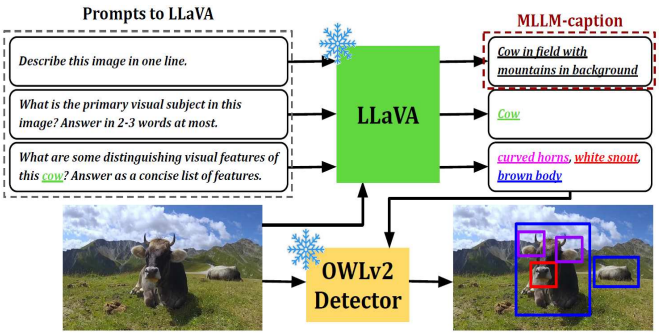

Figure 1. Overview of our two-stage annotation process: (1) prompting LLaVA for image descriptions and (2) acquiring corresponding region annotations from OWLv2.

- We empirically 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 objective designed to simultaneously learn local and global correspondences, obtained via weak supervision from Multimodal LLMs and open-vocabulary detectors.
- To drive this pre-training, we introduce an automated annotation engine to source fine-grained supervision signal.
- We demonstrate significant gains across a range of tasks, including image classification and retrieval, 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.
- Acknowledging the lack of image classification datasets containing novel examples, we collect and manually label a benchmark dataset termed *Products-2023*.

2. Related Works

Contrastive Language-Image Pretraining. Follow-up works on CLIP [22] and ALIGN [11] focus on improving the quality of learned representations by further introducing self-supervision or cross-modal alignment objectives [10, 18, 33]. Relevant to our focus, FILIP [31] introduces a cross-modal late interaction mechanism that explores token-wise maximum similarity between image and text tokens to improve fine-grained alignment. Recently, SPARC [1] proposes a sparse similarity metric between image patches and text tokens to learn fine-grained representations. While our paper shares motivation with these works, we address the fact that web-based captioning datasets [24, 25] contain noisy captions that lack descriptive information thereby limiting the gains achievable from such elaborate objectives. Instead, we source rich text descriptions and region annotations and design a pre-training objective to learn from them. This allows us to effectively use complementary information at test-time (in the form of

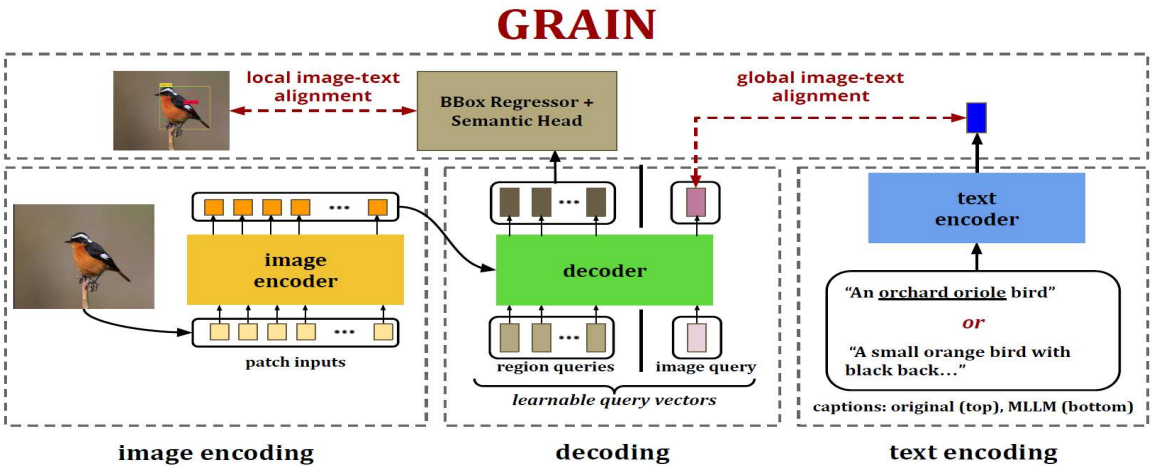

Figure 2. Architecture overview. Our method, GRAIN, aligns image representations to text captions at a global level while localizing salient image regions and aligning them to text descriptions at the local level.

LLM-generated descriptions) to recognize fine-grained or novel entities.

Improving CLIP using Generative Models. Recent works have explored the use of LLMs towards improving the downstream performance of CLIP. Menon *et al*. [16] and CuPL [20] 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. Different from these, LaCLIP [7] and VeCLIP [12] use 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 in the form on image regions and descriptions generated by a MLLM and an open-world detector to drive a novel pretraining strategy.

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 [3], to infuse the rich context from sub-image regions into learned visual representations. Alongside encoding the image into a semantic representation, our model predicts bounding boxes for salient image regions containing discriminative information. These localizations are then aligned 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 [25] (CC3M, CC12M) dataset and a 50M subset of LAION [24] (LAION-50M) 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 region-level 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 focusing on regions within the image and their descriptions in text as supervision for training our model. For generating descriptions and locating their corresponding regions, we leverage an instruction-tuned Multimodal Large Language Model, LLaVA[13]. We select LLaVA for its superior captioning capabilities and accessibility due to its openness; however, our approach is fundamentally compatible with any multimodal LLM. For our annotation purposes, we select the LLaVA v1.6 model which integrates a pretrained Vision Transformer Large (ViT-L) [6] as the visual encoder with the Vicuna-13B LLM [4]. It is worth noting that we only use LLaVA to describe regions/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 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.

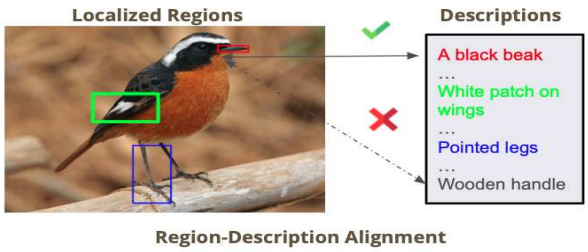

Figure 3. Contrastively align predicted regions with descriptions.

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.We present a qualitative analysis on this in the Appendix. This procedure provides us with a list of descriptions for each image. Additionally, we ask LLaVA to generate a short one-line description the image by prompting it with "*Describe this image in one line*". This description 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 description as the *MLLM-caption*, and the one from the pretraining 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 [17] 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. Subsequently, we eliminate redundant bounding box predictions using non-maximum suppression, retaining only the box with the highest confidence score for each region and discarding others with significant overlap.

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. We aim to release this dataset to benefit future research in this direction.

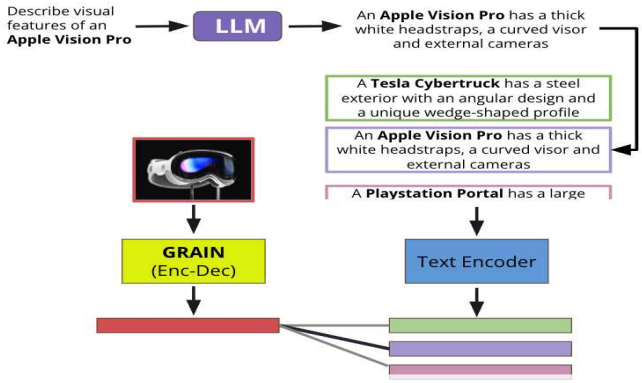

Figure 4. For zero-shot image classification, the image output embedding is compared with text embeddings of classnames enriched with descriptions for assignment.

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 [3], 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 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 is inactive during evaluation.

Bounding-Box Prediction. The region output embeddings

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 + CLIP	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[22]	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[16]	49.35	17.93	29.74	0.60	10.43	7.05	43.89	7.67	2.84	19.12	9.64	18.02	14.12
	CuPL[20]	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[16]	49.37	17.98	29.94	0.62	10.55	7.14	44.02	8.38	3.51	19.23	10.24	18.27	14.16
	CLIP* + CuPL[20]	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 [22]	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 [16]	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 [20]	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 [16]	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 [20]	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**
LAION-50M	CLIP	79.90	55.52	54.14	8.90	31.01	60.97	76.24	45.05	35.60	36.40	60.26	49.45	44.83
	Menon&Vondrick [16]	79.15	55.55	55.28	9.94	34.62	**62.36**	77.02	44.65	35.29	37.12	60.83	50.16	45.69
	GRAIN (Ours)	**86.48**	**64.55**	**58.86**	**12.56**	**40.42**	61.48	**79.15**	**46.65**	**35.79**	**37.44**	**61.48**	**53.17**	**48.44**

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 [26] to learn alignment between output region embeddings and descriptions. Here, the descriptions corresponding to ground truth 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-

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**
LAION-50M	CLIP	49.33	75.52	86.46	38.12	62.42	75.56	64.88	89.37	93.18	49.44	73.52	80.48
	GRAIN	**64.24**	**87.75**	**93.48**	**48.49**	**75.10**	**83.30**	**82.40**	**95.88**	**98.24**	**63.18**	**82.93**	**87.14**
	Δ	**+14.91**	**+12.23**	**+7.02**	**+10.37**	**+12.68**	**+7.74**	**+17.52**	**+6.51**	**+5.06**	**+13.74**	**+9.41**	**+6.66**

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/retrieval 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 inference; however, they can be used to provide valuable insights for interpreting the model's predictions. For zero-shot image classification (Tables 1, 5), we enhance class names by appending their descriptions, as illustrated in Figure 4. These descriptions are sourced from a LLM similar to [16, 20]. Leveraging the rich image-text correspondences learned during training, our model effectively uses these descriptions to recognize fine-grained and novel categories.

4. Experiments

The goal of our method is to learn fine-grained vision-language representations that can aid zero-shot visual recognition. 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 through experiments on cross-modal retrieval benchmarks. We compare against CLIP as our primary baseline, along with recent works like Menon & Vondrick [16] and CuPL [20], 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 and evaluate all approaches with a zero-shot evaluation protocol.

4.1. Experimental Setup

Model Architectures. For all models, we employ the ViT-B/16 [6] architecture for the vision encoders and the Transformer base model [27] for text encoders as described in CLIP [22]. We include results for additional ViT sizes in the Appendix. 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 [29].

Pretraining Setup. All models are pre-trained on two distinct image-text datasets that vary in scale: Conceptual Captions 3M (CC3M), Conceptual Captions 12M (CC12M) [25] and a 50M subset of LAION [24]. Training for all models is conducted using the AdamW optimizer [14] across 35 epochs, using a cosine learning rate schedule and weight decay regularization. While training GRAIN, we randomly choose between the original caption and the MLLM-generated caption as the text supervision.

Baselines. To ensure fair evaluation, all baselines were trained under conditions similar to GRAIN. The introduc-

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**

tion 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. Open-ended MLLMs like LLaVA are known to struggle with fine-grained visual recognition [34]. Hence, we propose a new inference strategy to evaluate LLaVA on classification tasks, providing a stronger baseline. Specifically, we first prompt LLaVA to predict a category for an image. Due to its open-ended nature, we cannot directly determine if the generated answer matches the ground truth. To address this, we use a pretrained CLIP text encoder to map LLaVA's generated answer to the closest category within the dataset's vocabulary. This mapped category is then used as the prediction to compute the top-1 accuracy. We refer to this baseline as **LLaVA + CLIP** in Table 1, representing a stronger and improved baseline over LLaVA alone. Despite possesing orders of magnitude more parameters and being trained on billion-scale datasets, our method manages to surpass LLaVA's performance, which shows that our improvements emerge from careful modeling decisions, rather than a simple increase in data volume or model size.

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 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 [8] 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%**.

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. Toward this end, we first prompted GPT-3 using class names from the downstream dataset's vocabulary to obtain descriptions. Next, instead of the traditional approach of encoding class names and computing similarities with images, we encoded the description corresponding to the class name (omitting the class name itself) to obtain the text representation and computed similarities with images. The class corresponding to the text 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.

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 [16]	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

4.5. Recognizing Novel Examples

It is desirable for open-vocabulary models to generalize to novel, unseen examples at test-time without requiring re-training. 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 this experiment, we aim to test our model's ability in recognizing novel entities that were absent from the training distribution. Toward this end, we collect 1500 images of products launched after 2023, manually filter these images for quality control and label them into 27 novel categories to form a new benchmark dataset. We call this the **Products-2023** dataset. These concepts are absent from our model's training distribution making them novel. We provide additional details on this dataset in Appendix. We evaluate our model along with CLIP and LLaVA on this dataset in Table 5. Again, LLaVA is evaluated using the same strategy described in Section 4. We observe superior results achieved by our approach against CLIP and even against the much larger LLaVA model confirming the efficacy of our approach in recognizing novel samples.

Accuracy (%)	**Products-2023**
CLIP	33.65
LLaVA	42.08
GRAIN	**45.24**

Table 5

4.6. 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. The outcomes of these ablations are reported as top-1 accuracy in Table 4.

Ablating the region-description alignment loss. This component is crucial to our framework as removing it causes an accuracy drop of 5% on all datasets on average. 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.

Ablating the role of test-time descriptions. In line with the approach of Menon & Vondrick [16], 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.

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. We demonstrate superior performance across 11 different classification datasets, including ones containing fine-grained and novel examples, 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 outperform LLaVA, which is over 13B parameters (compared to our $\sim$170M) and was trained on billions of data-points.

Acknowledgements

This material is based upon work partially supported by the National Science Foundation under Grant No. 2239292.

References

[1] Ioana Bica, Anastasija Ilić, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A. Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, and Jovana Mitrović. Improving fine-grained understanding in image-text pre-training, 2024. 2

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 3, 4

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 3

[5] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification, 2023. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 6

[7] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. In *NeurIPS*, 2023. 3

[8] Li Fei-Fei, Rob Fergus, and Pietro Perona. 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*, pages 178–178. IEEE, 2004. 7

[9] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *International Conference on Computer Vision*, 2023. 1

[10] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3922–3931, 2021. 2

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. 2

[12] Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, and Meng Cao. Veclip: Improving clip training via visual-enriched captions, 2024. 3

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[15] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. 1

[16] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3, 5, 6, 8

[17] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[18] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training, 2021. 2

[19] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland, 2022. Association for Computational Linguistics. 2

[20] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 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*, pages 15691–15701, 2023. 1, 3, 5, 6

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. 1

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5, 6

[23] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models, 2023. 2

[24] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training

next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2, 3, 6

[25] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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)*, pages 2556–2565, 2018. 2, 3, 6

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 5

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Cub. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1

[29] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. 6

[30] Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models, 2025. 2

[31] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training, 2021. 2

[32] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023. 2

[33] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 2

[34] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification?, 2024. 7

[35] Chenyang Zhao, Kun Wang, Janet H. Hsiao, and Antoni B. Chan. Grad-eclip: Gradient-based visual and textual explanations for clip, 2025. 2