# Enhancing Cognition and Explainability of Multimodal Foundation Models with Self-Synthesized Data

**Yucheng Shi[1], Quanzheng Li[2], Jin Sun[1], Xiang Li[2], Ninghao Liu[1]**

1.University of Georgia. 2. Massachusetts General Hospital and Harvard Medical School.

## Background

- Large Multimodal Models (LMMs) are good at general visual question answering.
- However, they fail at **fine-grained visual identification**.
- Moreover, fail to provide **valid explanations**.

**User:** Identify the bird species and explain your reason.

**LMM:** The bird species shown in this image is a *seagull*. The reason for this identification is that ***the bird is flying over the ocean***, which is a typical habitat for seagulls.

**Label:** *Black footed Albatross*

**LMM:** The bird species is a *black crow*. The reason for this identification is that ***the bird is perched on a tree branch***, which is a common behavior for crows.

**Label:** *Groove billed Ani*

## Challenge:

How to empower LMMs to provide **interpretable and accurate** visual identification?

**Visual Fine-tuning** sounds good!

A fine-tuning training sample.

**X:**

**Q**: What is this bird's species? Explain your reason.

**A:** Synthetic Answer

**#only label**
**A:** This is a Black Footed Albatross.

**#label with general label-level features**
**A:** ... Black Footed Albatross, because ...*black feet; long, narrow wings;*..., *hooked bill*..., *dark eyes*

**#label and specific image-level features**
**A:** ... a Black Footed Albatross based on its *large, hooked bill, dark feathers, small, dark eyes on either side of its head ....*

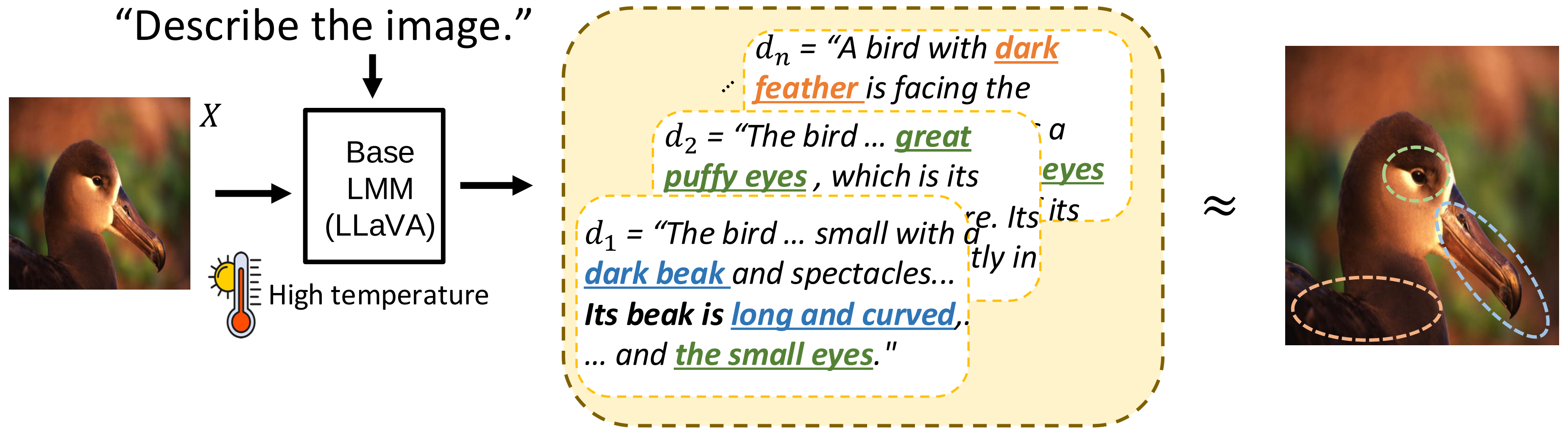
**Task**: Identify ***visual concepts*** in each image.

## Information Bottleneck (IB)

Given an image $X$, label-level concepts $Z$, we propose using **Information Bottleneck (IB)** principle to select image-level concepts $Z^* \subseteq Z$:

$$Z^* = \arg\max_{Z' \subseteq Z} [I(X; Z') - \beta I(Z'; Z)]$$

Intractable, $X$ is hard to capture.



"Describe the image."

$X$

Base LMM (LLaVA)

High temperature

$d_n$ = "A bird with dark feather is facing the ... a ... eyes ... its

$d_2$ = "The bird ... great puffy eyes, which is its ...re. Its ...tly in

$d_1$ = "The bird ... small with a dark beak and spectacles... Its beak is long and curved,.. ... and the small eyes."

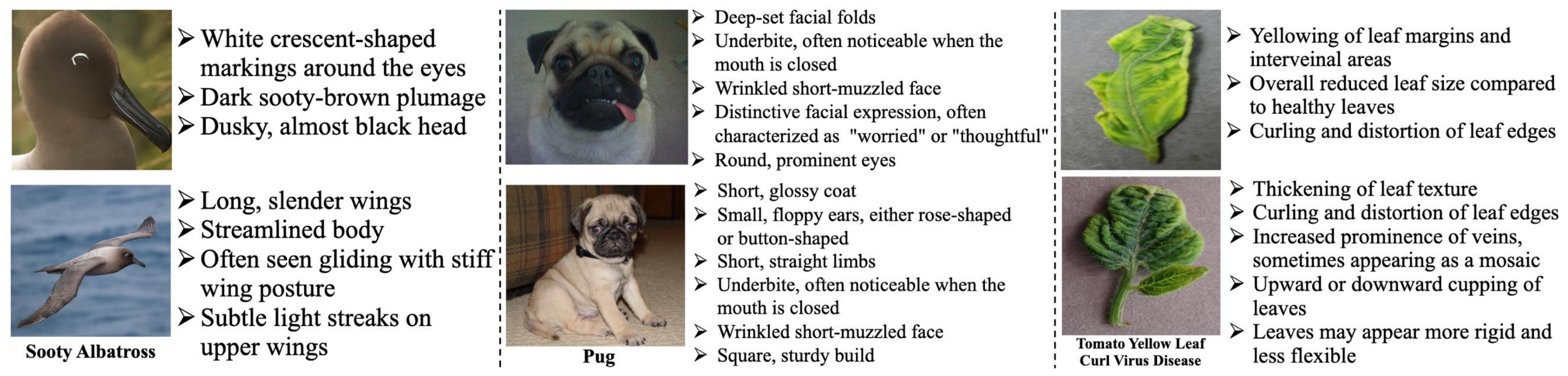
≈

## Quantitative Results

| Dataset | Method | Accuracy ↑ Per Iteration | | | | Explanation Quality | | | General Ability |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | EE ↑ | CS ↑ | FS ↓ | MMMU ↑ |
| CUB-200 | Base | 2.69 | – | – | – | 0.92 | 0.67 | 4.28 | 35.56 |
| | NL | 73.42 | 78.25 | 79.94 | 82.21 | 0.00 | – | – | 35.67 |
| | L+GE | 61.48 | 72.23 | 73.23 | 73.06 | 1.00 | 0.70 | 6.84 | 34.89 |
| | Ours | 80.24 | 83.76 | 84.69 | 85.02 | 1.00 | 0.82 | 6.53 | 35.00 |
| Stanford dogs | Base | 12.2 | – | – | – | 0.94 | 0.69 | 5.47 | 35.56 |
| | NL | 82.73 | 82.34 | 84.03 | 84.27 | 0.00 | – | – | 34.67 |
| | L+GE | 73.45 | 77.89 | 78.15 | 76.55 | 1.00 | 0.77 | 7.50 | 34.56 |
| | Ours | 85.29 | 86.75 | 86.86 | 86.91 | 1.00 | 0.86 | 7.41 | 34.56 |
| FGVC-A | Base | 3.00 | – | – | – | 0.97 | 0.42 | 5.39 | 35.56 |
| | NL | 83.47 | 87.28 | 87.82 | 87.73 | 0.00 | – | – | 35.56 |
| | L+GE | 72.13 | 79.87 | 82.45 | 82.69 | 1.0 | 0.76 | 8.59 | 35.56 |
| | Ours | 88.78 | 90.91 | 91.42 | 91.99 | 1.0 | 0.79 | 7.00 | 37.33 |
| PLD | Base | 0.00 | – | – | – | 0.95 | – | – | 35.56 |
| | NL | 89.38 | 94.52 | 94.29 | 93.95 | 0.00 | – | – | 34.78 |
| | L+GE | 24.03 | 25.27 | 24.56 | 24.90 | 1.00 | 0.76 | 10.45 | 35.44 |
| | Ours | 75.96 | 92.80 | 96.59 | 97.16 | 1.00 | 0.86 | 9.01 | 35.22 |
| HAM10000 | Base | 1.62 | – | – | – | 0.98 | 0.63 | 3.93 | 35.56 |
| | NL | 77.28 | 80.75 | 82.49 | 81.71 | 0.00 | – | – | 35.33 |
| | L+GE | 7.47 | 8.83 | 9.35 | 8.45 | 1.00 | 0.94 | 9.68 | 35.22 |
| | Ours | 79.37 | 82.29 | 83.69 | 85.06 | 1.00 | 0.87 | 7.43 | 35.89 |
| Chest X-ray Pneumonia (LLaVA-Med) | Base | 62.50 | – | – | – | 1.00 | 0.24 | 3.49 | – |
| | NL | 85.58 | 89.10 | 85.90 | 89.58 | 0.00 | – | – | – |
| | L+GE | 62.50 | 62.50 | 62.98 | 62.66 | 1.00 | 0.79 | 7.19 | – |
| | Ours | 97.60 | 96.31 | 99.04 | 98.72 | 1.00 | 0.87 | 8.25 | – |

*Base: original model; NL: only train with labels; L+GE: train with labels and general explanations*

## Qualitative Results



- White crescent-shaped markings around the eyes
- Dark sooty-brown plumage
- Dusky, almost black head

- Long, slender wings
- Streamlined body
- Often seen gliding with stiff wing posture
- Subtle light streaks on upper wings

**Sooty Albatross**

- Deep-set facial folds
- Underbite, often noticeable when the mouth is closed
- Wrinkled short-muzzled face
- Distinctive facial expression, often characterized as "worried" or "thoughtful"
- Round, prominent eyes

- Short, glossy coat
- Small, floppy ears, either rose-shaped or button-shaped
- Short, straight limbs
- Underbite, often noticeable when the mouth is closed
- Wrinkled short-muzzled face
- Square, sturdy build

**Pug**

- Yellowing of leaf margins and interveinal areas
- Overall reduced leaf size compared to healthy leaves
- Curling and distortion of leaf edges

- Thickening of leaf texture
- Curling and distortion of leaf edges
- Increased prominence of veins, sometimes appearing as a mosaic
- Upward or downward cupping of leaves
- Leaves may appear more rigid and less flexible

**Tomato Yellow Leaf Curl Virus Disease**

Our method can pick up **accurate and different** image-level visual concepts for **objects with the same label**.

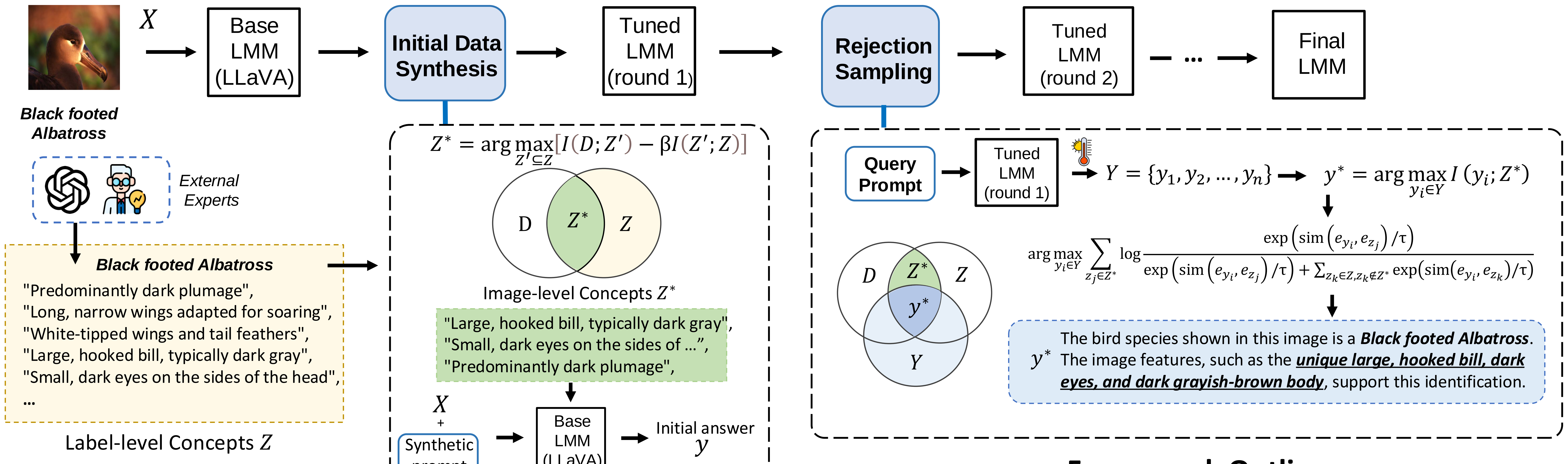
| | |
|---|---|
| User: | What is the name of this leaf disease? Can you give me your reason? |
| LMM: | The name of this leaf disease is **Bell Pepper Leaf with Bacterial Spot Disease**. The reason for this name is that the leaf has *brown or black necrotic spots*, which are characteristic of bacterial infections. These spots can coalesce, forming *larger necrotic areas*, and the leaf may exhibit *chlorosis or yellowing* of surrounding leaf tissue. The spots are often located on *the underside of the leaf*, and the leaf may become *crinkled or deformed* near affected areas. |
| User: | OK then, what chemical should I use to treat this disease? |
| LMM: | It is recommended to use a **copper-based fungicide** to treat this disease. |

| | |
|---|---|
| User: | Identify this bird. What features led to your conclusion? |
| LMM: | The bird is a **male Northern Cardinal** with a *bright red plumage* on its head, throat, and chest. The presence of *a crest on its head* and a *black "mask" around its eyes* further supports the identification of this bird. |
| User: | Where can I find this bird? |
| LMM: | You can find this bird in **the United States**, specifically in the **eastern and central parts** of the country. The Northern Cardinal is a common resident bird in areas with dense woodlands, shrubs, and gardens. |

Our tuned LMMs can provide **verifiable details**, enhancing **user trust**.

## Future Applications:

- Medical image diagnosis
- Detecting plant diseases
- Monitoring wildlife species…



$X$

Base LMM (LLaVA)

**Initial Data Synthesis**

Tuned LMM (round 1)

**Rejection Sampling**

Tuned LMM (round 2)

...

Final LMM

***Black footed Albatross***

*External Experts*

***Black footed Albatross***

"Predominantly dark plumage",
"Long, narrow wings adapted for soaring",
"White-tipped wings and tail feathers",
"Large, hooked bill, typically dark gray",
"Small, dark eyes on the sides of the head",
...

Label-level Concepts $Z$

$$Z^* = \arg\max_{Z' \subseteq Z} [I(D; Z') - \beta I(Z'; Z)]$$

Image-level Concepts $Z^*$

"Large, hooked bill, typically dark gray",
"Small, dark eyes on the sides of ...",
"Predominantly dark plumage",

$X$ + Synthetic prompt → Base LMM (LLaVA) → Initial answer $y$

**Query Prompt** → Tuned LMM (round 1) → $Y = \{y_1, y_2, \ldots, y_n\}$ → $y^* = \arg\max_{y_i \in Y} I(y_i; Z^*)$

$$\arg\max_{y_i \in Y} \sum_{z_j \in Z^*} \log \frac{\exp\left(\mathrm{sim}\left(e_{y_i}, e_{z_j}\right)/\tau\right)}{\exp\left(\mathrm{sim}\left(e_{y_i}, e_{z_j}\right)/\tau\right) + \sum_{z_k \in Z, z_k \notin Z^*} \exp\left(\mathrm{sim}\left(e_{y_i}, e_{z_k}\right)/\tau\right)}$$

$y^*$ The bird species shown in this image is a ***Black footed Albatross***. The image features, such as the ***unique large, hooked bill, dark eyes, and dark grayish-brown body***, support this identification.

## Framework Outline