Democratizing Fine-grained Visual Recognition with Large Language Models

ICLR

Mingxuan Liu[1], Subhankar Roy[4], Wenjing Li[3,6]*

Zhun Zhong[3,5]*, Nicu Sebe[1], Elisa Ricci[1,2]

[1]University of Trento [2]Fondazione Bruno Kessler [3]Hefei University of Technology

[4]University of Aberdeen [5]University of Nottingham [6]University of Leeds

* corresponding authors

To recognize a common object, we now can …

What is the name of the main object in this photo?

BLIP-2[1]: Pizza! ✔

LENS[3]: Pizza! ✔

GT: Pizza

LLaVA-1.5[2]: Pizza! ✔

MiniGPT-4[4]: Pizza! ✔

[1] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Arxiv*, 2023
[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Arxiv*, 2023
[3] William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. Towards language models that can see: Computer vision through the lens of natural language. *Arxiv*, 2023
[4] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *Arxiv*, 2023

But, let’s imagine a case …

A curious boy encountered a unique challenge when collecting several unlabeled images from a smartphone located in the Amazon jungle. Tasked with identifying the diverse bird species within these images, the boy faced a daunting task, especially without any prior knowledge of species names typically provided by ornithologists.

Can the modern VLM systems help him?

What is the name of the main object in this photo?

BLIP-2[1]:
Sparrow!

LENS[3]:
Vesper Sparrow!

LLaVA-1.5[2]:
Sparrow!

GT: Lincoln’s Sparrow

MiniGPT-4[4]:
White-throated Swainson Sparrow!

[1] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Arxiv*, 2023
[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Arxiv*, 2023
[3] William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. Towards language models that can see: Computer vision through the lens of natural language. *Arxiv*, 2023
[4] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *Arxiv*, 2023

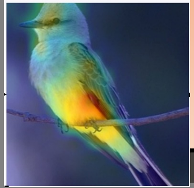
What is the name of the main object in this photo?

BLIP-2[1]**:**
Sparrow!

3:
r Sparrow!

Not good : (
Foundational VLMs
Struggle with identifying
Fine-grained concepts

Attention is needed!

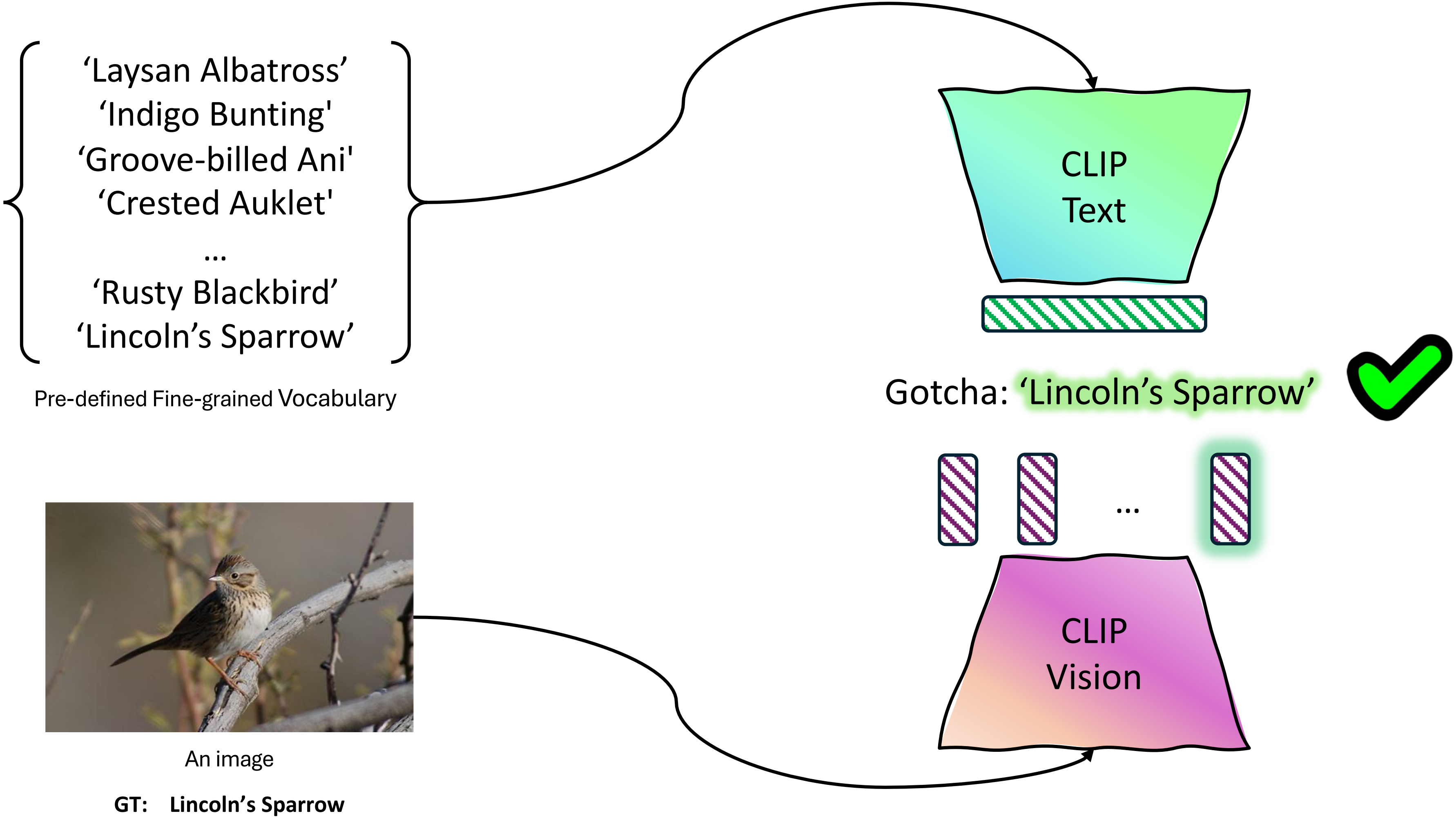


LLaVA-1.5[2]**:**
Sparrow!

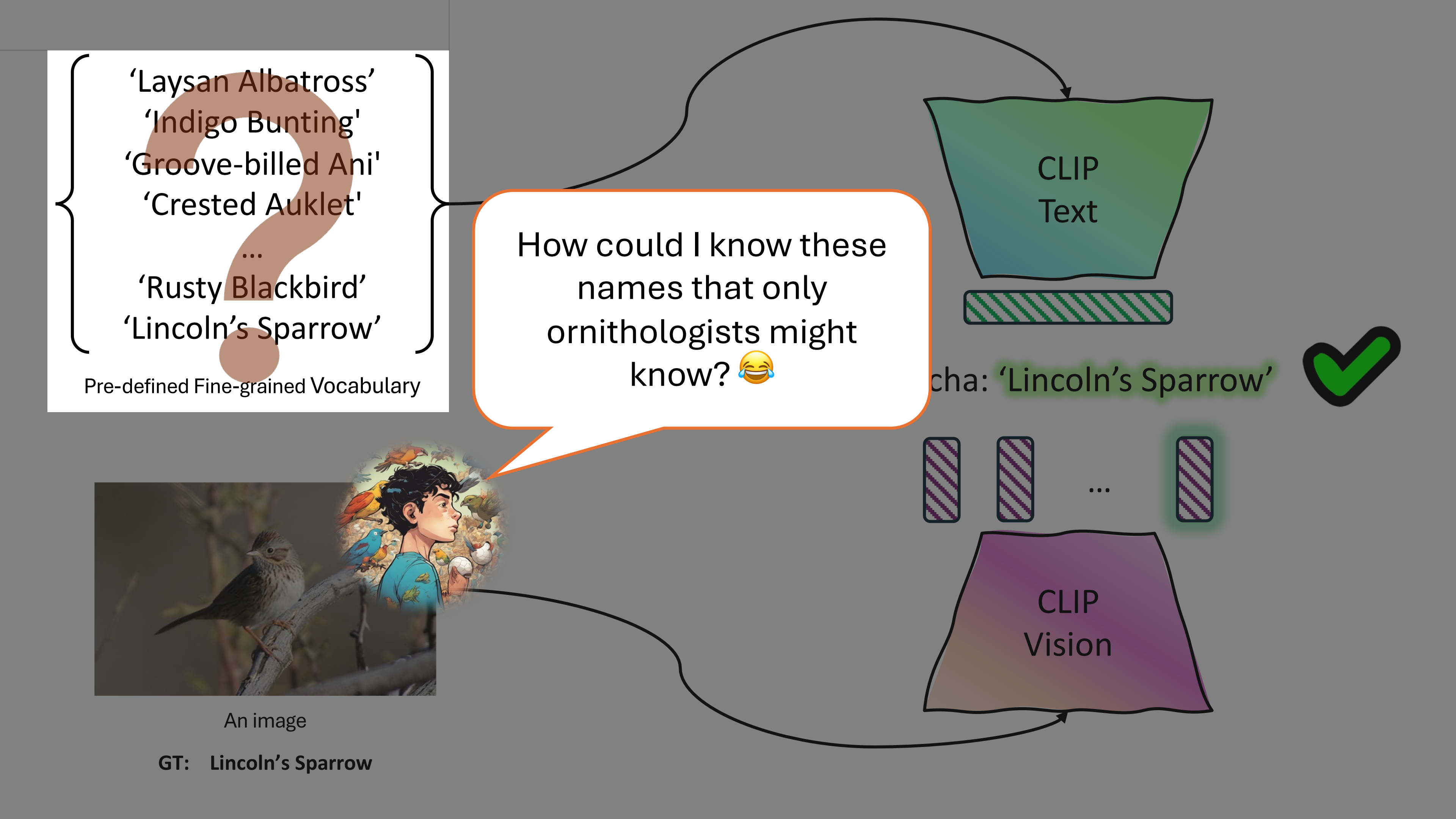
GT: Lincoln's Sparrow

MiniGPT-4[4]**:**
White-throated Swainson Sparrow!

[1] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Arxiv*, 2023
[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Arxiv*, 2023
[3] William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. Towards language models that can see: Computer vision through the lens of natural language. *Arxiv*, 2023
[4] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *Arxiv*, 2023



‘Laysan Albatross’
‘Indigo Bunting'
‘Groove-billed Ani'
‘Crested Auklet'
...
‘Rusty Blackbird’
‘Lincoln’s Sparrow’
Pre-defined Fine-grained Vocabulary
CLIP Text
Gotcha: ‘Lincoln’s Sparrow’
...
CLIP Vision
An image
GT: Lincoln’s Sparrow



'Laysan Albatross'
'Indigo Bunting'
'Groove-billed Ani'
'Crested Auklet'
...
'Rusty Blackbird'
'Lincoln's Sparrow'
Pre-defined Fine-grained Vocabulary
How could I know these names that only ornithologists might know? 😂
CLIP Text
cha: 'Lincoln's Sparrow'
...
CLIP Vision
An image
GT: Lincoln's Sparrow

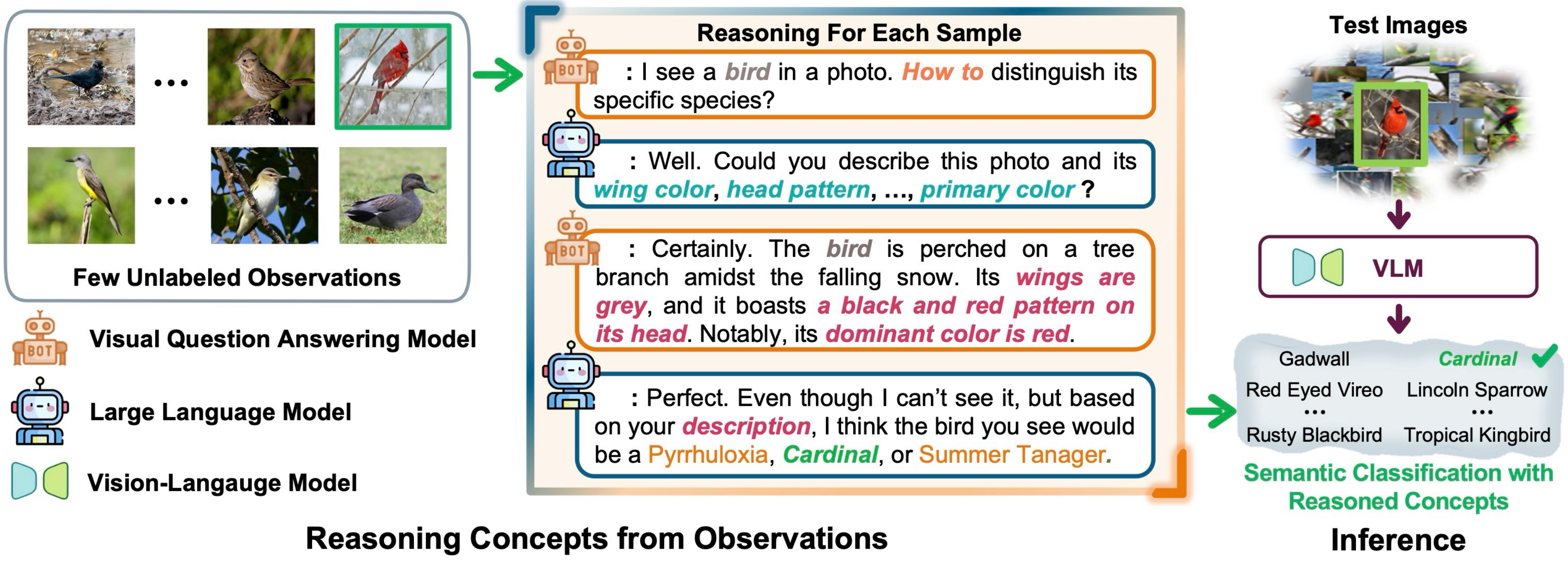
Can we build an effective system that can automatically discover fine-grained concepts (names) from few unlabeled observations and thereby classify them?

Problem Formulation

This is essentially a Vocabulary-free FGVR task with only few unlabeled samples as observation

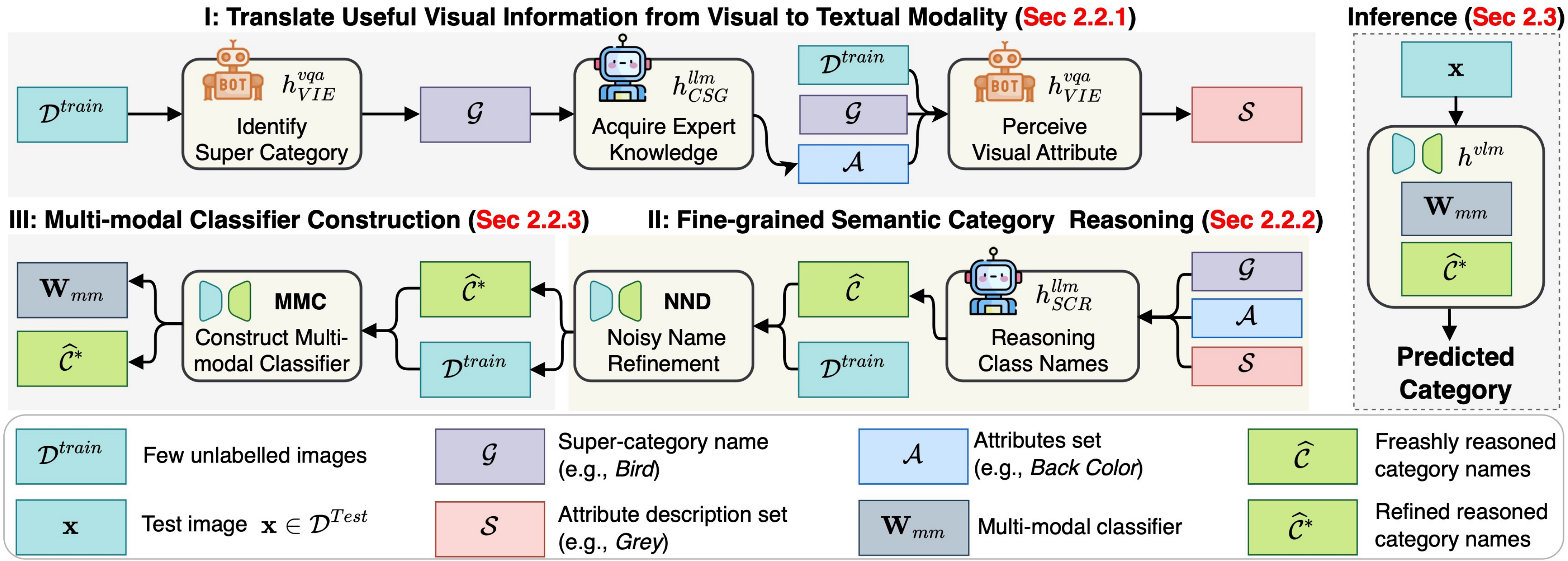
Method

We proposed **FineR** :
A Fine-grained Semantic Category Reasoning System with LLMs that reason fine-grained concepts from few observation and thereby facilitate vocabulary-free FGVR

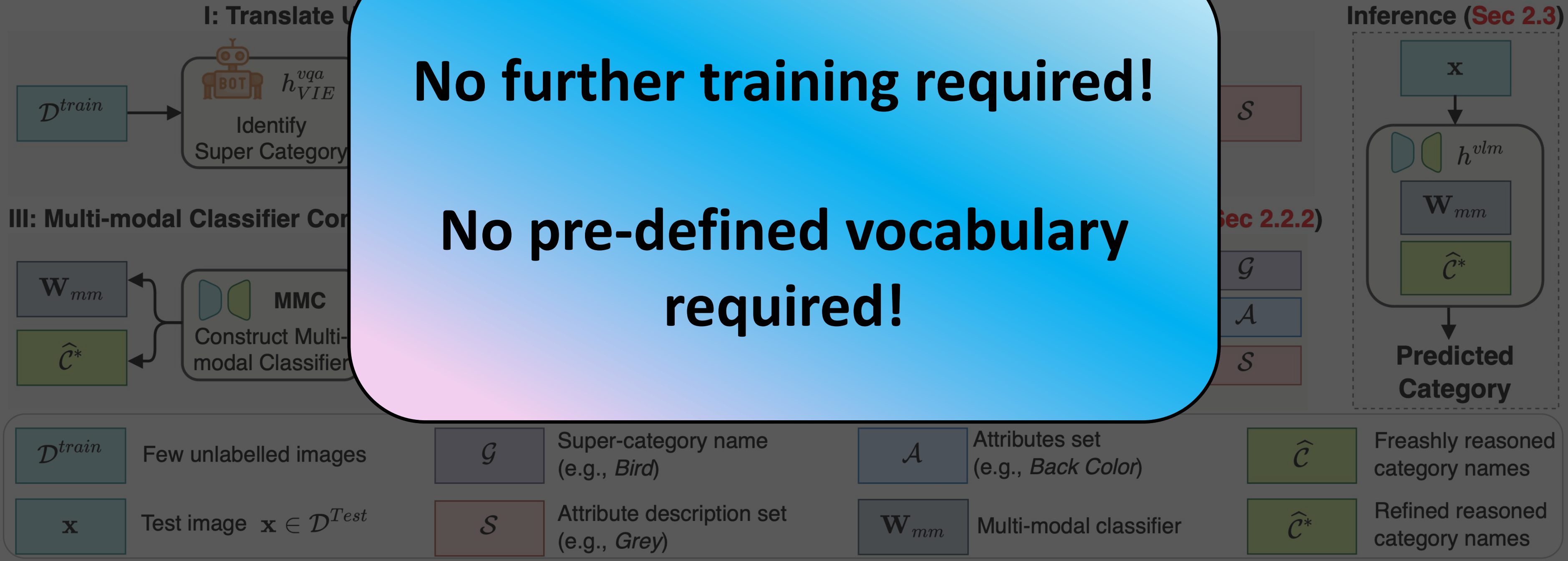


Just like what human would do …

Overview of FineR System



Overview of FineR System



No further training required!

No pre-defined vocabulary required!

Experimental Results

Evaluation Metrics

1. Are they semantically close?

Semantic Similarity (sACC):
Cosine similarity of embeddings of predicted label vs GT

Clustering Accuracy (cACC)
Hungarian match between clusters of predictions vs GT clusters

2. Do samples of the same category get predicted with the same label?

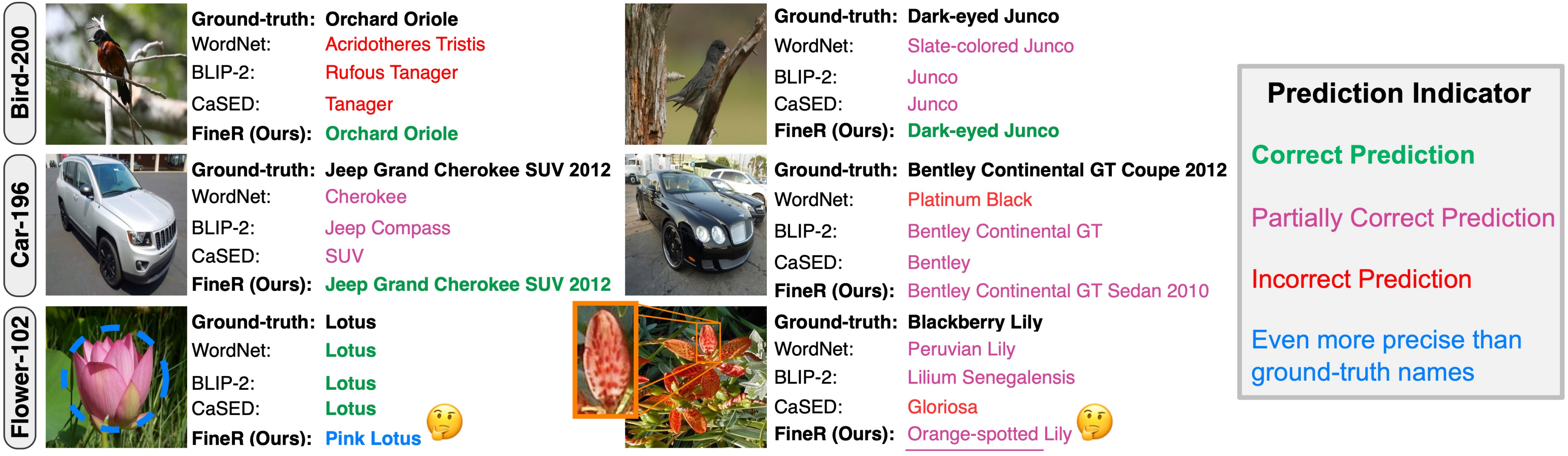
Quantitative Results

vs. SOTAs

	Bird-200		Car-196		Dog-120		Flower-102		Pet-37		Average	
	cACC	sACC	cACC	sACC	cACC	sACC	cACC	sACC	cACC	sACC	cACC	sACC
Zero-shot (UB)	57.4	80.5	63.1	66.3	56.9	75.5	69.7	77.8	81.7	87.8	65.8	77.6
CLIP-Sinkhorn	23.5	-	18.1	-	12.6	-	30.9	-	23.1	-	21.6	-
DINO-Sinkhorn	13.5	-	7.4	-	11.2	-	17.9	-	5.2	-	19.1	-
KMeans	36.6	-	30.6	-	16.4	-	66.9	-	32.8	-	36.7	-
WordNet	39.3	57.7	18.3	33.3	53.9	**70.6**	42.1	49.8	55.4	61.9	41.8	54.7
BLIP-2	30.9	56.8	43.1	57.9	39.0	58.6	61.9	**59.1**	61.3	60.5	47.2	58.6
CLEVER †	7.9	-	-	-	-	-	6.2	-	-	-	-	-
SCD †	46.5	-	-	-	**57.9**	-	-	-	-	-	-	-
CaSED	25.6	50.1	26.9	41.4	38.0	55.9	**67.2**	52.3	60.9	63.6	43.7	52.6
FineR (Ours)	**51.1**	**69.5**	**49.2**	**63.5**	48.1	64.9	63.8	51.3	**72.9**	**72.4**	**57.0**	**64.3**

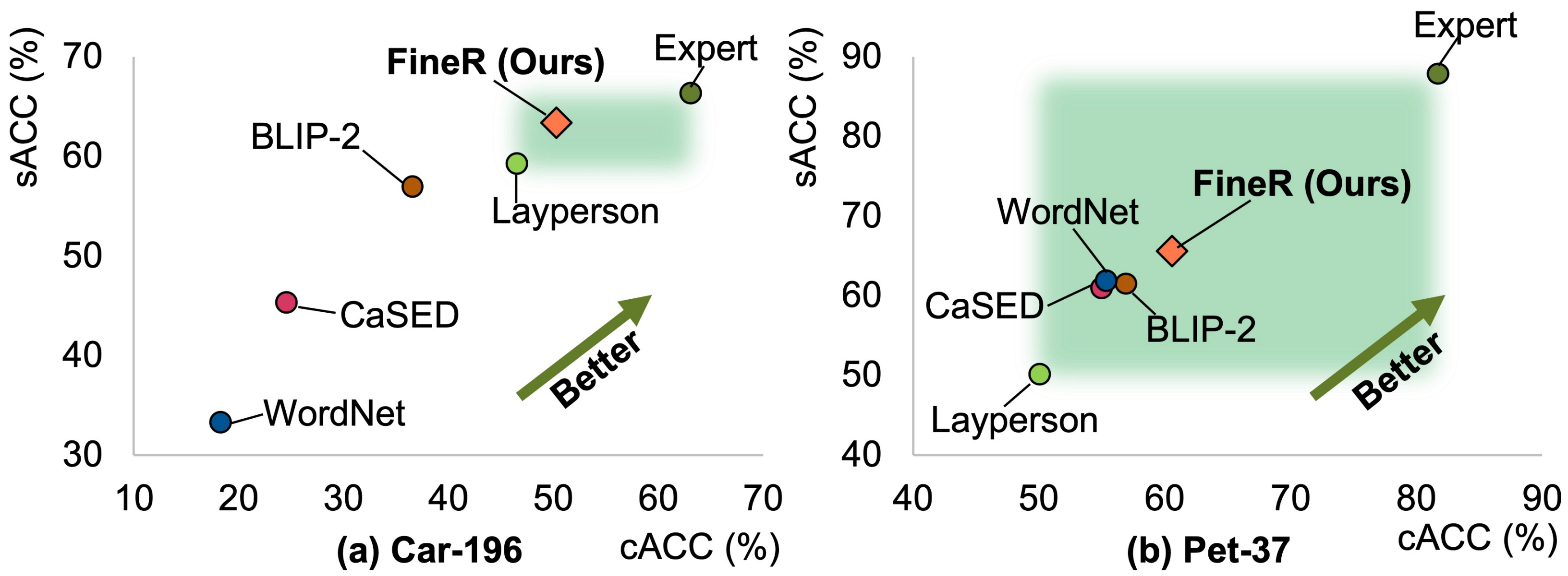
Table 1: cACC(%) and sACC (%) comparison on the five fine-grained datasets. $|\mathcal{D}_c^{\text{train}}| = 3$. Results reported are averaged over 10 runs. †: SCD and CLEVER results are quoted from original paper (SCD uses the *entire* dataset for class name discovery and assumes the number of classes known as *a-priori*). Best and second-best performances are coloured **Green** and Red, respectively. Gray presents the upper bound (UB).

- Given 3 images per class for discovery, FineR outperforms the 2nd-best model by **+9.8%** in cACC and **+5.7%** in sACC on the five fine-grained datasets



Qualitative Results:

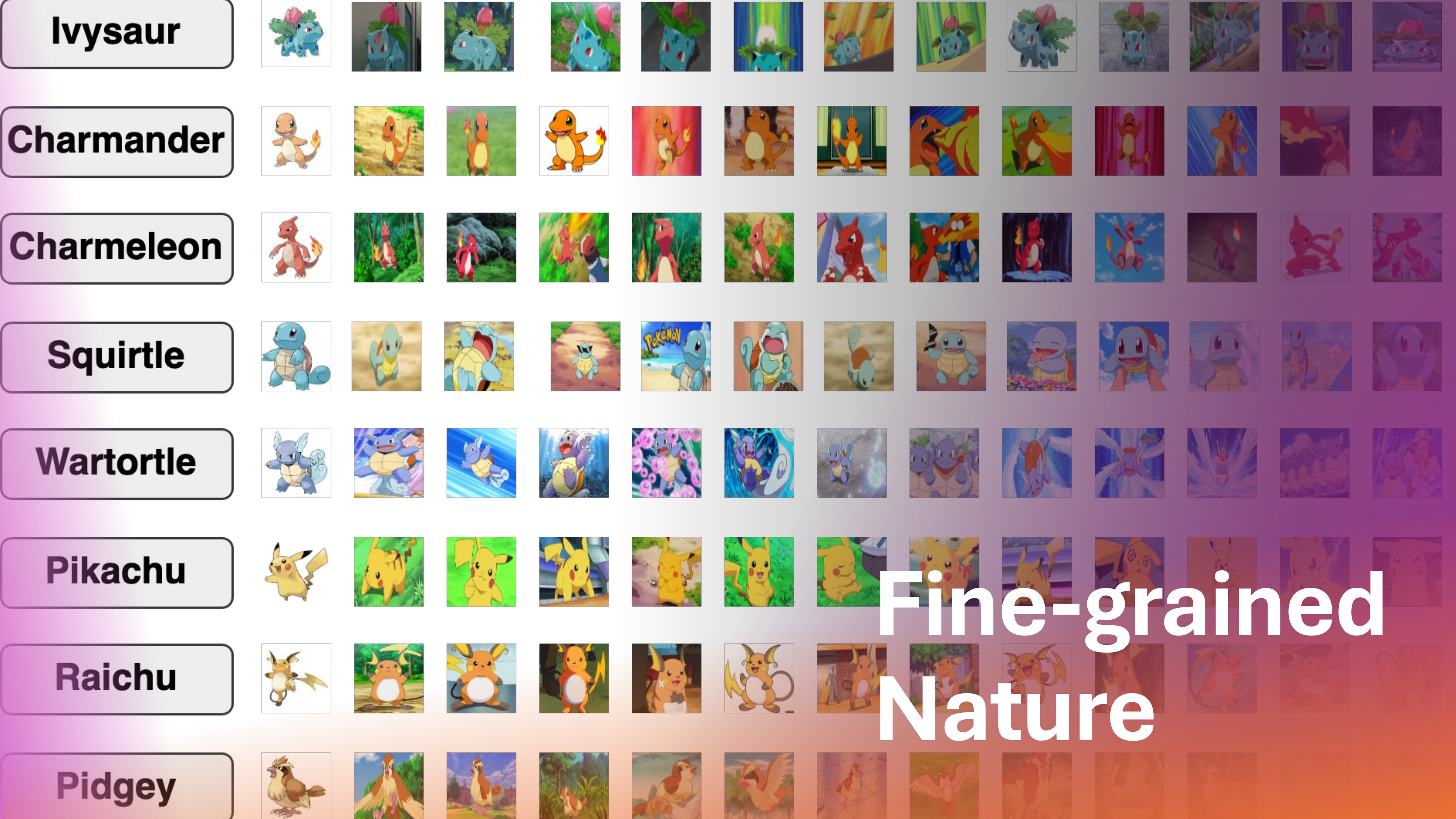
FineR not only shows better and finer predictions, but also demonstrates its semantic-awareness, therefore making better mistakes!



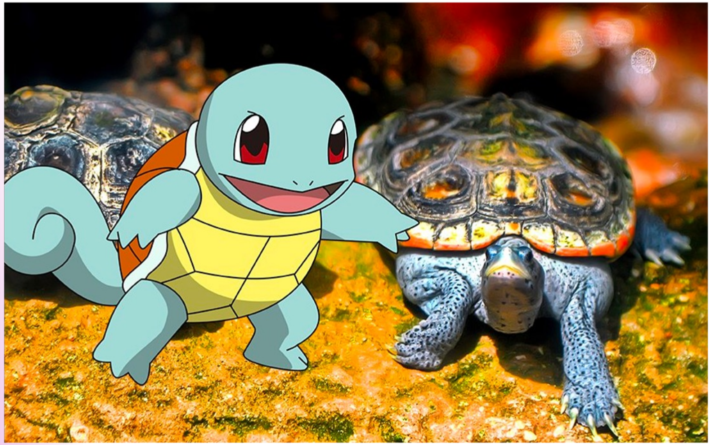
From layperson to expert - where do we stand?

A human study: FineR presents better performance than layperson on fine-grained Car and Pet recognition tasks.

What about virtual objects?


Ivysaur
Charmander
Charmeleon
Squirtle
Wartortle
Pikachu
Raichu
Pidgey
Fine-grained
Nature

New Challenge



Squirtle

vs.

Turtle

Rose-Breasted Grosbeak

vs.

Pidgey

Real-world Analogs

Comparison on the new Pokemon Dataset

Methods based on knowledge base (large corpus base like WordNet) retrieval are in effective for virtual concepts due to the real-world analogs (e.g., Turtle). However, FineR is still robust and approach upper-bound performance

Method	GT Hit Rate	Discovered Names
WordNet	0/10	Falkner, Turtler, Shiny Lyonia, Chicken Hawk, Gerfalcon, Pika, Garrison, Birdlime, Patrol, Tyto, Firedrake, Pokeweed, Archean Eon, Panduriform Leaf
BLIP-2	2/10	Sylveon Squirtle, Pikachu
CaSED	2/10	Interbreeding, Pikachu, Turtle, Plant, Pokemon, Bulbasaur, Bird
FineR (Ours)	**7/10**	Greenleaf Squirtle, Charmander, Charmeleon, Squirtle, Wartortle, Pikachu, Raichu,Pidgeotto, Pichu, Sadtail Pikachu, Flower Squirtle

Bulbasaur

(a) Discovered names and GT Hit Rate

Ivysaur

	Pokemon-10	
	cACC	sACC
Zero-shot (UB)	70.8	89.2
WordNet	34.6	33.1
BLIP-2	32.3	55.4
CaSED	39.2	55.7
FineR (Ours)	**70.8**	**81.6**

(b) Quantitative Results

Time to Wrap up

Conclusion

- We proposed a novel ***Vocabulary-free FGVR task*** with only few observations

- To achieve this challenging task, we designed ***FineR system*** that uses LLM to reason fine-grained semantic concepts from only few image observation

- FineR quantitatively and qualitatively demonstrates ***better performance*** on both real and virtual fine-grained benchmarks

UNIVERSITÀ DI TRENTO
FONDAZIONE BRUNO KESSLER
University of Nottingham UK | CHINA | MALAYSIA
UNIVERSITY OF LEEDS
UNIVERSITY OF ABERDEEN






Mingxuan LIU

Subhankar ROY

Wenjing LI

Zhun ZHONG

Nicu SEBE

Elisa RICCI

We thank you for your listening!

ICLR

Project page: