

生成式人工智能支持下看图说话试题图像生成方法研究

季瑞骅 *

石昱欣

张天

赵文书

南京大学, 中国

摘要

生成式人工智能正成为国际中文教育资源开发的重要支撑,但现有研究多关注生成效率与跨模态联动,对生成结果是否符合专业命题要求及其对学习者参与和表达动力的影响缺乏检验。本文围绕看图说话试题图像的自动生成,构建场景、人物、动作、表情、道具五要素生成模板,设计两种教师参与度不同的智能体,以 27 道试题为基准生成图片,从教师评价与学生体验两层面对评估。结果显示,不同方案教育应用效果差异显著,较优方案接近原图水平,较弱方案在教师评价及学生情感接受、话题联想和表达意愿方面均不足。研究表明,生成式人工智能支持国际中文教育资源开发切实可行,但成效不仅取决于模型能力,还与试题需求描述的组织方式及教师参与程度密切相关。

关键词

生成式人工智能, 看图说话, 图像生成, 学习者体验

1 引言

随着教育数字化进程持续推进,信息数字化作为数字化转型的基础阶段,已在教育领域得到广泛开展(汪琼、刘敏, 2023)。以计算机技术、多媒体技术和通信技术为代表的现代信息化技术,逐步成为教育发展与变革的关键影响因素与驱动力,信息化技术进步的同时推动了教育的发展与变革(郑艳群, 2023)。而信息化技术发展的影响同样传递到了国际中文教育领域。尤其近年来,以大模型为代表的生成式人工智能技术飞速发展,国际中文教育资源开发的速度、规模与形态类型均得到显著扩展(张黎、刘敏, 2026),教育资源生成的自动化与智能化日益成为该领域的重要发展方向。与此同时,国际中文教育研究也正在从单纯关注知识传授与资源供给,逐步转向更加重视学习者体验、情绪机制与学习环境建构的综合视角(赵欢等, 2024; 黄芳炜、赵杨, 2026)。

在这一背景下,生成式人工智能技术已经被应用到众多具体场景中,涵盖从生成式图片库建设到情境化视频制作的多元形态(何晴、饶高琦, 2025; 郑小军等, 2025)。其中,教育资源生成的自动化与智能化,是一线国际中文教师较为关注的应用场景之一。有效支持教育资

* 通讯作者。联系电邮: jrhluck@nju.edu.cn

源的精准自动化生成, 将有望缓解一线国际中文教师的备课负担 (Wu *et al.*, 2024)。现有研究重点关注生成效率的提升与跨模态内容的联动, 同时也发现了现有生成式人工智能在教育应用中存在事实性偏差、生成内容图文不符以及情感化、个性化支持不足等问题 (吕振华、叶军, 2024; 欧志刚等, 2024; 童莉莉等, 2024)。

在此基础上, 仍有必要进一步聚焦具体资源类型, 考察生成式人工智能是否能够真正支持国际中文教育资源的专业化生成。基于现有研究 (梁宇、刘根芹, 2024; 张黎、刘敏, 2026), 本文从国际中文教育的应用与测试场景出发, 以看图说话类试题图像的自动生成为研究对象, 利用大模型驱动的生成式人工智能技术, 探讨其是否能够有效支持教学资源的专业化生成。与一般性的图像生成任务不同, 看图说话试题不仅要求生成结果在视觉内容上与命题需求保持一致, 还需要进一步考察其作为考试材料的适用性, 以及其对学习者情感接受、话题联想和口语表达意愿的影响。基于此, 本文从教师专业评价与学生学习体验两个层面构建评估框架, 对不同生成方案的应用效果进行系统分析, 以期为生成式人工智能支持下的国际中文教育资源开发提供更具针对性的实证依据。

基于上述分析, 本文围绕以下三个研究问题展开。

研究问题 1: 基于五要素模板构建的看图说话试题图像生成智能体, 能否在命题准确度和考试适配度方面达到接近原始试题图像的水平?

研究问题 2: 在学习者体验层面, 智能体生成的试题图像与原始试题图像之间是否存在显著差异?

研究问题 3: 在智能体生成图像中, 教师参与需求描述的组织方式是否会对教育应用效果产生显著影响?

2 文献综述

国际中文教育资源已形成较完备的体系与数字基础 (马箭飞等, 2021; 吴应辉等, 2021; 曾君、陆方喆, 2021), 但仍存在供需失衡、更新滞后与适配不足等问题 (柯清超等, 2018; 彭锦维, 2022; 梁宇、刘根芹, 2024)。与此同时, 文本生成图像技术迅速发展, 其文图对齐能力近年来不断提高 (Ramesh *et al.*, 2021, 2022; Rombach *et al.*, 2022; Betker *et al.*, 2023)。相关研究综述表明, 生成式人工智能在提升资源开发效率的同时, 也面临事实性偏差与评价标准缺失等挑战 (Kasneci *et al.*, 2023; Lo, 2023)。在国际中文教育领域, 生成式人工智能已被用于生成式图片库、多媒体教材与情境视频开发 (Chen & Wu, 2024; Wu *et al.*, 2024; 何晴、饶高琦, 2025; 郑小军等, 2025), 但幻觉偏差、情感缺失等问题仍较突出 (吕振华、叶军, 2024; 童莉莉等, 2024; 魏小东、习鹏飞, 2024)。

自然语言处理领域已形成系统的提示词方法谱系 (Liu *et al.*, 2023; White *et al.*, 2023)。在文生图任务中, 提示词通常需要以特定格式组织, 并通过关键词、短语或修饰语控制风格、质量与画面效果 (Liu & Chilton, 2022; Oppenlaender, 2024); 同时, 非专业用户往往难以系统设计并评估有效提示词 (Zamfirescu-Pereira *et al.*, 2023), 教育应用中也需要提示词素养、教师培训与可操作提示策略支持 (Giray, 2023; Lo, 2023)。由此, 试题需求描述的组织方式与教师参与程度, 可能是决定试题图像生成质量的关键变量。

看图说话可视作以视觉材料为任务输入、引导考生进行口语产出的任务形式 (Luoma, 2004)。语言测试研究表明, 作为任务输入的视觉材料并非中性载体, 视觉信息的组织方式与考生对材料的熟悉度会影响口语产出质量, 并可能构成构念无关变异 (Xi, 2005, 2010)。同

时,任务形式、互动条件与技术介入也会影响口语能力构念的操作化与效度解释 (Nakatsuhara, 2013; Nakatsuhara *et al.*, 2017; Galaczi & Taylor, 2018)。这为本文从命题准确度、考试适配度与学习者体验三个层面评估生成图像提供了理论依据。

积极心理学视角下的学习者体验正日益受到国际中文教育和技术中介语言学习研究的关注 (赵欣等, 2026)。已有研究表明,技术接受度、学习愉悦、无聊情绪、教师热情、自我效能等因素会影响学习者的交际意愿、幸福感、课堂参与和学习表现 (Huang *et al.*, 2024; Huang & Zhang, 2025; Zhao & Wang, 2025; Zhao, Liu, *et al.*, 2026)。在数字化与AI中介语言学习情境中,学习动机、愉悦、自我效能、批判数字素养和学习者能动性也会影响学习参与,多模态 GenAI 还可能改变中文二语学习者的学习动机、兴趣和态度 (Liu *et al.*, 2024, 2025; Liu & Zhao, 2026; Zhao, Hossain, *et al.*, 2026; Zhao, Wang, *et al.*, 2026)。因此,教育资源生成结果的有效性不能仅以生成效率或内容一致性判断,还应考察其是否能够为学习者提供可接受、可进入、可表达的任务材料。

综合上述文献,现有研究存在三方面缺口。文生图教育应用研究重视生成效率,但对生成结果能否达到专业命题要求缺乏检验;语言测试研究已证实图像设计影响口语产出与考生体验,但尚未与生成式资源开发实践相结合;提示词工程研究表明需求描述的组织方式影响生成质量,但教师参与方式对试题图像生成效果的影响尚无实证证据。

3 看图说话试题图像生成智能体构建方法

3.1 构建思路与理论依据

在应用生成式人工智能开发国际中文教育资源的过程中,如何将源知识有效转化为适合教学与测评使用的目标资源,是核心问题之一。张黎和刘敏 (2026) 指出,知识资源要素需要对源知识进行系统编码,即进一步加工与结构化表达;知识编码不仅是形式转换过程,更应关注编码质量:一方面,编码后的知识内容应准确反映源知识的性质与特点;另一方面,还应充分考虑学习者的认知特点及认知加工难度,以促进知识的高效解码、记忆、内化与迁移。

基于上述认识,本文认为,看图说话试题图像生成的关键,并不在于简单调用生成模型输出图像,而在于如何将命题需求转化为结构清晰、要素明确、可供模型理解和执行的试题需求描述。换言之,试题图像生成本质上是一个由命题意图到图像表达的知识编码过程:若关键要素表达不充分或要素组织结构不清晰,即便底层生成模型能力较强,生成结果也可能偏离命题目标。因此,本文首先归纳抽象看图说话试题的常见图像要素,形成要素提示词模板,并在此基础上设计两种教师参与程度不同的智能体构建方案。

3.2 看图说话试题图像生成要素模板

为提高试题需求描述的规范性与可操作性,本文首先基于所收集的看图说话试题样本,对该类试题中常见的图像构成要素进行了系统梳理。在综合分析试题内容、场景类型和表达任务特点的基础上,本文将看图说话试题图像中的核心要素概括为五类,即场景、人物、动作、表情和道具。场景刻画空间环境,人物限定参与主体及人数关系,动作体现人物核心行为,表情补充情绪状态,道具强化场景信息与叙事线索。

3.3 两类智能体构建

在上述要素模板基础上,本文首先设计并实现了面向教师深度参与场景的看图说话试题图像生成智能体(智能体A),适用于教师希望充分表达命题意图、对生成结果进行较强控制的情境。在该模式下,教师结合要素提示模板,围绕试题目标编写完整细致的试题需求描述;智能体接收后仅进行格式识别与规范化处理,不再改写教师输入,直接提交文本生成图像模型输出试题图像。其基本逻辑在于以教师高质量输入保障需求表达质量。

为进一步探索试题图像生成的自动化可能性,本文又设计并实现了面向教师低参与度场景的看图说话试题图像生成智能体(智能体B),适用于希望降低教师编写负担、提高生成效率的情境。

在这一模式下,教师仅需提供较为简要的需求输入,随后由智能体参照要素提示模板自动对需求进行补全、重组与调整,再将处理后的需求文本提交至文本生成图像模型生成图像。换言之,与前一种模式相比,本模式将部分原本由教师承担的需求组织工作转移给了智能体。

从系统流程上看,该模式同样包括如下三个步骤。

1. 教师依据要素模板提供较简要的试题需求输入,即五类核心要素各自由一些关键词组成,例如场景(户外、街道)、人物(一人、女性)、动作(车辆维修)、表情(焦急)、道具(汽车、手机);
2. 智能体将用户依据要素模板给出的所有关键词汇总并交由大模型自动完成需求补全、结构整理与文本优化,如上述例子将会被自动组织为:

写实摄影风格,光线明亮,背景简单干净,画面清晰,适合汉语考试教材使用。
户外城市街道边,有车辆或行人,真实街景。一位女性,检查车辆故障或维修,表情焦急/无奈。画面中包含:汽车(引擎盖打开)、手机。

3. 将调整后的需求文本提交至文本生成图像模型,输出试题图像。

该模式强调由智能体承担更多中间处理任务,因此其表现高度依赖于智能体在需求组织与语义保持方面的能力。

4 实证研究

4.1 研究对象与实验材料

看图说话是语言测试与教学中的典型任务形式,也是 HSK 口语考试的重要题型。以 HSK 官网最新版口语考试样卷为例,HSK 三级口语考试第二部分为看图说话,每题配 1 张图片并要求考生进行描述;HSK 四级口语考试第二部分则含 3 张关联图片,要求整体描述(详见官网 <https://www.chinesetest.cn/>)。鉴于相关研究尚处探索阶段,且为提高实验可操作性,本文实证范围限定于 HSK 三级口语考试的单张图片类看图说话任务。

基于上述目标,本文构建了一个小规模看图说话试题数据集,作为后续实验的基础。该数据集共包含 27 道试题,来源于 HSK 官网公布的口语测试样卷,以及当前互联网可检索的历届口语测试真题和模拟题。这些试题主要来源于官方样卷及相关标准化备考资源,具有较强的规范性和题型代表性,能够为本文的实证研究提供较为可靠的任务基础。本文收集的真题和模拟题仅用于学术研究。

在实验系统方面,智能体 B 首先使用 DeepSeek 应用五要素下的关键词构建完整描述,再选用阿里万相系列文生图大模型 Wan2.6-T2I 生成看图说话试题;智能体 A 则直接使用 Wan2.6-T2I 完成看图说话试题生成。DeepSeek 与 Wan2.6-T2I 均使用系统默认参数,未额外调整随机种子等设置;两个智能体在实际应用时,每次针对输入内容各自生成一张图片作为试题。

在评估对象方面,本文采用教师组与学生组两类受试者对生成结果进行评价。教师组共 23 人,均具有不少于 2 年的国际中文教育教学经验;学生组共 16 人。两组成员均通过南京大学相关教学网络招募。

4.2 实证研究设计

为系统考察大模型驱动的看图说话试题生成效果,本文围绕试题需求构建、图像生成、问卷评估与结果分析四个环节展开。

第一步,将已有看图说话试题逆向构建为文本化的试题需求描述,这是本研究得以开展的关键基础。一般而言,看图说话试题的自动生成应以命题需求为起点;然而在 HSK 口语考试场景中,可获得的高质量试题资源主要为公开样题、真题或模拟题,不包含命题者关于出题目标与图像设计意图的说明。为建立生成任务的输入基础,本文首先对已有试题开展逆向构建,将现有图片试题转换为可供智能体使用的文本化需求描述。

第二步,将上述需求描述分别输入两类智能体完成试题图片生成,以获得可对照的实验材料。

第三步,将全部 27 道试题的原始图片与对应生成结果统一整理为问卷材料,并依据教师与学生两类评价主体的不同关注点分别设计题项:教师问卷关注图像与命题需求的匹配程度及考试适配度,学生问卷关注情感接受度、话题联想度与表达意愿激发度。

第四步,汇总问卷数据,结合统计分析方法对实验结果进行分析。

4.3 问卷设计与实施

为全面评估生成试题图像的应用效果,本研究依据两类受试者在试题使用中的不同关注重点分角色设计问卷。教师作为专业评价主体,更关注图像与命题要求的匹配程度及考试适用性;学生作为直接接受者,更关注情感接受度、话题引导作用与表达意愿激发。问卷填写前已向受试者说明研究用途,问卷采用匿名方式收集,且数据仅用于学术研究。

教师问卷设置两个题项,分别用于考察图像与文字需求的符合程度以及其适合用于试卷的程度,对应“命题准确度”与“考试适配度”两个评价维度。学生问卷设置三个题项,分别用于考察图像引发不适的程度、唤起相关话题联想的能力以及激发开口描述意愿的强弱,对应“情感接受度”“话题联想度”与“表达意愿激发度”三个评价维度。通过上述设计,本研究从教师专业评价与学生学习体验两个层面,对生成试题的教育应用效果进行综合考察,具体如表 1 所示。

需要说明的是,命题准确度与考试适配度虽同属教师专业评价,但关注环节不同。前者考察图像对文字需求的内容还原程度;后者考察图像作为正式考试材料的使用适宜性。学生问卷方面,情感接受度对应二语习得中的情绪研究,即外语愉悦、焦虑等情绪会影响学习者的投入与学习表现,而学习内容与材料特征正是此类情绪的重要来源 (Zhao, Hossain, *et al.*, 2026;

Zhao, Liu, *et al.*, 2026; 黄芳炜、赵杨, 2026)。话题联想度对应任务型语言教学中的任务复杂度视角, 即任务材料的熟悉度与具体性会影响学习者话题图式的激活与口语内容的组织, 这与资源开发应适配学习者认知特点的要求一致 (张黎、刘敏, 2026)。表达意愿激发度对应交际意愿理论, 即交际意愿具有情境性, 会随任务情境与材料特征而变化 (Huang & Zhang, 2025)。

表 1. 问卷评价维度与题项设置

评价主体	评价维度	题项内容	评价目的
教师	命题准确度	请根据图像试题与文字需求的符合程度, 从高到低进行排序	评估图像对题目需求的还原程度与内容匹配性
教师	考试适配度	请根据适合用于试卷的程度, 从高到低进行排序	评估图像作为正式考试材料的适用性
学生	情感接受度	请根据是否感到不适, 程度从低到高进行排序	评估图像是否引发负面情绪反应
学生	话题联想度	请根据是否能够很快联想到相关话题, 从高到低进行排序	评估图像对表达任务的引导作用
学生	表达意愿激发度	请根据看图后开口描述的意愿强烈程度, 从高到低进行排序	评估图像对学生开口表达意愿的促进作用

4.4 评估指标与统计分析方法

基于上述问卷设计, 本研究将教师和学生对三幅图像的排序结果转化为可分析的数据, 并主要采用下述统计方法比较分析。

在评估指标方面, 本研究采用平均名次与首位率描述各图像来源在不同维度上的表现: 平均名次反映整体排序水平, 首位率反映被视为最优选项的频率, 二者分别从总体水平与偏好分布呈现相对表现。对于学生问卷中的“不适感”题项, 数据整理阶段对排序方向进行了统一处理, 以保证所有题项均满足排序越靠前评价越优。

在统计分析方面, 考虑到本研究数据属于配对排序数据, 本文采用 Friedman 检验比较三类图片在各评价维度上的总体差异; 总体差异显著后, 进一步采用 Wilcoxon 配对符号秩检验进行两两比较, 并结合 Bonferroni 校正控制多重比较误差。为避免将同一受试者的多题重复评价视为独立样本, 显著性检验以受试者为分析单位, 先计算每位受试者在各维度下对三类图片的平均名次, 再进行统计检验。对于未完整填写的排序题组, 本文仅作题组层面剔除, 原 39 份问卷中实际纳入有效完整排序题组 2485 个。

4.5 实验结果与分析

4.5.1 描述性结果

由于本研究采用排序式问卷, 研究首先将受试者对三幅图像的排序结果转换为名次值, 其中第 1 位至第 3 位分别记为 1、2、3, 并据此计算平均名次和首位率。平均名次越小, 说明评价越优。

表 2 呈现了不同评价维度下三类图片的统计结果。学生组中, 三类图片在三个维度上呈现一致的排序, 即原图的平均名次均最低、效果最优, 智能体 A 生成图次之, 智能体 B 生成图最高。

智能体 B 的平均名次均接近 2.5、首位率仅为 13% ~ 15%，表明学生普遍认为其在情感接受、话题引导与表达促进方面明显弱于另两类图片。

教师组中，三类图片在两个维度上同样表现出较清晰的差异。原图与智能体 A 生成图的平均名次接近（1.620 ~ 1.774），智能体 B 生成图则明显落后（平均名次约 2.6，首位率不足 10%），教师评价中原图与智能体 A 已形成接近的表现层级。

就两个教师维度的关系而言，以全部“教师 × 试题 × 图片”观测计算的 Spearman 相关系数为 0.838，且同一教师对同一试题两维度给出完全相同排序的比例达 78.0%，表明二者高度相关。但两维度并不冗余，即两者仍有 22.0% 的排序在两维度间不一致，且在题目 6、10、17 等试题上，两维度的“最优图片”出现明显差异，即原图在内容还原上最优，却因早期在画质与呈现形式上的限制，在考试适配度上被生成图反超。这说明两维度捕捉了教师评价中相互关联但可分离的信息。

表 2. 不同评价维度下三类图片的平均名次与首位率

评价主体	评价维度	原图	智能体 A	智能体 B	原图	智能体 A	智能体 B
		平均名次	平均名次	平均名次	首位率	首位率	首位率
学生	不适感	1.652	1.853	2.495	51.9%	33.4%	14.7%
学生	话题联想	1.661	1.822	2.517	51.7%	35.1%	13.3%
学生	开口意愿	1.643	1.844	2.513	51.3%	34.5%	14.2%
教师	符合需求	1.620	1.774	2.607	53.0%	37.2%	9.8%
教师	试卷适配	1.656	1.738	2.605	51.5%	39.1%	9.4%

4.5.2 统计分析

表 3 进一步报告了 Friedman 检验及 Wilcoxon 事后比较结果。Friedman 检验显示，学生组三个维度上三类图片的总体差异均达到显著水平。事后比较表明，三个维度上原图与智能体 A 生成图之间差异均不显著，而原图与智能体 B 生成图、智能体 A 与智能体 B 生成图之间的差异在 Bonferroni 校正后仍达到显著水平。结合描述性结果，这意味着图像生成质量不仅影响图片是否“像”题目要求，也会进一步影响学生是否愿意进入口语表达状态。

根据表 3 的内容，教师组两个维度的 Friedman 检验同样达到显著水平，且事后比较呈现与学生组相同的模式：原图与智能体 A 生成图之间差异不显著，涉及智能体 B 的两组比较在校正后仍显著。这说明智能体 A 生成图在命题需求还原与试卷适配方面已达到与原图相近的水平，具备较强的教育应用潜力，而智能体 B 生成图在正式试题使用场景下仍存在明显不足。

表 3. 不同评价维度下三类图片的统计分析结果

评价主体	评价维度	Friedman $\chi^2(2)$	p	p (原图 vs 智能体 A)	p (原图 vs 智能体 B)	p (智能体 A vs 智能体 B)
学生	不适感	16.125	<0.001	0.140	<0.001	<0.001
学生	话题联想	18.875	<0.001	0.215	0.001	<0.001
学生	开口意愿	19.484	<0.001	0.109	<0.001	<0.001
教师	符合需求	35.565	<0.001	0.162	<0.001	<0.001
教师	试卷适配	34.967	<0.001	0.570	<0.001	<0.001

5 讨论

5.1 本文主要发现

本文实证结果显示,三类图片在教师评价与学生评价中均呈现稳定的层次结构。智能体 A 生成图与原图之间未检测到显著差异,智能体 B 生成图在全部五个评价维度上均显著落后。这表明生成式人工智能支持看图说话试题图像生成具有现实可行性,但其应用效果高度依赖试题需求描述的组织方式与教师专业参与程度。

5.2 与已有研究的关系

本文结果与已有生成式人工智能教育资源研究形成呼应与推进。一方面,可行性判断与 Wu *et al.* (2024)、何晴和饶高琦 (2025) 的结论相一致;本文进一步表明,这种可行性并非无条件成立,较优方案接近原图水平,较弱方案则在教师评价与学生体验各维度上全面落后,说明“如何组织生成”是决定资源质量的关键变量之一。另一方面,智能体 B 的要素级语义保持困难,为欧志刚等 (2024) 所报告的图文不符等事实性偏差及童莉莉等 (2024) 的生成风险治理议题提供了图像生成场景下的具体实证,对齐失败更多表现为要素关系结构的隐性流失,而非显性的内容错误,因而更难被常规质检发现。

本文结果也为张黎和刘敏 (2026) 的知识编码原理提供了实证支撑。在本文中,试题需求描述正是命题意图的编码形式。教师深度参与的描述以显式句式固定要素及其关系结构,编码质量高,生成结果随之接近原图水平;自动补全的描述在重组中稀释了要素级约束,编码质量下降,生成结果出现系统性偏移。这说明在生成式人工智能条件下,知识编码原理不仅仍然适用,其重要性还被进一步放大,底层模型能力越强,编码环节越可能成为资源质量的瓶颈。

在学习者层面,本文将任务材料质量与情感接受、话题联想、表达意愿实证地连接起来,呼应了学习者情绪与参与研究的基本主张 (Huang *et al.*, 2024; Liu *et al.*, 2024, 2025; Huang & Zhang, 2025; Liu & Zhao, 2026; Zhao, Hossain, *et al.*, 2026; Zhao, Wang, *et al.*, 2026; 赵忭等, 2026)。已有研究多从教学设计、教师行为与技术接受角度解释学习者的表达动力,本文则表明,生成材料本身的命题质量同样会传导至学习者的情感与表达层面,补充了来自材料一侧的证据。

5.3 智能体 B 生成图像失败案例分析

为深入理解智能体 B 表现较弱的具体原因,本文按教师评价中智能体 B 与智能体 A 的平均名次差,筛选出差距最大的 4 道代表性试题 (见表 4、图 1),对图像逐题对照,并结合题目文本需求分析失败原因。从图 1 可见,智能体 B 的失败并非简单的画面质量或风格问题,而是三类系统性的要素偏移。

第一,核心动作要素丢失或置换。题目 19 的“两人站在墙边比划挂画”被生成为两人在餐桌旁擦拭收拾,命题核心动作被替换为无关家务动作。第二,关键道具缺失或变异。题目 23 的核心道具“地图”被替换为书本,并凭空出现行李箱,画面叙事从“旅行看地图找路”偏移为“操场旁的阅读”。第三,人物数量与关系结构偏移。题目 13 的“一站一坐”被生成为两人并排而坐;题目 24 的“几位女生自拍”被生成为教室中的多人合影。值得注意的是,智能体 B 对表情与大致场景类别保持较好,丢失的主要是人物数量、姿态、空间关系、持有物与具体动作等要素级约束。

表 4. 代表性试题中三类图片的教师评价平均名次

题目	原图	智能体 A	智能体 B	B 与 A 差值
题目 19（墙边比划挂画）	1.61	1.43	2.95	1.52
题目 23（看地图找路）	1.64	1.43	2.93	1.50
题目 13（一站一坐交流）	1.75	1.39	2.86	1.48
题目 24（毕业自拍）	1.57	1.50	2.93	1.43

图 1. 代表性试题的智能体生成图



上述差异可结合两类智能体的构建方式加以解释。智能体 A 模式下，教师编写的需求描述以完整句式显式固定五类要素及其关系结构，模型只需照单执行；智能体 B 模式下，关键词经自动补全与重组，要素间的关系约束随之稀释，文生图模型在组合性生成上的固有困难被放大，转而依据典型场景先验补全画面。这也解释了失败为何集中于含复合要素与关系性约束的试题，而在单人物、原型化场景中两者差距明显缩小。

可见，智能体 B 的弱势主要源于需求描述自动组织过程中要素级约束流失所致的系统性语义保持失败。由此也可更深入地回答研究问题 3，教师深度参与的价值，主要不在于语言润色，而在于以显式描述保住命题要素及其关系结构。

5.4 研究局限

其一，试题需求描述的逆向构建在效度上存在局限。受限于 HSK 官方公开命题资料的稀缺性，本研究的试题需求描述由现有试题图像逆向构建而来，而非源自命题者的原始设计意图，与真实命题场景存在一定差距。为缓解这一问题，逆向构建过程由 3 名研究者依据五要素模板独立完成并交叉核对，对存在分歧的表述经集体讨论后达成一致。

其二，问卷样本规模有限且存在一定同质性。本研究仅以 27 道试题为实验材料，教师组（23 人）与学生组（16 人）均通过单一高校渠道招募，教学背景与语言背景较为集中，一定程度上限制了结论的外部效度。同时，较小的样本规模可能降低统计检验力，因此本文对未达显著的比较结果持谨慎态度。

需要说明的是，本研究属于探索性研究，旨在初步检验通用大模型支持看图说话试题图像生成的可行性。尽管样本存在上述限制，但排序式评价任务能够较直接地反映受试者对不同图像材料的相对判断，实验结果仍能在较大程度上支撑本文结论。

5.5 未来研究方向

未来研究可在四个方面推进: 扩大样本规模与背景多样性, 在真实命题情境中检验结论稳健性, 并为逆向构建补充评分者间一致性指标; 针对要素级约束流失问题, 探索自动补全环节的要素模板约束校验与教师轻量确认等人机混合机制; 将评估框架扩展至多图题型、其他等级考试及其他语言测试场景, 检验五要素模板的可迁移性; 开展纵向教学实验, 追踪生成试题在实际口语教学与备考中的使用效果。

6 总结

本文以国际中文教育中的看图说话试题图像自动生成为对象, 考察了大模型驱动的生成式人工智能在教育资源专业化开发中的应用效果。结果表明, 该技术具有现实可行性, 但不同生成策略之间存在明显差异, 教师深度参与模式下的生成结果整体更优。研究说明, 国际中文教育资源的智能化生成不仅取决于模型能力, 也与试题需求描述的组织质量及教师专业参与程度密切相关。

参考文献

- Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., & Ramesh, A. (2023). *Improving image generation with better captions*. OpenAI. <https://cdn.openai.com/papers/dall-e-3.pdf>
- Chen, X., & Wu, D. (2024). Automatic generation of multimedia teaching materials based on generative AI: Taking Tang poetry as an example. *IEEE Transactions on Learning Technologies*, 17, 1327–1340. <https://doi.org/10.1109/TLT.2024.3378279>
- Galaczi, E. D., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, 15(3), 219–236. <https://doi.org/10.1080/15434303.2018.1453816>
- Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, 51(12), 2629–2633. <https://doi.org/10.1007/s10439-023-03272-4>
- He, Qing (何晴), & Rao, Gaoqi (饶高琦). (2025). 国际中文教育生成式图片库的构建与应用 [Construction and application of generative image library for international Chinese language education]. *国际中文教育 (中英文)* [*International Chinese Language Education*], 10(4), 59–73.
- Huang, F., Wang, Y., & Zhang, H. (2024). Modelling generative AI acceptance, perceived teachers' enthusiasm and self-efficacy to English as a foreign language learners' well-being in the digital era. *European Journal of Education*, 59(4), Article e12770. <https://doi.org/10.1111/ejed.12770>
- Huang, F., & Zhang, H. (2025). Enhancing willingness to communicate in the second language through technology: The effects of community of practice, enjoyment, and boredom in online classes. *Computer Assisted Language Learning*, 1–33. <https://doi.org/10.1080/09588221.2025.2497496>
- Huang, Fangwei (黄芳炜), & Zhao, Yang (赵杨). (2026). 国际中文在线学习愉悦、焦虑情绪与阅读成绩的关系研究——在线小组活动参与度的中介作用 [Exploring the effects of online learning enjoyment and anxiety on reading achievement in Chinese as a second language: The mediating role of group activity engagement]. *西安外国语大学学报* [*Journal of Xi'an International Studies University*], 34(1), 30–37. <https://doi.org/10.16362/j.cnki.cn61-1457/h.2026.01.011>

- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Ke, Qingchao (柯清超), Wang, Pengli (王朋利), & Zhang, Jieqi (张洁琪). (2018). 数字教育资源的供给模式、分类框架及发展对策 [Supply mode, classification framework and developmental strategies of digital educational resources]. *电化教育研究 [e-Education Research]*, 39(3), 68–74, 81. <https://doi.org/10.13811/j.cnki.eer.2018.03.011>
- Liang, Yu (梁宇), & Liu, Genqin (刘根芹). (2024). “中文 + 职业技能”教学资源建设的现状与展望 [The reality and prospects of “Chinese + Vocational” teaching resource development]. *沈阳师范大学学报 (社会科学版) [Journal of Shenyang Normal University (Social Science Edition)]*, 48(1), 24–31. <https://doi.org/10.19496/j.cnki.ssb.20231124.001>
- Liu, G. L., Lee, J. S., & Zhao, X. (2025). Critical digital literacies, agentic practices, and AI-mediated informal digital learning of English. *System*, 134, Article 103797. <https://doi.org/10.1016/j.system.2025.103797>
- Liu, G. L., & Zhao, X. (2026). A scoping review of AI-mediated informal language learning: Mapping out the terrain and identifying future directions. *ReCALL*, 38(1), 111–130. <https://doi.org/10.1017/S0958344025100359>
- Liu, G. L., Zhao, X., & Yang, B. (2024). The predictive effects of motivation, enjoyment, and self-efficacy on informal digital learning of LOTE: Evidence from French and German learners in China. *System*, 126, Article 103504. <https://doi.org/10.1016/j.system.2024.103504>
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1–35. <https://doi.org/10.1145/3560815>
- Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (Article 384). Association for Computing Machinery. <https://doi.org/10.1145/3491102.3501825>
- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), Article 410. <https://doi.org/10.3390/educsci13040410>
- Luoma, S. (2004). *Assessing speaking*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511733017>
- Lyu, Zhenhua (吕振华), & Ye, Jun (叶军). (2024). 大语言模型赋能“中文 + 职业技能”教学资源建设研究——以 ChatGPT 为例 [The research on the construction of “Chinese + Vocational Skills” teaching resources empowered by large language models: Taking ChatGPT as an example]. *天津师范大学学报 (社会科学版) [Journal of Tianjin Normal University (Social Sciences)]*, (6), 1–8. <https://doi.org/10.20247/j.issn1671-1106.2024.06.001>
- Ma, Jianfei (马箭飞), Liang, Yu (梁宇), Wu, Yinghui (吴应辉), & Ma, Jianan (马佳楠). (2021). 国际中文教育教学资源建设 70 年：成就与展望 [Seventy years of construction of international Chinese language teaching resources: Achievements and prospects]. *天津师范大学学报 (社会科学版) [Journal of Tianjin Normal University (Social Sciences)]*, (6), 15–22. <https://doi.org/10.20247/j.issn1671-1106.2021.06.002>
- Nakatsuhara, F. (2013). *The co-construction of conversation in group oral tests*. Peter Lang. <https://doi.org/10.3726/978-3-653-03584-1>
- Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. D. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly*, 14(1), 1–18. <https://doi.org/10.1080/15434303.2016.1263637>

- Oppenlaender, J. (2024). A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, 43(15), 3763–3776. <https://doi.org/10.1080/0144929X.2023.2286532>
- Ou, Zhigang (欧志刚), Liu, Yuping (刘玉屏), Qin, Ke (覃可), Wang, Yi (王艺), & Li, Xinyi (李欣蕊). (2024). 人工智能多模态教学资源的生成与评价——基于AIGC在国际中文教育的应用 [Generation and evaluation of artificial intelligence multimodal teaching resources: Based on the application of AIGC in international Chinese language education]. *现代教育技术 [Modern Educational Technology]*, 34(9), 37–47. <https://doi.org/10.3969/j.issn.1009-8097.2024.09.004>
- Peng, Jinwei (彭锦维). (2022). 国际中文教育教学资源库建设综述 (2000 ~ 2022 年) [Review of international Chinese language education teaching resource libraries in China (2000–2022)]. *汉语国际教育学报 [Journal of International Chinese Language Education]*, (12), 31–48.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical text-conditional image generation with CLIP latents*. arXiv. <https://doi.org/10.48550/arXiv.2204.06125>
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8821–8831). PMLR. <https://proceedings.mlr.press/v139/ramesh21a.html>
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695). IEEE. <https://doi.org/10.1109/CVPR52688.2022.01042>
- Tong, Lili (童莉莉), Zeng, Jia (曾佳), & Di, Ying (底颖). (2024). AIGC 视域下数字教育产品的四维风险矩阵与治理框架 [Four-dimensional risk matrix and governance framework of digital education products from the perspective of AIGC]. *现代远程教育研究 [Modern Distance Education Research]*, 36(2), 12–19. <https://doi.org/10.3969/j.issn.1009-5195.2024.02.002>
- Wang, Qiong (汪琼), & Liu, Ming (刘敏). (2023). 美国高等教育的数字化转型进程 [Digital transformation of higher education in the United States]. *世界教育信息 [Journal of World Education]*, 36(5), 18–28.
- Wei, Xiaodong (魏小东), & Xi, Pengfei (习鹏飞). (2024). 大语言模型驱动下的虚拟数字人第二语言教育应用研究 [A study on application of virtual digital human for second language education driven by large language model]. *电化教育研究 [e-Education Research]*, 45(5), 75–82. <https://doi.org/10.13811/j.cnki.eer.2024.05.010>
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT*. arXiv. <https://doi.org/10.48550/arXiv.2302.11382>
- Wu, K., Ding, J., Li, J., Yang, Y., Zhang, C., & Cao, J. (2024). LLM-empowered image generation in the Neko Painter App: A preliminary application for producing teaching materials. In *Proceedings of ICET 2024* (pp. 110–114). IEEE. <https://doi.org/10.1109/ICET62460.2024.10868922>
- Wu, Yinghui (吴应辉), Liang, Yu (梁宇), Guo, Jing (郭晶), & Ma, Jianan (马佳楠). (2021). 全球中文教学资源现状与展望 [The current situation and prospects of global Chinese language teaching resources]. *云南师范大学学报 (对外汉语教学与研究版) [Journal of Yunnan Normal University (Teaching and Research on Chinese as a Foreign Language)]*, 19(5), 1–6. <https://doi.org/10.16802/j.cnki.ynsddw.2021.05.002>
- Xi, X. (2005). Do visual chunks and planning impact performance on the graph description task in the SPEAK exam? *Language Testing*, 22(4), 463–508. <https://doi.org/10.1191/0265532205lt305oa>
- Xi, X. (2010). Aspects of performance on line graph description tasks: Influenced by graph familiarity and different task features. *Language Testing*, 27(1), 73–100. <https://doi.org/10.1177/0265532209346454>

- Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Article 437). Association for Computing Machinery. <https://doi.org/10.1145/3544548.3581388>
- Zeng, Jun (曾君), & Lu, Fangzhe (陆方喆). (2021). 国际汉语数字化教学资源的概念、分类与体系 [The concept, classification and system of international Chinese digital teaching resources]. 云南师范大学学报 (对外汉语教学与研究版) [*Journal of Yunnan Normal University (Teaching and Research on Chinese as a Foreign Language)*], 19(3), 28–37. <https://doi.org/10.16802/j.cnki.ynsddw.2021.03.005>
- Zhang, Li (张黎), & Liu, Min (刘敏). (2026). 新技术条件下国际中文教育资源开发的基本原理解析 [Fundamental principles for developing International Chinese Language Education resources under new technological conditions]. 世界汉语教学 [*Chinese Teaching in the World*], 40(2), 147–159. <https://doi.org/10.13724/j.cnki.ctiw.2026.02.006>
- Zhao, X., Hossain, M. K., & Sun, P. P. (2026). The facilitating role of enjoyment and anxiety in shaping (AI-mediated) informal language learning and confidence: An explanatory sequential mixed-methods investigation. *Journal of Computer Assisted Learning*, 42(1), Article e70181. <https://doi.org/10.1002/jcal.70181>
- Zhao, X., Liu, G. L., Eerdemutu, L., & Shirvan, M. E. (2026). Do shyness and emotions matter in shaping classroom engagement? A study of multilingual learners of Chinese in Thailand. *Journal of Multilingual and Multicultural Development*, 47(6), 3827–3845. <https://doi.org/10.1080/01434632.2025.2539922>
- Zhao, X., & Wang, D. (2025). The role of enjoyment and boredom in shaping English language achievement among ethnic minority learners. *Journal of Multilingual and Multicultural Development*, 46(3), 668–680. <https://doi.org/10.1080/01434632.2023.2194872>
- Zhao, X., Wang, D., & Liu, G. L. (2026). Innovating Chinese vocabulary learning through multimodal GenAI: The motivational, interest, and attitudinal shifts among CSL learners. *The Asia-Pacific Education Researcher*. Advance online publication. <https://doi.org/10.1007/s40299-026-01098-x>
- Zhao, Xian (赵欣), Tian, Yan (田艳), & Sun, Peijian (孙培健). (2024). 线上明德中文教学模式的“变”与“不变” [Changes and constants of the “Ming De” Chinese teaching model online]. 国际中文教育 (中英文) [*International Chinese Language Education*], 9(3), 55–63.
- Zhao, Xian (赵欣), Wang, Danping (王丹萍), & Ren, Wei (任伟). (2026). 从理论建构到实证研究: 积极心理学在国际中文教育领域的探索与展望 [The exploration and prospects of positive psychology in International Chinese Language Education: From theoretical framework to empirical research]. 全球中文发展研究 [*Global Chinese Development Studies*], (6), 125–149.
- Zheng, Xiaojun (郑小军), Ma, Dongli (麻东利), Yang, Shangying (杨上影), & Yao, Siyu (姚思雨). (2025). 生成式 AI 赋能微课开发: “文—图—视频”链式制作方法研究 [Generative AI-empowered micro-lecture development: A study of the “text–image–video” chained production method]. 基础教育研究 [*Basic Education Research*], (15), 23–28.
- Zheng, Yanqun (郑艳群). (2023). 在教育变革和技术变革中思考国际中文教育的前景 [Comments on the prospect of international Chinese language education in the context of educational and technological change]. 天津师范大学学报 (社会科学版) [*Journal of Tianjin Normal University (Social Sciences)*], (2), 15–23. <https://doi.org/10.20247/j.issn1671-1106.2023.02.003>

投稿: 2025 年 8 月 29 日, 接受: 2026 年 5 月 20 日, 出版: 2026 年 8 月 28 日

作者简介

季瑞骅，南京大学助理研究员。研究方向：教学资源生成、智能化软件工程。

石昱欣，南京大学在读硕士生。研究方向：教学资源生成、国际中文教育。

张天，南京大学教授。研究方向：智能化软件工程、模型驱动软件工程。

赵文书，南京大学教授。研究方向：美国文学、语言教育。

Generative AI-Supported Image Generation for Picture-Description Speaking Test Items

Ruihua Ji

Yuxin Shi

Tian Zhang

Wenshu Zhao

Nanjing University, China

Abstract

Generative artificial intelligence (GenAI) is becoming an important support for resource development in international Chinese language education. However, existing research has focused mainly on generation efficiency and cross-modal content integration, with little examination of whether the generated outputs truly meet professional item-writing requirements or how they affect learners' participation and motivation for oral expression. Against this background, this study investigates the automatic generation of images for picture-description speaking test items in international Chinese language education. A generation template comprising five elements, i.e., scene, person, action, expression, and prop, was constructed, and two generation agents involving different degrees of teacher participation were designed accordingly. Using 27 picture-description speaking items as benchmarks, the two agents generated corresponding images, and an evaluation framework was established from two perspectives: professional teacher evaluation and learner experience. The results show significant differences among the generation schemes in educational effectiveness: the better-performing scheme was generally comparable to the original images, whereas the weaker scheme underperformed in teacher evaluations as well as in students' affective acceptance, topic association, and willingness to speak. The findings indicate that GenAI-supported resource development for international Chinese language education is feasible in practice, but its effective application depends not only on model capability but also on how item requirement descriptions are organized and on the degree of teachers' professional involvement.

Keywords

Generative artificial intelligence, picture description, image generation, learner experience

Ruihua Ji (Ph.D.), Assistant Research Fellow at Nanjing University, China. Research interests: generation of instructional resources; intelligent software engineering.

Yuxin Shi, Master's student at Nanjing University, China. Research interests: generation of instructional resources; international Chinese language education.

Tian Zhang (Ph.D.), Professor at Nanjing University, China. Research interests: intelligent software engineering; model-driven software engineering.

Wenshu Zhao (Ph.D.), Professor at Nanjing University, China. Research interests: American literature; language education.