Journal of Chinese Language Education and Evaluation

***JCLEE*, Vol. 2, No. 1, 2026, pp.33-49.**

Print ISSN: 3078-283X; Online ISSN: 3104-5030

Journal homepage: https://www.cleejournal.com

DOI：https://doi.org/10.64058/jc1ee260103miclt

原生多模态大语言模型驱动下的中文多维教学创新

王大亮（Wang Daliang）

摘要：2025 年底以来，以 GPT-4o、Gemini 系列等为代表的原生多模态大语言模型（Multimodal Large Language Models,MLLMs）实现了从文本生成向图像、音频、视频实时生成的跨越式升级，对于国际中文教育来说，MLLMs 的即时多模态生成能力，为教师提供了突破“汉字难认、语境难设、文化难懂”三重困境的方法。本文围绕汉字、语法、文学与文化四大内容，通过真实课堂案例阐述教师如何运用 MLLMs 生成即时、个性化的教学材料，用多模态学习理论、认知负荷理论、双重编码理论与具身认知理论构建四层理论框架，对各案例的作用机制进行系统解读，并为后续量化研究提供理论框架与案例参照。

关键词：中文教学；二语习得（SLA）；多模态教学化；跨文化教学

作者简介：王大亮，副教授，美国高点大学世界语言文学文化系系主任，研究方向：国际中文教学、二语习得。电邮：dwang@highpoint.edu。

Title: Multidimensional Innovations in Chinese Language Teaching Driven by Native Multimodal Large Language Models

Abstract: Since the end of 2025, native Multimodal Large Language Models (MLLMs), represented by the GPT-4o and Gemini series， have achieved an upgrade from text generation

Received: 08 Apr 2026 / Revised: 16 Apr 2026 / Accepted: 22 Apr 2026 / Published online: 30 Apr 2026 / Print published: 30 May 2026.

to the real-time generation of images, audio, and video. For International Chinese Language Education, the instantaneous multimodal generation capabilities of MLLMs provide teachers with a methodology to break through the "triple dilemma" of difficult character recognition, challenging context creation, and complex cultural understanding. Focusing on four core content areas, Chinese characters, grammar, literature, and culture, this paper uses real-world classroom cases to illustrate how teachers utilize MLLMs to generate immediate, personalized instructional materials. It constructs a four-layer theoretical framework based on Multimodal Learning Theory, Cognitive Load Theory, Dual Coding Theory, and Embodied Cognition Theory to systematically interpret the underlying mechanisms of each case. Ultimately, this study provides a theoretical framework and case references for subsequent quantitative research.

Keywords: Chinese Language Teaching; Second Language Acquisition (SLA); Multimodal Instruction, Cross-culture education

Author Biography: Wang Daliang, Associate Professor and Chair of the Department of World Languages, Literatures, and Cultures at High Point University, USA. Research Interests: International Chinese Language Education, Second Language Acquisition. E-mail: dwang@highpoint.edu .

一、引言

（一）研究背景：从生成式 AI 到多模态“智”变

近年来，人工智能（AI）尤其是生成式人工智能（GenAI）的迅猛发展，正在深刻地影响第二语言教学的生态系统。据 Derakhshan 和 Ghiasvand（2025）在其最新的特刊导论中指出，AI 已不再仅仅是辅助工具，而是逐渐演变为语言教学环境中的核心驱动力，教学正由“传统多模态教学”向“AI 驱动的动态多模态环境”转变。

然而，这一转变并非一蹴而就的，而是经历了一个不断的技术升级过程。早期的大语言模型，比如 GPT-3 及其同时期的产品，主要以文本生成为核心能力，其多模态功能有限。从 2024 年起，GPT-5，Gemini3.0, Claude4.5，以及国内的 DeepSeekV3.2 等新一代人工智能跨越式升级了“原生多模态”的能力。这些系统能够在单一架构内实时理解并生成文本、图像、音频与视频等多种符号资源，而非依赖模块拼接[1]。这一质变，而非量变，标志着 AI 辅助教学进入了全新阶段——教学互动的即时性与信息丰富度实现了前所未有的飞跃。

[1] 网络文献。Google DeepMind, *Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context* (Technical Report, 2024), https://arxiv.org/abs/2403.05530.

对于国际中文教育而言，这一技术突破具有特殊意义。汉语作为表意文字体系，其习得难点长期集中于三个方面：汉字字形与语义之间缺乏透明关系从而导致“汉字难认”；真实语境的构建费时费力从而使得“语境难设”；以及文化内涵的隐性传递使得“文化难懂”。Pan、Karataş 和 Kohnke（2025）对 GenAI 语言课堂应用的系统综述表明，现有研究虽已证实 AI 在提升写作、阅读和口语技能方面的优势，但如何系统利用原生多模态生成能力来针对性地解决上述三大难点，相关实证研究仍然相对匮乏。本文旨在回应这一研究缺口。

（二）目的与意义：AI 赋能，教师生成即时沉浸式教学内容

长期以来，国际中文教育面临的核心挑战，不仅在于教学内容的难度，更在于教学材料的生成与调整主要由教材编写者决定，而非课堂教师本人。教师即便对本班学生的学习状态、混淆难点与兴趣偏好了如指掌，也往往因为缺乏即时生成定制化材料的工具，而不得不依赖与课堂实际需求存在落差的预制内容。Siddiqi（2024）指出，缺乏具身感知的学习体验往往限制了跨文化理解的深度。而这一局限，在很大程度上源于教学材料无法针对学习者的具体认知状态进行精准调适。

MLLMs 的出现，从根本上改变了这一格局。其核心价值不在于生成了更好看的图片，而在于将即时定制的能力交到了教师手中：只有教师本人清楚学生学过哪些内容、哪里容易混淆、对什么风格更有共鸣。MLLMs 使教师得以将这种独有的课堂判断，即时转化为高度匹配的教学材料，真正做到有的放矢。

本文通过真实课堂案例，与同行分享 MLLMs 在国际中文教学四大内容中的具体应用方式与教学逻辑，并依托认知科学理论对其作用机制进行系统解读，为这一快速发展领域的后续研究提供实践参照与理论框架。具体而言，本文将围绕汉字、语法、文学与文化四大内容展开，每个内容用一至两个课堂案例，阐述 MLLMs 如何将教学材料的生成主导权还给教师，来即时生成丰富有趣的教学材料，为转型期的国际中文教学提供参考。

二、理论基础：多模态教学与二语习得的深度融合

在探讨具体教学应用之前，我们必须从理论层面厘清其作为“即时多模态引擎”如何介入二语习得（SLA）的认知过程。本章将结合经典的“可理解输入”理论与最新的认知心理学研究，探讨生成式 AI 如何通过降低认知负担与增强具身感知，从根本上改变中文作为第二语言（Chinese as a Second Language CSL）的习得机制。

（一）可理解输入的多模态重构：从“i+1”到“视觉脚手架”

语言学家克拉申（Krashen, 1982）的“可理解输入”理论认为，通过理解比自己当前水平略高（i+1）的语言材料（听或读）来自然习得语言，而非死记硬背规则。然而，在传统的中文教学中，

由于汉字与其语音、语义之间缺乏关联（不能望文生义或望文生音），单纯的文本输入往往导致学习者的学习负担过重，使得“i+1”无法有效利用。

原生多模态 AI 的出现，为破解这一困境提供了技术可能。Pan、Karataş 和 Kohnke（2025）对 GenAI 语言课堂应用的系统综述表明，MLLMs 能够即时将抽象的语言符号转化为图像或动态视频，这种“视觉辅助”充当了理解的脚手架（multimodal scaffolding），使原本可能超出学习者理解范围的文本输入得以降维为可理解输入。Yu、Song 和 Lu（2025）在针对多模态汉字资源的实证研究中进一步验证，AI 生成的“图—文”双模态输入，能够显著降低学习者在解码汉字时的认知焦虑，并提高汉字习得效率。

多模态脚手架并非仅仅是“图片配文字”的简单叠加。Vygotsky（1978）的最近发展区（Zone of Proximal Development，ZPD）理论强调，有效的学习发生在学习者独立能力与借助支持所能达到的水平之间。MLLMs 的关键优势在于其脚手架的“即时性”与“自适应性”。教师可以根据课堂实时反馈，即刻调整 AI 生成内容的复杂程度与呈现形式，从而将每位学习者的输入精准定位于其 ZPD 区间。这在传统教学中几乎无法实现，因为预制教材无法动态响应个体差异。

（二）认知负荷的系统性管理：多媒体学习原则的 AI 实现

理解输入的有效性，不仅取决于内容难度，还取决于学习者的认知资源是否被合理分配。Sweller（1988）提出的认知负荷理论（Cognitive Load Theory）区分了三类认知负荷：内在负荷（intrinsic load，由学习材料复杂性决定）、外在负荷（extraneous load，由不当的讲授方式引起）和相关负荷（germane load，学习者在构建新学习模态的有益投入）。有效的教学设计应当降低外在负荷、管理内在负荷，从而最大化相关负荷。

在传统中文课堂中，外在负荷往往被严重高估：教师以口头解释辅以板书，学生需同时处理听觉信息流与手写符号流，两者之间的整合本身就消耗大量认知资源。Mayer（2009）在多媒体学习理论中提出了著名的“多媒体原则”：当语言信息与对应的视觉图像同步呈现时，学习效果显著优于单一模态。其核心机制在于，视听双通道的并行处理能够降低单一通道的认知负担，释放更多资源用于深层的意义建构。

MLLMs 在这一框架下的价值在于：它能够自动生成符合多媒体原则的教学材料，而无需教师具备专业的视频制作技能。多项研究（Toyama & Hori，2025；Tao & Zeng, 2025）表明，当技术通过多模态营造出沉浸式学习氛围时，语言学习不再是单纯的脑力劳动，而是调动视觉、听觉与情感的综合体验，学习者的内在动机与注意力持续度均显著提升。这一发现与认知负荷理论的预测高度吻合：当外在负荷被技术手段系统压缩后，学习者得以将认知资源重新投入于语言意义的深度加工，有效提升对目标语言内涵的理解与内化效率。

（三）双重编码与具身认知：从“符号解码”到“意义体验”

认知心理学家 Paivio（1971）的双重编码理论（Dual Coding Theory）主张，人类大脑拥有两个功能独立的信息处理系统：言语系统（verbal system）负责处理语言符号，非言语系统（nonverbal system）负责处理图像与空间信息。当同一概念同时激活两个系统时（即“图—文”共现），记忆的编码路径增倍，提取的成功率也随之大幅提升。这一理论对汉字教学具有直接的方法论意义：孤立呈现汉字（纯言语编码）远不如将汉字与其对应的具象图像或动态情境同步呈现（双重编码）有效。

然而，双重编码理论的局限在于，它主要关注静态的记忆编码机制，而对学习者的“身体感知”关注不足。具身认知理论（Embodied Cognition）对此进行了补充：该理论认为，认知并非仅仅发生在大脑中，而是深刻根植于身体与环境的动态互动之中（Varela, Thompson & Rosch, 1991）。语言的习得，尤其是文化负载词汇与语用规则的习得，需要学习者建立起语言符号与具体身体感知经验之间的神经连接，而这恰恰是传统课堂教学中最难提供的要素。

Siddiqi（2024）在关于沉浸式技术的研究中指出，缺乏具身感知的学习体验往往限制了跨文化理解的深度。Abdallah（2025）提出的 C.H.A.T.S. 模型（Conversational, Holistic, Authentic, Transformative, Situated）进一步强调，AI 驱动的语言学习应当是“情境化”（Situated）与“真实性”（Authentic）的：通过 MLLMs 生成的高保真视听情境（如模拟真实中餐馆的点餐互动，或重现中国传统节日的仪式场景），学习者不再是语言材料的旁观者，而是被置于一个虚拟但高度逼真的语境中，语言符号因而与可感知的情境经验发生直接绑定。这种“准具身”（quasi-embodied）的学习体验，正是 MLLMs 相较于早期文本型 AI 最本质的教学优势所在。

（四）理论整合：构建 MLLMs 介入 CSL 习得的分析框架

综合以上四个理论，本文提出一个整合性框架，用以分析 MLLMs 在国际中文教育中的作用机制（见图 1）：

第一层：输入层。依托 Krashen 的可理解输入理论与 Vygotsky 的 ZPD 理论，MLLMs 通过即时的多模态脚手架，将文本输入转化为可理解输入，并实现对个体学习者 ZPD 的动态匹配。

第二层：加工层。MLLMs 结合 Sweller 认知负荷理论和 Mayer 多媒体原则，通过视听同步呈现，减少外部认知负担，提升深度学习资源利用效率。

第三层：编码层。依托 Paivio 的双重编码理论，MLLMs 实现图—文—声的多模态共现，激活言语与非言语的双重记忆编码路径。

第四层：内化层。依托具身认知理论与 C.H.A.T.S.模型，MLLMs 通过高保真情境的生成，为语言符号提供“准具身”的感知锚点，推动学习者从符号解码迈向意义体验。

这一四层框架将在后续各章节中作为分析工具，系统解释 MLLMs 在汉字，语法，文学，和文化四大教学内容中的作用机制，而不仅仅是对课堂活动的描述性呈现。

图 1：四个理论的整合框架

(Figure 1: integration of the four theories)

三、汉字教学：以 AI 图像生成构建视觉认知体系

汉字教学历来是中文教学中的难点，也是学生学得最吃力的地方。对于以字母文字为母语的学习者来说，汉字的学习障碍来自三个相互叠加的认知困境：一，汉字字形与语音之间缺乏透明的对应规则，学习者无法如拼音文字那样见字知音；二，字形与语义之间虽有关联，但脱离历史语境则往往无迹可寻；其三，形近字数量庞大，细微的笔画差异导致意思不同，极易混淆。传统教学对这三重困境的回应，长期依赖于偏旁、文字释义和机械练习，有点学习者还是“知其然而不知其所以然”，记忆效率低下且遗忘率极高（Everson, 2011）。

原生多模态大语言模型的介入，为上述三重困境提供了一条以视觉认知为核心的解决办法。MLLMs 不仅能够即时生成针对特定汉字的部件拆解图示、字源演变叙事与形声字家族对比图，更能够根据教师对学习者水平、语言背景与教学目标的具体描述，定制生成兼顾语言准确性与视觉叙事性的教学材料。本章通过三个递进式案例，阐述 MLLMs 如何在汉字教学的不同层次上重构学习者的视觉认知体系。

（一）部件语义解析与形近字辨析：以“体”与“休”为例

“体”与“休”是初级中文课程中同期出现的高频汉字，字形相近，学习者极易混淆。传统教学通常以笔画数量或偏旁名称加以区分，比如“体是人字旁加本，休是人字旁加木”。这种描述虽然准确，却停留于文字层面，学生仍然容易忘记。

利用 MLLMs，教师可以生成形象的对比解析图（见图 2）。两栏的视觉风格完全统一，使学习者在同一画面内完成对两个汉字的构成和逻辑的整体认知。图示所呈现的内容读者一目了然，无需逐一描述。传统的汉字教学，教师只能依赖已有的文本内容，而无法自由创作，将自己的意图生成生动有趣的对比图片，从而刺激学生的视觉系统，学生更容易识记。

依据 Paivio（1971）的双重编码理论，语言符号与对应视觉图像的同步激活能够成倍提升记忆效率；而两个形近字各自绑定不同场景图像的设计，进一步确保了两条记忆路径之间的清晰区隔，避免了单纯依赖字形辨别时容易发生的记忆干扰。Yu、Song 和 Lu（2025）的实证研究同样表明，AI 生成的图文双模态输入能够显著降低学习者在解码汉字时的认知焦虑，提高习得效率。

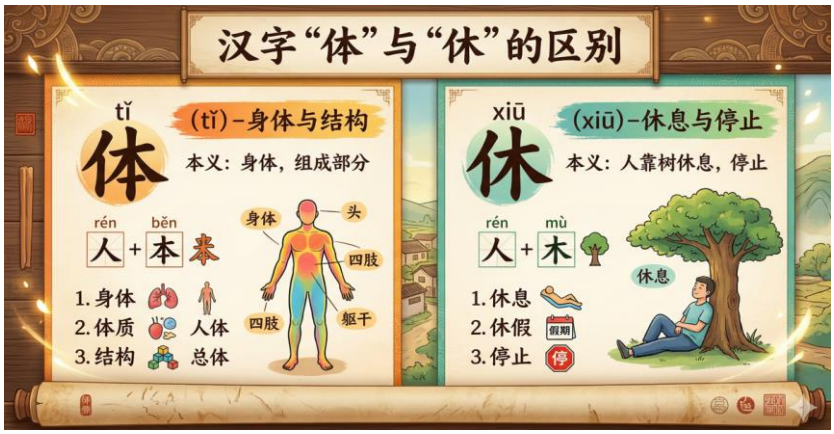

图 2：汉字“体”与“休”的区别

(Figure 2: the different between “体” body and “休” rest)

（二）部件深度解析：以“懂”字为例

“体”与“休”的教学价值在于字形字义的视觉对比，“懂”字教学示例所展示的，则是 MLLMs 在部件语义网络构建层面更为精密的教学能力（见图 3）。

“懂”字由三个部件构成：竖心旁（忄）、草字头（艹）与声旁“重”。教师利用 MLLMs 生成的部件深度解析图，以视觉化的方式将这三个部件的语义贡献逐层呈现：竖心旁（忄）代表心脏与内在情感；草字头（艹）原指草木生长，在此引申为外部观察与感知；声旁“重”除了表音，也有分量和深度的意思。三个部分结合，体现了“懂”字的意思：从内心深处增长理解力，最终有了深度和重量。图片进一步将“重”字解析，既帮学生复习以前学过的字，又能进一步记忆和书写“懂”这个复杂的汉字。

以前的教学，教师只能利用手头材料解释汉字，MLLMs 的介入从根本上改变了这一约束条件：教师可以针对课程中任意一个目标汉字，即时生成结构完整、语义准确、视觉清晰的部件解析图，将字理识字的方法论从少数精讲案例扩展为系统性的教学策略。这正是 Giannakos et al.

（2025）所描述的教师角色转型的具体体现：教师不再是知识的来源，而是通过精准的提示词，引导 MLLMs 成为汉字认知体系的即时构建工具。

从四层分析框架来看，本案例在编码层的作用尤为显著：部件的颜色编码（忄为红色、艹为绿色、重为蓝色）与对应的视觉图标（心脏、植物、天平）构成了三条独立的双重编码路径，分别对应“情感，感知，深度”三个语义维度，使学习者在记忆“懂”字时拥有三个相互强化的视觉锚点，而非单一的字形印象。

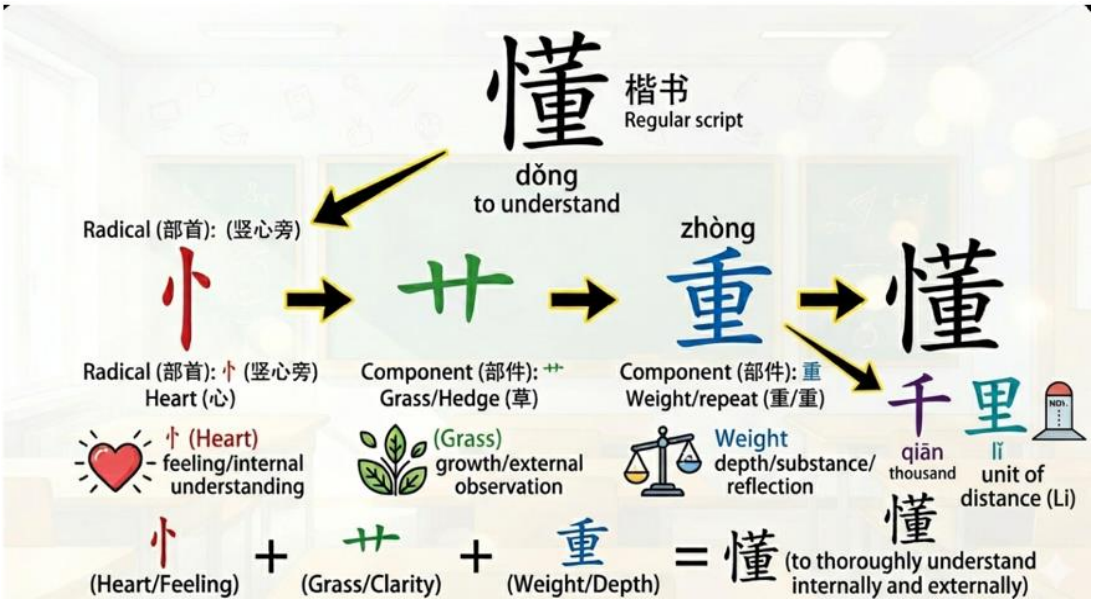

图 3：“懂”字的静态解析图

（Figure 3: Static Analysis Diagram of the Character “懂”）

（三）形声字家族的系统认知：以“青”字为例

汉字中约有八成以上属于形声字，即由表义的形旁与表音的声旁组合而成（Wang, 1998）。然而，传统教学往往将形声字作为孤立个体逐一讲授，未能充分利用声旁所构建的汉字家族网络。MLLMs 的图像生成能力，使教师得以将同一声旁的汉字家族作为认知系统而非孤立词条呈现，从根本上改变了学习者理解和记忆形声字的方式。

以汉字作为声旁的“青”为例，教师利用 MLLMs 生成了两种互补的教学图示。

第一种为系统性静态解析图（见图 4）。该图以“青”为核心声旁，横向并列呈现精、请、清、晴四个形声字，将每个字的形旁、声旁与语义以色彩与箭头清晰标注，使汉字的内部构造一目了然。这类图示的教学价值，并不在于图片所呈现的具体内容，而在于它实现了一件传统教学手段长期无法轻易做到的事：将汉字的内部结构、偏旁部首以直观、感官化的方式具象呈现，并将学生已学过的多个汉字在同一视觉框架内组合、复现。

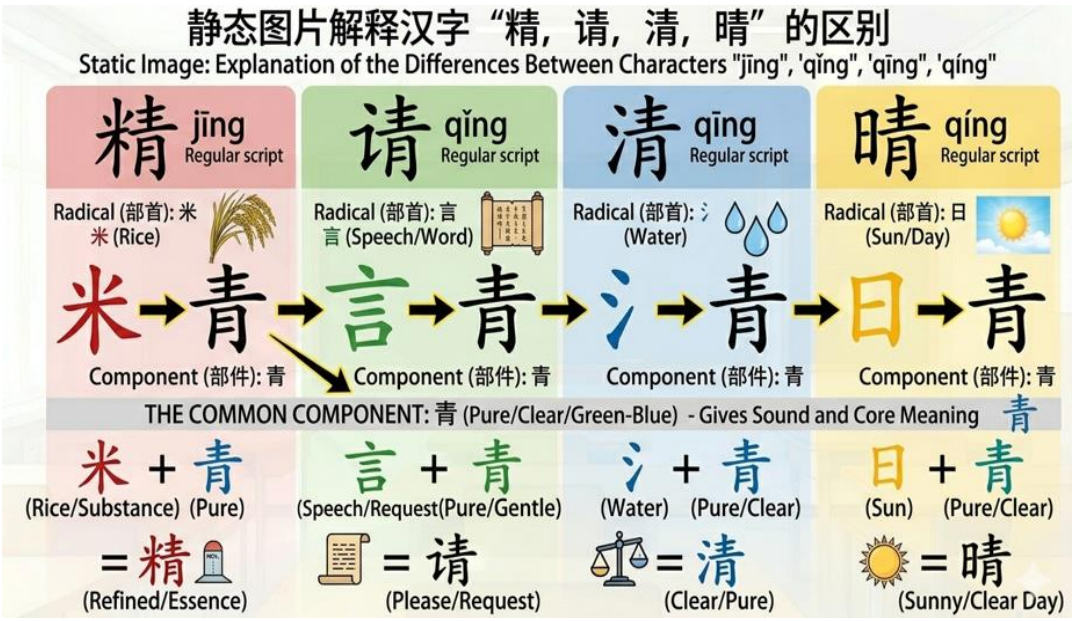

图 4：“青”字系统的解释图片

（Figure 4: the systematic explanation of characters related to “青”）

这种组合能力是 MLLMs 介入汉字教学，人机互动的很好的例子。教师本人清楚学生学过哪些汉字、哪些字容易混淆、哪些部件尚未建立认知，而 MLLMs 使教师得以将这种独有的课堂判断即时转化为个性化的教学，真正做到有的放矢。

第二种为叙事性漫画图（见图 5）。与静态解析图的结构化逻辑不同，漫画图为“青”字家族的每个成员构建了独立的故事场景，图像叙事的方式将汉字语义具象化、情境化。漫画的教学价值体现在两个层面：一，将枯燥的字义记忆转化为可感知的故事体验，学习者记住的不再是孤立的字形，而是与之绑定的画面与情节，记忆的提取路径因此大幅拓宽；二，这种呈现方式本身对学习者具有强烈的新鲜感与吸引力，能够有效激发学习动机。笔者在图 5 中要求使用宫崎骏的漫画风格生成作品，根据学生的喜好，选择最能引发共鸣的视觉风格，从而使教学材料在内容之外，在审美层面也与学习者建立情感连接。

更有意义的是，在教师示范之后，学生可以自行选择喜爱的风格，为自己觉得难记的汉字生成图片或漫画，并在课堂上互相分享、讨论。教师可以将学生生成的优质内容汇集整理，制作成班级专属的汉字学习手册。这一从教师到学生，再到集体共创的活动，将 MLLMs 的技术能力转化为真正以学习者为中心的教学：每一位学生既是学习者，也是教学过程的参与者。这种参与感与创造感，是任何教材所无法提供的。

从本文四层分析框架来看，“青”字家族案例在编码层实现了言语、视觉与情感三重编码的同步激活；在内化层，叙事性漫画为汉字语义提供了情境锚点，而学生自主生成内容的环节则进一步将语言符号与个人审美体验和创造行为深度绑定，使内化过程从被动接受转向主动建构。

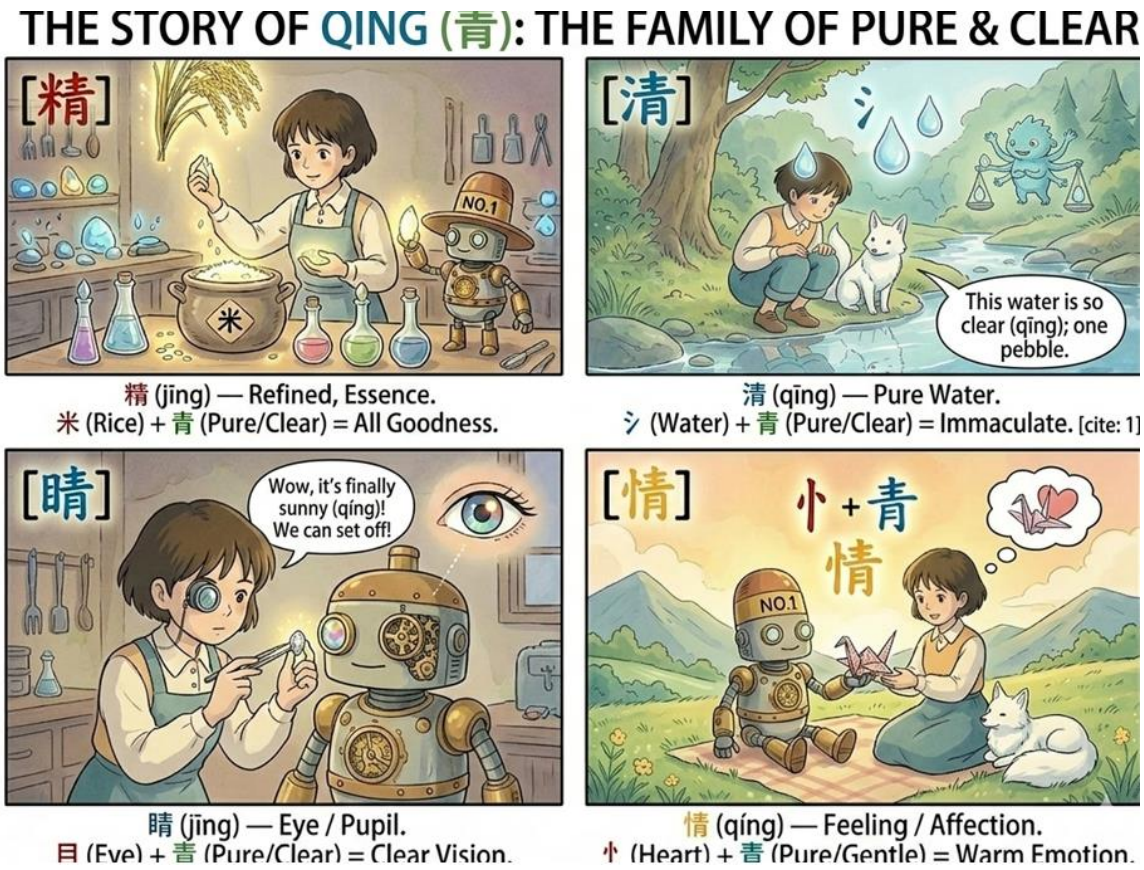

图 5：“青”字系统的漫画解释

（Figure 5:the cantonized explanation of characters related to 青）

（四）本章小结：MLLMs 重构汉字教学的三个层次

综合上述三个案例，MLLMs 对汉字教学的重构体现在三个递进层次上：在记忆策略层，AI 生成的对比图解析字源字形，加强记忆；在系统认知层，AI 可以将字形发音相似的形声字按照偏旁或发音组合在一起，将单个汉字串联记忆；在文化意义层，汉字的形意解析与叙事性漫画的结合使汉字从抽象符号转化为文化信息的载体，从而使得学习者对汉字的认知态度从记忆的负担，变为有意义的解码探索。

这三个层次的转变，也揭示了 AI 所带来的另一个转变，就是老师可以根据学生的学习情况，随时制定高质量，个性化的学习材料，而不限于教材或网上已有的材料。在 MLLMs 承担内容生成任务之后，教师的专业价值得以向更高层次迁移：从知识的搬运者，转型为学习体验的设计者。

四、语法与语用：动态可视化的教学构建

在传统的国际中文教学中，语法通常是一套学生需要记忆和运用的规则系统。教师的讲解依赖语言描述，教材的呈现依赖静态例句，而语用情境的建立则往往停留于课本对话的机械模仿。这种

教学模式的根本局限在于：语法规则本身是不可见的，而传统工具所提供的视觉辅助，比如二维图示、箭头、黑板板书等，与真实的空间感知和事件逻辑之间，始终存在难以弥合的认知鸿沟。

MLLMs 的图像与视频生成能力，为突破这一局限提供了新的可能。教师得以根据对本班学生学习状态与具体困难的判断，即时生成针对特定语法点的可视化材料，将“看不见”的语法规则转化为看得见、感受得到的视觉体验。本章通过两个案例，阐述这一能力如何介入语法教学的核心难点。

（一）抽象视点的具象化：以“简单趋向补语”为例

汉语中的简单趋向补语就直接 ctrl a 给它复制，然后呢，这些是我然后一篇一篇的，比如说像这样，我就把这上面的留着，如“上来，下去”和“进来，出去”严重依赖于说话人的空间立足点。“来”表示动作朝向说话人，“去”表示动作背离说话人，两者的区分本质上是一个第一人称空间感知的问题，而非单纯的语法规则记忆问题。传统教材通常以箭头示意方向，却无法提供真实的第一人称视角，导致学习者在实际运用中频繁混淆。

教师可以利用 MLLMs 的视频生成功能，通过控制镜头的为主来模拟说话人位置，生成四组对比强烈的短视频。如图 6 所示，这四个视频分别代表“上楼去”，“下楼去”，“上楼来”，和“下楼来”。老师可利用视频引导学生说出正确答案。

图 6：“来”和“去”做为趋向补语的视频截图

(Figure 6: Video Screenshots of 来 and 去 as Directional Complements)

这一教学设计为学生创造了一种“感同身受”的沉浸式体验。抽象的“来/去”语法规则就瞬间转化为了直观的视觉感知。教师展示视频后，无需给出答案，而是引导学生根据画面中人物与说话人的相对位置，自行判断并说出正确的趋向补语。这样，学习者从语法规则的被动接受者，转变为空间关系的主动判断者。这一教学设计的核心价值，与汉字章节中“形声字家族”案例的逻辑一脉相承：教师清楚地知道学生在哪个空间关系上最容易混淆，因此能够有针对性地生成专门对准这一混淆点的对比视频。在实际教学中，笔者可以明显地感受到学生理解这个语法的速度和准确度都提高了。有条件的时候，笔者也让学生在楼梯间互相给出指令，反复练习。这种因班制宜的即时定制，是预制教材无法做到的。抽象的“来/去”语法规则由此转化为直观的视觉感知，认知负担大幅降低，学生更有兴趣，学得也好（Zhang, 2025）。

（二）抽象句式的视觉具象化：以“把字句”为例

“把字句”是汉语教学中公认的经典难点之一。其语法核心在于：主语通过某一动作，对宾语施加影响并产生可见的结果，即“主语+把+宾语+动作+结果”的逻辑链条。对于母语中缺乏对应句式的英语学习者而言，这一逻辑链条的难点不在于结构记忆，而在于理解“动作必须导致结果”这一语义约束。

教师利用 MLLMs 生成分屏连续视频（见图 7），将把字句的因果逻辑转化为可观察的连贯事件序列：视频 1 中，猫跳上了桌子；视频 2 显示猫用爪子将杯子推至桌边；视频 3 呈现杯子落地打碎的结果状态。三段视频合在一起，将“把+宾语+动作+结果”的完整因果链以动态影像的方式直观呈现，让原本晦涩的语法逻辑活了起来。

视频相较于静态图示的优势在于，事件的动态展开过程本身就是语法意义的载体。学习者不是在看一个已经发生的结果，而是在实时目睹“动作导致结果”这一过程的发生，与把字句的语义逻辑形成同步感知。研究表明（Toyama & Hori, 2025），这种将语言特征视觉化、动态化的处理，能够帮助学习者在认知层面直接建立“动作导致结果”的逻辑图式，显著降低学习难度。

教师可以根据本班学生当前掌握的词汇与动补结构，即时调整视频中的动作对象与结果类型，确保视觉材料与课堂教学内容精准对应。

图 7：把字句连屏视频

(Figure 7: Split-screen Video of the 把 construction)

五、文学教学：从文字解码到意境体验

中国古典文学的教学，在国际中文课堂中长期面临一个结构性困境：古诗词的美学价值高度依赖“意境”的感知，而意境的传递本质上是一种跨感官的整体体验，是视觉、听觉、情感与文化记忆的同时激活。然而，传统课堂所能提供的手段极为有限：教师的口头讲解将意境还原为知识点的罗列，配套图片往往风格单一且与诗歌气韵相去甚远。对于中文水平有限的学习者而言，仅凭文字符号的解码，几乎不可能真正体会到古诗词所传递的情感与画面。

MLLMs 的视频与音频生成能力，为这一困境提供了新的解决办法。

以《江雪》为例（见图 8）。柳宗元这首五言绝句以极简的二十个字，构建了一个绝对孤寂与苍茫的意境世界：“千山鸟飞绝，万径人踪灭。孤舟蓑笠翁，独钓寒江雪。”对于初次接触古诗的学习者，“千山鸟飞绝”这样的文字所指向的那种绝对的空旷与寒冷，单靠语言解释几乎无法传递。

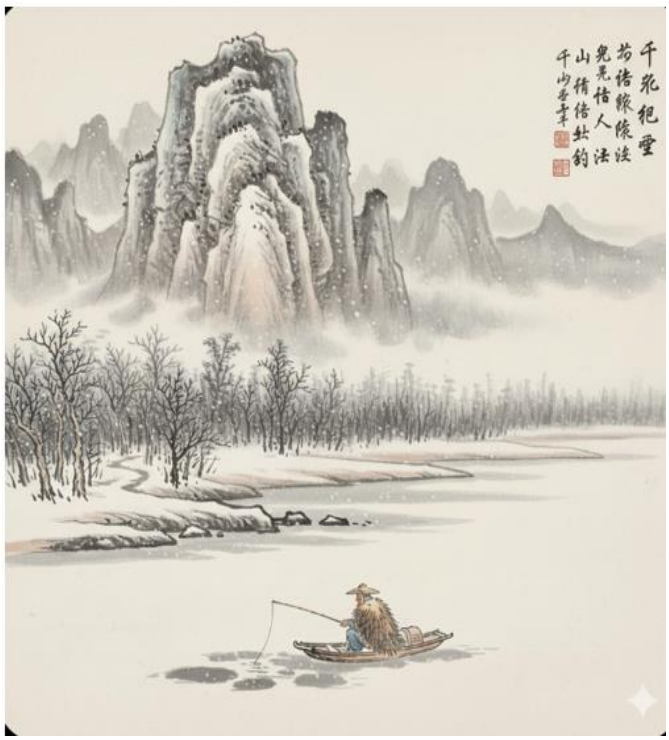

图 8：《江雪》示意图

(Figure 8: Schematic Diagram of “River Snow”)

教师利用 MLLMs 的视频与音频生成能力，将这首诗转化为视听通感的艺术体验：AI 生成动态水墨画风格的视频，画面随诗句逐句展开，配以风声与古琴声，将诗句的冷与静直观呈现。学生由此得以看诗、听诗、读诗，通过多感官的同步接收，理解诗句所构建的意境世界，而非仅仅记住字面意思。

教师可以根据所教诗词的具体气韵与风格，即时生成与之高度匹配的视听材料。古典诗词的意境各异：《江雪》的苍茫孤寂、《静夜思》的月下乡愁、《春晓》的清新明朗，每一首诗所需要的视觉风格、色调、音效都截然不同。传统课堂中，教师若要为每首诗配备与其气韵相符的视听材料，需要花费大量时间搜寻、筛选、剪辑，且结果往往差强人意。MLLMs 使教师得以用一句提示词，精准描述所需的画面风格、音乐氛围与镜头节奏，即时获得为这首诗、这个班级、这一教学目标量身定制的视听内容。

与汉字和语法章节的策略相呼应，文学教学同样可以将主动权交还给学生。教师展示《江雪》的视频案例后，鼓励学生自行选择喜爱的诗句，用 MLLMs 生成自己心目中这首诗的意境画面，并

在课堂上向同学解释自己的创作选择。比如，为什么选择这种色调、这种音乐、这种构图。这一过程将文学鉴赏从被动的欣赏推向主动的诠释，学习者对诗歌意境的理解在创作与表达的过程中得到深化。

从本文四层分析框架来看，文学案例在输入层通过视听材料将抽象的文字意象转化为可感知的多模态输入，大幅降低了古典文学对语言水平的门槛要求；在加工层，视觉画面与听觉氛围的同步呈现将意境的认知负担从纯语言理解转移至多感官感知；在编码层，动态水墨画与古琴声为每一句诗建立了独立的视听记忆锚点；在内化层，学生自主生成意境画面的创作过程，使诗歌的情感体验与个人审美判断深度融合，将文学感知从课堂活动转化为真正属于学习者自己的审美经验。

六、文化教学：从知识罗列到情感体验

在国际中文教学的四大内容中，文化教学长期面临一个独特的困境：文化知识的传递相对容易，但文化理解的深度内化却极为困难。比如，学生可以背诵“筷子不能垂直插在饭碗里”这条规则，却往往将其视为一条毫无逻辑的“怪规矩”，因为他们缺乏对这一禁忌背后完整文化语境的整体感知。这种“知其然而不知其所以然”的文化学习，导致学生对中国文化的认知停留于碎片化的知识点堆积，难以建立真正的跨文化理解。

传统文化教学的局限，源于呈现手段的单一性：教师的口头讲解将文化还原为信息，先成配套图片无法呈现文化行为的完整语境，而课本对话中的文化内容又往往经过高度简化，失去了真实文化情境的复杂性与情感张力。Siddiqi（2024）指出，真正有效的文化学习需要“整体语境感知”，就是说，学习者必须在一个完整的、可感知的情境中，同时接收文化行为的视觉信号、情感氛围与价值，而非孤立地接受单一文化知识的灌输。

MLLMs 的多模态生成能力，是教师可以为学生构建这种文化感知环境。

以“筷子禁忌”与“红白文化”的跨文化教学为例（见图 9）。教师利用 MLLMs 构建两个在视觉上截然对立的平行场景：一个是喜庆的节日家宴，红色主调、热闹氛围、筷子横置于碗边；另一个是肃穆的祭祀场景，素色主调、寂静氛围、筷子垂直插于供碗之中。两个场景通过颜色、服装、场合与人物神态的鲜明对比，使“筷子的摆放方式”成为区分“乐与哀”两种截然不同文化语境的视觉符号。

这种教学方法将一条孤立的文化禁忌，放在完整文化情境之中。学生看到两个场景的瞬间，感受到的不是一条需要记忆的规则，而是中国文化中“红白喜事”截然二分的整体色彩体系与行为规范，情感的构建自然过渡到记忆的持久。这正是 Siddiqi（2024）所描述的“沉浸式文化与情感唤起”机制的具体体现。AI 赋能教师的这种定制能力可以延伸至课程中的任意文化教学点：面子文化、长幼礼仪、送礼禁忌、餐桌座次等。每一个文化难点，都可以在教师的提示词引导下，生成与之对应的情境场景，实现从记忆知识点到感知文化情境的根本转变。与前面的教学案例一致，学生也可以

自己制作真实的“文化情境”，来参与和分享学习。笔者在期末文化展示中让学生用 AI 制作还原文化视频和图片，作为他们展示内容的辅助。学生都表现出很高的兴趣，积极参与制作个性十足的材料。这跟以前学生只是简单的从网上截取图片，将内容拼凑起来完全不同。同学之间的分享与讨论，则进一步在课堂内部形成了真实的跨文化交流语境。

从本文四层分析框架来看，文化案例在输入层通过完整情境场景的视觉呈现，将隐性的文化价值观显性化，使文化逻辑成为可感知的输入；在加工层，对比场景的并置设计将文化差异的认知负担从语言描述转移至视觉感知，系统降低了外在认知负荷；在编码层，色彩、服装、场合等多维视觉信号为文化禁忌建立了多条并行的非言语编码路径，使文化记忆的提取线索大幅增加；在内化层，情境的整体感知触发了情感唤起机制，使文化理解从知识层面深入至情感认同层面，而学生自主生成与解释文化场景的创作过程，则将这一内化推向最深层的主动建构。

图 9：红白喜事的文化对比图

(Figure 9: Cultural Contrast Diagram of Weddings and Funerals)

七、结论与局限

（一）结论

回顾全文，本文围绕汉字、语法、文学与文化四大内容，通过来自真实课堂的教学案例，系统探讨了原生多模态大语言模型（MLLMs）如何介入国际中文教学的核心难点，并依托多模态学习理论、认知负荷理论、双重编码理论与具身认知理论构建的四层分析框架，阐释其潜在的认知作用机制。

本文的核心论点在于：MLLMs 的即时多模态生成能力，能够将语言教学中高度抽象的符号材料，转化成直观，具象，具有情感体验的感性材料，从而在课堂中创造真正意义上的沉浸式学习体验。在这一过程中，教师对本班学生学习状态与具体需求的深度了解，使 MLLMs 的生成能力得以精准地制作有效的学习材料。正如研究所指出的，在 AI 承担内容生成任务之后，教师的角色从知识的搬

运者转型为学习体验的设计者（Giannakos et al., 2025）；而学生在教师引导下逐步参与内容的创造与分享，则使沉浸式体验从课堂活动延伸为真正属于学习者自己的认知建构过程。

（二）局限和后续研究

本篇文章具有以下局限性，需要特别说明：

第一，文中所呈现的课堂案例，均来源于真实教学实践，并在使用过程中得到了学生的积极认可与正面反馈。然而，这些案例目前尚未经过系统的量化研究验证，有待后续研究加以完善。因此，本文的案例分析所能支持的结论目前只是基于课堂经验。

第二，本文选择以理论框架而非文献综述作为分析基础，主要的原因主要是 MLLMs 原生多模态能力的重大跃升发生于 2025 年底，相关教学应用的实证研究尚处于起步阶段，针对中文教学情境的系统性文献积累十分有限。在这一背景下，以成熟的认知科学与二语习得理论为分析框架，是在实证文献相对匮乏的条件下为课堂实践提供理论支撑的合理选择。

正因如此，本文的定位始终是抛砖引玉：将笔者在教学中摸索、使用并得到课堂检验的方法与案例与同行分享，期待引发更广泛的实践探索与学术讨论。这一领域的真正成熟，有赖于更多教师将自身的课堂实践转化为可对话的研究成果，也有赖于后续研究在更大规模、更严格的实验设计下对本文提出的理论假设进行系统检验。

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

ORCID

Daliang Wang [ID] https://orcid.org/0009-0005-4238-2582

References

Abdallah, M. M. S. (2025, June). *The C.H.A.T.S. Model: A Framework for AI-Driven Language Learning in the Digital Age* (ED673561). ERIC. https://eric.ed.gov/?id=ED673561

Derakhshan, A., & Ghiasvand, F. (2025). “Multimodality in Language Learning and Teaching in the Age of AI and GenAI: Current Innovations and Emerging Trends.” *Computer Assisted Language Learning* (01):1–15. DOI: https://doi.org/10.1080/09588221.2024.2435721

Everson, M. E. (2011). “Best Practices in Teaching Chinese Character Recognition and Production.” In Cao, Y. (Ed.), Teaching and Learning Chinese as a Foreign Language, Hong Kong University Press: 51–68.

Gemini Team. (2023). *Gemini: A Family of Highly Capable Multimodal Models*. arXiv. DOI: https://doi.org/10.48550/arXiv.2312.11805

Krashen, S. D. (1982). *Principles and Practice in Second Language Acquisition*. Pergamon Press.

Mayer, R. E. (2009). *Multimedia Learning* (2nd ed.). Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511811678

Pan, X., Karataş, T., & Kohnke, L. (2025). "Generative AI (GenAI) in the Language Classroom: A Systematic Review." *Computer Assisted Language Learning* (38):335–359. DOI: https://doi.org/10.1080/10494820.2025.2498537

Paivio, A. (1971). *Imagery and Verbal Processes*. Holt, Rinehart and Winston.

Siddiqi, M. (2024). "Enhancing International Education with Generative AI: From Recruitment to Academic and Cultural Support." *American Journal of STEM Education* (01):48–60. DOI: https://doi.org/10.32674/5k0hbd11.

Sweller, J. (1988). "Cognitive Load During Problem Solving: Effects on Learning." *Cognitive Science* (02):257–285. DOI: https://doi.org/10.1207/s15516709cog1202_4

Rokkones, R. K., & Giannakos, M. (2025). "Toward Hybrid Teaching Intelligence: Investigating the Potential of Teacher–AI Collaboration Using Large Language Models." *Behaviour & Information Technology*:1–19. DOI: https://doi.org/10.1080/0144929X.2025.2564368

Tao, Y., & Zeng, L. (2025). "The Image-Text Relationship and Its Rationality Analysis of International Chinese Education Textbooks Under Multimodal Theory: Taking Experiencing Chinese Basic Course (Upper and Lower) as an Example." *Journal of Chinese Language Education and Evaluation* (01):53–66. DOI: https://doi.org/10.64058/JCLEE.25.1.04

Toyama, M., & Hori, T. (2025). "Technology-Enhanced Multimodal Approaches in Classroom L2 Pronunciation Training." *Frontiers in Education*. DOI: https://doi.org/10.3389/feduc.2025.1552470

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press. DOI: https://doi.org/10.7551/mitpress/6730.001.0001

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. DOI: https://doi.org/10.2307/j.ctvjf9vz4

Wang, J. (1998). "Reading Chinese Characters: Orthography, Phonology, and Semantics." In J. Wang, A. W. Inhoff, & H. C. Chen (Eds.), *Reading Chinese script: A cognitive analysis*, Erlbaum:1–34.DOI: https://doi.org/10.4324/9780203774618

Yu, J., Song, J., & Lu, Y. (2025). "Harnessing Generative Artificial Intelligence to Construct Multimodal Resources for Chinese Character Learning." *Systems* (08):692. DOI: https://doi.org/10.3390/systems13080692

Zhang, T. (2025). "Exploring the Use of Audiovisual Media in Teaching Chinese as a Foreign Language: A Comparative Perspective Based on DaF/DaZ Practices in German-Speaking Contexts." *Journal of Chinese Language Education and Evaluation* (01):39–52. DOI: https://doi.org/10.64058/JCLEE.25.1.03