Finer: Investigating and Enhancing Fine-Grained Visual Concept Recognition in Large Vision Language Models

Jeonghwan Kim **Heng Ji**
University of Illinois Urbana-Champaign
{jk100, hengji}@illinois.edu

Abstract

Recent advances in instruction-tuned Large Vision-Language Models (LVLMs) have imbued the models with the ability to generate high-level, image-grounded explanations with ease. While such capability is largely attributed to the rich world knowledge contained within the Large Language Models (LLMs), our work reveals their shortcomings in fine-grained visual categorization (FGVC) across six different benchmark settings. Most recent state-of-the-art LVLMs such as LLaVa-1.5, InstructBLIP and GPT-4V not only severely deteriorate in terms of classification performance, e.g., average drop of 65.58 in EM for Stanford Dogs for LLaVA-1.5, but also struggle to generate descriptive visual attributes based on a concept that appears within an input image despite their prominent zero-shot image captioning ability. In-depth analyses show that instruction-tuned LVLMs suffer from modality gap, showing discrepancy when given textual and visual inputs that correspond to the same concept. In an effort to further the community's endeavor in this direction, we propose a multiple granularity attribute-centric benchmark and training mixture, **FINER**, which aims to establish a ground to evaluate LVLMs' fine-grained visual comprehension ability and provide significantly improved explainability.

1 Introduction

In recent years, Large Vision-Language Models (LVLMs) that are able to generate image-grounded text have seen significant progress. Models such as InstructBLIP (Dai et al., 2023) and LLaVA (Liu et al., 2023b,a) have consistently exhibited strong zero-shot capability in generating image captions, visual reasoning and textual descriptions, and even leveraging external knowledge for complex question answering tasks (Marino et al., 2019; Schwenk et al., 2022). Such results across diverse benchmarks indicate that these models, most of them being built on large language models (LLMs)

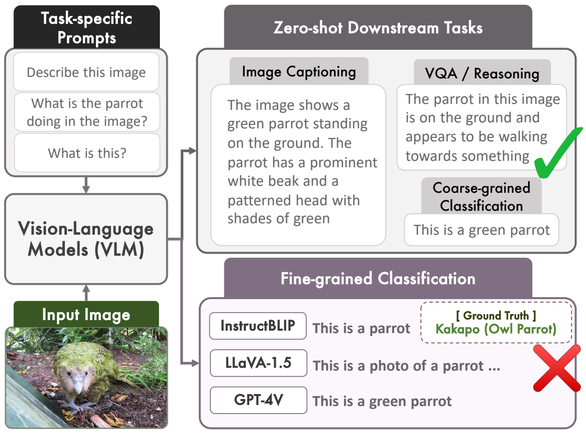


Figure 1: Current state-of-the-art LVLMs exhibit strong zero-shot downstream task solving abilities (e.g., image captioning, VQA, reasoning). However, when prompted to classify the fine-grained concepts, most of them fail to distinguish them into finer categories. Fine-grained classification prompt here is omitted for brevity.

like Vicuna (Chiang et al., 2023), Flan-T5 (Chung et al., 2022), Llama (Touvron et al., 2023), are already equipped with the ability to simultaneously leverage the interplay between the textual parametric knowledge acquired during pre-training and the image understanding ability acquired during instruction-tuning. Notably, all of these models exhibit strong *zero-shot task transferability* to multiple downstream tasks.

Conventionally, in the computer vision domain, many previous works on fine-grained visual classification (FGVC) (Wei et al., 2022b; Diao et al., 2022; Zhu et al., 2022; Yang et al., 2022; Wang et al., 2022) sought to accurately classify diverse images ranging from different types of birds, plants, animals (Van Horn et al., 2015, 2018) and artificial objects such as cars (Krause et al., 2013) and aircrafts (Maji et al., 2013). In this work, we investigate whether state-of-the-art LVLMs can combine their image understanding ability and rich textual knowledge acquired during pre-training to handle zero-shot FGVC. To our surprise, while the

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6187–6207
November 12-16, 2024

LVLMs perform almost perfectly, e.g., 98.43 for LLaVA-1.5 (13B) on iNaturalist, at superordinate-level granularity (e.g., *birds*, *jets*), their classification abilities do not extend to the coarse and finer-grained concepts (e.g., *bald eagle*, *F-22 Raptor*), exhibiting substantially deteriorated classification performance (§3.2); 46.91 for coarse-level and 1.56 for fine-level categories on iNaturalist. Our empirical analyses of these models reveal that these models suffer from *modality gap*. We empirically demonstrate that such discrepancy stems from LVLMs' limited ability to exploit the rich parametric knowledge given image input, to infer fine-grained concepts. We also show that such constraints lead to diminished fine-grained understanding of the image, preventing these models from generating accurate and detailed visual attributes of the concepts that appear within an image.

We also present an attribute-centric and multiple granularity classification benchmark and training mixture, **FINER**. Our benchmark constructs concept-indicative attributes for six conventional FGVC benchmarks like iNaturalist (Van Horn et al., 2018) and FGVC-Aircrafts (Maji et al., 2013) by (i) generating multiple granular concept labels for visual concept recognition, and (ii) constructing a set of visual attributes per fine-grained concept to measure the ability of LVLMs to accurately generate fine-grained concept descriptions given an image. To summarize, our contributions include:

- We highlight the lack of fine-grained image comprehension ability of instruction-tuned LVLMs across various real-life objects. To the best of our knowledge, we are the first to explore FGVC as an evaluation criteria for these models and their lack of ability thereof.
- We underscore the persistence of modality gap in state-of-the-art LVLMs by conducting an extensive per-modality-based probing, revealing the discrepancy in how the two modalities are processed by these models (§4).
- We construct a novel attribute-centric benchmark for FGVC to open up a new direction for future works to measure LVLMs' modality gap and their granular image understanding capability. Our **FINER** training mixture and newly proposed prompting technique **ATTRSEEK** enable substantially improved zero-shot FGVC performance for GPT-4V (§3.2) and LLaVA-1.5 (§5.3).

2 Related Work

2.1 Instruction-tuned Large Vision-Language Models

State-of-the-art LVLMs such as LLaVA (Liu et al., 2023a,b), BLIP-2 (Li et al., 2023), InstructBLIP (Dai et al., 2023), and closed-source models like GPT-4V (OpenAI, 2023; Yang et al., 2023) have brought to our attention their zero-shot task solving abilities, especially in downstream tasks such as Visual Question Answering (VQA), reasoning and image captioning, all of which require output generation conditioned on extensive knowledge of the real-world. Based on an intricate interplay between their parametric knowledge and image understanding ability, they are able to generate sensible outputs. However, many of them focus almost exclusively on image captioning and reasoning, most often disregarding the concept recognition tasks traditional computer vision tasks evaluate on.

2.2 Fine-grained Visual Categorization

Previous works approach FGVC with masked image modeling (He et al., 2022; Ryali et al., 2023), concept meta-information injection (Diao et al., 2022), or LLM-generated concept descriptions (Menon and Vondrick, 2023). However, learning of fine-grained visual categories in LVLMs and their ability to elaborate on the fine-grained details of the input image via text generation is yet to be explored. Furthermore, recent works in FGVC have also shown to disregard fine-grained details of images (Krojer et al., 2022) and work poorly on downstream tasks involving localization (Ranasinghe et al., 2023; Zhong et al., 2022) and object attributes (Yuksekgonul et al., 2022).

2.3 Modality Gap in Vision-Language Models

Different from instruction-tuned VLMs like LLaVA and InstructBLIP, contrastively trained VLMs like CLIP (Radford et al., 2021) and GLIP (Li et al., 2022) rely on directly minimizing the contrastive objective between visual and textual representations. A recent analytical work (Liang et al., 2022) on CLIP-like models reveals that there is a modality gap between the text and visual modalities, which imposes substantial implications on downstream task performances. Our work shows that such *modality gap* also exists in LVLMs like LLaVA and InstructBLIP, despite the noticeable architectural difference between models like CLIP and LVLMs discussed in this work.

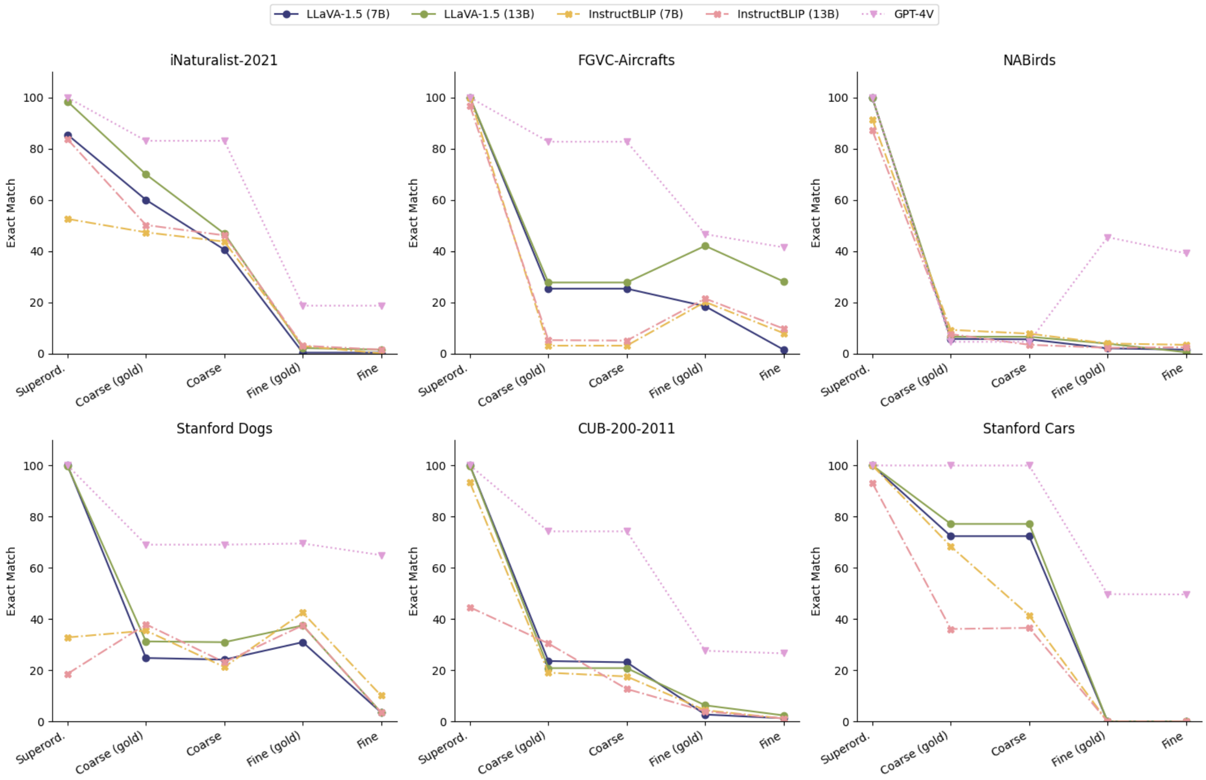


Figure 2: **State-of-the-art instruction-tuned LVLM zero-shot performance on fine-grained classification.** All the models exhibit strong classification capabilities when prompted to classify superordinate-level (e.g., *birds*, *cars*) and coarse-grained categories(e.g., *owls*, *SUVs*), but exhibit significant deterioration in performance when prompted to categorize more fine-grained categories on the same images. The `gold` tags for coarse- and fine-grained classifications denote the use of gold labels from the parent category in the prompt.

3 Fine-Grained Image Understanding in Vision-Language Models

We first evaluate the fine-grained visual categorization performance of five different instruction-tuned baselines on six different FGVC benchmarks. This section elaborates on the details of the experimental setup and models investigated in this work.

3.1 Evaluation Settings

Datasets Covering a wide range of real-world objects over various categories, existing FGVC benchmarks provide richly annotated set of image-concept pairs. As shown in Figure 2, we use iNaturalist-2021 (Van Horn et al., 2018), FGVC-Aircrafts (Maji et al., 2013), Stanford Dogs (Khosla et al., 2011), Stanford Cars (Krause et al., 2013), NABirds (Van Horn et al., 2015), and CUB-200-2011 (Wah et al., 2023). For each dataset, we divide the ground-truth concept label into three levels of granularity: superordinate, coarse and fine, as defined in a previous work (Hajibayova, 2013). Superordinate level refers to the highest taxonomic concepts (e.g., *bird*, *car*), coarse level refers to the lower-level granularity concepts (e.g., *parrot*, *SUV*), and fine level refers to the lowest, finer-level granularity (e.g., *owl parrot (Strigops habroptila), Hyundai Santa Fe 2018*). We discuss each benchmark and the construction of superordinate and coarse-grained labels in detail in Section 5.

Metrics We assess the accuracy of the generated concept labels given an image and a granularity-specific prompt asking the model to figure out the correct category the concept in the image belongs to. Following previous works on concept classification using auto-regressive models, we employ F1 and Exact Match (EM) scores; note that the EM score used in this work is a modified EM score that parses a sequence of generated text and considers the output label correct if the ground-truth label string exists within a pre-defined maximum number of tokens, m (we set $m = 20$).

Models The models used in this work are as follows: **LLaVA-1.5** (Liu et al., 2023b,a), **InstructBLIP** (Dai et al., 2023), and **GPT-4V**; the hyperparameter setting for each model is in Appendix

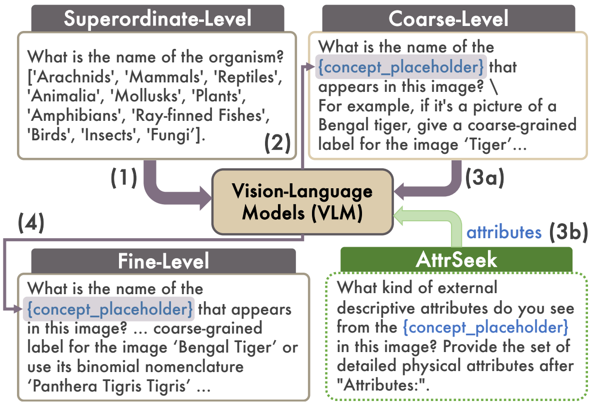


Figure 3: **Fine-grained classification pipeline.** At each level, an output from LVLM is injected into the next level prompt. **(1)** Superordinate-level prompt is used to predict the highest-level category (e.g., *bird*). **(2)** Coarse-level prompt is subsequently fed with the predicted output and fed back to the LVLM to generate the next output (e.g., *parrot*), and **(3a)** and **(4)** follow the same steps. **(3b)** illustrates **AttrSeek**, a newly proposed prompting scheme in this work, wherein the model is prompted to generate the visual attributes.

Models	Superordinate	Coarse	Fine
LLaVA-1.5 (13B)	96.872	**40.625**	1.562
+ CoT (0-shot)	-	35.910	**4.687**
+ CoT (3-shot)	-	34.925	2.159
+ **AttrSeek**	-	35.937	1.562
InstructBLIP (13B)	98.437	**50.221**	**3.714**
+ CoT (0-shot)	-	**31.805**	1.238
+ CoT (3-shot)	-	30.591	0.919
+ **AttrSeek**	-	30.177	3.020
GPT-4V	100.00	83.115	18.752
+ CoT (0-shot)	-	93.750	20.312
+ CoT (3-shot)	-	95.283	24.389
+ **AttrSeek**	-	**95.331**	**53.125**

Table 1: **Elicitive prompting results (EM;%) on iNaturalist.** Prompting techniques like Chain-of-Thought (CoT) on the fine-grained classification (**Fine**) is unable to improve performance in open-source models.

A.1. The open-sourced models like LLaVA-1.5 and InstructBLIP follow a generic pipeline of transforming an input image $\mathbf{X_v}$ with a frozen vision encoder such as CLIP ViT-L/14 (Radford et al., 2021) into an encoded image representation $\mathbf{Z_v}$. Then, these models either project $\mathbf{Z_v}$ into the language representation space through a learned projection layer $\mathbf{W}$, which becomes $\mathbf{H_v} = \mathbf{W} \cdot \mathbf{Z_v}$ as in Liu et al. (2023b), or attend over $\mathbf{Z_v}$ with learnable queries $\mathbf{Q}$ as in Li et al. (2023); Dai et al. (2023). Such transformed visual representations interact with $\mathbf{X_{instruct}}$, a language instruction, which attends over the image representations (or queries) within the self-attention layers of the LLM component to generate the final output sequence.

3.2 Brittleness of Vision-Language Models

Zero-shot Model Performance on FGVC To evaluate the fine-grained image recognition ability of LVLMs, we measure their classification performance per granularity by prompting the models to generate the correct label for a given concept image as shown in Figure 2. As illustrated in Figure 3, we assess the models' classification ability across three different granularity levels. In Figure 2, we evidence significant deterioration in terms of classification performance across all the five baselines, with some, e.g., iNaturalist-2021, even reaching near 0% in EM score. While models do perform very well for superordinate-level categories, often achieving 100% in EM, the finer granularity leads to substantially worsened classification performance. In terms of model size, larger models like LLaVA-1.5 (13B) and GPT-4V tend to perform better than smaller models like the 7B variants. For InstructBLIP, the 7B version performs better than the 13B version by a large margin, with the 13B version exhibiting less capable instruction-following ability than the 7B one, potentially due to 7B variant being less prone to overfitting and exhibiting efficiency in simple tasks like classification.

Elicitive Prompting for FGVC LLMs like Vicuna (Chiang et al., 2023) and LLama (Touvron et al., 2023) used as textual reasoning components of LVLMs are known to perform better when presented with elicitive prompts like Chain-of-Thought (CoT) (Wei et al., 2022a) that improve model's reasoning ability. A unifying thought along this line of prompting techniques is to break down a complex problem into a sequence of subproblems, i.e., divide-and-conquer. Inspired by these prompting techniques, we propose and evaluate our prompting technique for FGVC, **AttrSeek**. In this simple prompting strategy, we first (i) prompt the models to generate the most distinctive physical attributes visible in the concepts in an image, and (ii) feed the generated set of attributes along with the concept-asking prompt for final prediction as shown in Figure 3. We evaluate the 13B model variants and GPT-4V only on iNaturalist-2021 dataset due to budget constraint of evaluating on the full evaluation set. In Table 1, we do not see much improvement in terms of fine-grained concept classification in the open-source models, with neither CoT nor ATTRSEEK enhancing the classification performance. However, there is a substantial increase in fine-level performance for GPT-4V when prompted with our simple yet effective ATTRSEEK scheme. This result suggests that LLaVA-1.5 and InstructBLIP, while they exhibit strong image captioning and reasoning ability, are limited in terms of image-grounded attribute understanding even when provided with elicitive prompts; we further elaborate on the need to fine-tune the model according to the newly proposed prompting scheme in Section 5.3. This result also suggests that open-source LVLMs may lag behind in instruction-following abilities in comparison to GPT-4V. For additional details on few-shot prompting, see Appendix A.3.

4 Modality Gap: Discrepancy Between Textual and Visual Modalities

We hypothesize that the lack of zero-shot concept classification ability of LVLMs arises mainly due to the modality gap between textual and visual inputs, preventing the models from leveraging the existing concept-related parametric knowledge when an image of a concept is given. Note that modality gap studied in this work is different from the one previously identified in CLIP-like VLMs (Liang et al., 2022). In this section, we aim to delve into the details of how these models process visual and textual modalities by exploring how well they perform when given only textual descriptions of a concept (§4.1) and how they accurately elaborate what they see in a given image (§4.2). We also perform linear probing (§4.3) against projected and original vision encoder output embeddings to gauge the influence of projection and the subsequent loss of visual information on the modality gap.

4.1 Probing for Concept-Related Parametric Knowledge

With the drastic performance drop in Section 3.2, we first need to verify whether concept-attribute knowledge already exists within the LVLM parameters to make sure that the models have already acquired the knowledge necessary for zero-shot classification. The concept-attribute knowledge refers to the textual parametric knowledge related to specific concepts, e.g., a concept *Bengal Tiger* has visual attributes *dark brown or black stripes*.

In this experiment, we measure the classification performance of the models with two different input types: (i) **Text-only** input that consists of a concept's external visual attributes (details about the attribute extraction in §5), and (ii) **Image-only** input that consists of the image of the concept. The text-only input, $\mathbf{X_{txt}} = [I; Attr; C]$, is composed of Instruction (I), visual attributes ($Attr$), and coarse-grained labels (C); the image-only input is $\mathbf{X_v} = [I; X_{img}; C]$. The final output is the concept name. Note, for fair comparison, we also include C as input for image-only probing. Refer to Appendix B.2 for prompts. In Figure 4, the results show that even with text-only input that contains the detailed physical attributes of a concept, LVLMs are capable of solving fine-grained visual classification, outperforming the image-only input. The results imply two things: (i) concept-attribute knowledge exists in model parameters, and (ii) while the visual attributes are strongly correlated with the concept, the image modalities are incapable of leveraging the concept-attribute knowledge. We also see that the larger the model size, the better the performance, relating to the amount of parametric knowledge that resides in the LLMs.

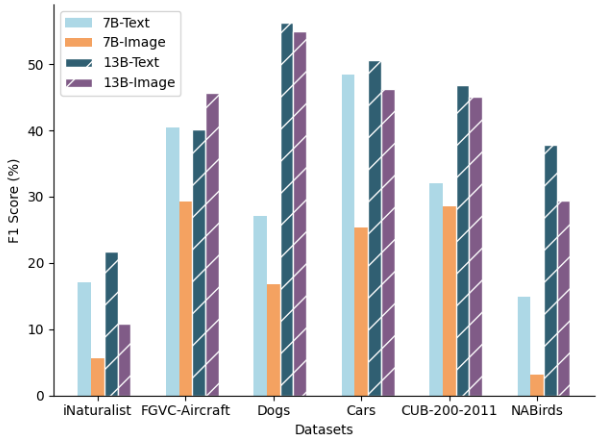


Figure 4: **Model Performance on Text-only vs. Image-only Inputs.** LLaVA-1.5 (`7B` and `13B`), when provided only the textual information (`7B-`, `13B-Text`) related to the ground-truth concept, outperforms the image-only input (`7B-`, `13B-Image`) counterpart.

4.2 Measuring the Modality Gap with Attribute Generation

Having observed the discrepancy between the visual and textual modalities, we now analyze whether LVLMs can observe and tell visually grounded physical attributes of an input image. We construct a set of Web-extracted concept attributes from Wikipedia documents (further elaborated in detail in Figure 6; §5) to be used as the

Base Model	LLM	Input Modality	Similarity Measurement Metrics				
			ROUGE-1	ROUGE-2	ROUGE-L	BertScore	AlignScore
LLaVA-1.5 (Liu et al., 2023a)	`Vicuna-7B`	`Text`	**22.343**	**11.023**	**21.161**	**85.429**	**18.935**
		`Image`	17.468	10.049	17.314	84.499	15.509
	`Vicuna-13B`	`Text`	**20.616**	10.902	**19.647**	**85.944**	**15.746**
		`Image`	19.661	**11.334**	19.114	85.704	13.319
		Δ `Avg.`	2.920	0.703	2.189	0.585	2.926
InstructBLIP (Dai et al., 2023)	`Vicuna-7B`	`Text`	**18.987**	9.192	**16.577**	**83.102**	**15.002**
		`Image`	16.018	**9.788**	11.258	82.681	10.186
	`Vicuna-13B`	`Text`	**13.731**	**7.541**	**16.446**	**80.549**	**13.884**
		`Image`	10.949	6.009	9.583	79.277	10.191
		Δ `Avg.`	2.875	1.064	6.091	0.846	4.254
GPT-4V (OpenAI, 2023)		`Text`	**24.320**	**10.179**	**22.748**	**87.457**	**10.524**
		`Image`	22.675	8.812	20.477	85.495	5.067
		Δ `Avg.`	1.644	1.367	2.271	1.961	5.456

Table 2: Measuring the modality gap via textual similarity against the Web-extracted concept attributes against the LVLM-generated attributes for the iNaturalist subset in **Finer**. The discrepancy between the attributes generated from `Text`-only input and `Image`-only input indicate that VLMs treat the two modalities of the same concept differently. Δ `Avg.` indicates the average difference between the `Text` and `Image` outputs against the reference.

reference texts. Then, we prompt the models to generate a set of external, discriminative physical attributes of a concept when given either an image or text input (see the detailed prompts in Appendix 11); the text input refers to the concept label along with a prompt that asks for the concept's visual attributes. The textual similarities between the LVLM-generated attributes and the web-extracted attributes are compared using five different scoring metrics that span both the token-level overlap and NLI model-based semantic similarity measure: **ROUGE-1, 2, L**, **BertScore** (Zhang et al., 2019), and **AlignScore** (Zha et al., 2023). Both the model-generated attributes and web-extracted attributes are linearized for textual similarity comparison (Appendix A.2).

As shown in Table 2, text-only inputs show greater textual similarity to reference attributes, indicating that while the concept-attribute knowledge is being used by the textual modality, the visual modality does not leverage such knowledge to a degree that matches the textual modality. Aside from the modality discrepancy, these models potentially fall short on accurately focusing on specific parts of the image, diluting away the fine-grained details such as stripes or patterns (Table 3) for a coarse-grained understanding of the image.

4.3 Visual Information Loss After Projection

We also evaluate the impact of projection from the visual embedding space to the textual space through linear probing. We use CLIP-ViT-L/14 as the image encoder and use LLaVA-1.5's projector. We freeze both the image encoder and the projector and finetune a multi-layer perceptron (MLP) layer on top for classification for 10 epochs (for experiment details refer to Appendix B.5). As shown in Figure 5, the loss of visual information encoded by the vision encoder leads to substantial drop in classification performance across the six FGVC tasks. The results strongly suggest that such loss of visual information further contributes to the modality gap between the two modalities, especially when performing tasks such as FGVC that rely heavily on fine-grained visual attributes.

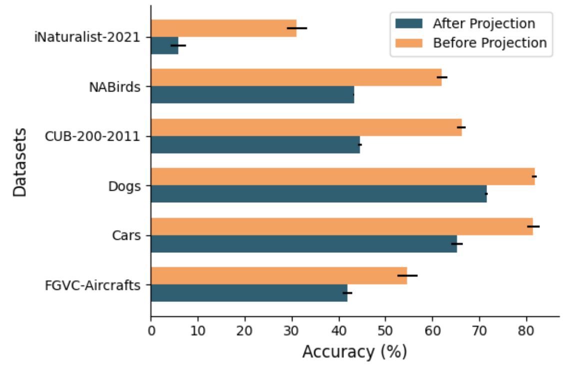


Figure 5: **Linear Probing on Projected Image Embeddings.** Classification accuracy (%) for before and after image embedding projection to textual space.

5 Finer: Fine-Grained Concept Recognition Benchmark

To facilitate research for fine-grained image understanding in LVLMs, we propose a new benchmark

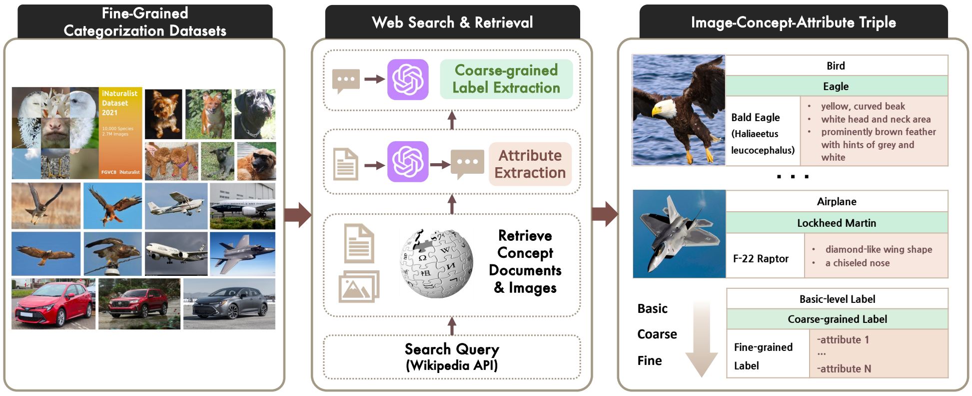


Figure 6: Depiction of the **FINER** benchmark construction pipeline. Following the aggregation of the six benchmarks in the FGVC domain, concept attributes and concept images are retrieved and extracted from Wikipedia documents.

and training mixture, **FINER**. **FINER** intends to evaluate the interplay between the concept image understanding and attribute knowledge, in addition to the training mixture that mitigates the modality gap and enhances fine-grained concept recognition.

5.1 Dataset Construction

We construct **FINER** based on six different FGVC datasets: iNaturalist-2021 (Van Horn et al., 2018), FGVC-Aircrafts (Maji et al., 2013), Stanford Dogs (Khosla et al., 2011), Stanford Cars (Krause et al., 2013), NABirds (Van Horn et al., 2015), and CUB-200-2011 (Wah et al., 2023). These datasets span a wide range of objects such as airplanes, insects, plants, birds, mammals and cars, challenging the models to cover a variety of fine-grained concepts.

We first crawl all Wikipedia documents and their main images via a search API[1]. We then extract external, visual attributes of a concept (i.e., concept-indicative attributes) with GPT-4V as our attribute extractor (OpenAI, 2023) for its strong zero-shot text span extraction ability (Huang et al., 2024). To briefly elaborate, we divide the extracted attributes into two different types: (i) required and (ii) likely attributes. The "required" attributes are the external, concept-indicative attributes that can be used for concept identification, e.g., *blue-tailed hawks with black thorax with a broad apple green stripe*, while the "likely" attributes are attributes that may co-occur with the concept but is not directly correlated with the concept, e.g., *blue-tailed hawks inhabit trees*; we provide the likely attributes as meta-information since existing models such as MetaFormer (Diao et al., 2022) has proven that meta-information associated with these fine-grained concepts are beneficial for more accurate FGVC performance; however, since the use of meta-information for better FGVC is not the main focus of this work, we leave it to future works to leverage this information. We also populate the dataset with superordinate and coarse-level concept labels for multi-granular concept recognition performance evaluation. For example, as shown in Figure 6, we assign *Airplane* as the superordinate-level label and *Lockheed Martin* as the coarse-level label since FGVC-Aircrafts dataset provides a concept granularity hierarchy, e.g., Boeing → Boeing 707 → Boeing 707 MAX. However, datasets like Stanford Dogs do not provide such granular hierarchy for their fine-grained concepts. We therefore few-shot prompt GPT-4V to generate the coarse-level labels and manually inspect their validity. We provide the dataset statistics in Table 9 and extraction prompts in Appendix B.3 and B.4.

5.2 Qualitative Analysis

In Table 3, we provide model-generated attributes as a case study on lack of visually-grounded generation. The attributes are from GPT-4V, but note that the generated attributes from other models, such as LLaVA-1.5, all exhibit a similar trend. When provided with image-only input, the model generates a set of attributes that pertain to the image, e.g., elongated body and two pairs of wings for dragonfly. However, the attributes from image-only input are non-discriminative compared to those from text-only input; furthermore, changing the prompting technique to elicit more fine-grained, detailed physical attributes from the LLMs either lead to hallucination or needlessly verbose outputs that describe non-concept related aspects of the input image. In other words, these attributes do not serve

[1]https://github.com/goldsmith/Wikipedia

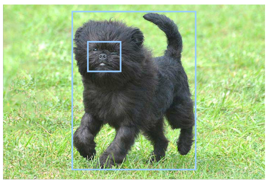
Image (Concept Names)	Attributes (Text-only)	Attributes (Image-only)	FINER Attributes
Dragonfly | Orthetrum Triangulare 	Blue tail; Broad wings; White throat; Dark head; Banded tail	Elongated body; Two pairs of wings; Six legs, Compound eyes, Segmented abdomen	Dark face; Bluish eyes; Black thorax with a broad apple green stripe on both sides; Black segments 1-2 and 8-10 on the abdomen; Remaining segments of the abdomen pruinosed with azure blue
Pinscher | Affenpinscher 	Black, gray, silver, or tan fur; Rough, shaggy coat; Distinct "monkey-like" facial expression; Prominent chin and jaw; Small, round, dark eyes; Ears set high and usually cropped to a point	Presence of fur; Four-legged stance; Distinct muzzle with a nose; Visible tail; Ears that are either erect or floppy	Harsh rough coat when not clipped; Shaggier coat over the head and shoulders forming a mane; Shorter coat over the back and hind quarters; Notable monkey-like expression; Coat is harsh and wiry in texture when properly maintained
Embraer | Embraer ERJ 145 	T-tail design; Straight wing design with no winglets; Mounted engines on the rear fuselage; Small cockpit windows compared to the body size; Three sets of landing gear	Fixed wings on either side of the fuselage; Cockpit windows at the front of the fuselage; Jet engines, either under the wings or mounted on the rear of the fuselage; Landing gear with wheels for taking off	T-tail configuration; Two rear-mounted Rolls-Royce AE 3007 series turbofan engines; Straight wing with no winglets; Narrow, tube-like fuselage; Short, nearly oval-shaped passenger windows.

Table 3: **Qualitative Analysis of the Text-only and Image-only generated attributes from GPT-4V against FINER.** The generated attributes based on the Text-only and Image-only dataset exhibits notable discrepancy. Juxtaposed with our **FINER Attributes**, Image-only attributes are not concept-indicative and generic compared to Text-only which are discriminative of the concept. We provide **FINER** attributes as a reference for comparison. We provide the coarse-level and fine-level labels along with the images, and the bounding boxes are only drawn to match the highlighted text and are not part of the dataset.

as useful knowledge to identify the input image as a specific concept. This again suggests that these models not only fail to properly observe the fine-grained details of a concept, but fail to leverage the knowledge contained within its own parameters as can be seen from the outputs of text-only inputs.

5.3 Enhanced Zero-Shot Transferability from Learning to Generate Attributes

To substantiate the effectiveness of our visual attributes in **FINER**, we construct an instruction-tuning mixture based on the ATTRSEEK prompting pipeline to improve the zero-shot attribute generation and FGVC performance (see Appendix A.4). The **FINER** mixture consists of six subsets, where each split is a 5 held-in and 1 held-out FGVC datasets for training and evaluation, respectively. For instance, for the iNaturalist subset, the iNaturalist set is not included in training for zero-shot evaluation. Each instance of the training mixture follows the ATTRSEEK pipeline (§3.2). In Table 4, we finetune LLaVA-1.5 (7B) on the training mixture and see that the **FINER**-tuned model outperforms the direct finetuned counterpart that was simply trained to directly predict the concept label. This implies that in FGVC, instruction-tuning LVLMs to attend to visible attributes in images by explicitly generating the attributes and then subsequently performing classification improves the model's performance. Our interpretation is that the generation of the attributes of concepts in the images allow the model to leverage its concept-attribute parametric knowledge (identified in Figure 4), and perform better zero-shot FGVC classification. It also underscores the effectiveness of the ATTRSEEK pipeline in model training to improve the inherent, zero-shot capability of LVLMs for fine-grained concept recognition. We demonstrate case studies in Table 5, where we present the zero-shot generated outputs of the **FINER** mixture-trained LLaVA-1.5 (7B) and the one that was only finetuned to directly predict the final concept label. Refer to Appendix A for additional details on training.

6 Discussion and Conclusion

In this paper, we provide an in-depth analysis on the substantial lack of fine-grained visual comprehension (FGVC) ability among instruction-tuned LVLMs. Our work discovers the presence of modal-

Models	CUB-200-2011	Stanford Dogs	NABirds	iNaturalist	Stanford Cars	FGVC-Aircrafts
LLaVA-1.5 (7B)	5.262	16.718	3.145	6.788	25.317	29.314
Direct Prediction	**21.071**	22.942	12.610	7.109	24.624	28.622
FINER	20.673	**36.297**	**13.692**	**7.530**	**29.974**	**32.293**

Table 4: **Zero-shot Performance on FGVC**. Finetuning on the **FINER** mixture significantly enhances the zero-shot performance on all six FGVC tasks. We choose LLaVA-1.5 (7B) for this experiment with 1 dataset held-out and 5 other datasets held-in for finetuning. Direct Prediction refers to the setting without the **ATTRSEEK** pipeline and the model directly predicting the final concept label without the intermediate attribute generation

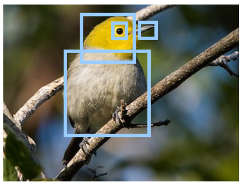
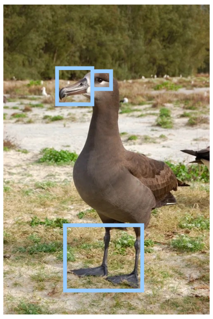
Image (Concept Names)	Attributes from FINER	Attributes from Direct Pred.
Warbler | Yellow-headed Warbler 	It has yellow round head, white body with a hint of gray, black eyes and black-colored tail.	In the image, the bird is perched on a wire, and it has a yellow beak. Its physical characteristics include a white body, a black tail, and a yellow beak.
Albatross | Black-footed Albatross 	Almost all black plumage, White markings around the base of the beak, White markings below the eye, Dark beak and feet, White undertail coverts in some adults.	The bird has Predominantly dark color with subtle variations between brown and gray. The face appears to have lighter shades of brown or gray.

Table 5: **Qualitative analysis of the zero-shot outputs generated by LLaVA-1.5 (7B) instruction-tuned on FINER and Direct Prediction dataset.** The generated attributes from the **FINER** trained LLaVA-1.5 (7B) generates accurate, image-grounded outputs when compared to the Direct Prediction counterpart. The attributes are indicated by their discriminative characteristics in contrast to more generic/hallucinated ones. We provide the coarse-level and fine-level labels along with the images, and the bounding boxes are only drawn to match the highlighted text and are not part of the dataset.

ity gap in LVLMs and propose a prompting scheme, ATTRSEEK, and training mixture, **FINER** to mitigate the gap and improve the zero-shot FGVC ability of LVLMs. We also reveal that the loss of visual information after projection hinder the effective cross-modal interplay from manifesting. Such discrepancy leads to LVLMs being unable to exploit the rich parametric knowledge and deteriorates performance in visual concept recognition. While this is the first study of FGVC among instruction-tuned LVLMs, we hope our work would further the research endeavors in this direction.

7 Limitations

Intra-Concept Variance in Images Images of a single concept can appear in various different forms. Some images may have the whole view of the concept, while other images may have certain parts of the concept (e.g., legs, wings) partially occluded. The attributes collected per concept in **FINER** are constructed to be visually-grounded based on the textual attributes extracted from Web documents that pertain to these concepts; nonetheless, such intra-concept variance among images may render an attribute obsolete for certain images. Other problems such as low-quality images may also lead to this issue. In future work, exploring the visual "ground-ability" of each attribute through image-text retrieval may be a plausible approach to identifying both the most discriminative attributes that pertains to an image and the fine-grained concept label.

Selection of Baseline Models While our work covers LVLMs that receive image and text as input, there are other VLMs such as Kosmos-2 (Peng et al., 2023), Shikra (Chen et al., 2023) and Ferret (You et al., 2023) out there, that receive bounding-box annotated images as input. However, our work only deals with un-marked image and prompt inputs, since the objective of this research is to see whether LVLMs without any referring markings (e.g., bounding boxes), can ground their generative capabilities on the input image.

Acknowledgement

We thank the anonymous reviewers for their suggestions and comments. We also would like to thank Ansel Blume, Derek Hoiem and Carl Vondrick for their ideas, feedback and support for this work. This research is based upon work supported by U.S. DARPA ECOLE Program No. #HR00112390060. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*.

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

Lala Hajibayova. 2013. Basic-level categories: A review. *Journal of Information Science*, 39(5):676–687.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jingwei Huang, Donghan M Yang, Ruichen Rong, Kuroush Nezafati, Colin Treager, Zhikai Chi, Shidan Wang, Xian Cheng, Yujia Guo, Laura J Klesse, et al. 2024. A critical assessment of using chatgpt for extracting structured data from clinical notes. *npj Digital Medicine*, 7(1):106.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer.

Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. 2013. Collecting a large-scale dataset of fine-grained cars.

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. 2022. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3426–3440, Dublin, Ireland. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. 2013. Fine-grained visual classification of aircraft. Technical report.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Sachit Menon and Carl Vondrick. 2023. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

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

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

Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. 2023. Perceptual grouping in contrastive vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5571–5584.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. 2023. Hiera: A hierarchical vision transformer without the bells-and-whistles. *ICML*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2023. The caltech-ucsd birds-200-2011 dataset.

Hao Wang, Junchao Liao, Tianheng Cheng, Zewen Gao, Hao Liu, Bo Ren, Xiang Bai, and Wenyu Liu. 2022. Knowledge mining with scene text for fine-grained recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4614–4623.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022a. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. 2022b. Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8927–8948.

Xuhui Yang, Yaowei Wang, Ke Chen, Yong Xu, and Yonghong Tian. 2022. Fine-grained object classification via self-supervised pose alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7399–7408.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803.

Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. 2022. Dual cross-attention learning for fine-grained visual categorization and object reidentification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4692–4702.

A Experiment Details

In this section, we elaborate in detail the experimental settings of our work, including the hyperparameter settings of the large vision-language models (LVLMs), and the large language models (LLMs), which are used as a major driving block of these LVLMs. In addition to the experiment settings, we also provide details on the training mixture used in 5.3 and the fine-tuning settings of LLaVA-1.5 (7B).

A.1 Hyperparameter Settings of Vision-Language Models

Inference Settings Our work deals with the inference-time fine-grained image understanding abilities of VLMs. The hyperparameter settings provided in the table are only for inference time (§3.2, 4, 4.1) and are not to be used for fine-tuning.

Hyperparameter	LLaVA	InstBLIP	BLIP-2	GPT-4
max_seq_len	256	256	256	256
top_p	1.0	0.95	0.95	1.0
temperature	0.2	0.75	0.75	0.2

Table 6: Hyperparameters of the Vision-Language Models (VLMs) studied in this work. **LLaVA** refers to LLaVA-1.5 and **InstBLIP** refers to the InstructBLIP model. All the VLMs use the same hyperparameter settings for both the 7B and 13B variants; the same applies to `Flan-T5-XL` and `Flan-T5-XXL` models, which consist of 3B and 11B parameter, respectively.

For LLaVA-1.5 models, we use `Vicuna-v1.5`, and for InstructBLIP models, we use `Vicuna-v1.1`; these are the original settings of both of these models.

Fine-Tuning Settings In Table 7, we also provide the hyperparameter settings of the instruction-tuned LLaVA-1.5 (7B) in Section 5.3. The fine-tuning was conducted with LoRA (Hu et al., 2021) on 4 V100 (16G) for approximately 28 hours on the training mixture, which is elaborated in detail in Section A.4.

Hyperparameter	LLaVA-1.5 (7B)
max_seq_len	256
top_p	1.0
temperature	0.2
lora_r	128
lora_alpha	256
gradient_accum.	16
batch_size	1
learning_rate	2e-4
lr_schedule	cosine decay
optimizer	AdamW
weight_decay	0.0
warmup_ratio	0.03
bf16	`False`
fp16	`True`

Table 7: Hyperparameters of LLaVA-1.5 (7B) for Instruction-tuning on the FINER training mixture.

A.2 Additional Details on Modality Gap Experiment

Using LLMs for Text-Only Setting In Section 4, we experiment on the LVLMs' ability to generate textual attributes based on either the text-only input or the image-only input. The caveat of using InstructBLIP and BLIP-2 models is that this models do not simply allow text-only input like LLaVA-1.5 or GPT-4V. We therefore opt to use their LLM components, Vicuna-7B and -13B, Flan-T5-XL and -XXL to generate the attributes for the text-only input. The validity of this setting holds since our goal is to compare the modality gap via how much the model stores the concept-related knowledge within its parameters.

Metrics For ROUGE, we used the Python's ROUGE API[2] for calculation, and we only use the F1 score in this work. As for the model-based metrics, for BertScore (Zhang et al., 2019) we use the `bert-base-uncased` and for AlignScore (Zha et al., 2023) we use `roberta-large`. These model-based metrics are state-of-the-art models for textual faithfulness evaluation, making them fit to evaluate both the faithfulness and textual consistency of the generated text.

Linearization of the Web-Extracted Attributes In this section, we elaborate on the linearization process of the Web-extracted attributes, which are used as ground-truth reference texts in Section 4. The linearization of the attributes is a simple process of concatenating the fine-grained concept name and the generated attributes of the concept. For example, for a dragonfly named *Orthetrum Triangulare*, we construct a linearized string that says, *Orthetrum Triangulare exhibits blue eyes, black thorax with a broad apple green stripe*. The format is as follows: [`<concept-name>`; `exhibits`; `attribute-1`; `attribute-2`; `...`; `attribute-j`]; the list is converted into a single sentence for evaluation;. The same process applies to both the LVLM-generated attributes and

[2]https://pypi.org/project/rouge/

Web-extracted attributes to compare their textual similarity in Section 4.

A.3 Enabling Multiple Image Inputs in LVLMs for Few-Shot Prompting

The three open-sourced LVLMs investigated in our work, LLaVA-1.5 (Liu et al., 2023b,a), InstructBLIP (Dai et al., 2023) and BLIP-2 (Li et al., 2023) were all pre-trained and instruction-tuned based on a single image input and a consecutive sequence of pertinent textual instructions. Nonetheless, these models are capable of receiving multiple images given its input format. For LLaVA-1.5, we simply provide the few-shot (k) examples sampled from iNaturalist dataset in by interleaving k `<image>` tokens along with the input prompt and their ground-truth concept labels. For InstructBLIP and BLIP-2, since they require attending over the input image via cross-attention with a pre-defined set of query embeddings, we first embed each of the k few-shot sample images with the vision encoder. Then, the image embeddings are passed through the Q-Former (Dai et al., 2023; Li et al., 2023) to generate an instruction-attended query embeddings that contains the image information, and we concatenate them together to use them as few-shot samples.

A.4 Construction of the Instruction-Tuning Mixture

To evaluate the validity of our proposed benchmark, **FINER**, we construct six different instruction-tuning mixtures on top of LLaVA-1.5's instruction-tuning mixture. For example, to build a training mixture to train the model for iNaturalist evaluation, i.e., the **FINER**'s iNaturalist subset, we hold out the iNaturalist dataset for evaluation and include the rest of the five other FGVC datasetes into the instruction-tuning mixture. We use all the six FGVC datasets to construct the mixture, sampling 2.5K instances from each of the datasets and their attributes into the training data; note, the 2.5k instances are sampled to follow a uniform distribution for the number of classes for each dataset. We structure each instruction-tuning instance into the ATTRSEEK pipeline format, with each instruction consisting of three turns: (i) Asking the model for the coarse-level concept category given the superordinate concept, *"Can you identify the bird shown in this image?"*; (ii) asking the model to generate a set of external, descriptive visual attributes of the coarse-level concept, *"What kind of external descriptive attributes do you see from the penguin"*, and finally (iii) predicting the fine-grained concept category given the coarse-level concept and the self-generated attributes set. For each of the three steps, we use GPT-4V to generate 15 possible paraphrases of the instruction in order to avoid biasing the model to specific textual instructions and to retain the model's instruction-following ability. We trained the models for 1 epoch each; the checkpoint for 1 epoch is the one we used to evaluate the FGVC performance in Table 4.

B Prompts

In this section, we provide the input prompts used in each of the experiments, including the fine-grained visual classification and the elicitive prompting of LVLMs in Section 3.2, probing for concept-related parametric knowledge 4.1, attribute generation 4.2

B.1 Fine-Grained Visual Classification Prompts

We structure our prompts as shown in Table 8. For datasets with less than 100 class categories, we provide them along with the instruction, allowing the models to choose from the provided list of classes. Therefore, for superordinate-levels and certain coarse-levels, we provide the categories as lists so that the models solve the class generation problem by choosing from the input prompt; this is analogous to a multiple choice setting. However, for fine-grained classes, it is difficult to feed in all the concept categories in the input prompt, since some of the datasets like the iNaturalist-2021 has 10,000 categories to choose from. In order to confine the generation scope for the fine-grained labels, we decided to input the coarse-level label as denoted in Table 8, to condition the generation of the fine-grained output within a specified category space.

B.2 Knowledge Probing Prompts

The knowledge probing prompts are shown in Table 10. We structure the prompt as explained in Section 4.1, where we input the Web-extracted textual attributes along with the coarse-level label for the text-only setting. Since LVLMs are good at identifying the superordinate and coarse-level concepts, we also provide the coarse-level labels as a prior for the text-only setting for a fair comparison in the analysis for fine-grained concept knowledge

Dataset: **iNaturalist-2021**

Superordinate-level
What is the name of the organism that appears in this image? Provide your answer after "Answer:" from one of the following categories: ['Arachnids', 'Mammals', 'Reptiles', 'Animalia', 'Mollusks', 'Plants', 'Amphibians', 'Ray-finned Fishes', 'Birds', 'Insects', 'Fungi'].

Coarse-level
What is the name of the {concept_placeholder} that appears in this image? For example, if it's a picture of a bengal tiger, give a coarse-grained label for the image 'Tiger'. Provide your answer after "Answer:".

Fine-level
What is the name of the {concept_placeholder} that appears in this image? For example, if it's a picture of a bengal tiger, give a fine-grained label for the image 'Bengal Tiger' or use its binomial nomenclature 'Panthera tigris tigris'. Provide your answer after "Answer:".

Dataset: **FGVC-Aircraft**

Superordinate-level
What is the name of the object that appears in this image? Provide your answer after "Answer:" from one of the following categories: ['Airplane', 'Car', 'Train', 'Bicycle', 'Cell Phone', 'Plants', 'Dogs', 'Birds', 'Trucks'].

Coarse-level
What is the manufacturer of the {concept_placeholder} that appears in this image? Provide your answer after "Answer:" from one of the following categories: ['Embraer', 'Lockheed Corporation', 'Douglas Aircraft Company', 'Cirrus Aircraft', 'Airbus', 'Antonov', 'de Havilland', 'Eurofighter', 'Cessna', 'Tupolev', 'Dornier', 'Yakovlev', 'Panavia', 'Robin', 'ATR', 'Beechcraft', 'Dassault Aviation', 'Fairchild', 'McDonnell Douglas', 'Fokker', 'Gulfstream Aerospace', 'Boeing', 'Saab', 'Canadair', 'Lockheed Martin', 'Supermarine', 'Ilyushin', 'British Aerospace', 'Piper', 'Bombardier Aerospace'].

Fine-level
What is the name of the airplane model made by {concept_placeholder} that appears in this image? For example, if it's a picture of a Boeing 787 Dreamliner, give a fine-grained label for the image 'Boeing 787 Dreamliner'. Provide your answer after "Answer:".

Dataset: **Stanford Dogs**

Superordinate-level
What is the name of the organism that appears in this image? Provide your answer after "Answer:" from one of the following categories: ['Arachnids', 'Dogs', 'Reptiles', 'Mollusks', 'Plants', 'Amphibians', 'Ray-finned Fishes', 'Birds', 'Insects', 'Fungi'].

Coarse-level
What is the name of the {concept_placeholder} that appears in this image? For example, if it's a picture of a Golden Retriever, give a coarse-grained label for the image 'Retriever'. Provide your answer after "Answer:".

Fine-level
What is the name of the {concept_placeholder} that appears in this image? For example, if it's a picture of a Golden Retriever, give a coarse-grained label for the image 'Golden Retriever'. Provide your answer after "Answer:".

Table 8: Prompts for fine-grained image classification. The {concept_placeholder} is replaced with upper-level concept labels as illustrated in Figure 3.

in the parametric knowledge space.s

B.3 Attribute Generation Prompts

The attribute generation prompts are shown in Table 11. We divide the attribute generation prompts to three different types: (i) Prompt that generates the Web-extracted attributes given the Wikipedia API retrieved concept documents, (ii) prompt that generates the attributes straight from the model given a text-only input, (iii) prompt that generates the attributes from the model given an image-only input. Note that the prompt variance between the text-only input and image-only input is intentionally minimized, i.e., minimal change in the input prompts, to more accurately isolate the effect of change in the input modalities. The prompts shown

```
Dataset:NABirds

Superordinate-level
What is the name of the organism that appears in this image? Provide your answer after
"Answer:" from one of the following categories: ['Arachnids', 'Mammals', 'Reptiles',
'Animalia', 'Mollusks', 'Plants', 'Amphibians', 'Ray-finned Fishes', 'Birds', 'Insects',
'Fungi'].

Coarse-level
What is the name of the {concept_placeholder} that appears in this image? For example, if
it's a picture of a Owl Parrot, give a coarse-grained label for the image 'Parrot'. Provide
your answer after "Answer:".

Fine-level
What is the name of the {concept_placeholder} that appears in this image? For example,
if it's a picture of a Owl Parrot, give a coarse-grained label for the image 'Owl Parrot'.
Provide your answer after "Answer:".
```

```
Dataset:CUB-200-2011

Superordinate-level
What is the name of the organism that appears in this image? Provide your answer after
"Answer:" from one of the following categories: ['Arachnids', 'Mammals', 'Reptiles',
'Animalia', 'Mollusks', 'Plants', 'Amphibians', 'Ray-finned Fishes', 'Birds', 'Insects',
'Fungi'].

Coarse-level
What is the name of the {concept_placeholder} that appears in this image? For example, if
it's a picture of a Owl Parrot, give a coarse-grained label for the image 'Parrot'. Provide
your answer after "Answer:".

Fine-level
What is the name of the {concept_placeholder} that appears in this image? For example,
if it's a picture of a Owl Parrot, give a coarse-grained label for the image 'Owl Parrot'.
Provide your answer after "Answer:".
```

```
Dataset:Stanford Cars

Superordinate-level
What is the name of the object that appears in this image? Provide your answer after
"Answer:" from one of the following categories: ['Airplane', 'Car', 'Train', 'Bicycle', 'Cell
Phone', 'Plants', 'Dogs', 'Birds', 'Trucks'].

Coarse-level
What is the name of the {concept_placeholder} that appears in this image? Provide your answer
after "Answer:" from one of the following categories: ['Sedan', 'SUV', 'Coupe', 'Convertible',
'Pickup', 'Hatchback', 'Van']

Fine-level
What is the name of the {concept_placeholder} that appears in this image? For example, if
it's a picture of a 2006 Honda Civic LX Coupe, give a fine-grained label for the image '2006
Honda Civic LX Coupe'. Provide your answer after "Answer:".
```

Table 8: Prompts for fine-grained image classification. The {concept_placeholder} is replaced with upper-level concept labels as illustrated in Figure 3.

in Table 11 are used for the measuring of the modality gap experiments in Section 4.2.

B.4 Coarse-Grained Label Generation Prompts

The coarse-grained label generation prompts are shown in Table 12. We only generate the coarse-level labels for the following three datasets: (i) Stanford Dogs, (ii) Stanford Cars, (iii) CUB-200-2011, (iv) iNaturalist-2021 because they do not provide concept hierarchy like the rest of the other three datasets. For iNaturalist-2021, although the benchmark does provide the granularity hierarchy, it does so in a taxonomic manner, e.g., order, family, genus, species, which makes it challenging for the model to classify the coarse-grained categories; therefore, we generate coarse-grained labels for the dataset as well. By randomly selecting few-shot examples to guide the coarse-grained label generation, we ensure that the generative model, in this

Dataset Name	Total # of Instances (Test / Training)	# of Superordinate Categories	# of Coarse Categories	# of Fine Categories	# of Images Annotation	Granularity Hierarchy	Partial Attribute Annotation
iNaturalist-2021	100K / 2.6M	11	1,103	10K	3,286,843	✓	×
CUB-200-2011	5,794 / 5,994	1	59	200	11,788	×	✓
FGVC-Aircraft	3,333 / 6,667	1	30	100	10,000	✓	×
Stanford Dogs	8,580 / 12,000	1	-	120	20,580	×	×
NABirds	24,633 / 2,929	1	146	404	48,562	✓	×
Stanford Cars	8,144 / 8,041	1	7	196	16,185	×	×
FINER	372K/2.63M	16	1,416	11,171	3,393,958	✓	✓

Table 9: **Overview of the FGVC benchmarks.** Our **FINER** dataset demonstrates richer set of attributes per concept that enables the evaluation of fine-grained image comprehension. We also augment the benchmarks without **Granularity Hierarchy** with **Superordinate Categories** and **Coarse Categories**.

case GPT-4V, sticks to the generation of a coarse-grained label. For the faithfulness of the generated coarse-grained labels, we manually evaluate them for datasets other than iNaturalist-2021. For iNaturalist-2021, there are 1,103 coarse-grained categories, which makes it challenging to evaluate each generated coarse-grained label. We therefore group them together with their corresponding `family`-level category, which serves as a grouping category for the coarse-grained labels. For instance, for *Euphaea fraseri*, we place its coarse-grained labels, *Damselfly* and *Dragonfly* under *Euphaeidae*. By doing so, we not only provide room for more encompassing coarse-grained prediction, e.g., Damselflies are also classified as Dragonflies, but also distinguish the granularity setting from the fine-grained level, which requires a more specific categorization of a given species.

B.5 Linear Probing Experiment

To evaluate the quality of output representations before and after the multimodal projection in LLaVA-1.5 (7B), we perform linear probing over the output representations of the vision encoder (CLIP-ViT-L/14 (Radford et al., 2021)) and the projected representations in the textual space. The experimental settings are 10 epochs for finetuning and we use accuracy as the evaluation metric. We train on 4x V100 16GB for 57 GPU hours.

Dataset:**iNaturalist-2021**

Knowledge Probe Prompt
Can you guess the specific name (specific epithet) of an organism in the following taxonomic category given its physical attributes? Provide your answer after "Specific Epithet:".
Physical Attributes: {attribute_placeholder}

Supercategory: {supercategory_placeholder}
Kingdom: {kingdom_placeholder}
Phylum: {phylum_placeholder}
Class: {class_placeholder}
Order: {order_placeholder}
Family: {family_placeholder}
Genus: {genus_placeholder}
Specific Epithet:

Dataset:**FGVC-Aircraft**

Knowledge Probe Prompt
Can you guess the specific name (specific type) of an Airplane in the following taxonomic category given its physical attributes? Provide your answer after "Specific Airplane:".
Physical Attributes: {attribute_placeholder}

Supercategory: {supercategory_placeholder}
Coarse-grained Category: {coarse_placeholder}
Specific Airplane:

Dataset:**Stanford Dogs**

Knowledge Probe Prompt
Can you guess the specific name (specific type) of a Dog in the following taxonomic category given its physical attributes? Provide your answer after "Specific Dog:".
Physical Attributes: {attribute_placeholder}

Supercategory: {supercategory_placeholder}
Coarse-grained Category: {coarse_placeholder}
Specific Dog:

Table 10: Prompts for concept-related knowledge probing. The {concept_placeholder} is replaced with upper-level concept labels as illustrated in Figure 3.

Attribute Gen. (Text-Only)

```
What are useful visual features for distinguishing {concept_placeholder} in a photo? Provide
the answer as lists of required and likely attributes. For example, for a bengal tiger (Felis
Tigris) you might say:

Required:
- yellow to light orange coat
- dark brown to black stripes
- black rings on the tail
- inner legs and belly are white
- 21 to 29 stripes
Likely:
- lives in mangrove, wooded habitat
- amber, yellow eyes
- large, padded paws
- long tail
- stout teeth

'Required' attributes are a set of external, physical attributes that allows a human to
distinguish it from other similar looking concepts.
'Likely' attributes are a set of attributes that may or may not be visible or are not one of
the most discriminative features of the concept.
In the required (Required:) set, do not include relative, non-visual attributes like size or
weight, only the external, visually distinguishable attributes.
Provide your response in the above format, saying nothing else. If there are no useful visual
features, simply write "none".
```

Attribute Gen. (Wikipedia Doc; Web-Extracted)

```
What are useful visual, external features for distinguishing {concept_placeholder} in a photo?
Given an input document (Document:) that may talk about {concept_placeholder}, provide the
answer as lists of required and likely attributes. For example, for a bengal tiger (Felis
Tigris) you might say:

Required:
- yellow to light orange coat
- dark brown to black stripes
- black rings on the tail
- inner legs and belly are white
- 21 to 29 stripes

Likely:
- lives in mangrove, wooded habitat
- amber, yellow eyes
- large, padded paws
- long tail
- stout teeth

'Required' attributes are a set of external, physical attributes that allows a human to
distinguish it from other similar looking concepts.
'Likely' attributes are a set of attributes that may or may not be visible or are not one of
the most discriminative features of the concept.
In the required (Required:) set, do not include relative, non-visual attributes like size or
weight, only the external, visually distinguishable attributes.
If no document is given, generate from what you already know about {concept_placeholder}.
Provide your response in the above format, saying nothing else. If there are no useful visual
features, simply write "none".
```

Table 11: Prompts for Attribute Generation. The {concept_placeholder} is replaced with coarse-grained concept labels for Image-only case and Web-Extracted cases; for Text-only case, use the fine-grained concept label since we want to extract the attributes stored in the parametric knowledge by using the fine-grained concept label as a query.

```
Attribute Gen.  (Image-Only)
What are useful visual features for distinguishing the {concept_placeholder} in the photo?
Provide the answer as lists of required and likely attributes.  For example, for a bengal
tiger (Felis Tigris) you might say:

Required:
- yellow to light orange coat
- dark brown to black stripes
- black rings on the tail
- inner legs and belly are white
- 21 to 29 stripes

Likely:
- lives in mangrove, wooded habitat
- amber, yellow eyes
- large, padded paws
- long tail
- stout teeth

'Required' attributes are a set of external, physical attributes that allows a human to
distinguish it from other similar looking concepts.
'Likely' attributes are a set of attributes that may or may not be visible or are not one of
the most discriminative features of the concept.
In the required (Required:)  set, do not include relative, non-visual attributes like size or
weight, only the external, visually distinguishable attributes.
Provide your response in the above format, saying nothing else.  If there are no useful visual
features, simply write "none".
```

Table 11: Prompts for Attribute Generation. The {concept_placeholder} is replaced with coarse-grained concept labels for Image-only case and Web-Extracted cases; for Text-only case, use the fine-grained concept label since we want to extract the attributes stored in the parametric knowledge by using the fine-grained concept label as a query.

```
Dataset:Stanford Cars
Generate a coarse-grained label for the following fine-grained car types.
The coarse-grained car types are as follows:  ["sedan", "SUV", "coupe", "convertible",
"pickup", "hatchback", "van"].

For example, if the car is a "Ford F-150 Regular Cab 2012" generate "pickup", and if the
car is "Chrysler 300 SRT-8 2010", generate "sedan".
Output format is as follows:

Fine-grained Car Name:  Ford F-150 Regular Cab 2012
Car Name:  pickup

Fine-grained Car Name:  Chrysler 300 SRT-8 2010
Car Name:  sedan

Fine-grained Car Name:  Hyundai Santa Fe 2008
Car Name:  SUV
```

```
Dataset:CUB-200-2011
Generate a coarse-grained label for the following fine-grained bird types.
For example, if the bird is a "bald eagle (Haliaeetus leucocephalus)" generate "Eagle", and if
the bird is "Pine grosbeak", generate "Finch".
Output format is as follows:

Fine-grained Bird Name:  Bald eagle
Bird Name:  Eagle

Fine-grained Bird Name:  Pine grosbeak
Bird Name:  Finch

Fine-grained Bird Name:  The black backed woodpecker
Bird Name:  Woodpecker
```

```
Dataset:Stanford Dogs
Generate a coarse-grained label for the following fine-grained dog types.
For example, if the dog is a "Cavalier King Charles Spaniel" generate "Spaniel", and if the
dog is "The Dear-Headed Chihuahua", generate "Chihuahua".
Output format is as follows:

Fine-grained Dog Name:  Cavalier King Charles Spaniel
Dog Name:  Spaniel

Fine-grained Dog Name:  Curly-coated retriever
Dog Name:  Retriever

Fine-grained Dog Name:  Newfoundland
Dog Name:  Newfoundland
```

Table 12: Prompts for Coarse-grained Label Generation. We generate the coarse-grained labels for each of the dataset, in case the dataset does not provide the concept hierarchy. The few-shot sampled are randomly sampled from the training instances.