

A Experimental Details

A.1 Dataset Details

We evaluate the performance on both coarse-grained and fine-grained out-of-distribution (OOD) detection tasks, categorized based on the semantic similarity between in-distribution (ID) and OOD samples.

A.1.1 Coarse-grained datasets

Coarse-grained OOD detection typically involves ID and OOD datasets from semantically disjoint domains. This setting is widely adopted in prior works, where large distributional shifts make detection easier for vision-language models. We use the following ID datasets: CUB-200 [40] (birds), Stanford-Cars [41] (vehicles), Food-101 [42] (food), Oxford-Pet [43] (pets), and ImageNet-1K [44] (general objects). The corresponding OOD datasets include iNaturalist [45], SUN [46], Places [47], and Textures [48], as compiled in MOS [39]. These dataset pairs are carefully selected to ensure no semantic overlap between ID and OOD classes.

ImageNet-1K [44] includes 1,000 object categories with 1.28M training, 50K validation, and 100K test images. We use the validation set as the ID data.

iNaturalist [45] contains 859K images of over 5,000 fine-grained species. Following [45], we sample 10,000 images from 110 plant categories not present in ImageNet as OOD.

SUN [46] is a scene dataset with 130,519 images across 397 categories. We sample 10,000 images from 50 nature-related classes not included in ImageNet.

Places [47] comprises over 2.5M images in 205 scene categories. We sample 10,000 images from 50 manually selected non-overlapping categories.

Textures [48] contains 5,640 images from 47 texture categories. The full dataset is used as OOD.

A.1.2 Fine-grained datasets

Fine-grained OOD detection poses a greater challenge, as ID and OOD samples belong to visually similar subcategories within the same high-level category (e.g., different cat breeds). To construct such datasets, we randomly split each dataset—CUB-200, Oxford-Pet, Food-101, and Stanford-Cars—into disjoint subsets: half of the classes are used as ID, and the remaining half as OOD, ensuring no class overlap.

Stanford-Cars [41] includes 16,185 images of 196 car types, categorized by make, model, and year. The data is split into 8,144 training and 8,041 testing images.

CUB-200 [40] contains 11,788 bird images across 200 species, with 5,994 for training and 5,794 for testing.

Oxford-Pet [43] includes 7,390 images from 37 cat and dog breeds, with approximately 200 images per class.

Food-101 [42] comprises 101 food categories, with 750 training and 250 test images per class, totaling 101,000 images.

A.2 Implementation Details

A.2.1 Experimental Environment

Refer-OOD does not require any training. All experiments are conducted under the zero-shot setting without training samples. All experiments are performed using PyTorch 2.0.1 and Python 3.10.4, running on an NVIDIA GeForce RTX 3090.

A.2.2 Evaluation Metrics

VLM Metrics. For VLM-based models, we adopt two commonly used OOD detection metrics: (1) **FPR95**, which denotes the false positive rate of OOD samples when the true positive rate of

Table 7: Computational cost of Refer-OOD.

(a) Refer-OOD-VLM			(b) Refer-OOD-MLLM		
Generate		Retrieve	Textual		Visual
Time	15.32s/100 labels	3.70s/image	Time	24.10s/sample	35.89s/sample

in-distribution (ID) samples reaches 95%. A lower FPR95 indicates better detection performance; (2) **AUROC**, the Area Under the Receiver Operating Characteristic curve, where a higher value reflects stronger discriminative capability.

MLLM Metrics. For MLLM-based models, we prompt the model to generate binary (yes/no) responses indicating whether a given sample belongs to the ID distribution. We evaluate performance using the following metrics: (1) **F1 Score**, which is the harmonic mean of precision and recall, reflecting the balance between false positives and false negatives; and (2) **Accuracy (ACC)**, which measures the proportion of correctly classified ID samples. Higher values of both metrics indicate better detection performance.

A.2.3 Efficiency

We report the computational cost of the Refer-OOD method. While Refer-OOD eliminates the need for pre-training, it introduces extra test time costs due to the generation (15.32 seconds/100 generated labels) and retrieval (3.70 seconds/retrieved image) processes for VLM-based pipeline. The whole MLLM-based pipeline costs 24.10 and 35.89 seconds per test image for the textual and visual references pipeline, respectively.

B More Results and Analysis

B.1 Detailed results

Detailed results for coarse-grained datasets. Table 8 summarizes the performance of various CLIP-based OOD detection methods on four coarse-grained ID datasets (CUB-200, Stanford-Cars, Food-101, Oxford Pets) against four OOD datasets (iNaturalist, SUN, Places, Texture). Due to the large semantic gap between Oxford Pets (pet images) and Texture (surface patterns), this setting is relatively less challenging, enabling most methods to achieve high OOD detection performance. Despite this, Refer-OOD-VLM consistently outperforms all baselines across datasets, achieving near-zero FPR95 (0.0–0.06) and AUROC close to 100%, demonstrating strong OOD discrimination with minimal false positives.

Detailed results for ImageNet-1k. Table 9 reports the performance comparison on ImageNet-1k. Compared to previous datasets, ImageNet-1k poses a greater challenge due to closer semantic overlap between ID and OOD samples, such as landscape categories shared with iNaturalist. Despite the increased difficulty, Refer-OOD-VLM achieves the best average performance across all datasets, demonstrating strong robustness under challenging conditions.

Detailed results for MLLM coarse-grained datasets. Table 10 compares Refer-OOD-MLLM with baseline methods on coarse-grained OOD detection. MLLM-based models show strong performance in detecting OOD samples with clear semantic differences from ID categories. For example, even baselines achieve high accuracy on datasets like iNaturalist when ID is "pets," reflecting the perceptual power of large models. Refer-OOD-MLLM further corrects baseline errors and improves ID recall—for instance, boosting CUB recall from 67.25% to 84.95%. These improvements in both precision and recall result in higher F1 scores across benchmarks.

B.2 Ablations

Effect of n_i^I . Figure 4 illustrates the impact of varying the number of visual reference samples per class on FPR95 across four datasets. As n_i^I increases from 1 to 20, FPR95 drops significantly for most datasets, indicating that more references lead to better representation and improved OOD detection. For example, on CUB and PET, performance improves notably when $n_i^I \leq 20$, then

Table 8: Performance comparison for VLM-based methods on coarse-grained datasets.

ID datasets	Methods	OOD datasets									
		iNaturalist		SUN		Places		Texture		Average	
		FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
CUB-200	CLIPN	6.58	98.78	2.13	99.51	4.29	99.07	7.18	98.58	5.04	98.98
	MCM	9.67	98.26	4.8	99.1	6.46	98.62	6.7	98.79	6.91	98.69
	EOE	0.06	99.98	0.03	100.0	0.34	99.89	0.0	100.0	0.11	99.97
	NegLabel	1.27	99.67	0.41	99.84	1.59	99.60	0.20	99.92	0.87	99.76
	CSP	0.96	99.76	0.26	99.93	1.59	99.63	0.05	99.98	0.72	99.83
	Refer-OOD-VLM	0.05	99.99	0.01	100.0	0.29	99.91	0.0	100.0	0.0	99.97
STANFORD-CARS	CLIPN	0.06	99.8	0.02	99.95	0.27	99.9	0.02	99.96	0.09	99.9
	MCM	0.04	99.78	0.0	99.96	0.24	99.9	0.02	99.96	0.08	99.9
	EOE	0.0	100.0	0.02	99.99	0.21	99.95	0.0	100.0	0.06	99.98
	NegLabel	0.03	99.99	1.17	99.68	3.38	99.18	1.95	99.57	1.63	99.60
	CSP	0.02	100	0.25	99.95	1.23	99.71	0.59	99.9	0.52	99.89
	Refer-OOD-VLM	0.0	100.0	0.01	100.0	0.04	99.99	0.0	100.0	0.0	100
Food-101	CLIPN	0.78	99.75	0.51	99.86	1.4	99.7	4.56	98.52	1.81	99.46
	MCM	0.68	99.79	0.9	99.76	1.85	99.59	4.11	98.64	1.89	99.44
	EOE	0.07	99.99	0.0	100.0	0.15	99.97	2.52	99.06	0.69	99.76
	NegLabel	0	100	0.2	99.94	0.5	99.85	3.24	99.17	0.99	99.74
	CSP	0	100	0.02	99.98	0.37	99.9	1.84	99.58	0.56	99.87
	Refer-OOD-VLM	0.0	100.0	0.0	100.0	0.0	100.0	0.21	99.96	0.05	99.99
Oxford Pet	CLIPN	1.37	99.51	0.85	99.79	1.37	99.67	0.69	99.82	1.07	99.7
	MCM	2.71	99.39	1.03	99.74	1.97	99.58	0.8	99.82	1.63	99.63
	EOE	0.0	100.0	0.02	99.99	0.14	99.96	0.12	99.97	0.07	99.98
	NegLabel	0.02	100	1.18	99.68	2.48	99.22	1.91	99.39	1.40	99.57
	CSP	0.02	99.99	0.59	99.82	3.50	98.98	1.44	99.48	1.39	99.57
	Refer-OOD-VLM	0.0	100.0	0.01	100.0	0.14	99.97	0.04	100.0	0.05	99.99
Average	CLIPN	2.2	99.46	0.88	99.78	1.83	99.59	3.11	99.22	2.0	99.51
	MCM	3.27	99.31	1.68	99.64	2.63	99.42	2.91	99.3	2.63	99.41
	EOE	0.03	99.99	0.02	100.0	0.21	99.94	0.66	99.76	0.23	99.92
	NegLabel	0.33	99.91	0.74	99.78	1.98	99.46	1.82	99.51	1.21	99.66
	CSP	0.25	99.93	0.28	99.92	1.67	99.55	0.98	99.73	0.79	99.78
	Refer-OOD-VLM	0.01	100.0	0.01	100.0	0.12	99.97	0.06	99.99	0.03	99.99

Table 9: Performance comparison for VLM-based methods on ImageNet-1k dataset.

ID datasets	Methods	OOD datasets									
		iNaturalist		SUN		Places		Texture		Average	
		FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
ImageNet-1k	CLIPN	29.73	94.86	34.54	93.11	41.55	90.52	64.59	82.27	42.6	90.19
	MCM	31.29	94.63	38.02	92.57	44.49	89.97	57.82	86.2	42.91	90.84
	NegLabel	8.06	98.22	29.76	93.97	39.3	90.77	52.98	86.79	32.27	92.44
	CSP	6.41	98.56	23.17	95.14	38.01	90.88	38.16	90.90	26.44	93.87
	EOE	15.1	97.09	21.14	95.72	29.68	93.17	57.75	86.03	30.92	93.0
	Refer-OOD-VLM	14.42	96.01	24.88	93.97	31.16	93.29	28.81	92.85	24.81	94.03

Table 10: Performance comparison for MLLM-based methods on coarse-grained datasets.

ID datasets	Methods	OOD datasets										
		iNaturalist		SUN		Places		Texture		Average		
		Precision	F1	Precision	F1	Precision	F1	Precision	F1	Precision	F1	Recall
CUB-200	Vanilla	100	80.42	98.70	80.00	96.2	79.16	100	80.42	98.72	80.00	67.25
	Refer-OOD-MLLM	100	91.86	100	91.86	96.96	90.56	100	91.86	99.24	91.27	84.95
Stanford-Cars	Vanilla	100	81.13	100	81.13	100	81.13	100	81.13	100	81.13	68.25
	Refer-OOD-MLLM	100	82.79	100	82.79	100	82.79	100	82.79	100	82.79	70.63
Food-101	Vanilla	100	93.51	100	93.51	100	93.51	98.29	92.76	99.57	93.32	87.81
	Refer-OOD-MLLM	100	94.36	100	94.36	100	94.36	98.87	93.86	99.71	94.02	89.34
OxfordPet	Vanilla	100	80.00	100	80.00	100	80.00	97.95	79.93	99.48	79.98	66.66
	Refer-OOD-MLLM	100	92.53	100	92.53	100	92.53	98.41	91.85	99.60	92.56	86.11

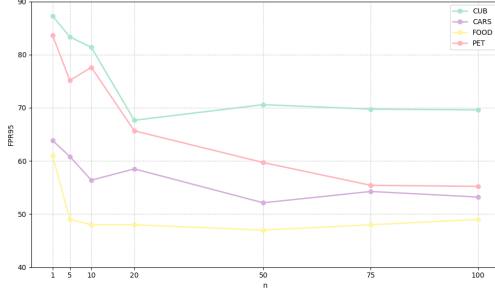


Figure 4: Comparison results on n_i^1 .

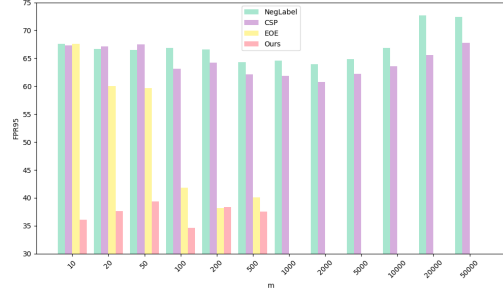


Figure 5: Comparison results on m .

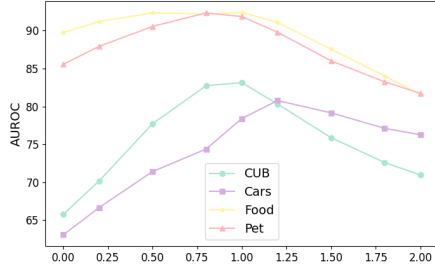


Figure 6: Comparison results on β .

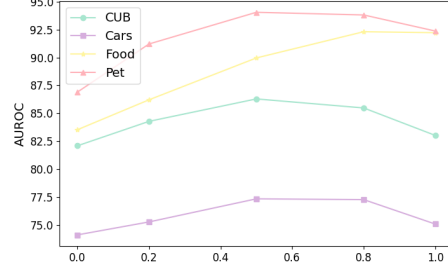


Figure 7: Comparison results on α .

897 stabilizes. This aligns with our theoretical analysis: a larger n_i^1 provides richer semantic coverage,
 898 reducing uncertainty in distinguishing OOD samples.

899 **Effect of m .** Figure 5 illustrates the impact of the candidate sample number m on FPR95 across
 900 four methods. For generative methods such as Refer-OOD and EOE, we limited the number of
 901 generated labels to a maximum of 500 to control API usage and time costs, while for mining-based
 902 methods, we tested up to 50000 labels. Both EOE and our method achieve strong performance even
 903 with small m , indicating superior sample efficiency. Our method consistently outperforms EOE. In
 904 contrast, mining-based approaches such as CSP and NegLabel require larger m to be effective and
 905 suffer performance degradation when m becomes too large due to the introduction of irrelevant labels.
 906 Overall, our method achieves the lowest FPR95 across all m , demonstrating strong robustness.

907 **Effect of β .** Figure 6 illustrates the effect of varying $\beta \in [0, 2]$ on OOD detection performance across
 908 different fine-grained datasets. When $\beta = 0$, Eq. 9 reduces to the traditional Maximum Softmax
 909 Probability (MSP)[50], relying solely on classification confidence. As β increases, performance
 910 consistently improves, highlighting the benefit of incorporating OOD references into the scoring
 911 function. The optimal performance is generally observed around $\beta = 1$, where ID and OOD
 912 components are equally weighted. This suggests that balancing both sources of information leads to
 913 more reliable detection.

914 **Effect of α .** Figure 7 analyzes the role of $\alpha \in [0, 1]$, which controls the fusion of textual and
 915 visual modalities in the final OOD score. When $\alpha = 0$, only textual reference is used, while $\alpha = 1$
 916 relies solely on visual similarity. Both extremes underperform due to the lack of complementary
 917 modality. By adjusting α , the model effectively integrates multimodal references, resulting in superior
 918 performance compared to using a single modality.

919 **Effect of τ .** As presented in Table 11, the optimal value of τ varies with dataset granularity. For
 920 coarse-grained datasets, the highest AUROC is achieved around $\tau = 0.80$, whereas fine-grained
 921 datasets perform best when τ is set closer to 0.95. This difference arises from the semantic similarity
 922 between ID and OOD samples. In fine-grained settings, the text labels associated with OOD samples
 923 are often semantically close to those of ID classes. As a result, a low τ may incorrectly filter out the
 924 correct OOD label, leading to degraded detection performance.

Table 11: Performance comparison(AUROC) with varying τ on coarse-grained and fine-grained datasets.

τ	Coarse-grained datasets				Fine-grained datasets			
	iNaturalist	SUN	Places	Texture	CUB	Stanford-Cars	Food	Oxford-Pet
0.70	95.55	77.86	70.97	59.04	61.41	47.99	64.91	87.64
0.75	97.27	90.24	83.88	81.69	65.44	56.89	75.68	88.45
0.80	96.86	94.39	90.57	92.37	72.99	68.16	88.20	88.88
0.85	95.68	93.68	90.38	92.95	73.45	70.54	88.87	91.09
0.90	94.78	93.95	90.38	92.95	80.53	73.57	88.77	90.90
0.95	94.45	93.73	89.91	93.07	81.89	74.03	89.09	91.65
1.00	94.37	93.65	89.80	92.95	82.18	72.79	87.68	91.55

Table 12: Performance comparison on score functions.

Method	CUB		Stanford-Cars		Food		Oxford-Pet		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
MSP	83.72	67.5	84.02	68.76	44.1	91.67	64.03	84.88	68.97	78.2
Energy	79.61	85.67	78.19	84.34	71.52	84.92	92.71	66.59	80.51	80.38
Maxlogits	59.74	89.77	63.79	87.56	59.81	87.17	85.2	72.58	67.14	84.27
Ours	60.64	80.88	58.74	75.45	34.40	91.50	38.19	92.28	47.99	85.02

Score functions Comparison. Table 12 compares different score functions, showing that our proposed method outperforms other baseline approaches. Table 13 further compares the summation and maximum strategies over logits in Equation 9 across various methods. The maximum strategy consistently outperforms summation, likely because summing logits dilutes discriminative semantics, especially when ID and OOD samples are semantically similar in fine-grained settings.

VLM Architecture Comparison. Table 14 compares our method with the baseline EOE across CLIP, ALIGN, and AltCLIP backbones on four fine-grained datasets. Refer-OOD-VLM consistently achieves lower FPR95 and higher AUROC than EOE. These results show that Refer-OOD-VLM generalizes well across architectures and effectively enhances OOD detection.

MLLM Architecture Comparison. Tables 15 and 16 demonstrate Refer-OOD’s effectiveness across MLLM architectures including GPT-4o, consistent with Qwens’ results.

Online Retrieval Comparison. As shown in Table 17, using multimodal references significantly outperforms using textual labels alone, improving both FPR95 and AUROC. The performance remains stable across different retrieval APIs (Bing, Google, Baidu), highlighting the robustness of Refer-OOD.

Table 13: Comparison of logit aggregation strategies across methods.

Score	NegLabel		CSP		EOE		Refer-OOD-VLM	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
Sum	91.03	64.75	91.09	65.25	92.14	67.81	92.83	60.49
Max	64.56	87.45	62.97	89.25	55.17	90.3	38.19	92.28

Table 14: Comparison results on different VLM backbones.

Backbone	Methods	CUB		Stanford-Cars		Food		Oxford-Pet		Average	
		FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
CLIP	EOE	74.13	73.18	77.6	70.98	39.66	91.54	55.17	90.3	61.64	81.5
	Refer-OOD-VLM	60.64	80.88	58.74	75.45	34.40	91.50	38.19	92.28	47.99	85.02
ALIGN	EOE	87.48	61.32	79.44	73.32	64.66	83.30	44.51	87.42	69.02	76.34
	Refer-OOD-VLM	86.27	68.99	51.06	81.37	57.58	85.93	85.07	71.91	69.99	77.05
AliCLIP	EOE	74.54	70.81	57.44	79.23	34.74	92.47	26.21	92.5	48.23	83.75
	Refer-OOD-VLM	57.84	86.76	53.19	84.19	36.25	92.36	38.81	89.60	46.52	88.22

Table 15: Comparison results on different MLLM for generation.

Retrieval	CUB		Stanford-Cars		Food		Oxford-Pet		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
EOE	87.48	61.32	79.44	73.32	64.66	83.30	44.51	87.42	69.02	76.34
Refer-OOD-VLM with Qwen	49.02	83.02	53.19	75.08	39.44	92.22	34.33	92.37	43.99	85.67
Refer-OOD-VLM with GPT-4o-turbo	67.65	65.72	38.30	85.10	23.18	94.54	32.84	91.86	40.49	84.30

C Case study

C.1 Specific example for MLLM generation prompt

The generated labels. To obtain OOD class labels automatically generated by the MLLM, we design a question-and-answer template and append a hyphen (-) to the end of the answer prompt to encourage a bullet-point list output. The prompt for Pet detection is shown in Fig 8.

C.2 Case study for Refer-OOD-MLLM

For the MLLM-based method, common errors of the baseline method can be summarized into three categories (1) misclassifying ID test samples as OOD; (2) identifying ID test samples as ID but with incorrect label prediction; (3) misclassifying OOD samples as ID. Refer-OOD-MLLM effectively corrects the above errors by introducing multimodal references. The specific method flow is shown in Figure 9.

Misclassifying ID. As shown in Figure 9a, the baseline method incorrectly detects the ID sample (Red-bellied Woodpecker) as OOD, while Refer-OOD-MLLM successfully generates reasonable category predictions by enhancing semantic understanding, thereby correctly identifying the ID sample.

Misclassifying OOD. As shown in Figure 9b. When the baseline method is limited by the limited ID label range and mistakenly identifies the OOD sample as ID, Refer-OOD-MLLM is able to generate the accurate OOD label (Black-footed Albatross), thereby achieving correct OOD sample detection.

Incorrect label prediction. As shown in Figure 9c, the baseline method misclassifies the ID test sample (House Wren) as Bewick’s Wren. Although both belong to the Wren category, the subtle semantic differences pose a challenge to the accuracy of the model. Refer-OOD-MLLM first generates an OOD label (Marsh Wren) and the max probability ID label (House Wren) respectively, then performs reasoning and finally correctly identifies the ID attribute of the image and accurately predicts the category label.

Table 16: Comparison results on different MLLM for detection.

Backbone	Methods	CUB			Stanford-Cars			Food			Oxford-Pet			Average		
		Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Qwen	Vanilla	82.35	60.43	69.70	86.17	59.12	70.12	91.83	64.74	75.94	80.64	70.42	75.18	85.24	63.67	72.73
	Refer-OOD-MLLM	85.29	63.50	72.80	76.59	79.12	77.83	79.59	83.87	81.67	85.48	92.98	89.07	81.73	79.86	80.34
GPT-4o-turbo	Vanilla	93.13	54.28	68.59	69.14	85.55	63.41	90.81	65.92	76.39	91.93	66.27	77.02	86.25	68.00	71.35
	Refer-OOD-MLLM	80.39	82.17	81.77	82.97	62.90	71.55	84.69	59.24	86.91	96.77	77.92	86.33	86.20	70.55	81.64

Table 17: Comparison results on retrieval Api.

Retrieval	CUB		Stanford-Cars		Food		Oxford-Pet		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
Textual	49.02	83.02	53.19	75.08	39.44	92.22	34.33	92.37	43.99	85.67
Bing	52.94	85.48	51.06	77.27	32.83	92.31	29.85	93.81	41.67	87.21
Baidu	56.00	85.03	45.45	78.16	31.52	92.03	35.82	92.24	42.19	86.86
Google	51.23	84.88	45.74	81.52	32.46	92.40	32.84	92.42	40.56	87.80

Question: Provide {nums} additional distinct categories that are visually similar in the showing picture, and haven't been mentioned in {test_labels}. You only need to tell the category, no explanation.

Answer: The {nums} additional categories are:

- category 1
- category 2
- category 2

- Mallard
- Wood duck
- American black duck

Figure 8: Example for MLLM label generation prompt.

Generated OOD label: Red bellied Woodpecker
The max probability ID label: Red bellied Woodpecker

Question: According to references, the image shows?
Answer: The bird in the picture is a Red-bellied Woodpecker. This identification is based on several key features:

1. **"Head and Neck":** The bird has a red cap on its head, which is characteristic of the Red-bellied Woodpecker.
2. **"Beak":** It has a long, chisel-like beak, typical of woodpeckers.
3. **"Body and Wings":** The body shows a mix of black and white barred patterns on the wings and back.
4. **"Belly":** Although not clearly visible in this image, Red-bellied Woodpeckers have a pale belly with a hint of red, which gives them their name.

These features match those of a Red-bellied Woodpecker.

Question: According to the analysis, is it in one of {ID labels}?

Answer: Yes. Red bellied Woodpecker.

(a) Misclassifying ID

Generated OOD label: Black footed Albatross
The max probability ID label: Laysan Albatross

Question: According to references, the image shows?
Answer: The bird in the picture is a "Black-footed Albatross". This identification can be made based on several key features:

1. **"Coloration":** The Black-footed Albatross has a predominantly dark brown body with a lighter face and neck, which is visible in this image.
2. **"Beak":** The beak of the Black-footed Albatross is pale pinkish-gray with a darker tip, as seen here.
3. **"Eye Ring":** There is a noticeable dark eye ring around the eye, which is characteristic of the Black-footed Albatross.

In contrast, the Laysan Albatross typically has a white head and upper body with a more uniformly colored beak that lacks the darker tip seen in the Black-footed Albatross.

Question: According to the analysis, is it in one of {ID labels}?

Answer: No.

(b) Misclassifying OOD

Generated OOD label: Marsh Wren
The max probability ID label: House Wren

Question: According to references, the image shows?
Answer: The bird in the picture is a House Wren. Here are some key features that help identify it:

1. **"Size and Shape":** The bird is small with a relatively large head, short neck, and an upright posture.
2. **"Coloration":** It has brown plumage with darker streaks on its back and wings, which is typical of House Wrens.
3. **"Beak":** The beak is short and slightly curved, suitable for catching insects.
4. **"Tail":** Although not fully visible in this image, House Wrens often hold their tails upright.

Marsh Wrens, on the other hand, have a more slender build, longer tail (often held upright), and a distinctive white eyebrow stripe, which is not present in this bird. Therefore, based on these characteristics, the bird in the picture is a House Wren.

Question: According to the analysis, is it in one of {ID labels}?

Answer: Yes. House Wren.

(c) Incorrect label prediction

Figure 9: Case study for Refer-OOD-MLLM