

# ENHANCING COGNITION AND EXPLAINABILITY OF MULTIMODAL FOUNDATION MODELS WITH SELF-SYNTHESIZED DATA

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Large multimodal models (LMMs) have shown impressive capabilities in a wide range of visual tasks. However, they often struggle with identifying domain-specific objectives and fail to explain their predictions reasonably. To address the above challenge, we propose a novel iterative visual fine-tuning framework to improve the effectiveness and explainability of LMMs using self-synthesized data. Specifically, visual fine-tuning requires images, queries, and target answers. Our approach begins by synthesizing interpretable answers that include human-verifiable visual features. These features are based on expert-defined concepts, carefully selected based on their alignment with the image content. After each round of fine-tuning, we apply a reward model-free filtering mechanism to select the highest-quality interpretable answers for the next round of tuning. This iterative process of data synthesis and fine-tuning progressively improves the model’s ability to generate accurate and reasonable explanations. Experimental results demonstrate the effectiveness of our method in improving both the accuracy and explainability of specialized visual classification tasks.

## 1 INTRODUCTION

Large Multimodal Models (LMMs) have recently demonstrated remarkable performance in visual descriptive tasks, excelling in applications such as image captioning and visual question answering (Liu et al., 2024a; Alayrac et al., 2022; Li et al., 2022; Liu et al., 2024b). However, recent research (Kim & Ji, 2024; Geigle et al., 2024) reveals a significant drop in their effectiveness when applied to knowledge-intensive, domain-specific visual classification tasks. We observe that even advanced models like LLaVA-1.5 (Liu et al., 2024a) can only achieve a mere 12.2% classification accuracy in the Stanford Dogs dataset (Khosla et al., 2011). Further investigation reveals that current LMMs struggle to utilize key visual features for reasoning in classification and fail to provide justifiable explanations for their predictions, as illustrated by the examples in Figure 1.

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



**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

Figure 1: LLaVA-1.5 struggles to utilize key visual features in images for reasoning and explaining predictions in classification tasks.

The core issue stems from insufficient domain-specific alignment, as the model struggles to recognize key visual features and link them to the correct labels. To address this problem, we propose enhancing the LMM’s domain-specific cognition and explainability through *fine-tuning* (Touvron et al., 2023; Gu et al., 2021). However, this approach is hindered by a lack of data, as creating high-quality, feature-level image annotations is both complex and resource-intensive (Liu et al., 2024c). While labeling images by category and identifying key features for each class independent of the image is manageable, annotating the *specific visual characteristics per image* requires an extensive workload. Moreover, this level of detailed annotation goes beyond the capacity of standard annotators and current LMMs (Chen et al., 2024a), making it impractical to scale.

The biggest challenge now is synthesizing high-quality training data, specifically interpretable target answers. Given a dataset with images and labels, a naive approach would be to use only labels as target answers. However, training on such data may result in shortcut learning, where models pick up spurious correlations instead of truly understanding key visual features (Geirhos et al., 2020). While including general label-associated features as target answers might seem beneficial, it often results in overly generic explanations that lack the *image-specific details* necessary for accurate interpretation. We illustrate these shortcomings with examples in Figure 2.

To tackle this challenge, we propose a framework that allows LMMs to self-synthesize interpretable answers without relying on explicit image-specific annotations. For a given image, we first leverage the LMM’s captioning ability to generate descriptions, which are then used to identify visual features relevant to that specific image. Each description may only cover part of the key features, but by collecting a large set of descriptions from the LMM, we can approximate the true distribution of the image’s features, reducing the incompleteness in individual descriptions. We provide a formal justification for this approach in Section 3.2. Moreover, to ensure precise identification, we apply an information bottleneck technique to select the most relevant features. Once the image-specific concepts are identified, they are rewritten into interpretable answers.

For the training procedure, we also design an iterative fine-tuning approach to further improve performance over a one-shot training scheme. We begin by extracting image-level features and transforming them into interpretable answers, which, together with the corresponding images and queries, form the initial training dataset. Fine-tuning on this data results in an updated model that can generate more accurate answers. The updated model is then used to repeatedly generate answers, with the best one selected for the next round of fine-tuning. This self-boasting process progressively improves the LMM’s ability to deliver reliable explanations.

In summary, our contributions are threefold: (1) We propose a novel framework that improves LMMs’ interpretable visual classification abilities without requiring extensive manual labeling, (2) We introduce an information-theoretic approach to select interpretable visual concepts for each image and a reward model-free filtering method to ensure high-quality data selection from synthesized outputs, and (3) We develop an iterative process of data synthesis and model fine-tuning to progressively enhance LMMs’ cognitive abilities and explainability.

## 2 PRELIMINARY

**Problem Statement.** We aim at developing a Large Multimodal Model (LMM) for explainable visual classification. Let  $f_\theta$  denote the LMM model,  $X$  be the input image, and  $q$  be the query prompt. The model’s answer is denoted as  $\hat{y} = f_\theta(X, q)$ , where  $\hat{y}$  is expected to correctly predict the label and explain its prediction by using the visual features observed in the image. To build such a model, a straightforward strategy is to *fine-tune* the LMM with a training dataset that contains the ground-truth answer for each input image. However, most available datasets  $\mathcal{D} = \{(X_i, c_i)\}_{i=1}^N$  only consist of images  $X_i$  and class labels  $c_i$  (Khosla et al., 2011). To solve this, we propose a data synthesis approach that transforms the raw dataset into an augmented dataset  $\mathcal{D}^* = \{(X_i, q_i, y_i)\}_{i=1}^N$  with queries and explainable answers. While generating queries  $q_i$  is straightforward, the challenge lies in synthesizing explainable answers  $y_i$ , which must include detailed visual features  $Z_i^*$  that humans can identify and use for explanations. Therefore, the key problem is developing a method to automatically annotate visual features  $Z_i^*$  for each image  $X_i$ , given their labels  $c_i$ .

**Visual Fine-Tuning.** Visual fine-tuning adapts a pre-trained LMM to understand specific visual information by training on image-text pairs. Typically, an LMM  $f_\theta$  consists of a vision encoder that extracts visual embeddings from an input image  $X$ , a projector that maps these embeddings into the language embedding space, and a language model that processes the combined visual and



Figure 2: Examples of synthesized answers for query  $Q$ . Training with the first two types leads to shortcut learning or overgeneralization.

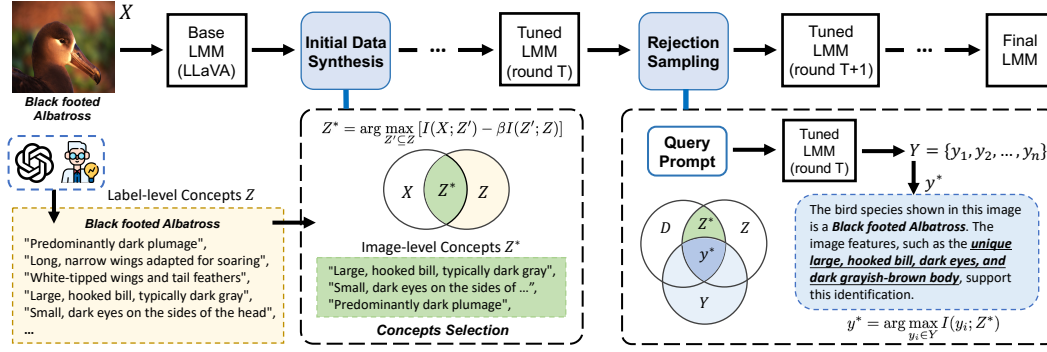


Figure 3: Our framework: An iterative approach of data synthesis and model fine-tuning.

textual information (Liu et al., 2024b;a; Lin et al., 2024). Formally, given a round of conversation containing an image  $X$ , a question  $q$  and an answer  $y$ , the model is trained to maximize the likelihood of generating the target answer:  $\mathcal{L}(\theta) = \sum_{i=1}^{|y|} \log p_{\theta}(y^i | X, q, y^{<i})$ . Here,  $y^i$  is the  $i$ -th token of the answer  $y$ , and  $|y|$  is the length of the answer. The fine-tuning process typically optimizes performance by freezing the pre-trained visual encoder to preserve learned visual representations while updating the projector and language model parameters to improve language understanding for visual inputs (Liu et al., 2024b; Lin et al., 2024). Recognizing the critical role of fine-tuning data quality in model performance, our research proposes synthesizing high-quality conversation data to improve performance.

### 3 METHODOLOGY

#### 3.1 OVERVIEW

Our approach leverages synthesized data for visual fine-tuning to enhance both the cognitive ability and explainability of LMMs, as illustrated in Figure 3. There are two major steps:

**1. Image-level Visual Concept Selection:** Given an image and its label, our first task is to extract a set of *image-specific concepts* that explain the connection between the image and label. We propose a selection method that identifies a subset of concepts most relevant to the image content while ensuring the subset is concise. Using these concepts, we prompt the LMM to rewrite them into textual answers that explains the label. These answers are used for the first round of fine-tuning.

**2. Reward Model-Free Rejection Sampling:** After the initial fine-tuning, the model can generate synthetic answers, which can be used for next-round fine-tuning, but their quality still varies. To filter out low-quality explanations and only select the best quality one as training data, we use previously selected concepts as filtering criteria, quantifying alignment between explanations and the concepts. The synthetic answer with the best-aligned explanation is then selected and combined with the corresponding image and query, forms a new *data pair* for subsequent rounds of fine-tuning.

#### 3.2 STEP 1: IMAGE-LEVEL VISUAL CONCEPT SELECTION

In this step, our goal is to identify the visual concepts present in a given image. Let  $X$  represent the true content of the image and  $c$  its class label. Each label class  $c$  is associated with a set of expert-defined visual concepts  $Z$ , which can be obtained by consulting domain experts or using a large language model. However, not all concepts in  $Z$  will necessarily be present in the image  $X$ . Therefore, we aim to select a subset of concepts,  $Z^* \subseteq Z$ , that are observable in the image  $X$ .

To achieve this, we propose concept selection by leveraging the Information Bottleneck (IB) principle, which seeks a compressed representation that preserves maximal information about another variable (Tishby et al., 2000). In our context, we need to find  $Z^*$  that maximizes its mutual information with the image content  $X$ , i.e.,  $I(X; Z^*)$ , while minimizing the redundancy by penalizing the mutual information between the selected concepts  $Z^*$  and the full concept set  $Z$ , i.e.,  $I(Z^*; Z)$ .

Formally, we define the optimization problem:

$$Z^* = \arg \max_{Z' \subseteq Z} [I(X; Z') - \beta I(Z'; Z)], \quad (1)$$

where  $\beta$  is a Lagrange multiplier that balances relevance and redundancy. However, directly computing  $I(X; Z')$  is intractable due to the high dimensionality and complexity of the image space. To address this, we introduce an intermediate variable: a set of image descriptions  $D = \{d_1, d_2, \dots, d_n\}$ , generated by prompting an LMM (i.e., the base LLaVA-1.5 (Liu et al., 2024a)), with instructions like “Please describe the image.” Each description  $d_i$  attempts to capture some aspects of the image content  $X$ . By increasing the number of collected descriptions  $n$  with different prompts, we aim to approximate the true distribution of  $X$ .

This approach is analogous to assembling pieces of a puzzle: Each description provides partial information about the image, and collectively, they form a more complete representation. Similarly, multiple potentially partial descriptions generated from different prompts can collectively approximate the true image content. Under this intuition, we formalize the approximation in Theorem 1, with proof provided in the appendix.

**Theorem 1.** *Let  $X$  be the true image content with label  $c$  and  $D = \{d_1, d_2, \dots, d_n\}$  be independent and identically distributed (i.i.d.) samples from  $P(D|X)$ . Let  $Z$  be an expert-defined concept list about label  $c$ . Under the assumptions of conditional independence and convergence (Assumptions 1 and 2), as  $n \rightarrow \infty$ , the mutual information  $I(D; Z)$  converges to  $I(X; Z)$ :*

$$\lim_{n \rightarrow \infty} I(D; Z) = I(X; Z).$$

In practice, we cannot sample an infinite number of descriptions, and LMMs may generate inconsistent or contradictory descriptions due to hallucinations or uncertainties. To mitigate this, we design high-quality prompts and encourage diverse responses to improve the reliability of the generated descriptions. By doing so, we assume that most of the descriptions will accurately reflect the image content. This assumption is further validated through experiments in Section 4.4. From another perspective, if a feature is not consistently covered by image descriptions, it means that the model is not certain about its presence, which will naturally result in lower MI scores for the associated concepts, reducing their likelihood of being selected. Using the set of descriptions  $D$ , we reformulate the IB objective in Equation. 1 as:

$$Z^* = \arg \max_{Z' \subseteq Z} [I(D; Z') - \beta I(Z'; Z)]. \quad (2)$$

However, computing mutual information in high-dimensional space directly remains challenging. Therefore, we employ the InfoNCE loss (Oord et al., 2018) as a lower-bound estimator of mutual information. For each concept  $z_j \in Z$ , we calculate an InfoNCE score  $s_j$ :

$$s_j = \sum_{d_i \in D} \log \frac{\exp(\text{sim}(\mathbf{e}_{d_i}, \mathbf{e}_{z_j}) / \tau)}{\exp(\text{sim}(\mathbf{e}_{d_i}, \mathbf{e}_{z_j}) / \tau) + \sum_{d_k \in \bar{D}} \exp(\text{sim}(\mathbf{e}_{d_k}, \mathbf{e}_{z_j}) / \tau)}, \quad (3)$$

where  $\text{sim}(\cdot, \cdot)$  denotes the cosine similarity between embeddings,  $\mathbf{e}_{d_i}$  and  $\mathbf{e}_{z_j}$  are the embeddings of description  $d_i$  and concept  $z_j$ , respectively, and  $\tau$  is a temperature parameter.  $\bar{D}$  are the descriptions for other images as negative samples. We can easily obtain the above embeddings through an off-the-shelf language embedding model (e.g., BERT (Devlin, 2018)). Next, we approximate  $I(D; Z^*)$  as the sum of InfoNCE scores for the selected concepts:  $I(D; Z^*) \approx \sum_{z_j \in Z^*} s_j$ . To minimize  $I(Z^*; Z)$ , we selectively include concepts to reduce redundancy. Given that  $Z^*$  is a subset of  $Z$ , we have:  $H(Z^*) = I(Z^*; Z) = -\sum_{z_i \in Z^*} p(z_i) \log p(z_i)$ , where  $H(Z^*)$  is the entropy of  $Z^*$ . A selected subset  $Z^*$  with a smaller size and higher probabilities for its  $z_i$  elements will result in lower entropy. Combining these approximations, our selection criterion for  $Z^*$  becomes:

$$Z^* = \{z_j \in Z \mid s_j > \mu + \hat{\beta}\sigma\}, \quad (4)$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation of the InfoNCE scores  $\{s_j\}_{j=1}^{|Z|}$ , respectively. The parameter  $\hat{\beta}$  controls the trade-off between including relevant concepts and avoiding redundancy. The selected concepts  $Z^*$  are not only relevant to the image but also capture the most informative features unique to the class label, providing strong evidence for the classification result and serving as reasonable explanations. Once we obtain  $Z^*$ , we can **generate an explainable answer** for a classification query on the image  $X$ . Specifically, we prompt the base LMM with these concepts to produce a coherent explanation. The prompts used are detailed in the appendix. After gathering the image and query-answer pairs, we can use them to **fine-tune our LMMs**.



### 3.3 STEP 2: REWARD MODEL-FREE REJECTION SAMPLING

After the initial round of fine-tuning with explainable visual query-answer pairs, the LMMs has significantly improved their ability to generate reasonable explanations. This improvement allows us to leverage the current fine-tuned model to generate new data for subsequent training rounds. However, the quality of newly generated data can vary considerably in terms of label accuracy and explanation quality. Training on low-quality data could lead to performance degradation. To address this issue, we propose a rejection sampling technique that filters out low-quality outputs.

Rejection sampling, also known as Best-of-N, is an inference-time strategy that generates multiple candidates and selects the best one for further tuning (Touvron et al., 2023; Stiennon et al., 2020). In our work, we adapt this idea for visual fine-tuning to iteratively improve LMM performance. Our proposed rejection sampling process begins by generating a series of answer candidates, using the fine-tuned model  $f_{\theta}^T$  from the last round  $T$ , and then identify the best answer from these candidates. For language-only conversations, this selection is typically performed by a reward model, which assigns higher rewards to answers aligning with desired criteria (Touvron et al., 2023). The answer with the highest reward is then selected and used for the next round ( $T + 1$ ) of training. However, in the visual domain, finding a reliable reward model remains challenging, as noted by (Chen et al., 2024a). To address this, we propose a *reward model-free* data filtering method to select the highest quality interpretable answers.

Specifically, we leverage the selected concept set  $Z^*$  from Section 3.2 as a reference to evaluate explanation quality. Our aim is to select the answer that best aligns with these relevant concepts. Formally, let  $Y = \{y_1, y_2, \dots, y_m\}$  represent the set of answers generated by the fine-tuned model for a given image. These answers are obtained by prompting the model with questions like “What is the  $\{item\}$  in this image? Please provide your reasoning.” The “*item*” here is set to be an coarse-level label, like *bird*, *airplane*. Our goal is to select the answer  $y^* \in Y$  that maximizes the mutual information with  $Z^*$ :

$$y^* = \arg \max_{y_i \in Y} I(y_i; Z^*). \quad (5)$$

We approximate  $I(y_i; Z^*)$  using the InfoNCE score:

$$s'_i = \sum_{z_j \in Z^*} \log \frac{\exp(\text{sim}(\mathbf{e}_{y_i}, \mathbf{e}_{z_j}) / \tau)}{\exp(\text{sim}(\mathbf{e}_{y_i}, \mathbf{e}_{z_j}) / \tau) + \sum_{z_k \in Z, z_k \notin Z^*} \exp(\text{sim}(\mathbf{e}_{y_i}, \mathbf{e}_{z_k}) / \tau)}, \quad (6)$$

where  $\mathbf{e}_{y_i}$  is the embedding of answer  $y_i$ . We select the answer with the highest score:  $y^* = \arg \max_{y_i \in Y} s'_i$ . The InfoNCE score provides a quantitative measure of how well the generated answer aligns with the relevant concepts. Therefore, our method eliminates the need for a separate reward model, which is particularly beneficial given the lack of reliable reward models for multi-modal data (Chen et al., 2024a). Additionally, we add another empirical constraint: the selected answer should contain the correct label  $c$ ; otherwise, it will be discarded.

In summary, our framework, outlined in Algorithm 1, enhances the model’s fine-grained classification capabilities by iteratively fine-tuning on diverse, high-quality, synthesized visual classification query-answer pairs. Step 1 identifies the most informative concepts in the image, while Step 2 selects the explanations that best align with these concepts. This two-step approach improves both the accuracy and interpretability of the model’s predictions, enabling it to perform more effectively in complex visual classification tasks.

### 3.4 THEORETICAL JUSTIFICATION

Our approach is grounded in information theory, specifically the maximization of mutual information among the image  $X$ , the expert-defined concepts  $Z$  for the whole class, and the generated answer  $Y$  with explanations. Theoretically, the problem of finding the best  $Y^*$  is equivalent to maximizing the three-way mutual information:

$$\max_{Y^*} I(X; Y; Z), \quad (7)$$

where  $Y^*$  captures the maximum amount of information from both the image and the expert-defined concepts. However, directly optimizing this objective is intractable due to the high dimensionality of  $X$  and  $Y$  (Poole et al., 2019). To make the problem manageable, we decompose it into two subproblems corresponding to our two-step method.

In Step 1, we select a subset of concepts  $Z^* \subseteq Z$  that maximizes  $I(X; Z^*)$ , capturing the most relevant concepts for the image. This aligns with the Information Bottleneck principle, where  $Z^*$  serves as a distilled subset of  $Z$  that retains maximal information about  $X$ . In Step 2, we select the generated explanation  $Y^*$  that maximizes  $I(Y^*; Z^*)$ , ensuring that the generated explanation closely aligns with the selected concepts. With this decomposition, we establish a lower bound on the three-way mutual information (proof provided in appendix).

**Theorem 2.** Let  $X$ ,  $Y$ , and  $Z$  be discrete random variables. Define:

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

$$Y^* = \arg \max_{Y' \subseteq Y} I(Y'; Z^*)$$

Then, the following inequality holds:

$$I(X; Y; Z) \geq I(Y^*; Z^*) + I(X; Z^*) - I(Z^*; Z).$$

This theoretical foundation justifies our two-step optimization process, ensuring that our method effectively captures the essential information shared among the image, concepts, and explanations.

## 4 EXPERIMENTS

We conduct experiments to address the following questions. **Q1:** How effectively does our method improve performance on fine-grained visual classification tasks? **Q2:** How does our method enhance the explainability of the fine-tuned model? **Q3:** How does our concept selection strategy compare to baseline methods? **Q4:** Does our training scheme mitigate shortcut learning? **Q5:** How usable is our trained model?

### 4.1 EXPERIMENTS SETTINGS

**Implementation Details:** We evaluated our approach on a variety of datasets, including fine-grained classification datasets (CUB-200 (Wah et al., 2011), Stanford Dogs (Khosla et al., 2011), FGVC-Aircraft (Maji et al., 2013)), medical datasets (HAM10000 (Tschandl et al., 2018), Chest X-Ray for Pneumonia (Kermany et al., 2018)), and the Plant Disease Dataset (PLD) (Vipooool, 2020). By testing on datasets from diverse domains, we demonstrate the versatility of our proposed framework with self-synthesized data. Our experiments used LLava-1.5-7B (Liu et al., 2024a) as the base LMM, and for the Chest X-Ray datasets, we employed its medical version (Li et al., 2024b). Additionally, E5 (Wang et al., 2022) served as the embedding model. We fine-tuned the LMMs using LoRA, focusing on all linear layers. Training was conducted on 8 H100 GPUs, utilizing HuggingFace (Wolf, 2019) and DeepSpeed frameworks for efficient distributed training and optimization. Further details are provided in the appendix.

**Baselines:** Given the novel problem of generating interpretable answers without image-specific annotations, we design the following baselines for comparisons: (1) **Base LMM:** Assesses the base multimodal model’s performance in the zero-shot setting. (2) **Naive Label Fine-tuning (NL):** Fine-tunes the base model using only class labels and a simple template (e.g., “This is a picture of {label}”). (3) **Label with General Explanations (L+GE):** We adopt the data synthesis approach

---

### Algorithm 1 Iterative synthesis and fine-tuning.

---

**Require:** Dataset  $\mathcal{D} = \{(X_i, c_i)\}_{i=1}^N$ , Concept Sets  $\mathcal{Z} = \{Z_i\}_{i=1}^{|C|}$  of every label, Pre-trained LMM  $f_\theta^0$ , Synthesized query set  $Q = \{q_i\}_{i=1}^N$ , Descriptive query  $q_d$

**Ensure:** Fine-tuned LMM  $f_\theta$  for accurate, interpretable explanations

- 1: Initialize training set  $\mathcal{D}^* \leftarrow \emptyset$
- 2: **for** each  $(X_i, c_i) \in \mathcal{D}$  **do**
- 3:    $Z \leftarrow \mathcal{Z}[c_i]$
- 4:    $D_i \leftarrow f_\theta^0(X_i, q_d)$
- 5:   **for**  $z_j \in Z$  **do**
- 6:      $s_{ij} \leftarrow \text{InfoNCE}(D_i, z_j)$
- 7:   **end for**
- 8:    $\mu_i \leftarrow \text{mean}(\{s_{ij}\}), \sigma_i \leftarrow \text{std}(\{s_{ij}\})$
- 9:    $Z_i^* \leftarrow \{z_j \in Z \mid s_{ij} > \mu_i + \beta \sigma_i\}$
- 10:    $y_i \leftarrow \text{assemble}(Z_i^*)$
- 11:    $\mathcal{D}^* \leftarrow \mathcal{D}^* \cup \{(X_i, q_i, y_i)\}$
- 12: **end for**
- 13: **FT:** Fine-tune  $f_\theta^0$  on  $\mathcal{D}^*$  to obtain  $f_\theta^1$
- 14: **repeat**
- 15:   **for** each  $(X_i, -) \in \mathcal{D}$  **do**
- 16:      $Y_i \leftarrow f_\theta^T(X_i, q_i)$
- 17:     **for**  $y_{ij} \in Y_i$  **do**
- 18:        $s'_{ij} \leftarrow \text{InfoNCE}(Z_i^*, y_{ij})$
- 19:     **end for**
- 20:      $y_i^* \leftarrow \arg \max_{y_{ij}} s'_{ij}$
- 21:      $\mathcal{D}^* \leftarrow \mathcal{D}^* \cup \{(X_i, q_i, y_i^*)\}$
- 22:   **end for**
- 23:   **Iterative FT:** Fine-tune  $f_\theta^T$  on  $\mathcal{D}^*$
- 24: **until** Max iterations  $T^m$
- 25: **return**  $f_\theta^m$

---

Table 1: Our method achieves superior accuracy and explanation quality across diverse datasets.

Dataset	Method	Accuracy $\uparrow$ Per Iteration				Explanation Quality			General Ability
		1	2	3	4	EE $\uparrow$	CS $\uparrow$	FS $\downarrow$	MMMU $\uparrow$
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*

introduced in the LLaVA paper (Liu et al., 2024b) and (Kim & Ji, 2024), where training data is generated using a language-only model (e.g., GPT-4), based on class labels and their corresponding label-level knowledge. More details, including the prompts, are in the appendix.

#### 4.2 RQ1: TRAINING ON OUR SYNTHESIZED DATA IMPROVES CLASSIFICATION

We evaluated our model’s classification capabilities using a multi-round, progressive rejection sampling training process comprising four iterations, each with two epochs. Classification accuracy was measured after each iteration, with success defined as the presence of the ground truth label in the model’s response (Kim & Ji, 2024). All trainable baselines were trained and evaluated under identical settings to ours for a fair comparison. More details on training are provided in the appendix.

Our proposed method achieves higher accuracy than both NL and L+GE baselines, as shown in Table 1. While the baselines, particularly NL, exhibit overfitting, our method improves accuracy with each iteration, demonstrating its effectiveness in fine-grained classification tasks. Our approach’s resilience to overfitting stems from the use of rejection sampling, which generates more diverse training dataset. This expanded data pool enhances the model’s generalization capabilities. Moreover, training with our dataset, which includes specific visual features, helps the model learn more detailed visual knowledge, contributing to further gains in classification performance. In contrast, the L+GE baseline struggles with certain datasets, notably HAM10000, likely due to irrelevant information in general explanations impeding effective learning. Moreover, we assessed the general ability of the models using the MMMU metric (Yue et al., 2024) evaluated by the LMMs-eval tool (Bo Li\* & Liu, 2024). The results indicate that all trained models maintain comparable general abilities to the base model, with negligible degradation. Note that the LMMs-eval tool does not support the LLaVA-Med model; therefore, the general ability metric is not reported. [More experiment results can be found in the appendix.](#)



Figure 4: Our generated answers contain detailed visual explanations.

#### 4.3 RQ2: OUR METHOD PROVIDES HIGH-QUALITY EXPLANATION

Assessing the quality of generated explanations is challenging, especially without case-by-case ground truth annotations (Ding et al., 2022; Schuff et al., 2022). Following Mohseni et al. (2021), we evaluate our explanation from three aspects: explanation existence, cognition level, and fluency.

**Explanation Existence (EE).** This metric assesses a model’s capability to generate explanations (Xu et al., 2023). Following Bills et al. (2023); Bricken et al. (2023), we employ an advanced proprietary LLM (GPT-4o) to determine whether each model-generated answer includes an explanation. Formally, for a set of generated answers  $Y = \{y_1, y_2, \dots, y_n\}$  queried by prompt like “What is the bird name? Provide your reason.”, EE is defined as:  $EE = \frac{1}{n} \sum_{i=1}^n e_i$ , where  $e_i = 1$  if the  $i^{\text{th}}$  answer  $y_i$  includes an explanation, and  $e_i = 0$  otherwise. We provide the evaluation prompt in the appendix.

**Results:** As shown in Table 1, our method achieves an EE of 1.00 across all datasets, indicating that it consistently produces explanations for its predictions. The base model can generate explanations at most times but fails to do so in some cases. In contrast, the NL baseline records an EE of 0.00, reflecting its inability to generate explanations due to training solely on class labels. The L+GE method also attains an EE of 1.00 but falls short in other quality metrics compared to our approach.

**Cognition Score (CS).** This metric evaluates the coherence and logical flow of generated explanations (Nourani et al., 2019; Fan et al., 2020). Following Liu et al. (2023); Bills et al. (2023); Lieberum et al. (2024), we employ an advanced proprietary LLM (GPT-4o) to assess the rational integrity of explanations by analyzing their alignment with expert knowledge. For an answer  $y_i$  containing a label and explanation, we first extract label-level concepts  $Z$  corresponding to the label. We then use an evaluation prompt to obtain a cognition score from LLM:  $cs_i = \text{LLM}(y_i, Z, \text{Eval.Prompt})$ . The Eval.Prompt is detailed in the appendix. Scores range from 0 to 1, with higher scores indicating better alignment with expert-predefined knowledge. We only assess correct answers. The total CS is calculated as:  $CS = \frac{1}{|Y_c|} \sum_{i=1}^{|Y_c|} cs_i$ , where  $Y_c$  is the set of correct answers.

**Results:** Our method generally achieves higher CS values than the baselines, demonstrating superior coherence and logical reasoning in the explanations. This improvement reflects the effectiveness of our approach in generating explanations that are more useful and trustworthy to human users. The reason behind our better cognition is that we select image-level concepts to build the training dataset, rather than using all label-level features. This ensures that our model focuses on relevant features present in the image and effectively connects the image content with expert-defined concepts, avoiding distractions from irrelevant features not shown in the image.

**Fluency Score (FS).** This metric evaluates the grammatical correctness and naturalness of the explanations (Liu et al., 2021). We use perplexity, a standard metric that measures how well a language model predicts the text, with lower values indicating better fluency.

**Results:** Our method demonstrates good fluency in the generated explanations. The base model receives the best fluency. This is because training on expert-level knowledge inevitably increases the complexity of explanations, which can lead to higher perplexity scores. However, our method generally achieves lower perplexity than the L+GE baseline, indicating that our explanations are more natural, highlighting the robustness of our approach.

In summary, our proposed method not only enhances classification accuracy but also consistently generates high-quality explanations that are coherent, logical, and fluent. We also show some quantitative results in our explanation in Figure 4. More results can be found in the appendix.



#### 4.4 RQ3: EVALUATION ON INTERPRETABLE CONCEPTS SELECTION

To evaluate the efficacy of our proposed model in selecting expert-defined features from images, we conducted a comprehensive study with human experts. Experts were invited to identify and annotate the top-4 relevant concepts for each image from a dataset of six bird species, providing ground truth annotations for our evaluation. Our investigation examined two primary aspects: the effect of varying the number of descriptions on concept selection performance, and a comparative analysis against several baseline methods. These baselines include GPT-4o for concept extraction, which relied on a carefully crafted prompt to guide the model in identifying and returning the four most probable concepts for each image. We applied the same approach to the LLaVA model for concept extraction. In contrast, for the CLIP model, concept selection was performed by identifying top-4 concepts with the highest CLIPScore (Hessel et al., 2021) relative to the target image.

We evaluated performance using precision, comparing model-selected concepts with human-annotated ground truth. Results demonstrated that our model outperforms strong baselines in the concept selection task. Its precision improves proportionally with the increased number of descriptions, peaking at 72.89% with 25 descriptions. In contrast, GPT-4o maintained a relatively high precision of 63.95%, while both LLaVA and CLIP showed weaker performances at approximately 55%. These findings highlight our model’s superior ability to leverage repeated sampled descriptions for more accurate concept selection. We show some qualitative results in Figure 6, with more in the appendix.

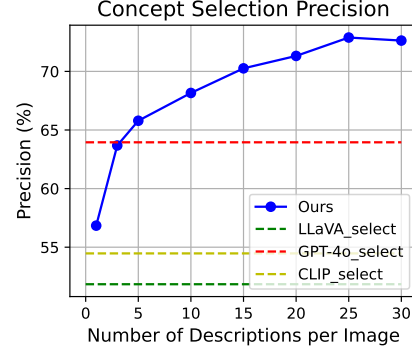


Figure 5: Our method demonstrates superior precision in concept selection compared to applying GPT-4o.

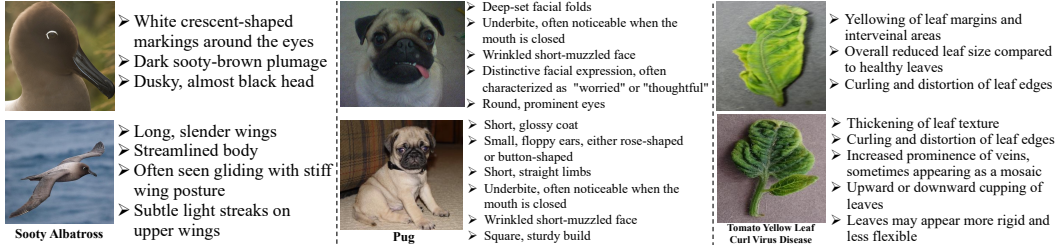


Figure 6: Different image-level visual concepts for objects with the same label.

#### 4.5 RQ4: VISUALIZATION OF VISUAL FOCUS OF LMMs

To interpret our model’s predictions, we visualize the outputs of LMMs using a gradient-based explanation method (Wu et al., 2023). This approach identifies the image regions most influential in generating the model’s answer by computing gradients of answer token probabilities with respect to image pixels. Figure 7 presents case examples demonstrating that training on our synthesized answers with cognitive explanations can effectively prevent the model from learning spurious features, whereas directly linking images to naive labels may result in shortcut learning.



#### 4.6 RQ5: EVALUATION ON USABILITY

Previous LMMs often struggle with specialized questions due to difficulty recognizing domain-specific objects. Our method enhances LMMs’ ability to identify specialized objects and provide verifiable visual details, as demonstrated in Table 10. In Case 1, the LMM accurately identifies a bell pepper leaf disease, describing specific visual cues such as brown necrotic spots and chlorosis. It then recommends an appropriate treatment. In Case 2, the LMM correctly identifies a male Northern Cardinal, noting distinctive features like its bright red plumage and black mask. These improvements enable LMMs to serve as **effective visual assistants** in specialized domains, offering users accurate and interpretable responses that can be verified against the provided images.



Figure 7: Visualization of heatmaps for different datasets. From left to right in each triplet: **Original Image, Heatmap of NL model, Heatmap of Ours.**

Table 2: LMMs fine-tuned with our method provide verifiable details, enhancing user trust.

	User:	What is the name of this leaf disease? Can you give me your reason?
	LMM:	The name of this leaf disease is <b>Bell Pepper Leaf with Bacterial Spot Disease</b> . The reason for this name is that the leaf has <i>brown or black necrotic spots</i> , which are characteristic of bacterial infections. These spots can coalesce, forming <i>larger necrotic areas</i> , and the leaf may exhibit <i>chlorosis or yellowing</i> of surrounding leaf tissue. The spots are often located on <i>the underside of the leaf</i> , and the leaf may become <i>crinkled or deformed</i> near affected areas.
	User:	OK then, what chemical should I use to treat this disease?
	LMM:	It is recommended to use a <b>copper-based fungicide</b> to treat this disease.
	User:	Identify this bird. What features led to your conclusion?
	LMM:	The bird is a <b>male Northern Cardinal</b> with a <i>bright red plumage</i> on its head, throat, and chest. The presence of a <i>crest on its head</i> and a <i>black "mask" around its eyes</i> further supports the identification of this bird.
	User:	Where can I find this bird?
	LMM:	You can find this bird in <b>the United States</b> , specifically in the <b>eastern and central parts</b> of the country. The Northern Cardinal is a common resident bird in areas with dense woodlands, shrubs, and gardens.

## 5 RELATED WORK: SELF-IMPROVEMENT AND DATA SYNTHESIS

In the field of LLMs, self-generated data has become a powerful tool for enhancing model performance (Liu et al., 2024d). Researchers have explored various techniques, such as rejection sampling (Touvron et al., 2023; Dubey et al., 2024), self-rewarding (Yuan et al., 2024), and self-play (Chen et al., 2024c), enabling models to improve using synthesized data. Recent studies (Snell et al., 2024; Brown et al., 2024) have proposed the “inference-time scaling law”, suggesting that increasing inference samples size boosts the likelihood of generating high-quality data. Data synthesis techniques have also been applied to LMMs, improving general vision tasks like visual question answering (Zhou et al., 2024a; Deng et al., 2024; Gao et al., 2023; Zhou et al., 2024b) and enhancing instruction-following capabilities (Luo et al., 2024). Our work extends this line of research by focusing on domain-specific visual classification to enable effective visual assistance in professional tasks. In contrast to existing methods, we address the unique challenges of data synthesis in specialized domains, thereby extending these techniques to support expert-driven applications.

## 6 CONCLUSION

In this work, we addressed LMMs’ limitations in domain-specific visual classification tasks by introducing a novel framework that enhances their cognition and explainability through iterative fine-tuning with self-synthesized interpretable answers. By applying the IB principle to select relevant visual concepts without extensive image-specific annotations, our approach significantly improved classification accuracy and explanation quality across various datasets. By enabling LMMs to generate accurate, interpretable explanations grounded in domain-specific visual features, our framework advances their applicability in specialized domains, paving the way for more reliable multimodal models in knowledge-intensive applications. Future work may explore more complex tasks than classification tasks, and refinements to enhance scalability and generalizability.

## REFERENCES

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. URL <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>. (Date accessed: 14.05. 2023), 2, 2023.
- Kaichen Zhang\* Fanyi Pu\* Xinrun Du Yuhao Dong Haotian Liu Yuanhan Zhang Ge Zhang Chunyuan Li Bo Li\*, Peiyuan Zhang\* and Ziwei Liu. Lmms-eval: Accelerating the development of large multimodal models, March 2024. URL <https://github.com/EvolvingLMs-Lab/lmms-eval>.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.
- Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024a.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024b.
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024c.
- Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *arXiv preprint arXiv:2405.19716*, 2024.
- Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Weiping Ding, Mohamed Abdel-Basset, Hossam Hawash, and Ahmed M Ali. Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*, 615:238–292, 2022.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Ming Fan, Wenying Wei, Xiaofei Xie, Yang Liu, Xiaohong Guan, and Ting Liu. Can we trust your explanations? sanity checks for interpreters in android malware analysis. *IEEE Transactions on Information Forensics and Security*, 16:838–853, 2020.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: a comprehensive evaluation benchmark for multimodal large language models. *corr abs/2306.13394 (2023)*, 2023.

- Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.
- Gregor Geigle, Radu Timofte, and Goran Glavaš. African or european swallow? benchmarking large vision-language models for fine-grained object classification. *arXiv preprint arXiv:2406.14496*, 2024.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.
- Daniel S Kermay, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.
- Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.
- Jeonghwan Kim and Heng Ji. Finer: Investigating and enhancing fine-grained visual concept recognition in large vision language models. *arXiv preprint arXiv:2402.16315*, 2024.
- Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024a.
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024b.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.
- Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.
- Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024.
- Haotian Liu, Chunyu Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.
- Haotian Liu, Chunyu Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.
- Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hananeh Hajishirzi. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*, 2021.



- Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. *arXiv preprint arXiv:2401.13837*, 2024c.
- Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*, 2024d.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.
- Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, et al. Mmevol: Empowering multimodal large language models with evol-instruct. *arXiv preprint arXiv:2409.05840*, 2024.
- S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.
- Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.
- Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pp. 97–105, 2019.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.
- Hendrik Schuff, Heike Adel, Peng Qi, and Ngoc Thang Vu. Challenges in explanation quality evaluation. *arXiv preprint arXiv:2210.07126*, 2022.
- Yucheng Shi, Yushun Dong, Qiaoyu Tan, Jundong Li, and Ninghao Liu. Gigamae: Generalizable graph masked autoencoder via collaborative latent space reconstruction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 2259–2269, 2023.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.
- Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.
- Vipooool. New plant diseases dataset. <https://www.kaggle.com/datasets/vipooool/new-plant-diseases-dataset>, 2020. Accessed: 2024-09-25.

- C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Cub-200. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.
- T Wolf. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. *arXiv preprint arXiv:2310.00492*, 2023.
- Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See-Kiong Ng, and Jiashi Feng. Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2023.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.
- Yuanhan Zhang Bo Li Songyang Zhang Wangbo Zhao Yike Yuan Jiaqi Wang Conghui He Ziwei Liu Kai Chen Dahua Lin Yuan Liu, Haodong Duan. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multi-modal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.
- Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024a.
- Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*, 2024b.

## A THEORETICAL PROOF ON THEOREM 1

**Theorem 3.** Let  $X$  be the true image content with label  $c$  and  $D = \{d_1, d_2, \dots, d_n\}$  be independent and identically distributed (i.i.d.) samples from  $P(D|X)$ . Let  $Z$  be an expert-defined concept list about label  $c$ . Under the assumptions of conditional independence and convergence (Assumptions 1 and 2), as  $n \rightarrow \infty$ , the mutual information  $I(D; Z)$  converges to  $I(X; Z)$ :

$$\lim_{n \rightarrow \infty} I(D; Z) = I(X; Z).$$

*Proof.* We proceed with the following steps:

### A.1 ASSUMPTIONS

**Assumption 1.** *Conditional Independence:*  $d_1, d_2, \dots, d_n$  are conditionally independent of  $Z$  given  $X$ .

**Assumption 2.** *Convergence of Mutual Information:* As  $n \rightarrow \infty$ , the mutual information between  $D$  and  $X$  converges to the entropy of  $X$ :

$$\lim_{n \rightarrow \infty} I(D; X) = H(X)$$

### A.2 PROOF STEPS

1. Mutual Information Expression: We start with the definition of mutual information:

$$I(D; Z) = H(D) - H(D|Z)$$

2. Expanding  $H(D|Z)$  Using the Chain Rule: Apply the chain rule for entropy:

$$H(D|Z) = H(D|Z, X) + H(X|Z) - H(X|D, Z)$$

Substituting this back into the mutual information expression:

$$I(D; Z) = H(D) - [H(D|Z, X) + H(X|Z) - H(X|D, Z)]$$

3. Applying Conditional Independence: By Assumption 1, we have:

$$H(D|Z, X) = H(D|X)$$

Therefore,

$$I(D; Z) = H(D) - H(D|X) - H(X|Z) + H(X|D, Z)$$

Recognizing  $I(D; X) = H(D) - H(D|X)$ , we have:

$$I(D; Z) = I(D; X) - H(X|Z) + H(X|D, Z)$$

4. Taking the Limit as  $n \rightarrow \infty$ : Apply the limit to both sides:

$$\lim_{n \rightarrow \infty} I(D; Z) = \lim_{n \rightarrow \infty} [I(D; X) - H(X|Z) + H(X|D, Z)]$$

By Assumption 2, we have:

$$\lim_{n \rightarrow \infty} I(D; X) = H(X)$$

For the term  $H(X|D, Z)$ , we argue that:

$$\lim_{n \rightarrow \infty} H(X|D, Z) = 0$$

This follows from Assumption 2, as it implies that  $D$  becomes a sufficient statistic for  $X$  as  $n \rightarrow \infty$ . Therefore, conditioning on  $Z$  does not add any information about  $X$  once we have  $D$ .

5. Conclusion: Substituting these limits into our equation:

$$\lim_{n \rightarrow \infty} I(D; Z) = H(X) - H(X|Z) + 0 = I(X; Z)$$

This demonstrates that as the number of sampled descriptions  $n$  increases indefinitely, the mutual information between the aggregated descriptions  $D$  and the human concepts  $Z$  converges to the mutual information between the true image content  $X$  and  $Z$ .  $\square$

### A.3 ADDITIONAL NOTES

#### A.3.1 CONVERGENCE OF $I(D; X)$ TO $H(X)$

The statement:

$$\lim_{n \rightarrow \infty} I(D; X) = H(X)$$

This is directly from Assumption 2. To understand its implications:

- Recall that mutual information is defined as:  $I(D; X) = H(X) - H(X|D)$
- For this equality to hold as  $n \rightarrow \infty$ , it must be true that:

$$\lim_{n \rightarrow \infty} H(X|D) = 0$$

- This means that as we gather more samples ( $D$ ), we eliminate all uncertainty about  $X$ .
- In other words, with infinite samples,  $D$  contains all information about  $X$ .

#### A.3.2 CONVERGENCE OF $H(X|D, Z)$ TO 0

The statement:

$$\lim_{n \rightarrow \infty} H(X|D, Z) = 0$$

This follows from the previous point. Here's the reasoning:

- We've established that as  $n \rightarrow \infty$ ,  $D$  contains all information about  $X$ .
- This means  $D$  becomes a *sufficient statistic* for  $X$ .
- A sufficient statistic contains all the information that the sample provides about the parameter (in this case,  $X$ ).
- Therefore, once we know  $D$ , knowing  $Z$  doesn't provide any additional information about  $X$ .
- Mathematically, this means:  $H(X|D, Z) = H(X|D)$
- But we know from the first part that  $\lim_{n \rightarrow \infty} H(X|D) = 0$
- Thus,  $\lim_{n \rightarrow \infty} H(X|D, Z) = 0$

## B THEORETICAL PROOF ON THEOREM 2

**Theorem 4.** Let  $X$ ,  $Y$ , and  $Z$  be discrete random variables. Define:

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

$$Y^* = \arg \max_{Y' \subseteq Y} I(Y'; Z^*)$$

Then, the following inequality holds:

$$I(X; Y; Z) \geq I(Y^*; Z^*) + I(X; Z^*) - I(Z^*; Z). \quad (8)$$

where  $I(\cdot; \cdot)$  denotes mutual information and  $H(\cdot)$  denotes entropy.

*Proof.* We prove this theorem using fundamental principles of information theory:

1. Recall the definition of multivariate mutual information:

$$I(X; Y; Z) = I(X; Y) - I(X; Y|Z) \quad (9)$$

2. By the chain rule of mutual information (Shi et al., 2023), we can rewrite this as:

$$I(X; Y; Z) = I(X; Z) + I(Y; Z) - I(X, Y; Z) \quad (10)$$



3. Consider our subsets  $Z^*$  and  $Y^*$ . By definition of mutual information and the data processing inequality:

$$I(Y^*; Z^*) \leq I(Y; Z^*) \quad (11)$$

$$I(X; Z^*) \leq I(X; Z) \quad (12)$$

4. Substituting these into our equation:

$$I(X; Y; Z) \geq I(X; Z^*) + I(Y^*; Z^*) - I(X, Y; Z) \quad (13)$$

5. For any random variables  $A$  and  $B$ :

$$I(A; B) \leq \min(H(A), H(B)) \quad (14)$$

Therefore:

$$I(X, Y; Z) \leq \min(H(X, Y), H(Z)) \leq I(Z^*; Z) \quad (15)$$

6. Applying this to our inequality:

$$I(X; Y; Z) \geq I(X; Z^*) + I(Y^*; Z^*) - I(Z^*; Z) \quad (16)$$

7. Therefore, we can conclude:

$$I(X; Y; Z) \geq I(Y^*; Z^*) + I(X; Z^*) - I(Z^*; Z) \quad (17)$$

□

## C PROMPTS FOR DIFFERENT TASKS

### EXAMPLE PROMPT FOR BIRD IMAGE DESCRIPTION IN CUB-200

<b>Prompt 1</b>	Focus solely on the bird shown in the image. Describe the bird’s appearance in detail, emphasizing its most prominent physical features. Avoid mentioning the background or other elements not related to the bird.
<b>Prompt 2</b>	Provide a focused analysis of the bird in this image, detailing its distinctive physical features. Concentrate exclusively on the bird and describe its appearance without referencing the surroundings or any extraneous details.
<b>Prompt 3</b>	Directly observe the bird depicted and offer a precise description of its visual attributes. Ensure your description is limited to the bird itself, detailing its primary features and omitting any unrelated background elements.

Table 3: Prompts for bird image analysis in CUB-200 dataset

### EXAMPLE REWRITE PROMPTS FOR DIFFERENT DATASETS

Dataset	Prompt
<b>cub-200 / stanford_dogs</b>	This is a picture of a {label} with the following visual features: {concepts_str}. Based on the information provided, please answer the following question. <b>Question:</b> '{query}'
<b>HAM10000</b>	This is a dermatoscopic image of {label} disease with the following visual features: {concepts_str}. Based on the information provided, please answer the following question. <b>Question:</b> '{query}'
<b>PLD / fgvc</b>	This is a picture of {label} with the following visual features: {concepts_str}. Based on the information provided, please answer the following question. <b>Question:</b> '{query}'
<b>chest-xray</b>	This is a chest-xray of {label} with the following visual features: {concepts_str}. Based on the information provided, please answer the following question. <b>Question:</b> '{query}'

Table 4: Answer rewrite prompts for different datasets.

### L+GE SYNTHESIZE PROMPT EXAMPLE

Dataset	Prompt
<b>L+GE synthesize prompt</b>	There is a picture of a {label}, which is known for the following characteristics: {concepts_str}. Act as if you can see the picture. Please answer the following question based on the above information. Make your answer concise. <b>Question:</b> '{query}' <b>Answer:</b>

Table 5: Prompt for answering questions based on image characteristics for general image datasets.

## EVALUATION PROMPTS FOR EXPLANATION EXISTENCE (EE) AND COGNITION SCORE (CS)

Prompt Type	Prompt
EE Prompt	Determine whether the following answer contains a valid explanation supporting its conclusion. Respond with only 'true' or 'false'.
	<b>Answer:</b> {answer}
CS Prompt	Contains an explanation?
	Evaluate the coherence and logical alignment of the following explanation with the provided concepts. Please note: the explanation does not need to fully encompass all concepts.
	Assign a consistency score between 0 and 1, where 1 indicates the explanation contains no irrelevant information to the listed concepts, and 0 indicates complete misalignment with entirely irrelevant information. Only give the score.
	<b>Concepts:</b> {concepts_formatted} <b>Explanation:</b> {explanation}
	<b>Consistency Score:</b>

Table 6: Prompts for Explanation Existence (EE) and Cognition Score (CS).

## CONCEPT EXTRACTION PROMPTS

To obtain label-level concepts, we used GPT-4o with prompts designed to elicit detailed visual features associated with each class label. An example prompt is:

Prompt Type	Prompt
Concept Extraction Prompt	Please provide a list of visual characteristics that are commonly associated with the bird species {Class Name}. Include features such as color patterns, shapes, and distinctive markings.

Table 7: Prompts for Concept Extraction.

## D EXPERIMENTS DETAILS

### D.1 TRAINING FRAMEWORK IMPLEMENTATION

Our experimental setup leverages state-of-the-art techniques in large model fine-tuning, utilizing the LLaVA-1.5-7B model as our foundation. We employ Low-Rank Adaptation (LoRA) for efficient parameter-efficient fine-tuning, along with advanced distributed training strategies to maximize computational efficiency.

#### D.1.1 MODEL INITIALIZATION AND LoRA CONFIGURATION

We initialize our model using the LLaVA-1.5-7B weights, sourced from the Hugging Face model hub (llava-hf/llava-1.5-7b-hf). To facilitate efficient fine-tuning, we implement LoRA using the following configuration:

- LoRA rank ( $r$ ) = 128
- LoRA alpha ( $\alpha$ ) = 256
- LoRA dropout rate = 0.1

This configuration targets all linear modules within the model, allowing for a balanced approach between adaptation capacity and computational efficiency.

#### D.1.2 TRAINING PROCESS AND OPTIMIZATION

Our training process is orchestrated using the SFTTrainer, which we configure for supervised fine-tuning. Key aspects of our setup include:

**Gradient Checkpointing:** We enable gradient checkpointing with `use_reentrant=True`, trading increased computation for reduced memory usage. This technique allows us to process larger batch sizes by not storing all activations in memory.

**DeepSpeed ZeRO Stage 3:** To fully utilize our multi-GPU setup, we implement DeepSpeed ZeRO Stage 3. This advanced parallelism strategy shards model parameters, gradients, and optimizer states across available GPUs, enabling the training of models larger than what a single GPU can accommodate.

#### D.1.3 MIXED PRECISION TRAINING

We utilize mixed precision training with BF16 (Brain Float 16) format.

#### D.1.4 GRADIENT ACCUMULATION

To optimize our effective batch size without increasing memory demands, we implement gradient accumulation. Our configuration uses:

- Gradient accumulation steps = 2
- Per-device train batch size = 16
- Number of GPUs = 1

This results in an effective batch size of 32 ( $2 * 16 * 1$ ), allowing us to simulate larger batch training while maintaining memory efficiency. Through these combined strategies, we aim to fully leverage the computational power of our GPUs.

### D.2 DATASET OVERVIEW AND EXPERIMENTAL RUNNING TIMES

We use the original labels and train/test split from the original datasets. Detailed train/test data statistics are reported in Table 8. Regarding the experiments running time, each iteration of rejection sampling and fine-tuning took approximately 2.5 hours on a single NVIDIA H100 GPU for datasets like CUB-200 and FGVC-Aircraft. A more detailed report can be found in Table 8.



Table 8: Summary of Selected Datasets

Dataset Name	Test / Training Cases	Label Categories	Running Time Per Iter
CUB-200-2011	5,794 / 5,994	200	143 minutes
FGVC-Aircraft	3,333 / 6,667	100	171 minutes
Stanford Dogs	8,580 / 12,000	120	310 minutes
Chest X-ray images	624 / 5,216	2	113 minutes
HAM1000	5,008 / 5,007	7	107 minutes
PLD	47,523 / 5,279	37	109 minutes

## E VISUALIZATION RESULTS

Table 9: Generated Answer Examples.










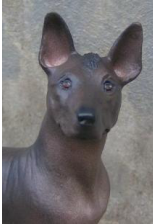
	User:	Based on the visual content, what is the specific name of this bird? Provide an explanation.
	LMM:	The specific name of this bird is the Yellow breasted Chat. This bird is known for its bright yellow throat and chest, olive-green back, wings, and tail, and short, rounded wings. It is a small bird that is commonly found in the Americas, particularly in the United States and Canada.
	User:	Based on the visual content, what is the specific name of this bird? Provide an explanation.
	LMM:	The specific name of this bird is the Blue Grosbeak. The Blue Grosbeak is a small, colorful bird with a deep blue overall plumage in males and a brown overall coloration with subtle blue tinges and faint streaking in females. The bird has a slightly glossy sheen to its blue feathers and is known for its melodious songs.
	User:	Based on the visual content, what is the specific name of this bird? Provide an explanation.
	LMM:	The specific name of this bird is the Orchard Oriole. The bird has a black head, back, throat, and upper chest, with a rich chestnut-brown underparts. It has a black tail and a black throat patch developing. The bird's wings are black with white wing bars, and it has a black bill. The Orchard Oriole is a small bird with a distinctive appearance, and it is known for its melodious songs.
	User:	Based on the visual content, what is the specific name of this bird? Provide an explanation.
	LMM:	The specific name of this bird is the Black throated Sparrow. This bird is known for its distinctive black throat patch and white stripe below its eyes. It is a small bird with a brown and white coloration, and it is often seen perched on branches or low vegetation.
	User:	Based on the visual content, what is the specific name of this bird? Provide an explanation.
	LMM:	The specific name of this bird is the Henslow Sparrow. This bird is known for its olive-colored head, short, pointed bill, and pale eye-ring. It is a small bird with a brown and white coloration, and it is often seen flitting low in dense grassy habitats.

Table 10: Selected Concepts Examples.

	Concepts: "Powerful, thick neck" "Broad and strong back" "Large and muscular build" "Well-sprung ribs"
	Concepts: "Pure white or cream-colored fur" "Well-feathered tail that blends with the body fur" "Strong, straight legs" "Broad head with a slightly rounded skull"
	Concepts: "Straight, arched tail that rests on the back" "Broad, flat skull" "Thick double coat, either rough or smooth"
	Concepts: "Deep-set, almond-shaped eyes" "Erect, triangular ears" "Thick ruff of fur around the neck" "Thick double coat, either rough or smooth"
	Concepts: "Ears: Large, bat-like ears that stand erect and move attentively." "Distinctive facial features: Elongated muzzle with a moderate stop and expressive eyes." "Whiskers: Short or absent due to the lack of hair follicles around the muzzle area."

## F ADDITIONAL EXPERIMENTS

### F.1 IMPACT ON GENERAL ABILITIES

To assess the effect of our proposed training method on the general abilities of the models, we conducted evaluations on four widely used multimodal benchmarks: MMStar, SEED-Bench-2 Plus, MMBench, and MME (Cognition). Table 11 summarizes the results.

- **MMStar** (Chen et al., 2024b): MMStar is a benchmark designed to evaluate vision-indispensable multimodal capabilities of MLLMs. It comprises 1,500 high-quality samples selected through a meticulous process from an initial pool of 22,401 samples. The benchmark assesses six core capabilities, each containing 250 samples, distributed across 18 detailed axes. This structure ensures a comprehensive evaluation of models' performance in tasks that require visual understanding.
- **SEED-Bench-2 Plus** (Li et al., 2024a): SEED-Bench-2 Plus is specifically designed to evaluate text-rich visual comprehension in MLLMs. It features 2,300 multiple-choice questions with precise human annotations, spanning three broad categories: Charts, Maps, and Webs. These categories encompass a wide spectrum of real-world text-rich scenarios, effectively simulating environments where visual and textual information are intertwined. The benchmark aims to assess models' proficiency in interpreting complex visual data embedded with textual content.

- **MMBench** (Yuan Liu, 2023): MMBench is a comprehensive benchmark that evaluates MLLMs across 20 ability dimensions. It includes approximately 3,000 multiple-choice questions, each with a single correct answer. MMBench addresses limitations of traditional benchmarks by focusing on fine-grained abilities and introducing robust evaluation strategies. The benchmark employs ChatGPT to match a model’s prediction with the choices of a question, providing a more reliable assessment of model performance.
- **MME (Cognition)** (Fu et al., 2023): MME (Cognition) benchmark evaluates the cognitive abilities of MLLMs through tasks requiring reasoning and understanding of visual and textual inputs. The cognition part includes four subtasks: Commonsense Reasoning, Numerical Calculation, Text Translation, and Code Reasoning. These tasks test the model’s ability to integrate multimodal information, such as interpreting visual scenes, performing arithmetic based on images, translating text in images, and reasoning about code snippets. The dataset consists of carefully curated images and instruction-answer pairs to ensure robust and fair assessment.

Table 11: General ability evaluation across additional benchmarks. The results demonstrate that our fine-tuned models not only retain their general abilities but also achieve overall improvements compared to the base LLaVA-1.5 model.

Model	MMStar	SEED-Bench-2 Plus	MMBench	MME (Cognition)	Overall Improvement
LLaVA-1.5 (Base)	34.46	41.81	63.05	334.28	–
Trained on CUB-200	33.40	41.78	63.14	355.00	3.2% ↑
Trained on Stanford Dogs	34.93	40.97	63.06	365.71	8.3% ↑
Trained on FGVC-Aircraft	35.14	40.14	63.23	348.57	2.1% ↑
Trained on PLD	35.30	40.89	63.14	337.14	1.1% ↑
Trained on HAM10000	34.46	41.11	64.08	378.21	12.9% ↑

As shown in Table 11, the fine-tuned models exhibit improved performance across multiple benchmarks. Notably, the model fine-tuned on HAM10000 achieves a significant overall improvement of 12.9%, indicating that our training method enhances domain-specific cognition without compromising and, in some cases, improves the models’ general abilities.

## F.2 EFFECTIVENESS OF FILTERING STRATEGIES

To evaluate the importance of our reward model-free rejection sampling method described in Section 3.3, we conducted an ablation study comparing our approach with a baseline that does not employ the filtering mechanism. In this baseline, the model generates the most probable responses during each iteration, which are used directly for training without any filtering. Table 12 presents the classification accuracy and cognition score (CS) across four iterations for both the baseline without filtering and our proposed method.

Table 12: Comparison of accuracy and cognition scores (CS) for the baseline without filtering and our proposed method across iterations.

Dataset	Method	Accuracy (Iter 1)	Accuracy (Iter 2)	Accuracy (Iter 3)	Accuracy (Iter 4)	CS Score
CUB-200	w/o Filtering	68.90	70.11	70.85	70.45	0.71
	Ours	80.24	83.76	84.69	85.02	0.82
FGVC-Aircraft	w/o Filtering	76.36	76.60	77.11	76.78	0.72
	Ours	88.78	90.91	91.42	91.99	0.79
Stanford Dogs	w/o Filtering	76.60	78.53	78.61	78.26	0.74
	Ours	85.29	86.75	86.86	86.91	0.86

From Table 12, it is evident that our filtering strategy significantly enhances both classification accuracy and explanation quality, as measured by the CS score. The baseline without filtering shows marginal improvements initially but fails to achieve comparable performance to our method. This demonstrates the critical role of our filtering mechanism in refining synthetic data and improving the model’s performance iteratively.

### F.3 IMPACT OF TEXT ENCODER

Our framework relies on a text embedding model for estimating mutual information during concept selection (Section 3.2). To assess the impact of different text encoders on concept selection accuracy, we compared three models: E5 (Wang et al., 2022), BERT-Large (Devlin, 2018), and BERT-Base (Devlin, 2018). The results are presented in Table 13.

Table 13: Concept selection accuracy using different text encoders.

Text Encoder	Concept Selection Accuracy (%)
E5	72.9
BERT-Large	71.4
BERT-Base	69.7

As shown in Table 13, the E5 model achieves the highest concept selection accuracy. This indicates that using a more powerful text embedding model improves the quality of concept selection, which in turn enhances the effectiveness of our overall framework.

### F.4 CROSS-DATASET TRANSFER PERFORMANCE

We conducted experiments to evaluate the cross-dataset transferability of our method. Specifically, we trained the model on the CUB-200 dataset and evaluated its performance on the Stanford Dogs dataset. The results are shown in Table 14.

Table 14: Cross-dataset transfer performance when training on CUB-200 and evaluating on Stanford Dogs.

Training Dataset	Evaluation Dataset	Accuracy (%)
None	Stanford Dogs	12.20
CUB-200	Stanford Dogs	16.60
Stanford Dogs	Stanford Dogs	86.91

The results in Table 14 indicate that training on CUB-200 provides a marginal improvement when evaluated on Stanford Dogs. However, the performance remains significantly lower than when the model is trained directly on Stanford Dogs. This suggests that while our method may improve general cognition to some extent, domain-specific fine-tuning is crucial for achieving high accuracy in specialized tasks.

### F.5 EXPLANATION QUALITY IMPROVEMENT ACROSS ITERATIONS

To assess the benefits of our iterative fine-tuning approach, we evaluated the Cognition Score (CS) of the models across four iterations for various datasets. The results are summarized in Table 15.

Table 15: Cognition Scores (CS) across iterations for various datasets. The CS Improvement represents the percentage increase from Iteration 1 to Iteration 4.

Dataset	CS Value (Iter 1)	CS Value (Iter 2)	CS Value (Iter 3)	CS Value (Iter 4)	CS Improvement
CUB-200	0.77	0.76	0.78	0.82	6.5% ↑
Stanford Dogs	0.82	0.84	0.83	0.86	4.9% ↑
FGVC-Aircraft	0.78	0.78	0.78	0.79	1.3% ↑
PLD	0.84	0.85	0.85	0.86	2.4% ↑
HAM10000	0.77	0.84	0.83	0.87	13.0% ↑
Chest X-ray	0.67	0.80	0.81	0.87	29.9% ↑

As observed in Table 15, the CS scores generally improve over iterations, indicating that our iterative fine-tuning process enhances the explanation quality of the models. The most significant improvements are seen in the HAM10000 and Chest X-ray datasets, with 13.0% and 29.9 % reported.