



刘祥根, 四川大学计算机学院副研究员, 博士研究生导师, 四川省重要人才计划入选者, 国家重点研发计划课题负责人, 清华大学博士。主要从事人工智能基础理论及其应用研究。在 CCF-A 类及中科院 1 区的权威学术期刊发表高水平论文 32 篇。主持国家自然科学基金、国家重点研发计划(课题)、国家重大科技专项(子课题)、工信部工业互联网创新发展工程(课题)、四川省重大专项等十余项国家级、省部级重要科技攻关项目, 突破了工艺大模型、分子优化大模型等多项关键技术。

基于多模态信息融合的工艺生成式设计算法

侯俊利¹, 李佳渊², 唐飞熊¹, 李 阳¹, 田 军³, 李 玥³,
张入夕², 刘祥根², 史建成¹, 吕建成²

(1. 西南电子设备研究所, 成都 610036;

2. 四川大学计算机学院, 成都 610065; 3. 东方电气(成都)创新研究有限公司, 成都 610213)

摘 要: 面向智能制造背景下工艺设计自动化的迫切需求, 本文指出当前研究方法存在明显局限性。现有基于视觉的工艺生成模型则过度聚焦于图像特征提取, 普遍忽视图纸中的关键文本标注(如尺寸公差、材料规格), 导致多模态信息协同不足。这些不足的深层次问题在于, 现有方法缺乏对工程图纸多模态特性(视觉-文本关联性)的有效建模, 且传统融合策略难以实现跨模态的渐进式对齐, 从而制约了工艺描述生成的准确性与逻辑一致性。针对上述不足, 本研究提出一种基于多模态信息融合的工艺生成式设计算法。该算法通过引入(Optical Character Recognition, OCR)文本特征与视觉特征的深度融合, 有效解决了工艺图纸中技术参数提取不完整、语义表达不准确等关键问题。本算法在双语评估替补指标(Bilingual Evaluation Understudy, BLEU)、面向召回率的摘要评估指标(Recall-Oriented Understudy for Gisting Evaluation, ROUGE)等指标上均优于传统视觉模型, 且未明显增加计算开销, 其为工艺设计智能化转型提供了可靠的技术支撑, 对推动制造业数字化转型具有重要实践意义。

关键词: 工艺描述生成; OCR; 工艺设计; 多模态融合; 门控注意力机制

中图分类号: TP391.4; TH166 **文献标志码:** A **DOI:** 10.19907/j.0490-6756.250348

A multimodal information fusion-based algorithm for generative process design

HOU Junli¹, LI Jiaquan², TANG Feixiong¹, LI Yang¹, TIAN Jun³, LI Yue³,
ZHANG Ruxi², LIU Xianggen², SHI Jiancheng¹, LÜ Jiancheng²

(1. Southwest China Research Institute of Electronic Equipment, Chengdu 610036, China;

2. School of Computer Science, Sichuan University, Chengdu 610065, China;

3. Dongfang Electric (Chengdu) Innovation Research Co., Ltd., Chengdu 610213, China)

Abstract: Facing the urgent demand for process design automation in the context of intelligent manufactur-

收稿日期: 2025-12-09 修回日期: 2026-01-26 接受日期: 2026-02-04

基金项目: 四川省宽频带电子产品智能制造工程技术研究中心基金(BEIM2023B001); 四川省自然科学基金(2024NSFTD0048); 四川省重大专项(2024ZDZX0003)

作者简介: 侯俊利(1970—), 男, 河南鹤壁人, 正高级工程师, 硕士, 主要研究方向为智能制造. E-mail: moonfangluna@126.com

通信作者: 刘祥根. E-mail: liuxianggen@scu.edu.cn

ing, this study identifies significant limitations in current research approaches. Existing vision-based process generation models tend to overemphasize image feature extraction while generally neglecting critical textual annotations in drawings (such as dimensional tolerances and material specifications), leading to insufficient multimodal information coordination. The underlying issues stem from the inadequate modeling of the multimodal characteristics (visual-textual correlations) inherent in engineering drawings by existing methods, and the inability of traditional fusion strategies to achieve progressive cross-modal alignment, thereby constraining the accuracy and logical consistency of generated process descriptions. To address these shortcomings, this study proposes a generative process design algorithm based on multimodal information fusion. By deeply integrating OCR-extracted textual features with visual features, the algorithm effectively addresses key challenges such as incomplete extraction of technical parameters and inaccurate semantic representation from process drawings. Our algorithm demonstrates superior performance over traditional vision-only models on metrics including BLEU and ROUGE, without significantly increasing computational overhead. It provides reliable technical support for the intelligent transformation of process design and holds significant practical value for advancing the digital transformation of the manufacturing industry.

Keywords: process description generation; OCR; process design; multimodal fusion; gated attention mechanism

近年来,多模态学习在计算机视觉和自然语言处理领域的融合应用取得了显著进展。从计算机视觉的角度来看,图像中的内容极具价值^[1]。以图像描述生成为代表的多模态任务,通过深度神经网络实现了视觉信息与语言描述的有效关联,成为连接图像理解与文本生成的重要技术桥梁。特别是在 Encoder-Decoder 架构的推动下,基于注意力机制的图像描述模型在通用场景中展现出令人瞩目的性能。然而,当这些通用模型迁移到工艺设计等专业领域时,其适用性面临严峻挑战。在工艺设计这一特殊应用场景中,工程图纸不仅仅是简单的二维形状集合,更是承载着丰富设计意图和技术规范的专业文档^[2],包含着大量丰富的文字标注与精确参数^[3]。然而,当前主流的计算机辅助工艺设计(Computer Aided Process Planning, CAPP)系统或基于图像的工艺生成方法在应对工程图纸的多模态特性时表现乏力,严重制约了智能化转型的进程。具体而言,现有 CAPP 系统多基于规则或模板,虽能处理结构化数据,但难以自动解析非结构化的工程图纸,尤其缺乏对图纸中视觉与文本信息的协同利用。工程图纸包含大量文本标注(如尺寸公差、材料规格和工艺要求等关键参数)。当前基于图像的工艺生成方法(如通用图像描述模型)主要聚焦于视觉特征的提取,往往忽略文本标注信息,导致生成的工艺描述丢失关键参数。例如,在复杂零件图纸中,尺寸公差(如“ $\varnothing 25 \pm 0.1$ ”)或材料要求(如“5A06-H112”)若未被有效提取,会引发工艺规划错误(如加工余量不足或

检测遗漏)。研究表明,单纯依赖视觉特征仅能捕捉图纸的几何轮廓,无法理解技术规范,亟需开发多模态融合方法以弥补这一语义鸿沟。

当前,基于 Encoder-Decoder 架构的通用图像描述模型在工艺设计任务中主要存在三方面局限:首先,多模态信息融合机制不完善,现有模型大多忽视 Optical Character Recognition 文本的集成^[4],无法有效协调图像特征与图纸标注信息;其次,长文本生成过程中的语义一致性难以保证^[5],关键工艺参数在生成序列时易出现遗漏或失真;此外,模型计算复杂度高^[6],难以满足实际工业应用对效率的严格要求。

针对上述挑战,本文提出了一种创新的工艺设计模型构建方法,该方法深度融合了二维图纸的视觉特征与 OCR 提取的文本信息。本工作的主要贡献包括 3 个方面:1) 设计了一种层次化的多模态融合架构,通过双阶段融合策略实现了图像特征与文本特征的有效协同;2) 构建了面向工艺设计领域的专业数据集,为该领域的研究提供了基准支持;3) 通过系统的实验验证,证明了所提方法在真实工业场景中的有效性、实用性及优越性。实验结果表明,相较于传统方法,本方法在保持较高推理效率的同时,提升了工艺描述的准确性和完整性。

1 相关工作

1.1 图像描述生成技术

图像描述生成技术的目标是构建视觉与文本

的精准映射,自动将图像内容转化为逻辑连贯、信息完整的自然语言描述,其技术演进始终围绕“特征提取-序列生成”的核心逻辑展开。早期研究以 CNN-RNN^[7] 架构为主流,利用卷积神经网络(Convolutional Neural Network, CNN)的视觉特征提取能力与循环神经网络(Recurrent Neural Network, RNN)^[8] 的序列建模优势:编码器选用 VGG^[9]、ResNet^[10] 等经典 CNN 模型,通过多轮卷积与池化操作,将原始图像转化为包含局部细节与全局语义的特征表示(如 ResNet-101 可输出 $14 \times 14 \times 2048$ 维度的图像编码张量);解码器则以长短期记忆网络(Long Short-Term Memory, LSTM)^[11] 为核心,有效缓解传统 RNN 的梯度消失问题,实现特征向量到文本序列的逐词解码。在应对长文本建模问题中,注意力机制发挥关键作用^[12]。随着自注意力机制的提出 Transformer^[13] 架构的普及,图像描述任务迎来重要突破——自注意力机制能够动态建模序列内部及图像区域间的依赖关系,显著提升长文本生成的语义一致性与逻辑连贯性,成为当前主流技术方向之一。

近年来,该领域持续向更高效的架构与更统一的范式演进。Li 等^[14] 通过引入冻结的预训练图像编码器与大语言模型进行高效协同,提升了模型的通用视觉语言理解与生成能力;Alayrac 等^[15] 通过引入门控交叉注意力高效融合大规模预训练模型,使模型具备出色的少样本视觉语言理解和生成能力;Lu 等^[16] 构建了能够统一处理多种模态的规模化自回归生成模型(Unified-IO2);Li 等^[17] 的团队进一步提出了基于冻结图像编码器和大型语言模型的高效视觉语言预训练框架(BLIP-2);Dai 等^[18] 通过指令调优机制增强模型的泛化能力,构建了 InstructBLIP 模型。这些架构在通用场景下取得了显著成功。

然而,这些架构仍存在明显的局限性:CNN 特征提取器对图像中的文本信息不敏感^[10],难以捕捉工程图纸中的关键参数标注;而 LSTM 解码器的序列建模能力有限,在生成长文本描述时容易出现语义不一致问题^[19];Transformer 的自注意力计算复杂度随序列长度呈平方级增长^[13],当处理包含大量文本的多模态输入时计算开销巨大。

1.2 OCR 技术在工业中的应用

OCR 技术是从图像中精准提取文本信息的核心技术,在工业领域中,其广泛应用于工程图纸数

字化、关键工艺参数提取(如尺寸标注、材料型号、公差要求)、生产文档智能解析等核心任务,为工业信息化与智能化转型提供基础数据支撑。近年来,文档智能领域取得显著进展,Ke 等^[20] 系统综述了文档图像分析的技术挑战,指出工程图纸理解需要结合视觉与文本的跨模态分析。传统 OCR 方法以 Tesseract 为代表,通过模板匹配与手工设计特征实现文本识别,Kirana 等^[21] 的比较研究证明了 Tesseract OCR 在规范场景下的实用性;而基于深度学习的 CRNN^[22] 模型则通过 CNN 与 BiLSTM^[23] 的融合,构建端到端识别框架,有效提升了复杂场景下的识别精度。当前主流方案(如 PaddleOCR)已能将工业图纸文本识别准确率提升至 98%^[24-25] 以上, Lu 等^[26] 通过 PaddleOCR 实现的书名识别工作展现了其优越性, Sarkar 等^[27] 在自动车牌字符识别中的应用进一步验证了其泛化能力。OCR 与图像描述模型的融合,本质上是实现“文本参数模态”与“视觉特征模态”的跨模态协同,通过将 OCR 提取的结构化文本与图像描述模型捕捉的视觉语义关联,为生成精准、详实的工业场景描述(如工艺设计方案)提供双重支撑。

OCR 与图像描述模型融合的核心优势在于能够弥补单一模态的信息短板:OCR 提供的精确文本参数(如“表面粗糙度 Ra6.3”)可强化描述的量化准确性,避免图像描述仅停留在定性层面;而图像描述模型的视觉理解能力则能为 OCR 文本赋予空间与语义上下文(如明确参数对应的零件区域),提升信息的关联性。

1.3 多模态融合方法

多模态融合是指将视觉、文本等不同来源的模态信息进行结构化整合,通过挖掘跨模态数据的内在语义关联,实现单一模态无法达成的精准信息建模,在工业智能、人机交互等领域具有核心应用价值。根据融合阶段的不同,现有方法主要分为早期融合、晚期融合和混合融合等 3 种范式。

早期融合方法通过在输入层或特征层直接拼接不同模态的特征,实现信息的初步整合。Baltrusaitis 等^[28] 提出的特征拼接方法计算效率高,易于实现,但在处理视觉与文本模态时,由于缺乏细粒度的交互机制,难以建立深层次的语义关联。Arevalo 等^[29] 在此基础上引入门控机制,通过可学习的权重调整各模态的贡献度,然而该方法仍然无法有效解决特征空间不一致性问题;晚期融合方法

则在各模态独立完成推理后,在决策层进行结果集成。Karpathy等^[30]通过加权平均策略融合视觉和文本分类结果,虽然保留了各模态的判别特性,但忽略了模态间的互补信息,在复杂任务中性能受限;基于注意力机制的混合融合方法近年来逐渐成为研究热点,Nagrani等^[31]提出的多模态注意力瓶颈架构,通过交叉注意力机制实现跨模态信息交互,提升了融合效果。然而,该方法计算复杂度较高,在处理长序列文本时面临效率挑战。

在工艺设计领域,多模态融合面临特殊挑战:

1) 模态对齐粒度不足:工程图纸中的文本标注与视觉元素具有精确的对应关系,要求融合方法能够建立细粒度的跨模态对齐^[32];2) 动态融合能力缺失:工艺描述生成需要同时保持技术参数的准确性和语义的连贯性,对融合策略的精确性提出更高要求^[29];3) 计算效率瓶颈:工业应用场景对计算效率的严格要求,也限制了复杂融合方法的实际部署^[33]。

为提升多模态融合的精细度,研究者们提出了多种创新架构。Liu等^[34]设计了一种基于自适应注意力的跨模态表示融合框架(ARIF),通过动态权重调节实现更灵活的特征整合。在工业应用层面,Mehta^[35]等开发了集成多模态Transformer的预测性行动指导系统,以优化复杂工业任务流程。此外,面向特定工业领域的多模态基础模型构建也受到关注,例如Ren等^[36]讨论了流程工业基础模型所面临的机遇,并强调了多模态数据与领域机理融合的重要性。这些工作表明,面向专业领域的、轻量且高效的自适应融合策略是当前的研究趋势之一,这与本研究聚焦于工艺图纸、采用双阶段门控融合以平衡精度与效率的思路相一致。

2 工艺设计模型构建方法

2.1 总体架构设计

本研究提出的模型总体架构核心为“Encoder-Decoder+Attention”的基础框架^[4],并创新性地融入OCR文本处理与动态融合机制,具体包括四大核心模块:

1) 图像编码器:采用预训练ResNet-101模型作为视觉特征提取骨干网络,通过移除顶层全连接层并添加自适应池化层,将输入图像编码为 $14 \times 14 \times 2048$ 维度的空间特征图,充分保留图像

的区域语义信息,为后续注意力机制提供细粒度视觉特征表示。

2) OCR文本处理器:新增的多模态输入模块,基于PaddleOCR工具从工程图纸中提取尺寸参数、材料型号等关键文本信息,经过分词处理后通过词嵌入层转换为稠密向量,再经由双向LSTM编码器生成2048维语义特征向量,实现文本特征与图像特征的维度对齐,为多模态融合奠定基础。

3) 多模态融合与注意力机制:采用双阶段融合策略实现视觉与文本特征的深度整合。在初始化阶段,通过特征拼接将图像全局平均特征(2048维)与OCR文本特征(2048维)组合为4096维融合向量,作为解码器初始状态;在解码过程中,引入门控融合机制,动态调节OCR特征对图像特征的贡献权重,同时运用注意力机制对图像空间特征进行动态聚焦,使模型能够自适应地平衡多模态信息流。门控机制的实现通过门控向量计算OCR特征的贡献度,具体实现见第3.5节融合策略消融实验。

4) 描述解码器:基于“LSTM+Attention”架构,将门控融合后的多模态特征与当前词嵌入向量拼接输入LSTM单元,通过逐时间步解码生成完整工艺描述,并采用束搜索策略优化序列生成质量,确保输出文本的流畅性与准确性。

该架构设计的核心优势在于“渐进融合、动态调整”——通过双阶段融合机制与门控权重调节以解决工艺图纸理解中的多模态表征学习挑战。从多模态表征学习理论来看,双阶段融合的设计灵感来源于跨模态对齐中的层次化交互机制。早期融合(如特征拼接)虽计算高效,但忽略了模态间的特征空间差异,容易导致语义混淆;晚期融合(如决策层加权)虽保留模态独立性,但缺乏细粒度交互,难以捕捉图纸中文本与视觉元素的精确对应关系。

为此,本模型采用初始化融合与解码阶段动态融合相结合的双阶段策略,通过初始化阶段和解码阶段的分离处理,实现了跨模态表征的渐进式对齐:在初始化阶段,将图像全局平均特征与OCR文本特征拼接,为解码器提供丰富的多模态初始状态,这类似于多模态学习中的特征级联策略,但通过维度对齐($2048 \text{ 维} + 2048 \text{ 维} = 4096 \text{ 维}$)避免了特征空间不匹配问题;在解码阶段,门控机

制动态调节 OCR 特征权重,其理论根基源于门控循环单元(Gated Recurrent Unit, GRU)中的信息流控制原理,通过可学习的门控向量自适应地平衡视觉与文本信息流,使模型能够根据当前解码上下文优先关注关键模态。这种设计优于简单融合的原因在于:工艺图纸中的文本标注(如尺寸公差)与视觉元素(如几何轮廓)具有强空间关联性,门控机制能够通过注意力权重动态聚焦于相关区

域(如螺纹标注对应的轴结构),而双阶段策略确保了模态交互的时序一致性,避免了长序列生成中的语义漂移问题,在保持原框架稳定性的基础上实现多模态信息的智能整合,既确保了模型迁移的便捷性,又通过精细化的融合策略提升了工艺参数描述的准确性与语义连贯性。模型架构图如图 1 所示。

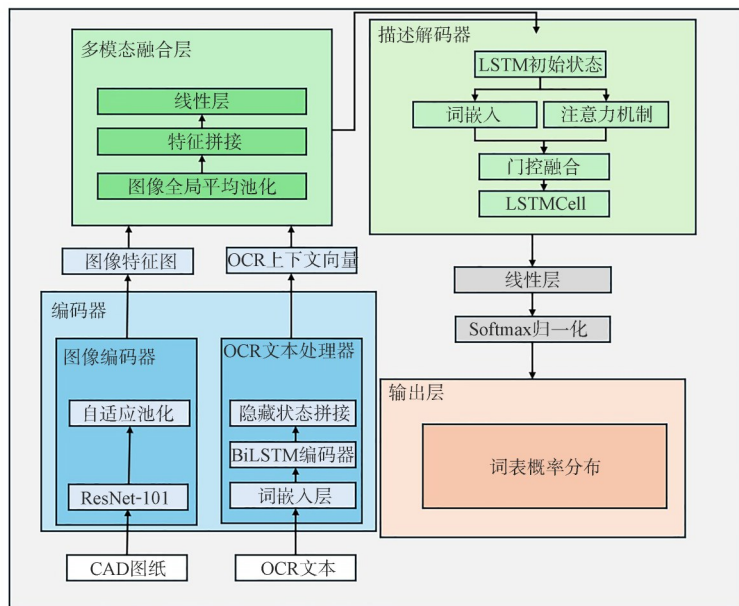


图 1 模型架构图

Fig. 1 Model architecture

2.2 OCR 文本集成策略

OCR 文本集成的核心目标是从工程图纸中精准提取结构化文本信息,如尺寸标注“ $\delta 30 \times 167 \times 168$ ”、材料说明“5A06-H112”、工艺参数“表面粗糙度 Ra6.3”等,并通过多级过滤机制确保文本质量,为后续多模态融合提供可靠输入。本研究采用 PaddleOCR 作为基础提取工具,其可微分二值化检测算法(Differentiable Binarization, DB)和卷积循环神经网络(Convolutional Recurrent Neural Network, CRNN)识别模型能高效定位和识别文本,基础准确率达 98%。经过 PaddleOCR 提取的 OCR 文本,会先经过层次化的 OCR 文本过滤处理,再输入模型中。该文本处理流程基于多模态表征学习中的跨模态对齐理论,旨在通过协同的过滤机制防止噪声干扰,同时保留高信息量的工程参数。零件图纸中不可避免会出现单个字符的噪声:比如对于每张零件图纸, CAD 图纸边框上会含有 A、B、C、D、E、F 及 1、2、3、4、5、6 等字符以进

行图幅分区,便于查找和定位,但是对工艺流程文本生成无太大帮助,反而这些字符在经过 OCR 提取后会变成模型难以理解的噪声,因此在对 OCR 文本进行基础清洗,剔除空 token 和冗余字符后,关键还需要对不同语言设置长度阈值:中文文本要求至少 2 个字符,英文文本需达到 4 个字符以上,数字组合至少需要 3 个字符,以避免单字与短词的干扰。

对于中文字符,在工艺制造领域,一些工艺术语十分简短,因此需要设置技术术语优先保留机制,建立包含材料术语(如“钢”、“铝”)、工艺术语(如“轴”、“铰”)的专业词典,保留这些术语以维护工程上下文完整性,避免因长度过滤导致语义丢失;而对于数字字符,需要通过规则模式识别有意义的规格参数,防止其被数字过滤规则误判。具体的方法是采用正则表达式和工程符号库(如“X”,“ $\pm$ ”等),识别尺寸标注与公差序列等结构化文本。

在进行完上述过滤策略后,还需要应用字符级分词策略将过滤后文本拆分为独立 token,每个字符作为基本单元,以确保所有字符都能被完整收录到词表中。后续通过词嵌入层将每个字符转换为 512 维向量,供多模态融合使用。该过滤机制在提升输入文本质量的同时,通过字符级分词构建了完备的工程字符词表,为 OCR 文本与图像特征拼接融合奠定基础,确保模型在计算高效的前提下实现可靠的多模态集成。

具体的 OCR 文本过滤示例如图 2 所示,在该案例中,原始 OCR 识别结果中包含了图中橙色荧光笔高亮出的“A”“B”“C”“2”“3”等无关字符。这

些字符实际上来源于 CAD 工程图纸的标准图幅边框——其设计用途是便于工程师在图纸上对位置进行网格化定位和索引。然而,对于“工艺描述生成”这一具体任务而言,这些纯定位功能的边框字符不承载任何工艺语义信息,若不经处理直接输入模型,反而会成为干扰模型理解核心工艺参数的噪声。因此,本研究设计了过滤算法来清洗此类噪声。经过过滤后,仅保留与零件特征、尺寸、公差、技术要求等相关的有效文本(如图 2 右下角所示)。这样的处理确保了输入模型的多模态信息更加纯净和聚焦,从而提升了后续融合与生成步骤的准确性和效率。

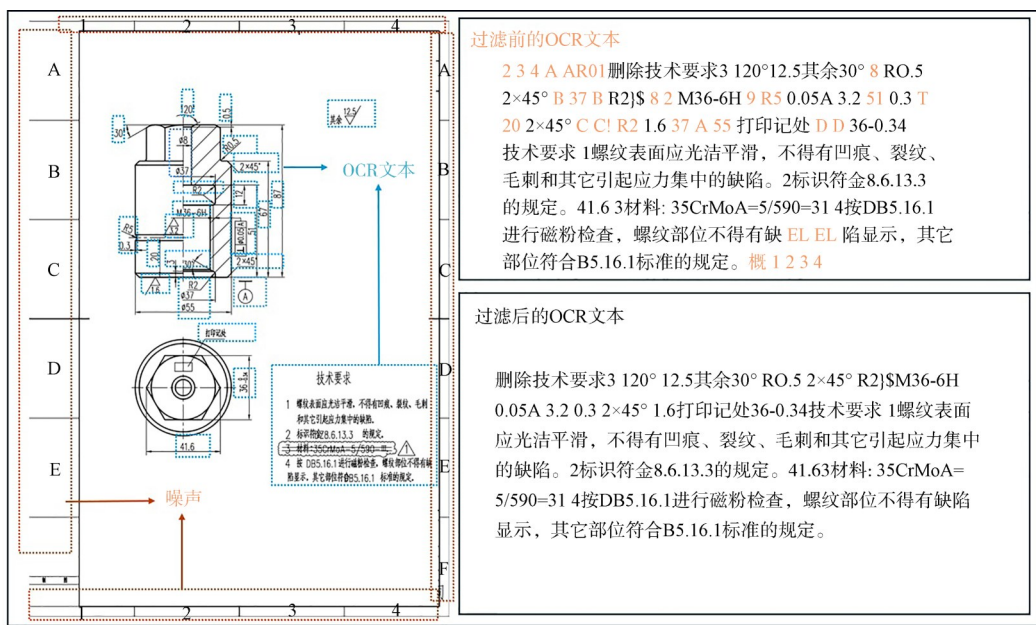


图2 OCR文本过滤示例

Fig. 2 Example of OCR text filtering

2.3 训练与优化策略

在训练过程中,采用了教师强制(Teacher Forcing)策略,即在解码器训练时使用真实的前一个词作为输入,而不是模型自身的预测输出,这有助于加速收敛、提高训练稳定性,并减少误差累积,确保多模态特征(图像与OCR文本)的有效融合。通过这些训练与优化策略的综合应用,模型能够在保证训练效果的同时,提高泛化能力和稳定性,更好地适应实际工艺设计任务的需求。

在优化器的选择上,采用Adam优化器^[37],其结合了Adagrad和RMSProp的优点,能够自适应地调整学习率,对于不同的参数采用不同的学习率更新策略。它在训练过程中能够快速收敛,并且对梯度的噪声具有较好的鲁棒性,能够有效地

更新模型的参数,使模型在训练过程中不断朝着损失函数最小化的方向优化。Adam优化器为解码器和编码器(当启用微调时)分别设置独立的学习率。学习率调整采用基于验证集性能的启发式调度:若验证集BLEU-4分数连续多个训练周期未提升,则对学习率进行衰减,以便模型在性能平台期进行更精细的参数调整。训练中引入了梯度裁剪机制,以抑制梯度爆炸,确保训练过程的数值稳定性。为了减少过拟合现象,提高模型的泛化能力,引入了dropout^[38]和权重衰减技术。dropout通过在训练过程中随机丢弃一部分神经元,使得模型在每次训练时看到的网络结构略有不同,从而减少神经元之间的协同适应,避免模型过度依赖某些特定的特征,增强模型的泛化能力。权重衰

减则是在损失函数中添加一个正则化项,对模型的权重进行约束,防止权重过大,使模型更加简单和泛化。详细的超参数设置(如学习率、批量大小、衰减周期等)见第 3.1 节实验设置。

3 实验结果及分析

3.1 数据集和实验设置

现有的工艺数据集规模都相对较小,包含的制造工序有限,且均未对外公开。本研究构建了一个新的大规模的工业领域的数据集。数据集包含 2502 张零件图纸和对应的生产工艺流程文本。这些数据由工厂直接提供,并且由资深的工艺师为每一张零件图纸进行标注,生成对应的工艺流

程。这些数据被分为训练集、验证集和测试集,其比例分别为 0.8、0.1 和 0.1。

数据示例如图 3 所示,具体而言,图纸中包含该零件的主视图与剖面图,并详细标明了尺寸等具体参数;而在工艺流程文本中,有着制备该零件的完整工序。

实验环境基于 PyTorch 2.9.0 和 CUDA 12.8 的开发环境下完成了模型结构的搭建与训练过程。具体的硬件环境与训练参数配置如表 1 所示。训练过程中采用 Adam 自适应矩估计优化器(Adaptive Moment Estimation, Adam),同时引入学习率调度及正则化方法,以提高模型稳定性、防止过拟合并增强泛化能力。

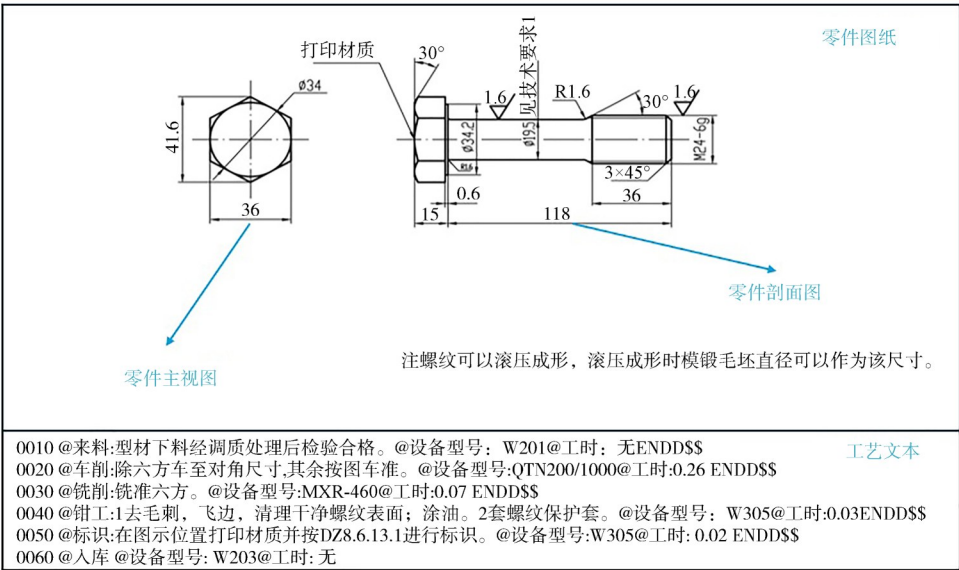


图 3 数据集示例图
Fig. 3 Dataset samples

表 1 实验参数设置	
Tab. 1 Experimental parameter setting	
参数	数值
CPU	AMD EPYC 7542 32-Core Processor
GPU	1 张 NVIDIA GeForce RTX 3090 GPU
训练周期 (Epochs)	65
批次大小 (Batch Size)	8
优化器	Adam
解码器学习率	4E-4
编码器学习率	1E-4
dropout	0.5
学习率衰减策略	验证集 BLEU-4 连续 8 个周期未提升时, 乘以 0.8 衰减
梯度裁剪阈值	5.0

3.2 评估指标

为全面评估模型在工艺流程生成任务中的性能表现,本研究构建了多维度的评估体系,从工序序列和工序文本评估指标方面考察模型的生成质量,具体如下:

1) 完全匹配率(Exact Match Rate)衡量生成序列与参考序列完全一致的比例,反映模型精确复现工艺序列的能力:

$$EM = \frac{\sum_{i=1}^N I(y_i = \hat{y}_i)}{N}$$

(1)

其中, N 为样本总数, y_i 为参考序列, $\hat{y}_i$ 为生成序列, I 为指示函数(序列完全一致时取 1, 否则取 0)。

2) BLEU(Bilingual Evaluation Understudy)^[39]

通过计算 n-gram 精度评估生成文本与参考文本的表面相似性,采用经典的四元组(BLEU-4)计算:

$$\text{BLEU}_{\text{seq}} = BP \cdot \exp\left(\sum_{n=1}^4 \frac{1}{4} \log p_n\right)$$

(2)

其中, p_n 为 n-gram 精度, BP 为简短惩罚因子 $BP = \min(1, \frac{\hat{y}_i}{y_i})$ 。

3) 精确率(Precision)与召回率(Recall)基于最长公共子序列(LCS)计算,从信息检索角度评估生成内容的准确性和覆盖度:

$$P_{\text{seq}} = \frac{\text{LCS}(y, \hat{y})}{|\hat{y}|}, \quad R_{\text{seq}} = \frac{\text{LCS}(y, \hat{y})}{|y|}$$

(3)

其中 $\text{LCS}(y, \hat{y})$ 为参考序列 y 与生成序列 $\hat{y}$ 的最长公共子序列长度。

4) 工序文本 BLEU:

$$\text{BLEU}_{\text{char}} = BP \cdot \exp\left(\sum_{n=1}^4 \frac{1}{4} \log p_n^{(\text{char})}\right)$$

(4)

其中, p_n 为 n-gram 精度, BP 为简短惩罚因子 $BP = \min(1, \frac{\hat{y}_i}{y_i})$ 。

5) ROUGE (Recall-Oriented Understudy for

Gisting Evaluation)^[40] 系列指标: ROUGE 与 BLEU 计算思路相近,均基于 n-gram 匹配原理,但 ROUGE 更侧重于召回率评估。本实验采用 ROUGE-1、ROUGE-2 和 ROUGE-L 等 3 个核心指标,分别从一元语法匹配、二元语法匹配和最长公共子序列三个维度,通过 F1 分数平衡精确率和召回率,全面评估生成文本的语义一致性和流畅度。

3.3 对比实验

为验证多模态融合策略在工艺流程生成任务中的实际效果,本研究构建了系统的对比实验框架。基于相同的训练数据集,相同的教师强制训练策略,将提出的融合 OCR 文本信息的 MultiMP 模型与仅依赖视觉特征的 Attention 基线模型、基于大规模预训练的 Qwen2. 5-VL-3B-Instruct-bnb-4bit 模型和 Deepseek-v1-1. 3B-chat 模型进行并行训练与测试。实验设计涵盖了从短期到完整的多个训练周期,确保对比结果的全面性和可靠性。通过在不同训练阶段对四类模型进行横向性能比较,重点考察多模态信息融合对生成质量的影响,从而为模型改进提供实证依据。具体实验结果如表 2 所示。

表 2 模型在工艺设计场景下的生成能力对比 (Epochs)
Tab. 2 Comparison of model generation capabilities in process design scenarios (Epochs)

模型(训练周期)	工序序列				工序文本			
	匹配率/%	BLEU	精确率	召回率	BLEU	ROUGE		
						Rouge-1	Rouge-2	Rouge-L
Attention (15)	3. 98	0. 402 6	0. 724 7	0. 723 5	0. 552 6	0. 569 5	0. 424 6	0. 562 3
Qwen (15)	5. 98	0. 345 3	0. 789 3	0. 657 9	0. 472 6	0. 450 5	0. 295 3	0. 443 2
Deepseek (15)	4. 38	0. 430 9	0. 731 4	0. 770 4	0. 608 0	0. 508 2	0. 358 7	0. 499 7
MultiMP (15)	9. 56	0. 479 3	0. 746 6	0. 758 4	0. 600 7	0. 616 4	0. 477 4	0. 611 1
Attention (25)	9. 96	0. 485 4	0. 743 0	0. 786 9	0. 614 1	0. 607 2	0. 470 3	0. 600 0
Qwen (25)	5. 98	0. 432 5	0. 787 8	0. 785 3	0. 553 6	0. 510 2	0. 359 1	0. 503 8
Deepseek (25)	5. 58	0. 486 4	0. 778 5	0. 812 5	0. 638 6	0. 537 8	0. 378 4	0. 529 1
MultiMP (25)	19. 52	0. 569 0	0. 799 0	0. 793 2	0. 662 5	0. 648 7	0. 520 5	0. 640 7
Attention (55)	20. 32	0. 562 0	0. 792 2	0. 823 7	0. 696 0	0. 654 7	0. 529 9	0. 645 3
Qwen (55)	7. 97	0. 465 4	0. 801 8	0. 756 8	0. 595 7	0. 549 0	0. 402 8	0. 540 8
Deepseek (55)	24. 30	0. 582 0	0. 811 1	0. 858 0	0. 713 2	0. 622 7	0. 490 1	0. 614 9
MultiMP (55)	24. 30	0. 606 4	0. 827 4	0. 816 1	0. 702 3	0. 681 8	0. 563 2	0. 674 9
Attention (65)	20. 72	0. 569 5	0. 796 9	0. 831 7	0. 702 3	0. 663 1	0. 539 5	0. 651 3
Qwen (65)	7. 97	0. 464 3	0. 813 4	0. 753 8	0. 581 1	0. 540 6	0. 393 8	0. 533 6
Deepseek (65)	22. 71	0. 570 3	0. 819 6	0. 851 7	0. 714 9	0. 628 8	0. 497 9	0. 620 7
MultiMP (65)	24. 70	0. 604 1	0. 819 0	0. 833 7	0. 724 1	0. 691 8	0. 571 1	0. 683 1

注:每一训练周期下,各评价指标的最优结果以加粗标出。
Note: For each training epoch, the best result of each evaluation metric is shown in bold.

从评估结果可以观察到,融合 OCR 文本信息的 MultiMP 模型在所有训练阶段均展现出显著优势。在工序序列完全匹配率方面,MultiMP 模型从训练初期就表现突出,在 15 个周期时达到 9.56%,显著高于 Attention 基线的 3.98%、Qwen 的 5.98% 以及 Deepseek 的 4.38%。

随着训练周期的增加,这一优势持续扩大,在完整的 65 轮训练后达到 24.70%,优于 Attention 基线的 20.72%、Qwen 的 7.97% 和 Deepseek 的 22.71%。

在自动评估指标方面,MultiMP 模型同样表现优异。BLEU-4 分数在各训练阶段均优于对比模型,在 25 个周期时达到 0.569 0,相比 Attention 基线的 0.485 4 提升 17.2%,同时显著高于 Qwen 的 0.432 5 和 Deepseek 的 0.486 4。这一结果凸显了多模态融合在提升生成文本与参考文本表面相似性方面的显著效果。

ROUGE 系列指标的对比结果进一步验证了多模态融合的有效性。在 65 个周期时,MultiMP 模型的 ROUGE-L 得分达到 0.683 1,显著高于 Attention 基线的 0.651 3、Qwen 的 0.533 6 和 Deepseek 的 0.620 7。这一结果表明,融合 OCR 信息有助于提升生成内容的准确性,并在一定程度上改善了文本的连贯性和语义一致性。

值得注意的是,Deepseek 以及 Qwen 模型在某些指标上表现接近 MultiMP 模型,这反映了大规

模预训练模型在语言理解方面的优势。然而,MultiMP 模型在关键指标如完全匹配率和 ROUGE 得分上仍保持领先,证明了针对工艺设计任务专门设计的多模态融合策略的有效性。

3.4 可解释性消融实验

图 4 展示了本文提出的面向工程图纸的 MultiMP 模型在生成“车”“削”二字时的注意力热力分布图。通过可视化模型决策时的特征聚焦区域,可直观验证其注意力机制的有效性与推理逻辑的合理性。

在“车”、“削”工艺对应的热力图中,模型的注意力(以红色高亮区域为标识)呈现高度集中的局部聚焦模式:其激活区域精准覆盖了零件右侧的多级阶梯轴结构。从制造工艺的专业逻辑来看,“车削”的核心加工对象是回转体特征(如轴、套、圆柱面),这类特征的工程表达通常以“同心圆轮廓+直径尺寸标注”为典型形式。模型的注意力分布与这一专业认知高度吻合,其并未随机激活图纸的无关区域(如标题栏、边框),而是主动锁定了阶梯轴的几何轮廓与尺寸标注的关联区域。这一现象表明,模型的决策并非基于“表面特征匹配”的机械记忆,而是构建了“几何特征→工艺类型”的因果推理链路:当输入包含“连续同心圆+多组直径尺寸”的特征时,模型会优先激活与“车削适配性”相关的特征权重,进而输出“车”的工艺判断。

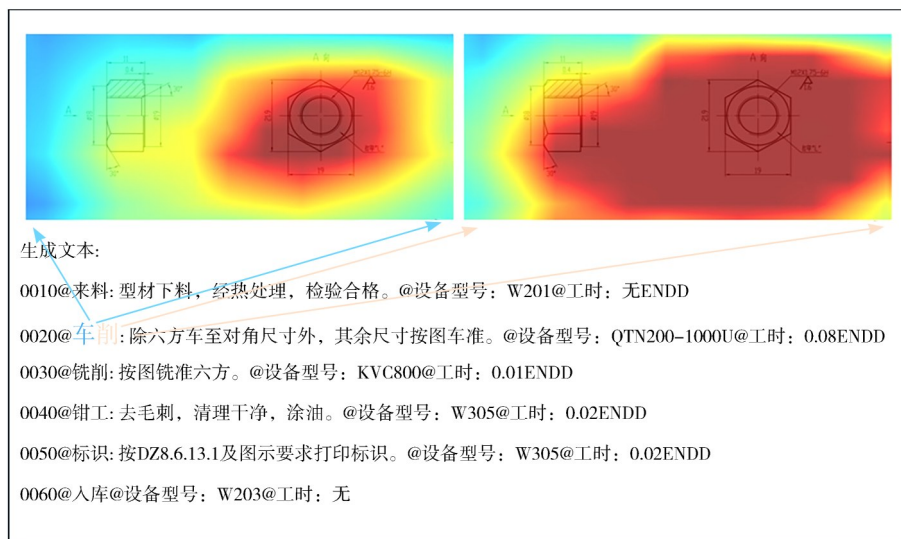


图 4 注意力热力图

Fig. 4 Attention heatmap

从“车”的集中聚焦到“削”的轻微扩散,注意力模式的演变印证了模型对工艺指令生成过程的

合理性:模型在完成图纸中右侧螺母基础轮廓识别后,进一步通过左侧螺母内部剖面图验证内部

特征的加工要求。这种注意力流动与制造实践中“先粗看外形,再精读细节”的图纸解读习惯吻合,说明模型能够在一定程度上模拟工程师阅读图纸时的认知流程。通过量化分析可发现,扩散至剖面图的注意力权重均值低于主视图聚焦区,该梯度分布正好反映了加工优先级差异,即外部轮廓为车削核心,内部结构为辅助特征。

3.5 融合策略消融实验

为验证门控机制在多模态融合中的有效性,本节设计了消融实验,对比分析简单拼接融合(Simple)与门控特征融合(MultiMP)两种策略在不同训练周期下的性能表现。实验旨在深入探究门控机制对工艺描述生成质量的影响。本消融实验采用控制变量法,保持模型主体结构不变,仅调整多模态特征融合策略:

1) 简单拼接融合(Simple):将 OCR 文本特征与图像视觉特征直接相加,公式如下:

$$v_i^{\text{fused}} = v_i^{\text{att}} + \lambda \cdot c^{\text{ocr}}$$

(5)

2) 门控特征融合(MultiMP):引入可学习的门控机制,动态调节 OCR 特征的贡献权重,公式可简化为:

$$g_t = \sigma(W_g h_t + b_g)$$

(6)

$$v_i^{\text{fused}} = v_i^{\text{att}} + \lambda \cdot g_t \odot c^{\text{ocr}}$$

(7)

其中, g_t 为当前时刻 t 的门控权重向量, σ 为 sigmoid 激活函数, h_t 为解码器 LSTM 的当前隐藏状态, W_g 和 b_g 为可学习参数。 v_i^{att} 为经过注意力加权的图像特征, c^{ocr} 为经过编码的 OCR 文本特征向量, λ 为调节融合强度的超参数(默认为 0.3), $\odot$ 表示逐元素相乘。该机制使模型能够在生成不同工艺步骤时,动态决定依赖视觉轮廓还是文本规范。

实验在相同的数据集和训练条件下进行,分别在第 15、25、55、65 个训练周期评估模型性能,从工序序列匹配率、BLEU 指标、精确率、召回率以及 ROUGE 系列指标等多个维度进行全面对比。表 3 展示了两种融合策略在不同训练周期下的性能对比结果。

表 3 简单拼接融合与门控特征融合的消融实验结果对比(Epochs)

Tab. 3 Ablation study results cmparison between simple concatenation fusion and gated feature fusion(Epochs)

模型(训练周期)	工序序列				工序文本			
	匹配率/%	BLEU	精确率	召回率	BLEU	ROUGE		
						Rouge-1	Rouge-2	Rouge-L
Simple(15)	9.16	0.485 3	0.784 9	0.752 7	0.593 2	0.618 9	0.483 7	0.613 4
MultiMP(15)	9.56	0.479 3	0.746 6	0.758 4	0.600 7	0.616 4	0.477 4	0.611 1
Simple(25)	15.54	0.524 8	0.779 0	0.761 8	0.610 1	0.642 3	0.512 8	0.637 9
MultiMP(25)	19.52	0.569 0	0.799 0	0.793 2	0.662 5	0.648 7	0.520 5	0.640 7
Simple(55)	23.90	0.589 6	0.811 6	0.805 9	0.662 3	0.698 1	0.585 5	0.691 7
MultiMP(55)	24.30	0.606 4	0.827 4	0.816 1	0.702 3	0.681 8	0.563 2	0.674 9
Simple(65)	23.90	0.573 8	0.837 4	0.772 4	0.655 0	0.688 2	0.572 0	0.681 8
MultiMP(65)	24.70	0.604 1	0.819 0	0.833 7	0.724 1	0.691 8	0.571 1	0.683 1

注:每一训练周期下,各评价指标的最优结果以加粗标出。
Note: For each training epoch, the best result of each evaluation metric is shown in bold.

实验结果可见,在工序序列完全匹配率方面,门控融合策略展现出持续优势。在训练初期(15 周期),MultiMP 的匹配率达到 9.56%,高于 Simple 策略的 9.16%。随着训练周期的增加,这一优势逐渐扩大,在 25 周期时差距最为明显,MultiMP 达 19.52%,相对 Simple 的 15.54% 提升达 25.6%。这表明门控机制能够更有效地学习工序间的逻辑关系,生成更加符合工程实际的工艺序列;在序列 BLEU 指标上,两种策略在不同训练阶段呈现交替领先态势。在训练初期(15 周期),Simple 策略在 BLEU 指标上略占优势,这可能是因为固定权重的拼接融合在训练初期收敛更快。

然而,随着训练深入,门控融合策略逐渐展现出其适应性优势,在 55 周期时 MultiMP 的 BLEU 得分达到 0.606 4,高于 Simple 的 0.589 6。而在 ROUGE 指标方面,门控融合在长文本生成的连贯性上表现更优。在 65 周期时,MultiMP 的 ROUGE-L 得分达到 0.683 1,优于 Simple 的 0.681 8,说明门控机制有助于生成语义更连贯、逻辑更合理的工艺描述。

实验验证了门控机制在多模态特征融合中的有效性。实验表明,与固定权重的简单拼接融合相比,门控机制能够根据当前解码上下文动态调节多模态特征的贡献度,这种自适应能力在复杂

的工艺描述生成任务中至关重要。门控机制通过注意力权重聚焦于与当前生成内容最相关的模式信息,不仅有效提升了生成文本的语义连贯性和逻辑一致性,还展现出良好的训练稳定性;尽管在训练初期收敛稍慢,但随着训练过程的深入,其性能提升更为显著,体现了更优的泛化能力与稳定性。

3.6 案例分析

图 5 展示的是一个罩螺母零件生成的案例,罩

螺母在工艺领域是一种关键紧固件,广泛应用于机械装配中,其核心作用是提供密封、防护和稳定连接,确保设备在高压或振动环境下长期可靠。图 5 左侧是该罩螺母零件的工艺图纸,图纸左上方为剖面图,展示罩螺母内部螺纹细节和孔结构;左下方为主视图,呈现零件外部轮廓、阶梯特征及尺寸标注;左侧右下角为技术要求栏,包含磁粉检测、表面粗糙度等关键工艺要求。

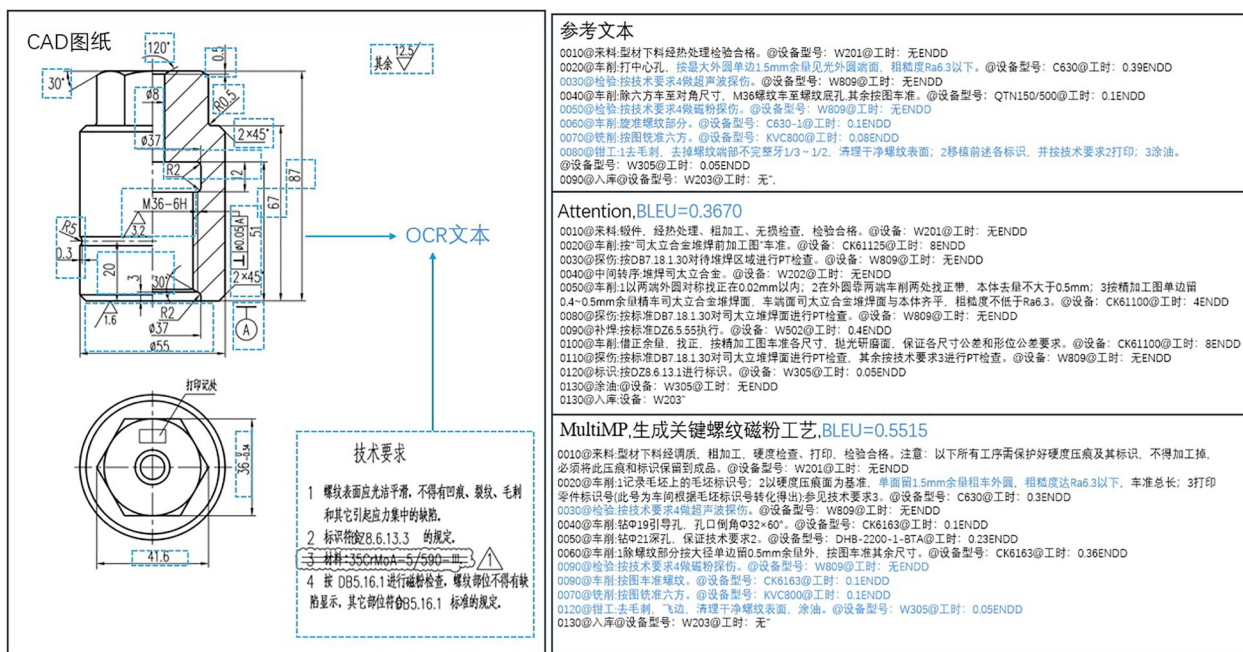


图 5 对于同一张图纸 Attention 模型与 MultiMP 的生成对比

Fig. 5 A comparative visualization of generation results between the attention model and MultiMP on the same engineering drawing

此案例的核心目标是通过图纸生成精准的工艺规划工序,要求模型识别罩螺母的加工特征(如车螺纹)和检测流程(如磁粉检测),并确保工序序列完整,逻辑合理。针对同一张包含螺纹结构和磁粉检测要求的工程图纸,基于纯视觉 Attention 的基线模型暴露了其模态单一的局限。由于缺乏 OCR 文本特征的深度融合,该模型仅依赖视觉特征进行判断,导致错误识别:将本应进行车削和螺纹加工的罩螺母错误识别为需要焊接的零件,生成了“焊接加工”等完全不相关的工序序列,其 BLEU 得分仅为 0.3670。相比之下,本文的 MultiMP 通过有效利用 OCR 提取的文本信息,成功生成了精确的制造指令,模型输出了包含精细参数的“粗车外圆时单面留 1.5 mm 余量,并控制粗糙度达 Ra6.3 以下”,并准确识别出关键的螺纹处理和磁粉检测工艺流程,生成了符合逻辑的“磁粉探

伤→车削螺纹→铣削六方→清理螺纹表面”一系列完整工序。获得了 0.5515 的 BLEU 得分,相比基线模型提升达 50.3%。

这一性能差距主要源于 MultiMP 方法的多模态融合机制,使其能够更全面地解析工程图纸中的信息。相较于传统方法在关联视觉模式与技术规格时存在的局限,本文提出的 MultiMP 方法在维持语义一致性与整合技术要素方面更具优势。该方法通过融合 OCR 识别的技术参数,生成了在流程上更符合规范、包含更详尽技术要素的工艺描述。在涉及多个制造工序和复杂几何公差场景下,MultiMP 方法所体现出的性能优势(如工艺序列的逻辑完整性与术语准确性)更为突出。通过视觉与文本模态的联合推理,该方法能够生成技术术语使用恰当、逻辑连贯的工艺序列,并在实验中成功解读了尺寸标注、表面粗糙度及检测标

准等关键信息。这些结果表明,MultiMP在处理实际工程图纸并生成可用于指导生产的工艺规划任务上具有潜力。

在案例分析的基础上,需要特别说明的是,本研究提出的多模态融合模型本质上是一个端到端的生成式系统,该模型的核心设计思路为引入 OCR 技术参数以优化工艺描述的生成效果,但当前阶段尚不能完全保证生成结果在严格物理约束下的逻辑准确性。从技术机制来看,端到端模型的特性决定了其依赖于数据驱动的模式学习,而非显式的逻辑规则推理。本节的案例分析表明,引入 OCR 特征后模型在 BLEU 指标上提升 50.3%,这证明了多模态融合在捕捉技术参数方

面的有效性。然而,工艺设计中的逻辑准确性涉及更深层次的工程知识推理,如加工顺序的合理性、工艺参数的匹配性等,这些需要结合领域专家的知识规则进行验证。在未来的研究工作中,我们计划引入专家知识库和规则引擎,开发约束满足算法,确保物理极限参数(如材料的加工公差极限)的严格遵守,与现有的多模态生成模型形成互补。

3.7 算法性能分析

在算法性能方面,本研究对比了 MultiMP 模型与对比模型在不同文本长度下的推理时间表现,实验结果如图 6 所示。

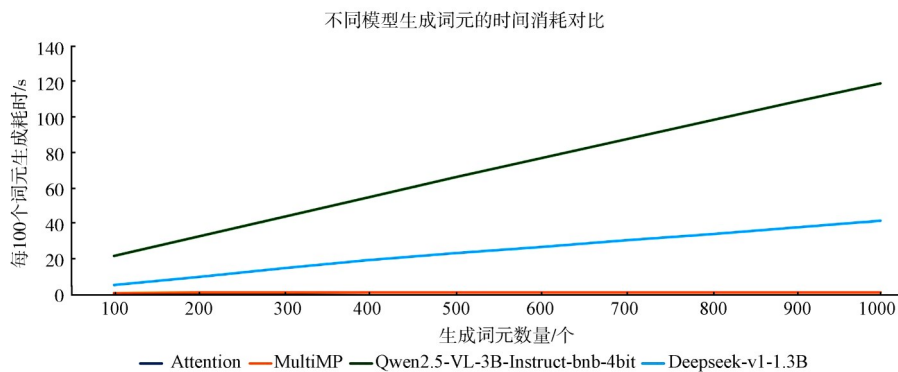


图 6 持续文本生成过程中每 100 个词元(Token)的处理时间

Fig. 6 Processing time for every 100 Tokens in the continuous text Generation process

在长文本生成任务中,不同模型呈现出明显的性能差异。Qwen2.5-VL-3B-Instruct-bnb-4bit 模型在生成 1000 个词元的连续文本时,平均每 100 个词元的推理时间达到 11.89 s,显示其较大的计算开销。Deepseek-v1-1.3B-chat 模型虽然参数量相对较少,但每 100 个词元仍需 4.12 s 的处理时间。相比之下,本研究提出的 MultiMP 模型凭借其轻量化架构设计,在相同条件下将平均每 100 个词元的推理时间控制在 0.10 s 以内,较基础 Attention 模型 0.094 s 的推理效率差距不大。

从趋势变化曲线可见,随着生成文本长度的增加,各模型表现出不同的时间增长特性。Qwen2.5-VL-3B-Instruct-bnb-4bit 与 Deepseek-v1-1.3B-chat 等通用大模型绝对推理时间显著高于专用模型。相比之下,MultiMP 与 Attention 模型的耗时随文本长度增长均保持线性趋势,且其增长斜率远小于通用模型,证明了在特定领域任务中轻量化设计的效率优势。值得注意的是,虽然 MultiMP 模型在处理多模态特征时增加了额外的

计算开销(比 Attention 模型约慢 10.8%),但其在工艺描述生成质量上的提升使得这一性能代价可以接受。

综合来看,MultiMP 模型通过优化的多模态融合机制,在保持与基础 Attention 模型相近推理效率的同时,提升了生成质量,为工艺设计领域的实时应用提供了可行的技术解决方案。

4 结 论

本研究针对工艺设计领域多模态信息融合不充分、长文本生成语义一致性差及关键工艺参数提取不足等核心问题,提出了一种基于图像描述与 OCR 文本融合的工艺设计模型构建算法。该算法以“CNN-LSTM+Attention”架构为基础,通过新增 OCR 文本处理与特征融合模块,实现了视觉特征与文本参数的深度融合。与简单的特征拼接不同,本算法采用双阶段融合策略:在解码器初始化阶段,将图像全局特征与 OCR 文本特征拼接为融合向量,为解码器提供丰富的多模态初始状态;

在解码过程中,引入门控融合机制,动态调整 OCR 特征在每一解码时间步的贡献,从而自适应地平衡视觉与文本信息。同时,引入 PaddleOCR 文本提取工具结合领域定制化预处理流程开展实验。实验表明,本算法在不显著增加计算开销的前提下,较基础 Attention 模型以及通用大模型在 BLEU 和 ROUGE 指标上均取得提升,证明了多模态融合的有效性。然而,当前模型在建立图像区域与 OCR 参数之间的显式语义关联方面仍有不足。未来工作将引入跨模态注意力等机制以进一步优化模型,继续扩展工艺设计专用语料库,并探索模型在实时工艺规划、智能图纸审核等工业任务中的应用,为工艺设计智能化提供更全面的技术支撑。

参考文献:

- [1] Kushwaha R, Biswas A. Hybrid feature and sequence extractor based deep learning model for image caption generation[C]//2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021: 1-6.
- [2] Gui H, Yang Z, Harish A R, *et al.* GenPattern: Dual-graph enhanced sewing pattern generation via multimodal large language model[J]. J Manuf Syst, 2025, 83: 822-838.
- [3] Sun C W, Hou J L, Liu X G, *et al.* Large language model prompt generation method for engineering drawing understanding[J]. J Comput Appl, 2025, 45(3): 801-807. [孙晨伟, 侯俊利, 刘祥根, 等. 面向工程图纸理解的大语言模型提示生成方法[J]. 计算机应用, 2025, 45(3): 801-807.]
- [4] Xu K, Ba J, Kiros R, *et al.* Show, attend and tell: Neural image caption generation with visual attention[C]//Proceedings of the 32nd International Conference on Machine Learning (ICML), 2015: 2048-2057.
- [5] See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017: 1073-1083.
- [6] TAN M, LE Q V. EfficientNet: Rethinking model scaling for convolutional neural networks[C]//Proceedings of the 36th International Conference on Machine Learning (ICML). Long Beach, USA: PMLR, 2019: 6105-6114.
- [7] McLaughlin N, Martinez del Rincon J, Miller P. Recurrent convolutional network for video-based person re-identification [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 1325-1334.
- [8] Elman J L. Finding structure in time[J]. Cogn Sci, 1990, 14(2): 179-211.
- [9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[PP/OL]. V1. (2014-09-04) [2025-05-06]. <https://arxiv.org/abs/1409.1556>
- [10] He K, Zhang X, Ren S, *et al.* Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.
- [11] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Comput, 1997, 9(8): 1735-1780.
- [12] Liu X G, Guo Y, Li Y, *et al.* A method for constructing Domain-Specific large language models oriented to manufacturing process design[J]. J Sichuan Univ Nat Sci Ed, 2025, 62(3): 513-521. [刘祥根, 郭彦, 李玥, 等. 面向工艺设计的领域大模型构建方法[J]. 四川大学学报: 自然科学版, 2025, 62(3): 513-521.]
- [13] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need[C]//Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS), 2017: 6000-6010.
- [14] Li J, Li D, Xiong C, *et al.* BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation [C]//Proceedings of the 39th International Conference on Machine Learning (ICML), 2022: 12888-12900.
- [15] Alayrac J-B, Barr I, Barreira R, *et al.* Flamingo: A visual language model for few-shot learning [C]//Advances in Neural Information Processing Systems 35, 2022: 23716-23736.
- [16] Lu J, Clark C, Lee S, *et al.* Unified-IO 2: Scaling autoregressive multimodal models with vision, language, audio, and action [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024: 26429-26445.
- [17] Li J, Li D, Savarese S, *et al.* BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models [C]//Proceedings of the 40th International Conference on Machine Learning, 2023: 19730-19742.

- [18] Dai W, Fung P N, Hoi S, *et al.* InstructBLIP: Towards general-purpose vision-language models with instruction tuning[C]//Advances in Neural Information Processing Systems 36. New Orleans, Louisiana, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 49250-49267.
- [19] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult [J]. IEEE Trans Neural Netw, 1994, 5(2): 157-166.
- [20] Ke W, Zheng Y, Li Y, *et al.* Large language models in document intelligence: A comprehensive survey, recent advances, challenges, and future trends [J]. ACM Trans Inf Syst, 2026, 44(1): 1-64.
- [21] Kirana K C, Kumalasari I, Oktagalu G Q, *et al.* Comparison of tesseract OCR, easy OCR, and transformer OCR on handwritten image [C]//2025 9th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), 2025: 1-6.
- [22] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Trans Pattern Anal Mach Intell, 2017, 39(11): 2298-2304.
- [23] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Netw, 2005, 18(5/6): 602-610.
- [24] Bautista D, Atienza R. Scene text recognition with permuted autoregressive sequence models [M]//Computer Vision-ECCV 2022. Tel Aviv, Israel: Springer Nature Switzerland, 2022: 178-196.
- [25] Atienza R. Data augmentation for scene text recognition[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021: 1561-1570.
- [26] Lu Y, Chen Y, Zhang S. Research on the method of recognizing book titles based on paddle OCR [C]//2024 4th International Signal Processing, Communications and Engineering Management Conference (ISPCEM), 2024: 1044-1048.
- [27] Sarkar O, Sinha S, Jena A K, *et al.* Automatic number plate character recognition using paddle-OCR [C]//2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET). Nagpur, India: IEEE, 2024: 1-7.
- [28] Baltrušaitis T, Ahuja C, Morency L-P. Multimodal machine learning: A survey and taxonomy[J]. IEEE Trans Pattern Anal Mach Intell, 2019, 41(2): 423-443.
- [29] Arevalo J, Solorio T, Montes-y-Gómez M, *et al.* Gated multimodal networks[J]. Neural Comput Appl, 2020, 32(14): 10209-10228.
- [30] Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions [J]. IEEE Trans Pattern Anal Mach Intell, 2017, 39(4): 664-676.
- [31] Nagrani A, Yang S, Arnab A, *et al.* Attention bottlenecks for multimodal fusion [C]//Proceedings of the 35th International Conference on Neural Information Processing Systems, 2021: 14200-14213.
- [32] Anderson P, He X, Buehler C, *et al.* Bottom-up and top-down attention for image captioning and visual question answering [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 6077-6086.
- [33] Tay Y, Dehghani M, Bahri D, *et al.* Efficient transformers: A survey [J]. ACM Comput Surv, 2023, 55(6): 1-28.
- [34] Liu C, Liu C, Luo Z, *et al.* ARIF: An adaptive attention-based cross-modal representation integration framework [C]//Artificial Neural Networks and Machine Learning-ICANN 2024, 2024: 3-18.
- [35] Mehta N K, Arvind, Prasad S S, *et al.* Optimizing multitask industrial processes with predictive action guidance [J]. IEEE Trans Automat Sci Eng, 2025, 22: 22833-22844.
- [36] Ren L, Wang H, Wang Y, *et al.* Foundation models for the process industry: Challenges and opportunities [J]. Engineering, 2025, 52: 53-59.
- [37] Kingma D P, Ba J. Adam: A method for stochastic optimization [C]//Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.
- [38] Hinton G E, Srivastava N, Krizhevsky A, *et al.* Improving neural networks by preventing co-adaptation of feature detectors [J]. Computer Science, 2012, 3(4): 212-223.
- [39] Papineni K, Roukos S, Ward T, *et al.* BLEU: A method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, 2001: 311.
- [40] Lin C-Y. ROUGE: A package for automatic evaluation of summaries [C]//Text Summarization Branches Out, 2004: 74-81.

(责任编辑:伍少梅)