# Integrating Query-aware Segmentation and Cross-Attention for Robust VQA

Wonjun Choi, Sangbeom Lee, Seungyeon Lee, Heechul Jung, Dong-Gyu Lee
Department of Artificial Intelligence, Kyungpook National University
Daegu, South Korea
{sangju, sb486868, statai3237, heechul, dglee}@knu.ac.kr

## Abstract

*This paper introduces a method for VizWiz-VQA using LVLM with trainable cross-attention and LoRA finetuning. We train the model with the following conditions: 1) Training with original images. 2) Training with enhanced images using CLIPSeg to highlight or contrast the original image. 3) Training with integrating the output features of Vision Transformer (ViT) and CLIPSeg features of the original images. Then, we ensemble the results based on Levenshtein distance to enhance the prediction of the final answer. In the experiments, we demonstrate and analyze the proposed method's effectiveness.*

## 1. Introduction

Visual Question Answering (VQA) tasks integrate vision and language, requiring a model to understand and answer questions about a given image. The conventional approaches in VQA have relied on direct mappings between visual content and question-answer pairs, often skipping the nuanced interplay between image-specific details and the context of the questions [4]. Enhancing the effectiveness of VQA models necessitates attention to global image features and localized, relevant visual cues that directly pertain to the questions asked. The VizWiz-VQA [2] focus on the parts of the image that correspond to the question's intent is crucial. In this paper, we introduce a parameter-efficient method using the Low-rank Adaptation(LoRA) and a cross-attention mechanism between text and visual input based on a large vision language model (LVLM). In addition, we integrate an image segmentation technique, called CLIPSeg [5] to improve visual understanding. We show results using validation/test-dev sets to demonstrate the effectiveness of our method.

## 2. Method

**Trainable cross-attention and LoRA tuning.** We use the Qwen-VL [1] model, which consists of ViT and QwenLM, as a backbone and set the cross-attention module to a learnable state to effectively learn specific parts of the image
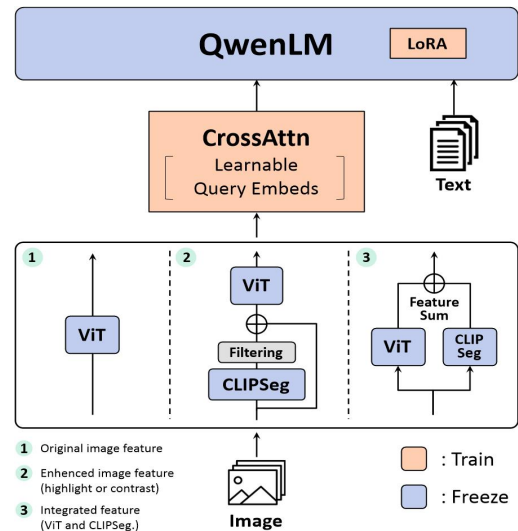


Figure 1. Our proposed fine-tuning using LoRA and cross-attention (FT-L/CA) method and three types of ViT configurations.

associated with the keywords in the question. The LoRA adapter [3], which has high parameter efficiency, is applied to fine-tune the QwenLM. In this configuration, only the cross-attention and LoRA application parts are set as trainable parameters, and the rest of the model remains with fixed parameters.

**Image segmentation for enhancement.** We use a query-aware segmentation feature, using the CLIPSeg model [5], which identifies and separates specific objects within an image based on text queries. CLIPSeg enhances image processing by dynamically isolating related objects and integrating them into the model. We enhance this with segmentation techniques highlighting specific image areas and contrast methods differentiating between masked and unmasked regions, providing more apparent visual distinctions.

**Ensemble using Levenshtein distance.** In our proposed method, we employ an ensemble of various predictive models trained to enhance accuracy. For the final answer generation, we utilize a method based on the Levenshtein distance to select the most accurate prediction. This approach

| Method | yes/no | number | other | unanswerable | overall |
|---|---|---|---|---|---|
| Zero-shot (original image, w/ instruct) | 76.87 | 50.42 | 56.26 | 93.45 | 69.05 |
| Zero-shot (segment-highlight, w/ instruct) | 64.15 | 42.50 | 49.83 | **97.23** | 65.59 |
| Zero-shot (segment-contrast, w/ instruct) | 67.08 | 46.88 | 53.08 | 95.96 | 67.39 |
| FT-L/CA | 81.23 | 52.71 | 58.53 | 93.60 | 70.73 |
| FT-L/CA (segment-highlight image) | 79.74 | 51.88 | 54.53 | 92.32 | 67.75 |
| FT-L/CA (segment-contrast image) | 79.64 | 48.96 | 57.16 | 93.13 | 69.62 |
| FT-L/CA + Segmentation | **81.59** | **56.25** | 58.84 | 91.99 | 70.47 |
| Ensemble | 81.38 | 52.71 | **60.30** | 94.10 | **72.01** |

Table 1. Experimental Results on validation set of VizWiz-VQA. FT-L/CA is that fine-tuning with LoRA and cross-attention. The ensemble result is perfomed using fine-tuned models.

| Method | yes/no | number | other | unanswerable | overall |
|---|---|---|---|---|---|
| FT-L/CA | 83.37 | **55.08** | 58.78 | 94.81 | 69.78 |
| FT-L/CA (segment-highlight image) | 81.07 | 49.05 | 54.49 | 93.07 | 66.25 |
| FT-L/CA (segment-contrast image) | 78.76 | 48.10 | 56.63 | 93.45 | 67.66 |
| FT-L/CA + Segmentation | 81.85 | 50.32 | 56.89 | 87.52 | 66.36 |
| Ensemble | **83.60** | 53.49 | **60.57** | **94.88** | **70.97** |

Table 2. Experimental Results on test-dev set of VizWiz-VQA.

leverages the edit distance metric to compare and evaluate the outputs of different models, thereby optimizing the selection of the final answer.

## 3. Experiments

**Experimental results.** Our experiment is conditioned under three conditions, as shown in Figure 1. (1) Input the original image into Qwen-VL's ViT(FT-L/CA). (2) Input the original image with two variants segmentation (highlight, contrast). The segment-highlight adds the segmentation output to the original image, and segment-contrast adjusts the brightness of the original image other than segmentation output. (3) Integrate ViT features and CLIPSeg features in the model. In our experiments, we perform experiments on zero-shot and fine-tuning. In the zero-shot setting, we confirmed without instructions, performance deteriorates significantly, thus we add and use instructions*. As shown in Table 1 and Table 2, integrating feature information directly in the model proves more effective than applying CLIPSeg output to the input image and transforming it, and particularly, the ensemble model yields the highest performance. The query-aware CLIPSeg actively improves correct answer rates across multiple categories.

## 4. Conclusion

In this paper, we present a method using LVLM with trainable cross-attention and LoRA tuning. We also use query-aware CLIPSeg to enhance focus on relevant image

---

*Instructions:* " The correct answer type is one of ['number', 'words', 'yes', 'no']. If it is impossible to answer an image-related question or there is no existing information, please reply as 'unanswerable'. Question: "

details. Our method ranked 3rd (team: KNU-HomerunBall) in the VizWiz challenge public leaderboard, demonstrating improved results through further experiments and ensembles.

## Acknowledgements

## References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1

[2] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 1

[3] 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. In *International Conference on Learning Representations*, 2022. 1

[4] Siyu Lu, Yueming Ding, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. Multiscale feature extraction and fusion of image and text in vqa. *International Journal of Computational Intelligence Systems*, 16(1):54, 2023. 1

[5] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096, 2022. 1