Can Large Vision Language Models Ground Fine-Grained Attribute?

Kazi Sajeed Mehrab
Virginia Tech
ksmehrab@vt.edu

Abhilash Neog
Virginia Tech
abhilash22@vt.edu

Gaurav Srivastava
Virginia Tech
gks@vt.edu

Aafiya Hussain
Virginia Tech
aafiyahussain@vt.edu

Abstract

Recent advancements in Large Vision-Language Models (LVLMs) have demonstrated impressive capabilities in generating visual-text alignments and performing complex vision-language tasks. However, most LVLMs struggle with fine-grained attribute localization, which is essential for precise visual grounding in applications such as fine-grained visual classification. In this work, we propose a novel framework that utilizes dual-scale attention maps to improve fine-grained attribute grounding. By combining coarse-grained and fine-grained attention maps, our approach captures both general object-level and detailed attribute-specific information. Additionally, we introduce an attention amplification mechanism that selectively enhances model focus on regions corresponding to target attributes, such as color or texture. Our method is evaluated across the Caltech-UCSD Birds-200-2011 dataset, showing that the dual-scale approach and targeted amplification significantly improve attribute localization accuracy, achieving higher modified point in region scores.

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 InstructBLIP [9], LLaVA [23], and CogVLM [35] 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 [7], Flan-T5 [8], and Llama [33], 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.

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 [24] and Finer [19], 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 ground or identify fine-grained attributes accurately.

In contrast, traditional research in fine-grained visual classification (FGVC) [5, 15] has placed emphasis on accurately classifying a wide variety of images, encompassing 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 [13, 14]. A significant challenge for visual models is the precise localization of these fine-grained attributes.

In this work, we focus on LVLMs, specifically LLaVA, which demonstrate a strong zero-shot performance for coarse-grained tasks such as captioning or question answering. However, it is observed that their ability to localize fine-grained attributes, such as specific colors, textures, and shapes, remains limited. We hypothesize that although these models may possess latent grounding capabilities, their current structure often fails to localize fine-grained attributes precisely. To address this, we propose a novel framework that leverages both coarse-grained and fine-grained attention maps to analyze whether attention maps produced by LVLMs can implicitly ground fine-grained attributes across multiple scales.

Prior approaches like Finer and VLM4Bio mainly focus on the generated token space to identify modality gaps. Furthermore, approaches such as VL-SAM [22] demonstrate that attention maps from LVLMs can be leveraged to ground coarse-grained objects in images. VL-SAM also uses attention maps from generated tokens to the visual tokens. We go beyond this by analyzing the attention maps from different layers and heads for fine-grained grounding.

Contributions: Our contributions are three-fold:

1. We demonstrate that specific attention maps fail to attend to the expected regions and propose a selection strategy that combines entropy-based filtering with maximally connected component filtering. This approach results in a more meaningful collection of attention maps for fine-grained grounding.

2. We introduce a method that integrates coarse-grained attention rollout with fine-grained attention maps using element-wise multiplication. This integration highlights fine-grained regions, suppresses irrelevant areas, and enhances focus on salient details.

3. We enforce a hierarchical constraint that ensures fine-grained objects (e.g., "eye") are localized within their corresponding coarse-grained objects (e.g., "bird"), improving the consistency of attention maps.

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 [29] and ALIGN [18] have achieved robust vision-language representations through large-scale pretraining. These early models were further enhanced with architectures like UNITER [6] and VinVL [40], which leverage object-semantics alignment for nuanced understanding. Oscar [21], 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 [23] and Instruct-BLIP [9] 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 [19], identifies a modality gap in these models' ability to detect fine-grained attributes. Similarly, VLM4Bio [24] tested these models using bounding boxes, highlighting their limitations in grounding tasks. Also, [28] 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 [17], GLAMM [30], and MoLMO [10] have advanced grounding abilities for textual phrases within images. CogVLM2, for instance, is trained on the LAION-40M-grounding dataset [31], 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 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. [4] 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 [37] introduces hard and soft attention mechanisms, enhancing fine-grained recognition in few-shot scenarios. Sun et al. [32] propose Multi-scale Attention Fusion, which integrates local and global features through a dual-stream network. Ouyang et al. [27] 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, [16] 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. [3] utilize a Dual Mixture of Experts (MoE) to balance grounding with image-language comprehension. Rasheed et al. [30] develop GLaMM, which achieves dense, pixel-level grounding using a hierarchical feature extraction framework. VideoGLaMM [26] 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 [36], 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 [38] 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, [30] advance fine-grained visual perception with "AnyRef," an MLLM model capable of generating pixel-level segmentations across modalities. Kirillov et al. [20] introduce a framework for progressively capturing multi-granular features, while [39] propose MMAL-Net, a multi-branch, multi-scale model that achieves fine-grained categorization without bounding box annotations.

2.6. Attention Flow in Transformers

Abnar et al. [1] introduced attention rollout and attention flow, techniques that model the network as a directed acyclic graph to quantify token contributions across layers. While attention rollout computes cumulative contributions, attention flow uses a max-flow algorithm to measure token influence, addressing the limitations of raw attention weights. Furthermore, Metzger et al. [25] extended this concept to general Transformer architectures, linking attention flow to Shapley value computations for measuring token impact while ensuring positional independence in autoregressive decoders. DeRose et al. [11] further developed a visual analytics tool, Attention Flows, to trace attention dynamics and compare pre-trained and fine-tuned models, providing insights into how attention adapts to downstream tasks.

3. Methodology

The proposed methodology utilizes the LLaVA model to enhance the localization of fine-grained visual attributes by analyzing and refining the attention mechanisms within the Large Vision Language Model. As illustrated in figure 1, the input to LLaVA consists of a structured sequence that begins with a system-defined prompt, followed by visual tokens, and culminating with textual query tokens. These visual and textual tokens are projected into a common representational space, enabling self-attention mechanisms across the LLM layers. By leveraging the self-attention interactions between the textual query tokens and the visual tokens, we extract attention maps that form the basis of our analysis. Our methodology involves a rigorous two-stage filtering process that refines these attention maps, focusing on head selection and layer-wise aggregation to improve the precision of fine-grained attribute localization in images.

3.1. Head Selection and Aggregation (HSA)

As seen from the initial findings in Figure 5, different heads capture different aspects and compute attention differently, resulting in certain heads not localizing (i.e. dispersed attention maps) correctly with the give query object. Therefore, we propose a mechanism to select the most informative attention maps. We hypothesize that an ideal attention map would have high attention points clustered together.

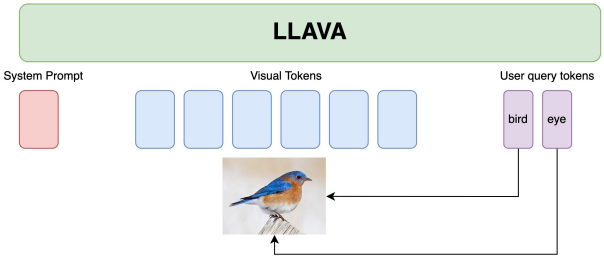


Figure 1. Input to LLaVA consists of a structured sequence that begins with a system-defined prompt, followed by visual tokens, and culminating with textual query tokens.

In HSA, we implement a two-stage filtering process (shown in figure 2) to enhance the attention maps at every layer of the LVLM, elaborated further in the below subsections.

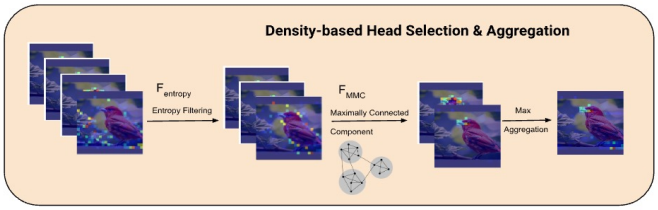


Figure 2. Entropy based filtering across heads followed by maximally connected component selection. Finally union is performed among the remaining attention heads

3.1.1 Entropy-based Filtering

Entropy is used to measure the randomness or dispersion in attention maps. Low entropy usually signifies that attention is concentrated in a small number of regions or elements, indicating focused attention. High entropy indicates that attention is spread across many regions or elements. Entropy (1) can be represented as,

$$H(X) = -\sum_{i} p(x_i) \log p(x_i) \tag{1}$$

where, $H(X)$ represents the entropy of the random variable X, $p(x_i)$ is the probability of the i^{th} state or outcome in the distribution of X

We use entropy to filter out attention heads that show high entropy values, indicating that their attention is too dispersed to be effective for precise localization. This leaves us with maps with more focused attention.

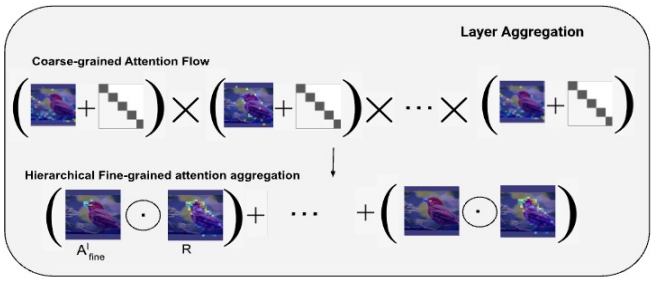


Figure 3. Layer aggregation and hierarchical constraint enforcement

The approach is implemented by computing a normalized entropy score which measures the spread of values in the attention map. If this score is less than a certain threshold, the attention map is filtered out.

3.1.2 Maximally-Connected Component filtering

In the second stage of our filtering pipeline, we employ Maximally Connected Component-Based Filtering. A Maximally Connected Component (MCC) in an attention map can be defined as a subset of the map where every pair of nodes (pixels or regions with high attention scores) is connected directly or indirectly and which is not part of a larger connected cluster. This approach helps to identify the largest clusters within the attention maps, focusing on those components that are densely connected (noisy regions usually do not have large clusters).

The compactness of the cluster is computed by calculating the ratio of the size of the largest connected component to the total number of active pixels on the binary map. A higher compactness value suggests that the largest connected component occupies is dense, and a lower value indicates that attention is weakly spread across the regions, hence, filtered out.

Post-MCC filtering, we take a max of the remaining attention maps to get the final attention map for a give layer. We perform a union over the remaining attention maps instead of averaging them out, as less activated regions reduce the effective attention scores of the highly activated regions if averaged.

3.2. Layer Aggregation

Once we select the best attention heads in every layer, we perform a head aggregation across all layers using the Layer Aggregation module. First, we compute the attention flow of the coarse-grained tokens across all layers. Then, the rolled-out attention flow is used to aggregate the fine-grained attention maps via an overlapping constraint. Figure 3 shows an illustration of this process.

3.2.1 Coarse-grained attention flow

We compute the attention flow of the coarse-grained tokens, which is consequently used to guide the aggregation of the attention maps of the fine-grained attention maps. However, this is based on the assumption that the coarse-grained maps generated are correct. The attention flow is computed using rollout.

The attention rollout R for coarse-grained flow is computed iteratively across all layers of the Transformer. Let L denote the total number of layers, and $A^{(l)} \in \mathbb{R}^{n \times n}$ represent the aggregated coarse-grained attention matrix at layer l, where n is the number of tokens. The computation is initialized with the identity matrix $I \in \mathbb{R}^{n \times n}$, which preserves the direct contributions of each token. The attention rollout R is expressed as shown in Equation 2.

$$R = \prod_{l=1}^{L} \left(I + A^{(l)} \right) \tag{2}$$

where $I + A^{(l)}$ accounts for both the residual connections (via I) and the attention contributions from layer l. The iterative product ensures that the flow of attention from earlier layers is propagated through subsequent layers, capturing the cumulative coarse-grained attention distribution.

Starting with the coarsest layer, we progressively combine the attention scores using an attention rollout technique, where each layer's output is fed as an input to the next, accumulating a composite map that represents an increasingly refined focus on relevant visual features. The rollout helps in understanding how general features (like the shape of a bird) are gradually refined into specific attributes (like the texture of feathers)

3.2.2 Hierarchically Constrained Aggregation

This step introduces a hierarchical constraint that ensures that fine-grained attention regions are always within the bounds of the relevant coarse-grained regions identified in the previous step. The constraint is based on the premise that attributes (such as the eye of a bird) should logically occur within specific broader regions (like the head of the bird).

Using the coarse-grained attention maps as a guide, we apply a masking technique where the fine-grained attention maps are element-wise multiplied by the coarse-grained maps. This operation ensures that attention for fine-grained features does not stray outside the relevant coarse-grained areas, thereby enhancing the precision of the localization.

To enforce hierarchical constraints, the rollout R is used as a coarse-grained mask, ensuring that fine-grained attention maps remain spatially bounded within the regions identified by R.

$$W_i = \text{Normalize}(A'_{\text{fine}} \circ R) \tag{3}$$

where,

A'_{fine} : Fine-grained token's attention map for layer l

R : Coarse-grained attention rollout

$\circ$: Element-wise multiplication

$$A_{\text{final}} = \sum_i W_i \cdot A'_{\text{fine}} \tag{4}$$

For every layer, we obtain a normalized fine-grained attention map weight(3), which is then used to perform a weighted summation of all attention maps across all the layers to get the final map (4).

4. Experiments

4.1. Datasets

We selected a richly annotated dataset with diverse samples within a given animal class to evaluate our proposed method for fine-grained attribute grounding. These datasets are chosen to ensure rigorous testing of our approach's fine-grained grounding capabilities. The datasets used in this study include the Caltech-UCSD Birds-200-2011 (CUB-200-2011).

4.1.1 CUB-200-2011

The Caltech-UCSD Birds-200-2011 (CUB-200-2011) [34] dataset consists of 11,788 images representing 200 bird species, each annotated with bounding boxes, part locations, and 312 binary attributes grouped into 28 categories. These annotations enable precise evaluations of part-based detection, attribute recognition, and fine-grained classification. The dataset's hierarchical structure organizes species by scientific taxonomy—order, family, genus, and species—and links each to a corresponding Wikipedia article for additional contextual details.

Each image in the dataset is annotated with a bounding box that specifies the spatial extent of the bird, facilitating object localization tasks. Additionally, the dataset provides pixel-level annotations for 15 body parts, such as the beak, wing, and tail, which were determined using the median position of five different Mechanical Turk workers. These part locations serve as critical ground truth for part-based localization experiments. Furthermore, 312 binary attributes—such as "wing color: red" and "bill shape: cone-shaped"—are grouped into 28 categories, enabling detailed attribute-based classification and recognition tasks.

The dataset also presents diverse statistical characteristics. Images were collected from Flickr and manually curated to ensure quality and relevance. Class distributions are approximately balanced, with most bird species represented by around 60 images. The average image resolution is approximately 500×500 pixels, with annotated bird regions occupying varying frame proportions. This diversity in pose, lighting, and background introduces significant challenges, as does the subtle visual similarity across some bird species.

4.2. Evaluation Metrics

Since the CUB-200-2011 consists of key points for different bird parts, we use a modified point-in-region metric to calculate the accuracy of the localized region. Point-in-region has a binary value, where the value is 1 if the key point lies within the predicted region; otherwise, it is 0. This could lead to extremely high accuracies in the case of very large predicted regions. Since we want compact regions, we add an additional score to penalize large regions. We follow the equation 1 to add a score to the existing point in the region score.

$$mPIR = \begin{cases} \frac{1}{P_A}, & \text{if predicted region overlaps with the keypoint} \\ 0, & \text{else.} \end{cases} \tag{5}$$

where $mPIR$ is the modified point in region and P_A refers to the area of the predicted region.

4.3. Results

In this section we present our layer-wise attention maps which show attention distribution across layers and results that demonstrate the function of the head selection and aggregation module and the layer aggegation.

4.3.1 Intermediate and Final layers

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 4, 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.

4.3.2 Early Layers

The attention in the first few layers is notably diffuse and appears haphazard, which aligns with the model's tendency

Figure 4. Attention maps from intermediate and final layers

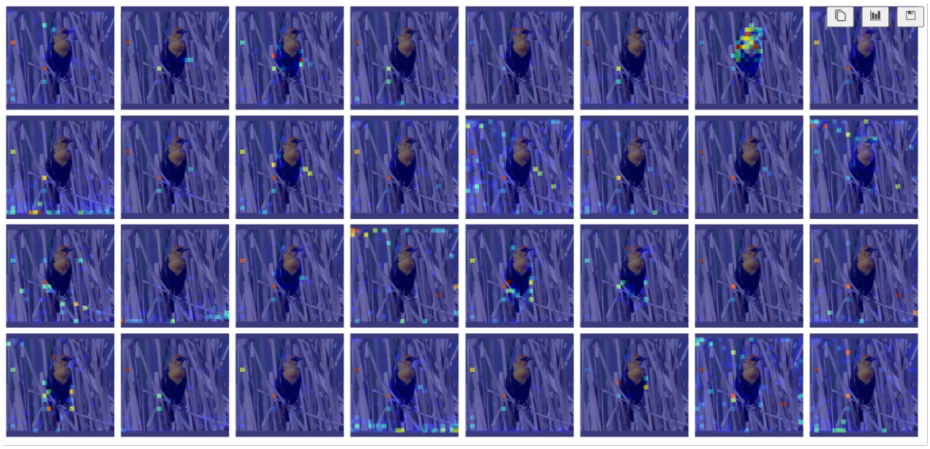


Figure 5. Attention from the word ”bird” to image for layer 10 of LLaVA

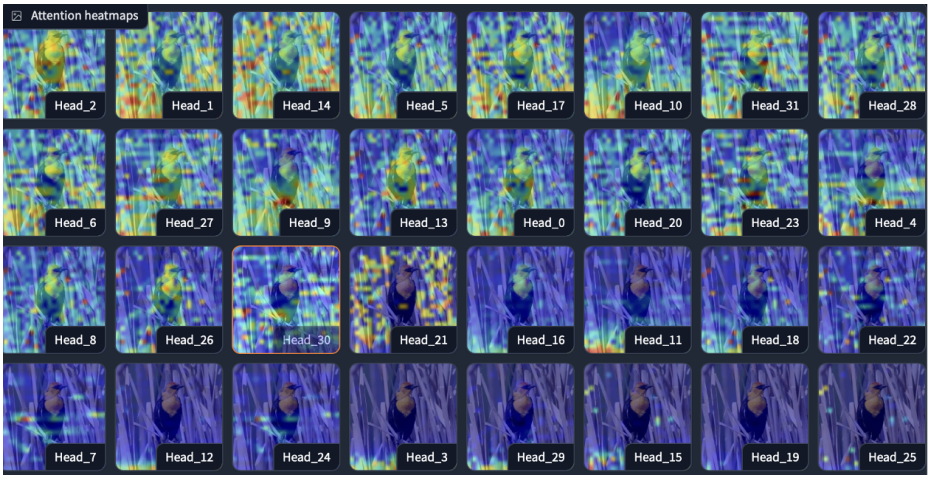
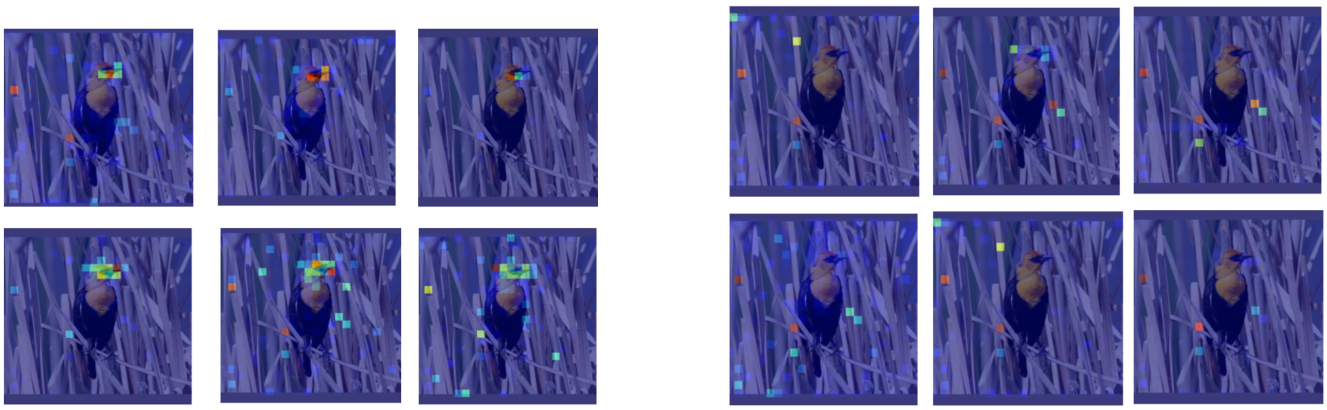


Figure 6. Attention maps from initial layer



(a) Attention maps selected by the HSA module for "head"

(b) Attention filtered out by the HSA module

Figure 7. Attention maps processed by the HSA module

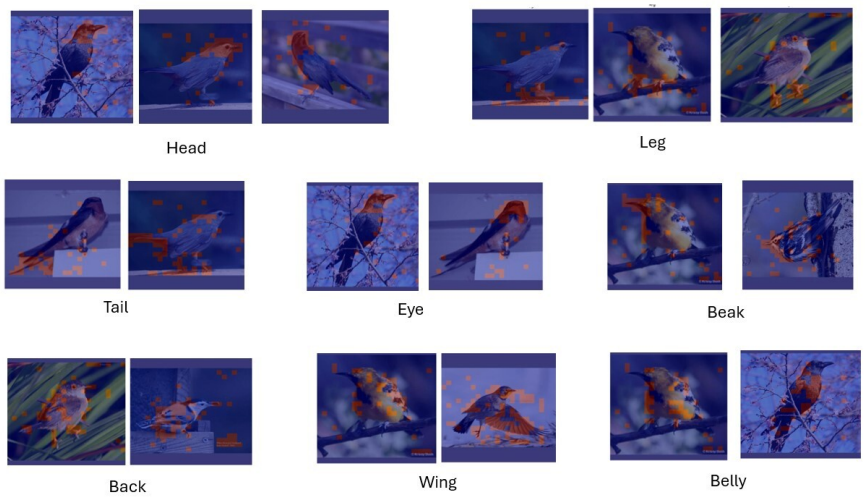


Figure 8. Outputs from layer aggregation for different queries

to incorporate broad contextual information at early stages. As seen in figure 6, 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.

4.3.3 Head Selection and Aggregation

The HSA picks out the most informative attention maps for a given layer and aggregates them into a final attention map. We set an entropy threshold of 0.8 and a compactness threshold of 0.5. The minimum threshold value for pixel to be considered as active is set at 0.05. Figure 7a shows the selected attention maps by the HSA module. It can be observed that selected maps consist of high attention pixels aggregated together. Figure 7b shows the attention maps filtered out by the HSA module. It can be seen that attention maps where high attention pixels are distributed throughout the image are filtered out.

4.3.4 Layer aggregation

The layer aggregation module gives the final localized regions on fine-grained attributes. The resulting outputs from the layer aggregation for different queries is shown in figure 8. The performance is relatively high for attributes like the head, leg and tail. However, for finer-grained attributes such as the eye and the beak, which fall within the head, we tend to get regions that are not localized enough. For body parts like the back, wing and the belly, we get localized regions over the body parts.

Table 1. Comparison of attribute localization scores obtained using VL-SAM* and our proposed method. Our method demonstrates significantly higher scores across all attributes, indicating improved localization ability.

Attribute	Score using VL-SAM*	Score using our method
beak	4.09	10.12
belly	3.76	7.53
breast	3.81	8.00
head	3.96	9.45
wing	3.19	6.90
eye	4.21	10.73
tail	3.75	7.61
throat	4.15	9.81
back	3.62	6.97
leg	3.88	8.82

5. Limitations and Future Work

While our framework demonstrates progress in leveraging LVLMs for fine-grained grounding, several limitations remain. Quantitative evaluation relies on an approximate scoring scheme, which, while informative, lacks the rigor of standardized metrics. Additionally, manual annotation for fine-grained segmentation is resource-intensive and limits scalability. Our method focuses primarily on aggregating informative attention heads, leaving spurious attention peaks unaddressed and failing to amplify connected attention regions for better spatial coherence. Moreover, the attention maps are not refined as 2D distributions, where lower-variance modes (disconnected regions) could be dampened, and higher-variance modes (connected regions) sharpened. These gaps may lead to noisy attention outputs, particularly for fine-grained traits requiring precise localization.

To address these limitations, future work will focus on improving the robustness and scalability of the framework. We plan to develop a fine-grained segmentation dataset based on the CUB-200-2011 dataset, leveraging annotated keypoints as spatial prompts for models like SAM. Additionally, refining attention maps by suppressing spurious peaks and amplifying connected regions will improve spatial precision. This involves treating attention maps as 2D distributions and using techniques like variance-based sharpening and temperature softmax. Another avenue involves modifying visual token attention scores directly within LLM layers to assess their impact on token generation and trait localization.

We also aim to extend the framework to other LVLMs, evaluating its generalizability and effectiveness compared to models explicitly trained for localization or segmentation, such as GLAMM (CVPR 2024). Preliminary analyses suggest that GLAMM struggles with fine-grained segmentation, providing an opportunity to further establish our method's strengths. By addressing these limitations, we aim to enhance LVLMs' ability to describe and localize subtle visual attributes with greater precision and reliability.

6. Conclusion

In this work, we presented a framework to improve fine-grained grounding in Large Vision-Language Models (LVLMs), which excel at coarse-grained tasks but struggle with fine-grained attribute localization. By integrating coarse- and fine-grained attention maps, we enhanced attention consistency and spatial focus through entropy-based filtering and hierarchical constraints. Our approach advances LVLMs' ability to localize fine-grained attributes, providing insights into improving attention mechanisms. In future work, we aim to address current limitations, extend the framework to other LVLMs, and evaluate its effectiveness on state-of-the-art localization models, ultimately enhancing LVLMs' precision in fine-grained reasoning tasks.

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] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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, 2

[10] 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. 2

[11] Joseph F DeRose, Jiayao Wang, and Matthew Berger. Attention flows: Analyzing and comparing attention mechanisms in language models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1160–1170, 2020. 3

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

[13] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, pages 153–168. Springer, 2020. 1

[14] Ruoyi Du, Jiyang Xie, Zhanyu Ma, Dongliang Chang, Yi-Zhe Song, and Jun Guo. Progressive learning of category-consistent multi-granularity features for fine-grained visual classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9521–9535, 2021. 1

[15] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise confusion for fine-grained visual classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 70–86, 2018. 1

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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. 3

[21] 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

[22] 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. 1

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2

[24] 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

[25] Niklas Metzger, Christopher Hahn, Julian Siber, Frederik Schmitt, and Bernd Finkbeiner. Attention flows for general transformers. *arXiv preprint arXiv:2205.15389*, 2022. 3

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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. 2, 3

[31] 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

[32] 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

[33] 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

[34] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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. 3

[40] 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