Feature-Level Knowledge Distillation from LMM for Enhanced Image Classification

Bumsu Jang
Department of Artificial Intelligence
Kyungpook National University
Buk-gu, Daegu, 41566, Republic of Korea.
`bsjang1208@gmail.com`

Heechul Jung
Department of Artificial Intelligence
Kyungpook National University
Buk-gu, Daegu, 41566, Republic of Korea.
`heechul@knu.ac.kr`

Abstract

Large Multimodal models (LMMs) leverage a vast number of parameters and large-scale training data to acquire extensive knowledge and exhibit strong reasoning capabilities. However, despite their generality, they often fall short of surpassing the performance of vision models that are specialized for traditional vision-centric tasks. Although recent efforts have been made toward developing smaller language models, they remain insufficient for visual reasoning in environments constrained by memory and communication resources. In this study, we investigate the transfer of prior knowledge from LMMs into vision models, and observe notable improvements in performance. Our experiments highlight the role of LMM-generated text in enhancing vision model training, providing new insights into improving vision models through multimodal knowledge transfer.

1 Introduction

Deep learning has achieved remarkable success in visual recognition tasks such as image classification, object detection, and segmentation, with models like ResNet [He et al., 2016] and ViT [Dosovitskiy, 2020] standing out in particular. Meanwhile, the advancement of large language models (LLMs) such as GPT-4 [Achiam et al., 2023] and LLaMA [Touvron et al., 2023] has demonstrated the potential of leveraging large-scale text data to enhance visual recognition performance. Furthermore, large multimodal models (LMMs), including GPT-4V [Ding et al., 2023] and LLaVA [Liu et al., 2024b], have shown that incorporating the knowledge embedded in language models can improve the understanding of visual representations and boost performance on vision tasks.

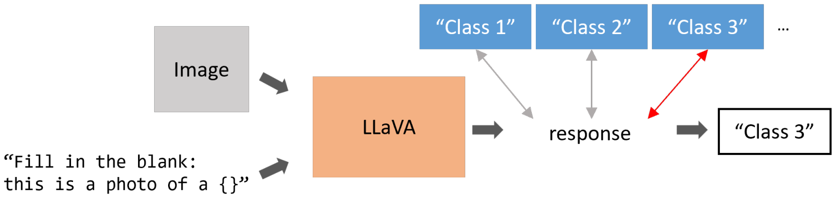


Figure 1: A simple procedure for using LLaVA for image classification involves selecting the class name with the highest cosine similarity to the LLaVA response.

However, LMMs often underperform in vision-only tasks—such as image classification, object detection, and semantic segmentation—when compared to conventional vision architectures. To illustrate this gap, we conducted a simple experiment evaluating image classification accuracy on three benchmark datasets (CIFAR-100 [Krizhevsky et al., 2009], Caltech-101 [Li et al., 2022a], and

CUB-200 [Li et al., 2022a]) using LLaVA and ResNet-50. Since LLaVA cannot directly output class labels, we derived predictions using the procedure illustrated in Figure1. As summarized in Table1, LLaVA's performance was significantly lower than that of ResNet-50, which is specifically optimized for classification.

In contrast, LMMs demonstrate outstanding capabilities in vision-language tasks such as Visual Question Answering (VQA) and image captioning [Liu et al., 2024b]. To mitigate their weaknesses in vision-only settings, several studies have explored fine-tuning strategies [Sun et al., 2025, Wang et al., 2024, Zhang et al., 2024], including low-rank adaptation (LoRA) [Hu et al., 2021, Sun et al., 2025] and visual prompt tuning (VPT) [Jia et al., 2022]. While these methods improve task-specific performance, they remain computationally expensive at inference time, limiting their practicality in resource-constrained environments. As a result, lightweight models are often favored in such scenarios, reinforcing the need to enhance the visual recognition ability of small-scale models.

Table 1: Classification accuracy on each dataset comparing Resnet-50 and LLaVA (zero-shot).

Dataset	ResNet-50	LLaVA (zero-shot)
CIFAR-100	84.59	49.05
Caltech-101	96.39	69.30
CUB-200	78.51	6.21

In this paper, we propose a knowledge distillation framework that transfers the rich knowledge embedded in Large Multimodal Models (LMMs) to lightweight vision models. LLaVA serves as a teacher for ResNet-50, enabling the student model to achieve high-quality predictions with lower computational cost. Specifically, we leverage LLaVA's ability to generate diverse textual descriptions of an image—even from identical prompts—to construct a broad distribution of representations in the CLIP [Radford et al., 2021] embedding space. These feature-level descriptions (e.g., background, color, shape, texture) provide richer supervision, allowing the student model to learn fine-grained visual attributes.

To align visual and textual representations, we maximize the cosine similarity between CLIP text embeddings and ResNet-50 image embeddings, encouraging the model to capture semantically consistent features. Furthermore, we adopt a progressively growing distillation rate that gradually increases the influence of the dissimilarity loss, balancing teacher guidance and the student's own representations.

Extensive experiments demonstrate that our framework consistently improves image classification accuracy, confirming that multimodal knowledge from LMMs can effectively enhance small-scale vision models.

Contributions of this work:

- A knowledge distillation framework that transfers multimodal knowledge from LMMs to lightweight vision models.
- Utilization of LLaVA's diverse text generation to create rich, feature-level CLIP embeddings for supervision.
- A progressive distillation strategy that adaptively balances teacher and student learning.
- Empirical validation showing consistent improvements in ResNet-50 classification accuracy.

2 Related Works

2.1 Multimodal Information for Visual Recognition

CLIP [Radford et al., 2021] represents a pioneering Vision–Language Model (VLM) that incorporates textual information for visual recognition tasks. The model employs a simple handcrafted template, such as "a photo of [class]", paired with the corresponding image, and identifies the class with the highest similarity between text and image embeddings. While CLIP demonstrated the potential of multimodal information for visual recognition, its ability to exploit textual knowledge was constrained by its reliance on simple class-based prompts and the relatively limited capacity of its language model compared to recent LLMs. To address this limitation, Tzelepi and Mezaris [Tzelepi and Mezaris, 2024] utilized ten LMM-generated descriptions per image, averaged their text embeddings, and concatenated them with image embeddings to improve classification performance. However, such averaging may restrict the expressiveness of text embeddings, leading to suboptimal solutions.

Meanwhile, several studies—including MMPT [Wang et al., 2024] and LLaMA-Adapter [Zhang et al., 2024]—have explored VPT-based fine-tuning to harness LMM knowledge for visual recognition. These approaches train learnable prompts that guide LMMs toward task-specific outputs, yielding strong results in image classification and optical character recognition (OCR). Nevertheless, they require the use of LMMs during inference, resulting in computational costs that are prohibitive in resource-constrained environments.

In this paper, we propose an efficient knowledge distillation framework that transfers the rich knowledge of LMMs into conventional vision models. Our method preserves the expressiveness of qLMM-generated text in the CLIP embedding space while maintaining the computational efficiency necessary for practical deployment.

2.2 Knowledge Distillation

Knowledge distillation, first introduced by Hinton *et al.*[Hinton, 2015], has become a fundamental technique for transferring knowledge from large models to smaller ones. This approach has been extensively studied in both unimodal and multimodal domains, leading to substantial improvements in the performance of compact models[Xie et al., 2020, Touvron et al., 2021, Yang et al., 2024, Xu et al., 2024]. Among these efforts, CLIP-KD [Yang et al., 2024] explored optimal strategies for multimodal knowledge distillation, particularly within CLIP architectures for image classification and cross-modal retrieval. Similarly, LLAVADI [Xu et al., 2024] conducted a comprehensive analysis of critical factors in LMM distillation, focusing on VQA tasks with LLaVA.

While these studies highlight the effectiveness of knowledge distillation in multimodal contexts, they either overlook the potential of diverse textual descriptions or restrict their scope to specialized tasks such as VQA. In contrast, our work demonstrates that leveraging the rich textual information generated by LMMs can substantially enhance vision-only tasks. We present comprehensive validation of LMM-based knowledge distillation in the context of image classification, showing that textual diversity plays a crucial role in improving performance.

3 Method

3.1 Model Architecture

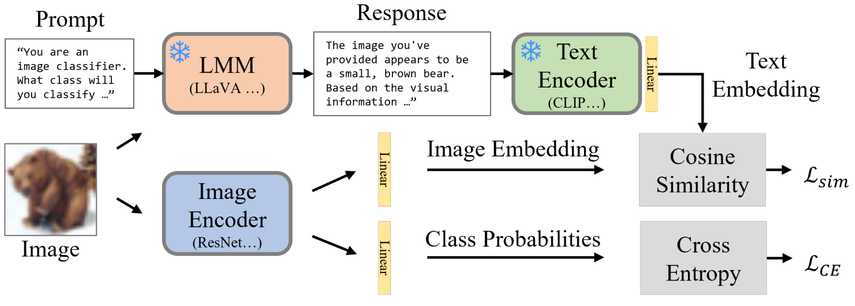


Figure 2: Overview of the proposed knowledge distillation framework. LLaVA generates diverse textual descriptions that are converted into CLIP text embeddings. ResNet-50 produces image embeddings, and both embeddings are aligned by combining cross-entropy loss with a dissimilarity loss.

We propose a knowledge distillation framework to transfer the prior knowledge of a LMM into a lightweight vision model. The overall pipeline is illustrated in Figure 2. In this framework, LLaVA [Liu et al., 2024b] serves as the teacher model and ResNet-50 [He et al., 2016] as the student model.

3.2 Framework

LLaVA, as a multimodal model, can generate diverse descriptions of an input image given a textual prompt. These generated texts are converted into embeddings using the CLIP [Radford et al., 2021] text encoder. Meanwhile, the student model, ResNet-50, produces visual embeddings from the same image. Both embeddings are then aligned into a shared representation space, enabling the student model to learn the semantic information contained in LLaVA.

3.3 Embedding alignment

The text embedding is obtained from the CLIP text encoder, and the image embedding is derived from the output of ResNet-50 through a linear transformation. Both embeddings are projected into the same dimensional space, allowing the student model to mimic text-based representations.

3.4 Loss function

Training is conducted with a combination of two loss functions. First, the cross-entropy loss is used for conventional classification training. Second, a dissimilarity loss based on the cosine similarity between text and image embeddings is introduced to encourage alignment between the two representations. Training is conducted with a combination of two loss functions. First, the cross-entropy loss ($\mathcal{L}_{CE}$) for conventional classification training. Second the dissimilarity loss ($\mathcal{L}_{sim}$) based on cosine similarity to encourage alignment between text and image embeddings. The final loss is defined as:

$$\mathcal{L}_{total} = (1 - \lambda)\mathcal{L}_{CE} + \lambda\mathcal{L}_{sim}.$$

where $\lambda \in [0, 1]$ is a weighting factor that balances the contributions of the two terms. Instead of keeping λ fixed, we adopt a progressive growing strategy in which λ starts at zero and increases linearly as training progresses. This gradual schedule prevents abrupt changes in the loss landscape when the student model is pre-trained with cross-entropy alone, and allows the model to smoothly adapt to the additional supervision provided by the teacher.

3.5 Feature-level description

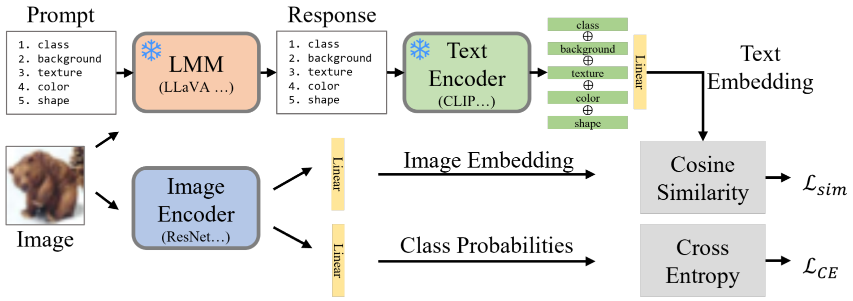


Figure 3: The Architecture of the model to use combined text embeddings of descriptions.

We also investigated the impact of both the number and the content of LMM responses on model performance. To this end, we categorized the prompts into five types, each instructing LLaVA to describe a different visual feature of the same image: **class**, **background**, **texture**, **color**, and **shape**. Among these, the class feature directly relates to the target task of image classification, and we specifically refer to it as the **task-specific feature**.

We first trained the model using only the LMM responses generated from each individual prompt type and analyzed their respective effects. Next, to examine the benefit of combining task-specific and additional descriptive features, we concatenated the CLIP [Radford et al., 2021] embeddings of multiple LMM responses and applied a linear transformation, as illustrated in Figure 3. We then evaluated the performance across different combinations of textual descriptions generated through this method.

4 Experiment

4.1 Models Used

4.1.1 LLaVA

LLaVA [Liu et al., 2024b] is built upon LLaMA [Touvron et al., 2023], an open-source LLM released by Meta. Unlike conventional VLMs such as Flamingo [Alayrac et al., 2022] and BLIP [Li et al., 2022b], which generate text solely from image–text pairs, LLaVA integrates a vision encoder with an LLM, enabling richer reasoning capabilities. In this work, we employ LLaVA-Next (also known as LLaVA-1.6) [Liu et al., 2024a], which demonstrates improved performance in reasoning, OCR, and world knowledge compared to earlier versions.

4.1.2 CLIP

CLIP [Radford et al., 2021] is a representative vision–language model consisting of separate text and image encoders. It is trained with a contrastive learning objective that maximizes cosine similarity between an image and its paired text embeddings. Various backbone architectures have been explored for its vision encoder. In this paper, we adopt OpenAI's CLIP ViT-B/32 model [Dosovitskiy, 2020] to extract text embedding vectors from LMM-generated responses. This version employs a transformer-based Vision Transformer (ViT) as its image encoder.

4.1.3 ResNet

Deep CNNs often suffer from performance degradation as the number of layers increases. To mitigate this issue, He *et al.* [He et al., 2016] proposed the residual block, which introduces shortcut connections that bypass intermediate layers and add their output to subsequent layers. In our experiments, we use ResNet-50, a 50-layer CNN pre-trained on ImageNet [Deng et al., 2009], both as an image feature extractor and as the baseline classifier.

4.2 Dataset

We employ three widely used image classification benchmarks. Below are the characteristics of the datasets used in our study.

The CIFAR-100 [Krizhevsky et al., 2009] dataset is a widely used benchmark for classification tasks, containing 100 classes. The dataset consists of 60,000 images, with an equal number of 600 images per class. However, the image resolution is relatively low, at 32 x 32 pixels, which makes it difficult for both humans and LMMs to accurately analyze the images, as they appear blurred.

The Caltech-101 [Li et al., 2022a] dataset is designed for image classification tasks and contains 101 classes. The number of images per class and the image resolutions are not consistent across the dataset. Nevertheless, since the image resolution is higher than that of CIFAR-100, the classification accuracy of ResNet-50 for Caltech-101 is generally higher compared to CIFAR-100.

The CUB-200 [Wah et al., 2011] dataset is intended for image classification and object detection tasks, containing 200 classes of bird species. The image resolution is similar to that of Caltech-101, but the CUB-200 dataset focuses specifically on bird species. Consequently, without fine-tuning on this dataset, it is challenging for LLaVA to effectively differentiate between the classes based only on visual information.

4.3 Experimental Settings

This section describes the experimental configurations, including the dissimilarity loss ratio, dataset size, and training hyperparameters.

4.3.1 Dissimilarity Loss Ratio

As described in Section 3.4, the total loss is computed as a weighted sum of cross-entropy loss and dissimilarity loss. The model's performance therefore depends on the dissimilarity loss ratio λ. To investigate this effect, we conducted experiments with $\lambda \in \{1, 0.75, 0.5, 0.25\}$ and identified the value yielding the best performance. We further compared two scheduling strategies: **fixed** λ, where the ratio remains constant across epochs, and **progressive growing** λ, where the ratio increases linearly from 0 at the start of training to its predefined maximum value (one of $\{1, 0.75, 0.5, 0.25\}$) by the final epoch. For clarity, we report only the best performance achieved among the tested λ values.

4.3.2 Dataset Size

We also examined the effect of training dataset size on model performance. To this end, we reduced the training data to fractions of $\{1.0, 0.5, 0.25, 0.1\}$ of the full dataset. Our hypothesis is that the additional descriptive information provided by LLaVA can mitigate the performance degradation caused by reduced training data. If the performance gap between the baseline and our proposed method increases as the dataset shrinks, this would support our hypothesis.

4.3.3 Hyperparameters

All models were trained for 50 epochs using the SGD optimizer. A grid search was conducted over learning rates $\{0.1, 0.01, 0.001, 0.0001\}$ and batch sizes $\{8, 16, 32, 64\}$. For each setting, we report the best performance achieved under the corresponding configuration.

4.3.4 Implementation Details

All experiments were conducted on a workstation equipped with an NVIDIA RTX A6000 GPU. Training was implemented in PyTorch with CUDA acceleration.

4.4 Results

4.4.1 Task-specific description.

Table 2 summarizes classification accuracy on CIFAR-100 [Krizhevsky et al., 2009], Caltech-101 [Li et al., 2022a], and CUB-200 [Wah et al., 2011]. Using class-based prompts, we compared our framework against zero-shot LLaVA and a ResNet-50 baseline trained with cross-entropy loss. The analysis focuses on how fixed and progressive λ schedules influence pre-trained and scratch models, and how dataset size affects the utility of LLaVA's knowledge. Overall, the proposed framework improved accuracy in both training settings.

For CIFAR-100 and Caltech-101, the progressive λ method consistently outperformed the fixed version in fine-tuning, although no clear relationship emerged between dataset size and improvement. Fine-tuning yielded only marginal gains compared to training from scratch, likely due to the difficulty of adapting parameters already optimized under cross-entropy loss. On CUB-200, performance fluctuated under both schedules, reflecting LLaVA's limited capacity for fine-grained species discrimination.

In the scratch setting (Table 2b), our method was generally effective, with fixed λ often surpassing progressive λ. This suggests that models without prior knowledge benefit from stronger LLaVA supervision early in training. Notably, scratch training improved CUB-200 accuracy, implying that even inaccurate class predictions from LLaVA can provide valuable descriptive guidance.

In summary, the effectiveness of the dissimilarity loss ratio depends on both pre-training status and dataset characteristics, indicating that optimal knowledge transfer strategies vary across scenarios.

4.4.2 Combining with other feature descriptions.

In addition to the results of using the task-specific prompt, we conducted experiments with four additional prompts requesting LLaVA's responses on specific image features (background, texture, color, shape). We then trained ResNet-50 using these responses to examine their effects. Furthermore, we evaluated performance when combining the task-specific description with other feature descriptions ("Class + One") and when combining all descriptions except one ("Class – One"), using the model architecture described in Section 3.5. All experiments in this section were conducted with pre-trained models using 100% of the data and the progressive growing λ method.

Table 2: Classification accuracy on each dataset. (a) Fine-tuned model from a pre-trained model, and (b) Model trained from scratch. For the fine-tuned model, the progressive growing λ method is outperforms the fixed λ method, while both methods show similar performance when training from scratch.

Dataset	Data Size (%)	LLaVA (zero-shot)	ResNet-50	Ours (Fixed λ)	Ours (Progressive Growing λ)
CIFAR-100	100	49.05	84.59	84.53 (-0.06)	85.26 (+0.67)
	50	–	81.70	81.82 (+0.12)	82.37 (+0.67)
	25	–	78.15	78.89 (+0.74)	79.08 (+0.93)
	10	–	73.69	73.15 (-0.54)	73.88 (+0.19)
Caltech-101	100	69.30	96.39	96.22 (-0.17)	96.79 (+0.40)
	50	–	95.05	95.79 (+0.75)	95.74 (+0.69)
	25	–	92.83	92.83 (-)	93.23 (+0.40)
	10	–	89.11	89.06 (-0.05)	89.56 (+0.46)
CUB-200	100	6.21	78.51	79.14 (+0.63)	79.60 (+1.10)
	50	–	73.48	72.80 (-0.68)	73.18 (-0.30)
	25	–	62.29	61.87 (-0.42)	61.53 (-0.76)
	10	–	44.55	44.64 (+0.08)	46.11 (+1.56)

(a) Model fine-tuned from pre-trained model

Dataset	Data Size (%)	LLaVA (zero-shot)	ResNet-50	Ours (Fixed λ)	Ours (Progressive Growing λ)
CIFAR-100	100	49.05	58.67	63.27 (+4.60)	63.78 (+5.11)
	50	–	45.60	53.70 (+8.10)	52.13 (+6.53)
	25	–	35.35	41.35 (+6.00)	42.01 (+6.66)
	10	–	22.07	27.20 (+5.19)	26.92 (+4.85)
Caltech-101	100	69.30	74.43	77.13 (+2.70)	75.29 (+0.86)
	50	–	62.27	65.67 (+3.40)	61.81 (-0.26)
	25	–	52.48	52.48 (-)	52.53 (+0.06)
	10	–	38.82	40.67 (+1.84)	40.32 (+1.50)
CUB-200	100	6.21	36.57	38.56 (+1.98)	38.05 (+1.48)
	50	–	18.54	22.97 (+4.43)	20.48 (+1.94)
	25	–	7.90	9.46 (+1.56)	9.76 (+1.86)
	10	–	3.93	3.93 (-)	4.10 (+0.17)

(b) Model trained from scratch

In these experiments, when multiple text embeddings were used, we hypothesized that the output dimension of the linear transformation should scale with the number of concatenated features. We therefore conducted preliminary experiments by varying n and setting the output dimension to $256 \times n$. Based on these experiments, we set $n = 2$ for the "Class + One" cases across all datasets, and $n = 4$ for CIFAR-100, $n = 6$ for Caltech-101, and $n = 6$ for CUB-200 in the "All – One" cases, and report performance under these settings.

When all descriptions were used, accuracy increased as the target dimension (n) grew. For Caltech-101 and CUB-200, the highest accuracy was achieved when $n = 6$, while for CIFAR-100, the highest accuracy occurred at $n = 3$ and $n = 4$. In contrast, When only the task-specific description was used, the highest accuracy was achieved at $n = 1$, and performance decreased as n increased.

Based on these findings, we applied $n = 2$ for 'Class + One' cases across all datasets, and $n = 4$ for CIFAR-100, $n = 6$ for Caltech-101 and $n = 6$ for CUB-200 in the 'All - One' case, to observe the accuracy.

As a result, unlike in Table 2, performance improvements compared to the baseline (ResNet-50) were observed in all cases for the CIFAR-100 and Caltech-101 datasets. Specifically, the 'All - Shape' case showed a significant improvement and achieved the highest accuracy on the CIFAR-100 dataset, as

Table 3: Accuracy for each description combination when the output dimension of the linear transform is changed.

	CIFAR-100	Caltech-101	CUB-200
ResNet-50	84.59	96.39	78.51
Class	85.26 (+0.67)	**96.79 (+0.40)**	**79.6 (+1.09)**
Class + Background	84.68 (+0.09)	96.67 (+0.28)	77.66 (-0.85)
Class + Texture	85.23 (+0.64)	96.62 (+0.23)	78.17 (-0.34)
Class + Color	84.69 (+0.10)	96.44 (+0.05)	79.31 (+0.8)
Class + Shape	84.77 (+0.18)	**96.79 (+0.40)**	78.84 (+0.33)
All - Background	85.22 (+0.63)	96.44 (+0.05)	78.89 (+0.38)
All - Texture	84.85 (+0.26)	96.50 (+0.11)	78.25 (-0.26)
All - Color	85.06 (+0.47)	96.55 (+0.16)	79.35 (+0.84)
All - Shape	**85.72 (+1.13)**	96.67 (+0.28)	78.42 (-0.09)
All	85.02 (+0.43)	96.62 (+0.23)	79.30 (+0.79)

shown in Table 3. In contrast, for the CUB-200 dataset, performance degradation was observed in approximately half of the cases, with no significant change.

4.5 Ablation Study

4.5.1 Unified Sentence

As can be seen above, our method functions to some extent. However, the question arises: is the use of the LMM necessary? To utilize the CLIP model for image classification, we defined captions in the image-text pairs as “The photo of *[class name]*" - we refer to this type of text as ‘unified text’. The class name corresponding to the most similar text to the image was selected [Radford et al., 2021, Zhou et al., 2022b,a]. Considering only the discriminability between the text embedding vectors, this method proves effective. However, the LMM model can provide a broader range of expressions. Therefore, we explored whether using these unified sentences alone would suffice or if the diversity and richness of expressions provided by the LMM would enhance the vision model’s generalization performance. The experiment was conducted using the CIFAR-100 dataset and the ‘Progressive Growing λ’ method, which showed the most significant improvement in the fine-tuning scenario.

Table 4: Classification accuracy comparison between unified text and LLM-generated text on CIFAR-100 dataset

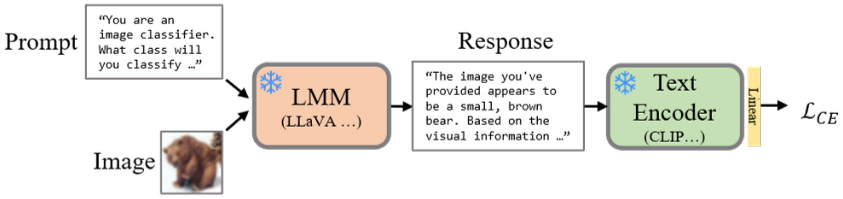
Amount of Data	Baseline	Unified Text	LMM-generated Text (ours)
100	84.59	84.66 (+0.04)	85.26 (+0.67)
50	81.70	82.41 (+0.74)	82.37 (+0.67)
25	78.15	78.98 (+0.83)	79.08 (+0.93)
10	73.69	73.58 (-0.11)	73.88 (+0.19)

As shown in Table. 4, most of the accuracies achieved with unified texts showed improvements compared to the baseline, but still underperformed when compared to the results using LMM-generated text, particularly in the case of using 50% of the data. This indicates that learning with text enhances the performance of the vision model. Additionally, text with greater diversity contributes more effectively than text that solely emphasizes discriminability. These results demonstrates the necessity of using LLM-generated text.

4.5.2 Classification of Text Embedding

Since the text embedding vectors of task-specific description in the CLIP embedding space form distinct clusters by class and effectively represent distances between images, it is possible to directly classify the text embedding vectors by training only the linear classifier. To achieve this, we designed the model as shown in Figure 4 and evaluated the classification performance.

As seen in Table 5, our method consistently demonstrates higher performance; however, for Caltech-101, where LLaVA can effectively describe the images, there is no significant difference. In contrast, due to LLaVA’s inability to accurately classify bird species, the classification accuracy of the text



"You are an image classifier. What class will you classify ..."

"The image you've provided appears to be a small, brown bear. Based on the visual information ..."

Figure 4: A pipeline for directly classifying the text embedding vectors of LLaVA responses.

embeddings for CUB-200 showed the largest discrepancy compared to our method. This result also explains why our method struggles with the CUB-200 dataset.

Table 5: Accuracy comparison between our method and direct classification of text embedding vectors. The performance of the progressive growing λ on finetuning scenario is shown for our method.

Dataset	Amount of Data	Ours	Text Embedding
CIFAR-100	100	85.26	60.43
	50	82.37	59.85
	25	79.08	58.54
	10	73.88	57.03
Caltech-101	100	96.79	89.69
	50	95.74	88.82
	25	93.23	87.67
	10	89.56	85.55
CUB-200	100	79.60	18.54
	50	73.18	16.05
	25	61.53	14.15
	10	46.11	11.99

4.5.3 Grad-CAM

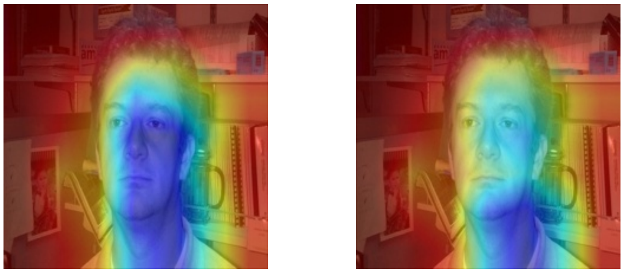

(a) class feature (b) background feature

Figure 5: Grad-CAM result of ResNet-50 using class and background feature

Grad-CAM [Selvaraju et al., 2017] visualizes the regions a CNN-based model attends to by using gradients from the final convolutional layer. We applied this method to verify whether our feature-level description approach effectively influences visual attention. Among the various image features requested from LLaVA, the performance difference between class and background features had the most significant impact on classification accuracy when using the CIFAR-100 dataset (Section 4.4.2). Thus, we compared Grad-CAM results of ResNet-50 models trained with each feature type (Figure 5). When using the background feature, the model often focused on irrelevant regions, leading to reduced accuracy. However, this issue was uncommon—most cases differed only in the degree of focus. Notably, even with the "background" description, the model rarely attended to the actual background. These observations suggest that improper feature descriptions may confuse the model, and identifying which features should be emphasized remains an open challenge.

5 Conclusion

5.1 Summary

In this study, we proposed a novel knowledge distillation framework that leverages LMM-generated text embeddings to enhance CNN-based vision models. Through extensive experiments, we demonstrated the effectiveness of our method in improving image classification performance, particularly on datasets like CIFAR-100 and Caltech-101. Our approach capitalizes on the diversity and richness of textual descriptions to train vision models with more comprehensive feature representations.

However, our findings also reveal that the effectiveness of LMM-generated text embeddings varies depending on dataset characteristics. For instance, while significant improvements were observed on general datasets, such as CIFAR-100, the method showed limited impact on fine-grained datasets, like CUB-200, where LMMs struggled to generate domain-specific discriminative features. Despite these limitations, our method provides a promising direction for bridging LMM knowledge with traditional vision models.

5.2 Future Works

Although our method has shown promising results, there are several avenues for future exploration:

1. **Domain-Specific Knowledge Integration**: Future research could focus on augmenting LMMs with domain-specific training or fine-tuning techniques to address the challenges observed in fine-grained datasets like CUB-200.
2. **Extending to Other Vision Tasks**: Our study primarily focused on image classification. Extending this framework to other vision tasks, such as object detection or semantic segmentation, would validate its general applicability.
3. **Optimization for Resource-Constrained Environments**: While our approach demonstrates computational efficiency, further optimization is necessary for real-time applications or deployment in resource-limited environments.
4. **Exploration of Other Model Architectures**: Incorporating other vision architectures, such as Vision Transformers (ViTs), could provide additional insights into the versatility of our framework.

By addressing these directions, future research could build upon the strengths of our method and overcome its current limitations, further advancing the integration of LMMs and vision models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Ning Ding, Yehui Tang, Zhongqian Fu, Chao Xu, Kai Han, and Yunhe Wang. Gpt4image: Can large pre-trained models help vision models on perception tasks? *arXiv e-prints*, pages arXiv–2306, 2023.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, 4 2022a.

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*, pages 12888–12900. PMLR, 2022b.

Haotian Liu, Chunyuan 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*, pages 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Guohao Sun, Can Qin, Jiamian Wang, Zeyuan Chen, Ran Xu, and Zhiqiang Tao. Sq-llava: Self-questioning for large vision-language assistant. In *European Conference on Computer Vision*, pages 156–172. Springer, 2025.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Maria Tzelepi and Vasileios Mezaris. Exploiting lmm-based knowledge for image classification tasks. In *International Conference on Engineering Applications of Neural Networks*, pages 166–177. Springer, 2024.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Taowen Wang, Yiyang Liu, James Chenhao Liang, Yiming Cui, Yuning Mao, Shaoliang Nie, Jiahao Liu, Fuli Feng, Zenglin Xu, Cheng Han, et al. Mmpt: Multimodal prompt tuning for zero-shot instruction learning. *arXiv preprint arXiv:2409.15657*, 2024.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

Shilin Xu, Xiangtai Li, Haobo Yuan, Lu Qi, Yunhai Tong, and Ming-Hsuan Yang. Llavadi: What matters for multimodal large language models distillation. *arXiv preprint arXiv:2407.19409*, 2024.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15952–15962, 2024.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

A Dataset Details

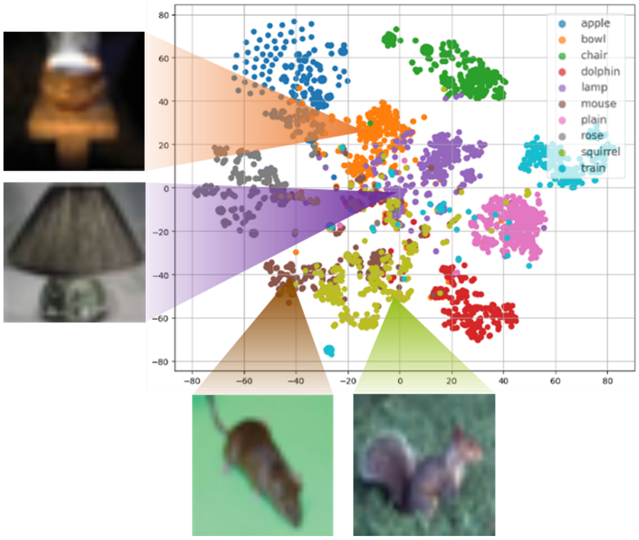


Figure A1: Text embedding vectors of LLaVA model's response

Figure A2: Text embedding vectors in 2D space and example images of CIFAR-100 dataset's 10 classes.Because LLaVA can generate diverse descriptions for each image, the vectors are distributed more evenly overall. However, clusters still form among similar classes, and images from classes with unclear boundaries exhibit similar patterns in their representations.

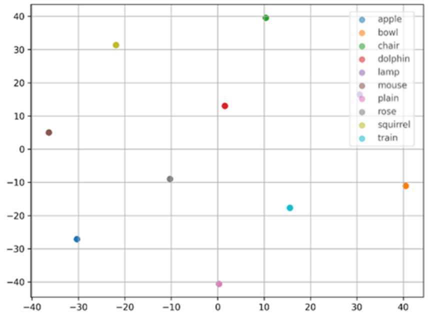
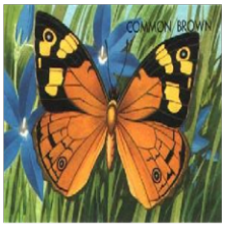


Figure A3: Embedding vectors of unified text from CIFAR-100's class name. They shows stronger class separability than the LLaVA-generated descriptions but lacks their diversity in representing individual instances.

Class: Butterfly

Image:



Prompt:
You are an image classifier. What class will you classify this image into?
Response:
The image you've provided appears to be a butterfly. Based on the visual characteristics such as the pattern on the wings and the overall shape, it can be classified as a butterfly.

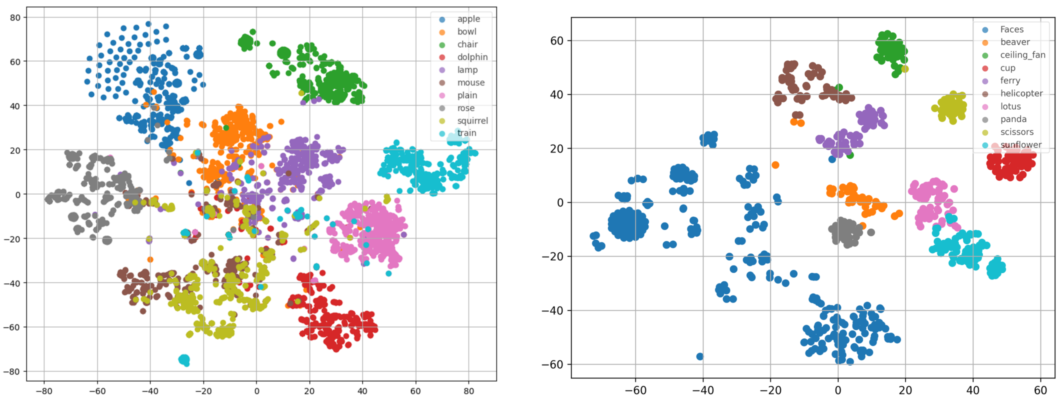
Prompt:
Please describe the background of this image
Response:
The background of the image features a field of blue flowers, which could be identified as forget-me-nots, set against a green grassy area. The flowers are in full bloom, and their bright blue color contrasts with the green

Prompt:
Please describe the texture of this image
Response:
The image features a butterfly with a patterned wing, which suggests a textured appearance. The texture of the butterfly's wings is likely to be a combination of smooth and rough surfaces, with the smooth areas

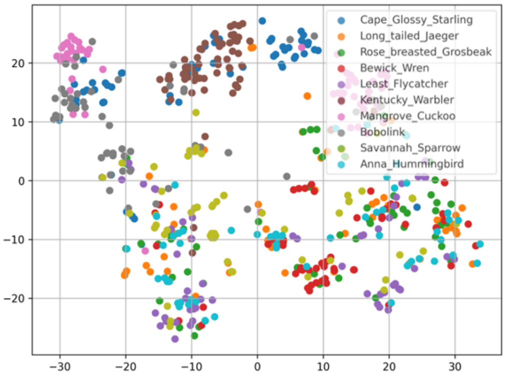
Prompt:
Please describe the color of this image
Response:
The image features a butterfly with a combination of orange and black colors. The butterfly's wings have a pattern of orange and black, and the body of the butterfly is also orange. The background consists of green grass and blue flowers

Prompt:
Please describe the shape of the object in this image
Response:
The object in the image is a butterfly. It has a symmetrical shape with two wings on either side of its body. The wings are elongated and have a pattern of colors, which is typical for many butterfly species.

Figure A4: Examples of LLaVA responses to different prompts for the butterfly class. The task-specific prompt generates a detailed class-level description, whereas feature-specific prompts (background, texture, color, shape) focus on individual attributes while still referencing the class name.



(a) CIFAR-100 dataset (b) Caltech-101 dataset



(c) CUB-200 dataset

Figure A5: Visualization of text embedding vectors for 10 classes from each dataset in 2D space. CIFAR-100 shows less distinct class boundaries due to low image resolution, while Caltech-101 exhibits clearer separation. In contrast, CUB-200 embeddings are highly mixed, reflecting LLaVA's limited ability to distinguish fine-grained bird species.