Grounding Descriptions in Images informs Zero-Shot Visual Recognition (Supplementary Material)

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

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 [26] and ALIGN [10] 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 [22, 37] 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 [35] 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 [9] 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 [30] or LAION [29] 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 [19, 25]) 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 [19] 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 [25] and Ren et. al. [28] use LLMs to generate descriptions in the form of long, cohesive sentences or via nuanced, hierarchy-aware comparisons. TAP [21] 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.

Improving fine-grained alignment with CLIP. Concurrent to our work, FG-CLIP [34] leverages synthetic data towards improving multimodal alignment achieving strong results. Our training recipe and test-time strategy is focused on achieving stronger zero-shot performance for recognizing fine-grained and novel entities through the use of complementary information available at test time. FG-CLIP is initialized from CLIP weights and trained on billion-scale images in addition to a specifically curated fine-grained data mixture, making direct comparison difficult. Our work does not utilize any additional data in terms of fine-grained or "hard" images and only uses the same images that other baselines like CLIP use. FineCLIP [11] and RO-ViT [13] focus on improving alignment for dense prediction tasks

*Correspondence to shalbe9@gatech.edu

Table A. Zero-shot top-1 accuracy (%) of different methods using the **ViT-L/14 backbone**.

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
CC12M	CLIP [26]	73.36	38.06	49.96	4.59	21.84	43.98	71.79	22.01	7.72	33.16	42.25	37.15	36.72
	Menon&Vondrick [19]	73.74	38.48	50.05	5.22	22.04	44.33	72.56	22.10	8.28	33.78	43.04	37.60	36.84
	CuPL [25]	73.53	38.55	50.46	5.14	21.96	43.28	73.08	23.12	8.65	32.48	42.96	37.56	37.05
	GRAIN (Ours)	**81.62**	**44.98**	**55.82**	**9.12**	**27.66**	**52.98**	**82.05**	**28.18**	**12.73**	**37.34**	**46.92**	**43.58**	**42.68**

like open-vocabulary object detection and do not evaluate on fine-grained/novel classification like ours.

Object-aware Vision-Language Pretraining. Encouraging object-oriented representations within a vision-language pretraining objective [2, 4, 5, 16, 18, 31, 38] 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 [3, 12, 15] 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 [6, 8] 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.

Multimodal Large Language Models. MLLMs like LLaVA [17], GPT-4V [23], Mini-GPT4 [39] integrate image tokens into LLMs, leveraging their powerful reasoning capabilities. MLLMs has been found useful in tasks such as scene understanding [27], story-telling [7] 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 [14, 24] have proposed using side information like attributes, hierarchical representations etc. to learn a generalizable mapping. More recent efforts [33, 36] 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.

B. Implementation Details

All baselines reported in the main paper (except LLaVA) utilize a ViT-B/16 architecture as the vision encoder. In Table A, we report results using the **ViT-L/14** architecture trained on CC12M. For encoding text, we utilize a 12-layer transformer network as used with CLIP [26]. 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 downscaled 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

In Table A, we report performance of GRAIN and competing baselines when using the **ViT-L/14** transformer back-

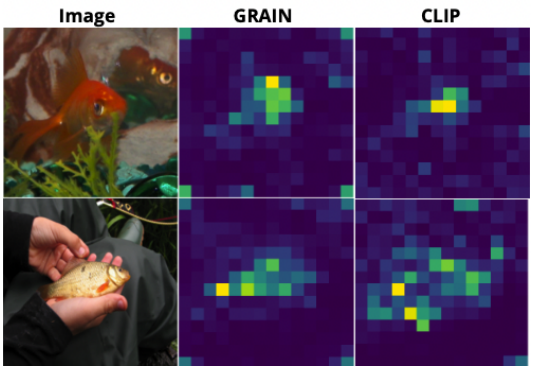


Figure A. Attention maps show more effective object localization by our model compared to CLIP.

bone. It shows that our approach is able to consistently outperform all baselines under various Vision Transformer backbones.

Further, as an ablation experiment, We report results by varying the number of decoder layers for GRAIN. The results are when the model is trained on CC3M and evaluated on Imagenet.

#layers	4	6	12
Accuracy (%)	21.52	23.34	24.08

Table B. Varying #layers.

D. Baselines

CLIP. We train the same ViT variant on the same Conceptual Captions (CC3M and CC12M) and LAION-50M datasets as our GRAIN model. For performing zero-shot testing on all reported datasets, we use the handcrafted prompts specific to each dataset as introduced in the official code-base [26]. 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 [19] to report performance for this baseline. In the main paper, we implement this baseline on top of the CLIP method as per the norm.

CuPL. Similarly, we implement the CuPL baseline leveraging official code [25] 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 + CLIP. We use a pretrained LLAVA v1.6 checkpoint from huggingface [32] 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 C. Next, we use a pretrained CLIP 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. We call this method **LLaVA + CLIP**. 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.

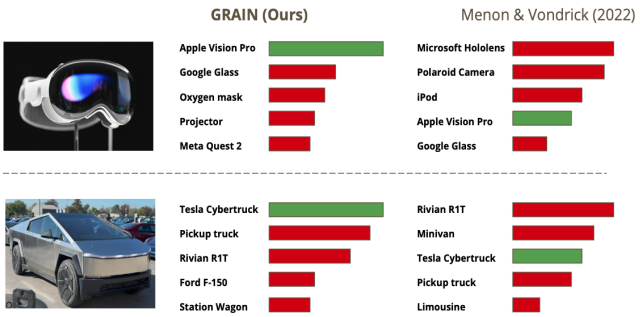


Figure B. Visualization of top-5 predictions of our model on novel entities alongside [19]. Our method consistently identifies the ground truth class as the top prediction.

FILIP. Although FILIP [35] 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.

E. Qualitative Analysis

Visualizing Attention Maps. We visualize attention maps of the penultimate encoding layer for our GRAIN model

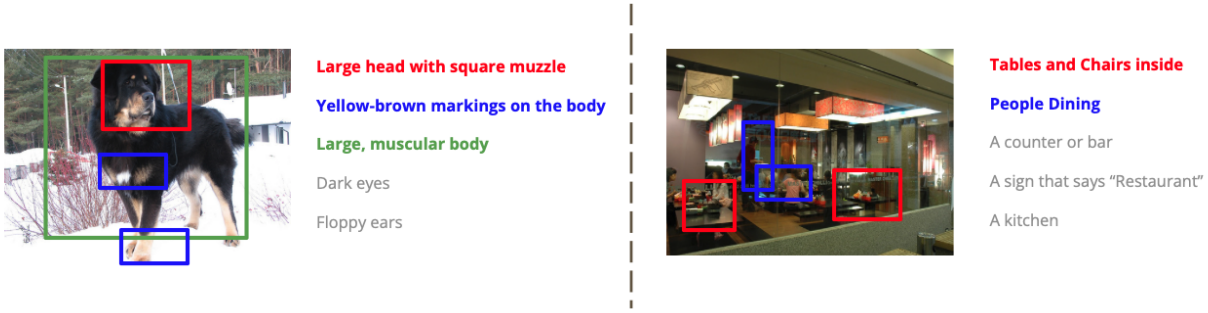


Figure C. Localization and region-description matching predictions made by our model on images from ImageNet.

and CLIP in Figure A. Our model is seen to effectively focus on the object regions in the image which stems from our localization and alignment objectives.

Recognizing Novel Classes. In this 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 the inference-time procedure discussed in the main paper, we present the top-5 predictions made by both our model and [19] in Figure B. 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.

Description Grounding. To showcase the efficacy of our grounding module, we present visualizations of its predictions in Figure C, with images from the Imagenet dataset. 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.

F. Pretraining Dataset Details

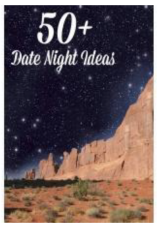
We train all models on the CC3M, CC12M and LAION-50M datasets. As explained earlier, to obtain description and localization annotations, we prompt LLaVA in two stages to extract the primary visual subject of the image and then gather image descriptions by asking LLaVA to focus on the identified visual subject. We obtain bounding boxes corresponding to each description by using OWLv2 [20], an off-the-shelf open-vocabulary object detector. We filter the predicted boxes using a confidence threshold of 0.3 to discard noisy predictions.

G. Products-2023 Dataset Details

To evaluate our approach's ability to recognize novel samples, we manually curated a dataset comprising 1,500 images spanning 27 distinct categories. These images were carefully filtered and labeled through a manual process. Specifically, we compiled a list of products launched after 2023, scraped corresponding images, and performed manual filtering and labeling. Since the pretraining datasets used in our setup (CC3M, CC12M, LAION) were finalized prior to 2024, this dataset represents novel examples. The full list of categories includes: [*Apple Vision Pro*, *Grimace Shake(drink from McD)*, *Starry (drink from Pepsi)*, *Playstation Portal*, *Apple Watch Ultra 2*, *Apple Watch Series 9*, *Samsung Galaxy Watch 6*, *Xiaomi Smart Band 8*, *Kia K4*, *Rivian R2*, *Honda Ye GT*, *Ferrari 12Cilindri*, *Renault 5 E-Tech*, *Toyota Tundra*, *Ford F-150 Lightning*, *Tesla Cybertruck*, *Xiaomi SU7*, *Lamborghini Revuelto*, *Hyundai Mufasa*, *Wordle*, *Asus ROG Ally*, *Meta Quest Pro*, *Microsoft Hololens*, *Google Glass*, *Prime (drink)*, *iPhone 15*, *Google Pixel 8*]. We intend to release this dataset with the final version of this paper.

H. Sample Annotations

In Figure D, we illustrate the annotations obtained using our two-stage LLaVA prompting followed by bounding box prediction using OWLv2. We randomly select im-



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

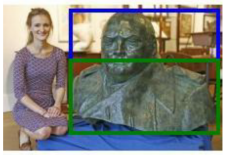
Descriptions: Arched entrance, Red roof, White walls, Multiple Windows, Fountain

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

Figure D. Sample annotations generated using our two-stage LLaVA prompting scheme followed by OWLv2 localization.

Table C. 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 {}

ages 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.

I. 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 E, 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.

J. 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 [19] 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 F.

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.

K. 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. Further integration of our approach with retrieval-based systems and multimodal LLMs is an interesting future direction.

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. 1, 3

[2] Kaylee Burns, Zach Witzel, Jubayer Ibn Hamid, Tianhe Yu, Chelsea Finn, and Karol Hausman. What makes pre-trained visual representations successful for robust manipulation?, 2023. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 2

[4] Brian Chen, Nina Shvetsova, Andrew Rouditchenko, Daniel Kondermann, Samuel Thomas, Shih-Fu Chang, Rogerio Feris, James Glass, and Hilde Kuehne. What, when, and where? – self-supervised spatio-temporal grounding in untrimmed multi-action videos from narrated instructions, 2023. 2

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020. 2

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

[7] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023. 2

[8] 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. 2

[9] 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. 1

[10] 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. 1

[11] Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. Fineclip: Self-distilled region-based clip for better fine-grained understanding. In *Advances in Neural Information Processing Systems*, pages 27896–27918. Curran Associates, Inc., 2024. 1

[12] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr – modulated detection for end-to-end multi-modal understanding, 2021. 2

[13] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers, 2023. 1

[14] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009. 2

[15] Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries, 2021. 2

[16] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks, 2020. 2

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[18] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. 2

[19] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3, 4, 6

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

[21] M. Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Horst Possegger, Rogerio Feris, and Horst Bischof. Tap: Targeted prompting for task adaptive generation of textual training instances for visual classification, 2023. 1

[22] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training, 2021. 1

[23] OpenAI. Gpt-4v(ision) system card. *OpenAI*, 2023. 2

[24] Devi Parikh and Kristen Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510, 2011. 2

[25] 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, 2, 3

[26] 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. 1, 2, 3

[27] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 2

[28] Zhiyuan Ren, Yiyang Su, and Xiaoming Liu. Chatgpt-powered hierarchical comparisons for image classification, 2023. 1

[29] 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. 1

[30] 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. 1

[31] Alex Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval, 2022. 2

[32] 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. 3

[33] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5542–5551, 2018. 2

[34] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment, 2025. 1

[35] 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. 1, 3

[36] Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. Zerogen: Efficient zero-shot learning via dataset generation. 2022. 2

[37] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 1

[38] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Helping hands: An object-aware ego-centric video recognition model, 2023. 2

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

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

Figure E. Qualitative comparison between one-stage (middle) and two-stage (right) LLaVA-based annotation schemes.

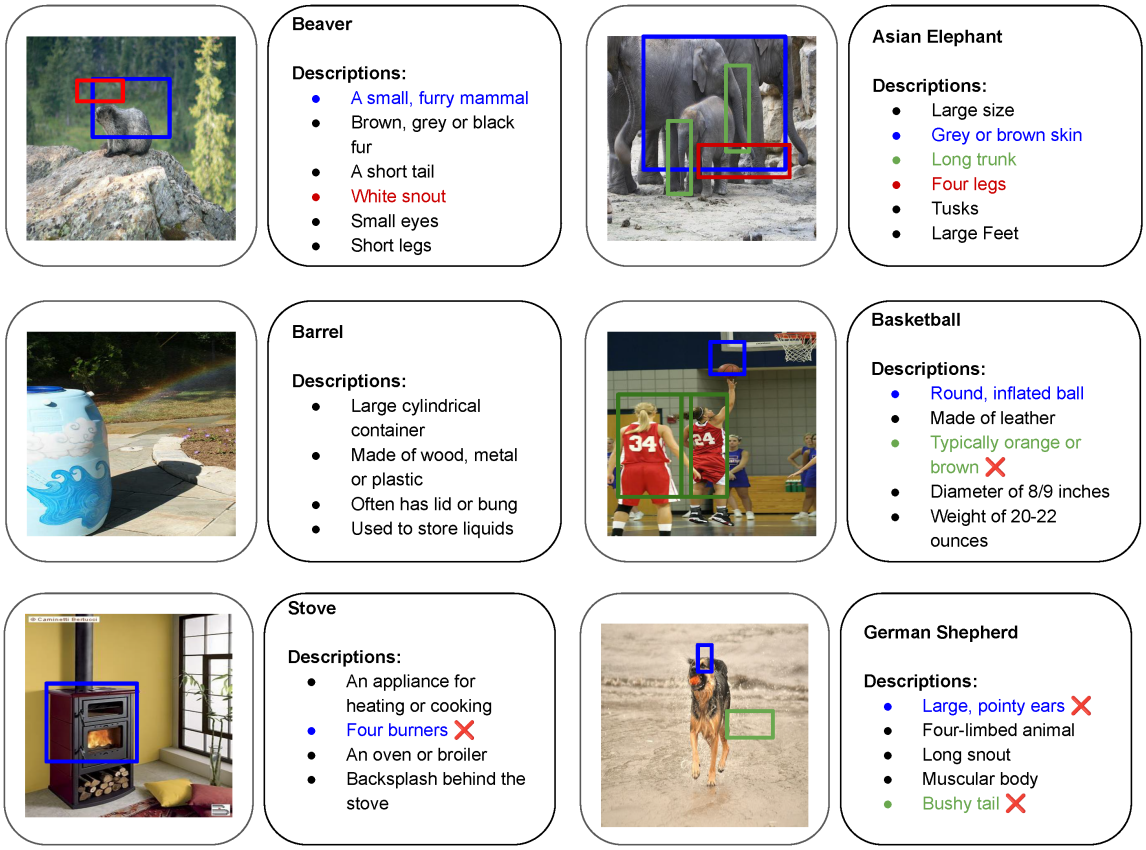


Figure F. Visualization of failure modes from our grounding module on ImageNet-1K.