ICLR

Reflexive Guidance: Improving OoDD in Vision-Language Models via Self-Guided Image-Adaptive Concept Generation

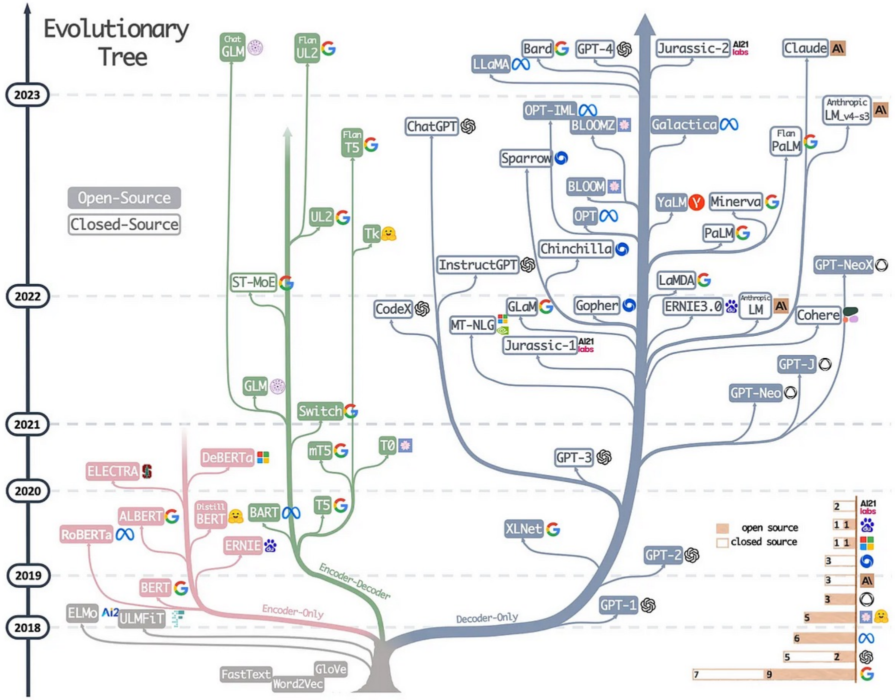
Jihyo Kim*, Seulbi Lee*, Sangheum Hwang (* Equal Contribution)
Seoul National University of Science & Technology

ICLR 2025

The Evolution of Foundation Models

DAINTLAB

- Foundation models, particularly vision-language models, have demonstrated their capabilities across diverse domains, from general tasks to specialized fields.



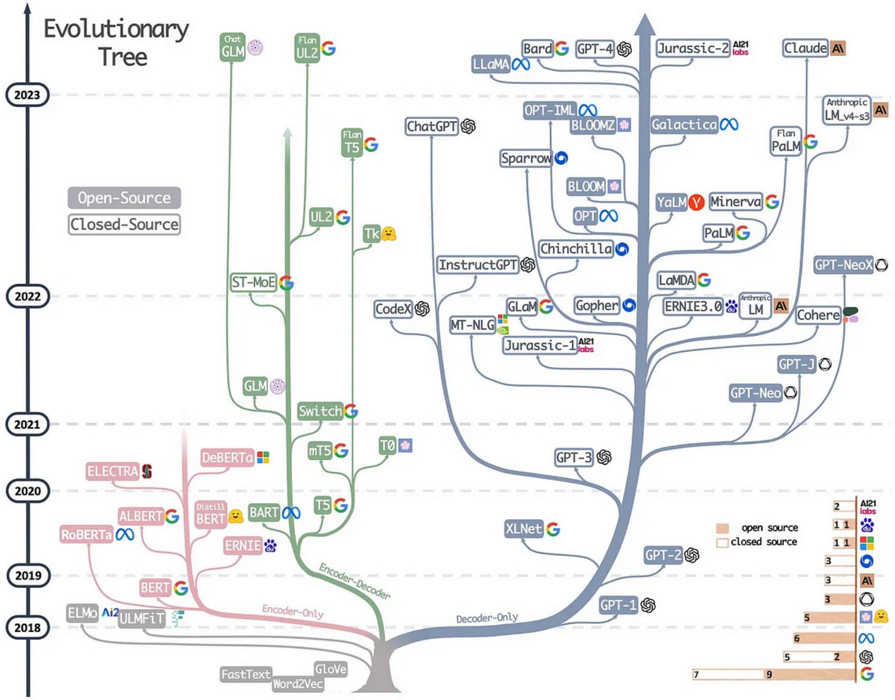
The evolutionary tree of foundation models [1]

[1] Berry D. (2023). Demystifying AI Foundation Models: A Comprehensive Guide to Large Language Models, ChatGPT, and Beyond. https://medium.com/@musicalchemist/demystifying-ai-foundation-models-a-comprehensive-guide-to-large-language-models-chatgpt-and-d69644299699

The Evolution of Foundation Models

DAINTLAB

- Foundation models, particularly vision-language models, have demonstrated their capabilities across diverse domains, from general tasks to specialized fields.



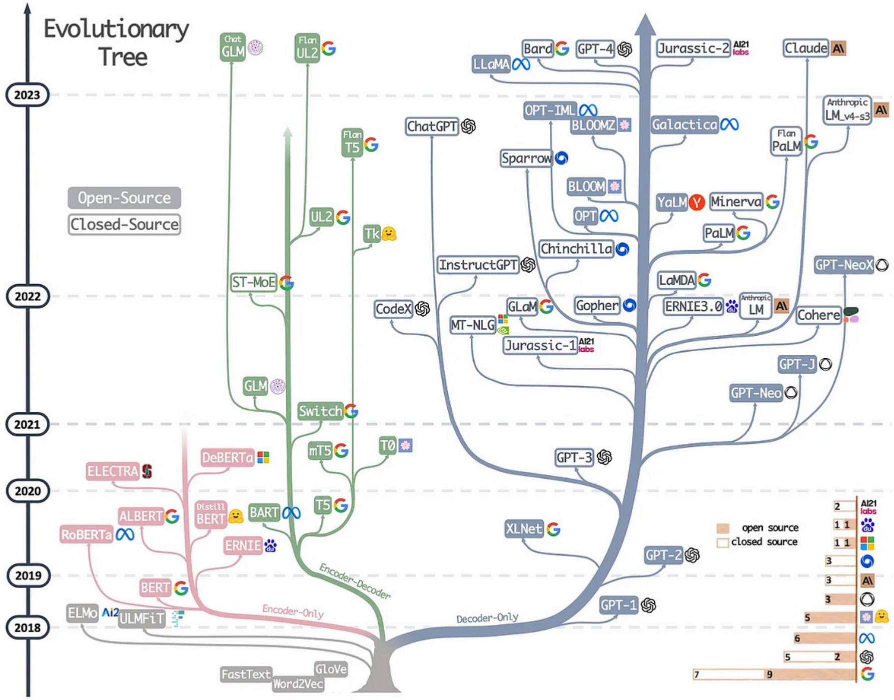
The evolutionary tree of foundation models [1]

- **Trustworthiness and reliability of large vision-language models (LVLMs) have not been adequately investigated** despite the widespread adoption of them.

[1] Berry D. (2023). Demystifying AI Foundation Models: A Comprehensive Guide to Large Language Models, ChatGPT, and Beyond. https://medium.com/@musicalchemist/demystifying-ai-foundation-models-a-comprehensive-guide-to-large-language-models-chatgpt-and-d69644299699

The Evolution of Foundation Models

- Foundation models, particularly vision-language models, have demonstrated their capabilities across diverse domains, from general tasks to specialized fields.
- **Trustworthiness and reliability of large vision-language models (LVLMs) have not been adequately investigated** despite the widespread adoption of them.



The evolutionary tree of foundation models [1]

[1] Berry D. (2023). Demystifying AI Foundation Models: A Comprehensive Guide to Large Language Models, ChatGPT, and Beyond. https://medium.com/@musicalchemist/demystifying-ai-foundation-models-a-comprehensive-guide-to-large-language-models-chatgpt-and-d69644299699

Contributions

- To bridge the gap, we

  1. **Evaluate and compare the OoDD capabilities of LVLMs**
     - Develop a framework for evaluating the OoDD capabilities of LVLMs
  2. **Propose a two-stage self-guided prompting approach called Reflexive Guidance (ReGuide)** to enhance the OoD detectability of LVLMs

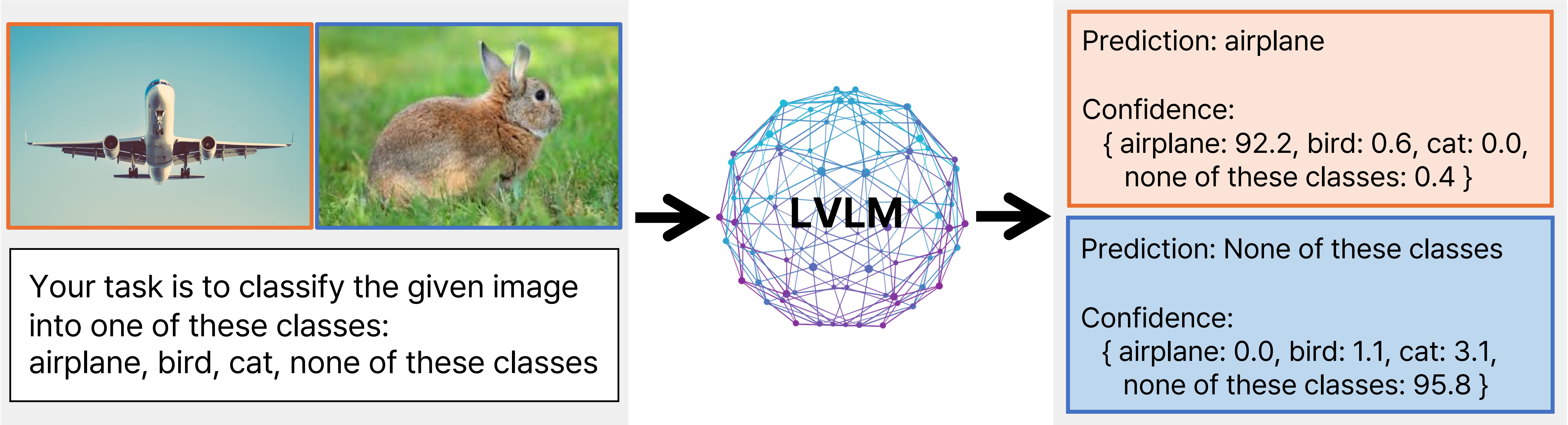
- From the results of our study, we can draw the following insights:

Despite the strong visual interpretation capabilities of LVLMs, which enable them to predict fine-grained classes of objects effectively, these models **tend to avoid generating responses that fall outside the given prompt categories**.

OoDD in LVLMs

Problem Definition

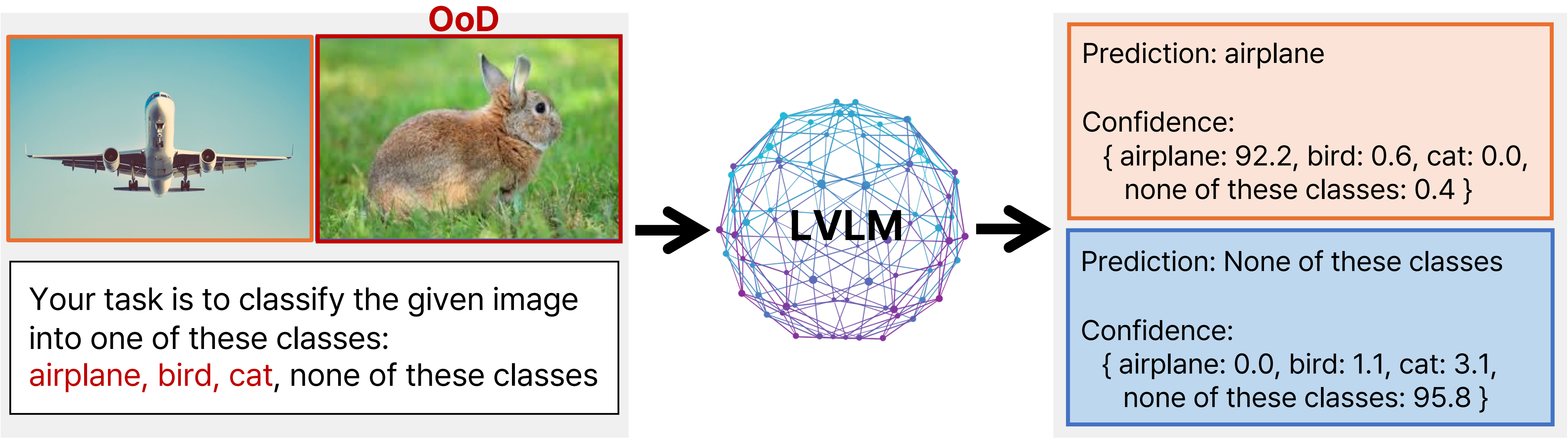
- Given the vast amount and broad domain coverage of data used to train LVLMs, this conventional definition faces challenges in its direct application to LVLMs.
  - To address this, we extend the zero-shot OoDD framework of CLIP to generative LVLMs: **the scenario where an in-distribution (ID) class words set does not contain the ground-truth label of an input image**.



OoDD in LVLMs

Problem Definition

- Given the vast amount and broad domain coverage of data used to train LVLMs, this conventional definition faces challenges in its direct application to LVLMs.
  - To address this, we extend the zero-shot OoDD framework of CLIP to generative LVLMs: **the scenario where an in-distribution (ID) class words set does not contain the ground-truth label of an input image**.



OoDD in LVLMs
Prompt Design

- Our prompt consists of four components: **a task description, an explanation of the rejection class, guidelines, and examples for the response format.**

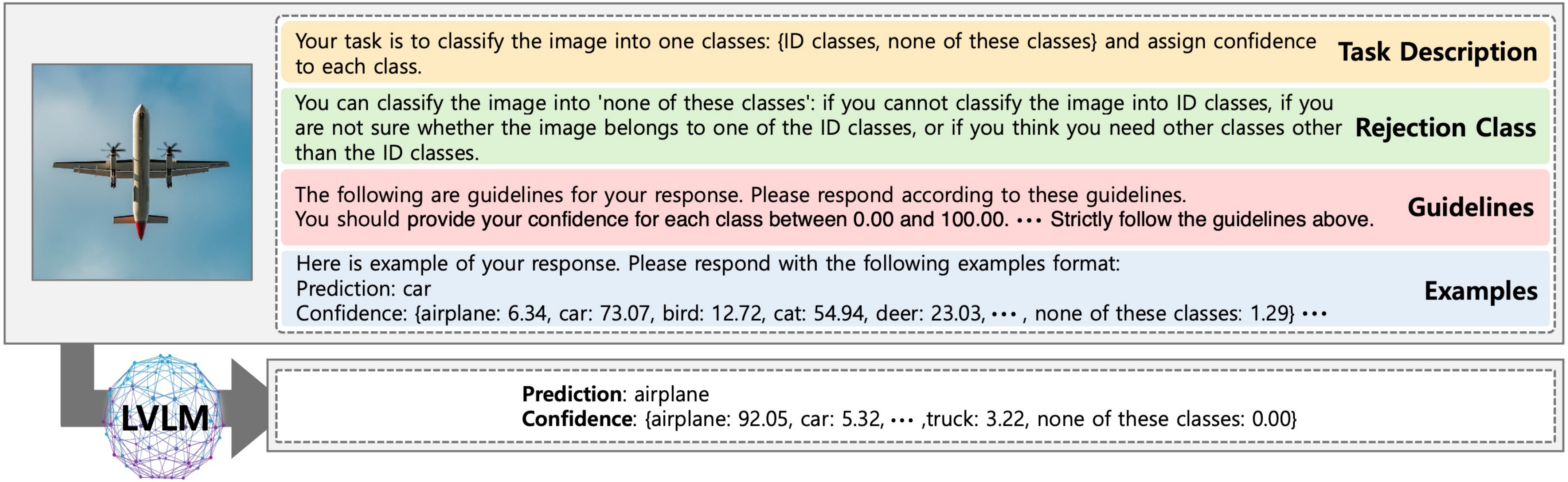


Figure 3: A simplified format of designed prompt for OoDD evaluation on LVLMs

OoDD in LVLMs

Prompt Design

- The task description provides a basic instruction, such as defining the model's objective.

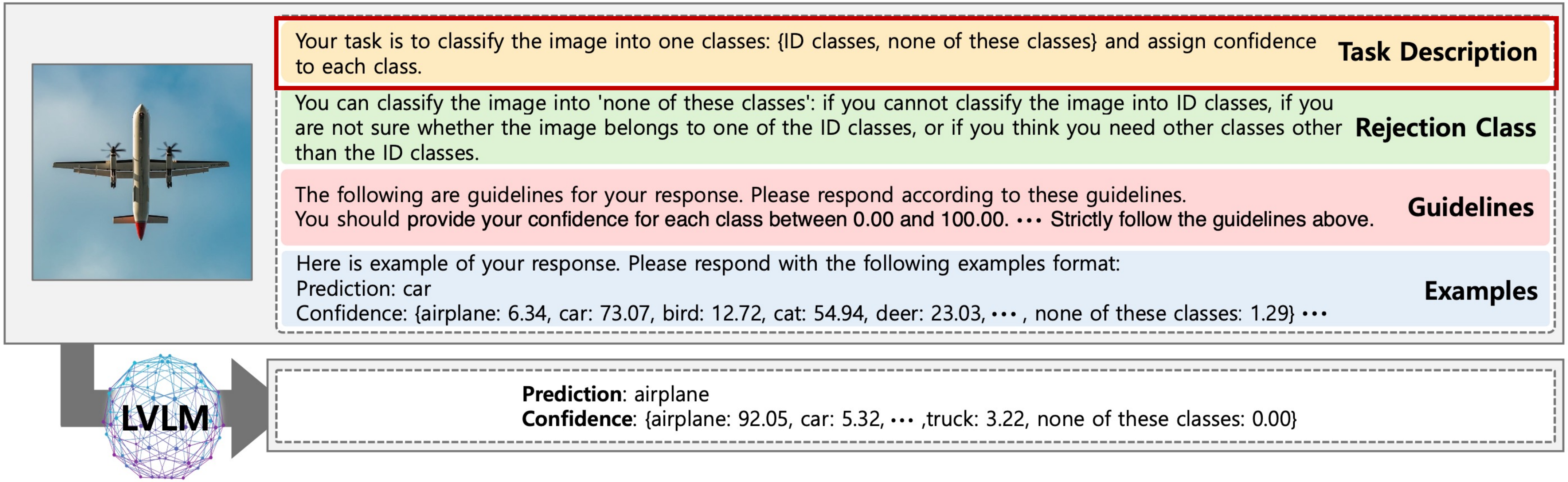


Figure 3: A simplified format of designed prompt for OoDD evaluation on LVLMs

OoDD in LVLMs

Prompt Design

- To mitigate failure cases, we enhance the prompt by adding the following components: **an explanation of the rejection class, guidelines, and examples for the response format**.

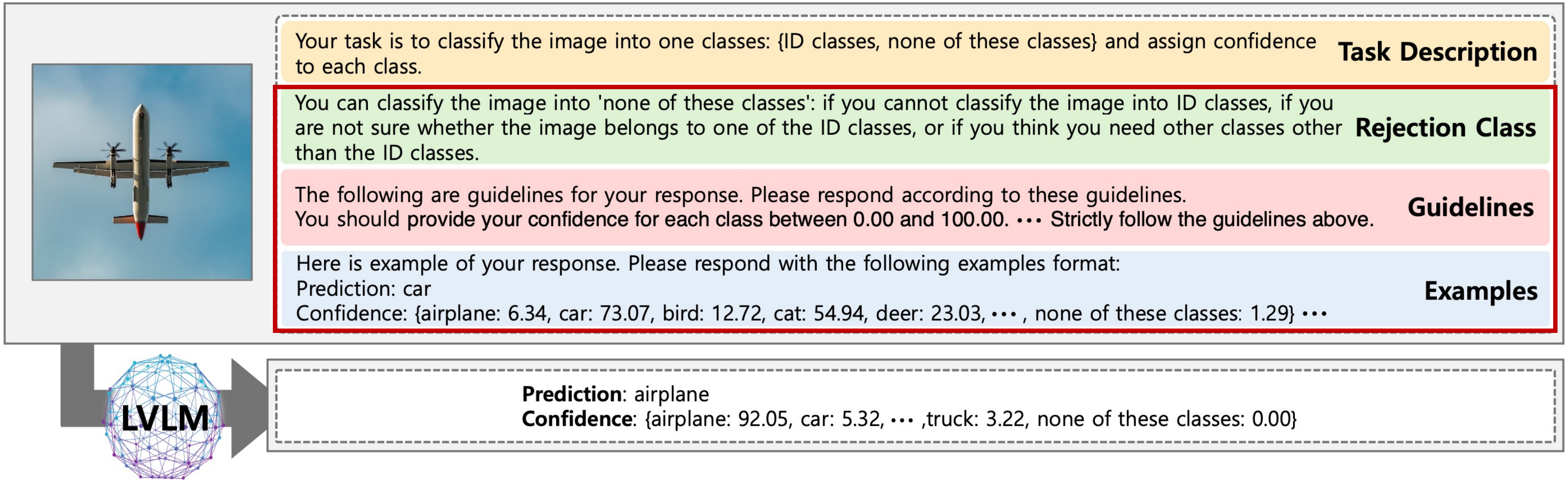


Figure 3: A simplified format of designed prompt for OoDD evaluation on LVLMs

OoDD in LVLMs

OoD Score Design

- We use **the maximum confidence score among the ID classes** as the OoD score.
  - The softmax function is applied to all confidence values to normalize them, including that of the rejection class.

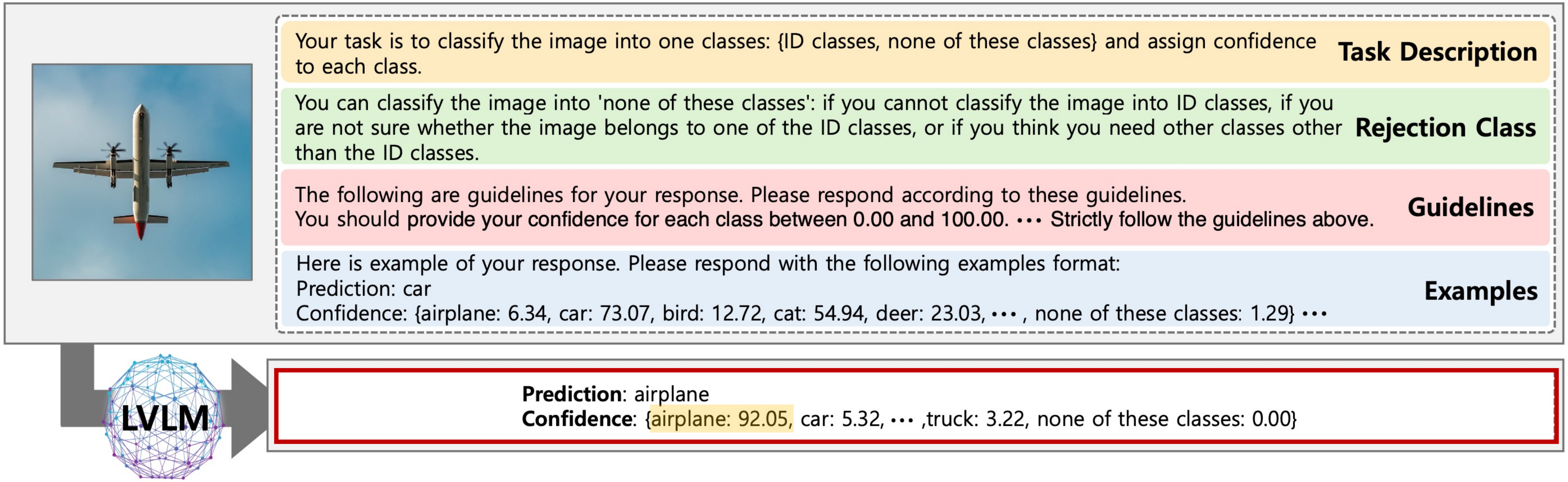


Figure 3: A simplified format of designed prompt for OoDD evaluation on LVLMs

OoDD in LVLMs

Experimental Results

- **Proprietary models outperform the open-source models in most cases**, with reasonable valid response rates.
- **Some open-source models** including LLaVa-v1.6 (Mistral-7B) and GLM-4v-9B **exhibit low OoDD performance despite achieving decent results on popular VLM benchmarks**.

Table 1: The comparison on the ImageNet200 benchmark

		ID	Near-OoD		Far-OoD			All OoD
		IN200	NINCO	SSB-Hard	iNaturalist	Textures	Openimage-O	
Models	Valid	ACC (↑)	FPR@95%TPR (↓) / AUROC (↑)					
SCALE**	-	86.37	84.84			93.98		-
fDBD**	-	86.37	84.27			93.45		-
AugMix+ASH**	-	87.01	55.83 / 85.74	71.22 / 80.00	19.14 / 95.81	21.00 / 95.67	31.06 / 92.51	-
OpenCLIP	100.00 (23,031)	87.41	62.27 / 85.31	71.48 / 78.36	42.76 / 92.49	47.83 / 89.62	47.47 / 90.68	61.42 / 83.54
GPT-4o	85.49 (19,689)	89.78	22.30 / **92.08**	**38.95 / 81.41**	2.06 / 97.58	7.45 / **95.85**	**3.78 / 97.17**	**23.50 / 88.50**
Claude 3.5 Sonnet	80.39 (18,515)	86.06	52.92 / 72.18	78.41 / 58.09	9.23 / 94.93	10.17 / 94.28	18.31 / 89.71	49.01 / 73.69
Gemini Pro 1.5	91.92 (21,170)	88.84	**21.55** / 89.03	55.24 / 77.40	**1.53 / 97.73**	**5.12** / 95.61	5.25 / 96.45	32.97 / 85.74
LLaVA-v1.6	71.63 (16,496)	2.45	100.00 / 50.85	100.00 / 48.95	100.00 / 50.05	100.00 / 59.26	100.00 / 49.23	100.00 / 50.11
GLM-4v	89.00 (20,498)	69.41	100.00 / 79.23	100.00 / 74.35	100.00 / 83.01	100.00 / 83.45	100.00 / 83.11	100.00 / 77.86
InternVL2-26B	62.68 (14,436)	**90.22**	82.59 / 58.32	94.21 / 52.51	36.69 / 81.26	28.08 / 85.56	50.89 / 74.16	75.95 / 61.63
InternVL2-76B	**97.36 (22,424)**	88.30	100.00 / 72.27	100.00 / 62.39	100.00 / 95.57	100.00 / 91.62	100.00 / 90.12	100.00 / 74.14

※ The results of QWEN-VL-Chat are omitted due to its exceptionally low ability to follow instructions, with a valid response rate of less than 1%. You can find QWEN-VL-Chat results on Table B.2.1 in Appendix B.2.

DAINTLAB

OoDD in LVLMs

Experimental Results

- **All compared models have more difficulty in detecting near-OoD than far-OoD.**

Table 1: Comparison on the ImageNet200 benchmark. Full model names are in the footnotes. 'Valid' indicates the ratio of valid responses out of a total of 23,031 image-prompt queries, with counts in brackets. **Bold** highlights the be[illegible]rformance among gene[illegible]e LVLMs.

Models	Valid	ID IN200 ACC (↑)	Near-OoD NINCO	Near-OoD SSB-Hard	Far-OoD iNaturalist	Far-OoD Textures	Far-OoD Openimage-O	All OoD
			FPR@95%TPR (↓) / AUROC (↑)					
SCALE**	-	86.37	84.84			93.98		-
fDBD**	-	86.37	84.27			93.45		-
AugMix+ASH**	-	87.01	55.83 / 85.74	71.22 / 80.00	19.14 / 95.81	21.00 / 95.67	31.06 / 92.51	-
OpenCLIP	100.00 (23,031)	87.41	62.27 / 85.31	71.48 / 78.36	42.76 / 92.49	47.83 / 89.62	47.47 / 90.68	61.42 / 83.54
GPT-4o	85.49 (19,689)	89.78	22.30 / **92.08**	**38.95 / 81.41**	2.06 / 97.58	7.45 / **95.85**	**3.78 / 97.17**	**23.50 / 88.50**
Claude 3.5 Sonnet	80.39 (18,515)	86.06	52.92 / 72.18	78.41 / 58.09	9.23 / 94.93	10.17 / 94.28	18.31 / 89.71	49.01 / 73.69
Gemini Pro 1.5	91.92 (21,170)	88.84	**21.55** / 89.03	55.24 / 77.40	**1.53 / 97.73**	**5.12** / 95.61	5.25 / 96.45	32.97 / 85.74
LLaVA-v1.6	71.63 (16,496)	2.45	100.00 / 50.85	100.00 / 48.95	100.00 / 50.05	100.00 / 59.26	100.00 / 49.23	100.00 / 50.11
GLM-4v	89.00 (20,498)	69.41	100.00 / 79.23	100.00 / 74.35	100.00 / 83.01	100.00 / 83.45	100.00 / 83.11	100.00 / 77.86
InternVL2-26B	62.68 (14,436)	**90.22**	82.59 / 58.32	94.21 / 52.51	36.69 / 81.26	28.08 / 85.56	50.89 / 74.16	75.95 / 61.63
InternVL2-76B	**97.36 (22,424)**	88.30	100.00 / 72.27	100.00 / 62.39	100.00 / 95.57	100.00 / 91.62	100.00 / 90.12	100.00 / 74.14

* OpenCLIP-ViT-B-32, GPT-4o (2024-08-06), LLaVA-v1.6-Mistral-7B, GLM-4v-9B, InternVL2-InternLM2-Chat-26B, InternVL2-LLaMA3-76B

** Results based on 100% of the benchmark from the OpenOOD v1.5 leaderboard.[3] Only the results available from the leaderboard are shown.

OoDD in LVLMs

Experimental Results

- **The proprietary models generally perform on par with or better than the single-modal SOTA OoDD models.**

Table 1: The comparison on the ImageNet200 benchmark

Models	Valid	ID IN200 ACC (↑)	Near-OoD NINCO	Near-OoD SSB-Hard	Far-OoD iNaturalist	Far-OoD Textures	Far-OoD Openimage-O	All OoD
			FPR@95%TPR (↓) / AUROC (↑)					
SCALE**	-	86.37	84.84		93.98			-
fDBD**	-	86.37	84.27		93.45			-
AugMix+ASH**	-	87.01	55.83 / 85.74	71.22 / 80.00	19.14 / 95.81	21.00 / 95.67	31.06 / 92.51	-
OpenCLIP	100.00 (23,031)	87.41	62.27 / 85.31	71.48 / 78.36	42.76 / 92.49	47.83 / 89.62	47.47 / 90.68	61.42 / 83.54
GPT-4o	85.49 (19,689)	89.78	22.30 / **92.08**	**38.95** / **81.41**	2.06 / 97.58	7.45 / **95.85**	**3.78** / **97.17**	**23.50** / **88.50**
Claude 3.5 Sonnet	80.39 (18,515)	86.06	52.92 / 72.18	78.41 / 58.09	9.23 / 94.93	10.17 / 94.28	18.31 / 89.71	49.01 / 73.69
Gemini Pro 1.5	91.92 (21,170)	88.84	**21.55** / 89.03	55.24 / 77.40	**1.53** / **97.73**	**5.12** / 95.61	5.25 / 96.45	32.97 / 85.74
LLaVA-v1.6	71.63 (16,496)	2.45	100.00 / 50.85	100.00 / 48.95	100.00 / 50.05	100.00 / 59.26	100.00 / 49.23	100.00 / 50.11
GLM-4v	89.00 (20,498)	69.41	100.00 / 79.23	100.00 / 74.35	100.00 / 83.01	100.00 / 83.45	100.00 / 83.11	100.00 / 77.86
InternVL2-26B	62.68 (14,436)	**90.22**	82.59 / 58.32	94.21 / 52.51	36.69 / 81.26	28.08 / 85.56	50.89 / 74.16	75.95 / 61.63
InternVL2-76B	**97.36 (22,424)**	88.30	100.00 / 72.27	100.00 / 62.39	100.00 / 95.57	100.00 / 91.62	100.00 / 90.12	100.00 / 74.14

※ The results of QWEN-VL-Chat are omitted due to its exceptionally low ability to follow instructions, with a valid response rate of less than 1%. You can find QWEN-VL-Chat results on Table B.2.1 in Appendix B.2.

OoDD in LVLMs

Experimental Results

- **The open-source models generally perform worse than the single-modal SOTA OoDD models.**

Table 1: The comparison on the ImageNet200 benchmark

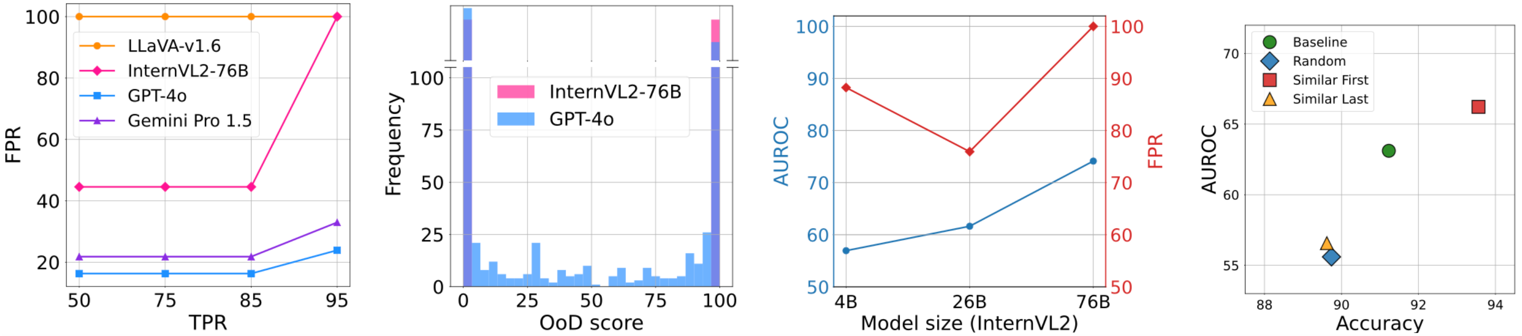
Models	Valid	ID IN200 ACC (↑)	Near-OoD NINCO	Near-OoD SSB-Hard	Far-OoD iNaturalist	Far-OoD Textures	Far-OoD Openimage-O	All OoD
			FPR@95%TPR (↓) / AUROC (↑)					
SCALE**	-	86.37	84.84			93.98		-
fDBD**	-	86.37	84.27			93.45		-
AugMix+ASH**	-	87.01	55.83 / 85.74	71.22 / 80.00	19.14 / 95.81	21.00 / 95.67	31.06 / 92.51	-
OpenCLIP	100.00 (23,031)	87.41	62.27 / 85.31	71.48 / 78.36	42.76 / 92.49	47.83 / 89.62	47.47 / 90.68	61.42 / 83.54
GPT-4o	85.49 (19,689)	89.78	22.30 / **92.08**	**38.95 / 81.41**	2.06 / 97.58	7.45 / **95.85**	**3.78 / 97.17**	**23.50 / 88.50**
Claude 3.5 Sonnet	80.39 (18,515)	86.06	52.92 / 72.18	78.41 / 58.09	9.23 / 94.93	10.17 / 94.28	18.31 / 89.71	49.01 / 73.69
Gemini Pro 1.5	91.92 (21,170)	88.84	**21.55** / 89.03	55.24 / 77.40	**1.53 / 97.73**	**5.12** / 95.61	5.25 / 96.45	32.97 / 85.74
LLaVA-v1.6	71.63 (16,496)	2.45	100.00 / 50.85	100.00 / 48.95	100.00 / 50.05	100.00 / 59.26	100.00 / 49.23	100.00 / 50.11
GLM-4v	89.00 (20,498)	69.41	100.00 / 79.23	100.00 / 74.35	100.00 / 83.01	100.00 / 83.45	100.00 / 83.11	100.00 / 77.86
InternVL2-26B	62.68 (14,436)	**90.22**	82.59 / 58.32	94.21 / 52.51	36.69 / 81.26	28.08 / 85.56	50.89 / 74.16	75.95 / 61.63
InternVL2-76B	**97.36 (22,424)**	88.30	100.00 / 72.27	100.00 / 62.39	100.00 / 95.57	100.00 / 91.62	100.00 / 90.12	100.00 / 74.14

※ The results of QWEN-VL-Chat are omitted due to its exceptionally low ability to follow instructions, with a valid response rate of less than 1%. You can find QWEN-VL-Chat results on Table B.2.1 in Appendix B.2.

OoDD in LVLMs

Experimental Results: Further Analysis

- **Scalability with image resolution**
  - Sec. 3.6, Appx. B.2
- **The trends of confidence scores**
  - Sec. 3.6, Appx. B.2
- **Reasoning for their response**
  - Sec. 3.6, Appx. B.6
- **Confidence scores on ID**
  - Sec. 3.6
- **Scaling law in terms of model size**
  - Sec. 3.6
- **Class order in the prompt**
  - Sec. 3.6, Appx. B.4
- **Response failure**
  - Sec. 3.6, Appx. B.5



Actual		Prediction: ID		Prediction: OoD	
ID	GPT-4o		• Ground-truth: Bald eagle • {**bald eagle**: **98.75**, vulture: 1.0, none of these classes: 0.25} • The bird in the image is ***a clear match to the bald eagle class***, recognizable by its distinctive ***white head, yellow beak, and large wingspan***. Hence, a high confidence is assigned to this class.		• Ground-truth: Centipede • {centipede: 40.00, snail: 45.00, **none of these classes**: **80.00**} • The image appears to ***resemble a millipede-like structure***, but none of the given 200 classes directly match this observation. Thus, the prediction is 'none of these classes' with high confidence.
ID	InternVL2-26B		• Ground-truth: Tree frog • {**tree frog**: **99.99**} • The image shows ***a green frog with a yellow eye, which is characteristic of a tree frog. The frog is perched on a leaf***, and ***its coloration and eye pattern are consistent with tree frogs.***		• Ground-truth: Bucket • {**none of these classes**: **100.00**} • The image shows ***a person in a swimsuit, bending over a bucket***, which appears to be in ***a grassy area***. The image does not clearly depict any of the 200 classes provided, so the prediction is 'none of these classes' with a confidence of 100.00.
OoD	GPT-4o		• Ground-truth: N/A (Openimage-O) • {goldfinch: 5.67, junco: 2.93, bald eagle: 1.42, **lorikeet**: **94.72**, toucan: 8.65, . . . , snail: 0.09, golden retriever: 0.14, ambulance: 0.18, space shuttle: 0.14, none of these classes: 0.13} • The bird in the image closely resembles ***a lorikeet, with its vibrant green plumage and colorful facial markings***. Other birds like ***toucans and goldfinches also show some resemblance***, but ***the lorikeet class has the highest confidence due to the specific combination of colors and features***.		• Ground-truth: N/A (Openimage-O) • {**none of these classes**: **99.99**} • The image shows ***purple ornamental flowers (Allium)***, which do not match any of the given 200 classes.
OoD	InternVL2-26B		• Ground-truth: Araneus gemma (NINCO) • {**spider web**: **99.99**} • The image shows ***a spider on a green plant***, with ***a web*** visible in the background. The ***spider's body and legs are clearly visible***, and the ***web's intricate pattern is evident***. The confidence in the prediction of 'spider web' is high ***due to the clear visibility of the web and the spider's position within it.***		• Ground-truth: Caracal (NINCO) • {**none of these classes**: **99.00**} • The animal in the image is a ***caracal***, which is ***a medium-sized wild cat native to Africa and parts of the Middle East. It is characterized by its reddishbrown coat, long legs, and distinctive blacktufted ears. The caracal is known for its agility and is often found in grasslands and savannas.*** The image shows an animal that matches these characteristics, making it a "none of these classes" with high confidence.

Reflexive Guidance (ReGuide)

Framework

- **Leveraging the LVLM itself** to obtain guidance for OoDD from its powerful zero-shot visual recognition capabilities
- A two-stage simple and model-agnostic prompting strategy

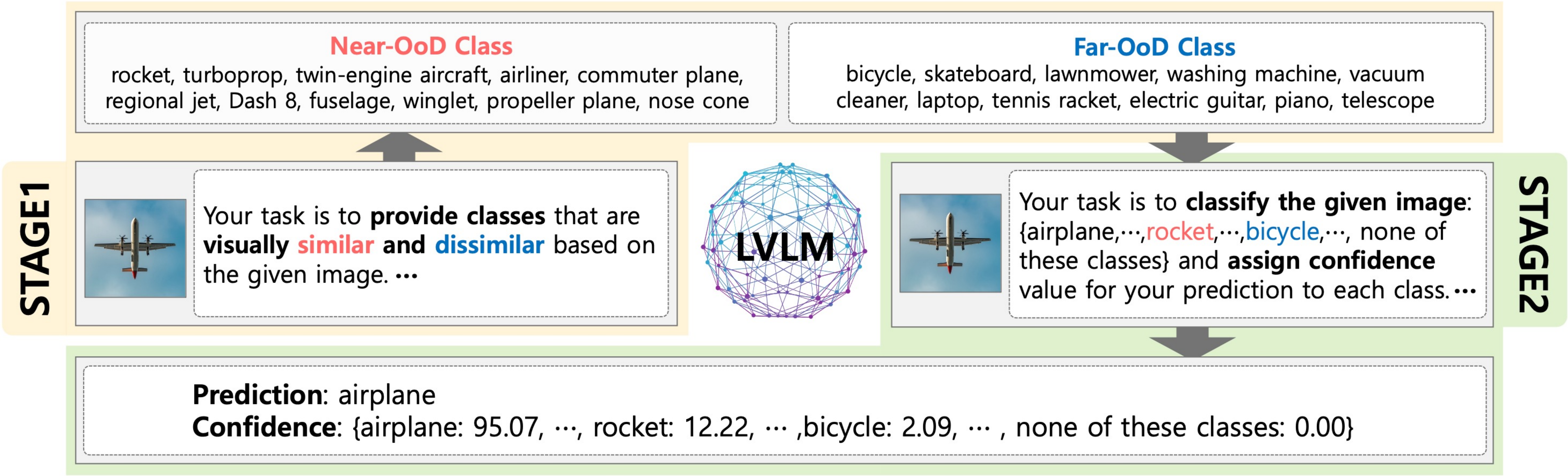
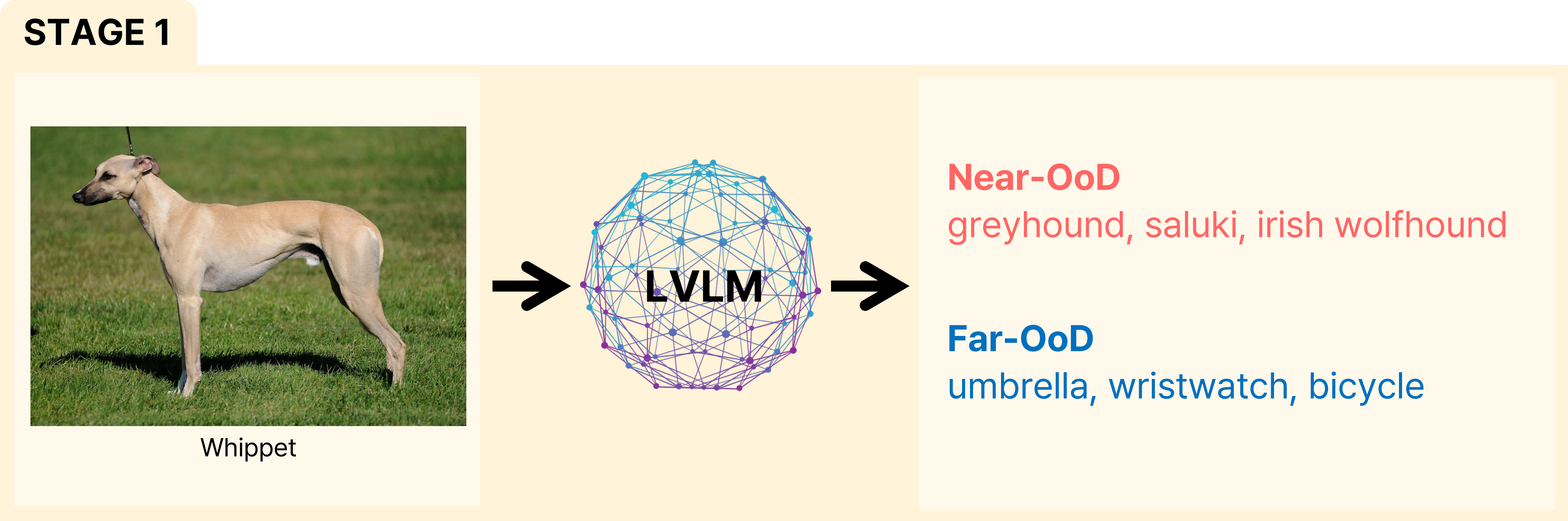


Figure 5: Framework of the proposed Reflexive Guidance for OoDD

Reflexive Guidance (ReGuide)

Stage 1: Image-adaptive Class Suggestions

- The LVLM is asked to **suggest $2N$ class names derived from the given image**.
  - If the input image is OoD, the suggested $2N$ class names can offer potential ground-truth label or class names closely related to the ground-truth label.



Reflexive Guidance (ReGuide)

Stage 2: OoDD with Suggested Classes

- **The suggested $2N$ classes are employed as auxiliary OoD classes.**
  - It is expected that OoD input images can be assigned higher confidence scores for the suggested $2N$ classes than for ID classes.
  - The rejection class ‘none of these classes’ is retained as a fallback.

STAGE 2

Whippet

airplane, bird, cat,
none of these classes,
greyhound, saluki, basenji,
umbrella, wristwatch, bicycle

→

→

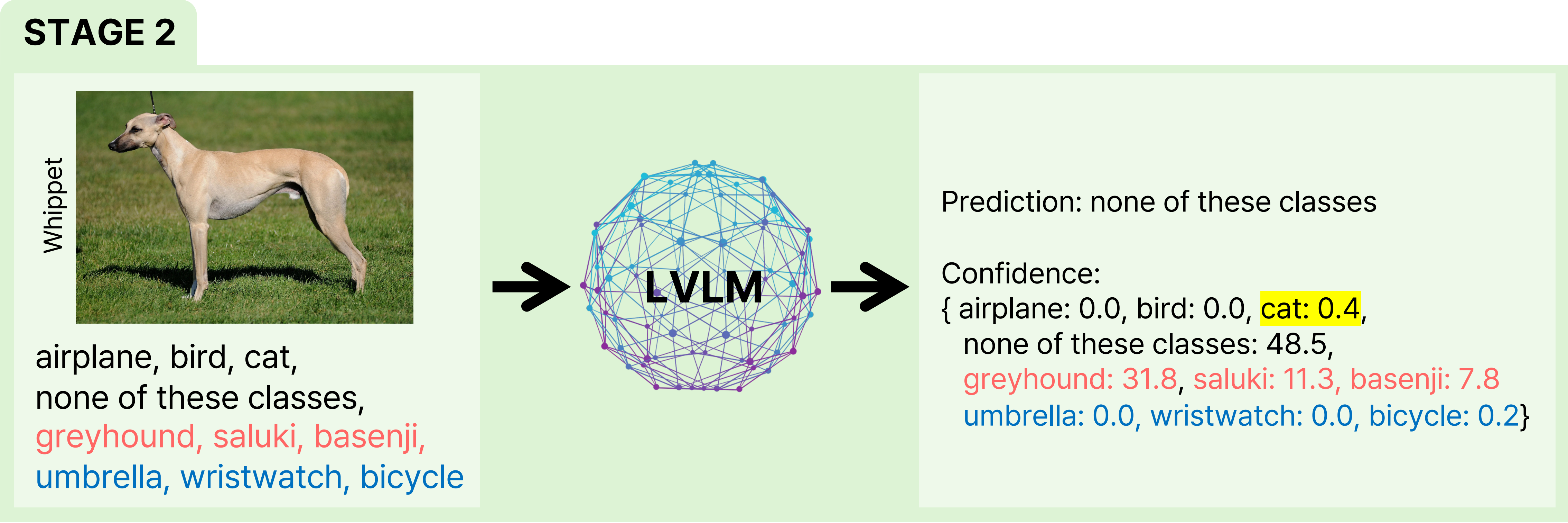
Prediction: none of these classes

Confidence:
{ airplane: 0.0, bird: 0.0, cat: 0.4,
none of these classes: 48.5,
greyhound: 31.8, saluki: 11.3, basenji: 7.8
umbrella: 0.0, wristwatch: 0.0, bicycle: 0.2}

Reflexive Guidance (ReGuide)

OoD Score Design

- **Same as the Baseline: the maximum confidence score among the ID classes**
  - For ID inputs, LVLMs should assign high confidence to one of ID classes.
  - For OoD inputs, LVLMs should assign high confidence to the suggested classes, including the rejection class.



Reflexive Guidance (ReGuide)

Experimental Results

- **ReGuide significantly improves various aspects of the LVLM's performance.**

Table 4: ReGuide effects on the ImageNet200 benchmark

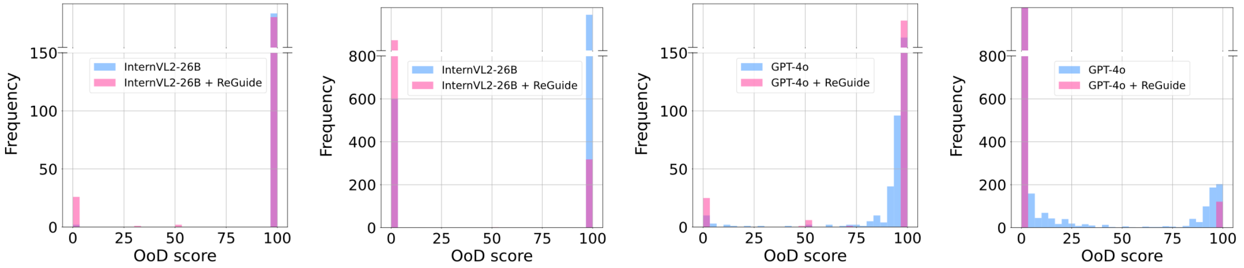
Models	Valid	ID IN200 ACC (↑)	Near-OoD NINCO	Near-OoD SSB-Hard	Far-OoD iNaturalist	Far-OoD Textures	Far-OoD Openimage-O	All OoD
			FPR@90%TPR (↓) / FPR@95%TPR (↓) / AUROC (↑)					
InternVL2-26B	61.01 (2,544)	91.23	82.73 / 82.73 / 58.31	94.34 / 94.34 / 52.51	38.03 / 38.03 / 80.66	28.86 / 28.86 / 85.25	47.91 / 47.91 / 75.72	73.12 / 73.12 / 63.11
+ GPT-text	69.88 (2,914)	89.58	69.44 / 69.44 / 62.17	85.65 / 85.73 / 53.55	26.82 / 28.00 / 84.72	29.20 / 29.20 / 83.51	39.39 / 39.39 / 78.10	62.41 / 62.64 / 65.88
+ ReGuide	**86.14 (3,592)**	**93.53**	**22.39 / 22.89 / 86.53**	**15.21 / 15.21 / 90.41**	**1.39 / 1.39 / 98.02**	**3.93 / 3.93 / 97.05**	**2.04 / 2.04 / 97.68**	**10.24 / 10.27 / 93.19**
InternVL2-76B	**97.26 (4,056)**	89.09	51.28 / **51.28** / 71.89	71.02 / 71.02 / 62.02	2.20 / **2.20** / **96.43**	10.76 / **10.76** / 92.15	14.27 / **14.27** / 90.40	44.46 / **44.46** / 75.30
+ReGuide	95.80 (3,995)	**90.93**	**8.05** / 56.36 / **91.35**	**14.58 / 66.65 / 87.65**	**0.00** / 59.75 / 95.35	**4.08** / 60.00 / **93.38**	**2.02** / 65.46 / **93.95**	**8.92** / 64.36 / **90.60**
GPT-4o	**87.58 (3,652)**	90.64	8.57 / **14.76** / 93.96	29.25 / 34.50 / 82.28	0.81 / **1.83 / 98.11**	5.60 / **6.47** / 95.37	1.21 / **3.63 / 97.82**	15.62 / **19.34** / 89.85
+ ReGuide	79.57 (3,318)	**91.59**	**0.49** / 18.72 / **96.76**	**7.53 / 31.17 / 92.56**	**0.00** / 17.05 / 97.08	**1.32** / 26.43 / **95.96**	**0.15** / 19.66 / 96.82	**4.02** / 25.66 / **94.61**

※ Due to computational and API costs, we evaluate ReGuide with GPT-4o and InternVL2-26B/-76B on a 5% subset of the ImageNet200 benchmark.

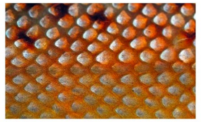
Reflexive Guidance (ReGuide)

Experimental Results: Further Analysis

- **The ratio of OoD inputs predicted to non-ID classes**
  - Sec 4.1
- **ReGuide results on the shared valid query set**
  - Appx. C.2
- **The analysis of suggested classes**
  - Sec 4.1, Appx. C.2
- **Text-adaptive vs. Image-adaptive**
  - Sec. 4.1, Appx. C.2
- **The OoD score distributions and misclassified inputs**
  - Sec. 4.1, Appx. C.3
- **Inference cost**
  - Appx. C.6



(a) InternVL2-26B ID (b) InternVL2-26B OoD (c) GPT-4o ID (d) GPT-4o OoD

		ReGuide suggested OoD classes: Near-OoD	ReGuide suggested OoD classes: Far-OoD	Prediction (GPT-4o / + ReGuide)
ImageNet200	 whippet	greyhound, lurcher, saluki, rhodesian ridgeback, vizsla, basenji, irish wolfhound, doberman, ibizan hound, deerhound, basque shepherd, great dane, galgo, pharaoh hound, borzoi	umbrella, wristwatch, bicycle, electric fan, soccer cleat, refrigerator, telescope, skateboard, basketball hoop, vacuum cleaner, washing machine, piano, coffee maker, laptop, microwave, tennis racket, desk chair, bookcase, alarm clock	greyhound / whippet
SSB Hard	 yellowhammer	oriole, bunting, grosbeak, vireo, sparrow, kinglet, finch, linnet, wren, pine siskin, nuthatch, phoebe, treecreeper, tanager, canary, starling, chickadee, warbler, titmouse	umbrella, wristwatch, bicycle, electric fan, soccer cleat, alarm clock, tennis racket, refrigerator, telescope, skateboard, basketball hoop, vacuum cleaner, washing machine, piano, laptop, microwave, desk chair, bookcase, coffee maker	goldfinch / none of these classes
Textures	 scaly	fish scales, coral, chainmail, turtle shell, reptile skin, armadillo shell, rocky surface, crocodile skin, bark texture, mesh fabric, mineral formation, mosaic, sandstone, fossil texture, lizard skin, dragon scales, pangolin scales, pebbles, snake skin, textured leather	umbrella, wristwatch, bicycle, electric fan, soccer cleat, alarm clock, tennis racket, refrigerator, telescope, skateboard, basketball hoop, vacuum cleaner, washing machine, piano, laptop, microwave, desk chair, bookcase, coffee maker	goldfish / fish scales

Conclusion

DAINTLAB

- **Address the lack of rigorous evaluation and comparison of the OoDD performance of LVLMs.**
- **Establish a framework to evaluate and compare various proprietary and open-source LVLMs.**
  - Overall, proprietary LVLMs outperform open-source LVLMs in both image classification and OoDD tasks.
  - Open-source LVLMs tend to be overconfident in their response, highlighting the need for confidence calibration.
- **Propose ReGuide, a self-guided prompting approach that enhances the OoDD capabilities of LVLMs by leveraging self-generated, image-adaptive concepts.**
  - ReGuide significantly boosts the OoDD performance of both proprietary and open-source LVLMs.
- LVLMs **tend to avoid generating responses that fall outside the given prompt categories. Simply leveraging their intrinsic abilities, it can effectively expand their scope of thinking.**

Thank you

For more results and a detailed analysis, please refer to the paper.

https://openreview.net/forum?id=R4h5PXzUuU

You can also find the sampled dataset we used and input prompt-generated response pairs on both Github and Huggingface.

https://github.com/daintlab/ReGuide

https://huggingface.co/datasets/daintlab/reguide