武漢大学

本科毕业论文（设计）

基于大模型的跨模态古诗创作

姓　　名：　张志东

学　　号：　2021302111480

专　　业：　软件工程

学　　院：　计算机学院

指 导 教 师：　朱卫平

二〇二五年四月

原创性声明

本人郑重声明：所呈交的论文（设计），是本人在指导教师的指导下，严格按照学校和学院有关规定完成的。除文中已经标明引用的内容外，本论文（设计）不包含任何其他个人或集体已发表及撰写的研究成果。对本论文（设计）做出贡献的个人和集体，均已在文中以明确方式标明。本人承诺在论文（设计）工作过程中没有伪造数据等行为。若在本论文（设计）中有侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

作者签名：　　指导教师签名：

日　期：　　年　　月　　日

版权使用授权书

本人完全了解武汉大学有权保留并向有关部门或机构送交本论文（设计）的复印件和电子版，允许本论文（设计）被查阅和借阅。本人授权武汉大学将本论文的全部或部分内容编入有关数据进行检索和传播，可以采用影印、缩印或扫描等复制手段保存和汇编本论文（设计）。

作者签名：　　指导教师签名：

日　期：　　年　　月　　日

摘　　要

TODO：中文摘要

关键词：古诗生成；大模型；跨模态

ABSTRACT

TODO：英文摘要

Keywords: Poetry Generation; Large Models; Cross-modal

目　　录

1 绪论

本章主要介绍研究背景与意义、中文古诗生成相关的研究现状，阐述本论文的研究思路与主要贡献，并介绍论文的组织架构。

1.1 研究背景和意义

文本生成任务（Text Generation）是自然语言处理（Natural Language Processing, NLP）的一个重要研究方向，需要在给定输入或上下文的条件下，输出符合要求的文本，涵盖机器翻译、文本摘要、对话生成、作品创作等多个应用方向。而在文本生成领域中，中国古代诗歌的生成更是一个困难的任务。

古诗是中华传统文化的瑰宝，作为最早形成的中国古代文学作品体裁之一，其措辞简洁、内涵丰富且韵律整齐。中国古诗的典故意象运用极具文化特色、用词凝练优雅，这些独特的艺术特色都为古诗的机器创作带来巨大的挑战。如何让机器模型深入理解并创作出具有文化内涵和艺术价值的古诗，继而为中华传统文化的再创造赋能，是一个富有挑战的有趣问题。

早期的古诗生成主要依赖于其他子领域的研究思路。例如，[1] 利用现有古诗作为模板，根据既定规则替换字词来生成新的古诗，这样生成的古诗虽在形式上合格，但表达力欠佳；[2] 将其看作是一个摘要生成任务，只是输入是作者的表达意图，且需要考虑中文古诗独特的韵律形式约束；[3] 则将其看作一个机器翻译任务，将古诗的上下句子分别看作是翻译的源语言和目标语言，利用统计机器翻译（Statistical Machine Translation, SMT）的方法来生成古诗。

随着深度学习技术的发展，近年的中文古诗生成大多将古诗生成视作是“从序列到序列”的预测任务（Sequence-to-Sequence），并由此出发训练循环神经网络（Recurrent Neuro Networks, RNNs），如编码器-解码器（Encoder-Decoder）模型 [4]，并以此为基础设置额外机制来增强语义表现 [5] 或韵律格式约束 [6, 7]。然而，这些方法对古诗内涵的掌握往往停留在上下文语义或是韵律对仗，无法进一步触达诸如典故、意象和全诗连贯性等更复杂的方面。所幸，大模型（Large Model, LM）展现出强大的语义理解与创作能力，而在中文领域也出现了诸如 ERNIE[8]、DeepSeek-R1[9] 这样的大语言模型（Large Language Model, LLM）和 DeepSeek-VL[10] 等跨模态大模型，为这一领域注入全新的活力。[11]

除了文本输入外，中国古诗往往蕴含着丰富的场景意象，其对应的视觉信息难以通过用户输入来精确描述。目前也有方案直接使用图像作为输入，如在循环神经网络外增加卷积神经网络（Convolutional Neuro Networks，CNNs）以处理图像信息，捕捉图像关键主体并把握整体氛围，最终生成古诗。[12, 13] 但相应地，这些方案放弃了文本输入能够具体描述要求的优势，输入图像所含信息的繁杂也导致生成古诗主题、内涵乃至风格的波动。现有的方案大都局限于或文本或图像的单一模态输入，要么局限于上下文语义或韵律的形式规律而无法触达更高的艺术层次，要么受制于图像信息的多变而无法实现更精细的输出控制。这种单一模态的输入限制了用户描述需求的可能，也使生成的古诗缺乏层次，这暗示着文本“跨模态”输入的研究方向，也是本选题希望探讨的内容。

另一个重要的问题是生成结果的“可解释性”（Interpretability），指系统以人类可理解的术语解释或呈现模型行为的能力。[14] 深度学习技术在带来更高表现的同时，其“黑盒”的特性也使得可解释性问题愈发突出。在古诗生成任务中，可解释性可体现为三个方面：

（1） 过程可解释性，即系统是如何从输入的文本和图像中提取信息，并进一步并生成古诗的，其中包含哪些中间步骤；
（2） 结果可解释性，即系统为何生成这样的古诗结果，其中遵守怎样的韵律形式、又运用了哪些典故意象；
（3） 反馈可解释性，即系统如何分析生成古诗的质量，如何让用户理解古诗“好坏在哪里”。

现有的古诗生成系统大多缺乏对生成过程的可解释性，用户难以理解系统是如何从输入信息中提取出关键信息并生成古诗的；而在结果可解释性方面，现有的系统也往往只提供了生成古诗的文本，而没有进一步解释其韵律、意象等方面的内涵；在反馈可解释性方面，现有的系统也缺乏对古诗质量评价维度的详细说明，用户需要具备较高的文学素养以理解和甄别对古诗质量的评价。因此，提升古诗生成系统的可解释性，将有助于增强用户对系统的信任感和使用体验，使得用户能够更好地理解和利用系统输出的古诗和相应的质量判断。

本文旨在探讨文图跨模态的中文古诗生成，开发了一个基于大模型的古诗创作系统。在给定用户两种模态输入的条件下，其能够通过图像的分析描述来

充分提取图像信息，结合用户输入的文本信息，生成符合古诗韵律和意象的古诗。此外，系统还将提供对古诗的分析文本、量化评分以及改进意见，涵盖韵律对仗、典故意象、主题思想、语言用词等多个赏析方面，并支持多轮迭代优化。

1.2 研究现状

1.2.1 中文古诗生成

近年来，中文古诗生成领域引起了广泛的研究兴趣。2016 年，[15] 使用修改的注意力结构的编码器-解码器模型来解决古诗生成过程中主题漂移的问题，但限制关键词的数量和顺序，降低了系统的灵活性。这一问题在 2018 年由 [13] 通过记忆网络（增加记忆组块的 RNNs）基于图像生成古诗解决。2020 年，[16] 将注意力机制引入了 Seq2Seq 模型，实现了基于关键字的自定义古诗生成。

古诗的生成可能会出现多方面的质量波动，包括主题、语言风格、字数格式、韵律对仗等等，需要在研究中纳入考量。2020 年，[6] 设计了一个基于 Transformer 的自回归模型，改进注意力机制并进一步收紧了包括中文古诗、歌词、英文十四行诗等特殊文体生成的格式要求。2021 年，[17] 从图像中提取物体关键词、情感和风格，以生成古诗。[18] 将数十万首古诗按照风格、情感、格式与主要关键词分类，并利用掩蔽自注意力机制来建立标签到诗句的关联，以此来生成情感与风格均可控的古诗。2022 年，[19] 将 GAN 中的判别器与生成器结构加入到 CVAE 中，实现对风格和情感的控制。2023 年，[20] 构建了一个古诗图像数据集，精确标注了其中的诗歌元素，并利用 GRU 网络来增强生成古诗的上下文关联。2024 年，[7] 首次使用扩散模型来生成古诗，以实现对语义与韵律的同时控制。

后来逐渐出现使用 LLM 生成古诗的方案。2024 年，[21] 通过强化学习算法 PPO 对 GLM 模型进行训练，提升其在古诗生成方面的表现。[11] 修正了 LLM 以 token 计数会导致输出格式错误的问题，改以字符计数，达到了极高的格式精度。[22] 则提出了一种图片输入的三段式绝句生成方法，将短语特征纳入考量，并构建了一个图像主题数据集。其将输入的图片映射到一个主题词，再随机选择与该主题词相关的短语，再通过一正一反两种方向的 LLM 来依次生成古诗的首行、标题和其他主体内容。

值得注意的是，鲜有研究关注文图跨模态的古诗生成。据调研，目前只有

[23] 一项研究同时包含文本和图像两种模态的输入，其通过 Clarifai 来将图像映射到两个具体的主题词，根据这两个主题词在已有短语库中进行检索拓展，再进一步将得到的短语通过一个自注意网络来生成古诗。这一系统允许用户来限制古诗的诗句前缀词，因此可广义地认为实现了图文的跨模态输入。

1.2.2 古诗质量评价

如何评价生成古诗的质量是一个难题，考虑到古诗体裁的艺术性，其质量评价往往依赖于人工评估。除此之外，也可使用以往文本生成的自动度量方法，如源自机器翻译领域基于 *n*-gram①的翻译文本评估方法 BLEU[24] 和 ROUGE[25]。其中，BLEU 指标计算生成文本中有多少 *n*-gram 出现在参考文本中，即使用精确率（Precision）来评估生成文本有多接近参考文本；相反，ROUGE 指标计算参考文本中有多少 *n*-gram 能够被生成文本包含，即使用召回率（Recall）来评估生成文本能否完整地覆盖参考文本。可见，BLEU 和 ROUGE 指标均依赖于与高质量参考文本的对比。此外，Distinct[26] 基于 *n*-gram 的多样性来评估生成古诗的多样性，即诗句里有多少不同的 *n*-gram。而为了评估上下文语义的一致性，Similarity[27] 使用词向量来计算句子之间的相似度。这些方法都能脱离参考文本独立地评价文本质量。

除了自动度量方法外，也有一些研究实现了对古诗的质量优化。2020 年，[28] 提出一个质量感知的掩蔽语言模型，实现一个可迭代的古诗优化框架，可用于判断古诗是否需要优化，并在润色时定位不恰当部分。2023 年，[29] 提出一种人机协作的古诗创作系统，能够在不同的约束条件下对古诗进行润色。2024 年，[30] 又提出了一个可迭代的古诗优化框架，基于 BiLSTM 和 CRF 构建用于检测低质量用词的检测网络、基于 BERT 模型构建用于修正用词的校正网络。这些研究均实现了“评价 + 优化”的流程功能，且效果良好。美中不足的是未提供对“评价”本身的解释，即“被选中的词为何是低质量的，而优化又依据着什么”的问题。

①表示文本中连续的 *n* 个词或字符。如“我爱你”中基于字符的 2-gram 集合为 [“我爱”、“爱你”]

1.2.3 大模型技术

近年来，大模型技术取得了显著的进步，其中以大语言模型最为典型，例如OpenAI 开发的生成预训练转换器（GPT）系列模型、百度开发的知识整合增强表示（ERNIE）系列模型 [8]。DeepSeek 开发了首个通过强化学习训练的大语言模型 DeepSeek-R1[9]，其采用混合专家模型（Mixture-of-Experts, MoE）架构，实现在相同计算成本下大幅提升模型的参数规模与推理能力。

而与 LLM 擅长自然语言类似，还有许多跨模态大模型适用于不同模态间数据的信息表征处理，如文本和图像。由 OpenAI 开发的 CLIP 模型 [31] 包含一个文本解码器和一个图像解码器，在大量图像及其对应的文本描述的数据集上进行预训练，因而能够把握视觉表征与文本之间的关联。MiniGPT-4[32] 是 GPT-4 模型的缩小版，它将一个与 BLIP-2 架构相同的视觉编码器与语言模型 Vicuna 通过一个单一的投影层链接起来，在图像描述生成方面展现出卓越的能力。DeepSeek 开发的 DeepSeek-VL[10] 的训练数据来自广泛的现实场景，其采用了一个混合视觉编码器，可在固定的 token 预算内有效处理高分辨率图像，同时保持相对较低的计算开销，这一设计选择确保了模型在各种视觉任务中捕捉关键语义和详细信息的能力。有趣的是，DeepSeek-VL 在一开始便整合了 LLM 的能力训练，促进视觉和语言两种模态能力的平衡整合，使其在能够捕捉视觉语义信息的同时，仍然具有强大的语言能力。在此基础上，DeepSeek-VL2[33] 进一步改进了视觉与文本的能力整合。视觉上融入动态分块的编码策略以处理不同比例尺寸的图像，文本上则引入了 MoE 架构，可动态选择专家模型完成不同的任务。

1.2.4 提示工程

提示工程逐渐作为一种通过自然语言来调整和控制 LLM 行为的技术。目前人们已提出多种提示词的设计原则与策略。Few-shot 框架 [34] 为模型提供少量示例样本，以有效地指示模型完成指定任务。思维链（CoT）框架 [35] 指导模型将任务分解为若干子任务来逐步完成，使模型能在无其他修改的情况下完成复杂的因果推理任务。自洽性（Self-Consistency）[36] 旨在小样本 CoT 中对多种推理路径进行采样，并在尝试生成之后选择最一致的答案，其有助于提高 CoT 提示在涉及算术推理和常识推理的任务中的性能。此外，开发者们也提出了十分多

样的提示词框架，例如 CRISPE、ICIO、BROKE，均经过了开源社区的检验，令人目不暇接。

1.3 研究思路和主要贡献

本选题旨在利用大模型的通用能力，设计并实现一个文图跨模态的中文古诗创作系统：1）分析输入的图像，生成易使用的描述文本，兼顾关键物体与整体氛围；2）设计评判标准，设计提示词分别用于古诗生成和分析，调用模型生成古诗、分析结果与改进意见；3）结合已有的质量评价维度和其他系统，设计实验验证古诗分析分数、跨模态生成对古诗质量影响的效度，并探索系统改进方向。

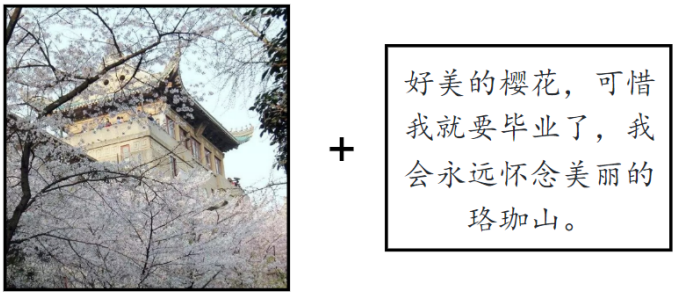

图 1.1　示例输入

为处理输入的图像模态，需要使用跨模态能力优良的视觉与文本编码模型。在先前的工作中，使用 CLIP 模型来识别图像中的物体，使用 MiniGPT4 模型来生成初步的图像描述，之后再由 ERNIE-4.0 结合二者生成最终的中文描述文本。诚然，CLIP 与 MiniGPT4 模型均具备极强的视觉与文本的映射能力，但二者均属英文模型，缺乏中文文化语境的训练，因而难以捕捉图像中潜在的文化意味与情感，这部分缺乏的信息难以在 ERNIE-4.0 模型的总结中恢复，从而影响古诗生成的质量。为此，本文采用在中文语境下训练的模型来处理图像模态，经初步对比，DeepSeek-VL 模型生成的图像描述文本更具诗意美感，且在古诗生成的质量上也更具优势，因而本文将使用 DeepSeek-VL 模型来处理图像模态。（输出如图 1.2）

古诗具有独特的韵律格式要求，如平仄、押韵、对仗等要求，且注重典故与意境协调。在先前的工作中，使用 ERNIE-4.0 模型来生成古诗，虽能基本遵守韵

这张图片展示了一座传统中式建筑，其屋顶装饰着精美的绿色琉璃瓦，显得庄重典雅。建筑主体为浅黄色，四周有栏杆围绕，似乎有游客站在阳台上欣赏风景。前景是繁茂的樱花树，粉白色的花朵覆盖了大部分画面，营造出浪漫而宁静的氛围。天空清澈，与花朵的柔和色彩形成鲜明对比，增添了画面的层次感。整幅画面让人感受到春天生机勃勃的气息，同时也透露出一种悠闲和诗意的情感。

图 1.2　图像描述（示例）

律格式要求，但无法兼顾典故运用和意境圆满。为此，本文使用基于强化学习训练的 DeepSeek-R1 模型（生成古诗如图 1.3），其链式的推理过程包含对多种候选推理路径的动态评估和筛选，能够在反复验证中确保生成古诗的韵律要求，并且也展现了对典故意象的更强运用能力。这一能力也有助于提高系统的可解释性，为用户输出更合理清晰的解释性文本，包括白话文翻译、古诗赏析和典故注释。（如图 1.4、 1.5和图 1.6）

珞珈山下琼英绽，朱阁檐前翠影摇。
万点飞花迷画槛，千重香雪覆虹桥。
琉璃映日浮金阙，玉砌连云接碧霄。
曲径曾携书卷过，长亭忍看落英飘。
东君不解离人恨，更遣残红逐晚潮。
粉泪暗随春水逝，青衿空对暮山遥。
他年若问芳菲事，犹记晴川鹤影娇。

图 1.3　生成古诗（示例）

珞珈山下樱花绽如琼瑶，朱红楼阁前翠影正轻摇。
万点飞花迷蒙了雕栏画柱，千层香雪压弯了求学虹桥。
琉璃瓦映日似浮动金殿，白石阶入云直通向碧霄。
曾捧书卷走过这林荫曲径，长亭外怎忍看落英飘摇。
春神不懂人间别离苦楚，偏让残红追着晚潮远漂。
粉泪默默随春水流逝去，青衫空对暮色里的山遥。
若问当年花开烂漫旧事，定记得晴川上鹤影正娇。

图 1.4　古诗白话文翻译（示例）

为有效地评估生成古诗的质量，本文设计了一个包含多维度的评分体系，以

这首七言排律以武大樱花为情感载体，通过工丽的对仗与典雅的意象，构建出毕业离别的双重意境。首联以”琼英”“翠影”点题，既暗合《诗经》”尚之以琼英乎而”的典故，又呼应图片中琉璃瓦与樱花的色彩映照。颔联”万点飞花”与”千重香雪”形成空间纵深感，化用岑参”千树万树梨花开”的意境而翻出新意。颈联转写建筑，“琉璃映日”直取图片意象，“玉砌连云”暗用李煜”雕栏玉砌”之典，将现代学府与传统意象完美融合。尾联”晴川鹤影”既化用崔颢”晴川历历汉阳树”，又暗喻学子如鹤高飞，留下余韵悠长的想象空间。全诗严守平水韵，中二联”迷画槛”对”覆虹桥”，“浮金阙”对”接碧霄”，工稳中见灵动，以物象变迁写时光流逝，最终在”鹤影”的意象中完成对母校记忆的诗意定格。

图 1.5　古诗赏析（示例）

1. 琼英：取自《诗经·齐风》“尚之以琼英乎而”，既指樱花如玉，又喻学子才俊
2. 朱阁：对应图片中浅黄色建筑，化用苏轼”转朱阁”意象，暗喻学术殿堂
3. 虹桥：既实指校园拱桥，又暗用《东京梦华录》虹桥典故，喻求学之路
4. 琉璃：特指图片中绿色琉璃瓦，呼应白居易”琉璃波面月”的华美意象
5. 玉砌：化用李煜”雕栏玉砌应犹在”，将现代建筑古典诗化
6. 东君：司春之神，此处双关东风与毕业季的时令特征
7. 青衿：源自《诗经》“青青子衿”，既指学子装束，又喻求知初心
8. 晴川鹤影：化用崔颢”晴川历历汉阳树”，”鹤”取《诗经》”鹤鸣于九皋”意象，喻志向高远
9. 香雪：借李商隐”匝路亭亭艳”咏樱花之意，特指白色樱花如雪
10. 画槛：对应图片中栏杆意象，暗用”画栋朝飞南浦云”的滕王阁典故

图 1.6　古诗注释（示例）

进行量化评分并促进优化方案的梳理。在之前的工作中，基于六个方面②设计提示词，其中对每个方面进行简要描述，并使用 ERNIE-4.0 模型打分。在测试时发现，这样的评分体系无法提供标准化的分数，输出的分数整体偏高，区分度差。因此，本文重新设计了一个包含五个方面③的评分体系，不同于之前对各方面的简要描述，其给出了各分数段的情况描述，并附上了具体例子（如表 A.1），有效地提升了评分结果的区分度和合理性，也为后续的优化方案提供了更清晰的方向。而作为补充，本文也将利用 BLEU、ROUGE、Distinct、Similarity 等指标

②包含“结构与形式”、“语言与风格”、“意象与主题”、“协调与一致”、“历史语境”、“创新性”

③包含“格律规范”、“意象意境”、“主题思想”、“语言锤炼”、“创新性”

辅助古诗质量的评估。

本研究的主要贡献在于探索了大模型在中文古诗生成中的应用潜力，结合文图两种模态来强化生成古诗与用户需求的契合度，期间提供用户友好的解释性文本，使得用户更易理解和信任机器生成的创作过程及其背后的文化内涵；同时，设计评分体系来输出量化结果，在促进古诗的多轮优化的同时进一步提高系统可解释性。本研究有助于拓展和弘扬中华传统文化的表达形式，推动诗词创作与人工智能技术的深度融合，具有重要的理论意义和实践价值。

1.4 论文组织结构

第一章是概论部分。

第二章介绍古诗生成任务的现有技术，对古诗的基本要求、自动度量方法、DeepSeek 大模型等进行概述。

第三章介绍本系统的设计与实现，对整体架构、各个模块的功能和实现方法进行概述。

第四章介绍开展实验的设计与结果分析，基于已有古诗数据集，结合自动度量方法来评估系统，并展开分析论述。

第五章是结语，对本文的工作内容进行总结，并探讨局限性与改进空间。

2 古诗生成与大模型

本章主要介绍古诗生成任务的基本要求、现有技术方案以及古诗质量的度量方法，简要介绍 DeepSeek-R1 与 VL2 两个大模型，为后续的系统设计方案作铺垫。

2.1 古诗生成任务概述

古诗生成任务要求模型能够根据需要生成符合要求的中国传统诗词，这要求模型的输出不仅能达到古诗的韵律、对仗等形式要求的建筑美，还要在内容上符合古诗的意境、情感等内涵要求。

古诗体裁多变，产生于汉朝前的古体诗形式自由、不拘泥于严格的格律，唐朝的近体诗（如律诗、绝句）与词则都格律严格。而在古诗生成任务中，往往会选择格律严格的近体诗或词作为生成目标，以便评估模型约束输出的能力。格律要求大概如下三点：

1. 押韵：指诗中某些句子的末尾字使用相同或相近的韵母，形成韵脚和谐的音韵效果。律诗和绝句均要求在偶数句的句末押韵，首句可押可不押，且通常押平声韵。在唐诗中，往往要求一韵到底，即整首诗只能使用同一个韵部的字来押韵，中途不可换韵。
2. 平仄：指汉字声调的高低，分为平声和仄声，在现代汉语中大致对应为一二声和三四声。在唐诗中，五言与七言各有不同的格律格式，如五言诗有四种基本的平仄类型，分别为“平平平仄仄”“仄仄平平仄”“仄仄仄平平”“平平仄仄平”，不同类型可在同一首诗中交替使用。此外，律诗的偶数句中的第二字须与前一句的第二字平仄相同，称为“粘连”。
3. 对仗：指律诗中颔联（第三、四句）与颈联（第五、六句）在词性、语义、平仄等结构上形成对称关系，这也是律诗重要的特征之一。如“大漠孤烟直，长河落日圆”，前后两句的词性、语义、平仄均形成对称关系，朗朗上口。

排律是格律的变体，其延续律诗严格的格律要求，但篇幅较长，通常在十句以上，不过押韵要求较为宽松，中途可换韵。排律的篇幅和结构更为复杂，十分考验创作者的文学功底和创作技巧，因而在传统诗歌中，排律的作品相对较少。

2.2 古诗质量评估

2.2.1 BLEU

BLEU（Bilingual Evaluation Understudy），又名双语替换测评，是一种用于评估机器翻译质量的指标，核心是通过比较机器翻译结果与参考翻译之间的 *n*-gram 匹配情况来评估翻译质量，并以精确率（Precision）作为衡量指标。[24]

具体而言，其计算候选句子中在参考句子中出现的 *n*-gram 的次数 $Count$。而为了避免导向无意义的 *n*-gram 重复，BLEU 对候选句子中的 *n*-gram 计数进行截断（clip），使同一个 *n*-gram 的计数不超过参考句子中该 *n*-gram 的最大次数，即 $Count_{clip} = \min\{Count, Max\_Ref\_Count\}$。

而在含有多个句子的长文本段落中，BLEU 的处理单元依旧是其中的句子。对候选段落中的每个句子 C，计算其所有的 *n*-gram 的截断后计数 $Count_{clip}$，再按句子累加在一起，整体除以同理得来的非截断计数 $Count$，于是得到一个归一化的分数，以适用于不同长度的文本比较。修正后的精确率如式（2.1）：

$$p_n = \frac{\sum_{C\in\{Candidate\}} \sum_{n\text{-}gram\in C} Count_{clip}(n\text{-}gram)}{\sum_{C\in\{Candidate\}} \sum_{n\text{-}gram\in C} Count(n\text{-}gram)} \tag{2.1}$$

为了结合不同 *n* 取值下 *n*-gram 的指标分数，对不同的 *n*-gram 分数进行加权几何平均，其中权重 $\sum \omega_i = 1$，如式（2.2）：

$$\exp\left(\sum_{n=1}^{N} \omega_i \log p_n\right) \tag{2.2}$$

进一步，为了避免机器翻译生成过短的句子来提高匹配的精确率，引入了简洁性惩罚（Brevity Penalty, BP），对候选文本 C 的长度 c 与参考文本 $\mathcal{R}$ 的长度 r 进行比较，计算方式如式（2.3）。

$$\mathrm{BP} = \begin{cases} 1, & c > r \\ e^{1-\frac{r}{c}}, & c \le r \end{cases} \tag{2.3}$$

最终可得 BLEU 的完整公式如式（2.4）：

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} \omega_i \log p_n\right) \tag{2.4}$$

BLEU 的值范围在 $[0, 1]$ 之间，值越大表示翻译质量越高，但绝对的数值意义不大，因此不需要追求接近于 1 的分数。

在实际使用中，权重 ω_i 通常设置为 $\frac{1}{N}$，即 *n*-gram 的加权平均。例如，BLEU-2

的权重设置为 $\omega_1 = \omega_2 = 0.5$。古诗中的词大多是一到两个字，因此古诗生成任务中通常使用 BLEU-1 和 BLEU-2。

2.2.2 ROUGE

与 BLEU 类似，ROUGE（Recall-Oriented Understudy for Gisting Evaluation）同样通过比较候选文本与参考文本之间的重叠 *n*-gram 来衡量文本翻译的准确率，但又更适用于文本摘要等注重信息提取和保留的任务。[25]

区别于 BLEU 适用准确率，ROUGE 使用召回率（Recall）作为衡量指标，计算候选文本 $\mathcal{C}$ 中重叠的 *n*-gram 的数量与参考文本 $\mathcal{R}$ 中 *n*-gram 的数量之比。如式（2.5）

$$\text{ROUGE-n} = \frac{\sum_{C\in\{Candidate\}} \sum_{n\text{-}gram\in C} Count_{clip}(n\text{-}gram)}{\sum_{\mathcal{R}\in\{Reference\}} \sum_{n\text{-}gram\in\mathcal{R}} Count(n\text{-}gram)} \tag{2.5}$$

此外，ROUGE 也支持多种变体，如 ROUGE-L 基于参考文本和候选文本之间的最长公共子序列（LCS）来计算召回率，如式（2.6）。

$$\text{ROUGE-L} = \frac{\sum_{C\in\{Candidate\}} \text{LCS}(\mathcal{C},\mathcal{R})}{\sum_{\mathcal{R}\in\{Reference\}} \text{Length}(\mathcal{R})} \tag{2.6}$$

2.2.3 Distinct

除了 BLEU 和 ROUGE 等基于参考文本对比的度量方法外，也有方法尝试独立地衡量文本自身的质量。其中一个例子是 Distinct 指标，其计算文本中独特的 *n*-gram 的数量与文本中所有 *n*-gram 的数量之比，以衡量文本中用词的多样性。而同理于 ROUGE，常常选取 1-gram 和 2-gram 来计算 Distinct 指标。[26, 37]

2.2.4 Similarity

为了衡量古诗中前后句子间的语义联系和一致性，也有研究尝试使用词向量，并基于余弦相似度来计算句子间的语义相似度。[28] 如在 [30] 和 [28] 中，由于输出固定为绝句（共四句诗），因此可固定计算头两句相似度Sim12、后两句相似度Sim34、以及二者的相似度Sim2L。但对于未约束体裁的古诗生成任务，如律诗、绝句、排律，这样静态的计算方法并不适用。

由此，本文使用了更通用的指标Sim_intra和Sim_inter，分别表示古诗中前后句子间的相似度和不同联之间的相似度。具体而言，对古诗中的句子两两分组，

Sim_intra计算组内两个句子的相似度并取平均（如第一句和第二句、第三句和第四句），Sim_inter计算组间的相似度再取平均（如第一二句和第三四句、第三四句和第五六句）。

这一方法的核心在于利用词向量模型在预训练阶段学习到的句子间的语义关系，因而指标分数十分依赖词向量模型的质量。为此，本文采用清华大学自自然语言研究中心开源的 Bert-CCPoem 词向量模型①，其在一个包含几乎所有中国传统诗词的数据集 CCPC-v1.0 上进行预训练，涵盖 926024 首古诗的 8933162 个句子，可提供高质量的古诗句子词向量。

2.2.5 人工评估

由于古诗体裁的高度的艺术性，过去的研究中往往会邀请人类评审来评估古诗的质量。评估往往会基于单独设计的分析角度进行，如“流畅性”、“艺术性”、“连贯性”等等，依赖于分析维度的先验设计。此外，人类评审的结果往往依赖于评审员自身的文化素养、个人品味与喜好，结果难有一致性。此外，招募具有高文学素养的人类评审员也是一个难题。

由此，本文认为可利用大模型的语言能力行使人类评审员的功能，设计一套严格的质量评估体系，作为提示词指导大模型，便可得到具有高解释性、高一致性的古诗质量评估结果。

2.3 DeepSeek 大模型

2.3.1 DeepSeek-VL2

DeepSeek-VL2[33] 采用了三阶段混合架构，包括视觉编码器、视觉-语言适配器和混合专家语言模型三大模块，如图 2.1。

1. 视觉编码器（Vision Encoder）：为了支持更大分辨率、不同比例尺寸的图像，在先前 VL1 模型 SigLIP 框架的基础上引入动态切片策略（dynamic tiling strategy），将高分辨率图像自适应分割为 384×384 的子图块，再与全局缩略图块组合。该策略在保证宽高比不变的前提下，通过填充面积最小化算

①https://github.com/THUNLP-AIPoet/BERT-CCPoem

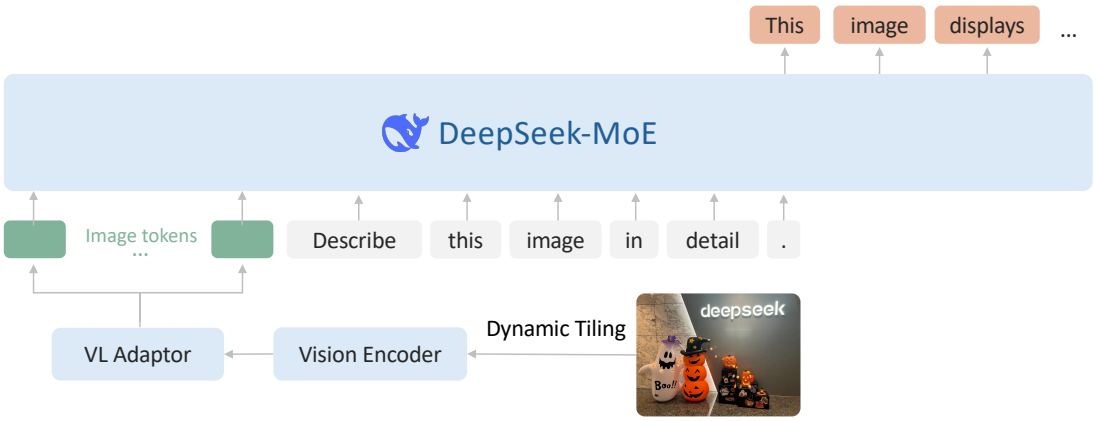

图 2.1　DeepSeek-VL2 模型架构 [33]

法选择最优分割方案，解决了传统固定分辨率编码导致的细节丢失问题，使模型支持最高 1152×1152 的分辨率输入。如图 2.2所示。

图 2.2　DeepSeek-VL2 动态切片策略 [33]

2. 视觉-语言适配器（VL Adaptor）：承接上一步中加入的<换行Token>（用户区分局部图块的结束）和<视觉分隔Token>（用于区分全局缩略图和局部图块），采用双层感知机与 2×2 像素混洗（pixel shuffle）操作，将视觉特征维度压缩映射到文本嵌入空间。该设计在保留局部细节特征的同时，实现了跨模态特征的高效对齐。
3. 混合专家语言模型：采用稀疏激活的专家混合架构，每个输入 token 动态激活 Top-2 专家网络。结合多头潜在注意力（Multi-head Latent Attention, MLA）机制，通过奇异值分解键值缓存压缩为潜在向量，使 4096 长度序列显存占用降低至传统架构的 6.7%

DeepSeek-VL2 的训练采用了三阶段训练范式，即先使用 120 万个图文对来

建立跨模态关联，再使用混合 70% 的图文数据和 30% 纯文本数据来进行预训练，最后再专注于 OCR 增强和文档理解方面的监督微调。值得一提的是，DeepSeek-VL2 的 MoE 架构包含有 64 个专家分组，每组都能处理特定的模态组合，并能在训练中通过负载均衡损失函数来优化专家激活分布，实现稀疏激活，从而大大提高推理效率（在总体 176B 的参数量下仅激活 4.5B 参数）。

2.3.2 DeepSeek-R1

为了训练模型的推理能力，过去的方法通常是在监督微调（supervised fine-tuning, SFT）后加入大量的思维链（CoT）范例数据，引导模型学会链式的推理思考。但 DeepSeek-R1[9] 则采用了强化学习 (RL) 的路径来训练模型推理能力。

在正式构建 DeepSeek-R1 前，DeepSeek 团队先尝试验证强化学习方向的可行性——直接在基底模型上应用强化学习，而不使用任何 SFT 的数据，通过准确性奖励和格式奖励来训练出 DeepSeek-R1-Zero。此外，为节省训练开销，采用相对策略优化（Group Relative Policy Optimization, LGRPO）算法，通过计算组内输出结果的得分均值来获得整个损失函数的期望值。

而 DeepSeek-R1 的训练过程分为四个阶段（如图 2.3）。

1. 冷启动 SFT：为了解决 RL 训练早期的不稳定性和语言混杂的问题，先采用较小规模的 CoT 数据集，对基底模型 DeepSeek-V3-Base 进行冷启动 SFT，作为 RL 训练的初始模型。这里使用数据是带有反思和验证的详细思考答案，由 DeepSeek-R1-Zero 生成。
2. 推理导向 RL：流程与训练 DeepSeek-R1-Zero 一致。此外，为了减少语言混杂的问题，引入了语言一致性奖励，以性能略微下降为代价，提高模型输出的可读性。
3. 拒绝采样 SFT：利用上一阶段 RL 训练过程中的 checkpoint 进行拒绝采样，生成多个候选的推理轨迹，再利用 DeepSeek-V3 充当奖励模型来进行打分，仅保留评分最高的样本，以此获得 60 万推理数据。此外，还纳入了 DeepSeek-V3 的部分训练数据，并对部分任务进行提示生成，获得 20 万非推理数据。于是模型在大小为 80 万的数据集上进行两轮微调。
4. 全场景 RL：为进一步实现人类偏好对齐，再次进行 RL 训练，使用基于规则的奖励来训练推理方面的学习，使用奖励模型来训练通用领域中较隐晦

的人类偏好。

此外，可直接使用“拒绝采样 SFT”中使用的 80 万数据集对其他较小模型（如 Qwen2.5-14B 和 Llama3.1-8B）进行 SFT，得到蒸馏模型 DeepSeek-R1-Distill。需要注意，这一过程并未包含强化学习训练。

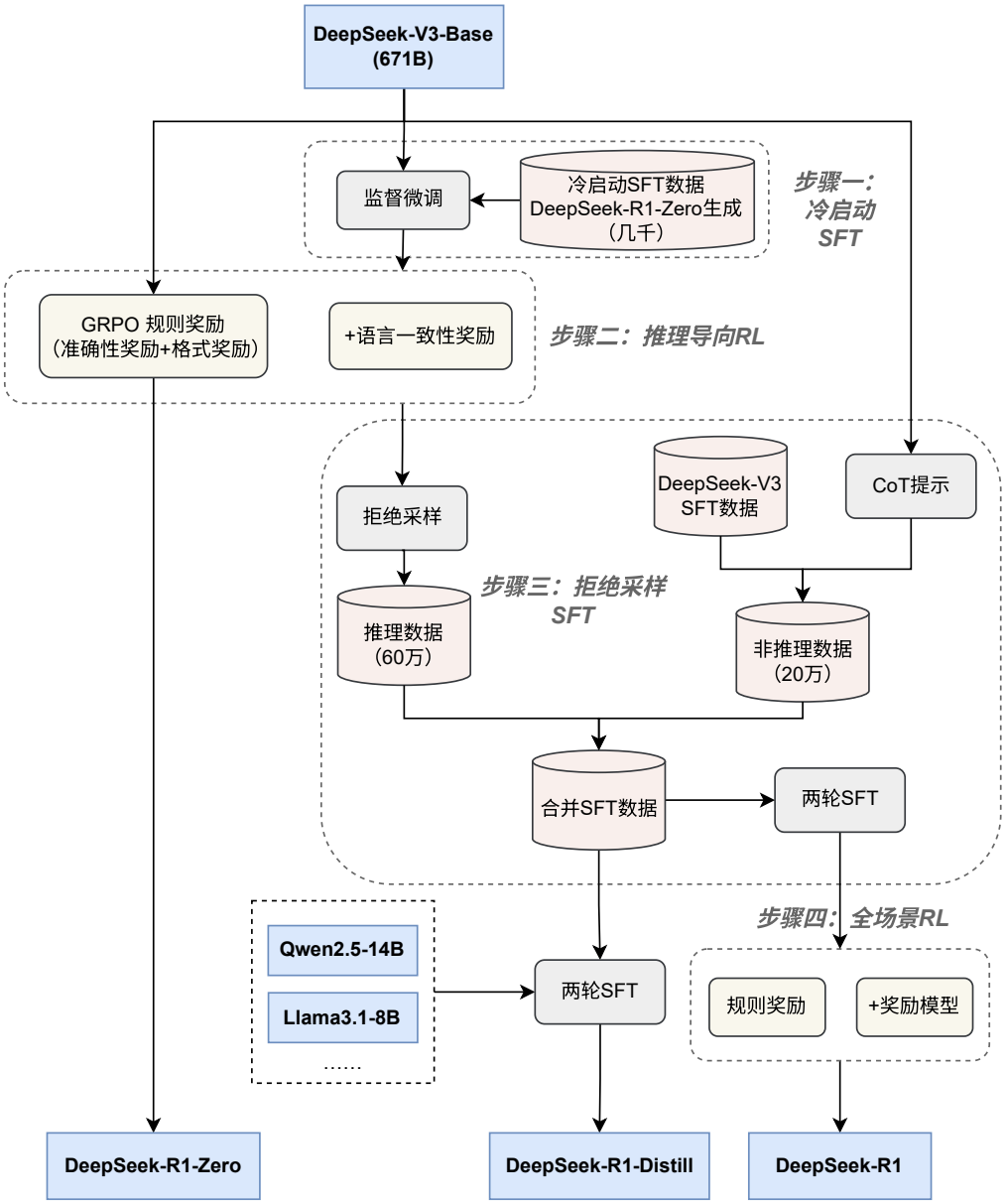

图 2.3　DeepSeek-R1 训练过程

3 系统设计与实现

本章主要介绍本系统的设计与实现，包括系统的整体架构、各个模块的功能和实现方法。

3.1 系统架构概述

本系统包含图像分析、古诗生成、古诗评价和古诗优化四个模块，基于 Python 语言开发，使用百度智能云提供的 API 接口来调用模型，整体架构如图 3.1所示。

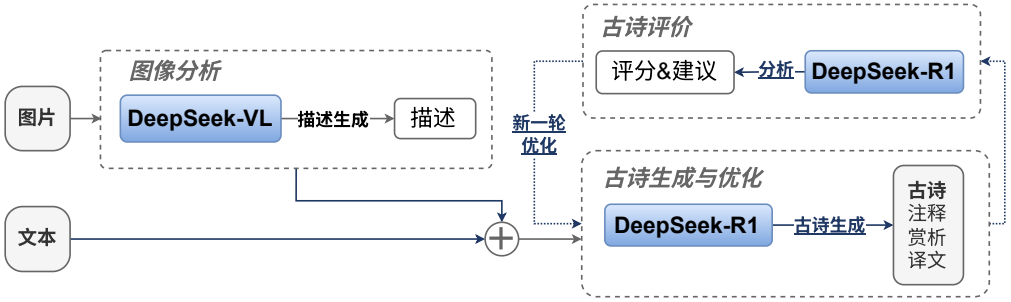

图 3.1 系统架构

3.2 图像分析

在先前的方案中，图像分析使用的是英文模型 CLIP 和 MiniGPT-4，尽管生成的描述较精确、但无法捕捉有助于古诗创作的中国文化联想素材与情感色彩。因此，本系统使用 DeepSeek-VL2[33] 替代之前的多模型组合方案，一步到位地为图像生成兼顾关键物体识别、整体场景信息和情感色彩的描述。（提示词见图 B.1）

在调用百度智能云的 API 接口时，需要提供提示词和图像的 URL 地址，因此还需要将用户提供的图像上传到云端，并获得可公开访问的 URL。为此，使用阿里云的对象存储 OSS 服务，利用 OSS Python SDK，将用户图像上传到云端后，生成带有过期时间的 GET 方法预签名 URL，供后续 API 调用使用。

3.3 古诗生成

在先前的方案中，古诗生成使用的是 ERNIE-4.0 模型，其无法在遵守韵律要求的同时充分运用经典典故意象，更别说提供使用典故的注释。因此，本系统使用 DeepSeek-R1[9]，提示词设计参考 CRISPE 框架（提示词见图 B.2）。

CRISPE 框架包括六个部分：

1. 能力与角色（Capacity and Role）：大模型应当具备的角色与能力。
2. 背景信息（Insight）：为完成任务，大模型应当知晓的背景知识信息，以及用户需求的上下文语境。
3. 指令（Statement）：大模型需要完成的任务。
4. 输出风格（Personality）：大模型输出回复的风格、特色以及规范。
5. 实验（Experiment）：尝试让大模型提供一些例子，以便更好地调试提示词。

其中，最后投入使用的只包括前五个部分，而最后一个部分“实验（Experiment）”只是作调试用，方便提示词的设计迭代。

在生成古诗时，系统接收用户的文本输入user_text，与用户输入图像的描述description结合，同时指定古诗的体裁poem_type（如五言绝句、七言律诗、不少于 8 句的排律等）。而除了古诗本身外，系统还会输出古诗的赏析、对古诗中典故的注释、以及白话文翻译，以便用户充分地理解古诗的意境、内涵和创作思路。

3.4 古诗评价

为了评价古诗的质量，本系统在自动度量方法的基础上，引入了 DeepSeek-R1 的评分机制。而为了使其能够给出合理、细致的评分，本文设计了一套包含五大维度的古诗评分体系（见表 A.1），在先前工作的基础上，这套体系强调了对子维度中分数段的详细划分，并提供对应的示例以进一步阐明评分标准。基于这套评分体系，系统将逐一分析古诗的每个维度，给出分数和评语。此外，系统还将依照这套体系，逐一地给出修改意见，以提高古诗的质量。

由于评分体系较复杂、所占文本较多，提示词的设计参考了 Few-shot 框架[34]，在完成需求说明后，给定两个输入与输出的范例，确保模型按照预期的格

式输出内容。（如图 **??**）

TODO 图例：评分、雷达图和优化建议

此外，系统还将结合自动度量方法（BLEU、ROUGE、Similarity、Distinct）来进行指标计算，以辅助古诗的质量评估。

3.5 古诗优化

为了对生成的古诗进行优化，系统基于之前分析得到的改进意见，对古诗进行进一步的迭代润色，基于评分体系中的薄弱部分针对性地提高分数。

在先前的工作中，古诗优化的输入只包含待改进的古诗poem和提供的修改意见suggestion，这会导致两个问题——其一，修改意见suggestion中的建议往往只包含那些明显不足的方面，并不能覆盖评分体系的所有维度，因此在优化时，模型很可能会忽略那些未被提及的维度，进而导致优化后的古诗在这些方面的分数下降，顾此而失彼；其二，在古诗优化时，模型只考虑了古诗的内容和结构，却没有考虑用户的原始输入，因此优化过程也可能会偏离用户的初衷，导致生成的古诗与用户的期望相去甚远。

因此，本系统在古诗优化时，除了原古诗poem和改进建议suggestion外，还增加了先前对原古诗的评分evaluation和用户的输入（文本输入user_text与图像输入描述description），在确保古诗优化有效性的同时，保留对用户需求的考量。

TODO 图例：优化后古诗、优化说明

3.6 项目结构

TODO：< 代码结构图 >

TODO：< 运行界面截图 >

4 实验及结果分析

为验证本系统在古诗生成、评价和优化方面的有效性，本文设计开展了相关实验，利用已有古诗数据集来检验系统性能。此外，结合自动度量方法对比了ERNIE-4.0 和 DeepSeek-R1 两种模型产出的结果。本章首先……

4.1 基于白话文的古诗生成实验

参考相关工作中评估模型表现的方法，基于相同的输入，将评估模型的输出古诗与原古诗进行对比，利用 BLEU 和 ROUGE 计算指标分数。在这一思路下，原古诗的质量应当足够好，以确保指标分数能够反映被评估输出的质量；且模型的输入要与原古诗的内容十分相关，以确保指标分数能够反映模型的生成能力。

为此，本文选择《唐诗三百首》中的名篇作为古诗原文，选择相应的白话文翻译作为文本输入。《唐诗三百首》是清代蘅塘退士孙洙编选的唐诗选集，问世不久便闻名遐迩，成为唐诗入门读物的首选，其中收录的唐诗均为名篇佳作。原书中除了律诗、绝句外，还收录有乐府、古体诗等形式多变的体裁，故不便于评估，本文仅选择律诗和绝句两种体裁的古诗进行测试，如表 4.1。此外，测试集选择的白话文翻译源自古诗文网①。

表 4.1 《唐诗三百首》测试数据集

体裁	数量
七言律诗	51
七言绝句	50
五言律诗	80
五言绝句	29
合计	210

测试发现，ERNIE-4.0 和 DeepSeek-R1 两种模型的输出均与参考古诗有极高的覆盖度，因而在 BLEU 和 ROUGE 指标上均有很高的得分，明显区别于以往工作中的结果（如BLEU-1=0.168, BLEU-2=0.002[30]），属于异常情况。作为大模型，两种模型的训练数据均包含了大量的古诗数据，尤其是对《唐诗三百首》这样的名篇，因而本实验并不能做到训练集与测试集的独立，并不具备验证效果。

①古诗文网: https://www.gushiwen.cn/gushi/tangshi.aspx/

测试统计结果见表 4.2和表 4.3。

表 4.2　白话文古诗生成实验结果（DeepSeek-R1）

	BLEU		ROUGE		
	BL-1	BL-2	R-1	R-2	R-L
七言律诗	0.583599	**0.432841**	**0.557353**	**0.349461**	0.527267
七言绝句	0.559833	0.415325	0.562665	0.344502	0.540417
五言律诗	**0.597726**	0.432494	0.549401	0.337500	**0.545573**
五言绝句	0.523605	0.332445	0.495130	0.245845	0.471983
平均	0.575037	0.414674	0.546996	0.329415	0.529737

表 4.3　白话文古诗生成实验结果（ERNIE-4.0）

	BLEU		ROUGE		
	BL-1	BL-2	R-1	R-2	R-L
七言律诗	0.765951	0.671162	0.770484	0.609392	0.743464
七言绝句	0.798136	0.702556	0.808346	0.649742	0.764375
五言律诗	0.691980	0.533006	0.670557	0.445450	0.646720
五言绝句	**0.839312**	**0.764698**	**0.834125**	**0.726052**	**0.812500**
平均	0.767420	0.658670	0.765072	0.596372	0.734993

4.2　评分功能实验

为了检验系统评分功能的有效性，本文尝试收集具有质量差异的古诗分组，通过先验的质量分层来验证系统评分的合理性。此外，也将使用 BLEU 等自动度量方法，作为实验的参考指标。

选择第六届“诗词中国”传统诗词创作大赛的公开获奖作品为测试集，测试系统评分功能的有效性和可信度。该大赛的评审专家均为古诗词领域的专家，且其评分标准公开透明，因而可以作为测试系统评分功能的参考。

测试发现，不同奖项组之间的评分存在差异但并不显著，且由于各奖项样本数量分布不均（一等奖 2 首，二等奖 8 首，三等奖 18 首，优秀奖 91 首），结论难有普适性。而往届获奖作品并不公开，因而无法继续这一方向的测试。

TODO 数据：获奖作品的评分实验

为此，本文扩大测试古诗范围，从古诗文网挑选“打油诗”，结合《唐诗三百首》的部分古诗（五言律诗和七言律诗），与原先的获奖作品一同作为测试集。

TODO 数据：唐诗、打油诗的评分实验

4.3 文图结合的古诗生成实验

为验证图片模态的必要性，需要设置文图输入的消融对比实验。之前的工作中表明，面对不同的输入模态组合，ERNIE-4.0 的输出古诗质量会有显著的差异。本文尝试对 DeepSeek-R1 进行相同的实验，即固定文本和图像输入，分别测试仅文字、仅图像、文字与图像三种模态输入，对比生产的古诗质量。古诗的生成质量依据系统自身的评分功能。

TODO 数据：模态组合的评分实验

实验发现，无论哪种模态组合，系统生成古诗的评分均较高，不存在显著差异。

图像输入的作用体现在其包含的视觉场景信息，能够作为用户文本输入的补充，帮助用户表达隐晦的场景情感，以提高用户的体验。但就古诗生成的质量而言，DeepSeek-R1 的能力足以生成高分数的古诗，这一点与用户的需求是不同的。

5 结语

参考文献

[1] Oliveira H G. PoeTryMe: a versatile platform for poetry generation[J]. Computational Creativity, Concept Invention, and General Intelligence, 2012, 1 : 21.

[2] Yan R, Jiang H, Lapata M, et al. I, Poet: Automatic Chinese Poetry Composition through a Generative Summarization Framework under Constrained Optimization[A]. IJCAI '13 : Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence[C], Beijing, China : AAAI Press, 2013 : 2197 – 2203.

[3] He J, Zhou M, Jiang L. Generating Chinese Classical Poems with Statistical Machine Translation Models[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2012, 26(1) : 1650 – 1656.

[4] Yi X, Li R, Sun M. Generating Chinese Classical Poems with RNN Encoder-Decoder[A]. Sun M, Wang X, Chang B, et al. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data[C], Cham : Springer International Publishing, 2017 : 211 – 223.

[5] Zhang X, Lapata M. Chinese Poetry Generation with Recurrent Neural Networks[A]. Moschitti A, Pang B, Daelemans W. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)[C], Doha, Qatar : Association for Computational Linguistics, 2014 : 670 – 680.

[6] Li P, Zhang H, Liu X, et al. Rigid Formats Controlled Text Generation[A]. Jurafsky D, Chai J, Schluter N, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics[C], Online : Association for Computational Linguistics, 2020 : 742 – 751.

[7] Hu Z, Liu C, Feng Y, et al. PoetryDiffusion: Towards Joint Semantic and Metrical Manipulation in Poetry Generation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(16) : 18279 – 18288.

[8] Zhang Z, Han X, Liu Z, et al. ERNIE: Enhanced Language Representation with Informative Entities[J], 2019(arXiv:1905.07129).

[9] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning[J], 2025(arXiv:2501.12948).

[10] Lu H, Liu W, Zhang B, et al. DeepSeek-VL: Towards Real-World Vision-Language Understanding[J], 2024(arXiv:2403.05525).

[11] Yu C, Zang L, Wang J, et al. CharPoet: A Chinese Classical Poetry Generation System Based on Token-free LLM[A]. Cao Y, Feng Y, Xiong D. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)[C], Bangkok, Thailand : Association for Computational Linguistics, 2024 : 315 – 325.

[12] Liu L, Wan X, Guo Z. Images2Poem: Generating Chinese Poetry from Image Streams[A]. MM '18 : Proceedings of the 26th ACM International Conference on Multimedia[C], New York, NY, USA : Association for Computing Machinery, 2018 : 1967 – 1975.

[13] Xu L, Jiang L, Qin C, et al. How Images Inspire Poems: Generating Classical Chinese Poetry from Images with Memory Networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1).

[14] Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning[J], 2017(arXiv:1702.08608).

[15] Wang Z, He W, Wu H, et al. Chinese Poetry Generation with Planning Based Neural Network[A]. Matsumoto Y, Prasad R. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers[C], Osaka, Japan : The COLING 2016 Organizing Committee, 2016 : 1051 – 1060.

[16] 王乐为, 余鹰, 张应龙. 基于 Seq2Seq 模型的自定义古诗生成 [J]. 计算机科学与探索, 2020, 14(6) : 1028 – 1035.

[17] Wu C, Wang J, Yuan S, et al. Generate Classical Chinese Poems with Theme-Style from Images[J]. Pattern Recognition Letters, 2021, 149 : 75 – 82.

[18] Shao Y, Shao T, Wang M, et al. A Sentiment and Style Controllable Approach for Chinese Poetry Generation[A]. CIKM '21 : Proceedings of the 30th ACM International Conference on Information & Knowledge Management[C], New York, NY, USA : Association for Computing Machinery, 2021 : 4784 – 4788.

[19] 李晓辰. 风格和情感控制的中国古诗生成 [D]. 哈尔滨 : 哈尔滨工业大学, 2022.

[20] Ren X, Chai X, Mao M. Generating Chinese Poetry from Images Based on Deep Learning[J]. Proceedings of the 2023 3rd International Conference on Big Data, Artificial Intelligence and Risk Management, 2023 : 134 – 138.

[21] 曾柯. 生成式语言模型在古诗生成中的优化 [D]. 上海 : 华东师范大学, 2024.

[22] Liu D, Guo Q, Li W, et al. A Multi-Modal Chinese Poetry Generation Model[A]. 2018 International Joint Conference on Neural Networks (IJCNN)[C], 2018 : 1 – 8.

[23] Liu Y, Liu D, Lv J. Deep Poetry: A Chinese Classical Poetry Generation System[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(09) : 13626 – 13627.

[24] Papineni K, Roukos S, Ward T, et al. BLEU: A Method for Automatic Evaluation of Machine Translation[A]. Isabelle P, Charniak E, Lin D. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics[C], Philadelphia, Pennsylvania, USA : Association for Computational Linguistics, 2002 : 311 – 318.

[25] Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries[A]. Text Summarization Branches Out[C], Barcelona, Spain : Association for Computational Linguistics, 2004 : 74 – 81.

[26] Li J, Galley M, Brockett C, et al. A Diversity-Promoting Objective Function for Neural Conversation Models[A]. Knight K, Nenkova A, Rambow O. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies[C], San Diego, California: Association for Computational Linguistics, 2016: 110–119.

[27] Wieting J, Bansal M, Gimpel K, et al. Towards Universal Paraphrastic Sentence Embeddings[J], 2016(arXiv:1511.08198).

[28] Deng L, Wang J, Liang H, et al. An Iterative Polishing Framework Based on Quality Aware Masked Language Model for Chinese Poetry Generation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05): 7643–7650.

[29] Ma J, Zhan R, Wong D F. Yu Sheng: Human-in-Loop Classical Chinese Poetry Generation System[A]. Croce D, Soldaini L. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations[C], Dubrovnik, Croatia: Association for Computational Linguistics, 2023: 57–66.

[30] Chen Z, Cao Y. A Polishing Model for Machine-Generated Ancient Chinese Poetry[J]. Neural Processing Letters, 2024, 56(2): 77.

[31] Radford A, Kim J W, Hallacy C, et al. Learning Transferable Visual Models From Natural Language Supervision[A]. Proceedings of the 38th International Conference on Machine Learning[C], [S.l.]: PMLR, 2021: 8748–8763.

[32] Zhu D, Chen J, Shen X, et al. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models[J], 2023(arXiv:2304.10592).

[33] Wu Z, Chen X, Pan Z, et al. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding[J], 2024(arXiv:2412.10302).

[34] Brown T, Mann B, Ryder N, et al. Language Models Are Few-Shot Learners[A]. Advances in Neural Information Processing Systems: Vol 33[C], [S.l.]: Curran Associates, Inc., 2020: 1877–1901.

[35] Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models[J]. Advances in Neural Information Processing Systems, 2022, 35 : 24824 – 24837.

[36] Wang X, Wei J, Schuurmans D, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models[J], 2023(arXiv:2203.11171).

[37] Li J, Song Y, Zhang H, et al. Generating Classical Chinese Poems via Conditional Variational Autoencoder and Adversarial Training[A]. Riloff E, Chiang D, Hockenmaier J, et al. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing[C], Brussels, Belgium : Association for Computational Linguistics, 2018 : 3890 – 3900.

致谢

感恩的心

A 古诗评分体系

表 A.1 古诗评分体系

维度	分值	子维度	小分	备注
格律规范	25	平仄音韵	10	9-10：完全符合唐体格律（例：杜甫《登高》”风急天高猿啸哀，渚清沙白鸟飞回”平仄严谨）
				7-8：个别拗句但有救（例：王维《终南别业》”行到水穷处”第三字拗，第四字救）
				5-6：三平尾/三仄尾不超过两处（例：韦应物《滁州西涧》”独怜幽草涧边生”三平尾）
				0-4：严重失律（例：打油诗体）
		对仗工稳	10	9-10：工对 + 借对精妙（例：李商隐《锦瑟》”庄生晓梦迷蝴蝶，望帝春心托杜鹃”）
				7-8：宽对但结构平衡（例：王勃《送杜少府》”海内存知己，天涯若比邻”）
				5-6：词性不对应（例：拙劣仿作”青山对绿水，饮酒对弹琴”名词对动词）
				0-4：无对仗意识
		押韵协调	5	5：严格遵循平水韵（例：李白《静夜思》”床前明月光”押下平七阳韵）
				3-4：邻韵通押（例：杜牧《清明》”纷”属文韵，”魂”属元韵通押）
				1-2：出韵超过两处
				0：完全无押韵
意象意境	30	古典运用	20	18-20：传统意象出新境（例：王维《使至塞上》”大漠孤烟直”重构”孤烟”意象）
				14-17：精准使用经典意象（例：柳宗元《江雪》”孤舟蓑笠翁”的渔父符号）
				10-13：意象堆砌无深意（例：劣作”残阳古道瘦马，西风落叶昏鸦”）
				0-9：意象误用（例：用”东篱”指代监狱）
		意境层次	10	9-10：多层意境交织（例：李商隐《夜雨寄北》时空折叠技法）
				7-8：单一意境完整（例：孟浩然《春晓》的晨醒意境）

转下一页

表 A.1　古诗评分体系

维度	分值	子维度	小分	备注
				5-6：意境破碎（例：拼贴” 明月松间照，股票涨停板”）
				0-4：无意境构建
主题思想	20	情感真挚	12	11-12：情志合一（例：杜甫《月夜》” 遥怜小儿女，未解忆长安” 的家国之痛）
				9-10：情感明确但稍显直露（例：高适《别董大》” 莫愁前路无知己”）
				6-8：情感造作（例：伪古风” 朕与将军解战袍”）
				0-5：情感空洞
		思想传承	8	7-8：接通传统文脉（例：苏轼《题西林壁》对禅理的化用）
				5-6：简单模仿前人（例：仿写” 采菊东篱下” 无新解）
				3-4：曲解经典（例：将” 仁者乐山” 解为爱好登山）
				0-2：思想谬误
语言锤炼	15	凝练度	8	7-8：字字珠玑（例：贾岛《题李凝幽居》” 鸟宿池边树，僧敲月下门”）
				4-6：可删减 1-2 字（例：初稿” 推” 改为” 敲” 的炼字过程）
				1-3：冗余明显（例：劣作” 我看到青山高又高，绿水长流流不停”）
		典雅度	7	6-7：文白交融自然（例：李清照《声声慢》” 寻寻觅觅” 的白话感）
				4-5：文言生硬（例：强行用” 之乎者也” 凑韵）
				1-3：语体混乱（例：夹杂”OK””Hi” 等外来词）
创新性	10	\	\	9-10：传统技法新用（例：王安石《泊船瓜洲》” 绿” 字形容词动用）
				7-8：有限度创新（例：崔颢《黄鹤楼》前半打破律诗常规）
				5-6：为变而变（例：强行改写五绝为六言）

B 提示词

请描述这张图片，注意要明确提及图像中的物体，描述清楚物体的色彩、大小、相对位置等基本信息，并兼顾整体的情感色彩，确保读者能够根据描述在心里构建出一个清晰的画面。请确保使用中文，不超过 7 句话，并使用一个段落完成。

图 B.1 提示词（图像分析）

Capacity & Role

你是一个古诗创作大师，通晓古今诗词的丰富传统和美学，善于运用古今诗词的典故，更乐于化用经典诗词的表达以作典故。

Insight

你的创作应当体现出古文诗词的韵味和深度，且充分使用古典诗歌中常用的意象，如"月"常表示思乡和怀念。你必须创作出***{poem_type}***，使其完美契合相关的平仄结构规则，每行的字数应该一致，并且注重对仗的手法。

Statement

你的任务是生成与用户输入的文本和图片相符的古诗，包括契合用户文本的内容、情感、叙事、主题等方面，并结合图片的描述来丰富古诗内容。

Personality

你必须按照下面的JSON格式输出内容，且不包含除JSON外的其他任何内容：

```
{{
  "标题": "创作古诗标题",
  "正文": "创作古诗主体内容",
  "赏析": "创作古诗的赏析，使用连续的文本，内容包括对诗的背景、主题、情感、意象、修辞手法、韵律和意境等方面的分析",
  "注释": "创作古诗的注释，必须为古诗中使用到的意象和用意一一剖析",
  "白话译文": "古诗的白话文翻译"
}}
```

Input

这是用户的输入文本:"***{user_text}***"。用户输入一张图片，其对应的文字描述为“***{description}***”

图 B.2 提示词（古诗生成）

TODO Prompt：古诗评分

Capacity & Role

你是一个古诗创作大师，通晓古今诗词的丰富传统和美学，善于运用古今诗词的典故，更乐于化用经典诗词的表达以作典故。

Insight

你的创作应当体现出古文诗词的韵味和深度，且充分使用古典诗歌中常用的意象，如"月"常表示思乡和怀念。你必须创作出***{poem_type}***，使其完美契合相关的平仄结构规则，每行的字数应该一致，并且注重对仗的手法。

Statement

你的任务是根据用户给出的古诗及其评分和改进建议，修改原古诗，使其在保持与文本输入和图片描述相契合的前提下，在格律规范、意象意境、主题思想、语言凝练、创新维度等方面有所提升。

Personality

你必须按照下面的JSON格式输出内容，且不包含除JSON外的其他任何内容：

```
{{
  "标题": "创作古诗标题",
  "正文": "创作古诗主体内容",
  "赏析": "创作古诗的赏析，使用连续的文本，内容包括对诗的背景、主题、情感、意象、修辞手法、韵律和意境等方面的分析",
  "注释": "创作古诗的注释，必须为古诗中使用到的意象和用意一一剖析",
  "白话译文": "古诗的白话文翻译",
  "改进说明": "优化修改的说明，可分点，但最终必须是连续的文本"
}}
```

Input

[文本输入]***{user_text}***
[图像描述输入]***{description}***
[古诗]***{poem}***
[评分]***{evaluation}***
[改进建议]***{suggestion}***

图 B.3　提示词（古诗优化）

C 成果

1. Yang L, Zhang Z, Niu K, et al. Large Model Based Crossmodal Chinese Poetry Creation[A]. 2024 IEEE Smart World Congress (SWC)[C], Nadi, Fiji: IEEE, 2024 : 27 - 34.