



UNIVERSITÀ  
DI TRENTO



University of  
Nottingham  
UK | CHINA | MALAYSIA



# Democratizing Fine-grained Visual Recognition with Large Language Models



**ICLR**

Mingxuan Liu<sup>1</sup>, Subhankar Roy<sup>4</sup>, Wenjing Li<sup>3,6\*</sup>

Zhun Zhong<sup>3,5\*</sup>, Nicu Sebe<sup>1</sup>, Elisa Ricci<sup>1,2</sup>

<sup>1</sup>University of Trento   <sup>2</sup>Fondazione Bruno Kessler   <sup>3</sup>Hefei University of Technology

<sup>4</sup>University of Aberdeen   <sup>5</sup>University of Nottingham   <sup>6</sup>University of Leeds

To recognize a common object, we now can ...

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

BLIP-2<sup>1</sup>: Pizza!



LENS<sup>3</sup>: Pizza!



GT: Pizza

LLaVA-1.5<sup>2</sup>: Pizza!




MiniGPT-4<sup>4</sup>: 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 ...

An illustration of a young boy with dark, curly hair and freckles, wearing a blue shirt. He is shown in profile, looking towards the right. He is holding a white, cracked bird egg in his hand. The background is a lush, colorful jungle filled with various birds, including parrots and smaller songbirds, perched on branches and flying. The style is a detailed, painterly illustration with vibrant colors and fine lines.

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<sup>1</sup>:  
Sparrow!



GT: Lincoln's Sparrow

LENS<sup>3</sup>:  
Vesper Sparrow!



LLaVA-1.5<sup>2</sup>:  
Sparrow!



MiniGPT-4<sup>4</sup>:  
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

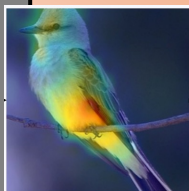
[3] William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. Towards language models that can see: Computer vision through the lens of natural language. *Arxiv*, 2023

[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<sup>1</sup>:  
Sparrow!

Not good : (  
Foundational VLMs  
Struggle with identifying  
Fine-grained concepts  
*Attention is needed!*



LLaVA-1.5<sup>2</sup>:  
Sparrow!



GT: Lincoln's Sparrow

MiniGPT-4<sup>4</sup>:  
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

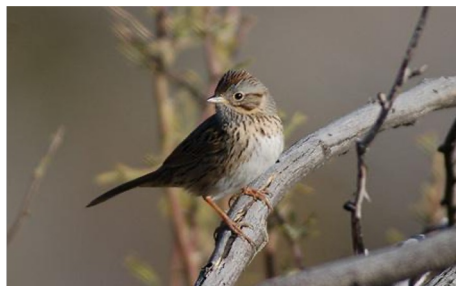
[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Arxiv*, 2023

[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

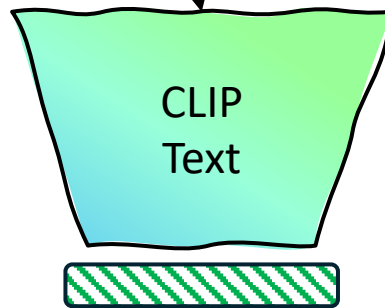
'Laysan Albatross'  
'Indigo Bunting'  
'Groove-billed Ani'  
'Crested Auklet'  
...  
'Rusty Blackbird'  
'Lincoln's Sparrow'

Pre-defined Fine-grained Vocabulary

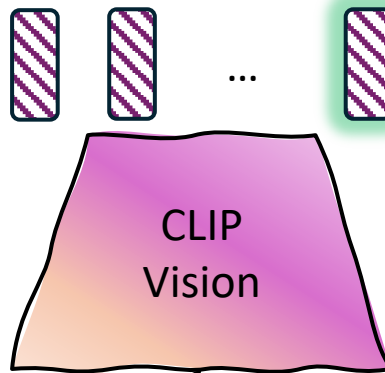


An image

GT: Lincoln's Sparrow



Gotcha: 'Lincoln's Sparrow'





'Laysan Albatross'  
'Indigo Bunting'  
'Groove-billed Ani'  
'Crested Auklet'  
...  
'Rusty Blackbird'  
'Lincoln's Sparrow'

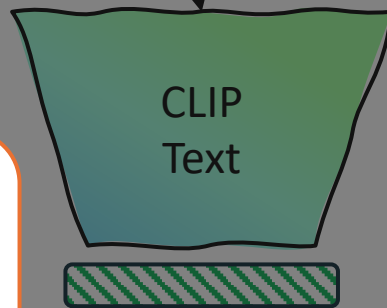
Pre-defined Fine-grained Vocabulary

How could I know these names that only ornithologists might know? 😂

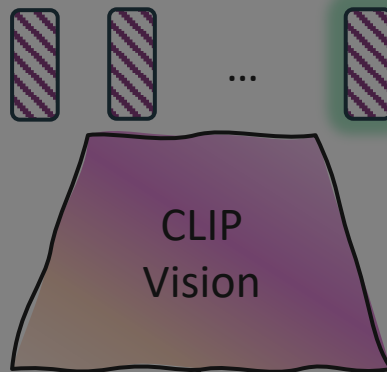


An image

GT: Lincoln's Sparrow



Output: '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?

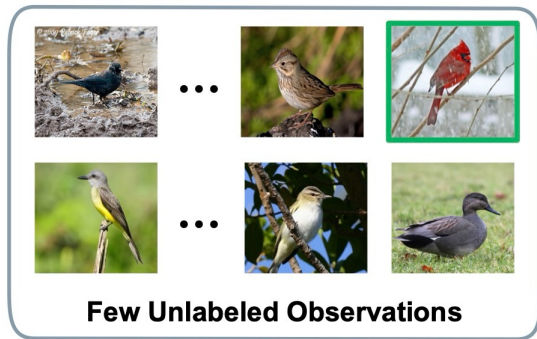
## Problem Formulation

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





Few Unlabeled Observations



### Reasoning For Each Sample

**BOT** : I see a *bird* in a photo. *How to* distinguish its specific species?

**BOT** : Well. Could you describe this photo and its *wing color, head pattern, ..., primary color* ?

**BOT** : Certainly. The *bird* is perched on a tree branch amidst the falling snow. Its *wings are grey*, and it boasts *a black and red pattern on its head*. Notably, its *dominant color is red*.

**BOT** : Perfect. Even though I can't see it, but based on your *description*, I think the bird you see would be a *Pyrrhuloxia, Cardinal, or Summer Tanager*.

Reasoning Concepts from Observations



Visual Question Answering Model



Large Language Model



Vision-Language Model



Test Images



Semantic Classification with Reasoned Concepts

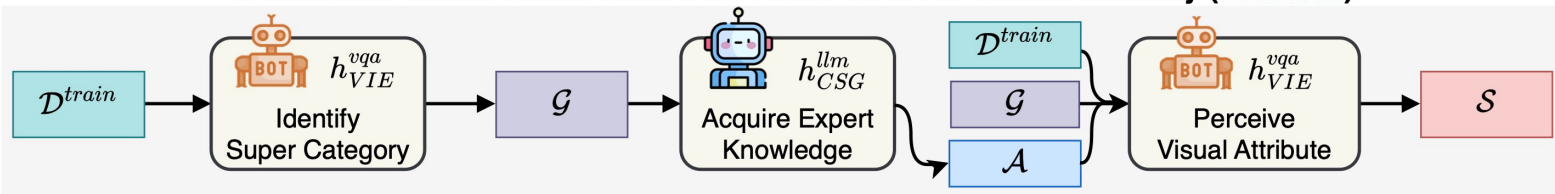


Inference

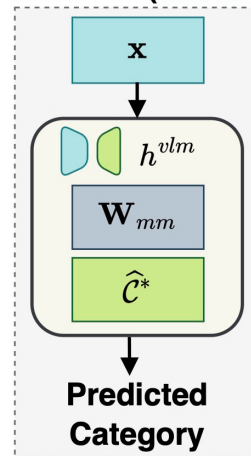
Just like what human would do ...

# Overview of FineR System

## I: Translate Useful Visual Information from Visual to Textual Modality (Sec 2.2.1)

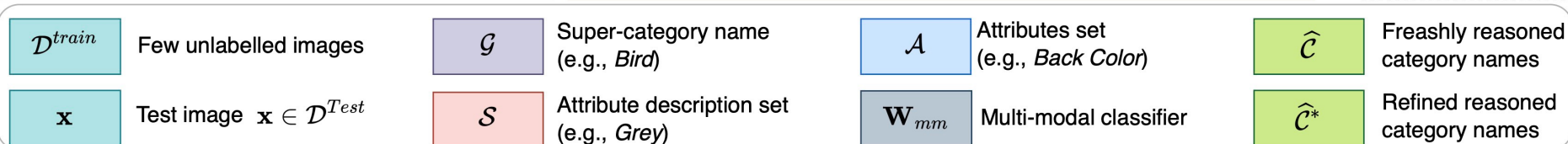


## Inference (Sec 2.3)



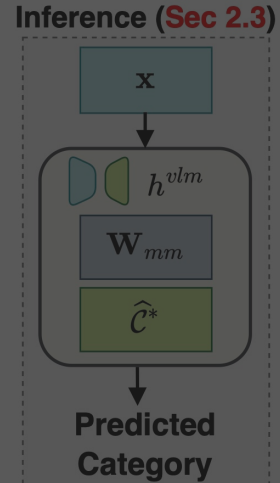
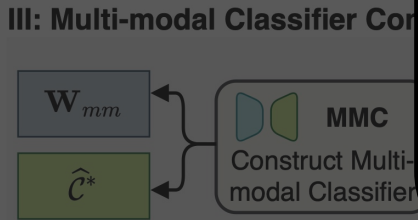
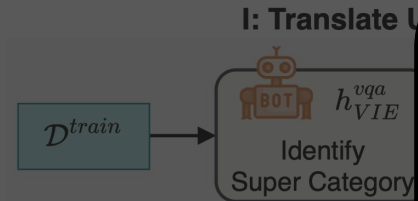
## III: Multi-modal Classifier Construction (Sec 2.2.3)

## II: Fine-grained Semantic Category Reasoning (Sec 2.2.2)



# Overview of FineR System

**No further training required!**  
**No pre-defined vocabulary required!**



$D^{train}$	Few unlabelled images	$\mathcal{G}$	Super-category name (e.g., <i>Bird</i> )	$\mathcal{A}$	Attributes set (e.g., <i>Back Color</i> )	$\hat{\mathcal{C}}$	Freshly reasoned category names
$x$	Test image $x \in \mathcal{D}^{Test}$	$\mathcal{S}$	Attribute description set (e.g., <i>Grey</i> )	$W_{mm}$	Multi-modal classifier	$\hat{C}^*$	Refined reasoned category names



# 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 2<sup>nd</sup>-best model by **+9.8%** in cACC and **+5.7%** in sACC on the five fine-grained datasets

Bird-200



**Ground-truth:** Orchard Oriole  
 WordNet: **Acridotheres Tristis**  
 BLIP-2: **Rufous Tanager**  
 CaSED: **Tanager**  
**FineR (Ours): Orchard Oriole**



**Ground-truth:** Dark-eyed Junco  
 WordNet: **Slate-colored Junco**  
 BLIP-2: **Junco**  
 CaSED: **Junco**  
**FineR (Ours): Dark-eyed Junco**

Car-196

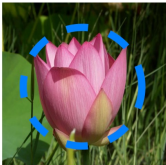


**Ground-truth:** Jeep Grand Cherokee SUV 2012  
 WordNet: **Cherokee**  
 BLIP-2: **Jeep Compass**  
 CaSED: **SUV**  
**FineR (Ours): Jeep Grand Cherokee SUV 2012**



**Ground-truth:** Bentley Continental GT Coupe 2012  
 WordNet: **Platinum Black**  
 BLIP-2: **Bentley Continental GT**  
 CaSED: **Bentley**  
**FineR (Ours): Bentley Continental GT Sedan 2010**

Flower-102



**Ground-truth:** Lotus  
 WordNet: **Lotus**  
 BLIP-2: **Lotus**  
 CaSED: **Lotus**  
**FineR (Ours): Pink Lotus** 🤔



**Ground-truth:** Blackberry Lily  
 WordNet: **Peruvian Lily**  
 BLIP-2: **Lilium Senegalensis**  
 CaSED: **Gloriosa**  
**FineR (Ours): Orange-spotted Lily** 🤔

### Prediction Indicator

Correct Prediction

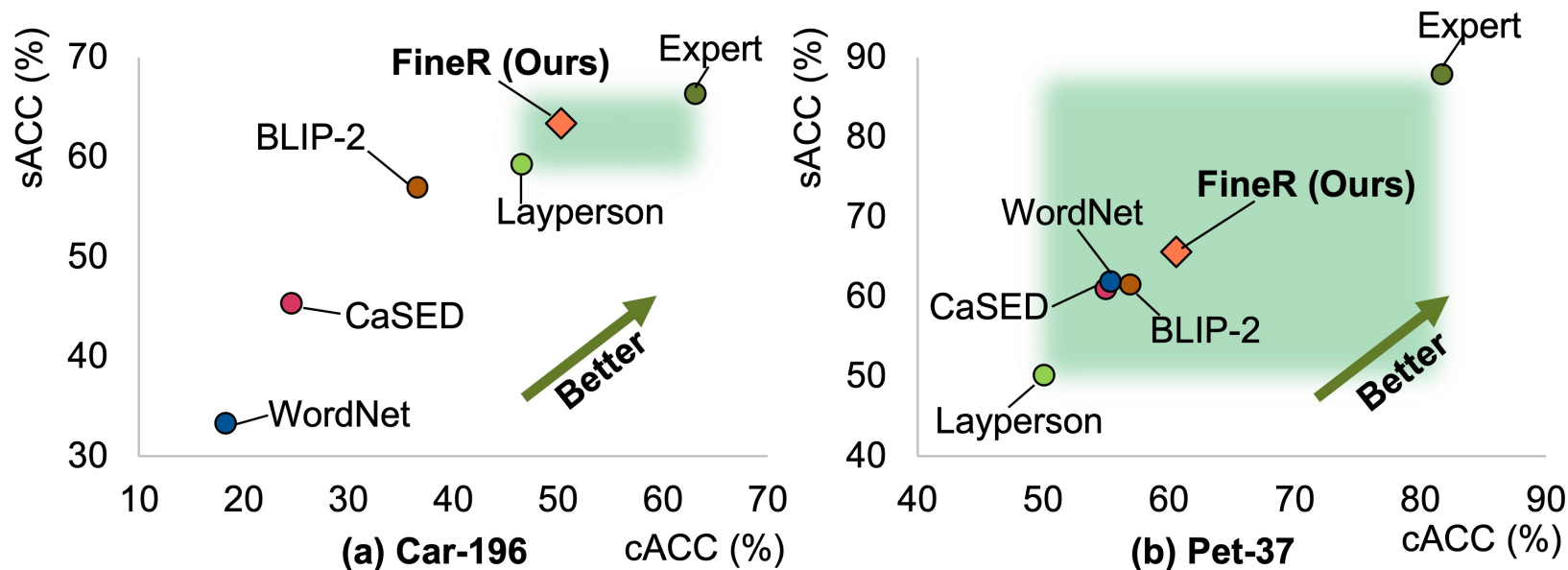
Partially Correct Prediction

Incorrect Prediction

Even more precise than ground-truth names

## 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?



Bulbasaur



Charmander



Squirtle



Pikachu



Pidgey



Ivysaur



Charmeleon



Wartortle

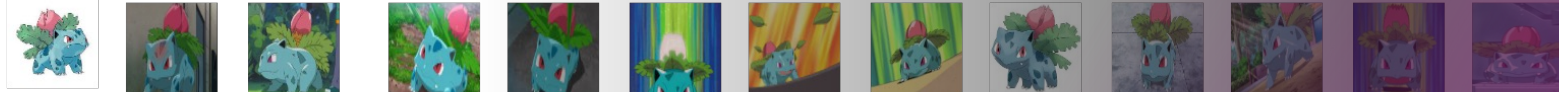


Raichu

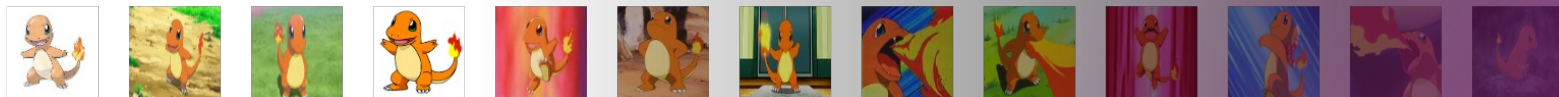


Pidgeotto

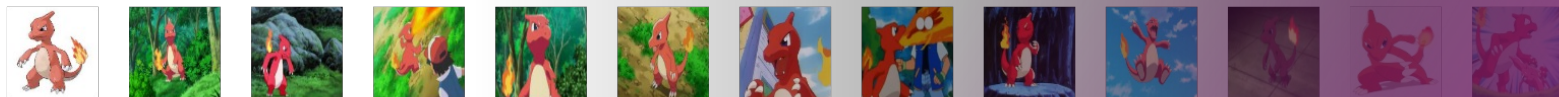
Ivysaur



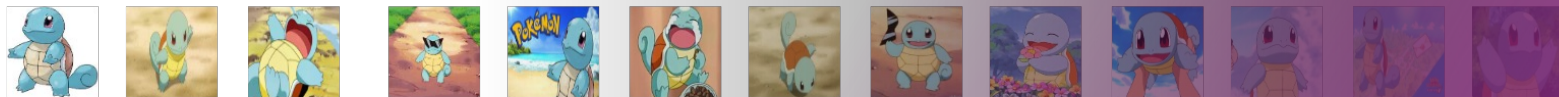
Charmander



Charmeleon



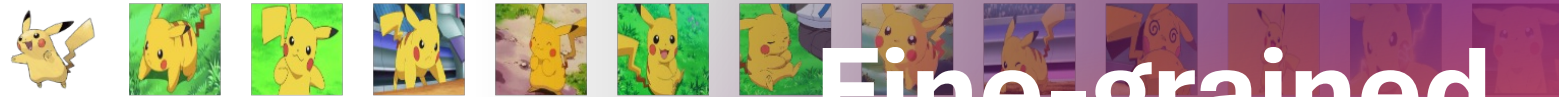
Squirtle



Wartortle



Pikachu



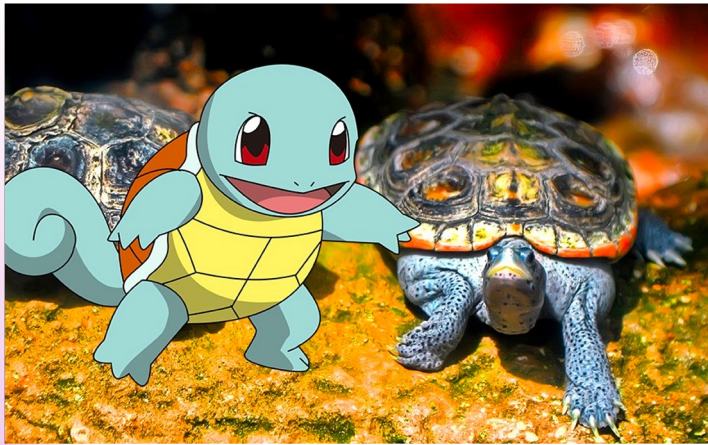
Raichu



Pidgey



Fine-grained  
Nature



Squirtle

vs.

**Turtle**

# New Challenge

**Rose-Breasted  
Grosbeak**

vs.

Pidgey



# Real-world Analog

# 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, Gerialcon, 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
<b>FineR (Ours)</b>	<b>7/10</b>	<u>Greenleaf Squirtle, Charmander, Charmeleon, Squirtle, Wartortle, Pikachu</u> Raichu, Pidgeotto, Pichu, Sadtail Pikachu, Flower Squirtle



(a) Discovered names and GT Hit Rate



	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
<b>FineR (Ours)</b>	<b>70.8</b>	<b>81.6</b>

(b) Quantitative Results

A man in a dark suit and tie is shown in profile, wearing a futuristic AR headset with a glowing purple lens. He is standing in a vibrant, stylized city street at night. The scene is filled with digital overlays of various Pokémon, each in a small, glowing frame with a label. The labels include 'Adorocctiiions', 'Pomor', 'Chracten', 'Anirsi', 'Doxdi..', and 'Foltrra'. In the background, there are tall buildings, neon signs, and other people walking. A sign on the right says 'APPO' and another on the left says 'SAFI'. The overall atmosphere is futuristic and digital.

# 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



University of  
Nottingham  
UK | CHINA | MALAYSIA



UNIVERSITY OF LEEDS



1495  
UNIVERSITY OF  
ABERDEEN



Mingxuan  
LIU



Subhankar  
ROY



Wenjing  
LI



Zhun  
ZHONG



Nicu  
SEBE



Elisa  
RICCI

We thank you for your listening!



**ICLR**

Project page:

