中图法分类号:TP391.41; TP183　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-23

论文引用格式:Wang Huazhen, Wu Luping, Hu Jiangang, Sun Jing. A progressive prompt expansion model for vocabulary image generation in international Chinese language education[J/OL]. Journal of Image and Graphics, XXXX:1-23. DOI: 10.11834/jig.260196.(王华珍, 吴鹭平, 胡建刚, 孙菁, 赵毅飞. 国际中文教育词汇图像生成的渐进式提示词扩展模型[J/OL]. 中国图象图形学报,XXXX:1-23. DOI: 10.11834/jig.260196.)
[DOI:10.11834/jig.260196]

国际中文教育词汇图像生成的渐进式提示词扩展模型

王华珍[1,2],吴鹭平[1,2],胡建刚[3*],孙菁[3],赵毅飞[3]

1. 华侨大学计算机科学与技术学院,厦门　361021; 2. 华侨大学计算机视觉与机器学习福建省高校重点实验室,厦门　361021; 3. 华侨大学华文学院,厦门　361021

摘　要: **目的**　图像辅助词汇习得是国际中文教育的重要教学策略,但现有提示词生成方法未能有效融合国际中文教育等级标准等领域专业知识,导致所生成图像在词义指向性、认知难度匹配性与教学适用性等方面存在显著不足。针对上述问题,本文提出一种面向国际中文教育词汇图像生成的渐进式提示词扩展模型(Progressive prompt expansion model for vocabulary image generation,PPEM-VIG)。**方法**　PPEM-VIG将词汇的认知等级、难度等级、词性、释义等离散教学要素统一建模为结构化教学要素集,并通过渐进式两阶段优化策略自动生成语义准确、结构规范且教学适用的图像提示词。监督学习阶段基于LLaMA模型以模板结构一致性损失与语言流畅性损失为约束,实现提示词的结构化生成;强化学习阶段基于监督微调LLaMA模型并增加模板内容相关性与基于CLIP的词图语义一致性奖励,对提示词进行深度优化,以实现提示词的拓展生成。**结果**　本文构建了首个面向国际中文教育场景的词汇图像提示词数据集(international chinese language education vocabulary image prompt dataset, ICLE-VIPD)。实验结果表明,PPEM-VIG在模板结构覆盖率、CLIP语义一致性得分、及教师主观评价等核心指标上较最优的基线模型分别提升了4.10%,3.13%,4.54%。**结论**　本文提出的渐进式监督-强化优化范式为词汇教学图像资源智能化生成上提供可借鉴的技术路线。

关键词: 国际中文教育;词汇教学;AIGC;图像生成;提示词模型

A progressive prompt expansion model for vocabulary image generation in international Chinese language education

Wang Huazhen[1,2], Wu Luping[1,2], Hu Jiangang[3*], Sun Jing[3]

1. College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China; 2. Fujian Provincial Key Laboratory of Computer Vision and Machine Learning, Huaqiao University, Xiamen 361021, China; 3. College of Chinese Language and Culture, Huaqiao University, Xiamen 361021, China.

Abstract: **Objective** Image-assisted vocabulary acquisition has long been regarded as an effective pedagogical strategy in international Chinese language education, especially for helping learners establish intuitive connections between lexical meaning, visual representation, and communicative context. Compared with purely verbal explanation, instructional images can reduce cognitive load, support dual coding, and enhance learners' retention of Chinese vocabulary. This is par-

收稿日期:2026-04-13;**修回日期**:2026-07-22

*通信作者:胡建刚 hujianganghw@126.com

基金项目:国家自然科学基金(项目编号:62476103);福建省社会科学基金重点项目(FJ2024A011);华侨大学中央高校基本科研业务费资助项目(项目编号:2024HQYJ01)

Supported by: National Science Foundation of China under Grant (62476103); Key Project of the Fujian Provincial Social Science Foundation (FJ2024A011); Fundamental Research Funds for the Central Universities of Huaqiao University (2024HQYJ01)

ticularly important because Chinese words often involve polysemy, contextual dependence, cultural connotation, and different levels of visual expressibility. With the rapid development of text-to-image generation models, such as DALL-E, Stable Diffusion, and Midjourney, teachers are now able to generate customized visual materials at a lower cost and with greater flexibility. However, current prompt generation and prompt expansion methods are mainly designed for general-purpose image generation. They usually emphasize visual details, artistic style, lighting, composition, or aesthetic quality, but pay insufficient attention to pedagogical constraints in international Chinese language education. As a result, images generated from ordinary prompts may fail to accurately represent the target word meaning, may not match learners' cognitive level or vocabulary difficulty, and may even introduce irrelevant or misleading visual elements. Existing methods also lack an effective mechanism for integrating domain-specific knowledge, such as vocabulary difficulty levels, cognitive stages, parts of speech, semantic explanations, and the Chinese proficiency grading standards for international Chinese language education. To address these limitations, this paper proposes a progressive prompt expansion model for vocabulary image generation, named PPEM-VIG, which aims to generate semantically accurate, structurally standardized, and pedagogically appropriate image prompts for Chinese vocabulary teaching. **Method** The proposed PPEM-VIG model formulates vocabulary image prompt generation as a structured mapping problem from pedagogical elements to image-generation prompts. Specifically, discrete teaching-related information, including cognitive grade, difficulty level, part of speech, and semantic explanation, is organized into a structured pedagogical element set. This representation enables the model to explicitly capture the instructional requirements of each vocabulary item rather than treating the word as an isolated text input. On this basis, PPEM-VIG adopts a progressive two-stage optimization framework. In the first stage, supervised learning is used to train a LLaMA-based structured prompt generation model. The model learns the mapping from the pedagogical element set to a structured prompt template consisting of several functional blocks, such as scene description, core vocabulary representation, primary visual features, auxiliary expression, pedagogical purpose, and style or cultural adaptation. To ensure that the generated prompt follows the expected template and remains readable, this stage introduces template structure consistency loss and language fluency loss. The former constrains the completeness and order of prompt blocks, while the latter encourages natural and coherent language generation. In the second stage, reinforcement learning is introduced to further optimize the supervised fine-tuned model. Unlike conventional prompt expansion methods that focus mainly on textual enrichment, this stage incorporates multiple reward signals related to pedagogical relevance and cross-modal alignment. The reward function includes template structure consistency, language fluency, template content relevance, and CLIP-based text-image semantic consistency. Among these rewards, the content relevance reward evaluates whether the generated prompt accurately reflects the input vocabulary and its teaching elements, while the CLIP-based reward measures whether the image generated from the prompt is semantically consistent with the vocabulary definition. Through this multi-objective reinforcement learning process, PPEM-VIG can refine prompts beyond surface-level template completion and improve their ability to guide text-to-image models toward educationally meaningful visual outputs. The progressive combination of supervised learning and reinforcement learning allows the model to first acquire stable structural generation ability and then enhance semantic alignment and visual teaching applicability. **Result** To support model training and evaluation, this study constructs the first vocabulary image prompt dataset specifically designed for international Chinese language education, named ICLE-VIPD. The dataset is built around vocabulary items selected from the Chinese proficiency grading standards for international Chinese language education and focuses on high-frequency nouns and verbs across different difficulty levels. Each sample contains a pedagogical element set and a corresponding structured prompt. The prompt annotation process considers the different visual representation principles of nouns and verbs. For nouns, the prompt design emphasizes core visual features, typical scenes, contrastive elements, and cultural appropriateness. For verbs, the prompt design focuses on action participants, action process, affected objects, result states, and contextual cues. Expert annotators with backgrounds in international Chinese language teaching participate in the construction and verification process to ensure that the prompts are both semantically accurate and suitable for classroom use. Experiments are conducted to evaluate PPEM-VIG from multiple perspectives, including template structure coverage, language fluency, block relevance, CLIP semantic consistency, and teachers' subjective evaluation. The results show that PPEM-VIG achieves superior performance compared with representative baseline models. In particular, compared with the strongest

baseline model, PPEM-VIG improves template structure coverage by 4. 10%, CLIP semantic consistency score by 3. 13%, and teachers' subjective evaluation score by 4. 54%. These improvements indicate that the proposed model not only follows the structured prompt template more reliably but also generates prompts that better support text-to-image models in producing semantically aligned teaching images. The teacher evaluation results further confirm that PPEM-VIG-generated prompts are more effective in terms of word-meaning accuracy, teaching adaptability, clarity of expression, and completeness of instructional information. Ablation analysis also demonstrates the necessity of the two-stage framework. The supervised learning stage substantially improves structural completeness and pedagogical element coverage, while the reinforcement learning stage further enhances cross-modal semantic consistency and practical teaching value. Removing individual reward components leads to performance degradation, which verifies the contribution of each reward signal to the overall model. **Conclusion** The proposes PPEM-VIG, a progressive prompt expansion model for vocabulary image generation in international Chinese language education. By incorporating structured pedagogical elements and combining supervised learning with reinforcement learning, the model effectively addresses the limitations of existing general-purpose prompt generation methods in educational scenarios. PPEM-VIG can generate prompts that are not only visually descriptive but also aligned with vocabulary meaning, learner cognition, difficulty level, and teaching objectives. The construction of ICLE-VIPD further provides a valuable benchmark resource for future research on vocabulary-oriented image generation, prompt engineering, and intelligent teaching material generation. Overall, the proposed progressive supervised-reinforcement optimization paradigm offers a practical technical route for the intelligent generation of vocabulary teaching images and contributes to the digital and intelligent transformation of international Chinese language education.

Key words: International Chinese Language Education; Vocabulary Teaching; AIGC; Image Generation; Prompt-based Model

0 引　言

近年来，随着文本生成图像(text-to-image, T2I)技术的迅猛发展，DALL-E、Stable Diffusion、Midjourney等文生图大模型在教育、设计、内容创作等领域展现出广阔的应用潜力(Song等, 2025)。词汇作为语言的基本单位，是连接句子和单字的桥梁。由于汉语词语独特的多义性与语境依赖性，图像在帮助汉语二语学习者建立词汇的直观映射、减轻认知负荷方面具有重要价值(赵雪婷，2021；饶静，2021)。郑艳群等人(2006)提出“视觉相关度与语义属性对应原则”，强调名词图像应突出核心视觉特征，并通过典型场景实现语境锚定。卢娜等人(2009)则围绕动词教学提出“动作特征与语义成分关联原则”，强调动作的主体、受事与结果状态的完整呈现。现有研究虽构建了部分视觉多媒体资源库，却难以实现从词汇图像的教学要素(认知等级、难度等级、词性、释义)到提示词的动态适配，限制了规模化应用。随着《国际中文教育中文水平等级标准》(以下简称《等级标准》)的发布，学者们开始关注难度等级、语义类型与图像呈现方式之间的对应关系(王军，2022)。这些原则为教学图像设计提供了理论指导，但多停留在人工设计层面，未与文生图大模型深度融合。

提示词是连接语言输入与图像生成的关键控制层，其质量直接影响生成图像的语义准确性与场景适用性。通常提示词采用要素组合框架形式，其要素包括主体描述、细节描述、修饰词、艺术风格和艺术家等(Bharne等，2025；Kadry等，2024)。由于词汇图像生成需体现词义的教学要素，要求认知等级、难度等级、词性、释义等多维度可控，国际中文教师在使用文生图模型时普遍面临提示词设计困难的问题(Cao，2024)。提示词扩展技术旨在通过分析提示词上下文语义对提示词要素进行重组、补充或模板化扩展，使其更加结构化和具象化，从而有效提升文生图模型的语义表达与图像生成质量。当前提示词扩展技术在通用场景下已取得一定进展，但仍无法将专业知识和领域语义有效注入提示词，以灵活适配国际中文教育中词汇的语义复杂性与教学要素的动态组合需求。因此，构建一种能够深度融合国际中文教育专业知识，实现词汇图像提示词结构化生成与语义精准对齐的提示词扩展模型，成为突破当前应用瓶颈的关键。

针对上述问题，本文提出一种面向国际中文教育词汇图像生成的渐进式提示词扩展模型（progressive prompt expansion model for vocabulary image generation，PPEM-VIG）。该模型将离散的认知等级、难度等级、词性、释义等教学要素，通过两阶段优化策略自动扩展出符合语义准确、结构规范且教学适用的的图像提示词。第一阶段为监督学习，模型利用模板结构一致性与语言流畅性损失等约束进行初步优化，以实现提示词的结构化生成；第二阶段为强化学习，模型引入模板内容相关性与基于CLIP的图像语义一致性等奖励信号对提示词进行深度优化，以实现提示词的拓展生成。本文的主要贡献包括三个方面：

1）提出面向国际中文教育词汇图像生成的渐进式提示词扩展框架，能对认知等级、难度等级、词性、释义等教学要素进行递进式优化，最终生成语义准确、结构规范且教学适用的的图像提示词。

2）设计双阶段优化策略，将监督学习的结构规范化与强化学习的语义一致性优化有机结合，有效缓解了教学提示词在语义匹配上的核心瓶颈。

3）构建首个面向国际中文词汇教学的图像提示词数据集ICLE-VIPD，为后续研究提供了重要的基准资源。

1 相关工作

1.1 文生图大模型技术进展

文生图（T2I）技术目标是将自然语言描述转换为语义一致、视觉真实的图像（Sudha等，2024）。早期的研究主要基于生成对抗网络（generative adversarial network，GAN）（Purwono等，2025；纪璎芮等，2026），如堆叠式生成对抗网络（stacked generative adversarial networks，StackGAN）、注意力生成对抗网络（attention generative adversarial network，AttnGAN）等模型，通过引入注意力机制与多阶段生成结构，在一定程度上提升了生成图像的分辨率与语义匹配度（Shafi等，2024；Wang，2024）。然而，GAN模型普遍存在训练不稳定和模式坍塌等问题，限制了其在复杂语义生成中的表现。随着稳定扩散模型（Stable Diffusion）等扩散模型的提出，T2I技术进入新的阶段。扩散模型通过多步去噪过程，从高斯噪声中逐渐恢复出与输入语义相符的图像，具备更高的生成稳定性和图像质量（Ho等，2020）。典型模型如OpenAI的DALL-E 2图像生成模型（Ramesh等，2022）、Google的Imagen图像生成模型（Saharia等，2022）、Stability AI的Stable Diffusion稳定扩散模型（Rombach等，2021）等，均在大规模图文对齐数据上进行训练，实现了文本到图像的高保真转换。这些模型依赖对比语言—图像预训练模型（CLIP）等词图语义一致性学习框架，将语言与视觉特征映射到共享的语义空间，从而实现高效的跨模态对齐。

提示词工程（prompt engineering，PE）作为生成式模型中的核心环节，极大影响模型的生成结果（Schneider等，2024）。PE不仅影响生成图像的主题、风格和细节，还直接决定了图像与文本之间的语义对齐性能（Yang等，2023）。PE采用模板化设计，通过结构化框架实现生成控制。例如，将提示词拆分为“语义核心”“视觉属性”“场景背景”“风格限定”等模块，从而提高可复用性（Chen等，2024；Liu等，2022）。现有的研究主要集中在如何通过设计精确结构的提示词来优化生成模型的输出，以实现更高质量的图像生成（Cao等，2023；卢裕弘等，2025）。手动设计的模板在面对大量复杂和多样化的任务时效率低下，且难以保证生成的图像与语义的高度一致性。

1.2 提示词扩展技术研究

随着大语言模型和图像生成模型的发展，PE逐渐从手动设计转向自动化扩展（Fernando，2025），成为提升生成质量的关键手段（Kitrungrotsakul等，2025）。在通用技术层面，Google团队提出的提示词扩展框架（prompt expansion，PE）通过多阶段扩充策略，实现了从简单文本到高质量、多样化图像描述的自动化跃迁（Datta，2024）；针对中文语境的提示词增强（promptboost，PB）等方法则通过免标注的扩展策略，显著提升了中文提示词在生成模型中的语义对齐效果（Leong等，2025）。个性化提示重写技术（Personalized Prompt Rewriting）通过融合用户历史偏好数据实现了生成内容的定制化输出（Li等，2024）。

尽管通用领域的提示词技术日趋成熟，然较少关注如何结合特定领域（如教育、医疗等）中的专业知识进行定制化设计（Nori等，2023）。理想的教学辅助图像不仅需要视觉美感，更需精准承载词汇的教学语义，符合特定的教学情境以及适应不同语言

水平学习者的认知特点(Liu 等，2023；Ahmadi 等，2020)。现有扩展技术多侧重于画面光影、风格渲染等视觉元素的调控，忽略了教学目标、学习者认知特征及文化锚定等深层语义的整合，这一矛盾在国际中文教育场景中尤为突出(Choi 等，2025)。

针对多维约束下的生成难题，强化学习(reinforcement learning，RL)作为一种基于反馈机制的优化范式，为提示词工程提供了新的解决思路。与传统的监督学习方法不同，强化学习通过引入奖励信号(Reward Signal)与策略优化(Policy Optimization)，使模型能够在多个生成步骤中自主调整，以逼近最优的提示词设计效果(Shi 等，2025)。近年来，RL 在提示词扩展中的应用主要集中在语义一致性优化方面，即根据生成图像的质量以及图像与文本的语义匹配度设计奖励函数，驱动模型在训练过程中不断进化(Ramesh 等，2021)。但在国际中文教学垂直领域，相关研究仍缺少构建有效的、包含领域特征的奖励机制(如引入“教学适用性”或“文化准确性”奖励)，以引导模型生成兼具视觉表现力与教学实用性的高质量提示词(Qi 等，2024)。

1.3 国际中文教育词汇视觉化

国际中文教育中的词汇教学一直是学界关注的焦点，而视觉化手段在促进词汇习得方面的有效性已得到认知心理学理论的广泛支持。Mayer 等人(2005)提出多模态学习理论(multimodal learning theory)与 Paivio 等人(1990)提出双编码理论(dual coding theory)，说明人类认知系统在同时处理言语与图像信息时，能够通过建立指代联系降低认知负荷，从而实现深层理解。Shen 等人(2010)进行实证研究进一步证实，在汉语二语习得中，图文结合的双编码策略利用了图像优势效应，能有效辅助学习者克服汉语高象形特征带来的认知障碍，促进词汇的长时记忆与提取。同时，情境认知理论强调语言习得的社会文化属性，指出在具体视觉情境中构建词汇意义比单纯的语义解释更为有效。Tong 等人(2020)研究指出，缺乏动态演示与语境关联的图片容易导致学习者对抽象概念产生误解。此外，涉及“龙”、“面子”等具有深厚文化图式的文化负载词，现有教材插图常因缺乏深度文化表征而导致跨文化交际中的语用失误(Xiong 等，2021)。

近年来，随着生成式人工智能(AIGC)技术的爆发，特别是 Midjourney、Stable Diffusion 等文生图模型的成熟，为上述教学痛点提供了全新的技术路径(Gou 等，2025)。Yu 等人(2025)指出 AIGC 技术打破了传统教学资源的获取瓶颈，使教师能够低成本、高效率地生成定制化图像，实现从“寻找图片”到“创造情境”的范式转变。高质量的 AI 生成图像在多义词辨析与近义词习得中表现出显著优势(Attygalle 等，2025)。Fan 等人(2025)针对中国英语学习者的认知研究发现，AIGC 生成的精细化图像能显著提升学习者对生词的加工深度与注意保持。而在文生图模型优化研究视角，有学者强调通过融合图像和文本特征作为输入来指导图像描述的生成模型(史秀聪，2020)，也有学者探讨基于文本感知和非重复单词生成的图像语义理解方法(杨晨露等，2023)，提出需要研制教育提示语工程的技术规范(赵晓伟等，2023)。

然而，当前文生图提示词的设计与优化多聚焦于通用图像生成的精准度与美学效果，却未深度融入《国际中文教育中文水平等级标准》等国际中文教育领域的专业知识体系。这使得文生图技术难以真正实现从“创造情境”到“精准赋能词义习得”的跨越，无法充分发挥其在词汇教学中的技术优势，甚至可能因语义偏差误导学习者的词义认知，制约国际中文词汇教学数字化、智能化转型的质量。

2 模型介绍

2.1 问题描述

面向国际中文教育场景下教学图像需精准承载词汇核心词义、契合分级教学目标、适配学习者认知水平等核心需求，提示词扩展技术将针对离散的教学要素(如认知等级、难度等级、词性、释义等)进行重组、补充或模板化扩展，使其输出符合教学要求、语言自然且语义一致的图像提示语句，进而指导文生图模型绘制符合教学语境的图像。设教学要素集合 X，文生图提示词定义为 Y，国际中文词汇文生图提示词扩展模型旨在学习从教学要素集合到文生图提示词的映射关系 F，表达式如式(1)：

$$Y = \mathrm{F}(X) \#(1)$$

2.2 模型架构

针对现有提示词生成方法未能有效融合国际中文教育等级标准等领域专业知识、所生成图像在词义指向性等方面存在显著不足的问题，本文提出一

种面向国际中文教育词汇图像生成的渐进式提示词扩展模型PPEM-VIG。该模型创新性在于，相较于当前主流文生图提示词扩展技术常侧重于增加图像的细节丰富度或艺术风格表达，本研究尝试引入监督微调和强化学习两阶段架构来增强国际中文教育这一特殊场景的领域适应性。另外，本研究引入跨模态的视觉对齐能力作为反馈信号以增强扩散模型的图像生成质量。具体而言，PPEM-VIG模型主要由2个核心模块组成，如图1所示：第一个是结构化提示词生成模块，负责学习输入与结构化模板之间的结构映射关系，初步生成结构化提示词，实现结构化模板中各个语块的识别与完整表达。第二个是结构化提示词扩展模块，通过多维奖励函数包括模板结构一致性、语言流畅性及图像语义一致性等迭代优化，生成结构化扩展提示词。

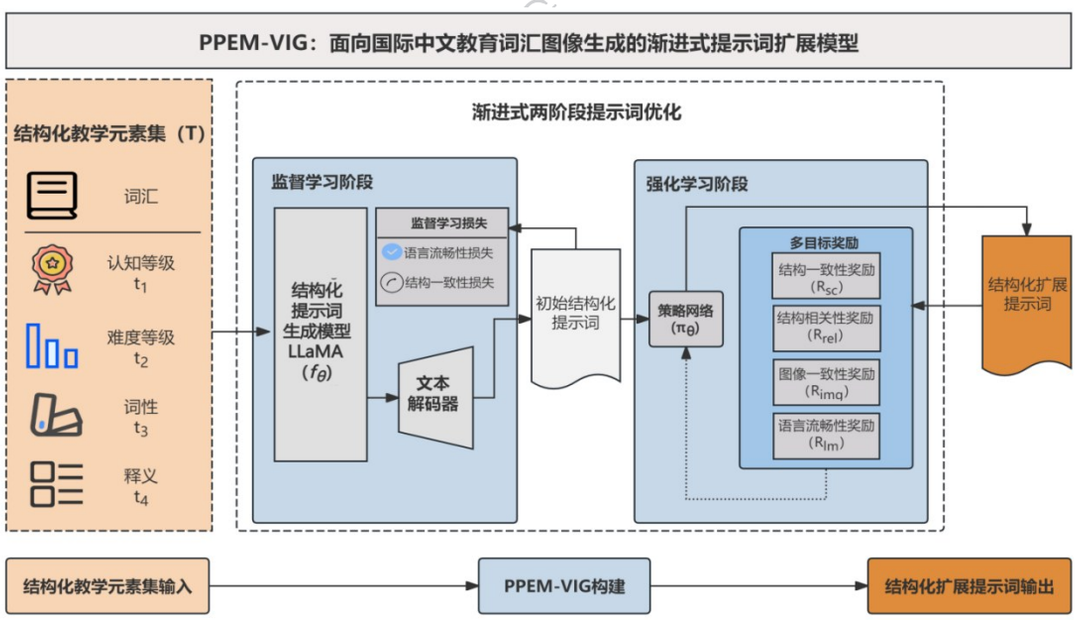

图1　PPEM-VIG模型框架图

Fig. 1　Framework of the PPEM-VIG model

2.3　结构化提示词生成模块

该模块执行PPEM-VIG模型的有监督学习阶段，以确保输入的教学要素集得以在提示词模板中被正确表达。模型将学习从教学要素集$\boldsymbol{T}$到文生图结构化提示词模板Y各个语块的映射关系。具体地，设参数化映射函数为f_θ，其目标是最大化给定要素集条件下目标提示词序列的条件概率，表达式如式(2)：

$$\theta^* = argmax_\theta \sum_{(\boldsymbol{T},\boldsymbol{Y})\in\mathcal{D}} \log P_\theta(\boldsymbol{Y}|\boldsymbol{T}) \#(2)$$

式中，θ表示结构化提示词生成模块的可训练参数集合，$\mathcal{D}$为训练数据集，包含教学要素集-结构化提示词组对，θ^*表示在训练数据集$\mathcal{D}$上通过最大化条件概率目标所得到的最优模型参数。教学要素集$\boldsymbol{T}=\{t_1,t_2,\ldots,t_J\}$，其中，J为教学要素的数量(本研究设定为4)，包含认知等级、难度等级、词性、释义等信息。$\boldsymbol{Y}=(y_1,y_2,\ldots,y_K)$表示输出的结构化生成提示词序列，其中，K为提示词要素语块的数量(本研究设定为6)分别对应场景语块(SCENE)、核心语义语块(CORE_vocab)、主要特征语块(PRIMARY_FEATURES)、辅助描述语块(AUXILIARY_EXPRESSION)、教学目的语块(PEDAGOGY)与风格文化语块(STYLE_CULTURE)。任一要素语块均以键值对形式组成：其键名为要素语块类别；其值为要素语块文本。以上语块设计实现文本教学指令向多模态视觉特征的精准映射与全维覆盖(郑艳群等，2006；卢娜，2009)。

进一步地，条件概率P_θ采用自回归方式进行建模，其分解形式如式(3)：

$$P_\theta(\boldsymbol{Y}\,|\,\boldsymbol{T}) = \prod_{k=1}^{K} P_\theta(y_k|\,y_{<k},T) \#(3)$$

本研究采用LLaMA预训练模型作为结构化提

示词生成模块的基础架构。在多条件上下文建模方面，本文需要综合认知等级、难度等级、词性及释义等多维教学属性进行连续语义生成。LLaMA 采用 Decoder-only Transformer 架构，能够在统一语义空间内完成自回归条件建模。相较于传统 Seq2Seq 模型，其在长距离上下文依赖建模与文本语义连贯性方面具有较好表现（Touvron 等，2023；Brown 等，2020），能够为教学语境下的提示词生成提供更加稳定的语义表示能力。此外，本研究所设计的结构化提示词由六类语块组成，模型不仅需要生成自然流畅的文本内容，还需要保持各语块之间的结构完整性与逻辑一致性。已有研究表明，LLaMA 在指令微调任务中具有较好的结构遵循能力（Taori 等，2023），能够有效维持生成过程中语块顺序的稳定性，降低语块缺失与语义漂移问题，从而与本研究的结构化提示词生成任务形成较高适配性。同时，考虑到国际中文教育领域标注数据规模相对有限，本文进一步结合 LoRA 等参数高效微调（Parameter-Efficient Fine-Tuning，PEFT）方法对模型进行领域适配。LLaMA 具有较好的模块化结构与参数扩展性，能够较好兼容低秩参数更新机制（Hu 等，2021），仅需训练少量可学习参数即可实现领域知识迁移，从而有效降低训练成本与显存消耗，并缓解全参数微调可能带来的灾难性遗忘问题。最后，考虑到第二阶段需要进一步引入结构一致性、语义相关性以及跨模态一致性等多目标奖励信号，LLaMA 的开放参数结构与良好的可扩展性能够为后续强化学习优化提供稳定基础。其生成式策略网络形式能够较自然地与策略梯度类强化学习算法结合（Ouyang 等，2022），从而支持模型在保持结构稳定性的基础上，进一步提升提示词的教学语义表达能力与跨模态一致性。

为适配 LLaMA 输入层格式，利用序列拼接操作符 $\oplus$ 构建自然语言输入序列 $S_{in} = \text{Word} \oplus p_1 \oplus t_1 \oplus p_2 \oplus t_2 \oplus p_3 \oplus t_3 \oplus p_4 \oplus t_4$ 其中 $\{p_1, \cdots, p_4\}$ 为教学要素键值对的键名，即“认知等级”、“难度等级”、“词性”以及“释义”。这一过程不仅能够捕捉词汇的上下文语义信息，还能有效提取句子级别的语义表示，从而为后续任务提供高质量的特征输入。

为实现计算资源受限条件下的高效训练，我们采用低秩适配对 LLaMA 模型进行参数高效微调（陈诗琪，2026）。我们冻结预训练权重 $\boldsymbol{W} \in \mathbf{R}^{d \times k}$，并引入两个可训练的低秩矩阵 $\boldsymbol{B} \in \mathbf{R}^{d \times r}$ 和 $\boldsymbol{A} \in \mathbf{R}^{r \times k}$，其中秩 $r \ll \min(d, k)$，在此前提下，对于自然语言输入序列 S_{in} 经分词与嵌入层映射后得到的第 t 个 token 表示 $\boldsymbol{x}_t \in \mathbf{R}^{k \times 1}$，其微调过程的前向传播如式（4）所示：

$$h^{d \times 1} = \boldsymbol{W}_0^{d \times k} \boldsymbol{x}_t^{k \times 1} + \boldsymbol{B}^{d \times r} \boldsymbol{A}^{r \times k} \boldsymbol{x}_t^{k \times 1} \#(4)$$

式中，第一项 $\boldsymbol{W}_0 \boldsymbol{x}_t$ 表示冻结的预训练模型输出，第二项 $\boldsymbol{BAx}_t$ **为低秩参数增量项，即** $\boldsymbol{\Delta W x}_t$，其中 $\boldsymbol{\Delta W} = \boldsymbol{BA} \in \mathbf{R}^{d \times k}$。在训练初期，矩阵 $\boldsymbol{B}$ 初始化为零矩阵，$\boldsymbol{A}$ 采用随机高斯初始化，从而确保 $\boldsymbol{\Delta W} = \boldsymbol{BA} = 0$ 在训练起点处为零。在反向传播过程中，仅更新 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的参数，而 $\boldsymbol{W}_0$ 保持冻结。这使得可训练参数量从 $d \times k$ 降至 $2 \times d \times r$，显著降低显存占用并缓解灾难性遗忘问题。

有监督学习阶段的训练目标为，融合结构一致性和语言流畅性两种损失，以确保生成的提示词既符合模板结构要求，又具有自然流畅的语言表达。其中，设计结构一致性损失用以衡量模型输出的结构化提示词与预设模板结构的契合程度。若某个语块缺失、错误生成或者语块顺序混乱，都会导致结构一致性损失增加。具体而言，结构一致性损失的表达式如式（5）所示：

$$L_{\text{sc}} = \frac{1}{K} \sum_{k=1}^{K} I\left(k(z_k) \neq k(y_k)\right) \#(5)$$

式中，K 为提示词要素语块的数量，$k(\text{z}_\text{k})$ 表示结构化提示词生成模块解码输出的第 k 个要素语块 z_k 中的要素语块类型，$k(y_k)$ 表示训练集中期望要素语块中的要素语块类型。当解码要素语块类型与期望要素语块类型不同时惩罚为 1；如果一致，则惩罚为 0。通过最小化该损失，模型能够有效维持结构化提示词的模板稳定性与语块完整性。

另一方面，设计语言流畅性损失用于衡量结构化生成提示词文本的可读性和自然度，其计算如式（6）所示。

$$L_{\text{lm}} = \text{PPL}(\boldsymbol{Y}) = \exp\left(-\frac{1}{N} \sum_{i=1}^{N} \log P_\theta(y_i | y_{<i}, S_{in})\right) \#(6)$$

式中，$\text{PPL}(\cdot)$ 为困惑度。困惑度越低，表示生成文本越自然、连贯，有利于作为下游文生图模型的有效提示。N 表示生成提示词序列 Y 中的 token 数量，y_i 表示第 i 个 token，$\text{y}_{<\text{i}}$ 表示当前位置已生成的 token 序列，$P_\theta(y_i | y_{<i}, S_{in})$ 表示在由教学要素集合

T经过序列拼接构建自然语言输入序列S_{in}条件下，模型生成第i个token的条件概率。

综合起来，监督学习损失函数L_{sl}定义如式(7)：

$$L_{\mathrm{sl}} = L_{\mathrm{llm}} + \lambda_{\mathrm{sc}} L_{\mathrm{sc}} + \lambda_{\mathrm{lm}} L_{\mathrm{lm}} \#(7)$$

式中λ_{sc}和λ_{lm}为权重超参数且$\lambda_{\mathrm{sc}} + \lambda_{\mathrm{lm}} = 1$。通过最小化$L_{\mathrm{sl}}$，模型旨在实现结构稳定、语言自然的结构化生成提示词生成$\boldsymbol{Y}$。

2.4 结构化提示词扩展模块

该模块执行PPEM-VIG模型的强化学习阶段，以进一步提升结构化生成提示词的语义合理性、语义相关性以及跨模态一致性。模型将以结构化生成提示词为优化对象，将多维质量指标纳入奖励函数，通过强化学习算法优化生成策略，实现多目标协同优化下的提示词质量提升，从而得到结构化扩展提示词P。

在强化学习阶段，将有监督学习阶段微调后的生成模型f_θ，作为策略网络π_θ。对于给定要素集$\boldsymbol{T}$，构建对应的自然语言输入序列S_{in}，策略网络基于该输入生成结构化扩展提示词序列$\boldsymbol{P} = [w_1, w_2, \ldots, w_{\mathrm{N}}]$，其中$N$表示生成序列的token数量，$w_N$表示第$N$个生成的token，其生成过程表示如式(8)：

$$P \sim \pi_\theta(\cdot \mid S_{in}) = \prod_{t=1}^{T} \pi_\theta(w_N \mid S_{in}, w_{<N}) \#(8)$$

式中，θ表示PPEM-VIG模型在监督学习阶段微调后得到的可训练参数集合，π_θ表示以参数θ为条件的生成策略，$\pi_\theta(w_N|S_{in}, w_{<N})$表示在输入条件与历史生成序列约束下生成当前token的条件概率，该生成方式能够有效建模提示词序列内部的上下文依赖关系，从而提高生成结果的语义连贯性与结构稳定性。强化学习的目标是优化策略参数θ，使得生成序列的期望奖励最大化，表达式如式(9)：

$$J(\theta) = \mathrm{E}_{P \sim \pi_\theta(\cdot|S_{in})}[R(\boldsymbol{T},\boldsymbol{P})] \#(9)$$

式中，$J(\theta)$表示强化学习阶段的期望奖励目标函数，$R(T, P)$励函数，用于评估结构化扩展提示词序列的质量。

本研究采用多目标奖励函数设计以实现多维度的质量控制。奖励函数由四个核心维度构成，分别为结构一致性奖励、语言流畅性奖励、内容相关性奖励以及图像语义一致性奖励四部分。结构一致性奖励对应第一阶段监督学习的结构一致性损失，但经过归一化和极性变换得到，用于衡量生成结果与目标模板在结构层面的匹配程度。其表达式如式(10)所示。

$$R_{\mathrm{sc}}(P) = 1 - \frac{1}{K}\sum_{k=1}^{K} \mathrm{I}\big(\boldsymbol{k}(P_k) \neq \boldsymbol{k}(y_k)\big) \#(9)$$

语言流畅性奖励对应第一阶段监督学习的语言流畅性损失，但经过极性变换得到，其表达式如式(10)所示。

$$R_{\mathrm{lm}}(\boldsymbol{P}) = \exp\left(-\frac{1}{N}\sum_{N=1}^{N} \log P_\theta(w_N|S_{in}, w_{<N}) \right)^{-1} \#(10)$$

结构相关性奖励着重评估强化学习解码输出的结构化扩展提示词$\boldsymbol{P}$的文本内容是否真正与输入词汇及其教学要素集相关。其表达式如式(11)所示。

$$R_{\mathrm{rel}}(\boldsymbol{T},\boldsymbol{P}) = S_{\mathrm{rel}}(\boldsymbol{T},\boldsymbol{P}) \#(11)$$

式中相关性评分$S_{\mathrm{rel}}(\cdot,\cdot)$由全参数大语言模型作为外部评估器进行打分取值范围为$[0, 1]$。评分越高，表示生成提示词与输入教学要素之间的语义相关性越强。其通过知识蒸馏将教师大模型的语义理解能力对学生模型(即PPEM-VIG)生成模板中的各结构进行细粒度评估。

图像语义一致性奖励用于衡量结构化扩展提示词$\boldsymbol{P}$能否驱动文生图模型生成与词汇释义一致的视觉内容，从而确保PPEM-VIG模型在跨模态层面上的对齐能力。其计算过程是首先将结构化扩展提示词$\boldsymbol{P}$输入到预训练的文生图模型生成图像$\boldsymbol{I}_P$；然后利用跨模态语义对齐模型CLIP分别对词汇释义要素t_4和图像$\boldsymbol{I}_P$进行编码，得到其在共同语义空间中的向量表示，记为$\boldsymbol{V}_{\mathrm{t}}$和$\boldsymbol{V}_{\mathrm{i}}$。进而计算两者的余弦相似度，如式(12)所示：

$$\mathrm{sim}(\boldsymbol{V}_{\mathrm{t}},\boldsymbol{V}_{\mathrm{i}}) = \frac{\boldsymbol{V}_{\mathrm{i}}^{T}\boldsymbol{V}_{\mathrm{t}}}{// \boldsymbol{V}_{\mathrm{i}} /// \boldsymbol{V}_{\mathrm{t}} //} \#(12)$$

并将该相似度作为图像语义一致奖励，如式(13)所示：

$$R_i(\boldsymbol{P}, t_4) = \mathrm{sim}(\boldsymbol{V}_{\mathrm{t}},\boldsymbol{V}_{\mathrm{i}}) \#(13)$$

最终，强化学习奖励函数为四者的加权和，并引入自适应权重调整机制，在训练过程中根据各指标收敛情况动态调整权重。表达式如式(14)所示。

$$R(\boldsymbol{P},\boldsymbol{T}) = \alpha_{\mathrm{sc}} R_{\mathrm{sc}}(P) + \alpha_{\mathrm{lm}} R_{\mathrm{lm}}(P) + \alpha_{\mathrm{rel}} R_{\mathrm{rel}}(T,P) + \alpha_{\mathrm{i}} R_{\mathrm{i}}(P,t_4) \#(14)$$

式中$R(\boldsymbol{P}, \boldsymbol{T})$为强化学习奖励函数，$\alpha_{\mathrm{sc}}, \alpha_{\mathrm{lm}}, \alpha_{\mathrm{rel}}, \alpha_{\mathrm{i}}$为对应奖励的非负权重超参数，相加总和为1，用于调节不同奖励项对整体优化过程的相对影响。即从模板结构、语言质量、教学语义相关性以及跨模态图像

一致性四个维度对生成提示词进行联合优化。与传统单目标文本生成方法相比，该多目标奖励机制能够更加有效地适配国际中文教育场景下的结构化提示词生成需求。

本研究采用策略梯度法优化目标函数$J(\theta)$，其梯度计算如式(15)所示：

$$\nabla_\theta J(\theta)=\mathrm{E}_{P\sim\pi_\theta(\cdot|S_{in})}[\nabla_\theta\log\pi_\theta(\boldsymbol{P}|S_{in})\cdot R(\boldsymbol{P},\boldsymbol{T})]\#(15)$$

式中，$\nabla_\theta J(\theta)$代表期望奖励目标函数的梯度，$\nabla_\theta\log\pi_\theta(\boldsymbol{P}|S_{in})$代表对数概率关于参数$\theta$的梯度。

在训练中，采用蒙特卡洛采样策略近似计算上述期望梯度。对于包含B个结构化扩展提示词序列$\{\mathrm{P}^{\mathrm{b}}\}_{\mathrm{b}=1}^{\mathrm{B}}$的训练批次，计算对应的奖励$R(P^b,T)$，其近似经验梯度$\nabla_\theta\hat{J}(\theta)$计算表达式如式(16)所示：

$$\nabla_\theta\hat{J}(\theta)=\frac{1}{B}\sum_{b=1}^{B}\left[\nabla_\theta\log\pi_\theta(P^{(b)}|S_{in})\cdot R(P^b,\boldsymbol{T}))\right]\#(16)$$

2.5 优化目标与收敛

结合有监督学习损失函数与强化学习奖励函数，PPEM-VIG模型的总损失函数可定义为式(17)：

$$L_{\mathrm{tl}}(\theta)=L_{\mathrm{sl}}(\theta)-\beta\cdot J(\theta)\#(17)$$

式中β为平衡系数。

通过迭代优化，策略网络π_θ逐步收敛至最优策略，使得生成的结构化扩展提示词满足式(18)：

$$\theta^*=\arg\max_\theta \mathrm{E}_{P\sim\pi_\theta(\cdot|S_{in})}[R(P,T)]$$
$$P^*\sim\pi_{\theta^*}(\cdot|S_{in})\#(18)$$

最终得到结构化拓展提示词P^*，其在语义合理性、语义相关性以及跨模态一致性均得到优化。训练完成得到优化后的PPEM-VIG模型参数，可用于后续推理任务。

3 实验与分析

3.1 数据集

国际中文教育领域尚无面向词汇图像提示词生成的现成数据集可供使用，本研究基于任务特性与领域现状，自建国际中文教育词汇图像提示词数据集(international chinese language education vocabulary image prompt dataset，ICLE-VIPD)。该数据集的构建本身也构成本研究贡献的重要组成部分。ICLE-VIPD数据集所有目标词汇均严格选自《国际中文教育中文水平等级标准》，聚焦高频使用的名词与动词，覆盖一至九的难度等级，确保数据集的教学场景多样性。ICLE-VIPD每条样本包含教学要素集、结构化提示词两个字段，服务于有监督学习阶段的训练需求。针对每一候选词汇的教学要素集合为{认知等级、难度等级、词性、释义}，系统刻画该词汇的教学要素信息，其中认知等级参考《国际中文教育中文阅读分级标准》(2025)，细分为低龄段(5—12岁)、中龄段(12—18岁)及高龄段(18岁以上)。

对于结构化提示词的标注，针对名词与动词的差异，参照郑艳群等(2006)提出的名词图片展示框架与卢娜(2009)提出的动词图片展示框架，分别设计要素语块文本，构建对应的结构化提示词模板。针对名词“西瓜”和动词“喊”，结构化提示词实例分别如表1所示。

如表1可见，名词的要素语块文本强调基于语义类别选择适合场景法、衬托法、象征法等多种图片表达方式，并强调图片释义的相关度与文化适切性；动词语块文本重点覆盖在动作结构、施受关系、情状特征等方面具有清晰视觉化线索、适合用图片呈现的动词语义类，从而保证动词样本在动作过程与结果状态的图像表达上具有较强的可操作性。

在结构化提示词标注时，邀请具有国际中文教学背景的教师专家进行撰写。标注过程遵循教学有效性与跨模态一致性双重校验机制：专家将所撰写的结构化提示词输入文生图大模型生成对应图像，评估图像能否准确传达词汇语义、是否符合国际中文课堂教学场景需求。数据标注工作由3名专家独立完成，标注前统一标注规范与细则，对存在分歧的标注结果经集体讨论后确定最终标签。

通过上述多阶段构建流程，ICLE-VIPD数据集实现了教学知识嵌入(要素集)、结构规范性(结构化提示词模板)与视觉语义对齐(结构化提示词图像校验)的统一。最终构建的ICLE-VIPD数据集规模为5682，其包含名词数据3400条样本，动词数据2282条样本。数据集按照8∶1∶1的比例划分为训练集、验证集和测试集。划分时保持名词与动词在各子集中的分布均衡，以确保模型在训练与评估阶段具备良好的覆盖性与泛化能力。

3.2 实验设置

本实验在搭载NVIDIA A100-SXM4-80GB GPU的服务器上进行，训练使用PyTorch框架(版本2.1.2)。以Meta-Llama-3-8B-Instruct作为基础模型，文本编码长度设为512。采用优化器AdamW进

表1　结构化提示词实例

Table 1　Example of a Structured Prompt

语块要素类型	名词的要素语块文本	动词的要素语块文本
SCENE （场景语块）	请绘制一张图片，清晰展示【梳子】的整体形象，或其最具代表性的部分	绘制一张图片，在一个教室中展示【夺冠】相关的动作或状态
CORE_vocab （核心语义语块）	展示【梳子】的整体形象	展示【夺冠】相关的动作或状态
PRIMARY_FEATURES （主要特征语块）	例如不同颜色和材质的梳齿。如果适用，可以在一个典型的环境中展示该【梳子】，如理发店的工作台上。	可以展示与【夺冠】相关联的其他动作或进行【夺冠】所涉及到的其他事物。例如一名老师高举奖杯，周围的学生们欢呼雀跃，有的学生举起胜利的手势，有的学生手持庆祝的标语
AUXILIARY_EXPRESSION （辅助描述语块）	同时，考虑使用相关或相反的衬托元素来突出显示该【梳子】的关键特征或用途，例如剪刀、电推子等美发工具	若无法直观展示【夺冠】，可以展示其对动作主体的影响。还可以对比展示【夺冠】与“失败”的不同动作。
PEDAGOGY (教学目的语块)	从而使用户更好地理解【梳子】的意思	从而使用户更好地理解【夺冠】的意思
STYLECULTURE]URE (风格文化语块)	根据目标受众调整图片风格，并确保所使用的象征和场景易于接受和理解，符合不同文化背景。画面人物为典型的中国面孔，背景展现浓郁的中国特色与生活气息。	根据目标受众调整图片风格，并确保所使用的象征和场景易于接受和理解，符合不同文化背景。画面人物为典型的中国面孔，背景展现浓郁的中国特色与生活气息。

注：【】需要生图的教学词汇。

行参数优化，初始学习率设置为1×10^(-4)，权重衰减系数为0.05，总训练步数为4000，批大小设为2。在参数高效微调方面，采用LoRA微调方法，LoRA的秩设为16，缩放系数为32，Dropout比例为0.1。针对公式(6)的加权系数分别设置为$\lambda_{sc}=0.3$、$\lambda_{lm}=0.7$。该参数配置参考了相关研究中的常用取值范围，并结合本任务对结构表达与语言自然性的平衡需求进行设定。参数取值基于验证集开展小规模参数敏感性实验获得。所有基线模型均在相同的数据划分与训练轮数下进行微调，以保证对比结果的公平性。

在第二阶段的强化学习实验中，我们继续以第一阶段微调完成后的Meta-Llama-3-8B-Instruct作为初始化模型采用优化器AdamW进行参数优化，初始学习率设置为1×10^(-4)，权重衰减系数为0.05，总训练步数为4000，批大小设为2。在参数高效微调方面，采用LoRA微调方法，其LoRA的秩设为8，缩放系数为16。针对公式(14)，总奖励由语块一致性奖励、语言流畅性奖励、语块相关性奖励以及图像语义一致奖励加权组合而成，对应权重系数分别为$\alpha_{sc}=0.2$，$\alpha_{lm}=0.2$，$\alpha_{rel}=0.3$，$\alpha_{img}=0.3$上述奖励权重同样通过验证集上的小规模参数敏感性实验确定。为降低梯度方差，我们对奖励进行标准化并引入baseline进行减均值处理。所有对比方法在第二阶段使用相同的奖励设计与训练配置，从而确保强化学习带来的性能提升具有可比性与可复现性。

为全面评估模型性能，本文从结构一致性、语言质量以及跨模态对齐能力三个方面选取多项评价指标进行综合分析。具体包括语块覆盖率、语言困惑度、语块相关性得分、CLIP语义一致性得分(CLIP Score)以及教师主观评分。其中，语块覆盖率用于衡量模型对结构模板语块要素的遵循程度，PPL用于评估生成文本的语言流畅性与自然程度，语块相关性得分反映各语块内容与教学目标之间的语义一致性，CLIPScore用于衡量扩展提示词生成的图片与词汇释义的一致性，而教师主观评分从词义准确性、教学适配性、表达清晰性以及内容完整性四个维度对生成提示词进行人工评价得到。其中，词义准确性用于衡量提示词是否准确体现目标词汇的核心语义；教学适配性用于评估提示词内容是否符合国际中文教学场景下的认知层级与教学目标；表达清晰性用于衡量提示词文本是否自然流畅、逻辑清晰且易于理解；内容完整性则用于评价提示词是否能够较完整地覆盖场景、特征及教学属性等关键信息，各

维度均采用五分制评分方式。邀请2名具备国际中文教育背景且具备相关教学经验的教师作为评估者采用双专家独立盲审模式进行人工评估。本研究取双专家人工评估结果的平均值作为教师主观评分结果，模型测试取重复三次平均值为报告结果。

3.3 实验结果

3.3.1 模型性能比较

现有提示词扩展研究均面向通用图像生成场景（如风格描述、美学增强），与本文所关注的等级匹配、词义指向、教学场景适配等核心需求存在本质差异，直接套用此类模型既无任务对齐性，亦无公平比较的前提。因此本文转向选取在中文语言能力、模型规模与开放程度上具有代表性的大语言模型作为基线，通过"同底座未微调—同规模异架构—强通用闭源"的三维对照体系，系统验证本文提出模型PPEM-VIG的效能。实验选取Meta-Llama-3-8B-Instruct、Qwen2.5-7B-Instruct、Baichuan2-7B-Instruct、QwenMax、GPT-4.5作为基线模型进行对比，概要如下：

1）Meta-Llama-3-8B-Instruct：Llama预训练模型是当前最具影响力的开源大语言模型之一，经过高质量指令微调，具备稳定的指令跟随能力，广泛应用于SFT与RLHF相关研究，具有良好的学术可复现性与社区基础。作为本文PPEM-VIG的骨干底座模型，旨在考察本文提出模型PPEM-VIG相对于基础语言模型的增益幅度。

2）Qwen2.5-7B-Instruct：基于Qwen2.5技术路线构建的7B规模开源大语言模型，在18T高质量语料上进行预训练，并经过多阶段监督微调和强化学习对齐，在中英双语通用任务上表现优异，被广泛用作中文场景下的通用基准模型。区别于以英文语料为主要预训练来源的LLaMA系列模型，Qwen2.5-7B在中文语料方面具有更强的建模能力。因此，选用该模型有助于进一步验证本文PPEM-VIG模型在不同语言背景下的适应性与泛化能力。

3）Baichuan2-7B-Instruct：该模型由百川智能提出，属于Baichuan2系列中的7B参数规模开源大语言模型。其在约2.6T多语言高质量语料上从头进行预训练，并通过监督微调与对齐技术进一步提升模型性能。在MMLU、CMMLU、GSM8K等常用基准上Baichuan2-7B-Instruct相较于同规模开源模型表现出较强竞争力，可作为中文场景下的强通用基线模型。Baichuan2-7B-Instruct与Qwen2.5-7B-Instruct在规模上保持对齐，二者共同构成开源模型的对照组，有效排除参数规模差异对实验结论的干扰。

4）QwenMax：是阿里巴巴推出的通义千问Qwen系列旗舰级闭源大语言模型，基于大规模多语种高质量语料进行预训练，并通过多阶段对齐进一步强化推理、代码生成与多轮对话能力，在中文理解与生成、通用推理以及编程等任务中表现出较强竞争力，可作为云端服务形态下中文场景的强通用基线模型。QwenMax作为闭源通用大语言模型的代表，一方面反映了当前工业级模型在教学提示词生成任务上的性能参考上界，另一方面也用于验证面向特定教学场景的专门化训练策略在领域适配性方面相对于通用模型的有效性。

5）GPT-4.5：GPT-4.5是OpenAI推出的新一代闭源通用大语言模型，基于大规模多语种高质量语料进行预训练，并结合强化学习对齐与多阶段指令优化，在复杂推理、长文本生成、多轮对话以及语义理解等任务中表现出较强能力。相较于传统开源模型，GPT-4.5在教学语义理解与指令遵循方面具有更高稳定性，能够较准确地完成结构化教学提示词生成任务。因此，选用GPT-4.5作为闭源强基线模型，一方面能够反映当前国际先进通用大语言模型在教学Prompt生成任务中的性能水平，另一方面也用于进一步验证本文提出的PPEM-VIG模型在结构约束建模、视觉语义对齐以及教学场景适配等方面相对于通用大模型的优势。

对比实验结果如表2所示：

从表2可知，预训练模型Meta-Llama-3-8B-Instruct由于未经过特定指令微调，其在结构化教学提示词生成任务中的表现较为有限，其语块覆盖率仅为18.54%，语块相关性得分为0.8542，表明模型在模板结构完整性和语义对齐方面均存在明显不足。尽管其语言困惑度为12.96，这主要是因为其生成内容无语块约束使其更接近自然语言分布，但CLIP语义一致性得分仅为0.6027，且教师主观评分仅为2.1，进一步说明纯基础轻量级语言模型无法生成具有严格结构约束的多模态教学提示词生成任务。

经过指令微调的大语言模型在各项指标上均表

表2　PPEM-VIG模型的对比实验结果

Table 2　Comparison Results of the PPEM-VIG Model

模型	语块覆盖率（%）↑	语言困惑度PPL ↓	语块相关性得分 ↑	CLIP语义一致性得分 ↑	教师主观评分 ↑
Meta-Llama-3-8B-Instruct	18.54	**12.96**	0.8542	0.6027	2.1
Baichuan2-7B-Instruct	82.4	28.34	0.9415	0.6203	3.6
Qwen2.5-7B-Instruct：	88.6	22.45	0.9620	0.6196	3.9
QwenMax	95.20	18.66	**0.9985**	0.6288	4.3
GPT-4.5	94.2	17.65	0.9947	0.6207	4.4
PPEM-VIG(ours)	**99.10**	15.82	0.9935	**0.6485**	**4.6**

注：加粗字体为每列最优值，↑代表分数越高模型表现越好，↓代表分数越低模型表现越好。

现出一定程度的提升。Baichuan2-7B-Instruct的语块覆盖率提升至82.40%，语块相关性达到0.9415，同时CLIP语义一致性得分提高至0.6203，教师评分达到3.6。Qwen2.5-7B-Instruct的语块覆盖率达到88.60%，语义一致性和整体结构完成度均优于Baichuan2-7B-Instruct。然而，受制于模型参数量级较小，这两种大语言模型在生成复杂结构化教学模板时，其语言困惑度分别达到28.34和22.45，反映出生成文本中存在一定程度的语句拼接以及部分教学要素表达不连贯等问题。

作为当前性能较强的商业闭源模型，QwenMax在各项基准指标上整体优于开源模型。其语块覆盖率与语块相关性分别达到95.20%和0.9985，语言困惑度降低至18.66，在保持较高语义一致性的同时实现了更好的语言流畅性。并且CLIP语义一致性得分和教师主观评分分别达到0.6288和4.3，表明全参数通用大模型即使未针对特定任务进行专门优化，也能够生成结构较完整、语义合理且具备一定跨模态对齐能力的教学提示词。GPT-4.5在各项基准指标上也整体优于开源模型。其语块覆盖率达到94.2%，语块相关性得分达到0.9947，说明模型能够较好地理解结构化教学模板并准确捕捉教学语义信息。同时，其语言困惑度降低至17.65，教师主观评分达到4.4，表明生成结果在语言流畅性与教学适配性方面具有较高质量。然而，GPT-4.5和，QwenMax的CLIP语义一致性得分均低于本文提出的PPEM-VIG模型。这表明，通用大语言模型虽然具备较强的语义理解能力，但其生成结果更偏向自然语言描述，在跨模态对齐能力方面仍存在一定局限。相比之下，PPEM-VIG通过结构约束与视觉语义一致性优化，进一步提升了教学提示词与生成图像之间的跨模态匹配效果。

本文提出模型PPEM-VIG方法在多个核心指标上取得最优性能。PPEM-VIG的语块覆盖率达到99.10%，较于最优基线模型QwenMax进一步提升约4.10%，表明模型已充分内化教学要素的结构约束；其语言困惑度进一步降低至15.82，相较于GPT-4.5降低约10.37%，在保证结构规范性的同时有效缓解了传统指令微调中常见的流畅度下降问题。在面向图像生成质量的关键指标上，PPEM-VIG的CLIP语义一致性得分提升至0.6485，相较于QwenMax提高约3.13%，为所有对比方法中最高，表明生成的提示词能够更有效地驱动扩散模型生成符合教学目标的图像内容。同时，在保持0.9935高语块相关性的前提下，其教师主观评分达到4.6，进一步验证了模型在教学适配性与实际应用价值方面的优势。综上所述，PPEM-VIG不仅能够有效建模结构化教学意图，还在文本提示与视觉生成结果之间实现了更优的跨模态对齐。

3.3.2　消融分析

为了验证本研究提出的两阶段模型中各个关键模块以及强化学习多维奖励设计对PPEM-VIG都有一定的提升作用，本文设计模块级消融与细粒度奖励级消融去除不同成分，然后进行实验评估。

1）　模块级消融变体

（1）W/O-RL-SFT：去除监督学习与强化学习两个阶段，即退化为基础模型（Meta-Llama-3-8B-Instruct）。在该设置下，模型不具备结构化模板约

束，仅进行无约束的自然语言生成。

（2）W/O-RL：仅保留监督学习阶段（SFT-only），不引入强化学习优化。在该设置下，模型通过结构一致性与语言流畅性损失学习生成符合教学模板的提示词，但不进行跨模态对齐优化。

1） 阶段级消融变体

（1）W/O-R_{sc}：去除强化学习阶段的结构一致性奖励，旨在验证该奖励对维持模板语块完整性的贡献。

（2）W/O-R_{lm}：去除语言流畅性奖励，旨在验证该信号对控制生成文本自然度（PPL）的作用。

（3）W/O-R_{rel}：去除结构相关性奖励，旨在验证大模型知识蒸馏评估在维持词义与教学目标一致性中的必要性。

（4）W/O-R_{img}：去除图像语义一致性奖励，旨在验证基于 CLIP 的跨模态奖励在提升最终文生图质量对齐中的核心贡献。

（5）PPEM-VIG：完整模型，包含监督学习与强化学习两个阶段。其中监督学习阶段用于学习结构化提示词模板，强化学习阶段进一步通过多奖励信号优化跨模态生成能力。

消融实验结果如表3所示：

表3 PPEM-VIG模型的消融实验结果

Table 3 Ablation Study Results of the PPEM-VIG Model

消融方案	模型	语块覆盖率（%）↑	语言困惑度 PPL↓	语块相关性得分↑	CLIP语义一致性得分↑	教师主观评分↑
模块消融	W/O-RL-SFT	18.54	**12.96**	0.8542	0.6027	2.1
	W/O-RL	97.82	14.96	0.9958	0.6342	4.3
奖励权重消融	W/O-R_{sc}	71.36	21.72	0.8124	0.6164	2.9
	W/O-R_{lm}	94.42	34.58	0.9237	0.6233	3.6
	W/O-R_{rel}	83.27	19.41	0.7428	0.6282	2.4
	W/O-R_{img}	94.53	21.84	0.9715	0.6016	3.3
完整模型	PPEM-VIG	**99.10**	15.82	**0.9935**	**0.6485**	**4.6**

注：加粗字体为每列最优值，↑代表分数越高模型表现越好，↓代表分数越低模型表现越好。

表3展示本文提出的PPEM-VIG模型去除不同模块下以及各个奖励权重的消融实验结果。从模块级消融实验数据显示，与基础模型相比，引入监督学习阶段后，模型在语块覆盖率和语块相关性方面均获得显著提升。其中语块覆盖率由18.54%提升至97.82%，语块相关性得分由0.8542提升至0.9958，同时教师主观评分也由2.1提升至4.3。这表明监督学习阶段能够有效引导模型学习结构化教学模板，从而生成符合教学要求的提示词。在此基础上进一步引入强化学习后（PPEM-VIG），模型在CLIP语义一致性得分上由0.6342提升至0.6485，教师评分提升至4.6，说明强化学习阶段能够在已有结构基础上进一步增强文本与图像之间的语义一致性，从而提升生成提示词的跨模态表达能力。需要注意的是，引入强化学习后，语块覆盖率和语块相关性指标整体保持稳定，说明强化学习阶段并未破坏监督学习阶段所建立的结构表达能力，而是在此基础上进行补充优化。其中语块相关性得分从0.9958下降到0.9935，因为在监督学习阶段模型更倾向于学习训练数据中的固定模板表达与高频语义映射，因此在基于关键词匹配与语义相似度计算的相关性指标上更容易取得较高分数。而在引入强化学习后，模型不仅需要优化文本语义相关性，还需要同时兼顾图像语义一致性、语言流畅性以及结构完整性等多个优化目标，因此生成的提示词内容会更加多样化与开放化。该过程虽然可能导致语块相关性指标出现轻微波动，但能够有效提升模型整体的跨模态表达能力与生成质量。

3.3.3 案例分析

为进一步验证PPEM-VIG在实际教学场景中的生成效果，本文选取典型名词“梳子”与动词“夺冠”作为案例，在不同年龄段设置条件下，对PPEM-VIG

模型生成的结构化拓展提示词及其对应图像进行定性分析。这部分同时展示QwenMax生成的结构化拓展提示词及其对应图像。两者均采用同一个文生图大模型进行图像生成。针对名词“梳子”，PPEM-VIG模型生成结果如表4所示，QwenMax模型生成结果如表5所示。

表5　通义千问模型的名词案例展示

Table 5　Noun-Based Case Examples of the Qwen Model

教学要素集合	模型生成提示词
"vocab":"梳子" "type":"名词" "age":"5-12" "level":"7~9" "definition":"一种整理毛发的用具成上面有齿，用木头、塑料等制"	[SCENE]: A bright and cozy corner of a child's dressing table, soft natural morning light, blurred background with colorful toys, shallow depth of field. [CORE_vocab]: A close-up of a wooden comb as the main subject, clearly showing the structure with neat teeth, designed for hair care. [PRIMARY_FEATURES]: Smooth cherry wood texture, rounded teeth safe for scalp, ergonomic handle, high detail, 8k resolution. [AUXILIARY_EXPRESSION]: The comb is gently resting on a strand of smooth hair, implying the action of grooming and tidying hair. [PEDAGOGY]: Contrast between the solid texture of the wood and the softness of the hair, warm tones of the comb against a cool pastel background. [STYLE_CULTURE]: 3D Pixar animation style, educational illustration for children, vibrant colors, friendly atmosphere, high quality, macro photography style.
"vocab":"梳子" "type":"名词" "age":"12-18" "level":"7~9" "definition":"一种整理毛发的用具成上面有齿，用木头、塑料等制"	[SCENE]：放置在浅色橡木纹理的桌面上，清晨的自然光从左侧窗户射入。 [CORE_vocab]：一把精致的宽齿檀木梳子。 [PRIMARY_FEATURES]：深褐色的木质纹理清晰可见，梳齿打磨圆润光滑，梳背有微微的人体工学弧度，表面泛着温润的光泽。 [AUXILIARY_EXPRESSION]：梳子旁边散落着几根刚刚梳理过的柔顺长发，空气中漂浮着细微的尘埃粒子，展现静谧的氛围。 [PEDAGOGY]：深色木梳与浅色桌面的色彩对比，木质的哑光与头发表面的光泽对比。 [STYLE_CULTURE]：8k分辨率，微距摄影，浅景深，极致细节，国家地理杂志静物风格。

表5续表

教学要素集合	模型生成提示词
"vocab":"梳子" "type":"名词" "age":"18以上" "level":"7~9" "definition":"一种整理毛发的用具成上面有齿，用木头、塑料等制"	1. [SCENE] – 场景 现代简约风格的浴室洗手台或充满阳光的卧室梳妆台一角。背景不再是模糊的玩具，而是具有生活气息的细节，如大理石台面、散落的几根发丝、或是一瓶极简设计的洗发水，营造出清晨洗漱或睡前护理的真实生活氛围。 2.[CORE_vocab]– 核心词汇 一把极具质感的梳子作为绝对视觉中心。它不仅仅是一个工具，更是一件设计品。重点展示其作为“整理毛发用具”的功能性结构，强调梳背的弧度与梳齿的排列规律，体现其在日常生活中的实用美学。 3. [PRIMARY_FEATURES] – 主要特征 材质表现需达到高保真级别。若是木梳，需清晰展现细腻的木纹肌理和温润的色泽（如檀木或榉木）；若是塑料梳，则需表现磨砂或高光的现代工业质感。梳齿部分需刻画精细，根部圆润，排列紧密而整齐，体现出精湛的制造工艺。 4. [AUXILIARY_EXPRESSION] – 辅助表达 通过细节暗示“整理毛发”的动作。例如，梳齿间缠绕着一两根自然的发丝，或者梳子正放置在散开的长发之上，亦或是手持梳子正在梳理的瞬间抓拍。这种辅助表达增加了画面的叙事性，让定义中的“整理”一词具象化。 5.[PEDAGOGY]– 辅助教学 利用材质的对比来增强视觉冲击力。例如，坚硬光滑的梳子材质与柔软蓬松的头发形成质感对比；或者梳子深沉的木色与明亮洁净的大理石台面形成色彩明暗对比。光影上，使用侧逆光勾勒梳子的轮廓，使其从背景中立体地浮现出来。 6. [STYLE_CULTURE] – 风格与文化 采用高清晰度的商业产品摄影风格或超写实主义风格。色调偏向莫兰迪色系或自然暖调，追求光影的细腻过渡和材质的真实触感。画面需干净、高级，符合青少年及年轻成人的审美偏好，类似于生活方式杂志的配图。

注：【】需要生图的教学词汇。

PPEM-VIG模型生成结果体现出较强的结构化语义建构能力与分层认知引导特征。从结构化扩展提示词文本来看，该模型围绕“梳子”构建了清晰的

(a)低认知等级生成图 (b)中认知等级生成图 (c)高认知等级生成图

((a) low cognitive level; (b) moderate cognitive level; (c) high cognitive level)

图2　梳子不同认知等级案例图-PPEM-VIG

Fig. 2　Examples of Different Cognitive Levels in the PPEM-VIG Comb Task

表4　PPEM-VIG模型的名词案例展示

Table 4　Noun-Based Case Examples of the PPEM-VIG Model

教学要素集合	结构化扩展提示词
"vocab":"梳子" "type":"名词" "age":"5-12" "level":"7~9" "definition":"一种整理毛发的用見成上面有齿,用木头、塑料等制"	[SCENE] 请绘制一张图片,清晰展示【梳子】的整体形象,或其最具代表性的部分 [CORE_vocab] 展示【梳子】的整体形象 [PRIMARY_FEATURES] 例如梳子齿和手柄设计可爱,可以用不同颜色或图案来装饰,吸引儿童注意。如果适用,可以在一个典型的场景中展示该【梳子】,如可以展示梳子在梳理一个娃娃或一个小熊玩偶的头发,让儿童感到亲近和有趣。 [AUXILIARY_EXPRESSION] 同时,考虑使用相关或相反的衬托元素来突出显示该【梳子】的关键特征或用途,例如柔软的头发和精致的发饰 [PEDAGOGY] 从而使用户更好地理解【梳子】的意思 [STYLE_CULTURE]根据目标受众调整图片风格,并确保所使用的象征和场景易于接受和理解,符合不同文化背景。画面人物为典型的中国面孔,背景展现浓郁的中国特色与生活气息。
"vocab":"梳子" "type":"名词" "age":"12-18" "level":"7~9" "definition":"一种整理毛发的用見成上面有齿,用木头、塑料等制"	[SCENE] 请绘制一张图片,清晰展示【梳子】的整体形象,或其最具代表性的部分 [CORE_vocab] 展示【梳子】的整体形象 [PRIMARY_FEATURES] 例如梳子齿和手柄设计简约大方,可以用纯色或几何图案来装饰,符合青少年的审美。如果适用,可以在一个典型的场景中展示该【梳子】,如可以展示梳子在梳理一个时尚发型或一个运动发型,展现青少年的活力和个性。 [AUXILIARY_EXPRESSION] 同时,考虑使用相关或相反的衬托元素来突出显示该【梳子】的关键特征或用途,例如柔软的头发和精致的发饰 [PEDAGOGY]从而使用户更好地理解【梳子】的意思 [STYLE_CULTURE]根据目标受众调整图片风格,并确保所使用的象征和场景易于接受和理解,符合不同文化背景。画面人物为典型的中国面孔,背景展现浓郁的中国特色与生活气息。
"vocab":"梳子" "type":"名词" "age":"18以上" "level":"7~9" "definition":"一种整理毛发的用見成上面有齿,用木头、塑料等制"	[SCENE] 请绘制一张图片,清晰展示【梳子】的整体形象,或其最具代表性的部分 [CORE_vocab]展示【梳子】的整体形象 [PRIMARY_FEATURES] 例如梳子设计典雅大气,可以选用优质木材或金属材质,并搭配精美的雕刻或镶嵌,彰显成人的品味和身份。如果适用,可以在一个典型的场景中展示该【梳子】,如可以展示梳子在梳理一个成熟稳重的发型,展现成人的优雅和自信。 [AUXILIARY_EXPRESSION] 同时,考虑使用相关或相反的衬托元素来突出显示该【梳子】的关键特征或用途,例如柔软的头发和精致的发饰 [PEDAGOGY] 从而使用户更好地理解【梳子】的意思 [STYLE_CULTURE]根据目标受众调整图片风格,并确保所使用的象征和场景易于接受和理解,符合不同文化背景。画面人物为典型的中国面孔,背景展现浓郁的中国特色与生活气息。

注:【】需要生图的教学词汇。

名词语义框架,在[PRIMARY_FEATURES]中细化材质(如木质,金属材质)、梳子的实际使用场景等关键视觉特征提示。在低认知等级段,如图2(a)所示,图像以卡通化表达为主,如“儿童为玩偶梳头”,

(a)低认知等级生成图 (b)中认知等级生成图 (c)高认知等级生成图

((a) low cognitive level; (b) moderate cognitive level; (c) high cognitive level)

图3　梳子不同认知等级案例图-通义千问

Fig. 3　Examples of Different Cognitive Levels in the Qwen Comb Task

突出工具与用途的对应关系；中认知等级段则强化真实人物使用情境，如“青少年梳理发型”，如图2(b)所示，增强物体与行为之间的关联；如图2(c)所示，高认知等级段呈现材质细节与审美属性“木梳纹理、工艺质感”并结合生活化环境“梳妆台、自然光线”，体现出从物体识别向功能理解与审美感知的递进。整体来看，PPEM-VIG实现了从静态实体到情境化使用的多层语义展开，使提示词与图像在“结构—功能—情境”三个层面保持高度一致。

教师主观评估表明，PPEM-VIG模型生成的结构化扩展提示词结果能够有效支持学习者对名词“梳子”的多维理解。图像不仅呈现梳子的物体外观，还通过与头发、人物及环境的互动，帮助学习者理解梳子用途以及使用场景。在低认知等级段，卡通形象降低认知门槛，强化形状与用途的直观联系；中年龄段通过真实人物的使用行为，使学习者建立“工具—动作”的语义连接；高认知等级段则通过材质、光影与生活场景的刻画，引导学习者关注细节与文化语境。这种由“形态识别—功能理解—情境扩展”的视觉递进路径，与文本中的结构化语义模块形成良好对齐，使名词语义从单一标签扩展为具有功能与情境的认知单元。

QwenMax模型生成的提示词在“梳子”的提示词与图像生成中呈现出一定的语义简化倾向。从文本角度看，其提示词虽同样包含[SCENE]与[PRIMARY_FEATURES]等结构，但语义重心更偏向对“梳子外观”的静态描述(如材质、光泽、分辨率等)，对“使用过程”与“功能情境”的刻画相对有限，不同语块之间存在一定程度的内容重复与功能弱化。例如，[AUXILIARY_EXPRESSION]中虽提及“梳齿间缠绕着一两根自然的发丝”，但更多停留在视觉摆放层面，缺乏明确的动作指向与事件结构组织。在对应图像中，这种倾向表现为“物体中心化”的呈现方式：梳子多以特写形式出现，强调纹理与质感，整体更接近产品摄影或静物展示。尽管在不同认知等级段引入了不同风格(如3D卡通、微距摄影、生活场景)，但语义变化主要体现在视觉风格层面，而非认知层级的递进。如图3(a)所示，儿童场景中虽有柔和色彩与玩具背景，但梳子仍以孤立物体为主体；高认知等级段，如图3(c)所示虽出现洗漱台等环境元素，但这些元素更多作为装饰性背景存在，未能与“梳理头发”的功能形成紧密关联。

这种“强结构细节、弱功能语义”的表达方式，使名词“梳子”的多模态呈现偏向静态物体展示，而缺乏“工具—行为—情境”的系统建构。其结果是，学习者虽能够识别梳子的外观特征，但较难通过图像理解其使用方式及实际语境，尤其在分级教学中缺乏明显的认知递进与语义深化路径。。

针对动词“夺冠”，PPEM-VIG模型生成结果如表6所示，QwenMax模型生成结果如表7所示。

相比于具体名词，动词“夺冠”的案例具有更强的抽象性与动态语义，对跨模态建模能力提出了更高要求。在动词“夺冠”的多模态表达中，PPEM-VIG模型生成结果展现出显著的过程语义建模能力与情境递进特征。从提示词文本来看，该模型能够围绕“夺冠”构建完整的动作语义框架，不仅在[PRIMARY_FEATURES]中明确刻画“高举奖杯”“表情兴奋”等核心动作，还通过[AUXILIARY_EXPRESS

ION]与[CONTRAST]语块引入“队友互动”“失败对比”等辅助语义，从而将动词由单一行为扩展为包含参与者、情境与结果的事件结构。这种结构化

表6　PPEM-VIG模型的动词案例展示

Table 6　Verb-Based Case Examples of the PPEM-VIG Model

教学要素集合	结构化扩展提示词
"vocab":"夺冠" "type":"动词", "age":"5-12" "level":"7~9" "definition":"在比赛中获得冠军"	[SCENE]绘制一张图片,在一个颁奖典礼中展示【夺冠】相关的动作或状态 [CORE_vocab]可以展示与【夺冠】相关联的其他动作或进行【夺冠】所涉及到的其他事物。 [PRIMARY_FEATURES]例如儿童运动员胸前戴着闪亮的大金牌,双手努力举起一个比自己头还大的金色奖杯,脸上露出天真烂漫、极度兴奋的笑容[AUXILIARY_EXPRESSION]若无法直观展示【夺冠】,可以展示其对动作主体的影响,例如展示可以飞扬的五彩纸屑和彩色气球,周围的小伙伴们都在欢快地拍手跳跃。 [PEDAGOGY]还可以对比展示【夺冠】与"落选"的不同动作。 [STYLE_CULTURE]根据目标受众调整图片风格,并确保所使用的象征和场景易于接受和理解,符合不同文化背景。画面人物为典型的中国面孔,背景展现浓郁的中国特色与生活气息。
"vocab":"夺冠" "type":"动词", "age":"12-18" "level":"7~9" "definition":"在比赛中获得冠军"	[SCENE]绘制一张图片,在一个颁奖典礼中展示【夺冠】相关的动作或状态 [CORE_vocab]可以展示与【夺冠】相关联的其他动作或进行【夺冠】所涉及到的其他事物。 [PRIMARY_FEATURES]例如儿童运动员胸前戴着闪亮的大金牌,双手努力举起一个比自己头还大的金色奖杯,脸上露出天真烂漫、极度兴奋的笑容[AUXILIARY_EXPRESSION]若无法直观展示【夺冠】,可以展示其对动作主体的影响,例如展示可以飞扬的五彩纸屑和彩色气球,周围的小伙伴们都在欢快地拍手跳跃。 [PEDAGOGY]还可以对比展示【夺冠】与"落选"的不同动作。 [STYLE_CULTURE]根据目标受众调整图片风格,并确保所使用的象征和场景易于接受和理解,符合不同文化背景。画面人物为典型的中国面孔,背景展现浓郁的中国特色与生活气息。
"vocab":"夺冠" "type":"动词", "age":"18以上" "level":"7~9" "definition":"在比赛中获得冠军"	[SCENE]绘制一张图片,在一个颁奖典礼中展示【夺冠】相关的动作或状态[CORE_vocab]可以展示与【夺冠】相关联的其他动作或进行【夺冠】所涉及到的其他事物。 [IPRIMARYFEATURES]例青少年获奖者高举奖杯,脸上洋溢着自信和自豪,向台下观众挥手致意 [AUXILIARY_EXPRESSION]若无法直观展示【夺冠】,可以展示其对动作主体的影响,例如展示可以展示获奖者与队友们热烈拥抱,分享成功的喜悦。 [PEDAGOGY]还可以对比展示【夺冠】与"落选"的不同动作。 [STYLE_CULTURE]根据目标受众调整图片风格,并确保所使用的象征和场景易于接受和理解,符合不同文化背景。画面人物为典型的中国面孔,背景展现浓郁的中国特色与生活气息。

注:【】需要生图的教学词汇。

语义在图像中得到有效映射:如图4(a)所示,低认知等级段侧重直观动作呈现(儿童举杯与欢呼),如图4(b)所示,中认知等级段强化群体互动与庆祝氛围,如图4(c)所示,高认知等级段则进一步融入情感张力与社会语境(如团队拥抱、观众反馈等)。由此形成从"动作识别"到"情境理解"的递进式表达路径,实现提示词与图像之间的高度一致。

教师主观评估指出,PPEM-VIG模型生成的图像能够通过图文协同呈现"过程性语义",有效支持学习者对动词意义的动态建构与迁移理解。图像并非停留于单一"获胜结果"的呈现,而是围绕"夺冠"这一动作构建了完整的事件链条:在低认知等级段,如图5(a)所示,图像以儿童高举奖杯、周围同伴欢呼的场景为主,突出"获得胜利"的直观认知;如图5(b)在中认知等级段,进一步强化群体互动与情绪共享,如队友围绕、集体庆祝,使"夺冠"从个体行为扩展为社会性事件;如图5(c)在高认知等级段,则通过人物姿态、团队拥抱及环境氛围的刻画,呈现出更具张力的胜利瞬间与情感释放。这种由"结果呈现—情境扩展—情感深化"的视觉递进路径,与文本中对动作、参与者及环境的结构化描述形成高度一致,使动词语义从静态标签转化为动态过程表达。

QwenMax模型生成的提示词在提示词与图像表达上呈现出明显的语义压缩倾向。从文本角度看,其提示词虽在结构上较为完整,但语义重心集中于"获胜后的状态描述",对"夺冠"这一动作过程的展开不足,导致不同语块之间存在一定程度的语义重叠与功能弱化。其对应的生成图像,虽在视觉细节

表7　通义千问模型的动词案例展示

Table 7　Verb-Based Case Examples of the Qwen Model

教学要素集合	通用模型扩展提示词
"vocab":"夺冠" "type":"动词", "age":"5-12" "level":"7~9" "definition":"在比赛中获得冠军"	[SCENE]:一个阳光明媚的户外体育场或舞台，背景中有模糊的欢呼人群和飘落的彩色纸屑，营造出胜利和庆祝的宏大氛围。 [CORE_vocab]:一个可爱的卡通儿童（约7-9岁）双手高高举起一个巨大的金色奖杯，动作充满力量感和喜悦。 [PRIMARY_FEATURES]:奖杯金光闪闪，细节清晰，上面刻有“冠军”或“第一名”的字样；儿童穿着运动服，脸上洋溢着灿烂、自豪的笑容，眼睛笑成了弯弯的月牙。 [AUXILIARY_EXPRESSION]:周围有队友在鼓掌，或者孩子头上戴着花环，金色的奖杯在阳光下反射出耀眼的光芒，象征着荣誉的获得。 [PEDAGOGY]:孩子鲜艳的服装与金色奖杯形成色彩呼应，与背景中稍暗的人群形成明暗对比，突出了胜利者的主体地位。 [STYLE_CULTURE]:3D皮克斯动画风格，色彩鲜艳明快，光影柔和，画面充满正能量和童趣，适合儿童教育认知。
"vocab":"夺冠" "type":"动词", "age":"12-18" "level":"7~9" "definition":"在比赛中获得冠军"	[SCENE]:大型体育场馆的领奖台中央，背景是虚化的观众席和漫天飞舞的金色纸屑，聚光灯从上方打下，营造出万众瞩目的焦点氛围。 [CORE_vocab]:一位少年运动员（约12-14岁）身着国家队或校队队服，双手紧紧抱着一座巨大的、雕刻精美的冠军奖杯，将其举至胸前或头顶。 [PRIMARY_FEATURES]:奖杯的金属质感强烈，反射着光芒，象征着至高无上的荣誉；少年的面部表情混合着激动的泪水与自豪的微笑，汗水顺着脸颊滑落，细节刻画入微。 [AUXILIARY_EXPRESSION]:脖子上挂着金牌，队友在身后模糊的背景中欢呼拥抱，或者少年紧闭双眼沉浸在胜利的喜悦中，强化“获得冠军”的结果。 [PEDAGOGY]:奖杯冰冷坚硬的金属光泽与少年温热充满活力的皮肤质感形成对比；前景人物的清晰锐利与背景喧闹人群的虚化形成景深对比，突出“夺”得的荣耀瞬间。 [STYLE_CULTURE]:电影级写实摄影风格，高动态范围，色彩饱满且具有戏剧性，类似体育杂志封面的质感，强调真实感与英雄主义色彩。
"vocab":"夺冠" "type":"动词", "age":"18以上" "level":"7~9" "definition":"在比赛中获得冠军"	[SCENE] 激烈的决赛现场，聚光灯汇聚的中心舞台或竞技赛场，背景是模糊但沸腾的观众席，空气中弥漫着紧张与兴奋的氛围。 [CORE_vocab]一位15-20岁的青年运动员，在赢得决定性胜利的瞬间，展现出“夺取”冠军的动作与神态。 [PRIMARY_FEATURES] 人物面部表情坚毅且充满激情，眼神中透露出胜利的渴望与释放；肌肉线条因发力或激动而紧绷，汗水清晰可见；手中紧握象征冠军的奖杯、奖牌或旗帜。 [AUXILIARY_EXPRESSION] 周围队友激动地冲上来拥抱，或者对手失落地站在背景中；漫天飘落的彩带或闪烁的镁光灯，强化了夺冠时刻的辉煌与荣耀。 [PEDAGOGY] 主角清晰锐利的面部特写与背景中狂喜或失落的人群形成虚实对比；胜利者激昂的情绪与赛场残酷的竞技氛围形成情感对比；明亮的聚光灯与周围较暗的环境形成光影对比。 [STYLE_CULTURE] 高清晰度的体育摄影风格或写实主义数字绘画，色彩饱和度高，光影具有戏剧性，强调力量感、速度感与青春的张力，符合青少年审美。

注：【】需要生图的教学词汇。

与表现力上具有一定水准，但整体更倾向于对“夺冠结果”的单点强化。尽管在不同年龄段中采用了卡通风格、体育摄影等多样视觉形式，但语义层面变化有限，主要围绕“胜利者形象”展开，缺乏对动作过程与情境发展的系统刻画。此外，文本中虽引入辅助元素（如观众、队友）但更多作为背景装饰存在，未能形成围绕动作展开的语义组织。这种“强结果、弱过程”的表达方式，使得动词“夺冠”的语义被简化为静态成就展示，而非动态行为建构。其难以支持学习者对“夺冠”这一动作过程的深入理解，尤其在分级教学场景中缺乏明显的语义递进与认知引导。

(a)低认知等级生成图 (b)中认知等级生成图 (c)高认知等级生成图
((a) low cognitive level; (b) moderate cognitive level; (c) high cognitive level)

图4　夺冠不同认知等级案例图-PPEM-VIG

Fig. 4　Examples of Different Cognitive Levels in the PPEM-VIG Championship Task

(a)低认知等级生成图 (b)中认知等级生成图 (c)高认知等级生成图
((a) low cognitive level; (b) moderate cognitive level; (c) high cognitive level)

图5　夺冠不同认知等级案例图-通义千问

Fig. 5　Examples of Different Cognitive Levels in the Qwen Championship Task

4　结　论

本文针对文本生成图像技术在国际中文词汇教学场景中面临的提示词语义失配与教学知识缺位问题，提出面向国际中文教育词汇图像生成的渐进式提示词扩展模型PPEM-VIG。该模型将词汇的认知等级、难度等级、词性、释义等离散教学要求统一建模为结构化教学要素集，通过监督学习与强化学习的渐进式两阶段优化，实现领域知识嵌入、语言结构规范化与词图语义一致性的协同优化，为教学图像的自动化精准生成提供一套完整的方法论框架。在数据资源建设方面，本文构建首个面向国际中文教育场景的词汇图像提示词数据集ICLE-VIPD，涵盖5682个严格对齐《国际中文教育中文水平等级标准》的目标词汇，覆盖初、中、高三等九级，以三元组形式组织标注数据，为本领域后续研究提供可复用的基准资源。

实验结果表明，PPEM-VIG在模板结构覆盖率、CLIP语义一致性得分、图像匹配度及教师主观评价等核心指标上超越Qwen2.5-7B、Baichuan2-7B及QwenMax等基线模型，验证了渐进式监督-强化优化范式在词汇教学图像生成任务中的有效性。然而，本文仅聚焦名词与动词两类词性，未来可将本框架迁移至抽象词汇、其他小语种等教学场景的图像语义表达，进一步提升模型在真实教学场景中的鲁棒性和泛化能力。

参考文献(References)

Song F and Abdul Rabu S N. 2025. Trends, advantages, and challenges: a systematic literature review of artificial intelligence in design education. Journal of Educational and Social Research, 15(4): 401-417 [DOI: 10.36941/jesr-2025-0147]

Zhao X T. 2021. On the important position of vocabulary teaching in international Chinese teaching. Xin Jishi, (9): 45-47 (赵雪婷. 2021. 试论词汇教学在国际中文教学中的重要地位. 新纪实, (9): 45-47)

Rao J. 2021. Effects of images and subtitles on incidental vocabulary acquisition of eighth-grade students under audio-visual input. Changsha: Hunan University (饶静. 2021. 视听输入下图像和字幕对初二学生词汇附带习得的影响. 长沙: 湖南大学)

Zheng Y Q and Chen W H. 2006. A study on picture expression methods of HSK nouns. Chinese Teaching in the World, (4): 107-115 (郑

艳群，陈文慧. 2006. HSK名词图片表达方法研究. 世界汉语教学，(4)：107-115)［DOI:10.13724/j.cnki.ctiw.2006.04.016］

Lu N. 2009. Analysis on the picture expressibility of HSK verbs from semantic features. Beijing: Beijing Language and Culture University (卢娜. 2009. 从语义特征分析HSK动词的图片可表达性. 北京：北京语言大学)

Wang J. 2022. Alignment and adaptation: vocabulary teaching strategies based on the Chinese Proficiency Grading Standards for International Chinese Language Education. Journal of International Chinese Teaching, (4): 10-19 (王军. 2022. 对接与调适：基于《国际中文教育中文水平等级标准》的词汇教学策略. 国际汉语教学研究，(4)：10-19)

Bharne S, Sapkale P, Sarda E, Salunkhe S and Padiya P. 2025. Towards sustainable image synthesis: a comprehensive review of text-to-image generation models. International Research Journal of Multidisciplinary Technovation, 7(5): 94-120［DOI: 10.54392/irjmt2557］

Kadry K, Gupta S, Nezami F R and Edelman E R. 2024. Probing the limits and capabilities of diffusion models for the anatomic editing of digital twins. npj Digital Medicine, 7(1)［DOI:10.1038/s41746-024-01332-0］

Cao Y. 2024. Study on the utilization of pictures in vocabulary instruction for international Chinese language education. Journal of Education and Educational Research, 9(1): 19-22［DOI: 10.54097/dbkgdg2

Sudha L, Aruna K B, Sureka V, Niveditha M and Prema S. 2024. Semantic image synthesis from text: current trends and future horizons in text-to-image generation. EAI Endorsed Transactions on Internet of Things, 11［DOI:10.4108/eetiot.5336］

Purwono P, Wulandari A N E, Ma'arif A and Salah W A. 2025. Understanding generative adversarial networks: a review. Control Systems and Optimization Letters, 3(1): 36-45［DOI:10.59247/csol.v3i1.170］

Ji Y R, Wang C H, Chen J B, Yue A Z, Xi Z H and Chen J S. Parameter-efficient diffusion model adaptation and spectral consistency learning for controllable multispectral remote sensing image generation[J/OL]. Journal of Image and Graphics: 1-16 (纪璎芮，王晨昊，陈静波，岳安志，席智浩，陈建胜. 多光谱遥感图像可控生成的扩散模型参数高效适配与光谱一致性学习[J/OL]. 中国图象图形学报：1-16)［DOI:10.11834/jig.260089］

Shafi S. 2024. Text-to-image generation using stack generative adversarial networks and stable diffusion models. International Journal for Research in Applied Science and Engineering Technology, 12(11): 1426-1429［DOI:10.22214/ijraset.2024.65350］

Wang R. 2024. A comparative analysis of StackGAN and AttnGAN in text-to-image generation. Applied and Computational Engineering, 105(1): 9-15［DOI:10.54254/2755-2721/105/2024tj0055］

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models[EB/OL].［DOI:10.48550/ARXIV.2006.11239］

Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M. 2022. Hierarchical text-conditional image generation with CLIP latents［EB/OL].［DOI:10.48550/ARXIV.2204.06125］

Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, et al. 2022. Photorealistic text-to-image diffusion models with deep languageunderstanding[EB/OL].［DOI:10.48550/ARXIV.2205.11487］

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2021. High-resolution image synthesis with latent diffusion models［EB/OL].［DOI:10.48550/ARXIV.2112.10752］

Schneider J. 2024. Explainable generative AI: a survey, conceptualization, and research agenda. Artificial Intelligence Review, 57(11)［DOI:10.1007/s10462-024-10916-x］

Yang L, Zhang Z, Song Y, Hong S, Xu R, Zhao Y, et al. 2023. Diffusion models: a comprehensive survey of methods and applications. ACM Computing Surveys, 56(4): 1-39［DOI:10.1145/3626235］

Chen Z, Zhang L, Weng F, Pan L and Lan Z. 2024. Tailored visions: enhancing text-to-image generation with personalized prompt rewriting//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE/CVF: 7727-7736［DOI:10.1109/CVPR52733.2024.00792］

Liu V and Chilton L B. 2022. Design guidelines for prompt engineering text-to-image generative models//Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. New Orleans, USA: ACM: 1-23［DOI:10.1145/3491102.3501825］

Cao T, Wang C, Liu B, et al. 2023. Beautifulprompt: towards automatic prompt engineering for text-to-image synthesis//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track. Singapore: Association for Computational Linguistics: 1-11［DOI: 10.18653/v1/2023.emnlp-industry.1］

Lu Y H, Feng Y C J, Zhu L, Zhou H Y, Zhu H, Yu C H, et al. 2025. PromptVis: prompt-based interactive visual analysis method for text-to-image creation. Journal of Computer-Aided Design & Computer Graphics, 37(4): 688-696 (卢裕弘，封颖超杰，朱琳，周海怡，朱航，喻晨昊，等. 2025. PromptVis：面向文本生成图片的提示词的交互式可视分析方法. 计算机辅助设计与图形学学报)，37(4)：688-696［DOI:10.3724/SP.J.1089.2023-00344］

Fernando DVA, Atharva T, Patil A, Sinha D and Ludup T.2025. Harnessing stable diffusion model for high-resolution text-to-image synthesis. International Journal of Scientific Research in Engineering and Management, 9(5): 1-9［DOI:10.55041/ijsrem48372］

Kitrungrotsakul T, Xu Y and Srichola P. 2025. Knowledge distillation meets reinforcement learning: a cluster-driven approach to image processing. Sensors, 26(1): 209［DOI:10.3390/s26010209］

Datta S, Ku A, Ramachandran D, et al. 2024. Prompt expansion for adaptive text-to-image generation//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand: Association for Computational Linguistics: 3449-3476［DOI:10.18653/v1/2024.acl-long.189］

Leong H H M, Chen Q, Ye Z, et al. 2025. PromptBoost: annotation-free prompt expansion for Chinese text-to-image generation//Proceedings of the 3rd International Workshop on Large Generative Models Meet Multimodal Applications: 3-10 [DOI: 10.1145/3728422.3762144]

Li C, Zhang M, Mei Q, Kong W and Bendersky M. 2024. Learning to rewrite prompts for personalized text generation//Proceedings of the ACM Web Conference 2024. Singapore: ACM: 3367-3378 [DOI: 10.1145/3589334.3645408]

Nori H, Lee Y T, Zhang S, Carignan D, Edgar R, Fusi N, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine [EB/OL]. [DOI: 10.48550/arXiv.2311.16452]

Liu B, Wang L, Lyu C, Zhang Y, Su J, Shi S, et al. 2023. On the cultural gap in text-to-image generation [EB/OL]. [DOI: 10.48550/arXiv.2307.02971]

Ahmadi M, Zarei A A and Esfandiari R. 2020. Learning L2 idioms through visual mnemonics. Journal of English Language Teaching and Learning, 12(26): 1-27 [DOI:10.22034/elt.2020.11438]

Choi W C and Chang C I. 2025. A survey of techniques, key components, strategies, challenges, and student perspectives on prompt engineering for large language models in education[EB/OL]. [DOI: 10.20944/preprints202503.1808.v1]

Shi J, Qi M, Zhang L, Wang D, Zhao Y, Li Z, et al. 2025. Collaborative text-to-image generation via multi-agent reinforcement learning and semantic fusion [EB/OL]. [DOI: 10.48550/ARXIV. 2510.10633]

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. 2021. Zero-shot text-to-image generation [EB/OL]. [DOI: 10.48550/ARXIV.2102.12092]

Qi Y and Guo D. 2024. Enhancing text-to-image generation with diversity regularization and fine-grained supervision. Highlights in Science, Engineering and Technology, 122: 1-9 [DOI: 10.54097/42m6by18]

Mayer R E. 2005. The Cambridge handbook of multimedia learning. Cambridge: Cambridge University Press (Mayer R E. 2005. 多媒体学习剑桥手册. 剑桥: 剑桥大学出版社)

Paivio A. 1990. Mental representations: a dual coding approach. Oxford: Oxford University Press (Paivio A. 1990. 心理表征:双重编码法. 牛津: 牛津大学出版社)

Shen H H. 2010. Imagery and verbal coding approaches in Chinese vocabulary instruction. Language Teaching Research, 14(4): 485-499 [DOI:10.1177/1362168810375370]

Li J T and Tong F. 2020. The effect of cognitive vocabulary learning approaches on Chinese learners' compound word attainment, retention, and learning motivation. Language Teaching Research, 24(6): 834-854 [DOI:10.1177/1362168819829923]

Xiong T and Peng Y. 2021. Representing culture in Chinese as a second language textbooks: a critical social semiotic approach. Language, Culture and Curriculum, 34(2): 163-182 [DOI: 10.1080/07908318.2020.1797079]

Gou D. 2025. The potential, challenges, and pathways of generative artificial intelligence in empowering the professional development of international Chinese language teachers. Journal of Current Social Issues Studies, 2(2): 104-115 [DOI:10.56397/JCSIS.2025.02.13]

Yu J, Song J and Lu Y. 2025. Harnessing generative artificial intelligence to construct multimodal resources for Chinese character learning. Systems, 13(8): 692 [DOI:10.3390/systems13080692]

Attygalle N T, Kljun M, Quigley A, et al. 2025. Text-to-image generation for vocabulary learning using the keyword method//Proceedings of the 30th International Conference on Intelligent User Interfaces. Cagliari, Italy: ACM: 1381-1397 [DOI: 10.1145/3670653.3703358]

Fan K. 2025. Generative AI and second language vocabulary processing: a cognitive study of Chinese EFL learners. Journal of Humanities and Social Sciences Studies, 7(7): 103-111 [DOI:10.32996/jhsss.2025.7.7.12]

Shi X C. 2020. Research on Chinese description generation technology for automobile evaluation images. Guangzhou: Guangdong University of Technology (史秀聪. 2020. 汽车评测图像中文描述生成技术研究. 广州: 广东工业大学)

Yang C L, Wan W G, Wang X Z, Sun X T and Zhang Z. 2023. Image semantic understanding based on text perception and non-repeating word generation. Industrial Control Computer, 11: 105-106, 109 (杨晨露, 万旺根, 王旭智, 孙学涛, 张振. 2023. 基于文本感知和非重复单词生成的图像语义理解. 工业控制计算机, 11: 105-106, 109

Zhao X W, Zhu Z T and Shen S S. 2023. Prompt engineering in education: constructing an epistemological new discourse in the era of digital intelligence. China Distance Education, 43(11): 22-31 (赵晓伟, 祝智庭, 沈书生. 2023. 教育提示语工程:构建数智时代的认识论新话语. 中国远程教育, 43(11): 22-31) [DOI:10.13541/j.cnki.chinade.2023.11.003]

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, et al. 2023. LLaMA: open and efficient foundation language models [EB/OL].[2026-05-17].[DOI:10.48550/arXiv.2302.13971]

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J D, Dhariwal P, et al. 2020. Language models are few-shot learners//Proceedings of the 34th International Conference on Neural Information Processing Systems. Virtual: Curran Associates, Inc.: 1877-1901

Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, et al. 2023. Stanford alpaca: an instruction-following LLaMA model[EB/OL].[2026-05-17]. https://github.com/tatsu-lab/stanford_alpaca

Hu E J, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. 2022. LoRA: low-rank adaptation of large language models//Proceedings of the 10th International Conference on Learning Representations. Virtual: OpenReview.net

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. 2022. Training language models to follow instructions with human feedback//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates, Inc.: 27730-27744

Chen S Q, Yang X, Zhu R Q, Liao N and Zhao W W. 2026. Parameter-efficient fine-tuning for remote sensing image interpretation: a survey. Journal of Image and Graphics, 31(1): 0212-0242 (陈诗琪, 杨学, 朱荣强, 廖宁, 赵卫伟. 2026. 面向遥感图像解译的参数高效微调研究综述. 中国图象图形学报, 31(1): 0212-0242) [DOI:10.11834/jig.250105]

Center for Language Education and Cooperation. 2025. Chinese graded readers standards for international Chinese language education. Beijing: Beijing Language and Culture University Press (中外语言交流合作中心). 2025. 国际中文教育中文阅读分级标准. 北京: 北京语言大学出版社)

作者简介

王华珍，女，副教授，研究方向为自然语言处理、知识图谱、人工智能教育。 E-mail：wanghuazhen@hqu.edu.cn

吴鹭平，男，硕士研究生，主要研究方向为人工智能。E-mail：wlp0628@163.com

胡建刚：通讯作者，男，教授，主要研究方向：海外华文教育、国际中文教育。E-mail：hujianganghw@126.com

孙菁：女，讲师，主要研究方向：语言学及应用语言学。E-mail：sunjingaza@126.com

赵毅飞：女，助教，主要研究方向：国际中文教育。E-mail：yifei@hqu.edu.cn