# System Report for CCL25-Eval Task 11: Enhancing Chinese Character Handwriting Evaluation with Multimodal Large Language Models

**Xiaoqing Hong***
School of Data Science and Engineering
East China Normal University
Shanghai, China
xqhong@stu.ecnu.edu.cn

**Yunhan Li***
School of Data Science and Engineering
East China Normal University
Shanghai, China
yhli@stu.ecnu.edu.cn

**Lyu Ni**†
School of Data Science and Engineering
East China Normal University
Shanghai, China
lni@dase.ecnu.edu.cn

## Abstract

With the development of smart devices, students' ability to handwrite Chinese characters has generally been decreasing. Chinese character handwriting receives increasing attention because the standardization of Chinese character handwriting is one of the most important components of national education in China. Due to inadequate professional teachers and labor-intensive evaluation means, it is difficult to provide large-scale, personalized, and low-latency evaluation feedback in Chinese character handwriting education. Recently, large language models (LLMs) have made outstanding achievements in natural language understanding and generation. Thus, the multimodal large language model(MLLM) is an efficient method to resolve the difficulties. We introduce an enhanced neural network architecture, referred to as ACBAM-VGG16, which is developed by augmenting the CBAM-VGG16 framework with adversarially generated examples. Leveraging this model, we propose customized training and inference mechanisms for MLLMs, specifically targeting two downstream tasks: quality assessment of handwritten Chinese character images and generation of descriptive textual comments. We introduce an effective inference strategy that allows an MLLM to maintain high performance in scenarios where limited training data are available for model fine-tuning, resulting in the average $F_1$ score can be improved by 6.74%. Moreover, we design a hierarchical MLLM fine-tuning framework to ensure the precision and diversity of generated comments. In the comparison of various MLLMs, the proposed framework increases the weighted average of ROUGE-1, ROUGE-2, and ROUGE-L by 2.33%-9.94%.

**Keywords**: Evaluation of handwriting Chinese characters, Multimodal large language model, Deep learning, Comment generation

## 1 Introduction

As the nationally standardized and widely used writing system, standard Chinese characters play a central role in education. Consequently, the ability to handwrite these characters is considered a foundational skill for students at the primary and secondary levels and represents a core element of contemporary national educational curricula (Office of the Ministry of Education, 2024). The widespread use of smart devices has been associated with a gradual decline in student proficiency in handwritten Chinese characters. The significance of standardized Chinese character writing education has recently gained full recognition. However, Chinese character handwriting education faces challenges due to a limited number of teachers and labor intensive evaluation approaches, making it difficult to achieve large-scale, personalized, and low-delay evaluation feedback, thus failing to improve students' handwriting skills.

Large language models (LLMs) have powerful capabilities in natural language understanding and generation. The objective of this paper is to explore the capabilities of Multimodal Large

Xiaoqing Hong and Yunhan Li contribute equally to this work.
Lyu Ni is the corresponding author.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 437-443, Jinan, China, August 11-14, 2025.

Language Models (MLLMs) in performing image comprehension and generating descriptive comments. To be specific, this paper concentrates on the following two specific subtasks.

**Subtask 1: To Rate a Handwritten Chinese Character** It aims to develop a scoring model designed to classify images of handwritten Chinese characters. Each handwritten Chinese character is divided into three grades: *Ace* (A), *Borderline* (B) and *Crude* (C).

**Subtask 2: To Generate Comments for Handwritten Chinese Characters** It focuses on building MLLMs that generate textual feedback resembling that of professional teachers to assess handwritten Chinese characters.

To resolve the two subtasks, we introduce innovative fine-tuning and inference techniques for Multimodal Large Language Models (MLLMs), leveraging a pre-trained neural network as a guiding framework. In the first stage, we construct the ACBAM-VGG16 model by integrating adversarial training into the CBAM-VGG16 framework, which leads to significantly improved performance in assessing the quality of handwritten Chinese characters. In the second phase, we introduce an ensemble inference approach that integrates the predicted scores from both the fine-tuned Multimodal Large Language Model and the ACBAM-VGG16 model, which effectively addresses the issue of insufficient training data when fine-tuning multimodal large language models. In the final phase, we implement a hierarchical fine-tuning approach for MLLMs, where models are fine-tuned separately for each grade's handwritten Chinese characters. A Large Language Model is also fine-tuned to combine comments from multiple MLLMs, generating more diverse and informative feedback—especially for borderline cases.

## 2 ACBAM-VGG16: a Deep Neural Network for Handwritten Chinese Character Image Assessment

### 2.1 Methodology

The commonality between the two sub-tasks lies in using handwritten Chinese character images as input. Convolutional neural networks are vital for image recognition, VGG16 being a popular choice (Simonyan and Zisserman, 2015). The VGG model captures complex features in an image by stacking convolutional blocks of different sizes，which can capture local features in handwritten images of Chinese characters. Furthermore, CBAM-VGG16 optimizes the network architecture by introducing an attention mechanism to the convolutional blocks (CBAM)(Praharsha and Poulose, 2024).

FGSM (Goodfellow et al., 2015) is a classic technique used to generate adversarial examples. It adds signed perturbations along the direction of the gradient of the loss function on the input,

$$x^{adv} = x + \epsilon \cdot \text{sgn}(\nabla x J(\theta, x, y)), \quad (1)$$

where $x$ denotes the original input sample, $\epsilon$ is a hyperparameter controlling the intensity of the perturbation, $J(\theta, x, y)$ is the loss function, and $\nabla_x J(\theta, x, y)$ denotes the gradient of the loss function over the input. The method is capable of generating effective perturbation samples at a low computational cost, thereby enhancing the model's robustness against input perturbations.

We combine the FGSM-based adversarial sample generator with CBAM-VGG16, denoted as ACBAM-VGG16 (shown in Figure 1), in order to significantly increase model accuracy. By incorporating these adversarial samples into the training process, the CBAM-VGG16 network not only enhances its robustness against small perturbations but also improves its generalization performance to a certain extent. The network becomes more resilient to variations in handwriting styles in realistic settings, leading to more stable and reliable quality ratings for Chinese character handwriting.

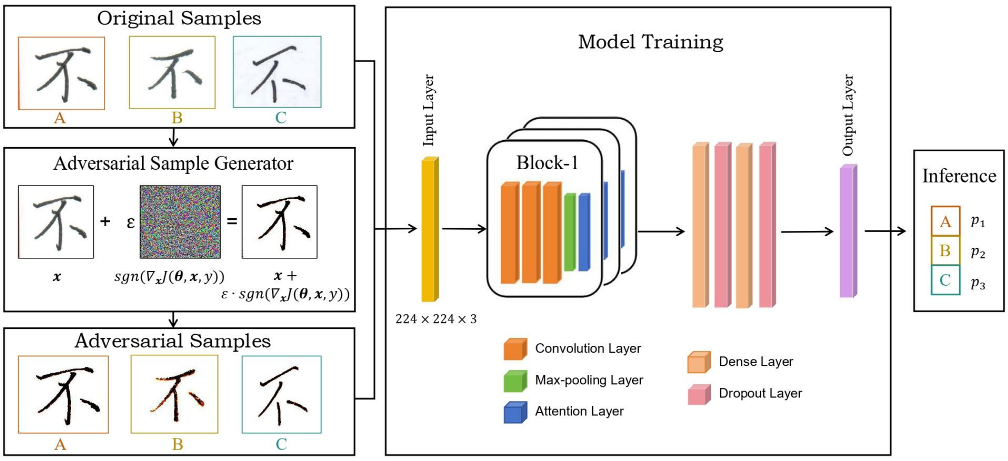


Figure 1: The framewrok of ACBAM-VGG16

### 2.2 Experimental Results

We compare ACBAM-VGG16 with some frequently-used convolutional neural networks, attention mechanisms and loss functions. The five-fold cross-validation results are shown in Table 1. Compared with other models, the ACBAM-VGG16 outperforms the others in terms of prediction result $F_1$-scores at each grade. Besides, we validate the effectiveness of the CBAM attention module and adversarial sample generation separately. Compared to models with SE attention and those without any attention mechanisms, the CBAM attention module improved performance by 4.42% and 6.97%, respectively. In comparing focal loss and categorical cross-entropy loss for the loss function, the latter demonstrated a 1.35% improvement over the former. After adding the adversarial sample generation model training method, the improvement is 2.40%.

| Model | Accuracy | Categories | Precision | Recall | $F_1$ score |
|---|---|---|---|---|---|
| VGG16 | 86.00 | A | 89.13 | 82.00 | 85.42 |
| | | B | 82.86 | 87.00 | 84.88 |
| | | C | 86.41 | 89.00 | 87.69 |
| VGG16 with SE attention | 88.67 | A | 97.53 | 79.00 | 87.29 |
| | | B | 83.81 | 88.00 | 85.85 |
| | | C | 86.84 | 99.00 | 92.52 |
| CBAM-VGG16 | 93.00 | A | 95.00 | 95.00 | 95.00 |
| | | B | 91.67 | 88.00 | 89.80 |
| | | C | 92.31 | 96.00 | 94.12 |
| CBAM-VGG16 with focal loss | 91.97 | A | 90.57 | 96.00 | 93.21 |
| | | B | 89.47 | 85.00 | 87.18 |
| | | C | 94.95 | 94.00 | 94.47 |
| **ACBAM-VGG16** | 95.33 | A | 97.92 | 94.00 | **95.92** |
| | | B | 90.57 | 96.00 | **93.21** |
| | | C | 97.96 | 96.00 | **96.97** |

Table 1: Comparison with ACBAM-VGG16 with other models (the **bold** is the best result)

## 3 ACBAM-VGG16 Enhancing Inference Approach to Multimodal Large Language Models in Handwriting Chinese Character Images

Multimodal Large Language Models (MLLMs) offer robust representations that effectively capture the nuances of both visual and textual data. Based on the images of handwritten

Chinese characters, we use LoRA (Hu et al., 2022) to fine-tune a MLLM (e.g., Qwen2.5-VL-3B-Instruct). In comparison to ACBAM-VGG16, the fine-tuned MLLM demonstrates inferior performance in evaluation tasks since the number of images are insufficient. We would like to consider ACBAM-VGG16 as a domain expert to improve the inference of the MLLM. Then, we design a weighted-sampling-based ensemble approach as shown in Figure 2.

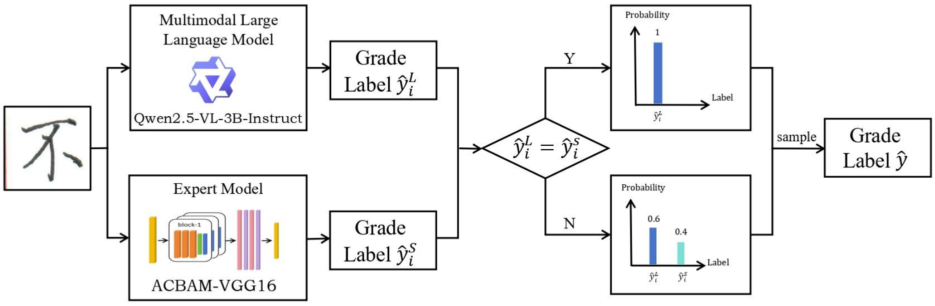


Figure 2: Weighted-sampling-based ensemble approach to generating image grades

Specifically, we obtain two grade labels, denoted as $\hat{y}_i^L$ and $\hat{y}_i^S$, respectively from a MLLM and an expert model, ACBAM-VGG16. When two predicted labels agree, i.e. $\hat{y}_i^L = \hat{y}_i^S$, we construct a degenerate probability model as $P(y = \hat{y}_i^L) = 1$. Thus, the ensemble grade label is actually $\hat{y}_i^L$. When the two predicted labels agree do not agree, i.e., $\hat{y}_i^L \neq \hat{y}_i^S$, we construct a Bernoulli model

$$P(y = \hat{y}_i^L) = p_1, P(y = \hat{y}_i^S) = p_2, \text{where } p_1 + p_2 = 1. \tag{2}$$

The ensemble label is sampled from this Bernoulli distribution, $p_1$ and $p_2$ are two hyperparameters. The adjusted accuracies of the validation set are used as the hyperparameters. In practice, $\alpha_1$ and $\alpha_2$ are respectively the accuracies of a MLLM and ACBAM-VGG16, and then $p_1 = \alpha_1/(\alpha_1 + \alpha_2)$, and $p_2 = 1 - p_1$.

Next, we verify the effectiveness of the proposed ACBAM-VGG16 enhancing inference approach for a MLLM and choose Qwen2.5VL-3B-instruct as an example. In the Table 2, the ensemble label is more accurate. With the assistance of ACBAM-VGG16, the accuracy rate increase by 7.33% and the $F_1$ scores has an improvement of 4.49% - 8.80%.

| Model | Accuracy | Categories | Precision | Recall | $F_1$ score |
|---|---|---|---|---|---|
| Without | 78.67 | A | 82.11 | 78.00 | 80.00 |
| | | B | 77.59 | 72.00 | 74.69 |
| | | C | 78.43 | 88.00 | 82.94 |
| With | **86.00** | A | **90.80** | **79.00** | **84.49**(↑4.49) |
| | | B | **81.13** | **86.00** | **83.49**(↑8.80) |
| | | C | **86.92** | **93.00** | **89.86**(↑6.92) |

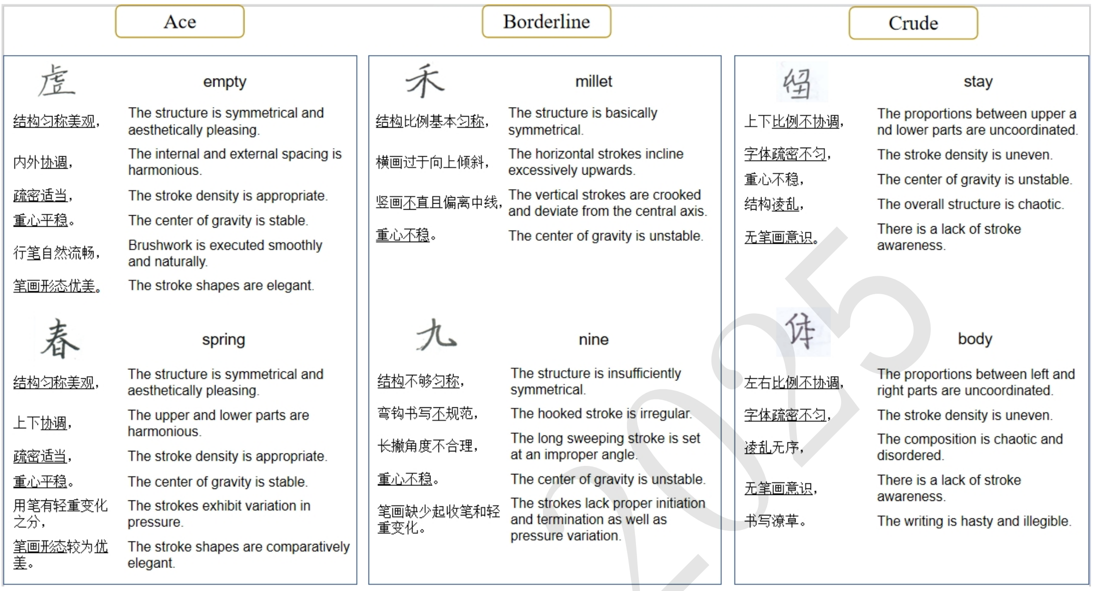
Table 2: Experimental results of Qwen2.5-VL-3B-Instruct with and without the assistance of ACBAM-VGG16

## 4 ACBAM-VGG16 Enhancing Fine-tuning Strategy for Multimodal Large Language Models in Chinese Character Comment Generation

### 4.1 Hierarchical Fine-tuning Strategies for Multimodal Large Language Models

We would like to fine-tune a multimodal large language model to generate comments for handwritten Chinese character images. LoRA(Hu et al., 2022) is a frequently-used effective fine-tuning approach for large language models, which only takes advantage of a few trainable parameters without changing the weights of the original model. Compared with some other fine-tuning methods, such as Adapter Tuning(Houlsby et al., 2019) and Prefix Tuning(Li and Liang, 2021), LoRA markedly decreases the demands on computational resources and storage capacity, which satisfies the resource requirements of generating comments on handwritten Chinese characters.

It is obvious that the comments from handwritten Chinese characters of a certain grade are similar as shown in Figure 3. Specifically, A-level handwritten characters have the most similar comments, C-levels are next, and B-levels are the least similar. Thus, we utilize ACBAM-VGG16 to classify the handwritten Chinese character image into three grades and independently fine-tune MLLMs for the training data on each grade. The similarity of comments across samples in A-level and C-level enables accurate and coherent feedback generation.



Ace

empty

| | |
|---|---|
| 结构匀称美观， | The structure is symmetrical and aesthetically pleasing. |
| 内外协调， | The internal and external spacing is harmonious. |
| 疏密适当， | The stroke density is appropriate. |
| 重心平稳。 | The center of gravity is stable. |
| 行笔自然流畅， | Brushwork is executed smoothly and naturally. |
| 笔画形态优美。 | The stroke shapes are elegant. |

spring

| | |
|---|---|
| 结构匀称美观， | The structure is symmetrical and aesthetically pleasing. |
| 上下协调， | The upper and lower parts are harmonious. |
| 疏密适当， | The stroke density is appropriate. |
| 重心平稳。 | The center of gravity is stable. |
| 用笔有轻重变化之分， | The strokes exhibit variation in pressure. |
| 笔画形态较为优美。 | The stroke shapes are comparatively elegant. |

Borderline

millet

| | |
|---|---|
| 结构比例基本匀称， | The structure is basically symmetrical. |
| 横画过于向上倾斜， | The horizontal strokes incline excessively upwards. |
| 竖画不直且偏离中线， | The vertical strokes are crooked and deviate from the central axis. |
| 重心不稳。 | The center of gravity is unstable. |

nine

| | |
|---|---|
| 结构不够匀称， | The structure is insufficiently symmetrical. |
| 弯钩书写不规范， | The hooked stroke is irregular. |
| 长撇角度不合理， | The long sweeping stroke is set at an improper angle. |
| 重心不稳。 | The center of gravity is unstable. |
| 笔画缺少起收笔和轻重变化。 | The strokes lack proper initiation and termination as well as pressure variation. |

Crude

stay

| | |
|---|---|
| 上下比例不协调， | The proportions between upper a nd lower parts are uncoordinated. |
| 字体疏密不匀， | The stroke density is uneven. |
| 重心不稳， | The center of gravity is unstable. |
| 结构凌乱， | The overall structure is chaotic. |
| 无笔画意识。 | There is a lack of stroke awareness. |

body

| | |
|---|---|
| 左右比例不协调， | The proportions between left and right parts are uncoordinated. |
| 字体疏密不匀， | The stroke density is uneven. |
| 凌乱无序， | The composition is chaotic and disordered. |
| 无笔画意识， | There is a lack of stroke awareness. |
| 书写潦草。 | The writing is hasty and illegible. |

Figure 3: Some examples of different-grade comments

In particular, the comments of B-level handwritten Chinese characters are diverse with fine-grained descriptions and a fine-tuned MLLM hardly offers the desired comments. Thus, we first fine-tune two MLLMs and derive two comments and then fuse them to fine-tune a large language model to carry out a fine-grained alignment in the textual domain. In the subsequent experiments, we choose Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct as two MLLMs, and Qwen2.5-7B-Instruct as a LLM.

We adopt a distinct fine-tuning approach for different-grade handwritten Chinese character images. Therefore, this method is called the **hierarchical fine-tuning strategy**, see Figure 4.

### 4.2 Experimental Results

To validate the effectiveness of the hierarchical fine-tuning strategy, we compare three fine-tuning approaches: (1) fine-tuning a single MLLM across all grades, (2) independently fine-tuning separate MLLMs for each grade, and (3) fine-tuning MLLMs using the proposed hierarchical strategy. Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-3B-Instruct and LLaVA-1.5-7B-Chat are considered as the reference MLLMs. We calculate three metrics ROUGE-1, ROUGE-2, and ROUGE-L based on the generated comments and define a composite score as

$$\text{score} = 0.4 * \text{ROUGE-L} + 0.3 * \text{ROUGE-2} + 0.3 * \text{ROUGE-1}.$$

The experimental results are shown in the Table3.

When compared to a uniformly fine-tuned MLLM, the use of independently fine-tuned MLLMs results in composite score increases between 1.17% and 11.38%, indicating improved perfor-

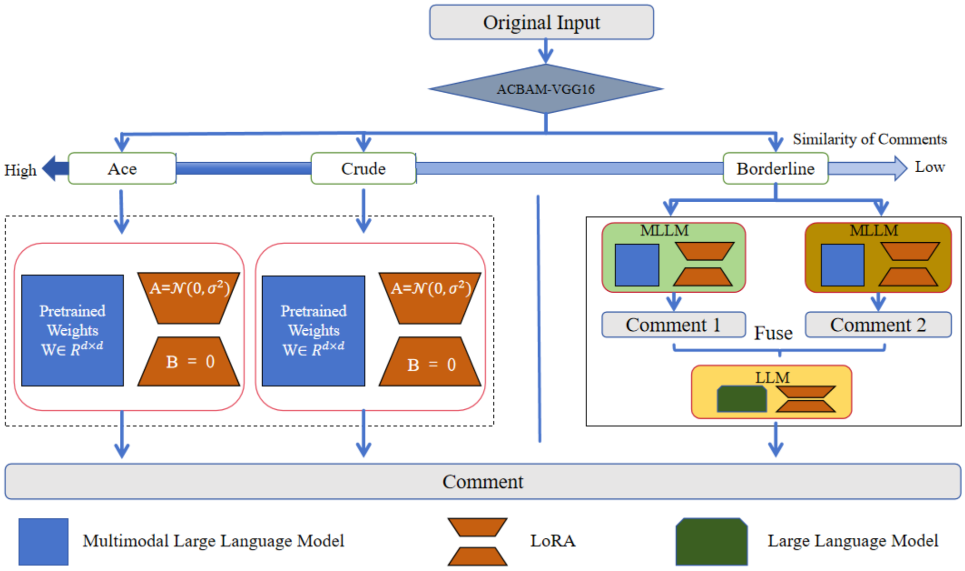

Original Input
ACBAM-VGG16
Similarity of Comments
High
Ace
Crude
Borderline
Low
Pretrained Weights W∈ $R^{d\times d}$
A=$\mathcal{N}(0,\sigma^2)$
B = 0
Pretrained Weights W∈ $R^{d\times d}$
A=$\mathcal{N}(0,\sigma^2)$
B = 0
MLLM
MLLM
Comment 1
Fuse
Comment 2
LLM
Comment
Multimodal Large Language Model
LoRA
Large Language Model

Figure 4: The hierarchical fine-tuning strategy for MLLMs and a LLM in comment generation

mance through specialized model tuning. In particular, different from LLaVA-1.5-7B-Chat, both Qwen2.5-VL-7B-Instruct and Qwen2.5-VL-3B-Instruct have obvious progress in the composite score. Thus, we choose Qwen2.5-VL-7B-Instruct and Qwen2.5-VL-3B-Instruct as the candidates of multimodal large language models, since the generated comments are highly qualified. The proposed hierarchical fine-tuning strategy achieves improvements in the composite score of 1.17% to 12.11% over the independent fine-tuning approach, indicating a notable enhancement in the coherence between the generated comments and the ground truth.For specific parameters of each model in this strategy, please refer to the appendix4.

| **Strategy** | **Base model** | **ROUGE-1** | **ROUGE-2** | **ROUGE-L** | **score** |
|---|---|---|---|---|---|
| Identical | Qwen2.5-VL-7B-Instruct | 55.74 | 40.34 | 56.28 | 51.34 |
| | Qwen2.5-VL-3B-Instruct | 55.14 | 40.40 | 55.48 | 50.85 |
| | LLaVA-1.5-7B-Chat | 61.31 | 45.64 | 55.23 | 54.18 |
| Independent | Qwen2.5-VL-7B-Instruct | 67.22 | 53.53 | 66.23 | 62.72 |
| | Qwen2.5-VL-3B-Instruct | 66.79 | 54.41 | 66.51 | 62.96 |
| | LLaVA-1.5-7B-Chat | 62.03 | 47.08 | 56.53 | 55.35 |
| Hierarchical | / | **69.24** | **57.11** | **68.46** | **65.29** |

Table 3: Experimental results of Task 2

## 5 Conclusion

In this paper, we conduct a study on the automatic evaluation of handwritten Chinese character images from two perspectives. Firstly, we construct a deep network model ACBAM-VGG16 to classify a handwritten Chinese character into three grades, including Ace, Borderline and Crude. Second, we implement fine-tuning of a multimodal large language model and introduce an ensemble inference approach supervised by ACBAM-VGG16, which enhances classification performance in scenarios with limited image data. Finally, we propose a hierarchical fine-tuning strategy that integrates two MLLMs and a LLM, tailored for handwritten Chinese character images across different grades determined by ACBAM-VGG16, with the aim of generating high-quality and teacher-style comments. The deployment of multiple large language models frequently incurs significant computational inefficiencies during inference, making their enhancement a critical yet challenging research problem.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China (62307014). We are very grateful to the reviewers and editors for their review and valuable comments on our paper. We have carefully read and fully adopted the review comments, and have revised and improved the content of the paper accordingly.

## References

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Xianliang Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Office of the Ministry of Education. 2024. Notice on further strengthening the education of standardised chinese characters in primary and secondary schools. Official website of the Ministry of Education. Education and Language Office [2024] No. 1, published on 2024-10 -22, https://www.moe.gov.cn.

Chittathuru Himala Praharsha and Alwin Poulose. 2024. Cbam vgg16: An efficient driver distraction classification using cbam embedded vgg16 architecture. *Computers in Biology and Medicine*, 180:108945.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

## A Appendix

### A.1 Parameters for fine-tuning each model in Task 2

| Parameters | Ace | Crude | Borderline1 | Borderline2 | LLM |
|---|---|---|---|---|---|
| Learning Rate | 5e-5 | 5e-5 | 2e-5 | 5e-5 | 5e-5 |
| Learning Rate Scheduler | cosine_with_restarts | cosine | cosine | cosine | cosine |
| Batch Size | 6 | 6 | 4 | 4 | 4 |
| LoRA Rank | 8 | 8 | 8 | 8 | 8 |
| Scaling Factor | 16 | 16 | 16 | 8 | 16 |
| Calculation Type | Bf16 | Bf16 | Fp32 | Bf16 | Bf16 |

Table 4: Parameters for each fine-tuning model