CCL25-Eval任务11系统报告：基于大模型微调的汉字硬笔书写质量自动评价

孔露露，眷红英，宋金旺，刘海芯，李一帆，罗哲伟
郑州大学，计算机与人工智能学院
{kll,jwsong,lhxin,lyfan}@gs.zzu.edu.cn, iehyzan@zzu.edu.cn，zwluo@stu.zzu.edu.cn

摘要

本技术报告探讨了通过微调本地视觉语言模型，实现汉字硬笔书写质量自动评价的技术方案。针对传统评价方法难以提供准确性反馈的问题，我们团队采用精心设计的prompt并结合微调的方式构建了一个高效的汉字硬笔书写质量自动评价系统。我们采用Qwen2.5-VL-7B-Instruct模型作为基础，通过LoRA微调技术实现了汉字书写质量等级分类（子任务一）和个性化评语生成（子任务二）的功能。系统地融合了视觉特征分析与语言生成能力，在训练过程中采用了梯度检查点、BF16混合精度训练等技术优化显存使用，并设计了针对性的损失函数和评估指标。实验结果表明，我们的方法能够有效实现汉字书写质量的细粒度评价。

关键词： 多模态大语言模型；LoRA微调；汉字书写质量评级；评语反馈

CCL25-Eval Task 11 System Report: Automatic Evaluation of Hard-Pen Chinese Handwriting Quality Based on Large Model Fine-Tuning

Lulu Kong, Hongying Zan, Jinwang Song, Haixin Liu, Yifan Li, Zhewei Luo
School of Computer and Artificial Intelligence，Zhengzhou University
{kll,jwsong,lhxin,lyfan}@gs.zzu.edu.cn，iehyzan@zzu.edu.cn，zwluo@stu.zzu.edu.cn

Abstract

This technical report explores a technical solution for automatically evaluating the quality of hard-pen Chinese character handwriting by fine-tuning a local vision-language model. Addressing the challenge that traditional evaluation methods struggle to provide accurate feedback, our team developed an efficient automatic evaluation system for hard-pen Chinese handwriting quality by employing carefully designed prompts combined with fine-tuning. We adopted the Qwen2.5-VL-7B-Instruct model as the foundation and utilized LoRA fine-tuning technology to achieve the functions of handwriting quality grade classification (Sub-task 1) and personalized comment generation (Sub-task 2). The system integrates visual feature analysis and language generation capabilities in a structured manner. During the training process, techniques such as gradient checkpointing and BF16 mixed-precision training were employed to optimize GPU memory usage, alongside the design of targeted loss functions and evaluation metrics. Experimental results demonstrate that our method can effectively achieve fine-grained evaluation of Chinese handwriting quality.

Keywords: Multimodal large language models , LoRA fine-tuning , Chinese handwriting quality evaluation , Feedback Generation

1 引言

汉字书写能力作为大学生人文素养的重要体现，其培养成效直接关系到文化传承效果与语言应用能力的提升。在数字化时代背景下，大学生书写能力普遍呈现下降趋势，而传统书法教学受限于人工评价的高成本与低效率，难以提供及时、个性化的书写指导。发展智能化的书写质量评价技术，已成为提升高等教育语言文字教学质量的关键突破口。现有汉字书写质量评价方法主要分为两类：基于传统图像处理的方法通过提取笔画形态、结构比例等手工特征进行评估，这类方法虽可解释性强但适应性有限；基于深度学习的方法利用卷积神经网络自动学习书写特征，在评级准确率上取得显著提升，但普遍存在"黑箱"决策问题，无法生成人类可理解的改进建议。这两类技术共同面临的核心瓶颈在于如何将视觉特征分析转化为符合教学逻辑的语言描述，实现从"评分"到"评语"的跨越。

多模态大语言模型的出现为上述问题提供了创新解决路径。这类模型同时具备视觉理解与语言生成能力，既能解析汉字图像的细部特征，又能基于教学知识库生成专业级评语。其核心优势在于：通过端到端学习建立书写质量特征与语言反馈的映射关系，模仿人类教师的"眼-脑-手"评价过程，先观察书写细节，再结合标准进行诊断，最后形成指导性文字。

本研究利用多模态大语言模型实现汉字硬笔书写质量的自动评价，来提供个性化的细粒度评价与反馈。通过对子任务一汉字书写质量进行评级，将汉字图片分类为"优秀"、"中等"、"不合格"三个等级；在子任务二中根据评级来实现自动性评语生成，为书写质量生成个性化、细粒度的反馈评语。我们在数据、训练和推理阶段优化了模型，为了测试模型效果，我们在任务中利用从训练集切分出来的数据集来进行评估，结果表明在子任务一中F1指标可达到0.954，并在子任务2中获得综合得分score 0.6764，取得了具有竞争力的结果。

2 相关工作

汉字书写自动评价研究经历了从人工规则到数据驱动的演进过程。早期系统主要依赖专家制定的评价标准，通过计算笔画与模板的匹配度 (Kamada, 2015)进行评分，这种方法虽直观但泛化能力有限。随着计算机视觉技术的发展，基于特征工程的方法开始采用Gabor滤波、方向梯度直分解 (Tran et al., 2017)等算法提取书写特征，结合传统机器学习分类器实现质量判断。这类方法在规范字体的评价上表现尚可，但对个性化书写风格的适应性较差。

深度学习的兴起推动了端到端评价模型的探索。研究者先后尝试了CNN、LSTM等架构来自动学习书写特征 (Saidaoui et al., 2020)，并在公开数据集上取得了优于传统方法的性能。特别地，一些工作引入了注意力机制 (Liang et al., 2021)来增强对关键笔画区域的关注，进一步提升了模型的判别能力。然而，这些方法仍存在明显局限，一方面，单纯的视觉模型难以构建书写质量与语言描述的关联；另一方面，独立训练的评级和评语生成模块无法实现端到端的协同优化。

视觉-语言大模型通过在大规模跨模态数据上的预训练，建立了图像与文本的深层语义关联。研究表明，这类模型在细粒度视觉描述生成任务中展现出接近人类的表现。在书写评价领域，已有初步尝试利用多模态模型生成简单评语 (Doostmohammadi et al., 2023)，但尚未系统探索其在完整评价流程中的应用潜力，特别是在评级与评语生成的联合优化方面仍缺乏深入研究 (Lu, 2023)。

本研究创新性地构建了基于Qwen2.5-VL-7B-Instruct的端到端评价框架，通过统一的模型架构同时解决评级和评语生成两个子任务。与现有工作相比，我们的方法具有三个显著优势：首先，采用参数高效的微调策略，在有限标注数据下实现模型能力的精准适配；其次，设计了任务特定的指令模板，有效引导模型生成符合教育规范的评语；最后，通过多任务协同训练，使视觉特征提取和语言生成模块相互促进，从而提升整体性能。

3 方法

我们的方法概述如图1所示。

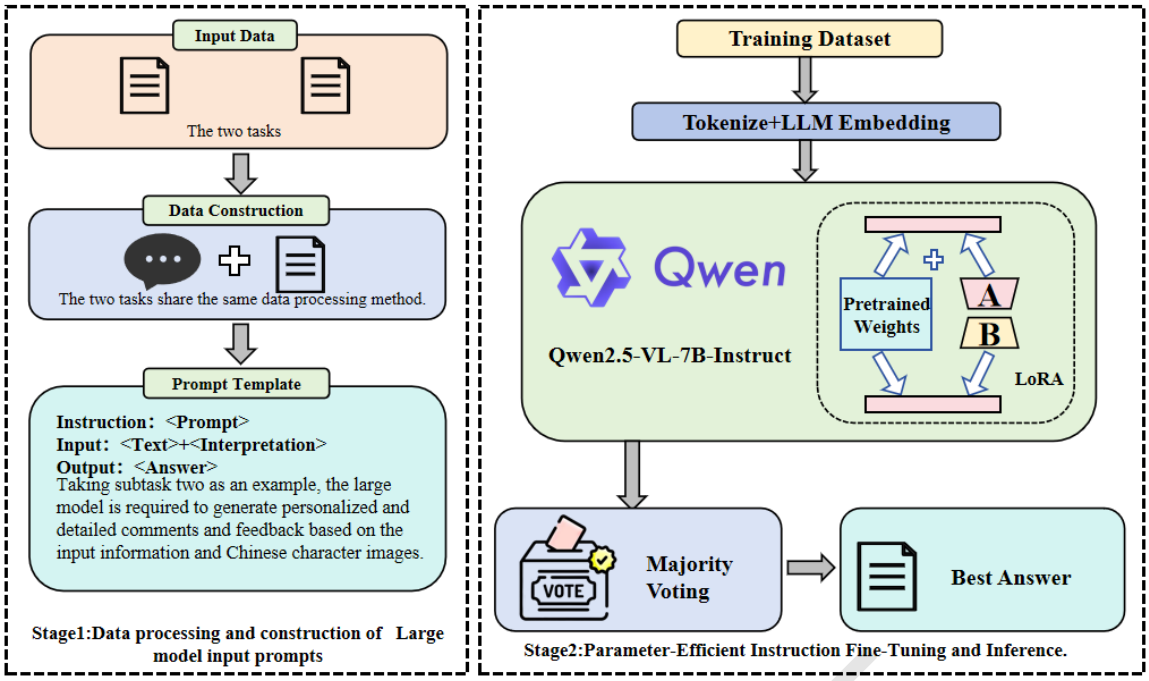

Figure 1: 方法概述图

3.1 数据处理

我们将评估任务提供的数据集处理为统一的指令-输出格式数据 (Ahn et al., 2022)，将图像输入与文本标注（评级标签或评语）结合，使其适配多模态模型的输入要求。数据处理过程通过特定的类实现，该模块支持训练、验证和测试三种模式下的差异化处理。在训练模式下，系统会将原始文本指令中的图像占位符替换为模型特定的视觉标记 (Li et al., 2023)，并采用动态掩码技术 (Raffel et al., 2019)确保损失函数仅对目标输出部分进行计算；在推理模式下，则保留完整的对话结构以供模型生成预测结果。针对两个子任务的不同特性，系统实现了差异化的评估指标，任务一采用F1值衡量分类性能，任务二则使用ROUGE系列指标评估生成文本质量 (Lin, 2004)。整个处理流程充分考虑了多模态数据的对齐问题，通过精心设计的批处理策略实现了图像和文本特征的高效融合 (Whitehouse et al., 2023)，我们为大模型设计的提示模板如图2所示。

```
"system": "你是一位书法鉴赏专家。你的任务是基于汉字书写的结构和笔画形态两个核心维度，为书写作品生成质量评语反馈。\n评价指标维度：\n\n###结构（宏观层面）：\n疏密\n匀称（独体结构、左右结构、左中右结构、上下结构、上中下结构、全包围、左上包围、右上包围、左下包围、左包右、上包下、下包上、品字结构）\n重心\n###笔画（微观层面）：\n轻重变化\n形态（点、横、竖、撇、捺、挑与提、折、钩）",
  "instruction": "<image>",
  "image": [
   "TRAIN-001.jpg"
  ],
  "output": "结构匀称美观，左右平衡，疏密适当，重心平稳。行笔自然流畅，笔画形态优美，字体遒劲妍美。"
```

Figure 2: 大模型输入模板

3.2 视觉语言模型的LoRA指令微调

我们使用LoRA方法对视觉语言模型进行指令微调。LoRA（Low-Rank Adaptation）(Hu et al., 2021)是一种高效的微调大规模预训练语言模型的方法，特别是在计算和内存资源有限的情况下。其核心是通过低秩分解逼近权重更新，给定预训练权重矩阵$W_0 \in \mathbb{R}^{a\times k}$，LoRA通过两组低秩矩阵$B \in \mathbb{R}^{a\times r}$ 和$A \in \mathbb{R}^{r\times k}$（其中$r \ll \min(a,k)$）进行参数更新，这里$k$ 为输入维度，a 为输出维度，r 是预定义的秩。这种低秩近似显著减少了可训练参数数量，在保持模型性能的同时显著降低了训练开销，适用于在资源受限的情况下对模型进行高效微调。这种低秩近似显著减少了可训练参数数量，前向传播过程表示为：

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{1}$$

在微调过程中，只有矩阵B和A被更新，而W_0保持冻结。该方法只需要添加少量参数就能实现适配，适用于在资源受限的情况下对模型进行微调。

模型架构设计方面采用视觉-语言多模态大模型框架，基于LoRA方法实现参数高效微调 (Verma et al., 2023)。系统整合了视觉编码器和语言解码器，其中：视觉编码器采用改进的ViT结构提取笔画、结构和章法等书写特征 (Huang et al., 2022)，语言解码器通过交叉注意力机制实现多模态特征融合 (Choma et al., 2020)。采用因果语言建模（CLM）技术进行训练：

$$\mathcal{L}_{\text{CLM}}(\Theta) = \mathbb{E}_{x\sim D}\left[-\sum_i \log p(x_i|x_{<i};\Theta)\right] \tag{2}$$

其中Θ为可训练参数（即B和A），$x_{<i}$表示历史标记序列$(x_0,...,x_{i-1})$。为适应不同子任务需求，模型引入了任务特定的提示标记，并采用LoRA微调策略在关键网络层（如q,k,v投影矩阵）注入低秩适配器，在保持模型性能的同时显著降低了训练开销。

3.3 训练优化及端到端评估

训练优化策略采用AdamW8bit优化器 (Dettmers et al., 2021)，集成Flash Attention 2 (Dao, 2023)优化注意力计算，采用BF16混合精度策略提升显存利用效率。训练过程启用梯度检查点技术，在不显著增加显存消耗的情况下支持更大模型训练 (Chen et al., 2016)。针对不同子任务特性，评级任务优化分类交叉熵损失函数，评语生成任务则优化语言建模损失目标。

本研究在模型训练过程中实现了端到端的动态评估机制，通过周期性验证确保模型性能的持续优化，并保存最好指标的lora权重。对于评级任务，系统每10个更新步数自动触发验证集评估，计算F1指标，其中F1达到0.9以上时自动保存最优模型；对于评语生成任务，则采用ROUGE-1、ROUGE-2和ROUGE-L的加权综合得分进行评估，当综合得分超过0.59时保留检查点。评估过程采用混合精度推理和动态批处理技术。系统设计了显存优化策略，包括梯度检查点激活和CUDA缓存清理机制。该评估方案通过早停机制和自动保存策略，在保证模型性能的前提下显著提升了训练效率。

3.4 模型集成

投票策略通过集成多个模型的预测结果或同一模型在不同推理路径上的输出，能够有效减少随机误差和模型偏差，显著提升系统的整体鲁棒性。在模型评估阶段，我们让每个模型独立地对测试集进行推理，确保预测结果的多样性。为了进一步增强预测的稳定性，我们采用集成学习中的投票机制（Voting Ensemble），通过多数表决或加权投票的方式融合不同模型的输出，从而确定最终预测结果。这种策略不仅能够降低单一模型在特定数据分布上的过拟合风险，还能充分利用不同模型的优势，提高在未知数据上的泛化能力。此外，投票集成还能有效缓解异常预测的影响，使模型在面对噪声或分布偏移时表现更加稳健。实验结果表明，经过投票策略优化后，我们的模型在验证集上的性能得到显著提升，最终取得了最优的提交成绩，验证了该方法的有效性。

4 实验

4.1 数据集

本任务使用的数据集主要用于评估汉字书写质量。数据集分为两个子任务：汉字书写质量评级和汉字书写质量评语反馈。

第二十四届中国计算语言学大会论文集，第430页-第436页，济南，中国，2025年8月11日至14日。

子任务一的目标是基于给定的汉字图片，对其书写质量进行等级分类，依据书写结构与笔画形态进行评判，评定等级分为优秀、中等、不合格三类。此任务的数据集包含1500个训练样本和300个测试样本，从1500条训练集中随机划分300条用于验证集。子任务二基于给定的汉字图片，对其书写质量生成个性化、细粒度的评语与反馈，关注于汉字书写质量的评语反馈，主要从结构与笔画形态两个维度进行评价，包含600个训练样本和100个测试样本，从600条训练集中随机划分60条用于验证集。

4.2 评估指标

4.2.1 子任务一指标

对于子任务一汉字书写质量评级，采用分类任务标准指标，包括每类别的精确率（Precision）、召回率（Recall）和F1值，重点关注三类别的均衡表现。其中，P为该类正确预测的样本数/预测为该类的样本数，R为该类正确预测的样本数/该类样本数，F1为精确率和召回率的调和平均数，用于综合评价分类器的性能。

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{5}$$

式中：真正例（True Positive, TP）是模型正确地将正样本预测为正样本的数量，假正例（False Positive，FP）是模型错误地将负样本预测为正样本的数量，假反例（False Negative，FN）是模型错误地将正样本预测为负样本的数量。

4.2.2 ROUGE指标

对于子任务二汉字书写质量评语反馈，使用文本生成任务的ROUGE指标（包括ROUGE-1、ROUGE-2和ROUGE-L），最终得分按0.4*ROUGE-L + 0.3*(ROUGE-1 + ROUGE-2)的加权公式计算，全面评估生成评语与专家标注在词汇、短语和语义层面的匹配程度。

$$\text{ROUGE-1} = \frac{\text{重合词语数}}{\text{参考答案词语数}} \tag{6}$$

$$\text{ROUGE-2} = \frac{\text{重合2-gram数}}{\text{参考答案2-gram数}} \tag{7}$$

$$\text{ROUGE-L} = \frac{\text{LCS长度}}{\text{参考答案长度}} \tag{8}$$

$$\text{Score} = 0.4 \times \text{ROUGE-L} + 0.3 \times \text{ROUGE-2} + 0.3 \times \text{ROUGE-1} \tag{9}$$

其中：重合词语表示预测结果与参考答案共现的词语集合，重合2-gram表示共现的连续二元词组集合，LCS（Longest Common Subsequence）表示最长公共子序列。

4.3 参数设置

我们采用了多项技术优化手段以提升训练效率和资源利用率。首先，我们使用Flash Attention 2来加速注意力计算，同时结合8bit AdamW优化器、梯度检查点和线性融合损失，以实现对预训练模型的高效适应。此外，通过梯度累积策略模拟更大批量的训练效果，并对原始模型梯度积累中的偏差问题进行了修正。

我们的实验环境为单张24GB VRAM的NVIDIA消费级显卡。在损失计算环节，我们特别优化了输入序列的处理方式，仅针对输出部分（评语或评级标签）计算交叉熵损失，而忽略指令部分的冗余计算，这一策略进一步提升了计算效率。在验证与保存机制上，我们设定了严格的评估频率，每训练10步即在验证集上进行一次全面评估。根据评估结果，我们会保存当前在

超参数	参数设置
Epoch	5
Batch size	4
Gradient_accumulation	2
Learning rate	1.8e-5
LoRA rank	64
LoRA Alpha	512
Max_grad_norm	0.5

Table 1: 训练所用超参数

子任务一中Macro-F1指标最高或在子任务二ROUGE综合得分最优的模型检查点。这一整套机制保证了模型性能的持续优化，我们最后基于Qwen2.5-VL-7B-Instruct 模型进行LoRA 微调，具体参数设置如表1所示。

4.4 实验结果

Model	track1_F1	track2_Score
Qwen2.5-VL-7B-Instruct	95.33	**67.64**
InternVL2_5-8B	94.25	66.31
Ovis1.6-9B	93.50	67.10
Valley-Eagle-7B	92.76	65.28
Voting Ensemble	**95.42**	-

Table 2: 模型效果

在任务中我们测试了多个模型效果，并在每个子任务的训练集上进行指令微调，结果如表2所示。我们在训练中评估所用的验证集从训练集中随机分割得到。我们观察到在Qwen2.5-VL-7B-Instruct上具有优越的性能，在子任务1中F1可达到0.9533，并在子任务2中获得综合得分score 0.6764。为进一步提升模型鲁棒性，我们对任务1中的三个结果较好的模型（Qwen2.5-VL-7B-Instruct、InternVL2_5-8B和Ovis1.6-9B）进行了模型集成。集成方法通过聚合多个模型的预测结果来实现性能增强，实验数据证实该方法有效提升了子任务1的评估指标。

5 结论

在本次CCL2025大学生汉字硬笔书写质量评测中，我们精心设计了prompt并结合微调的方式来激发多模态大语言模型的图像理解与文本生成能力，构建了一个基于视觉语言模型(VLMs)的综合优化框架，来更好地根据给定的汉字图片进行评级并生成详细、个性化的评价意见，以此弥补现有评价方法在提供个性化的细粒度评价与反馈方面的不足。我们采用模型投票集成策略提升模型的表现效果。最终，我们在比赛中获得了第一名的成绩。

致谢

本研究由国家自然科学基金重点项目资助（项目编号：U23A20316）。

参考文献

T. Kamada. 2015. The issues of automated driving vehicle and the expectations for 3D integration technology. In *2015 International 3D Systems Integration Conference (3DIC)*, pages KN2.1–KN2.4. IEEE.

S. Tran, X. Zhang, and Y. Li. 2017. The Recurrent Temporal Discriminative Restricted Boltzmann Machines. *arXiv preprint arXiv:1710.02245*.

H. Saidaoui, J. Chen, and L. Wang. 2020. Direct calculation of the kinetic energy functional derivative using Machine Learning. *arXiv preprint arXiv:2003.00876*.

S. Liang, M. Zhang, and K. Qiu. 2021. Parallel Rectangle Flip Attack: A Query-based Black-box Attack against Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7677–7687. IEEE.

E. Doostmohammadi, T. Norlund, M. Kuhlmann, and R. Johansson. 2023. Surface-Based Retrieval Reduces Perplexity of Retrieval-Augmented Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 521–529. ACL.

W. Lu. 2023. Clifford Algebra Cl(0,6) Approach to Beyond the Standard Model and Naturalness Problems. *International Journal of Geometric Methods in Modern Physics*.

M. Ahn, H. Kim, and S. Lee. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *Conference on Robot Learning*.

J. Li, D. Li, C. Xiong, and S. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning (ICML)*.

C. Raffel, N. Shazeer, and A. Roberts. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21:140:1–140:67.

C. Whitehouse, A. Smith, and B. Lee. 2023. WebIE: Faithful and Robust Information Extraction on the Web. *arXiv preprint arXiv:2305.14293*.

C.-Y. Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the ACL Workshop on Text Summarization Branches Out*, pages 74–81. ACL.

J. E. Hu, L. Shen, and W. Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

V. K. Verma, R. Gupta, and P. Singh. 2023. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Z. Huang, Y. Wang, and Q. Liu. 2022. Multilingual Knowledge Graph Completion with Self-Supervised Adaptive Graph Alignment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

N. Choma, S. Monti, and M. Bronstein. 2020. Track Seeding and Labelling with Embedded-space Graph Neural Networks. *arXiv preprint arXiv:2007.00149*.

T. Dettmers, M. Lewis, and L. Zettlemoyer. 2021. 8-bit Optimizers via Block-wise Quantization. *arXiv preprint arXiv:2110.02861*.

T. Dao. 2023. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. *arXiv preprint arXiv:2307.08691*.

T. Chen, B. Xu, and Z. Zhang. 2016. Training Deep Nets with Sublinear Memory Cost. *arXiv preprint arXiv:1604.06174*.