Retrieval-driven Reasoning for Deliberative Visual Classification

Jianye Xie[1,2], Lianyong Qi[1,2*], Fan Wang[3], Anqi Wang[1,2], Wenjuan Gong[1,2*], Danxin Wang[1,2], Wanchun Dou[4], Yang Cao[5,6], Shichao Pei[7], Xiaokang Zhou[8,9]

[1]College of Computer Science and Technology, China University of Petroleum (East China), China
[2]Shandong Key Laboratory of Intelligent Oil and Gas Industrial Software, China
[3]College of Computer Science, Zhejiang University, China
[4]State Key Laboratory for Novel Software Technology, School of Computer Science, Nanjing University, China
[5]School of Computing and Information Technology, Great Bay University, China
[6]Great Bay Institute for Advanced Study, Great Bay University, China
[7]Department of Computer Science, University of Massachusetts Boston, USA
[8]Faculty of Business and Data Science, Kansai University, Japan
[9]RIKEN Center for Advanced Intelligence Project, Japan
b24070008@s.upc.edu.cn, lianyongqi@upc.edu.cn, fanwang97@zju.edu.cn, z24070105@s.upc.edu.cn,
{wenjuangong,wangdx}@upc.edu.cn, douwc@nju.edu.cn, charles.cao@ieee.org, shichao.pei@umb.edu, zhou@kansai-u.ac.jp

Abstract

Vision-Language Models (VLMs) have demonstrated remarkable capabilities in visual classification tasks. Existing methods for enhancing VLMs on this task often rely heavily on direct category-to-image matching, which limits generalization and results in suboptimal performance. In addition, these methods provide no understanding of why a specific category is chosen. To address these limitations, we introduce a new deliberative visual classification task that decomposes the classification process into multiple deliberative steps and leverages Large Language Models (LLMs) to perform explicit reasoning before the final decision. Specifically, we propose a Retrieval-driven Reasoning model (RdR) with two components, i.e., retrieval database construction and deliberative category prediction. The first component leverages LLMs to extract category-relevant descriptors and constructs a retrieval database for effective image–descriptor matching. The second component facilitates multiple deliberative steps and performs explicit reasoning based on the retrieved descriptors to augment the category prediction. Extensive experiments on multiple datasets demonstrate that RdR consistently outperforms strong baselines, highlighting its robustness and generalization ability.

Introduction

Vision-Language Models (VLMs) have demonstrated remarkable capabilities across a wide range of visual recognition tasks (Chen et al. 2024; Zhang et al. 2024). By pretraining on large-scale image-text pairs using contrastive learning objectives, VLMs achieve strong zero-shot and few-shot generalization, enabling them to tackle various downstream tasks, including visual classification (Chandra and Bedi 2021; Conti et al. 2023). However, VLMs still face challenges in adapting to visual classification due to their reliance on large-scale pretraining without task-specific supervision (Evain et al. 2021; Yuan et al. 2022).

Existing methods to enhance the performance of VLMs on visual classification can be broadly categorized into two main approaches: manual prompt-based approaches and prompt learning approaches. Manual prompt-based approaches involve manually designing textual prompts, such as "a photo of a [class]" to guide the model's behavior for specific tasks, often requiring fine-tuning of model parameters (Wen et al. 2023). However, this approach is labor-intensive and highly sensitive to slight prompt modifications. For instance, adding or removing a single word in the prompt can lead to significant fluctuations in accuracy, as demonstrated in CoOp (Zhou et al. 2022b). Inspired by prompt learning in natural language processing (Hirschberg and Manning 2015; Chowdhary and Chowdhary 2020), recent works (Zhou et al. 2022b,a) utilize prompt learning to improve VLMs on visual classification tasks. These methods treat prompts as learnable parameters, keep the VLM backbone frozen, and optimize prompts to minimize the distance between visual and textual features. This approach reduces the need for manual tuning. However, these methods still face two core challenges: **CH1: Limited generalization due to over-reliance on direct category-to-image matching.** Most of these methods directly match image representations with category-specific prompts. This often results in poor generalization to unseen or rare categories, particularly in few-shot learning scenarios (Wang et al. 2020; Kang and Cho 2022). Moreover, relying solely on category names fails to capture the comprehensive visual and contextual information required for accurate recognition. **CH2: Lack of explainable reasoning for the predicted category.** Both manual and learnable prompt-based methods make predictions based on visual-textual similarity and provide no intermediate understanding of why a specific category is chosen, resulting in limited interpretability.

To address the aforementioned issues, we propose a new

*Corresponding Author.

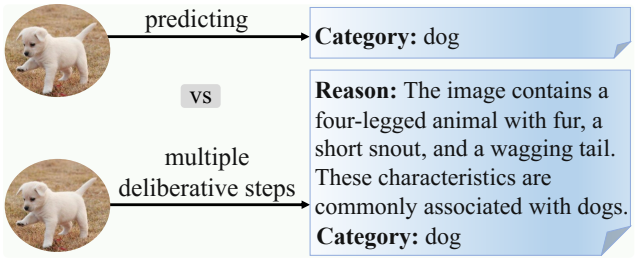

Figure 1: Comparison of conventional visual classification with deliberative visual classification.

Deliberative Visual Classification task. This task aims to decompose the classification process into multiple deliberative steps and perform explicit reasoning before making a final prediction. As illustrated in Figure 1, instead of making a direct prediction, the classifier generates a step-by-step reasoning process and leverages it to guide its final decision. A key to this formulation is identifying the category-relevant information associated with the input image, which serves as a foundation for the reasoning process.

To tackle this task, we propose a Retrieval-driven Reasoning model (**RdR**) for solving the deliberative visual classification task. To mitigate the reliance on direct category-to-image matching (solving **CH1**) and provide explainable reasoning for the predicted category (solving **CH2**), we utilize two main components in RdR, i.e., retrieval database construction and deliberative category prediction. Retrieval database construction aims to solve **CH1** by leveraging MLLMs to obtain descriptors for category-relevant information (visual information and contextual information) and construct a database to support effective image-descriptor matching. Deliberative category prediction aims to solve **CH2** by decomposing the classification process into multiple deliberative steps (descriptors retrieval, descriptors deliberation, and deliberative reasoning) and performing thoughtful reasoning to guide the final prediction. By combining these two components, we can achieve more accurate visual classification while maintaining a transparent and interpretable decision-making process that explains how and why a specific category is chosen, as illustrated in Figure 2. We first construct the retrieval database offline. Then, for a given target image, we retrieve the most relevant descriptors from the database and refine them. Finally, we perform explicit reasoning guided by these refined descriptors to augment the classification for better performance.

The key contributions of our work are summarized as follows: (1) We propose RdR, a novel retrieval-driven reasoning method for deliberative visual classification by decomposing the visual classification into multiple distinct deliberative steps and performing explicit reasoning to augment classification. (2) We clearly decompose image content into visual information and contextual information, and leverage an MLLM to generate precise descriptors for this information and construct a database to facilitate effective image-descriptor matching. (3) We novelly propose deliberative category prediction, which provides a transparent and interpretable decision-making process from feature matching to final classification. By generating a reasoning chain along with the final category prediction, RdR offers clear insights into how and why a specific category is chosen, significantly enhancing the model's interpretability. (4) Extensive experiments demonstrate that RdR outperforms strong baselines across various datasets and settings, highlighting its robustness and generalization capability.

Related Work

Large Language Models. The progress in LLMs can be largely attributed to the transformer architecture (Vaswani et al. 2017). Based on this foundation, models such as BART (Lewis et al. 2019) and Chinchilla (Hoffmann et al. 2022) have demonstrated remarkable capabilities. To improve alignment between outputs and user intent, InstructGPT (Ouyang et al. 2022) incorporates reinforcement learning from human feedback (Christiano et al. 2017), while GPT-4o (Islam and Moushi 2024) extends multimodal capabilities. The open-source community further drives progress with models such as Llama2 (Touvron et al. 2023), and Llama3 (Grattafiori et al. 2024).

Vision-Language Models. VLMs have garnered increasing attention for their remarkable capabilities in cross-modal understanding (Li et al. 2020). For example, CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) demonstrated the effectiveness of contrastive learning frameworks for vision-language pretraining. Other models, such as SimVLM (Wang et al. 2021), and FLAVA (Singh et al. 2022), adopt masked reconstruction techniques to enhance multimodal integration.

Prompt Learning for Visual Classification. Prompt learning (Liu et al. 2023) has emerged as an efficient approach to fine-tuning CLIP in visual classification. CoOp (Zhou et al. 2022b) is the first to introduce learnable prompts into VLMs for few-shot classification. Building upon this, CoCoOp (Zhou et al. 2022a) improves generalization to unseen categories by incorporating instance-level image features. To mitigate overfitting in prompt learning, PromptSRC (Khattak et al. 2023) proposes a self-regularization framework. PromptKD (Li et al. 2024) extends this line of work by applying domain-specific knowledge distillation.

Methodology

In this section, we will introduce the details of our proposed method RdR. First, we describe the notations and problem definition. Suppose we have a pretrained VLM C consisting of a visual encoder $V(\cdot)$ and a text encoder $T(\cdot)$, along with a dataset $D \subset X \times Y$, where each sample consists of an image x and its corresponding category label y. Given an image x from D, our goal is to classify it into one of the visual categories y. Instead of directly predicting y, our approach leverages an LLM L to first generate a reasoning chain R that identifies and explains the visual and contextual evidence supporting the category prediction. Based on R, L then determines the final category, ensuring that the decision is interpretable and grounded in explicit reasoning.

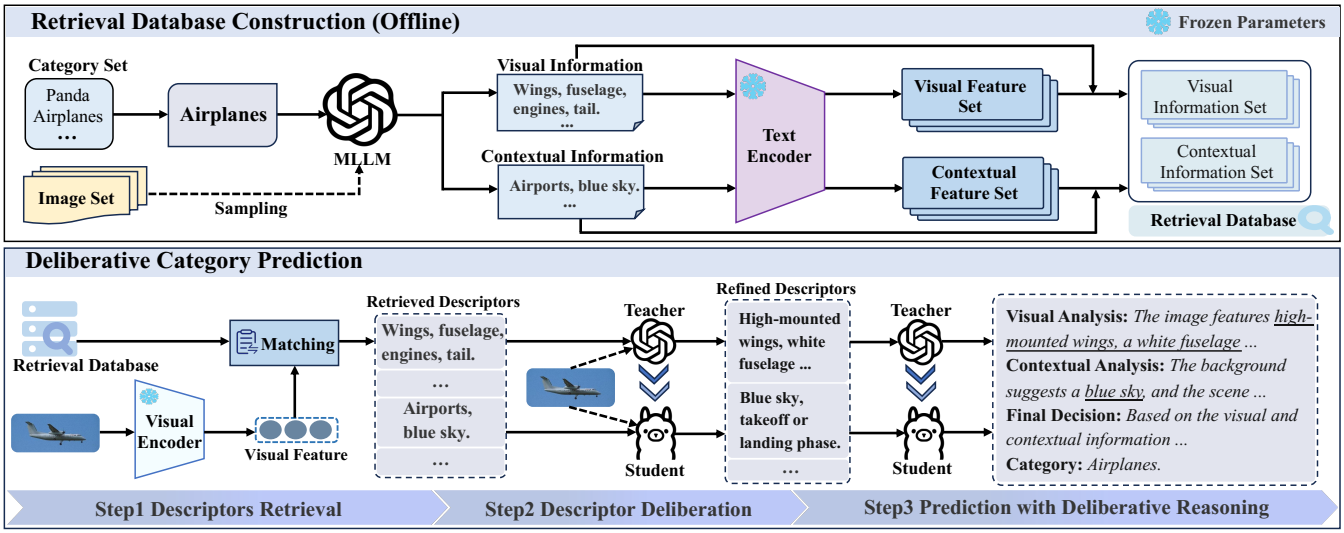

Figure 2: The framework of our proposed RdR model.

We then introduce the main components of our proposed RdR, as illustrated in Figure 2. RdR mainly has two components, i.e., retrieval database construction and deliberative category prediction. Retrieval database construction is set to construct a structured retrieval database to mitigate the reliance on direct category-to-image matching and to facilitate effective image-descriptor matching that provides a foundation for subsequent reasoning. Deliberative category prediction aims to decompose the visual classification into multiple distinct deliberative steps and enables the LLM to perform explicit reasoning before determining the final category. By incorporating retrieval database construction with deliberative category prediction, we can achieve more accurate results on visual classification while maintaining transparency and interpretability in decision-making. We will introduce the model details in the following subsections, and the preliminaries are provided in the Appendix.

Retrieval Database Construction

To address the limitations of existing methods that rely heavily on direct category-to-image matching and to facilitate effective image-descriptor matching, we construct a structured retrieval database in our framework. This database systematically organizes two distinct feature types (visual information and contextual information), enabling comprehensive utilization of all available category-related information.

Decomposition of Image Information. We decompose image information into two complementary types: visual information and contextual information. Visual information refers to distinctive attributes that can be directly extracted from the image, such as color, shape, and specific object features that are characteristic of a given category. For instance, in an image of a dog, visual attributes such as its shape or fur type are typically indicative of the "dog" category. Contextual information refers to the broader environmental and semantic context in which the object appears. This includes contextual relationships such as the setting or interactions involving the object, which reflect the broader background knowledge or situational details related to the object. For example, a "dog" might be associated with the context of being in a park, as dogs are commonly found in such environments. Contextual information provides supplementary details that complement the visual information. This decomposition enables RdR to explicitly consider both visual and contextual information for classification, fully utilizing all available category-related information.

MLLM-based Descriptor Extraction. We employ an MLLM for descriptor extraction through In-Context Learning (ICL) to efficiently generate scalable descriptors for both visual and contextual information. In this approach, we treat the MLLM as an implicit knowledge base, leveraging its extensive pretrained knowledge and strong generative capabilities to generate typical and representative category-relevant descriptors based on structured prompts. Instead of relying solely on category names, we design a structured prompt that includes multiple sample images and the category name, enabling the MLLM to better capture the category's visual distribution and generate descriptors that are grounded and less biased. Figure 3 illustrates the designed prompt, which also incorporates explicit examples to demonstrate the expected response format and content. By providing the MLLM with category-specific examples, we ensure that the generated outputs align with our expectations, accurately capturing both distinctive visual attributes and relevant contextual information.

Our clear definition and decomposition of image information, combined with carefully designed prompts, enable the MLLM to generate precise descriptors for each visual category. This effectively addresses the limitations of previous methods, such as CBD (Menon and Vondrick 2022), which often produce unrelated features. To further enhance the quality and reliability of the descriptors, we introduce a post-processing filtering step based on CLIP image-text similarity. For each descriptor, we compute its similarity to sam-

Given the following inputs:
[category]: Rose
[images]: [image1], [image2], [image3]
Generate commonly observed visual and contextual information of [category], based on the description framework and provided examples.

Description Framework:
1. **Visual Information:** Describe the distinctive visual attributes of an image of a [category]. What are the key features that characterize this object?
2. **Contextual Information:** What common environmental settings are associated with a [category]? Describe the typical context or background where this object is often found.

Example:
Category: Cyclamen
Visual Information: Delicate, nodding flowers in pink, white, or purple with marbled leaves.
Contextual Information: Prefers shaded, well-drained environments.

Output:

Figure 3: Prompt for descriptor extraction.

ple images from the category using CLIP's visual and textual encoders. Descriptors with low average similarity scores across sample images are considered inconsistent or hallucinated, and are consequently filtered out. This filtering step ensures the final descriptors are grounded in visual content, improving both quality and robustness. The details of this filtering step are provided in the Appendix.

Database Construction. After we obtain the textual descriptors for each category, we encode them with the text encoder $T(\cdot)$ to derive the corresponding text features. These text features, along with their associated descriptors, are stored in two distinct sets: a visual information set and a contextual information set. The visual information set captures features that directly describe object attributes, ensuring category-level identification, while the contextual information set provides supplementary details that enhance visual understanding. By structuring the retrieval database in this dual-feature manner, our approach enables fine-grained and context-aware retrieval during classification.

Offline Construction. To minimize runtime latency, we construct the retrieval database entirely offline. In this pipeline, descriptors are pre-extracted, and their embeddings are precomputed. Once constructed, this database supports efficient and scalable retrieval during both subsequent training and inference, significantly reducing runtime overhead.

Deliberative Category Prediction

After constructing the retrieval database, we will introduce how deliberative category prediction integrates with it to improve classification accuracy. Deliberative category prediction includes three main steps, i.e., descriptors retrieval, descriptors deliberation, and prediction with deliberative reasoning. Descriptors retrieval aims to identify the most relevant descriptors from the retrieval database for a target image, to support subsequent reasoning. Descriptors deliberation is set to fine-tune an expert to refine the retrieved descriptors. Prediction with deliberative reasoning aims to fine-tune another expert to generate a deliberative reasoning process based on refined descriptors to augment category predictions. By combining these three steps, our model can achieve more accurate and interpretable classification decisions through multiple distinct deliberative steps.

Step 1: Descriptors Retrieval. First, we leverage the visual encoder $V(\cdot)$ to encode the input image x to extract its visual features u as follows:

$$u = V(x). \tag{1}$$

The extracted visual features u are then used to retrieve the most relevant text features along with their associated descriptors from the retrieval database. Specifically, retrieval is performed separately from two distinct sets: a visual information set and a contextual information set, each retrieving the top N relevant descriptors. The retrieval process computes the cosine similarity scores:

$$s = \frac{u \cdot t_i}{\|u\|\|t_i\|}, \tag{2}$$

where t_i denotes the textual feature of the i-th descriptor stored in the retrieval database. The s are computed and ranked accordingly to select the most relevant descriptors, which include the retrieved visual descriptors i_v and contextual descriptors i_c. Unlike traditional approaches that rely solely on direct category-to-image matching, our method explicitly separates the matching process into visual and contextual information matching. This separation allows the model to consider not only the visual appearance of objects but also their semantic and contextual relationships. For instance, in addition to recognizing that an image contains a "dog", our framework can retrieve contextual information such as the dog's common environments where it might be found. This matching process significantly enhances the model's ability to generalize, particularly in scenarios where visual features alone may be ambiguous or insufficient.

Step 2: Descriptor Deliberation. After we obtain the retrieved descriptors i_v and i_c, we fine-tune an MLLM-based expert to refine the retrieved descriptors. The expert takes as input the retrieved descriptors i_v and i_c, along with the original image x, to generate more detailed and accurate descriptors r_v and r_c. For instance, if the retrieval phase identifies a "dog" in the image, the MLLM might augment this information by describing the dog's posture or the setting (e.g., a park or a living room), while filtering out or redefining features unrelated to the image. Specifically, we guide the expert through the following template:

Template: *Please combine the retrieved visual and contextual information along with the image to obtain detailed and precise information about the identified object. Focus on attributes such as visual details (e.g., skin color, texture) or contextual details (e.g., common locations). Ensure accuracy and keep the response concise while filtering out or redefining features unrelated to the image.*

You are an expert vision-language model tasked with classifying an image based on both visual and contextual information. You are given the following inputs:
Visual information: …
Contextual information: …
A predefined category set: …

Your task is to reason step by step and provide a final classification label. First, analyze the given visual features and describe what they suggest about the object in the image. Then, incorporate contextual information to refine your understanding. If multiple categories are possible, explain why certain options are more likely than others. Finally, conclude with the most probable category and a concise explanation for your decision.

Format your response as follows:
1. **Visual Analysis:** Describe the key visual features and what they indicate.
2. **Contextual Analysis:** Incorporate additional contextual details to refine the classification.
3. **Final Decision:** State the most likely category and explain why this choice is the most appropriate.
4. **Category:** Provide the final predicted category label.

Ensure that your reasoning is logical, transparent, and interpretable.

Figure 4: Prompt for deliberative category prediction.

This step can be formally expressed as:

$$r_v, r_c = \text{MLLM}(x, i_v, i_c, \text{Template}). \tag{3}$$

Here, r_v and r_c represent the refined visual and contextual descriptors. This deliberation step ensures that the descriptors are both relevant and informative, enhancing the retrieved descriptors for more accurate classification.

Step 3: Prediction with Deliberative Reasoning. Finally, we leverage deliberative reasoning combined with the refined descriptors to enhance visual classification. We fine-tune an LLM-based expert, which is prompted to generate both a deliberative reasoning process and the final category prediction. The input to the expert includes the refined descriptors obtained in **Step 2**, which contain visual and contextual information and are ordered based on the relevance determined in **Step 1**. It also includes the predefined category set from the dataset, as well as a carefully designed instruction to guide the reasoning process. Figure 4 illustrates the prompt for deliberative category prediction, and further details about the prompt design are provided in the Appendix. Based on this prompt, the expert constructs a transparent and deliberative decision-making process that explains how and why a specific category is chosen. This reasoning process not only enhances the model's classification performance but also provides transparent insights into its decisions, thereby improving the interpretability and trustworthiness of the overall framework. This step can be formally expressed as:

$$R, k = \text{LLM}(r_v, r_c, S, P), \tag{4}$$

where r_v and r_c denote the refined descriptors from **Step 2**, S represents the predefined category set, and P is the instruction designed to guide the reasoning process. Here, k is the final predicted category, and R is the structured reasoning chain generated by the expert. Figure 5 shows an example of the structured reasoning output generated by the expert, illustrating how visual and contextual information contribute to the final category prediction.

Visual Analysis: The image contains a large, four-legged animal with thick gray skin, a long trunk, and prominent tusks. These distinctive physical characteristics are commonly associated with elephants.
Contextual Analysis: The background appears to resemble a savanna or grassland environment, which is a typical natural habitat where elephants are commonly found.
Final Decision: Based on the visual and contextual information, the object in the image is most likely an elephant.
Category: Elephant.

Figure 5: An example of the structured reasoning output.

Expert Fine-tuning. To balance performance and efficiency, we leverage the powerful closed-source model (e.g., GPT-4o) to guide the fine-tuning of two specialized open-source experts. Specifically, one expert is trained for **Step 2** to refine the retrieved descriptors, while the other is trained for **Step 3** to perform deliberative reasoning. We prompt the teacher model to generate high-quality outputs, which serve as supervision signals to train the corresponding student experts via supervised fine-tuning (SFT) (Ouyang et al. 2022), implemented using LoRA adapters (Hu et al. 2022). To mitigate hallucinations, we adopt a generation-then-filter strategy: the teacher model first generates multiple candidate outputs, which are then filtered based on relevance and consistency. The selected high-quality outputs are used to supervise student learning through SFT. A detailed description of this strategy is provided in the Appendix.

Experiments

Setup

We conduct experiments to evaluate our proposed method RdR on the visual classification task under the in-distribution (ID) setting. For fair comparison, we assess RdR against baseline methods using CLIP with different backbones (ViT-B/16 (Dosovitskiy et al. 2020), ViT-B/32 (Dosovitskiy et al. 2020), and ResNet-50 (He et al. 2016)). All baselines follow the experimental settings for few-shot learning evaluation as specified in their respective papers.

Baselines. We compare RdR with four baseline methods: **CLIP** (Radford et al. 2021), **CoOp** (Zhou et al. 2022b), **PromptKD** (Li et al. 2024), and **CBD** (Menon and Vondrick 2022). The first baseline, zero-shot CLIP, relies on hand-crafted prompts. CoOp and PromptKD represent recent state-of-the-art prompt learning methods. The most related work, CBD, performs classification by computing category scores with CLIP, based on the similarity between images and class-conditioned textual prompts.

Datasets and Evaluation Metric. Following (Radford et al. 2021; Zhou et al. 2022b), we evaluate the model's performance on 11 publicly available recognition datasets, including ImageNet (Deng et al. 2009), Caltech101 (Fei-Fei, Fer-

Method	ImageNet			Caltech101			OxfordPets			StanfordCars		
	ViT-B/16	ViT-B/32	RN-50	ViT-B/16	ViT-B/32	RN-50	ViT-B/16	ViT-B/32	RN-50	ViT-B/16	ViT-B/32	RN-50
CLIP	64.05	58.46	58.18	86.19	85.43	86.10	81.88	79.94	83.19	55.98	56.43	56.10
CoOp	71.92	66.85	61.91	90.30	89.90	91.99	85.67	85.29	87.02	72.92	73.40	73.60
PromptKD	73.60	68.73	64.25	91.15	90.63	93.02	86.33	86.30	87.76	73.27	74.97	74.58
CBD	68.03	62.97	62.39	85.60	85.57	85.97	86.92	83.46	84.10	58.60	59.57	59.02
RdR	**75.05**	**70.28**	**68.19**	**91.96**	**90.95**	**93.45**	**87.03**	**86.59**	**87.85**	**75.15**	**76.32**	**75.22**
Method	Flowers102			Food101			FGVCAircraft			SUN397		
	ViT-B/16	ViT-B/32	RN-50	ViT-B/16	ViT-B/32	RN-50	ViT-B/16	ViT-B/32	RN-50	ViT-B/16	ViT-B/32	RN-50
CLIP	65.83	66.10	66.15	85.61	79.31	79.07	17.08	17.43	17.16	59.03	59.43	59.92
CoOp	93.60	93.07	94.49	74.93	75.06	74.48	31.34	32.30	31.43	67.65	67.90	68.36
PromptKD	94.07	94.50	95.05	77.32	76.50	77.10	32.83	34.57	33.19	68.72	68.57	69.50
CBD	66.67	68.55	68.32	88.50	83.63	80.65	21.80	21.57	21.62	62.60	62.97	64.10
RdR	**94.52**	**94.80**	**96.07**	**90.02**	**85.76**	**84.60**	**33.07**	**34.89**	**34.45**	**70.32**	**70.53**	**71.85**
Method	DTD			EuroSAT			UCF101			Average		
	ViT-B/16	ViT-B/32	RN-50	ViT-B/16	ViT-B/32	RN-50	ViT-B/16	ViT-B/32	RN-50	ViT-B/16	ViT-B/32	RN-50
CLIP	41.78	40.93	41.80	43.36	43.84	43.10	60.69	61.43	61.20	60.13	58.98	59.27
CoOp	62.44	61.58	62.51	82.95	83.17	83.69	74.55	76.05	76.90	73.48	73.14	73.31
PromptKD	63.37	62.37	63.65	83.92	83.77	84.32	76.10	77.83	77.95	74.61	74.43	74.58
CBD	46.59	45.62	46.02	48.82	48.94	47.75	63.50	64.57	64.09	63.42	62.49	62.18
RdR	**66.15**	**64.20**	**66.79**	**84.19**	**84.62**	**85.06**	**78.29**	**79.15**	**79.60**	**76.89**	**76.19**	**76.65**

Table 1: Comparison with state-of-the-art methods across 11 datasets using different backbones.

gus, and Perona 2004), OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Van Gool 2014), FGVCAircraft (Maji et al. 2013), SUN397 (Xiao et al. 2010), DTD (Cimpoi et al. 2014), EuroSAT (Helber et al. 2019), and UCF101 (Soomro, Zamir, and Shah 2012). Consistent with prior works (Radford et al. 2021; Zhou et al. 2022b; Khattak et al. 2023), we use classification accuracy as the evaluation metric.

Implementation Details. To construct the retrieval database, we employ GPT-4o for descriptor extraction, with the temperature parameter set to 0 to maintain consistency. For each dataset, we build a corresponding retrieval database accordingly. We set $K = 2$ as the number of retrieved descriptors. For the descriptor deliberation step, we use GPT-4o as the teacher model and fine-tune Llava-v1.6-mistral-7B as the expert using the LoRA technique. For the deliberative reasoning step, we similarly use GPT-4o as the teacher and fine-tune Llama3-8B as the expert with LoRA. For the teacher model used in both steps, we set the temperature to 1 and max_tokens to 4096.

Comparison with Baseline Methods

Table 1 presents the performance comparison between CLIP, CoOp, PromptKD, CBD, and RdR with different backbones (ViT-B/16, ViT-B/32, and ResNet-50). RdR consistently outperforms all baselines across 11 publicly available recognition datasets. Specifically, on the challenging ImageNet dataset with ViT-B/32 backbone, RdR achieves an accuracy improvement of 11.82% over CLIP, 3.43% over CoOp, 1.55% over PromptKD, and 7.31% over CBD. Notably, compared to the most related work, CBD, RdR demonstrates a significant performance gain. We attribute the observed performance gains across all comparisons to two key factors: (1) RdR leverages image-descriptor matching rather than direct category-to-image matching, allowing for more flexible and robust alignment and thereby improving the model's generalization to diverse instance images. (2) We enhance the visual classification by introducing multiple deliberative steps and performing explicit reasoning to augment predictions, leading to more accurate and reliable final predictions. These results validate the effectiveness and robustness of our approach in visual classification tasks.

Method	K=1	K=2	K=3	K=4	K=5
ViT-B/16	74.52	**75.05**	74.39	74.02	73.36
ViT-B/32	69.76	**70.28**	69.52	69.18	69.05
ResNet-50	66.95	**68.19**	66.17	64.20	63.09

Table 2: Ablation study on the impact of the number of retrieved descriptors on classification performance.

Ablation Study

Effect of the Number of Retrieved Descriptors. We conduct an ablation study to analyze how the number of top-K retrieved descriptors affects performance. Table 2 presents the classification accuracy on the challenging ImageNet dataset, illustrating the impact of different K values. We ob-

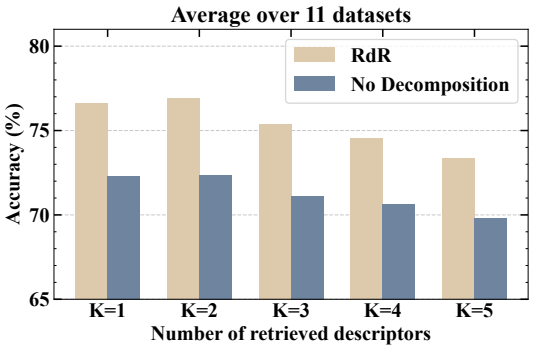

Figure 6: Ablation study on the effect of image information decomposition into visual and contextual descriptors.

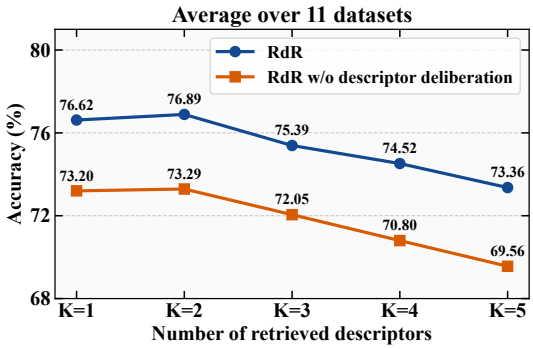

Figure 7: Performance comparison of RdR with and without the descriptor deliberation.

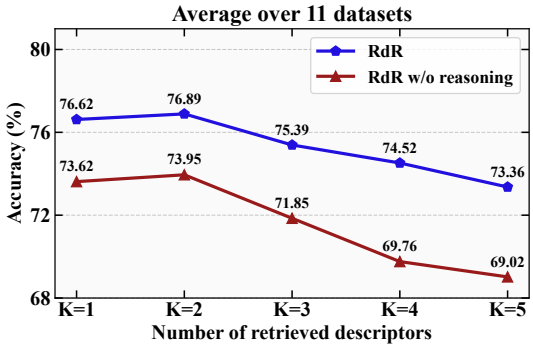

Figure 8: Performance comparison of RdR with and without the reasoning process in category prediction.

serve that retrieving too few descriptors ($K = 1$) provides limited information, resulting in suboptimal performance. Additionally, an overly large K introduces significant irrelevant information, which weakens the impact of relevant descriptors and causes a noticeable drop in classification accuracy. Our results show that selecting an optimal value for K ensures the retrieved descriptors are both relevant and informative. Based on our experiments, we set $K = 2$ as the default value for all evaluations.

Information Decomposition and Separate Retrieval. To evaluate the impact of information decomposition, we compare RdR with the baseline ("**No Decomposition**" in Figure 6) that stores all descriptors in a unified set without separation. We adopt ViT-B/16 as the backbone and vary the number of retrieved descriptors for evaluation. For fairness, the baseline retrieves the top-$2K$ descriptors, matching the total number retrieved from both sets. Figure 6 presents the average classification accuracy across 11 datasets. The results show that RdR outperforms the **No Decomposition** baseline across different K values. This demonstrates that explicit decomposition combined with separate retrieval improves the quality of retrieved information, thereby enhancing the classification performance of RdR.

Effect of Descriptor Deliberation. We conduct an ablation study to evaluate the effect of descriptor deliberation, comparing RdR with and without this step. In the baseline (w/o descriptor deliberation), descriptors are used directly for classification. Experiments are conducted with ViT-B/16 under different numbers of retrieved descriptors. Figure 7 reports the average classification accuracy across 11 datasets. The results show that incorporating descriptor deliberation consistently improves performance. Without this step, the retrieved descriptors often include irrelevant or misleading information due to retrieval noise. These findings highlight the importance of integrating this step to ensure that the final descriptors are both relevant and informative, thereby improving classification accuracy and generalization.

Effect of Prediction with Deliberative Reasoning. We assess the contribution of prediction with deliberative reasoning by comparing RdR with a baseline that directly predicts categories from refined descriptors without reasoning. Experiments are conducted with ViT-B/16 under varying numbers of retrieved descriptors. Figure 8 presents the average classification accuracy across 11 datasets, demonstrating that incorporating the reasoning process consistently enhances classification performance. Without reasoning, the model relies solely on direct feature matching, making it more prone to errors when categories are visually similar or overlapping. These findings underscore the importance of deliberative reasoning in enhancing not only classification accuracy but also the interpretability of model predictions.

Conclusion

In this work, we address the limitations of existing methods by introducing a novel deliberative visual classification task that decomposes the visual classification into multiple deliberative steps and performs explicit reasoning to augment classification. To achieve this task, we propose a Retrieval-driven Reasoning model (**RdR**). RdR comprises two components: retrieval database construction and deliberative category prediction. The first component leverages MLLMs to extract category-relevant descriptors and constructs a database for effective image–descriptor matching. The second component facilitates multiple deliberative steps and performs reasoning to augment the category prediction. We also conduct extensive experiments to demonstrate the superior performance of our proposed RdR model.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62572486), Natural Science Foundation of Shandong Province (No. ZR2023QF101, No. ZR2023MF007, No. ZR2023MF041).

References

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 446–461. Springer.

Chandra, M. A.; and Bedi, S. 2021. Survey on SVM and their application in image classification. *International Journal of Information Technology*, 13(5): 1–11.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Chowdhary, K.; and Chowdhary, K. 2020. Natural language processing. *Fundamentals of artificial intelligence*, 603–649.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Conti, A.; Fini, E.; Mancini, M.; Rota, P.; Wang, Y.; and Ricci, E. 2023. Vocabulary-free image classification. *Advances in Neural Information Processing Systems*, 36: 30662–30680.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Evain, S.; Nguyen, M. H.; Le, H.; Boito, M. Z.; Mdhaffar, S.; Alisamir, S.; Tong, Z.; Tomashenko, N.; Dinarelli, M.; Parcollet, T.; et al. 2021. Task agnostic and task specific self-supervised learning from speech with lebenchmark. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, 178–178. IEEE.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Hirschberg, J.; and Manning, C. D. 2015. Advances in natural language processing. *Science*, 349(6245): 261–266.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D. d. L.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Islam, R.; and Moushi, O. M. 2024. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Kang, D.; and Cho, M. 2022. Integrative few-shot learning for classification and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9979–9990.

Khattak, M. U.; Wasim, S. T.; Naseer, M.; Khan, S.; Yang, M.-H.; and Khan, F. S. 2023. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15190–15200.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2020. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*.

Li, Z.; Li, X.; Fu, X.; Zhang, X.; Wang, W.; Chen, S.; and Yang, J. 2024. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26617–26626.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9): 1–35.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Menon, S.; and Vondrick, C. 2022. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 722–729. IEEE.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15638–15650.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Yao, Q.; Kwok, J. T.; and Ni, L. M. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3): 1–34.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Wen, Y.; Jain, N.; Kirchenbauer, J.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36: 51008–51025.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Yuan, W.; Zhang, Z.; Wang, C.; Song, H.; Xie, Y.; and Ma, L. 2022. Task-level self-supervision for cross-domain few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3215–3223.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.