北京大學 PEKING UNIVERSITY

北京科技大学 University of Science and Technology Beijing

Analyzing and Boosting the Power of Fine-Grained Visual Recognition for Multi-modal Large Language Models

ICLR

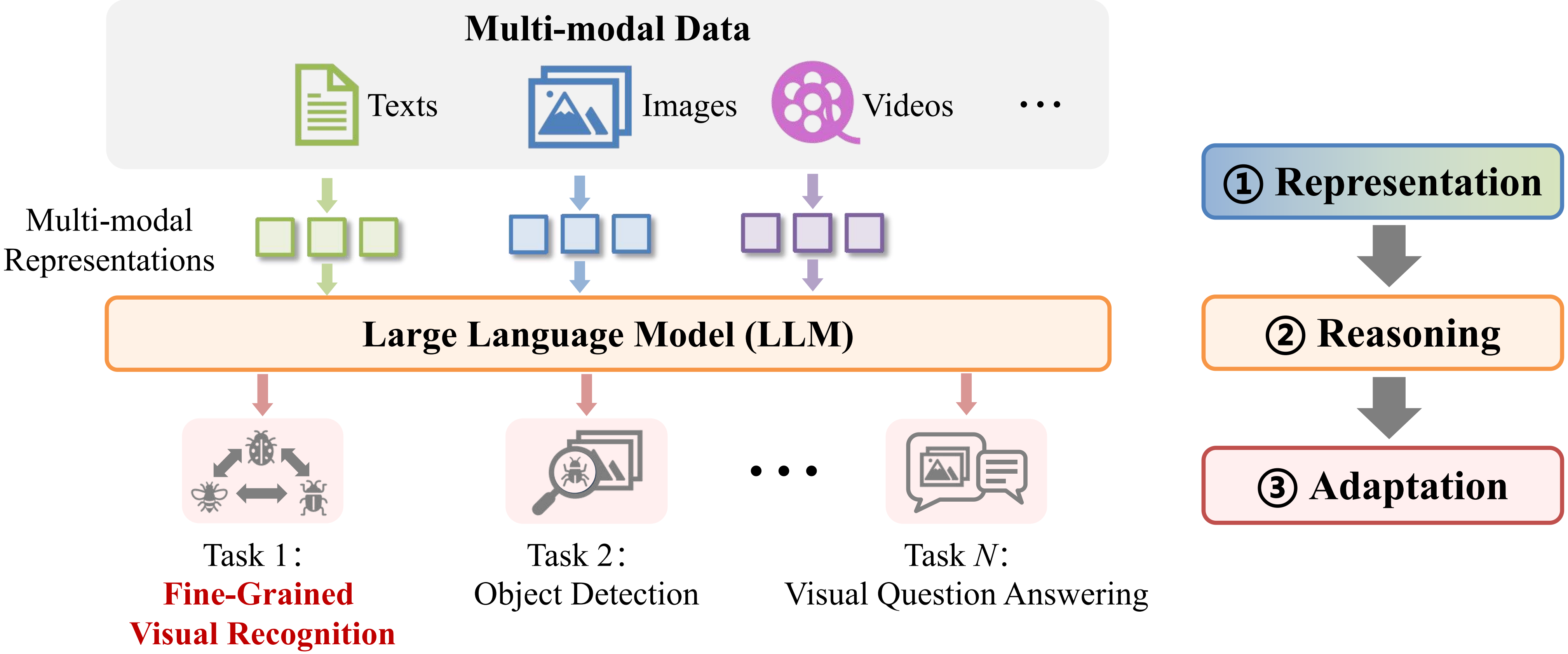
Hulingxiao He[1], Geng Li[1], Zijun Geng[1], Jinglin Xu[2], and Yuxin Peng*[1]

[1]Wangxuan Institute of Computer Technology, Peking University

[2]School of Intelligence Science and Technology, University of Science and Technology Beijing

Multi-modal Large Language Models

- **Multi-modal Large Language Models** (MLLMs) refer to foundational models that extract and integrate **representations** from multi-modal data such as texts, images, and videos, perform **reasoning** through Large Language Models (LLMs), and are fine-tuned to **adapt to** various downstream tasks like **Fine-Grained Visual Recognition**



Fine-Grained Visual Recognition

- **Fine-Grained Visual Recognition** (FGVR) aims at identifying **subordinate-level categories**, such as specific bird species and aircraft model

Poor FGVR Performance of MLLMs

- The recognition ability of MLLMs relies heavily on extensive training data. Due to **the high cost of annotating subordinate-level categories** in training data, MLLMs often **lack FGVR capabilities**

Which of these **birds** is shown in the image?

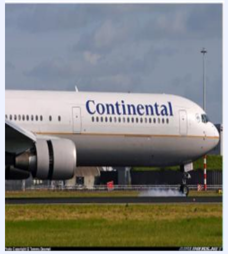
A. Common Raven
B. Groove billed Ani
C. American Crow
D. Shiny Cowbird

C. American Crow — **Idefics2**

Fail to identify the subordinate-level category (Ground Truth: B. Groove billed Ani) ✗

Which of these **aircrafts** is shown in the image?

A. 767-200
B. 767-300
C. 767-400
D. 767-500

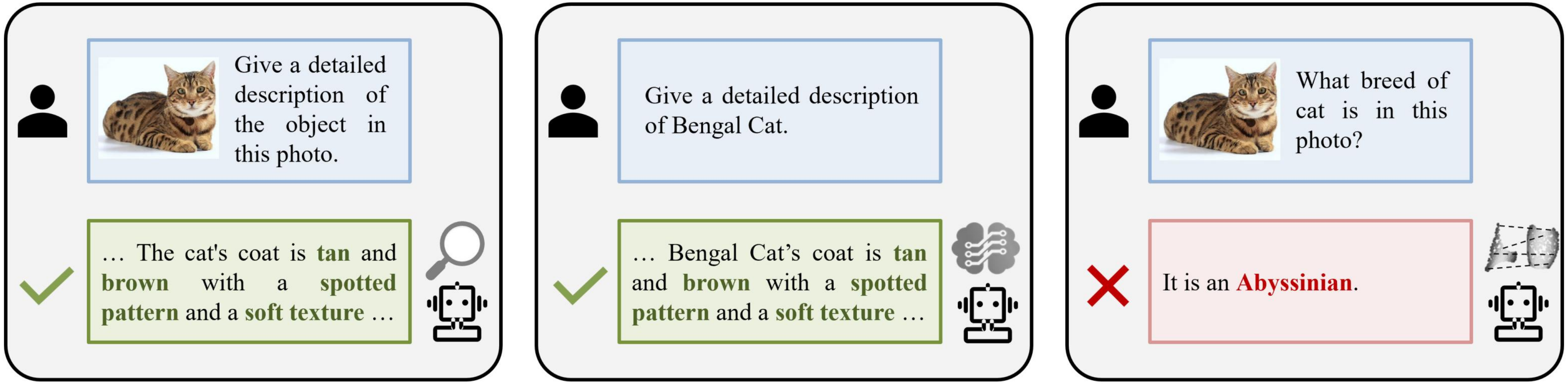


B. 767-300 — **Idefics2**

Fail to identify the subordinate-level category (Ground Truth: C. 767-400) ✗

Problem Analysis (1/4): Three Quintessential Capabilities

- We revisit **three quintessential capabilities** of MLLMs for FGVR
  (a) **Object Information Extraction**: Accurately and fully extracting the necessary information for distinguishing objects
  (b) **Category Knowledge Reserve**: Reserving sufficient knowledge of subordinate-level categories
  (c) **Object-Category Alignment**: Aligning visual objects and category names in the representation space to enhance classification performance



(a) Object Information Extraction (b) Category Knowledge Reserve (c) Object-Category Alignment

Problem Analysis (2/4): Object Information Extraction

(a) Object information lost exists between Vision Language Models (VLMs) and MLLMs but is not the bottleneck

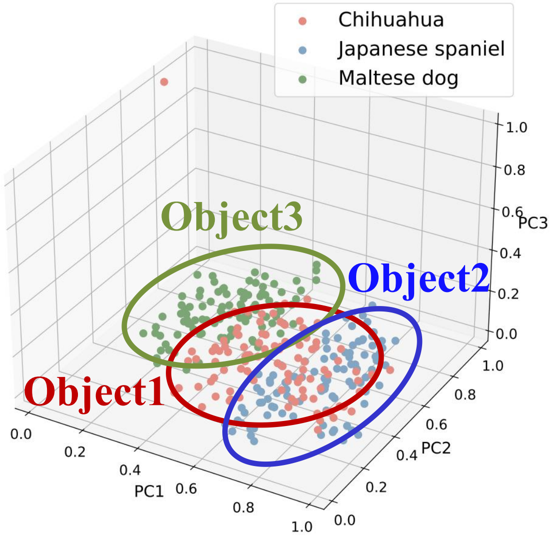
(a) Object features.

Model	Feature Type	Acc.
Idefics2	Last Avg.	94.99 90.24
SigLIP	CLS Avg.	95.28 94.44

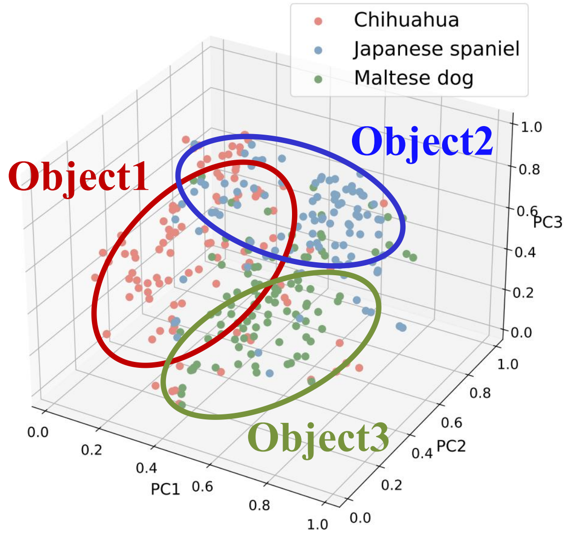
Idefics2 → Object features passed through the vision encoder, modality connector, and LLM

SigLIP → Object features output from last layer of vision encoder

Object dist. of SigLIP (VLM)



Object dist. of Idefics2 (MLLM)



Problem Analysis (3/4): Category Knowledge Reserve

(b) Category knowledge is sufficient, but category names can't fully capture semantics

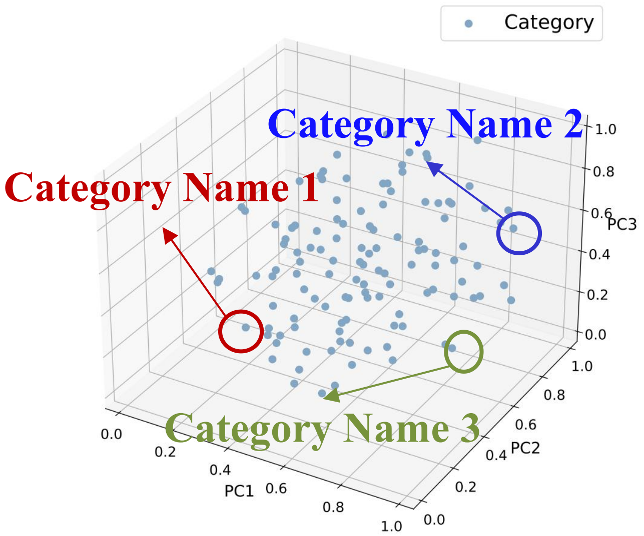
(b) Category description features.

Model	Feature Type	Acc.
Idefics2	Last Avg.	92.51 90.41
SigLIP	CLS Avg.	84.70 87.78

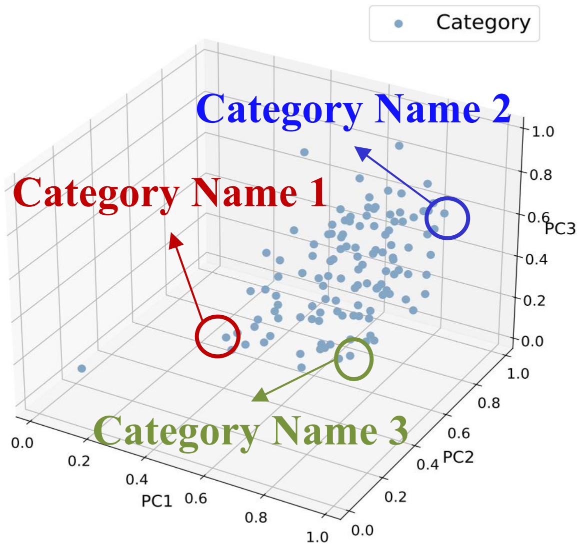
Category description features passed through the vision encoder, modality connector, and LLM

Category description features output from last layer of vision encoder

Category dist. of SigLIP (VLM)



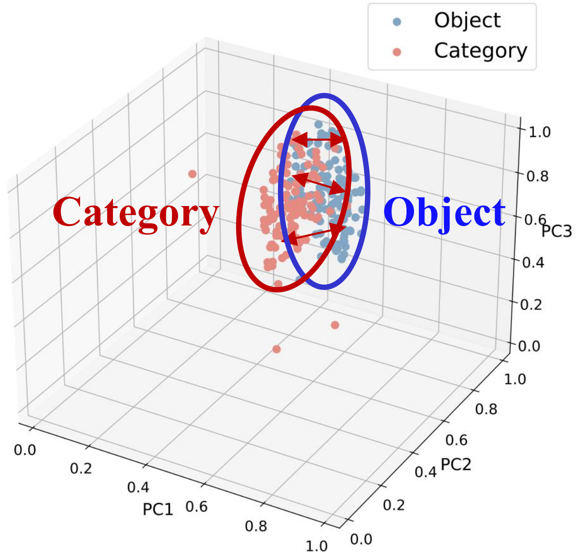
Category dist. of Idefics2 (MLLM)



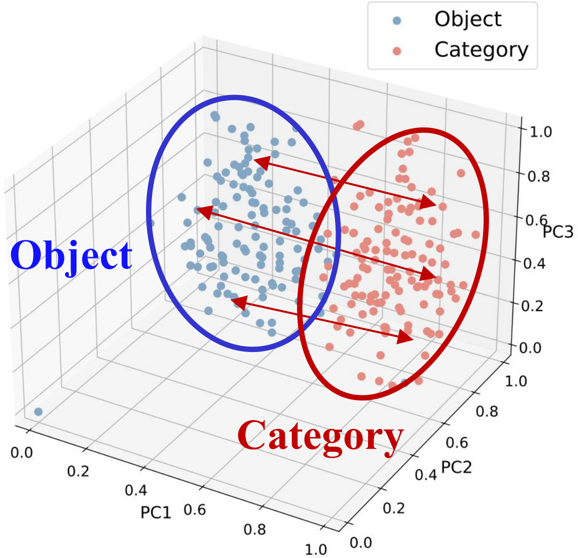
Problem Analysis (4/4): Object-Category Alignment

(c) **Misalignment between the visual object and category name** leads to underperformance

Object-category dist. of SigLIP (VLM)



Object-category dist. of Idefics2 (MLLM)



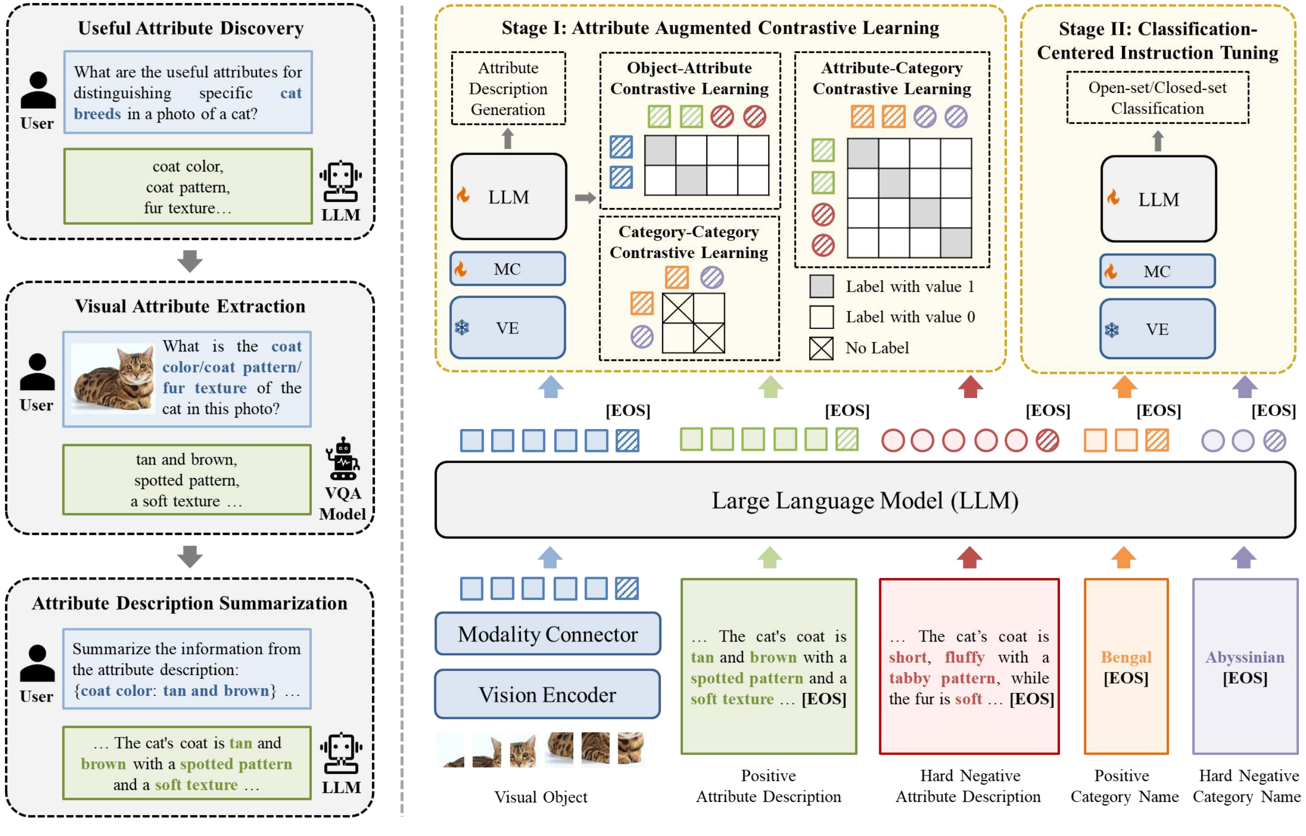
(a) Object Information Extraction ✓

(b) Category Knowledge Reserve ✓

(c) Object-Category Alignment ✗

Method (1/4): The Framework to Build Finedefics

- (a) **Attribute Description Construction**, which aims to obtain informative attribute descriptions of objects. (b) **Attribute Augmented Alignment**, which aims to use constructed attribute descriptions to bind visual objects and category names, thus enhancing the model's FGVR capability via a two-stage training paradigm



(a) Attribute Description Construction (b) Attribute Augmented Alignment

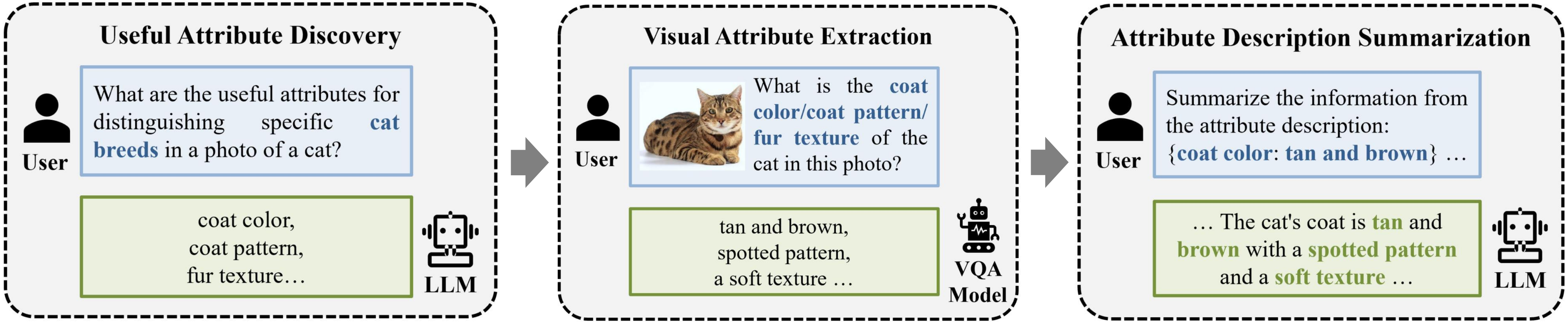
Method (2/4): Attribute Description Construction

- Extracting useful attribute information that can distinguish different categories

  ①**Useful Attribute Discovery**: Obtaining useful attributes for distinguishing subordinate-level categories

  ②**Visual Attribute Extraction**: Extracting attribute key-value pairs for the visual object in the image

  ③**Attribute Description Summarization**: Summarizing the attribute key-value pairs into detailed attribute descriptions

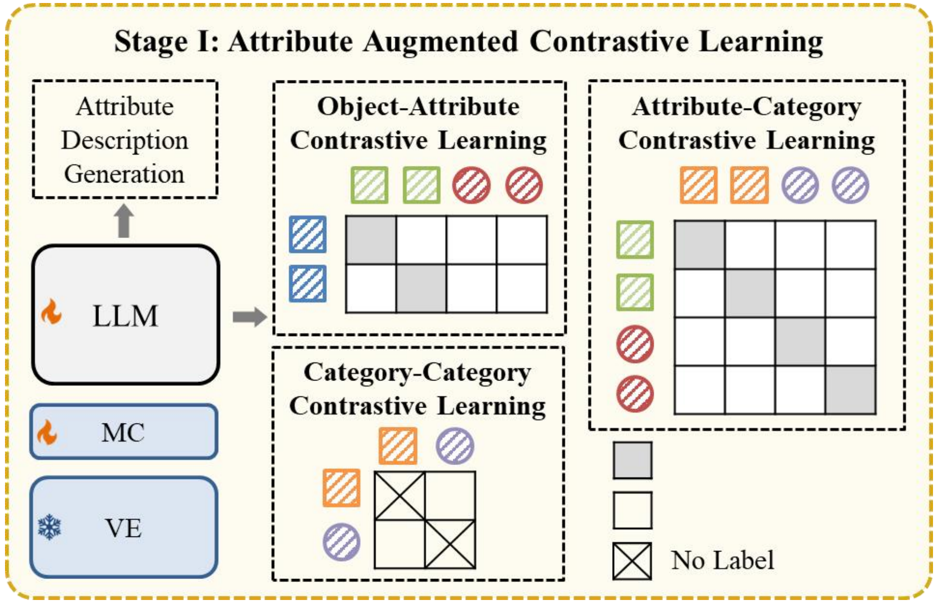


Method (3/4): Attribute Augmented Contrastive Learning

- Use Object-Attribute, Attribute-Category, and Category-Category Contrastive loss to **bind visual objects and categories names** in the representation space of LLMs with **attribute descriptions as an intermediate point**

Total Loss: Attribute Description Generation loss

$$\mathcal{O}^{\mathrm{I}}_{\beta,\theta} = \arg\min_{\beta,\theta} \boxed{\mathcal{L}_G^{\mathrm{att}}} + (\boxed{\mathcal{L}^{hn}_{OAC}} + \boxed{\mathcal{L}^{hn}_{ACC}} + \boxed{\mathcal{L}_{CCC}})/2,$$



$$\mathcal{L}^{hn}_{OA} = \sum_{(\hat{o}^i,\hat{a}^i,\hat{c}^i)\in\mathcal{B}} -\log \frac{\exp^{Sim(\hat{o}^i,\hat{a}^i)}}{\sum_{\hat{a}^j\in\mathcal{B}} \exp^{Sim(\hat{o}^i,\hat{a}^j)} + \sum_{\hat{a}^w\in\mathcal{A}^i_{hn}} \exp^{Sim(\hat{o}^i,\hat{a}^w)}},$$

$$\mathcal{L}_{AO} = \sum_{(\hat{o}^i,\hat{a}^i,\hat{c}^i)\in\mathcal{B}} -\log \frac{\exp^{Sim(\hat{o}^i,\hat{a}^i)}}{\sum_{\hat{o}^k\in\mathcal{B}} \exp^{Sim(\hat{o}^k,\hat{a}^i)}},$$

Object-Attribute Contrastive loss

$$\boxed{\mathcal{L}^{hn}_{OAC}} = (\mathcal{L}^{hn}_{OA} + \mathcal{L}_{AO})/2,$$

$$\mathcal{L}^{hn}_{AC} = \sum_{(\hat{o}^i,\hat{a}^i,\hat{c}^i)\in\mathcal{B}} -\log \frac{\exp^{Sim(\hat{a}^i,\hat{c}^i)}}{\sum_{\hat{c}^j\in\mathcal{B}} \exp^{Sim(\hat{a}^i,\hat{c}^j)} + \sum_{\hat{c}^w\in\mathcal{C}^i_{hn}} \exp^{Sim(\hat{a}^i,\hat{c}^w)}},$$

$$\mathcal{L}^{hn}_{CA} = \sum_{(\hat{o}^i,\hat{a}^i,\hat{c}^i)\in\mathcal{B}} -\log \frac{\exp^{Sim(\hat{a}^i,\hat{c}^i)}}{\sum_{\hat{a}^j\in\mathcal{B}} \exp^{Sim(\hat{a}^j,\hat{c}^i)} + \sum_{\hat{a}^w\in\mathcal{A}^i_{hn}} \exp^{Sim(\hat{a}^w,\hat{c}^i)}},$$

Attribute-Category Contrastive loss

$$\boxed{\mathcal{L}^{hn}_{ACC}} = (\mathcal{L}^{hn}_{AC} + \mathcal{L}^{hn}_{CA})/2,$$

Category-Category Contrastive loss

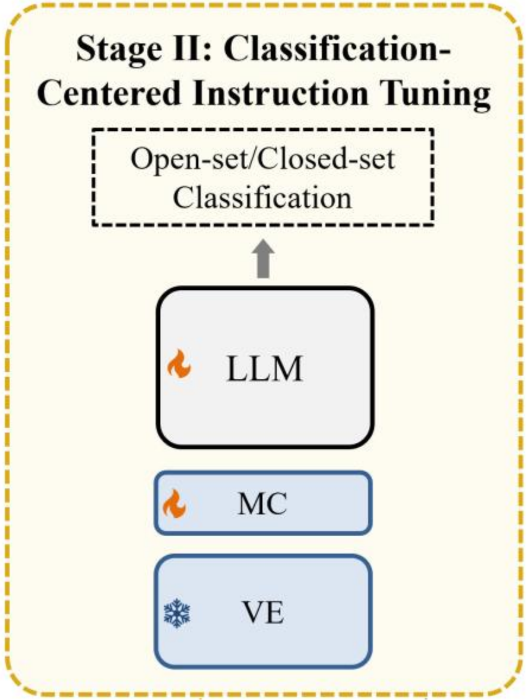
$$\boxed{\mathcal{L}_{CCC}} = \sum_{(\hat{o}^i,\hat{a}^i,\hat{c}^i)\in\mathcal{B}} -\log \frac{1}{\sum_{\hat{c}^k\in\mathcal{C}^i_{hn}} \exp^{Sim(\hat{c}^i,\hat{c}^k)}}.$$

Method (4/4): Classification-Centered Instruction Tuning

- Formulate FGVR datasets as **open-set QA data** and **closed-set multiple-choice data**, and finetune the model using this **classification-centered instruction tuning** data

the generation loss of classification-centered instruction tuning data

$$\mathcal{O}^{\mathrm{II}}_{\beta,\theta} = \arg\min_{\beta,\theta} \boxed{\mathcal{L}_G^{\mathrm{cls}}}$$



What is the species of the bird shown in the image?

Groove billed Ani

open-set QA data

Which of these birds is shown in the image?

A. Common Raven
B. Groove billed Ani
C. American Crow
D. Shiny Cowbird

B

closed-set multiple-choice data

Experiments (1/3): Main Results

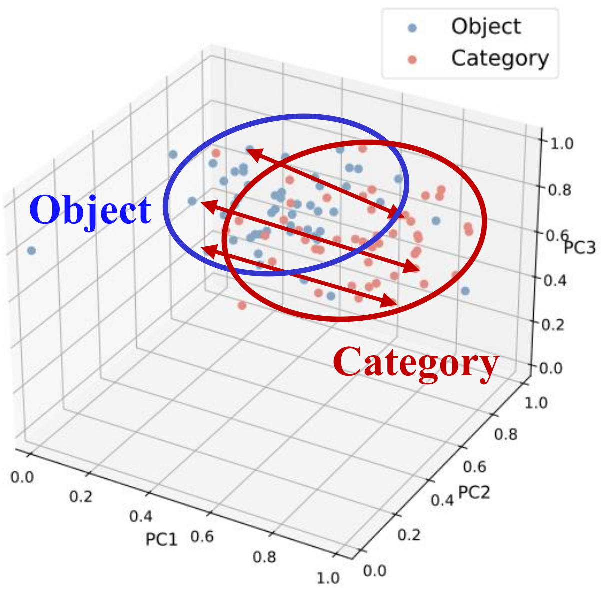
- On six FGVR datasets, the average accuracy of Finedefics reached **76.84%**, which is a **9.43%** improvement compared to **Qwen-VL-Chat** released by **Alibaba** in January 2024

Model	#P	Dog-120	Bird-200	Aircraft-102	Flower-102	Pet-37	Car-196	Avg.
LLaVA 1.5	7B	38.96	35.24	34.71	51.37	52.25	46.92	43.24
LLaVA-Next (Mistral)	7B	38.86	34.88	32.49	43.91	53.72	49.48	42.22
MobileVLM v2	7B	39.92	33.90	35.01	54.89	53.69	46.29	43.95
InstructBLIP Vicuna	7B	41.60	32.78	31.68	50.90	54.92	48.25	43.36
InstructBLIP Flan-T5-XL	4B	47.10	32.15	29.19	62.29	59.99	64.58	49.22
Phi-3-Vision	4B	39.80	37.63	42.33	51.59	56.36	54.50	47.04
BLIP2 Flan-T5-XL	4B	46.17	33.70	32.94	64.32	65.00	67.68	51.64
InternLM XComposer 2	7B	41.47	37.42	40.53	54.25	63.23	53.89	48.47
Pali-Gemma	3B	51.68	36.62	39.87	69.64	75.42	64.64	56.31
Idefics1	9B	39.74	36.50	34.62	51.70	48.51	29.42	40.08
Idefics2	8B	57.96	47.17	<u>56.23</u>	72.78	81.28	<u>80.25</u>	65.95
Qwen-VL-Chat	10B	<u>66.18</u>	<u>52.30</u>	45.96	<u>75.95</u>	<u>87.82</u>	76.23	<u>67.41</u>
Finedefics (ours)	8B	**72.86** (+6.68)	**57.61** (+5.31)	**63.82** (+7.59)	**89.88** (+13.93)	**92.18** (+4.36)	**84.67** (+4.42)	**76.84** (+9.43)

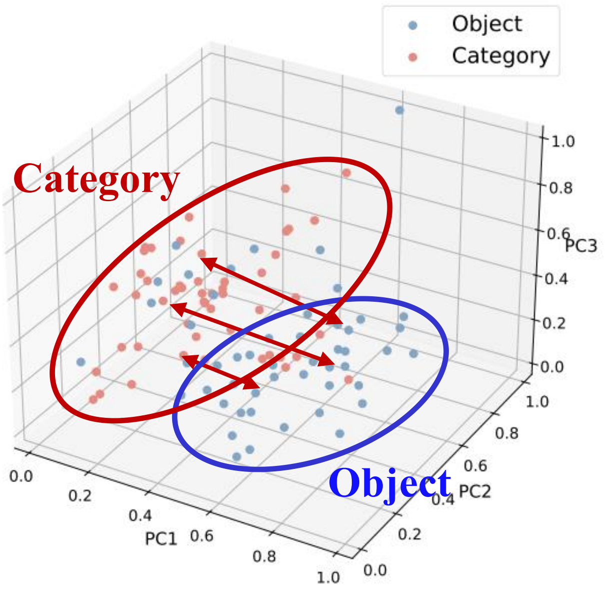
Experiments (2/3): Visualization

- With the usage of **contrastive learning on object-attribute-category triples**, Finedefics effectively **aligns visual objects and category names**, thus boosting FGVR accuracy

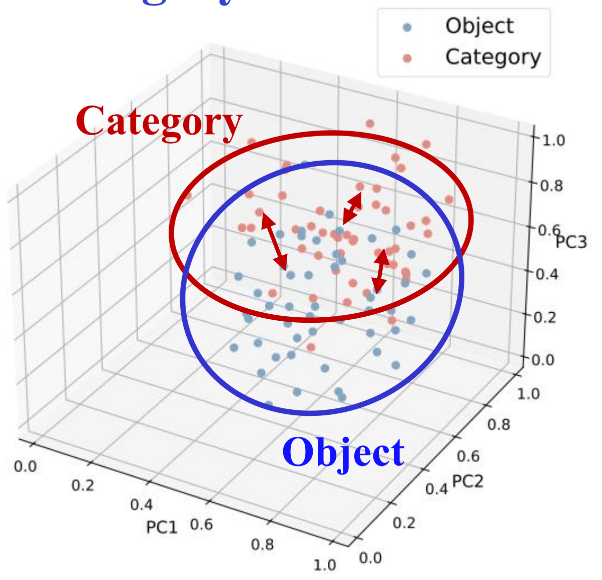
①**Finetune** only



②Incorporate **Object-Category** Contrastive loss



③Incorporate **Object-Attribute**, **Attribute-Category**, **Category-Category** Contrastive loss



Distribution Gap > **Distribution Gap** > **Distribution Gap**

Experiments (3/3): Case Study

- Compared to **Idefics2** released by **Hugging Face**, Finedefics successfully **captures the nuance of the object features**, setting them apart from visually similar subordinate-level categories

Options: Common Raven
Groove billed Ani
American Crow
Shiny Cowbird

Idefics2: **American Crow**
Finedefics (ours): **Groove billed Ani**

(a) Bird-200

Options: Pomeranian
keeshond
Samoyed
Eskimo dog

Idefics2: **Eskimo dog**
Finedefics (ours): **Samoyed**

(b) Dog-120

Options: Audi TT Hatchback 2011
Audi A5 Coupe 2012
Audi TTS Coupe 2012
Audi S5 Coupe 2012

Idefics2: **Audi A5 Coupe 2012**
Finedefics (ours): **Audi TTS Coupe 2012**

(c) Car-196

Options: 767-200
767-300
777-300
767-400

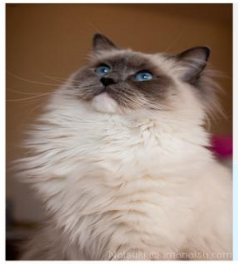
Idefics2: **767-300**
Finedefics (ours): **767-400**

(d) Aircraft-102

Options: tiger lily
fire lily
sword lily
peruvian lily

Idefics2: **fire lily**
Finedefics (ours): **peruvian lily**

(e) Flower-102



Options: Ragdoll
Birman
Maine Coon
Siamese

Idefics2: **Birman**
Finedefics (ours): **Ragdoll**

(f) Pet-37

Conclusion

- We investigate why MLLMs struggle with fine-grained visual recognition (FGVR) and position the problem as **the misalignment** between visual objects and category names
- We propose **Attribute Augmented Alignment**, designed to use attribute descriptions as an intermediate point to bind visual objects and category names
- Based on the aligned representation space, we build **Finedefics**, a new MLLM adept at identifying the subordinate-level category of the visual object.
- Our experiments conducted on six popular FGVR datasets, demonstrate the **remarkable performance** of Finedefics.

Thank you for listening!

- Code and model are available now, welcome to follow our work!
  - **Paper**： https://arxiv.org/abs/2501.15140
  - **Code**： https://github.com/PKU-ICST-MIPL/Finedefics_ICLR2025
  - **Model**： https://huggingface.co/StevenHH2000/Finedefics
  - **Lab**: https://www.wict.pku.edu.cn/mipl

【Paper】 【Code】 【Model】 【Lab】