Multi-Perspective LVLM Prompting for Robust Zero-Shot Image Classification

Wonjun Choi, Jeong-Cheol Lee, Jun-Hyeok Seo, Dong-Gyu Lee
Department of Artificial Intelligence, Kyungpook National University
Daegu, South Korea
{sangju, jclee2716, ssam2s, dglee}@knu.ac.kr

Abstract

In this paper, we present a multi-perspective prompt-based framework for zero-shot image classification using LVLMs. Real-world image classification often faces challenges that hinder accurate classification, particularly when images contain multiple objects or ambiguous intent. Our method integrates complementary viewpoints through novel prompting strategies without task-specific fine-tuning. The framework consists of prediction and voting stage. In the prediction stage, keywords are generated through dialogues with CLIP-filtered candidate classes, followed by LVLM predictions. In the voting stage, these predictions are aggregated through majority voting. These two stages collectively yield the final class prediction. Experimental results show our intent-aware prompt achieves the highest single-model accuracy, while ensemble methods improve performance to 67.65% on the test-dev set.

1. Introduction

Large Vision-Language Models (LVLMs) [2, 4] are trained on large-scale image-text pairs, enabling them to effectively align visual and linguistic information. These models can perform zero-shot classification without task-specific fine-tuning. A common approach is to provide the model with a textual description of the image with a set of candidate classes, enabling it to match the image to predefined classes. However, tasks such as the VizWiz-Classification [1], which contain images with multiple objects or ambiguous intent, require reasoning over multiple possible interpretations to determine the most reasonable answer.

To address these challenges, we introduce a multi-perspective prompt-based framework for zero-shot image classification using LVLMs, designed to facilitate reasoning over diverse aspects of the image through prompting strategies. We design object-aware, intent-aware, and reason-aware prompts, allowing the model to perform classification from diverse perspectives and integrate these perspectives for improved performance. We demonstrate the effectiveness of our method with results on the test-dev set.

2. Method

2.1. Framework

Our proposed framework is illustrated in Figure 1, comprising two stages: prediction and voting. In the prediction stage, multi-perspective prompts direct the LVLM to analyze the image and extract relevant information. The model generates a sentence containing multiple keywords based on the resulting dialogue, which is subsequently compared with a predefined set of class labels using the CLIP [3] text encoder. The top 20 semantically relevant candidate classes are identified. The final phase of this stage involves the identification of the optimal class from the candidate set, which is subsequently established as the prediction. In the voting stage, predictions derived from multiple prompts are aggregated. When at least two prompts predict the same class, that class is designated as the final prediction. If all prompts yield different predictions, CLIP is utilized to recalculate the image-text similarity for each predicted class, and the class achieving the highest similarity score is selected as the final output.

2.2. Prompt Design

Object-aware prompt instructs the model to generate descriptions by considering transformations such as distortions and abstract patterns, directing the model to focus on key objects within the image. It extracts candidate classes based on these considerations subsequently.

Intent-aware prompt instructs the model to generate descriptions based on photographic compositional cues, assuming the image was intentionally captured by a photographer. The integration of intent analysis allows the model to prioritize visually salient and contextually significant objects to extract the most probable subjects of interest as candidate classes.

Reason-aware prompt instructs the model to integrate previous object-aware and intent-aware predictions to guide the final decision-making process. The model utilizes candidate classes and predictions to generate reason and subsequently predicts the most reasonable class.

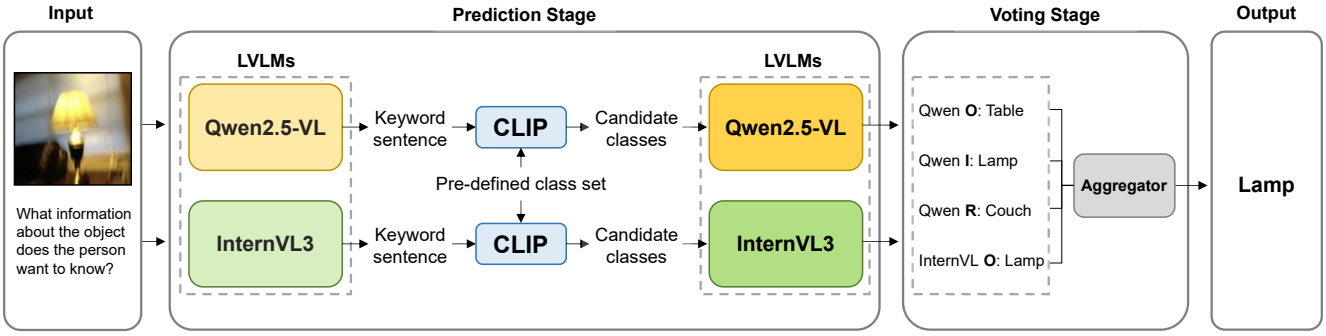

Figure 1. The framework of the proposed method. O, I, and R denote object-aware, intent-aware, and reason-aware prompts, respectively.

Method	Prompt	VizWiz								
		BLR	BRT	FRM	ROT	OBS	DRK	Corr	Clean	Total
Qwen2.5-VL-7B	O	58.40	59.49	60.15	56.48	64.29	59.09	59.65	65.89	62.70
Qwen2.5-VL-7B	I	60.50	63.29	61.09	56.81	57.14	63.64	60.41	65.71	63.45
Qwen2.5-VL-7B	R	60.22	63.29	60.53	57.81	50.0	56.82	58.11	64.64	62.35
InternVL3-8B	O	60.78	67.08	63.16	53.49	64.29	**72.73**	63.59	64.46	62.7
Ensemble (Qwen)	O+I+R	63.45	65.82	64.47	61.46	59.52	62.5	62.87	69.46	66.85
Ensemble (Total)	O+I+R	**64.57**	**68.35**	**65.88**	**61.79**	**66.67**	67.05	**65.72**	**70.18**	**67.65**

Table 1. Experimental results on VizWiz test-dev. O, I, and R denote object-aware, intent-aware, and reason-aware prompts, respectively. Used evaluation metric is accuracy (%).

3. Experiments

In this paper, we evaluate LVLMs (Qwen2.5-VL-7B [2] and InternVL3-8B [4]) and prompting strategies for the VizWiz-Classification. Table 1 shows the experimental results on the VizWiz-Classification test-dev.

Considering individual prompting methods, Qwen2.5-VL-7B with the intent-aware prompt achieved the highest overall accuracy of 63.45%, which means incorporating the intent may be more effective than focusing solely on visual object cues. In contrast, reason-aware produced the lowest performance among the three prompting strategies, with an accuracy of 62.35%. A particularly noteworthy finding is the variation in performance across different image conditions, depending on the prompting strategy. The object-aware prompt performed best under obstructions (OBS, 64.29%), whereas the intent-aware prompt excelled under too bright (BRT, 63.29%). Meanwhile, the reason-aware prompt demonstrated relatively better performance in rotated views (ROT, 57.81%). These findings indicate that multi-perspective prompting strategies are complementary across different image conditions.

To exploit their complementary strengths, we employ an ensemble method that consolidates the outputs of various prompts and models. The ensemble of three prompts using the Qwen2.5-VL-7B model achieved an accuracy of 66.85%. Furthermore, the complete ensemble combining Qwen2.5-VL-7B and InternVL3-8B yielded the highest accuracy of 67.65% with an improvement of approximately 4.2% over the best performing single model. The ensemble outperformed individual models in the final score, highlighting its robustness for real-world tasks such as VizWiz.

4. Conclusion

We propose a multi-perspective prompting framework for zero-shot image classification, addressing challenges in the VizWiz-Classification dataset. The combination of object-aware, intent-aware, and reason-aware prompts improves the robustness and generalizability of method. Results show that the multi-perspective approach provides complementary contributions, and their ensemble enhances performance. Future work includes incorporating diverse LVLMs to expand model-level perspectives.

Acknowledgements

This work was supported by the IITP(Institute of Information & Coummunications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT) (IITP-2025-RS-2020-II201808, 50%), (IITP-2025-RS-2024-00437718, 50%)

References

[1] Reza Akbarian Bafghi and Danna Gurari. A new dataset based on images taken by blind people for testing the robustness of image classification models trained for imagenet categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16261–16270, 2023. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2

[3] 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. 1

[4] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1, 2