Can Large Vision Language Models Ground Fine-Grained Attribute?

Abhilash Neog
Virginia Tech
abhilash22@vt.edu

Gaurav Srivastava
Virginia Tech
gks@vt.edu

Aafiya Hussain
Virginia Tech
aafiyahussain@vt.edu

Kazi Sajeed Mehrab
Virginia Tech
ksmehrab@vt.edu

Abstract

Recent advancements in Large Vision-Language Models (LVLMs) such as InstructBLIP, LLaVA, and CogVLM have demonstrated strong zero-shot performance in generating text grounded in images and solving complex visual tasks. However, these models still struggle with fine-grained visual attribute localization, a key challenge in fine-grained visual classification (FGVC). In this work, we investigate the potential of instruction-tuned LVLMs like GLAMM and MoLMO to localize fine-grained attributes by analyzing text-to-image attention maps. We hypothesize that LVLMs possess latent visual grounding capabilities, even if not reflected in their textual outputs. Our approach utilizes attention map intersections as an implicit grounding method to enhance attribute localization and provide insights into improving LVLM grounding through human-guided attention during instruction tuning.

1. Introduction

There has been a notable increase in interest and advancement in Large Vision-Language Models (LVLMs) capable of generating text grounded in images. Models like [10] InstructBLIP, LLaVA [21] and CogVLM [32] have demonstrated impressive zero-shot performance in generating image captions, conducting visual reasoning, producing textual descriptions and solving complex question-answering tasks. The strong results observed across various benchmarks suggest that these models, primarily based on large language models (LLMs) such as Vicuna [8], Flan-T5 [9], and Llama [30], are well-equipped to exploit the relationship between the textual knowledge gained during pre-training and the image understanding developed during instruction tuning. Additionally, these models display robust zero-shot transferability to a wide range of downstream tasks.

Traditionally, research in fine-grained visual classification (FGVC) [6], has focused on accurately classifying a wide variety of images, including different species of birds, plants, animals, and artificial objects like cars. FGVC poses a greater challenge because it requires the recognition of subtle distinctions in images, such as variations in **eye shape, flipper morphology, and tail coloration** among birds. A significant challenge for visual models is the precise *localization of these fine-grained attributes.*

Given the visual-text reasoning and understanding capabilities of Large Vision-Language Models (LVLMs), an important question arises: *Can the knowledge acquired by these models during large-scale training be utilized to localize or ground textual descriptions of fine-grained attributes with the corresponding visual pixel space?* Recent studies, such as VLM4Bio [22] and Finer [17], have explored the ability of general-purpose LVLMs to ground or classify the presence of fine-grained attributes in images. These studies demonstrate that while LVLMs perform well with coarse-grained objects, such as birds, they struggle to accurately ground or identify fine-grained attributes. However, these investigations do not consider more recent LVLMs like GLAMM [27] and MoLMO [11], which are specifically instruction-tuned for grounding textual phrases with the visual pixel space. Furthermore, existing analyses are primarily limited to the generated token space, focusing on whether LVLMs can produce the correct textual responses for grounding or identifying fine-grained attributes.

In this work, we first aim to investigate LVLMs such as GLAMM, MoLMO, and CogVLM2, which have been specifically instruction-tuned to segment, point, or ground visual objects to language phrases, assessing their capability to handle fine-grained attributes. We hypothesize that even general-purpose LVLMs, like LLaVA and InstructBLIP, may possess sufficient visual grounding capabilities to internally localize or ground fine-grained attributes, despite potentially failing to reflect this in their textual responses.

To explore this, we propose a framework that prompts LVLMs in a visual question-answering [3] format to identify fine-grained attributes. Rather than solely relying on the generated textual responses for grounding, we will analyze the intersection of text-to-image attention maps within the LVLMs to determine if the models focus on the relevant fine-grained attributes. If the attention maps are correctly localized, our approach can serve as an implicit grounding method using LVLMs. Conversely, if the attention maps

do not correspond to the correct regions, this discrepancy can offer valuable insights for enhancing LVLM grounding related to attention mapping. We propose that such improvements could be obtained through human-guided attention techniques during few-shot instruction tuning of the LVLMs, which may serve as a future direction of our work.

2. Related Work

Recent advances in Vision-Language Models (VLMs) have significantly improved the ability to understand and align visual and textual information. Foundational models like CLIP [26] and ALIGN [16] have achieved robust vision-language representations through large-scale pretraining. These early models were further enhanced with architectures like UNITER [7] and VinVL [37], which leverage object-semantics alignment for nuanced understanding. Oscar [19], in particular, introduced object-level semantics to enhance model comprehension of specific visual attributes.

2.1. General-purpose Large Vision Language Models

Models like LLaVA and Instruct-BLIP demonstrate strong zero-shot performance across various vision-language tasks, such as Visual Question Answering (VQA), reasoning, and image captioning. These models generate outputs informed by learned representations and image analysis, enabling coherent reasoning. Despite these capabilities, fine-grained object identification remains a challenge. Recent work, such as Finer [17], identifies a modality gap in these models' ability to detect fine-grained attributes. Similarly, VLM4Bio [22] tested these models using bounding boxes, highlighting their limitations in grounding tasks. Also, [25] present KOSMOS-2, a multimodal model that leverages grounded image-text pairs to enhance phrase grounding, VQA, and captioning tasks, marking a step toward more integrated multimodal systems. These findings suggest that most general-purpose LVLMs lack fine-tuning for fine-grained attribute grounding.

2.2. LVLMs with Grounding Abilities

Recent LVLMs like CogVLM2 [15], GLAMM [27], and MoLMO [11] have advanced grounding abilities for textual phrases within images. CogVLM2, for instance, is trained on the LAION-40M-grounding dataset [28], which includes bounding box annotations, allowing for improved object localization. GLAMM has been specifically instruction-tuned to map text phrases to pixel locations, while MoLMO utilizes landmark points to enhance object identification within pixel space. However, these models have yet to be comprehensively evaluated for fine-grained attribute grounding.

2.3. Multi-Scale Attention Mechanisms

In multi-scale attention research, VL-SAM [20] provides a training-free approach that combines generalized object recognition and localization models using attention maps as prompts for segmentation. As observed in VL-InterpreT [2], early attention layers produce broad, diffused maps, which become more specific in later layers. Additionally, Chefer et al. [5] indicate that deeper layers convey more semantic information, which aligns with the dual-scale approach of capturing both general and specific features.

DUAL ATT-NET [34] introduces hard and soft attention mechanisms, enhancing fine-grained recognition in few-shot scenarios. Sun et al. [29] propose Multi-scale Attention Fusion, which integrates local and global features through a dual-stream network. Ouyang et al. [24] further refine multi-scale attention in the Efficient Multi-Scale Attention (EMA) module by capturing both short- and long-range dependencies, improving spatial understanding.

2.4. Visual Grounding

In visual grounding, [14] introduce Parameter-efficient Fine-tuning for Medical Visual Grounding (PFMVG), a two-stage fine-tuning process for medical image captioning and grounding. Bhowmik et al. [4] utilize a Dual Mixture of Experts (MoE) to balance grounding with image-language comprehension. Rasheed et al. [27] develop GLaMM, which achieves dense, pixel-level grounding using a hierarchical feature extraction framework. VideoGLaMM [23] expands this approach to video grounding by using a spatio-temporal pixel decoder, generating object masks based on user queries.

2.5. Fine-Grained Visual Grounding

HiVG, proposed by [33], is a hierarchical multimodal framework that bridges visual and linguistic features for fine-grained visual grounding. Using adaptive cross-modal bridges, HiVG improves cross-modal alignment, achieving superior results across benchmarks. ViGoR by [35] enhances grounding in LVLMs with fine-grained reward modeling, incorporating feedback to reduce visual hallucinations and improve grounding accuracy. Dey et al. [12] propose AsphaltNet, a 3D grounding model utilizing offset and span losses to promote verbo-visual fusion for accurate object localization.

Furthermore, [27] advance fine-grained visual perception with "AnyRef," an MLLM model capable of generating pixel-level segmentations across modalities. Kirillov et al. [18] introduce a framework for progressively capturing multi-granular features, while [36] propose MMAL-Net, a multi-branch, multi-scale model that achieves fine-grained categorization without bounding box annotations.

2.6. Attention Flow in Transformers

In attention flow, [1] quantify information flow across layers by modeling attention as a directed graph. This method provides higher interpretability by correlating attention weights with ablation scores, improving diagnostic insights.

3. Proposed Methodology

3.1. Multi-Scale Attention Flow with Progressive Sharpening

Transformers have been shown to capture hierarchical representations across layers, with early layers focusing on broad, coarse-grained regions and later layers refining these regions to capture fine-grained details [5]. Building on this insight, we propose a *multi-scale attention flow* approach that leverages this progression to ground attention at both object and attribute levels.

In our approach, attention flows are generated separately for a selected coarse-grained token and a fine-grained token. Coarse-grained attention maps are derived from early layers, where attention naturally tends to capture general object-level features. In contrast, fine-grained attention maps are extracted from deeper layers, where attention focuses on specific attributes or regions. This multi-scale structure allows the model to capture both a high-level understanding of the scene and precise details within it.

To capture information at both coarse and fine scales, we employ a *multi-scale attention flow* that progressively refines the attention from general object regions to specific attributes across the layers of the model. This approach leverages the natural progression of transformers, where earlier layers capture high-level information, and later layers focus on finer details [5]. We achieve this by computing and aggregating attention maps across layers, using attention rollout to establish a cohesive attention flow for each scale.

Attention Rollout Attention rollout, as introduced by [5], aggregates attention across multiple layers to form a cumulative attention map that reflects the model's overall focus for each token. Given the attention matrix $A^{(l)} \in \mathbb{R}^{n \times n}$ at layer l, which represents the attention scores among n tokens, the cumulative attention rollout R for each token is computed as:

$$R = \prod_{l=1}^{L} (I + A^{(l)}) \tag{1}$$

where I is the identity matrix, and L denotes the number of layers. The identity matrix I ensures that the initial focus is on the input tokens themselves, and the product over layers accumulates the attention across all layers. This rollout gives a final attention map for each token, capturing both direct and indirect connections through the layers.

Coarse-Scale Flow For the coarse-grained token, which captures high-level object information, we extract attention maps from the earlier layers, where attention tends to focus on larger, general regions. By applying attention rollout on the initial layers up to layer L_c, we compute a cumulative coarse-scale attention map R^{coarse} as follows:

$$R^{\text{coarse}} = \prod_{l=1}^{L_c} (I + A^{(l)}) \tag{2}$$

This coarse-scale attention flow provides a broad focus that aligns with the general object's region, serving as an initial map for grounding the coarse-grained token.

Fine-Scale Flow For the fine-grained token, we seek a more precise focus in later layers. We compute the cumulative attention rollout from layer L_f to the final layer L to obtain a fine-scale attention map R^{fine}, capturing attribute-level details:

$$R^{\text{fine}} = \prod_{l=L_f}^{L} (I + A^{(l)}) \tag{3}$$

This fine-scale attention map provides a concentrated focus on the specific attribute region, refining attention to finer details as required by the fine-grained token.

To further enhance the model's ability to focus on fine-grained attributes, we apply *progressive sharpening* on the fine-grained token's attention maps at each successive layer. This sharpening mechanism reduces the spread of attention, concentrating it within the target attribute's region as the model progresses through layers. The entropy of the fine-grained token's attention map is minimized at each layer to encourage this compact and localized focus.

To achieve this, we apply *progressive sharpening* on the fine-grained token's attention maps at each successive layer, concentrating attention in finer regions. Progressive sharpening is implemented by minimizing the entropy of the "beak" attention map in deeper layers, thus encouraging a compact and localized attention flow.

$$\text{Entropy}(A_l^{\text{beak}}) = -\sum_i A_{l,i}^{\text{beak}} \log(A_{l,i}^{\text{beak}}) \tag{4}$$

The *progressive sharpening loss* $L_{\text{sharpening}}$ across layers is defined as:

$$L_{\text{sharpening}} = \sum_l \text{Entropy}(A_l^{\text{beak}}) \tag{5}$$

By minimizing this loss, we ensure that the fine-grained attention for "beak" becomes increasingly concentrated as the model progresses to deeper layers.

3.2. Hierarchical Consistency Loss for Coarse-to-Fine Attention Alignment

To maintain consistency between the coarse-grained focus on "bird" and the fine-grained focus on "beak", we enforce a *hierarchical consistency loss*. This loss ensures that the fine-grained attention map for "beak" is spatially contained within the broader, coarse-grained attention region of "bird".

For each layer l, let A_l^{bird} and A_l^{beak} represent the attention maps for "bird" and "beak" respectively. The hierarchical consistency loss $L_{\text{hierarchical-consistency}}$ is formulated as follows:

$$L_{\text{hierarchical-consistency}} = \sum_i \left(A_i^{\text{beak}} \cdot (1 - A_i^{\text{bird}}) \right) \quad (6)$$

This term penalizes attention in "beak" that falls outside the coarse-grained region of "bird", ensuring that fine-grained attention remains within the boundaries of the coarse-grained focus.

3.3. Weighted Head Aggregation for Focused Attention

Each layer consists of multiple attention heads, and attention may vary across these heads. To produce a coherent attention map at each layer, we aggregate attention across heads in a weighted manner.

Head Compactness Weighting We prioritize heads with more focused (compact) attention flows by assigning higher weights to heads with lower entropy (indicating less spread). The weight for head h in layer l, $w_{l,h}$, is calculated based on the entropy of the attention distribution for that head:

$$w_{l,h} = \frac{1}{\text{Entropy}(A_{l,h}) + \epsilon} \quad (7)$$

where ϵ is a small constant to prevent division by zero. The aggregated attention map A_l for layer l is then computed as:

$$A_l = \sum_h w_{l,h} \cdot A_{l,h} \quad (8)$$

This weighted aggregation ensures that attention heads with compact focus are emphasized, producing a cleaner and more interpretable attention map for each layer.

In this work, we focus on a single coarse-to-fine pair (example, "bird" and "beak") to validate our hierarchical and progressive attention approach. Future work will explore grouping and pairing strategies for handling multiple tokens in complex descriptions. We propose using part-of-speech tagging to categorize tokens into coarse and fine-grained groups, followed by a pairing mechanism based on hierarchical relationships. Additionally, a scalable weighting approach for head aggregation will be extended to address larger token sets in diverse contexts.

4. Experimental Setup

4.1. Datasets

To perform fine-grained visual classification we will use richly annotated datasets which consist of samples from diverse species for a given animal class.

4.1.1 NABirds

The NABirds dataset contains 48,562 images of North American bird species across 555 categories [13]. Each image is annotated with part annotations and bounding boxes. The dataset was curated with the help of citizen scientists, experts from the Cornell Lab of Ornithology, and Mechanical Turk workers, who provided annotations for bird parts and bounding boxes. Citizen experts were identified using eBird and could use a tool to give labels.

The images are organized based on species taxonomy and include diverse visual categories, such as different bird plumages, sex, and age attributes, making it particularly suitable for challenging computer vision tasks. Citizen expert intervention is shown to improve the quality of labels over other datasets that are annotated only by Mechanical Turk workers.

4.1.2 CUB-200-2011

The Caltech-UCSD Birds-200-2011 (CUB) [31] dataset includes 11,788 images across 200 bird species, each annotated with detailed information such as 15 part locations, 312 binary attributes, and bounding boxes. Attributes are divided into 28 different groups and divided into 312 binary attributes. Part locations were annotated using pixel location. These annotations were done by Mechanical Turk workers.

4.2. Evaluation Metrics

We plan to use quantitative and qualitative metrics to evaluate the performance of multimodal fine-grained attribute classification and localization. In the first step, we will focus on the accuracy of our localization approach comparing the predicted locations of fine-grained attributes with the ground truth bounding boxes. Intersection over Union (IoU) is used as the primary metric for this. For the attribute classification task, we will use standard metrics such as Precision, Recall, and F1-score to measure the model's ability to distinguish between fine-grained attributes.

As part of our evaluation, we will also conduct a comparative analysis of different Vision-Language Models (VLMs) and measure each model's performance in zero-shot classification tasks. Also, we plan to include a qualitative analysis by visually inspecting the results of the localization task. Additionally, we will analyze failure cases to understand the model's limitations better and identify areas for improvement.

4.3. Results/Inital Observations

Our initial experiments reveal several noteworthy patterns in the attention distributions across layers of the Vision-Language Model (VLM) when attempting to ground fine-grained attributes. These observations highlight key differences in the attention mechanisms at various stages of the model, underscoring potential areas for refining attribute localization.

- **Intermediate and Final Layers – Attention Drift from Target Attributes:** In the intermediate and final layers, attention does tend to concentrate around fine-grained attributes, such as the beak of a bird. However, the highest attention often peaks outside the actual attribute region and, in many cases, even falls outside the boundary of the coarse-grained object (e.g., outside the bird). As shown in Figure 3, the regions with the highest attention (highlighted in red) frequently miss the attribute of interest. This drift suggests that while deeper layers start to capture attribute-specific details, they do not yet fully align with the precise localization required for fine-grained grounding.

- **Early Layers – Dispersed and Contextual Attention:** The attention in the first few layers is notably diffuse and appears haphazard, which aligns with the model's tendency to incorporate broad contextual information at early stages. As seen in Figures 3 to 11 (particularly Figures 10 and 11), attention in these initial layers lacks a coherent focus on specific objects or attributes. Given this observation, disregarding attention from the early layers is beneficial when localizing fine-grained attributes, as their role is more suited to establishing general context rather than precise localization.

- **Coarse-Grained Attention Regions as Constraints for Fine-Grained Attributes:** By examining attention directed from the coarse-grained object token (e.g., 'bird') to the image, we observe that this attention spans broader regions within the visual input (Figures 16 to 18). These broader attention maps can serve as valuable constraints for localizing fine-grained attributes. Specifically, by enforcing that the attention for a fine-grained attribute, like 'beak,' falls within the boundaries of the coarse-grained object attention, we can significantly narrow down the search space, potentially improving localization accuracy. This hierarchical constraint approach could reduce misalignment and bring the model's attention closer to the actual attribute regions.

4.4. Group logistics

Our group will collaborate regularly to ensure smooth communication and efficient progress on the project. We will use a combination of tools such as Slack and email for day-to-day communication. Weekly meetings will be scheduled on Zoom to discuss the project's status, troubleshoot any challenges, and ensure that we are on track with our goals. Additional meetings will be scheduled as needed, particularly during key phases of implementation and experimentation.

Each member will take responsibility for different aspects of the project based on their expertise:

1. **Kazi Sajeed Mehrab** will lead the attention map analysis and assist in integrating different Vision-Language Models (VLMs) into the framework.

2. **Abhilash Neog** will be responsible for developing the framework to analyze attention maps and ensure proper evaluation metrics are implemented.

3. **Gaurav Srivastava** will handle the comparative analysis of LVLMs, focusing on how well they perform in localizing fine-grained attributes.

4. **Aafiya Hussain** will lead the dataset preparation and assist in running experiments on fine-grained visual classification datasets such as CUB-200.

All members will contribute to writing the report, reviewing results, and preparing the final presentation. We plan to hold progress updates weekly, ensuring everyone is aligned and contributing effectively.

References

[1] S Abnar and W Zuidema. Quantifying attention flow in transformers. arxiv 2020. *arXiv preprint arXiv:2005.00928*, 2022. 3

[2] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 21406–21415, 2022. 2

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[4] Aritra Bhowmik, Mohammad Mahdi Derakhshani, Dennis Koelma, Martin R Oswald, Yuki M Asano, and Cees GM Snoek. Learning to ground vlms without forgetting. *arXiv preprint arXiv:2410.10491*, 2024. 2

[5] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021. 2, 3

[6] Huazhen Chen, Haimiao Zhang, Chang Liu, Jianpeng An, Zhongke Gao, and Jun Qiu. Fet-fgvc: Feature-enhanced transformer for fine-grained visual classification. *Pattern Recognition*, 149:110265, 2024. 1

[7] 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. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 1

[9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 1

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1

[11] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1, 2

[12] Sombit Dey, Ozan Unal, Christos Sakaridis, and Luc Van Gool. Fine-grained spatial and verbal losses for 3d visual grounding, 2024. 2

[13] Qishuai Diao, Yi Jiang, Bin Wen, Jia Sun, and Zehuan Yuan. Metaformer: A unified meta framework for fine-grained recognition. *arXiv preprint arXiv:2203.02751*, 2022. 4

[14] Jinlong He, Pengfei Li, Gang Liu, and Shenjun Zhong. Parameter-efficient fine-tuning medical multimodal large language models for medical visual grounding. *arXiv preprint arXiv:2410.23822*, 2024. 2

[15] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 2

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[17] Jeonghwan Kim and Heng Ji. Finer: Investigating and enhancing fine-grained visual concept recognition in large vision language models. *arXiv preprint arXiv:2402.16315*, 2024. 1, 2

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[19] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 2

[20] Zhiwei Lin, Yongtao Wang, and Zhi Tang. Training-free open-ended object detection and segmentation via attention as prompts. *arXiv preprint arXiv:2410.05963*, 2024. 2

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[22] M Maruf, Arka Daw, Kazi Sajeed Mehrab, Harish Babu Manogaran, Abhilash Neog, Medha Sawhney, Mridul Khurana, James P Balhoff, Yasin Bakis, Bahadir Altintas, et al. Vlm4bio: A benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images. *arXiv preprint arXiv:2408.16176*, 2024. 1, 2

[23] Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Khan, and Salman Khan. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. *ArXiv*, 2024. 2

[24] Daliang Ouyang, Su He, Guozhong Zhang, Mingzhu Luo, Huaiyong Guo, Jian Zhan, and Zhijie Huang. Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[25] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2

[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. 2

[27] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 1, 2

[28] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo

Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[29] Huixin Sun, Yunhao Wang, Xiaodi Wang, Bin Zhang, Ying Xin, Baochang Zhang, Xianbin Cao, Errui Ding, and Shumin Han. Maformer: A transformer network with multi-scale attention fusion for visual recognition. *Neurocomputing*, 595:127828, 2024. 2

[30] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[31] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4

[32] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1

[33] Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. Hivg: Hierarchical multimodal fine-grained modulation for visual grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5460–5469, 2024. 2

[34] Shu-Lin Xu, Faen Zhang, Xiu-Shen Wei, and Jianhua Wang. Dual attention networks for few-shot fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2911–2919, 2022. 2

[35] Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. *arXiv preprint arXiv:2402.06118*, 2024. 2

[36] Fan Zhang, Meng Li, Guisheng Zhai, and Yizhao Liu. Multi-branch and multi-scale attention learning for fine-grained visual categorization. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part I 27*, pages 136–147. Springer, 2021. 2

[37] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 2