



Fine-R1: Make Multi-modal LLMs Excel in Fine-Grained Visual Recognition by Chain-of-Thought Reasoning



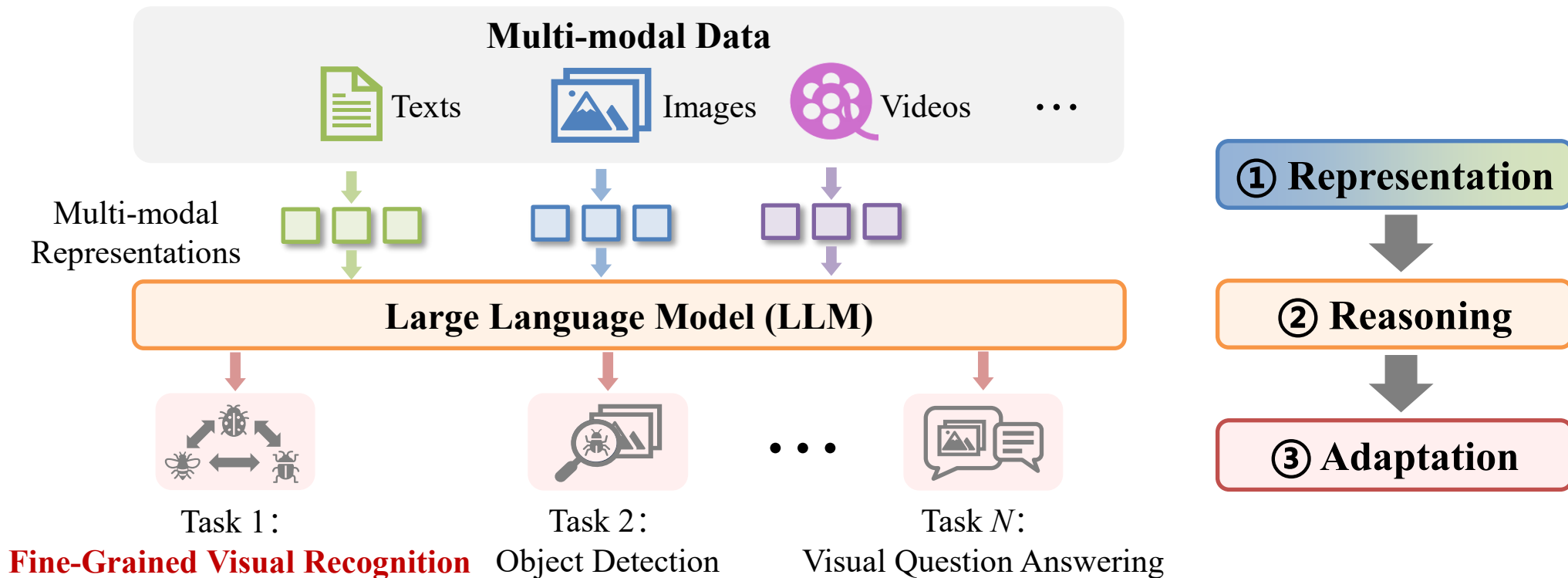
Hulingxiao He, Zijun Geng, and Yuxin Peng*

Wangxuan Institute of Computer Technology, Peking University

hehulingxiao@stu.pku.edu.cn, gengzijun2024@163.com, pengyuxin@pku.edu.cn

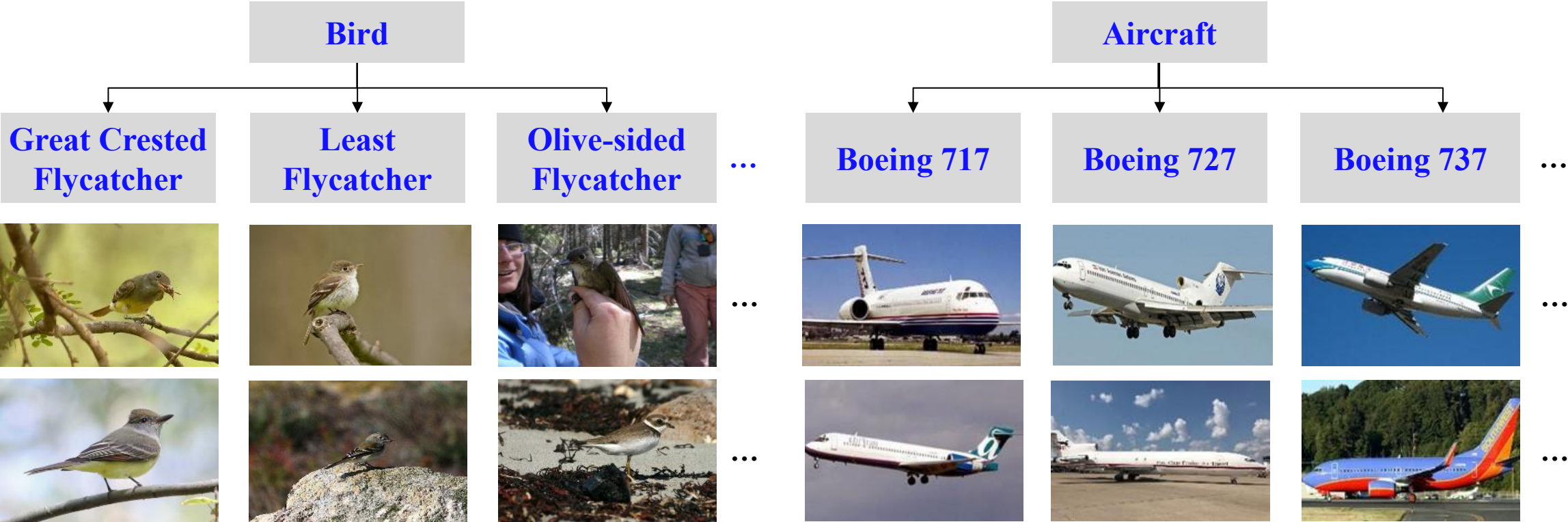
1. Multi-modal Large Language Models

- **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



2. Fine-Grained Visual Recognition

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



3. Poor FGVR Performance of MLLMs (1/2)

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



GPT-5.4
released by
OpenAI
in March 2026

**Correct in identifying
coarse-grained category**

**Wrong in identifying
fine-grained category**



Answer: It is a **bird**



Answer: Due to the limitations in image quality and details, the species **cannot be accurately confirmed.**



(**Correct answer:** The bird species in the image is Olive-sided Flycatcher.)



Question: What is the **animal?**

Question: What is the **bird species?**

3. Poor FGVR Performance of MLLMs (2/2)

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



Gemini 3.1
released by **Google**
in February 2026

Error in identifying the fine-grained category



Answer: The registration number D-AVVI is clearly visible on the rear fuselage. This is the prototype aircraft of an **Airbus A319neo**.

(**Correct answer:** The aircraft model in the image is the A320.)

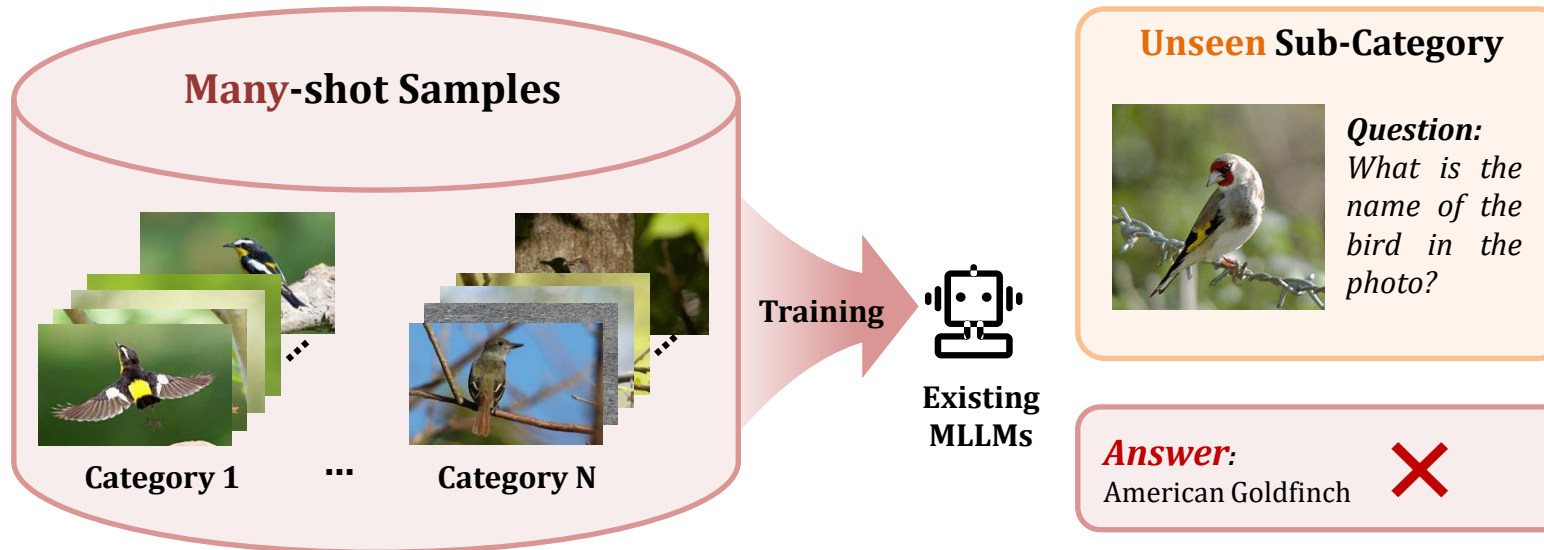


Question:

What is the aircraft model?

4. Motivation

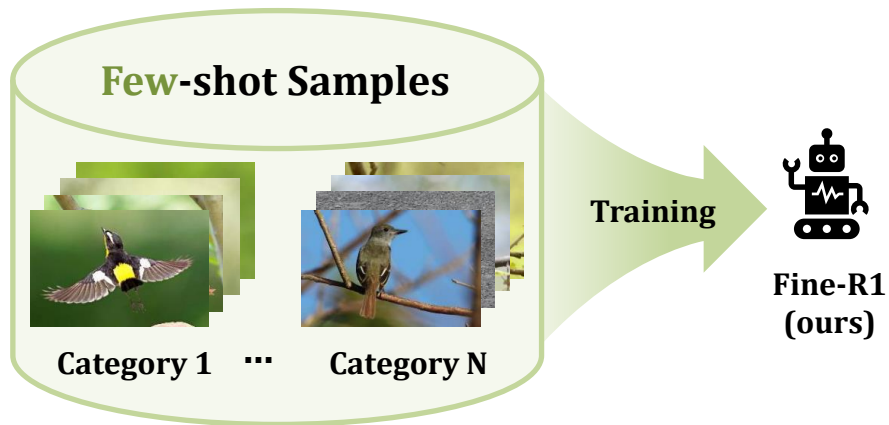
- Existing MLLMs for FGVR (e.g., Finedefics) have two main limitations:
 - High data requirements:** Annotating fine-grained subcategories is costly, resulting in limited samples per class and insufficient data for training large models
 - Limited coverage:** Models trained on a fixed set of subcategories can recognize seen classes but struggle to generalize to unseen ones



He et al., "Analyzing and Boosting the Power of Fine-Grained Visual Recognition for Multi-modal Large Language Models", *ICLR*, 2025.

5. Key idea

- ① **Long chain-of-thought reasoning:** Helps models better use latent fine-grained knowledge via step-by-step reasoning, **easing category name decoding**
- ② **Reinforcement Fine-tuning:** Encourages exploration of transferable strategies compared to SFT, improving **out-of-distribution generalization**



Thinking Process:

(Visual Analysis) The image shows a small bird perched on a wire, with distinctive features that can be analyzed to identify its species:

1. Size: The bird appears to be medium-sized, consistent with the description of **European Goldfinches** ...

(Candidate Subcategories) Based on the visual features observed, the following candidate subcategories within the super-category of "finches" are considered:

1. **European Goldfinch:** Known for its bright red forehead patch, pale gray body, and yellow wing patches ...

4. **American Goldfinch:** Similar in size to the European Goldfinch but lacks the bright red forehead patch and has a more subdued coloration.

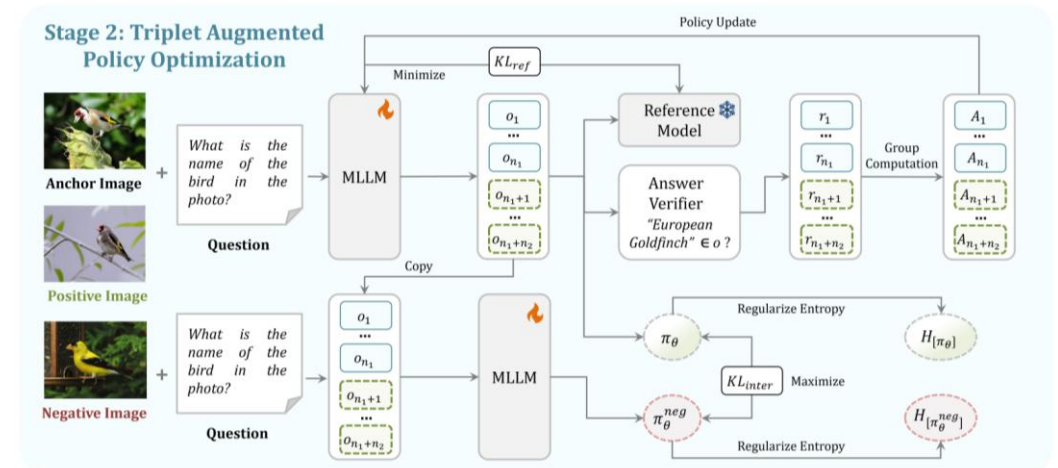
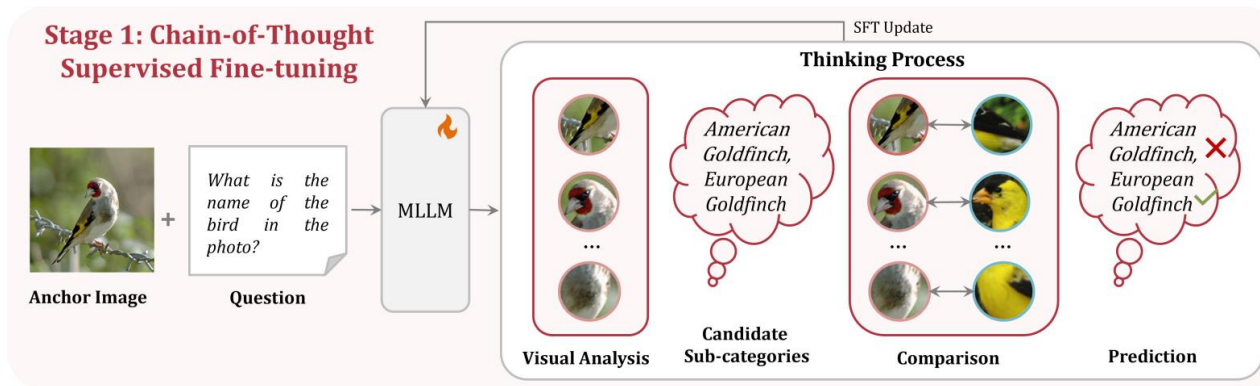
(Comparison) ... **European Goldfinch vs. American Goldfinch:** The American Goldfinch also lacks the bright red forehead patch and has a more subdued coloration, making it less likely to be the correct identification. The bright red forehead patch, pale gray body, and yellow wing patches are highly characteristic of the European Goldfinch ...

(Prediction) Based on the detailed analysis of the visual features, the bird in the image is most likely an **European Goldfinch**.

Answer: ✓
European Goldfinch

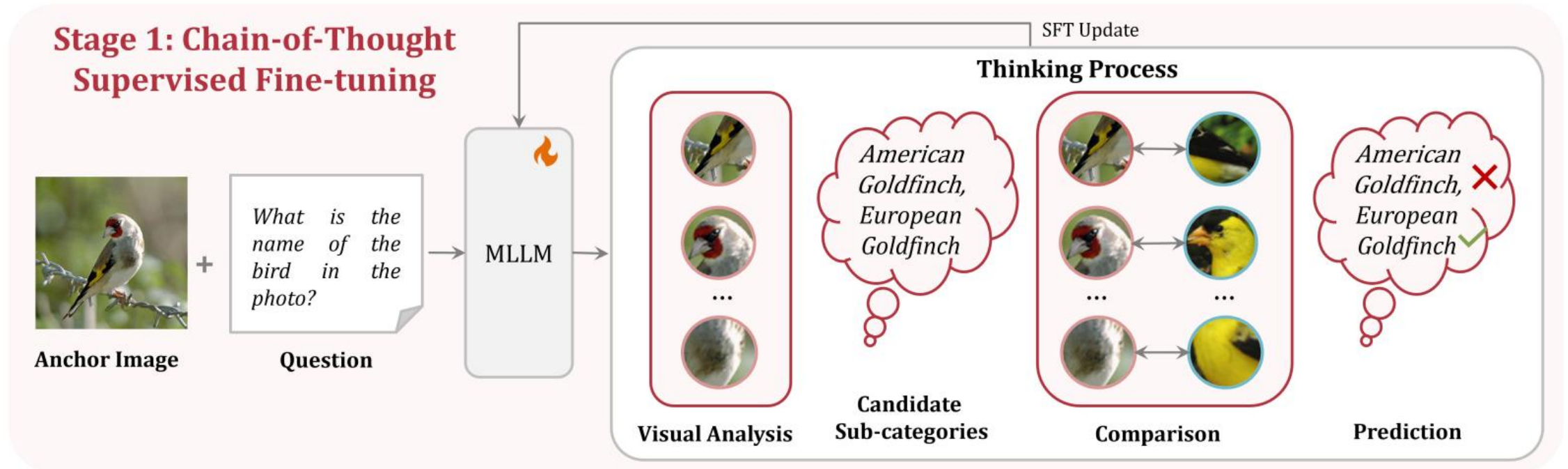
6. Method (1/3)

- **Stage 1 (Chain-of-Thought Supervised Fine-tuning):** learning structured reasoning process for FGVR by imitation
- **Stage 2 (Triplet Augmented Policy Optimization):** improving inter-class discrimination and intra-class robustness by exploration



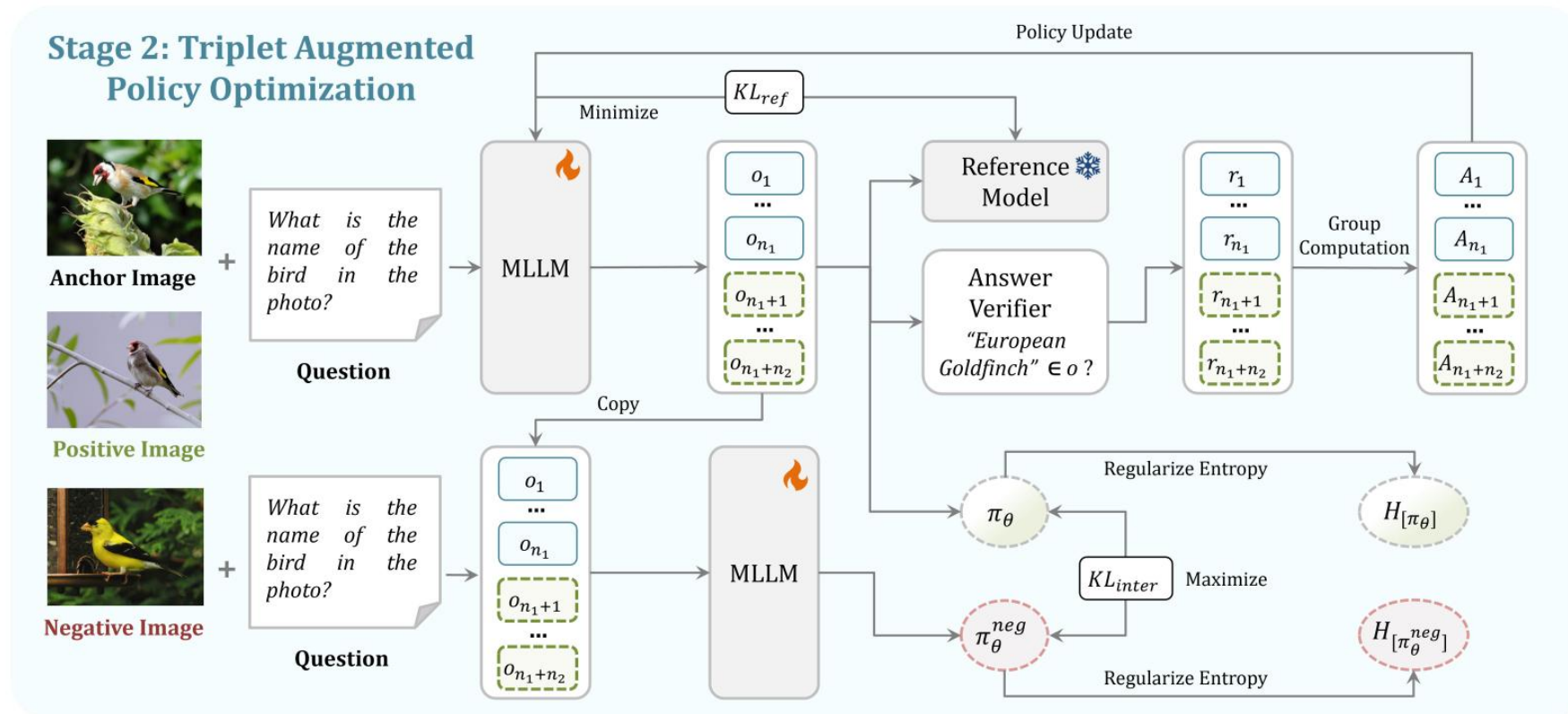
6. Method (2/3)

- ① **CoT SFT**: We mimic human reasoning with **a structured chain** (visual analysis → candidate classes → comparison → prediction), **enabling rapid reasoning** via few-shot tuning



6. Method (3/3)


- ② **TAPO**: During RL fine-tuning, use **positive** (same class) and **negative** (different class) samples. Adding positive rollouts improves **intra-class robustness**, while maximizing divergence from negatives enhances **inter-class discrimination**



7. Experiments (1/5)

- Evaluate the FGVR ability of MLLMs under two settings:
 - ① **Closed-world**: Candidates are predefined; the model selects from them. (**multiple choice**)
 - ② **Open-world**: No predefined candidates; the model directly outputs the category. (**QA**)

Closed-world




Q: What is the species of the bird?

A. Acadian Flycatcher B. Great Crested Flycatcher
C. Least Flycatcher D. Olive-sided Flycatcher

Multiple choice
(Candidates are predefined)

D. Olive-sided Flycatcher

Open-world



Q: What is the species of the bird?

QA
(No predefined candidates)

The bird is
Olive-sided Flycatcher.

7. Experiments (2/5)

- ① **Closed-set (multiple choice):** With **only 4 training images per class**, Fine-R1 outperforms discriminative models such as **OpenAI’s CLIP** and **DeepMind’s SigLIP**, demonstrating **the potential of generative MLLMs for discriminative tasks**

Models	Seen Categories							Unseen Categories							Avg.
	Air.	Bird	Car	Dog	Flower	Pet	Avg.	Air.	Bird	Car	Dog	Flower	Pet	Avg.	
CLIP Models															
CLIP-L	47.95	73.96	80.50	77.12	87.59	95.45	77.10	45.68	62.35	79.81	73.58	57.24	89.22	67.98	72.54
EVA-G	40.96	81.37	90.06	76.02	81.70	94.21	77.39	45.83	63.39	87.06	76.66	54.83	89.82	69.60	73.49
SigLIP-L	67.08	85.10	96.03	86.18	97.59	98.02	88.33	69.35	74.17	92.97	84.44	69.07	93.24	80.54	84.44
SigLIP2-L	43.36	71.26	93.43	78.51	89.05	92.14	77.96	40.87	58.03	<u>90.00</u>	77.15	61.39	93.03	70.08	74.02
General MLLMs															
Idefics2-8B	49.90	52.37	90.85	58.24	82.10	81.85	69.22	48.53	40.20	84.67	44.79	60.54	83.86	60.43	64.83
Idefics3-LLaMA3-8B	43.31	34.51	75.18	47.50	65.12	72.98	56.43	48.38	37.47	70.59	43.80	55.30	68.59	54.02	55.23
LLaVA-v1.6-mistral-7B	49.60	55.96	71.02	45.34	62.86	62.59	57.90	47.71	50.32	67.62	37.20	46.16	65.71	52.45	55.17
LLaVA-Onevision-7B	32.07	54.81	71.38	70.72	73.83	76.52	63.22	30.43	52.92	67.21	53.78	48.94	66.31	53.27	58.24
InternVL2.5-2B	36.71	65.23	63.40	53.70	65.39	75.46	59.98	40.57	62.74	63.53	42.62	46.35	60.21	52.67	56.33
InternVL2.5-4B	38.31	29.54	51.44	34.65	61.44	51.01	44.40	39.29	31.29	44.95	31.99	35.83	54.25	39.60	42.00
InternVL2.5-8B	46.70	53.14	62.17	54.26	68.25	76.65	60.20	45.38	47.12	61.15	48.82	47.95	62.83	52.21	56.20
Qwen2-VL-2B	66.18	60.15	94.28	64.67	91.93	85.39	77.10	65.51	44.92	87.89	58.60	50.21	79.84	64.50	70.80
Qwen2-VL-7B	78.27	67.41	94.60	71.70	93.84	91.04	82.81	<u>79.86</u>	56.25	89.63	66.16	67.47	78.30	72.95	77.88
Qwen2.5-VL-3B	64.24	65.40	86.70	70.51	94.24	83.46	77.43	68.29	58.98	80.65	67.10	68.74	87.88	71.94	74.68
Qwen2.5-VL-7B	74.28	70.54	90.75	80.19	96.20	91.91	83.98	71.60	66.29	84.02	77.54	65.63	<u>93.44</u>	76.42	80.20
Reasoning MLLMs															
DeepPerception-7B	83.52	74.16	<u>94.89</u>	80.40	97.05	91.91	86.99	86.48	61.19	89.72	77.85	72.80	87.41	79.24	83.12
Fine-R1-3B (ours)	76.87	<u>86.79</u>	92.14	<u>87.85</u>	96.25	93.89	<u>88.97</u>	75.73	<u>79.10</u>	87.40	80.93	<u>73.93</u>	91.36	<u>81.41</u>	<u>85.19</u>
Fine-R1-7B (ours)	<u>82.32</u>	90.50	94.03	90.11	<u>97.22</u>	<u>96.05</u>	91.71	77.91	87.54	87.99	89.71	74.12	96.92	85.70	88.71

Ours

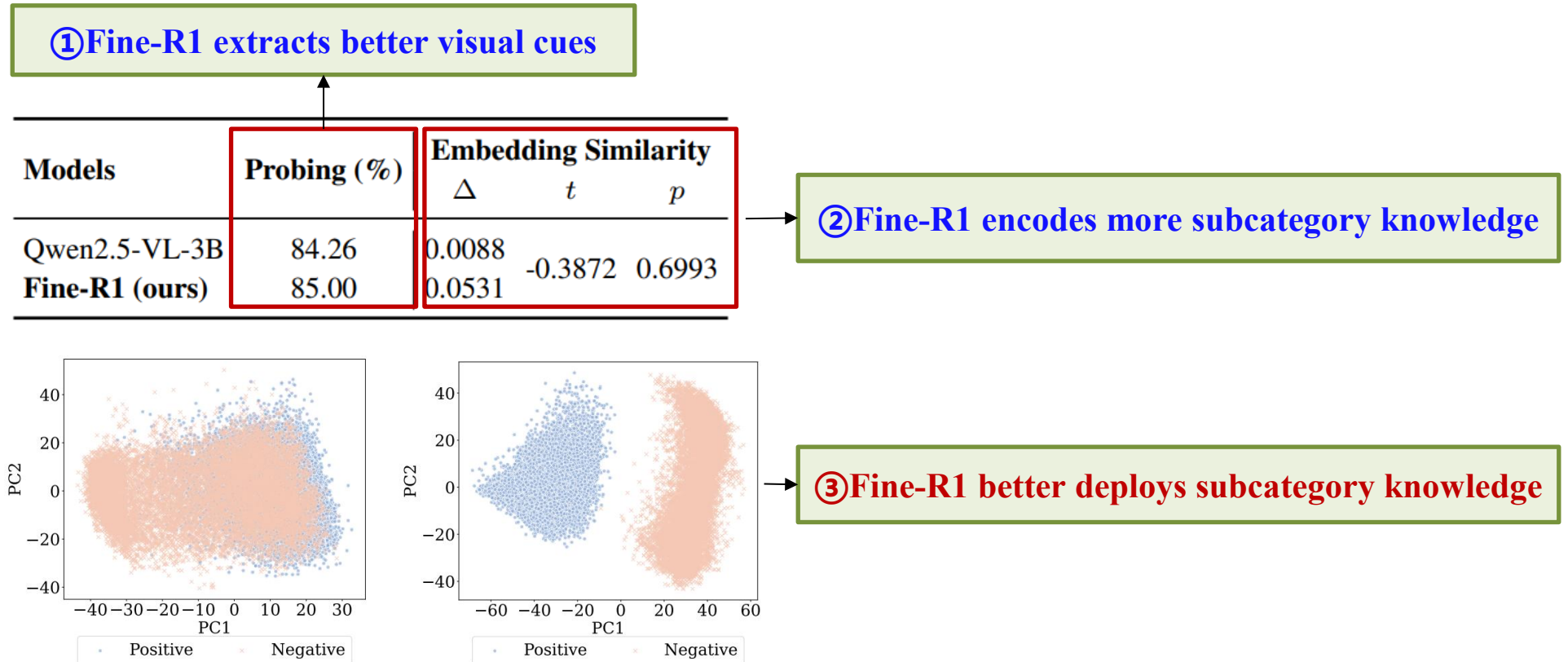
7. Experiments (3/5)

② **Open-world (QA)** : Similarly, **with only 4 training images per class**, Fine-R1 outperforms **mainstream general MLLMs and reasoning MLLMs** on both seen and unseen subcategories

Models	Seen Categories							Unseen Categories							Avg.
	Air.	Bird	Car	Dog	Flower	Pet	Avg.	Air.	Bird	Car	Dog	Flower	Pet	Avg.	
General MLLMs															
Idefics2-8B	3.64	19.68	19.54	10.03	14.94	2.50	11.72	3.69	15.35	20.81	10.19	5.84	2.47	9.73	10.72
Idefics3-LLaMA3-8B	9.66	27.72	22.96	35.39	40.92	20.17	26.14	7.09	24.93	22.08	27.67	21.99	24.84	21.43	23.79
LLaVA-v1.6-mistral-7B	2.73	16.02	21.75	19.35	10.33	12.47	13.78	2.51	17.40	23.24	18.16	8.20	10.62	13.36	13.57
LLaVA-Onevision-7B	9.90	31.74	21.35	30.13	45.24	17.87	26.04	7.47	29.81	19.56	25.44	16.53	19.35	19.69	22.87
InternVL2.5-2B	7.26	21.32	27.29	25.87	23.08	26.56	21.90	5.22	20.12	24.73	24.84	12.59	24.55	18.68	20.29
InternVL2.5-4B	14.71	25.41	32.19	33.13	23.84	28.19	26.25	12.64	23.91	29.11	30.80	13.10	26.32	22.65	24.45
InternVL2.5-8B	23.76	28.44	30.08	27.11	21.73	29.03	26.69	20.34	24.38	27.55	24.73	13.56	26.23	22.80	24.75
Qwen2-VL-2B	47.49	48.72	52.95	51.22	66.56	19.88	47.80	48.88	39.32	49.33	45.16	33.87	22.43	39.83	43.82
Qwen2-VL-7B	56.47	56.46	55.31	67.03	75.02	36.97	57.88	<u>52.75</u>	41.17	52.46	61.01	32.57	30.74	45.12	51.50
Qwen2.5-VL-3B	56.98	66.77	52.49	65.12	<u>68.96</u>	26.27	56.10	52.50	48.09	51.75	59.02	34.19	28.78	45.72	50.91
Qwen2.5-VL-7B	<u>58.86</u>	65.97	56.94	59.02	62.61	35.59	56.50	48.62	45.26	55.39	54.59	32.74	36.98	45.60	51.05
Reasoning MLLMs															
DeepPerception-7B	44.24	47.63	54.14	49.16	47.30	40.90	47.23	40.03	37.10	52.27	49.05	28.38	35.57	40.40	43.82
Ours Fine-R1-3B (ours)	54.36	<u>78.90</u>	<u>82.46</u>	<u>78.21</u>	64.55	<u>81.60</u>	<u>73.35</u>	46.43	<u>58.00</u>	<u>74.60</u>	<u>70.08</u>	<u>39.54</u>	<u>79.11</u>	<u>61.29</u>	<u>67.32</u>
Ours Fine-R1-7B (ours)	73.53	86.12	90.73	80.71	81.46	83.14	82.62	65.21	60.69	82.19	70.97	40.74	82.04	66.97	74.80



7. Experiments (4/5)

- Experimental analysis shows that Fine-R1 improves accuracy mainly by **enhancing the model's ability to deploy subcategory knowledge**, rather than by extracting better visual cues or encoding more subcategory knowledge



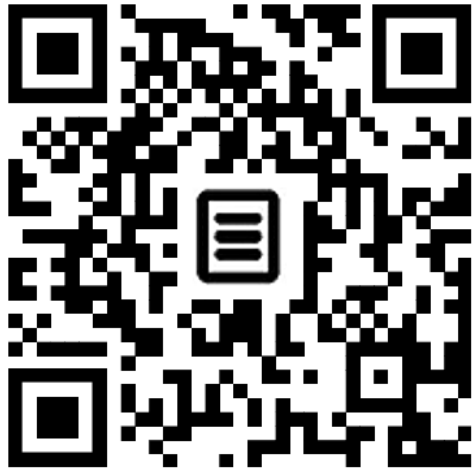
7. Experiments (5/5)

- Compared to **Alibaba's Qwen2.5-VL-3B**, Fine-R1 breaks the reasoning process into “**visual analysis** → **candidate classes** → **comparison** → **prediction**”, leveraging knowledge for step-by-step fine-grained subcategory recognition

 <p>What kind of car is this? Options: [Mercedes-Benz SL-Class Coupe 2009, Aston Martin Virage Coupe 2012, Chevrolet Corvette ZR1 2012, Aston Martin V8 Vantage Coupe 2012]</p>	<p>Qwen2.5-VL-3B</p> <p>Answer: Aston Martin ✗ Virage Coupe 2012</p>	 <p>Can you identify the aircraft in this image? Options: [BAE 146-300, BAE 146-200, Fokker 100, Fokker 70]</p>	<p>Qwen2.5-VL-3B</p> <p>Answer: Fokker 100 ✗</p>
<p style="text-align: center;">Fine-R1-3B</p> <p>Thinking Process: (Visual Analysis) The image shows a sleek, low-slung coupe with several distinctive features that can be analyzed to identify its name and type: 1. Roofline: The car has a long, flowing roofline that tapers towards the rear, creating an aerodynamic profile ... (Candidate Subcategories) Based on the visual features observed, the following candidate subcategories (all within the super-category of "sports coupes") are considered: 1. Mercedes-Benz SL-Class Coupe 2009: Known for its sleek design, long roofline, and fastback silhouette. 2. Aston Martin Virage Coupe 2012: Also features a fastback design but typically has a more rounded and less aggressive profile compared to the Mercedes-Benz ... (Comparison) Mercedes-Benz SL-Class Coupe 2009 vs. Aston Martin Virage Coupe 2012: Both have fastback designs, but the Mercedes-Benz typically has a more streamlined and elegant appearance, while the Aston Martin Virage tends to have a more rounded and less aggressive look ... (Prediction) Based on the detailed analysis of the visual features, including the roofline, side profile, and overall design, the car in the image is most likely a Mercedes-Benz SL-Class Coupe 2009. Answer: Mercedes-Benz SL-Class Coupe 2009 ✓</p>		<p style="text-align: center;">Fine-R1-3B</p> <p>Thinking Process: (Visual Analysis) The image shows an aircraft with the following visual features 1. Fuselage Length: The fuselage appears to be relatively short compared to larger regional jets, indicating it is designed for shorter routes ... (Candidate Subcategories) Based on the visual features observed, the candidate subcategories within the super-category "regional jet" are: 1. BAE 146-300 2. BAE 146-200 3. Fokker 100 4. Fokker 70 These options are chosen because they represent different regional jet models known for their T-tail configurations and rear-mounted engines. (Comparison) ... BAE 146-300 vs. Fokker 100/Fokker 70: The BAE 146 has a more compact and streamlined design compared to the Fokker 100 and Fokker 70, which have a more boxy appearance. The T-tail configuration and rear-mounted engines align better with the BAE 146's design ... (Prediction) Based on the analysis and comparison, the aircraft in the image most closely matches the BAE 146-300. The T-tail configuration, rear-mounted engines, and overall size are consistent with this model. Additionally, the presence of the "buzz" livery suggests it is operated by a regional airline, which aligns with the BAE 146's role as a regional jet. Answer: BAE 146-300 ✓</p>	

Thank you for listening!

- Code and model are available now, welcome to follow our work!
 - **Paper**: <https://arxiv.org/abs/2602.07605>
 - **Code**: https://github.com/PKU-ICST-MIPL/FineR1_ICLR2026
 - **Model**: <https://huggingface.co/collections/StevenHH2000/fine-r1>
 - **Lab**: <https://www.wict.pku.edu.cn/mipl>



【Paper】



【Code】



【Model】



【Lab】

More related work

- Finedefics (ICLR 2025): revisiting three quintessential capabilities of MLLMs for FGVR and position of the root cause as a misalignment problem
 - **Paper**: <https://arxiv.org/abs/2501.15140>
 - **Code**: https://github.com/PKU-ICST-MIPL/Finedefics_ICLR2025
- TARA (CVPR 2026): using fine-grained category tree to boost hierarchical visual recognition capability of MLLMs
 - **Paper**: <https://arxiv.org/abs/2603.00431>
 - **Code**: https://github.com/PKU-ICST-MIPL/TARA_CVPR2026