CARPE: CONTEXT-AWARE IMAGE REPRESENTATION PRIORITIZATION VIA ENSEMBLE FOR LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) are typically trained using autoregressive language modeling objectives, which align visual representations with linguistic space. While effective for multimodal reasoning, this alignment can weaken vision-centric capabilities, causing LVLMs to underperform their base vision encoders on tasks such as image classification. To address this limitation, we propose Context-Aware Image Representation Prioritization via Ensemble (CARPE), a lightweight framework that integrates raw vision features with aligned LLM representations through vision-integration layers and a context-aware ensemble mechanism. This design enhances the model's ability to adaptively weight visual and textual modalities and enables the model to capture various aspects of image representations. Extensive experiments demonstrate that CARPE improves performance on both image classification and diverse vision-language benchmarks. Our results suggest that modality balancing plays a critical role in multimodal generalization by improving representation utilization within autoregressive LVLMs.

1 INTRODUCTION

Large vision-language models (LVLMs) have become increasingly popular in the research community, as they serve as foundational building blocks towards general-purpose assistants (Liu et al., 2024a; 2023; Li et al., 2023a; Dai et al., 2023; Zhu et al., 2023; Wang et al., 2024; Ye et al., 2023; Chen et al., 2024b). While existing LVLMs exhibit impressive performance across various vision-language tasks, recent studies have highlighted their limitations in image classification (Zhai et al., 2023b; Zhang et al., 2024; Mitra et al., 2024). Notably, Zhai et al. (2023b) and Zhang et al. (2024) reveal that LVLMs significantly underperform CLIP (Radford et al., 2021) on standard image classification benchmarks such as ImageNet (Deng et al., 2009), despite CLIP being their base vision encoder, indicating that LVLMs do not fully preserve the generalization properties of their underlying vision encoders.

This underperformance in image classification presents a significant bottleneck for LVLMs. Although LVLMs are primarily designed for generative tasks, many vision-language tasks inherently rely on robust classification capabilities. The internal reasoning process often involves recognizing and categorizing visual elements before generating an answer. For instance, a sample from the TextVQA (Singh et al., 2019) benchmark presents the question, "*What company made this?*" along with an image of a laptop. If the model fails to first classify the object as a laptop, subsequent reasoning steps are likely to be incorrect. Consequently, enhancing classification performance could naturally lead to broader improvements in LVLMs' overall capabilities.

Building on the findings of Zhang et al. (2024), we investigate whether fine-tuning LVLMs can enhance their general image classification performance. Our experiments demonstrate that while fine-tuning LLaVA1.5 (Liu et al., 2024a) on ImageNet improves accuracy within the dataset, it simultaneously hurts the model's general capabilities. As shown in Table 1, this fine-tuning approach results in decreased performance across multiple benchmarks. The declines in CV-Bench (Tong et al., 2024a) and MMVP(Tong et al., 2024b) are particularly notable, as they are vision-centric

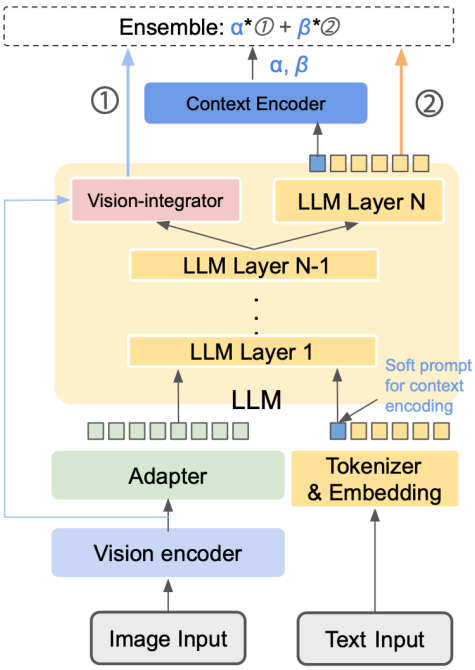


Figure 1: CARPE architecture incorporates vision-integrator with a context-aware ensemble approach, dynamically combining vision representations from diverse perspectives.

Model	Classification				Vision-language Benchmarks				
	ImageNet	Caltech	Flowers102	Food101	SQA	MME	MMB	CV-Bench	MMVP
LLaVA1.5-7B (Liu et al., 2024a)	26.4	55.4	6.7	29.5	69.4	1862.7	64.7	46.0	63
+ ImageNet fine-tune (Zhang et al., 2024)	78 (+51.6)	54.4 (-1.0)	6.7 (-)	22.9 (-6.6)	66.8 (-2.6)	1744.2 (-118.5)	62.1 (-2.6)	37.9 (-8.1)	56.6 (-6.4)

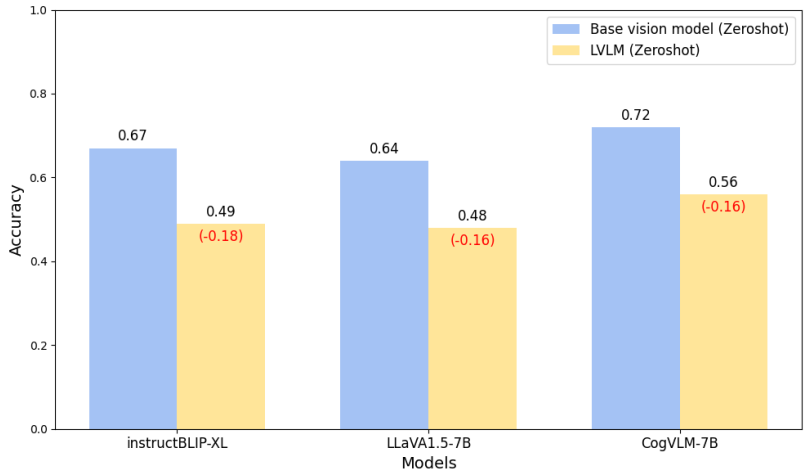
Table 1: Performance of vanilla LLaVA1.5-7B and its ImageNet fine-tuned checkpoint on classification and vision-language benchmarks. The numbers in parentheses indicate the change in performance compared to the vanilla model.

benchmarks where classification was anticipated to provide benefits. This indicates that simply fine-tuning LVLMs on classification does not effectively generalize to other vision-language tasks, nor does it lead to consistent improvements in general visual understanding.

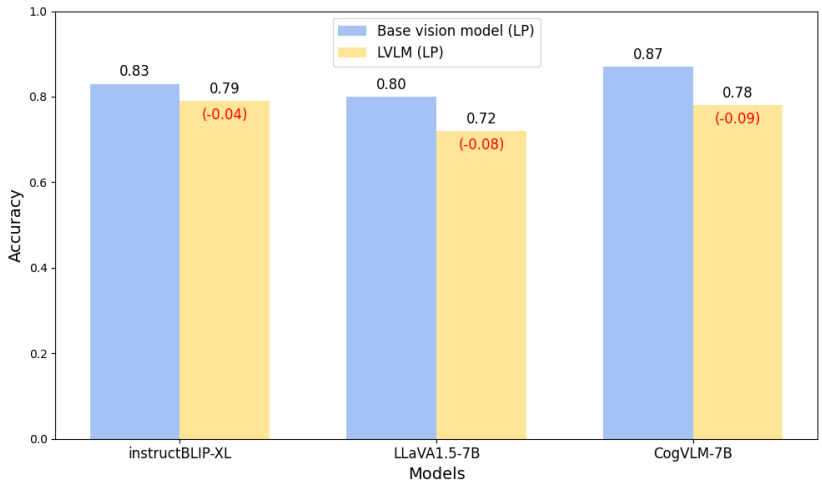
This observation leads us to investigate the following question: *How can we enhance the general visual understanding of LVLMs by improving their classification ability?*

Our key insight, supported by experimental results in Figure 2, is that classification-relevant information is largely retained within the LVLMs' latent space, despite being diminished in the final generated output. When evaluating three LVLMs on four classification datasets in a zero-shot setting, we observe a significant performance drop compared to their base CLIP models. However, when evaluated in a linear probing setting, they substantially close the performance gap with CLIP. These results indicate that classification-relevant visual information is largely retained in the LVLMs' latent space with minimal forgetting, as Zhang et al. (2024) found earlier. However, it becomes suboptimally aligned for downstream discriminative tasks during alignment with the LLM.

Based on this insight, we hypothesize that LVLMs struggle to adaptively discern when to prioritize image representations versus relying on language-based reasoning in a context-dependent manner. For example, in vision-centric tasks (e.g., image classification), which requires a strong focus on vision inputs, LVLMs may struggle to appropriately prioritize visual information over language model's reasoning capabilities due to the misalignment. This inability to balance visual and textual modalities based on the context could undermine their performance across diverse tasks.



(a) Zero-shot classification accuracy of base vision models (CLIP) and LVLMs.



(b) Linear probing classification accuracy comparing base vision models output and LVLMs final output.

Figure 2: Image classification performance analysis of LVLMs in zero-shot and linear probing settings. Accuracies are averaged across four classification datasets, Caltech101, Flowers102, DomainNet and Mini-ImageNet.

To address this, we propose a dynamic ensemble approach that integrates two embeddings: (1) raw vision encodings directly from the vision encoder, and (2) final LLM outputs. These embeddings provide different visual perspectives, as prior studies have found that when vision features are aligned with language, they tend to focus on different aspects of information (Radford et al., 2021; Chen et al., 2023a; Tong et al., 2024b; Chen et al., 2024a; Gao et al., 2022). For example, Radford et al. (2021) highlights that image-caption pairs emphasize high-level semantics over detailed descriptions. Therefore, by leveraging embeddings before and after LLM alignment, we aim to extract and utilize a richer set of visual features, enabling more effective utilization of suboptimally aligned visual representations.

Specifically, we fuse these complementary embeddings through a vision-integrator coupled with a context-aware ensemble mechanism. The vision-integrator captures different aspects of vision features, by combining vision encoder's output with LLM output to complement the visual information that may become misaligned during LLM's alignment process. Furthermore, the outputs of the vision-integrator and the final LLM output are combined via a context-aware ensemble mechanism guided by a context encoder. This context encoder dynamically adjusts ensemble weights based on

context from text input, allowing the model to prioritize the vision-integrator output when pre-LLM alignment information is more relevant or the final LLM output when language-based reasoning is more required.

Thus, we introduce this framework as **C**ontext-**A**ware Image **R**epresentation **P**rioritization via **E**nsemble (CARPE), which significantly improves performance on image classification benchmarks and extends benefits to various multimodal zero-shot tasks. Extensive experiments demonstrate that CARPE's context-aware ensemble design effectively integrates visual information from diverse perspectives to enhance their general capabilities. CARPE's model-agnostic framework can be seamlessly integrated with a wide range of open-source LVLMs that comprise a vision encoder and a language model comprising a vision encoder and a language model, ensuring broad applicability and ease of deployment.

Our contributions in this work are as follows:

- We empirically analyze modality imbalance in LVLMs and show that classification-relevant information is largely preserved in latent representations but becomes underutilized after language alignment.
- We introduce CARPE, a novel framework that adaptively integrates multiple perspectives of visual features to enhance general image understanding in a context-dependent manner.
- We demonstrate that our vision-integrator and context-aware ensemble improve generalization by enhancing classification performance.

2 Related Works

Large Vision-Language Models Recent LVLMs are predominantly built on pre-trained vision and large language models (Liu et al., 2023; 2024a; Dai et al., 2023; Zhu et al., 2023; Bai et al., 2023; Wang et al., 2024; Ye et al., 2023; Chen et al., 2024b). These models are often connected using different types of adapter modules, such as MLPs (Liu et al., 2023; 2024a; Lin et al., 2023; Chen et al., 2023b), and Q-former (Li et al., 2023a; Dai et al., 2023) to integrate the different modalities. Models like Qwen2-VL (Wang et al., 2024) and InternVL (Chen et al., 2024b) have showcased impressive capabilities in instruction-following and visual reasoning tasks, while some models are specifically designed to enhance visual understanding ability (Lin et al., 2023; Wang et al., 2023). SPHINX (Lin et al., 2023) employs an ensemble of various vision backbones to extract robust visual representations from different aspects, and CogVLM (Wang et al., 2023) introduces a visual expert module, doubling its parameters in language model to improve visual understanding abilities. In comparison, our work leverages light-weight vision-integrator to efficiently enhance image comprehension.

Limitations of LVLMs in image classification Although LVLMs showcase strong performance on many vision-language tasks, recent research underscores their shortcomings in image classification (Zhai et al., 2023b; Zhang et al., 2024; Mitra et al., 2024; Cai et al., 2025). For example, Zhai et al. (2023b) and Zhang et al. (2024) reveal that LVLMs fail to inherit the generalizability of CLIP on standard image classification tasks. Cai et al. (2025) identify this failure as a cross-modality competency problem, where LVLMs struggle to fairly assess information across modalities. In contrast, our framework introduces a dynamic weighting mechanism that adaptively prioritizes visual information, improving overall classification ability.

Vision features aligned to language overlook visual details Many studies have found that text-aligned image features emphasize high-level content while overlooking fine-grained details (Radford et al., 2021; Chen et al., 2023a; Tong et al., 2024b; Chen et al., 2024a; Gao et al., 2022). For instance, Radford et al. (2021) suggests that image-caption pairs focus on high-level semantics description rather than visual details, leading to image representation primarily capture global features. To address this issue, Gao et al. (2022) proposes constructing three visual embeddings from different semantic levels for more accurate alignment between image and text in vision-language pre-training. In this work, we leverage the visual representation both before and after LLM alignment to effectively extract different aspect of visual information.

3 METHODS

3.1 ARCHITECTURE

In this section, we introduce CARPE's three key components: a pre-trained LVLM, vision-integrators and a context-aware weighting modules consisting of a context prompt and a context encoder. The overall framework of CARPE is summarized in Figure 1.

Pre-trained LVLM LVLMs typically consists of three parts: a vision encoder, which is commonly based on CLIP, a large language model (LLM), and an adapter that connects the two modalities, typically implemented as an MLP or a Q-former. Our framework is designed to be compatible with any LVLM that incorporates both vision and language components.

Vision-integrator We introduce a vision-integrator to effectively combine three types of visual information: (1) raw vision features from the base vision encoder (e.g., CLIP), (2) the LLM representations prior to the final vocabulary projection.

The motivation for this integration stems from previous studies indicating that aligning image features with language shifts the model's focus towards semantic level while losing fine-grained visual details. Based on this finding, we assume that each of these feature types —before and after LLM alignment —encodes complementary perspectives of the image. Thus, the vision-integrator is designed to efficiently merge these two forms of information, to enhance the model's comprehensive visual understanding.

Specifically, vision-integrator consists of a multi-head cross-attention layer followed by multi-head self-attention layer and an MLP. The queries originate from the LLM's second-to-last output, while the keys and values are derived from the raw vision features. Since the raw vision features are not initially aligned with language space, they are first projected into the LLM's dimension using a newly introduced MLP adapter.

Context encoder and context prompt We introduce a context-aware weighting mechanism to enable the model to distinguish between vision- or language-centric contexts, rather than simply adding the two embeddings in an ensemble. We assume different tasks may require different weighting of these embeddings.

To explicitly encode this ability, we incorporate a context prompt—a learnable soft prompt appended to the input text embeddings—motivated by prompt tuning (Lester et al., 2021). The context prompt is processed by the LLM as standard text input, and subsequently passed through the context encoder. The context encoder generates two probability values that sum to 1.0, which serve as ensemble weights between the vision-integrator and the final LLM representations.

We ensure that the context prompt is influenced solely by text inputs, as the distinction between vision-centric and language-centric tasks is determined by the instruction rather than the image content. In our implementation, the context encoder is a single linear layer followed by a softmax function, and the context prompt consists of a learnable embedding of length one. The final prediction is obtained by a weighted sum of the vision-integrator and LLM logits (see Appendix A.1 for formulation).

Mixture of Experts Inspired by recent successes in Mixture-of-Experts (MoE) architectures (Jiang et al., 2024; Lin et al., 2023), we introduce CARPE-MoE. As shown in Figure 3, this extension is designed to capture more robust visual representations through an ensemble of different vision encoders.

Beyond the LVLM's base CLIP model, we add three pre-trained vision encoders as experts: SigLIP (Zhai et al., 2023a), DINOv2 (Oquab et al., 2023), and a CLIP-MoE model from the CuMo (Li et al., 2024) checkpoint. Each of the four backbones is paired with a dedicated two-layer MLP adapter, which projects its unique visual features into the LLM's common embedding space. A lightweight vision router dynamically selects the most suitable expert for a given input. To maintain our framework's context-aware nature, the routing decision is conditioned on the hidden state of the learnable context prompt token. Based on this textual context, the router performs top-1 gating to direct the image to a single, most appropriate expert.

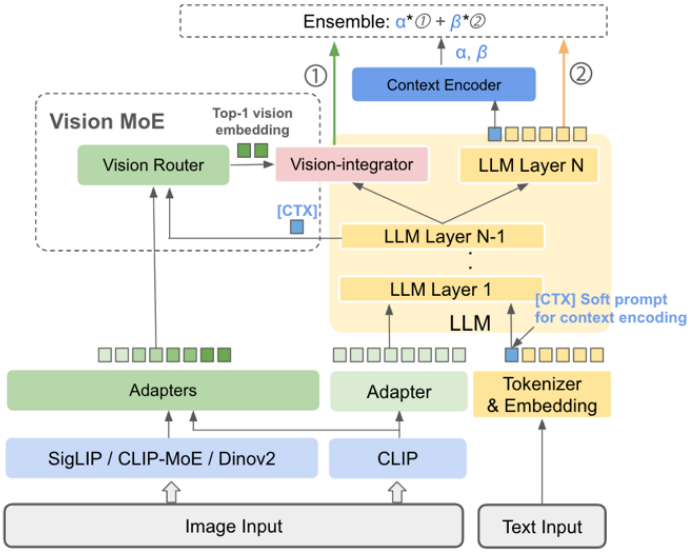


Figure 3: The architecture of CARPE-MoE. It extends the base CARPE model by incorporating a Vision MoE module that dynamically selects an expert from multiple vision backbones.

4 Experiments

4.1 Experimental Setup

We utilized LLaVA-Instruct-665K (Liu et al., 2024a), a publicly available instruction-tuning dataset, and combine it with Imagenet (Deng et al., 2009) to improve both classification ability and overall visual image comprehension. To avoid degrading language ability, we fix the mixing ratio at 1:7 (ImageNet: LLaVA-Instruct). For ImageNet prompting, we uniformly sample one of 20 classification prompt templates (i.e. 'Identify the object in this image:', 'What object can you spot in the picture?') per example. We mix open- and closed-world prompts 50/50 (without vs. with label lists) to reduce prompt overfitting and improve generalization.

In our experiments, we use LLaVA1.5-7b (Liu et al., 2024a) as our base model to evaluate our framework. During training, we keep the adapter, final output projection head, vision-integrator, context encoder and context prompt trainable while freezing all remaining parts. We train the base CARPE model for 2 epochs and the CARPE-MoE model for 3 epochs, using batch size of 64 for all experiments. We set the learning rate to 2e-5 for the adapter and 2e-4 for the other trainable components. To stabilize the learning process, we freeze the context encoder and context prompt during the first epoch and unfreeze them in subsequent epochs.

4.2 Baselines

We compare CARPE with four baselines. As a classification fine-tuning baseline, we use the ImageNet-fine-tuned LLaVA-1.5-7B checkpoint (Zhang et al., 2024). We also include two ensemble baselines, WiSE-FT (Wortsman et al., 2022) and LEVI (Roh et al., 2024). WiSE-FT (Wortsman et al., 2022) linearly interpolates the parameters of a zero-shot model and a fine-tuned model. In our setup, we mix the pre-trained LLaVA1.5-7B weights with the ImageNet-fine-tuned checkpoint (Zhang et al., 2024) using a coefficient of 0.5. LEVI (Roh et al., 2024) adaptively ensembles a pre-trained model layer-wise with a small task-specific model to improve generalization in fine-tuning. To apply LEVI to generative LVLMs, we replace the task-specific branch with adapter outputs from the vision side and attach five adapting layers to the last five LLM layers. Each adapting layer performs multi-head cross-attention with queries from the corresponding LLM hidden states and keys and values from the adapter outputs, followed by multi-head self-attention and an MLP. The five adapted hidden states are averaged and projected onto the vocabulary space, producing final logits.

Finally, we evaluate SPHINX (Lin et al., 2023) as a visually enhanced LVLM baseline that mixes model weights, training objectives, enriched visual embeddings, and high-resolution sub-images to improve overall capability.

We evaluate CARPE on four image classification benchmarks—ImageNet, Caltech101, Flowers102, and Food101—and seven vision-language benchmarks, including SQA, TextVQA , POPE , MME , MMBench , CV-Bench , and MMVP (see Appendix A.2 for details).

	Model	ImageNet	Caltech101	Flowers102	Food101	**Average**
Baseline	LLaVA1.5-7B (Liu et al., 2024a)	26.4	55.4	6.7	29.5	29.5
	+ Imagenet Fine-tune (Zhang et al., 2024)	78.0	54.4	6.7	22.9	40.5
	WiSE-FT (Wortsman et al., 2022)	48.1	56.2	6.6	26.9	34.4
	LEVI (Roh et al., 2024)	73.3	43.5	3.7	20.8	35.3
	SPHINX-13B (Lin et al., 2023)	32.3	49.7	17.7	36.3	34.0
Ours	CARPE	73.4	60.4	15.4	32.7	45.4
	CARPE-MoE	64.5	65.6	16.7	37.7	**46.1**

Table 2: Classification Accuracy (%)

	Model	General VL Benchmarks					Vision-centric VL		**Average**
		SQA	TextVQA	POPE	MME	MMBench	CV-Bench	MMVP	
Baseline	LLaVA1.5-7B (Liu et al., 2024a)	69.4	58.3	85.9	1862.7	64.7	46.0	63.0	68.6
	+ Imagenet Fine-tune (Zhang et al., 2024)	66.8	57.0	85.9	1744.2	62.1	37.9	56.6	64.7
	WiSE-FT (Wortsman et al., 2022)	68.0	57.8	85.1	1803.5	64.0	46.1	59.0	67.1
	LEVI (Roh et al., 2024)	69.4	49.2	84.5	1752.1	64.0	47.4	61.3	66.2
	SPHINX-13B (Lin et al., 2023)	69.3	51.6	80.7	1798.3	66.9	61.3	66.6	69.4
Ours	CARPE	68.4	55.8	85.2	1826.5	64.8	58.8	64.0	69.7
	CARPE-MoE	68.0	57.4	84.7	1861.7	64.0	58.8	65.0	**70.1**

Table 3: Performance on seven vision-language benchmarks, including both general-purpose and vision-centric tasks. MME scores are scaled to 100 for averaging; SQA refers to the image subset of ScienceQA; POPE is reported with F1 score; all others are accuracy.

5 Results

5.1 Classification Benchmarks

As shown in Table 2, CARPE improves performance not only on the in-distribution dataset (ImageNet) but also on all out-of-distribution (OOD) classification benchmarks compared to the base LLaVA1.5-7B (Liu et al., 2024a) model. The ImageNet fine-tuning baseline (Zhang et al., 2024) yields a substantial gain on the in-distribution dataset, but it causes a notable performance drop on OOD datasets such as Caltech101 and Food101. In contrast, both CARPE and CARPE-MoE increase accuracy across both in-distribution and all OOD datasets. Notably, CARPE-MoE achieves the highest average classification accuracy, demonstrating the benefit of integrating diverse visual representations from multiple backbones for classification tasks.

When compared to other ensemble-based baselines such as WiSE-FT (Wortsman et al., 2022) and LEVI (Roh et al., 2024), CARPE demonstrates clear superiority. Remarkably, despite SPHINX-13B (Lin et al., 2023) having nearly twice the model size, CARPE still surpasses it on three out of four benchmarks, highlighting the parameter efficiency of our design.

5.2 Vision-Language Benchmarks

Table 3 shows that both of our models outperform the baselines, with CARPE-MoE achieving the highest average performance across vision-language benchmarks. While ImageNet fine-tuning (Zhang et al., 2024) leads to severe degradation in several benchmarks, CARPE and CARPE-MoE preserve performance on general benchmarks and deliver substantial improvements on vision-centric benchmarks. In particular, compared to the base model, CARPE-MoE improves CV-Bench

and MMVP scores by 12.2% and 2.0%, respectively, whereas the fine-tuning baseline suffers large drops. Unlike LEVI, which relies on adapter outputs already projected into the language space, CARPE directly leverages raw vision features from the encoder, mitigating information loss during alignment.

These results support our hypothesis that visual features extracted from the vision encoder are partially distorted or lose generalizability during alignment with the LLM. By introducing the vision-integrator to incorporate raw vision features while retaining their original granularity and balancing them with LLM outputs in a context-dependent manner, CARPE enables more effective utilization of visual information that may be under-emphasized during LLM alignment.

Benchmark Type	**Vision Weight**	**Language Weight**
Classification	0.26	0.74
Vision-language Benchmark	0.13	0.87

Table 4: Average vision and language weights of CARPE-MoE assigned by the context encoder across classification and vision-language benchmarks.

5.3 Context-Aware Balancing of Vision and Language

Our design assumes that the optimal weighting between vision and language components depends on the task. Vision-centric tasks such as image classification should rely more heavily on raw vision features, while general vision-language tasks often require stronger utilization of the LLM's reasoning ability. To enable this adaptability, we introduced a learnable context prompt and a lightweight context encoder to dynamically adjust the ensemble weights according to the text instruction.

Table 4 shows that the vision-to-language weight ratio assigned by the context encoder differs between classification and vision-language benchmarks. Specifically, the average vision weight is 0.26 for classification benchmarks and 0.13 for vision-language benchmarks, indicating that the model allocates a relatively greater proportion of attention to vision features in classification tasks compared to vision-language tasks. This difference likely arises because classification tasks depend more heavily on precise visual recognition, whereas many vision-language benchmarks place greater emphasis on language-based reasoning and instruction-following. Such adaptive weighting reflects CARPE 's ability to adjust the balance between vision and language components according to task requirements, contributing to its strong performance across both classification and vision-language evaluations.

6 Conclusion

In this work, we addressed the challenge of enhancing the general visual understanding of LVLMs by improving their classification capability. We proposed CARPE, a lightweight vision-integration module combined with a context-aware dynamic ensemble strategy. Furthermore, we introduced CARPE-MoE, an extension that incorporates a Mixture of Experts framework to leverage multiple, diverse vision backbones.

Extensive experiments demonstrate that our methods effectively recover visual information that may be lost during LLM alignment, leading to performance gains on both classification and vision-language benchmarks. In particular, the CARPE-MoE variant demonstrate superior generalization, achieving the highest average performance across all evaluated tasks. These results confirm that improving general visual understanding also benefits broader vision-language capabilities. Finally, we showed that different tasks require different weighting of vision and language components, and by adaptively balancing the two based on context, our approach enhances the overall multimodal reasoning ability of LVLMs.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics[1]. Our research focuses on analyzing and mitigating modality imbalance in large vision-language models through a context-aware integration framework. We aim to conduct and report our research transparently, minimize potential harm, and consider the broader societal implications of multimodal systems. However, as with other LLM research, we acknowledge that there remains potential risk that such systems may generate biased or harmful outputs embedded in pretrained models. All datasets used in our experiments (e.g., ImageNet, Caltech101, Flowers102, Food101, and standard vision-language benchmarks) are publicly available and widely adopted in the research community.

REPRODUCIBILITY STATEMENT

We designed our experiments to be fully reproducible. Our implementation builds on the publicly available LLaVA codebase[2], and all experiments are conducted using publicly available pretrained vision-language models and widely adopted benchmark datasets. Experimental settings, implementation details, and hyperparameters are specified in Section 4 and the appendix. We will release the complete implementation of CARPE, including training scripts, evaluation pipelines, and configuration files upon acceptance.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models. *arXiv preprint arXiv:2505.19616*, 2025.

Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models, 2023a.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26540–26550, 2024a.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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*, pp. 24185–24198, 2024b.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf.

[1] https://iclr.cc/public/CodeOfEthics

[2] https://github.com/haotian-liu/LLaVA

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*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop*, 2004.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. URL https://arxiv.org/abs/2306.13394.

Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining, 2022. URL https://arxiv.org/abs/2204.14095.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021. URL https://arxiv.org/abs/2104.08691.

Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts, 2024. URL https://arxiv.org/abs/2405.05949.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023b. URL https://arxiv.org/abs/2305.10355.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models, 2023. URL https://arxiv.org/abs/2311.07575.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

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 (CVPR)*, pp. 26296–26306, June 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024b. URL https://arxiv.org/abs/2307.06281.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL https://arxiv.org/abs/2209.09513.

Chancharik Mitra, Brandon Huang, Tianning Chai, Zhiqiu Lin, Assaf Arbelle, Rogerio Feris, Leonid Karlinsky, Trevor Darrell, Deva Ramanan, and Roei Herzig. Sparse attention vectors: Generative multimodal model features are discriminative vision-language classifiers, 2024. URL https://arxiv.org/abs/2412.00142.

M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Yuji Roh, Qingyun Liu, Huan Gui, Zhe Yuan, Yujin Tang, Steven Euijong Whang, Liang Liu, Shuchao Bi, Lichan Hong, Ed H. Chi, and Zhe Zhao. LEVI: Generalizable fine-tuning via layer-wise ensemble of different views. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42666–42690. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/roh24a.html.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019. URL https://arxiv.org/abs/1904.08920.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models, 2022. URL https://arxiv.org/abs/2109.01903.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023a. URL https://arxiv.org/abs/2303.15343.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models, 2023b. URL https://arxiv.org/abs/2309.10313.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=MwmmBg1VYg.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Appendix

A.1 Using ensemble

To effectively combine the embeddings from the vision-integrator and the LLM representations, we employ an ensemble strategy that integrates their logits.

Formally, Let X_{txt} be the input text sequence, X_{img} the input image, and Y a target token. Let $H_{vision}, H_{llm} \in \mathbb{R}^{N \times d}$ denote the hidden representations obtained from the vision integrator and the final LLM layer, where N is the sequence length and d is the hidden dimension. The shared output projection to the vocabulary is denoted by $W_{head} \in \mathbb{R}^{V \times d}$, where V is the vocabulary size.

We first compute the logits from each representation as follows:

$$Z_{vision} = H_{vision} W_{head}^T$$

$$Z_{llm} = H_{llm} W_{head}^T$$

To determine context-aware ensemble weights, we introduce a learnable context prompt token, which is appended to the input and processed by the LLM. Let $H_{context} \in \mathbb{R}^d$ denote the final hidden state of the context prompt token, and let $W_{context} \in \mathbb{R}^{2 \times d}$ be the context encoder that projects this hidden state to a two-dimensional weight vector.

The ensemble weights are computed as:

$$\alpha, \beta = \text{Softmax}(W_{context} H_{context}^T)$$

Using these weights, the final logit is computed as a weighted sum:

$$Z = \alpha \cdot Z_{vision} + \beta \cdot Z_{llm}$$

A.2 Evaluation datasets

To validate the effectiveness of CARPE, we evaluated the models on four classification datasets and seven vision-language benchmarks. The classification datasets include ImageNet (Deng et al., 2009), Caltech101 (Fei-Fei et al., 2004), Flower102 (Nilsback & Zisserman, 2008), and Food101 (Bossard et al., 2014). Flowers102 (Nilsback & Zisserman, 2008) and Food101 (Bossard et al., 2014) comprise 102 and 101 categories, respectively, and are used to evaluate fine-grained visual understanding. Caltech101 (Fei-Fei et al., 2004), comprises 101 object classes and serves as a benchmark for assessing classification performance at a semantic level.

The vision-language benchmarks includes two academic-task-oriented datasets: image subset of ScienceQA (Lu et al., 2022) and TextVQA (Singh et al., 2019), which evaluates zero-shot generalization on scientific question answering and text-rich visual question answering, respectively. Benchmarks for instruction-following LVLMs include five datasets: POPE (Li et al., 2023b), MME (Fu et al., 2024), MMBench (Liu et al., 2024b) CV-Bench (Tong et al., 2024a), and MMVP (Tong

et al., 2024b). These benchmarks assess various LVLMs abilities, including object hallucination, OCR perception, language generation, mathematical reasoning, scene understanding, and object counting. In particular, CV-Bench and MMVP are vision-centric benchmarks. CV-Bench (Tong et al., 2024a), evaluates classic vision tasks in multimodal settings, including spatial relations and depth ordering. MMVP (Tong et al., 2024b) targets CLIP-blind pairs, where CLIP judges visually distinct images as similar.