中北大学学报(自然科学版)
JOURNAL OF NORTH UNIVERSITY OF CHINA(NATURAL SCIENCE EDITION)

文章编号：1673-3193(XXXX)XX-0001-13

面向中文多场景文本图像生成的高质量多模态数据集MojiTextCN的构造方法

张嘉伟[1]，杨秋翔[1]，高　宁[1]，韩伟明[1]，张富为[2*]

（1. 中北大学 软件学院，山西 太原 030051；2. 中北大学 计算机科学与技术学院，山西 太原 030051）

摘　要：针对当前多模态大模型在中文密集文本图像生成与理解任务中，因缺乏覆盖复杂排版与多样字体的高质量数据集而导致性能受限的问题，本文构建了面向多场景的大规模多模态数据集MojiTextCN。本文设计了一套融合生成式大模型与显式中文排版规则的自动化数据构造方法，该方法不局限于单一的API黑盒拼接范式，通过“合成＋真实”双轨策略，结合视觉语言模型与自动化标注技术，并引入《通用规范汉字表》进行底层字形硬约束与多级校验，有效保障了字形准确与排版一致。数据集共包含CleanChar(纯文本，约201万张)、PicTextLayout(图文混排，约41.9万张)、SlideDoc(真实幻灯片，约8万张)及SceneTextSynth(场景合成)四个优势互补的子集，总规模逾300万张。为了全面验证数据的质量，本文引入了多维度交叉评估体系。在字面可读性上，基于OCR的评测显示复杂场景子集的字符错误率(CER)低至2.40%；在图像超分辨率任务中，基于本数据集训练的模型取得了0.37%的极低CER，进一步证明了构建原生中文语料以克服中英文领域分布差异的必要性。此外，基于视觉大模型的零样本排版评估分数为4.27(满分5.0)。测评结果表明当前主流模型在长文本渲染与交错文档生成任务中仍面临字符扭曲与布局错位等局限。MojiTextCN为提升模型在复杂中文语境下的跨模态生成与理解能力提供了有效的数据基础与评测标准，具有一定的研究与应用价值。

关键词：多模态数据集；中文文本图像；文本生成；密集文本；跨模态学习

中图分类号：TP391.41　　**文献标识码：**A　　**doi:** 10.62756/jnuc.issn.1673-3193.2026.03.0014

引用格式：张嘉伟,杨秋翔,高宁等.面向中文多场景文本图像生成的高质量多模态数据集MojiTextCN的构造方法[J].中北大学学报(自然科学版),DOI:10.62756/jnuc.issn.1673-3193.2026.03.0014.

Zhang Jiawei, Yang Qiuxiang, Gao Ning, et al. Constructing Method of a High-Quality Multimodal Dataset MojiTextCN for Generating Chinese Text Images Across Multiple Scenarios[J]. Journal of North University of China(Natural Science Edition),DOI:10.62756/jnuc.issn.1673-3193.2026.03.0014.

Constructing Method of a High-Quality Multimodal Dataset MojiTextCN for Generating Chinese Text Images Across Multiple Scenarios

Zhang Jiawei[1], Yang Qiuxiang[1], Gao Ning[1], Han Weiming[1], Zhang Fuwei[2*]

(1. School of Software, North University of China, Taiyuan 030051, China;
2. School of Computer Science and Technology , North University of China, Taiyuan 030051, China)

Abstract: Current large multimodal models (LMMs) are often constrained in Chinese dense text-image

收稿日期：2026-03-25

基金项目：山西省基础研究计划项目(202403021222152)

作者简介：张嘉伟(2000－)，男，硕士生，主要从事多模态大模型合成数据的研究。

*通信作者：张富为(1991－)，男，讲师，博士，主要从事视频理解与图像增强方面的研究。E-mail：20240048@nuc.edu.cn。

generation and understanding tasks by the scarcity of high-quality datasets that encompass complex layouts and diverse fonts. To address this limitation, we introduced MojiTextCN, a large-scale, multi-scenario multimodal dataset. Moving beyond conventional black-box API concatenation paradigms, we designed an automated data construction pipeline that integrated generative large models with explicit Chinese typesetting rules. By adopting a "synthetic + real" dual-track strategy, combining Vision-Language Models (VLMs) with automated annotation techniques, and incorporating the Table of General Standard Chinese Characters for low-level glyph hard-constraints and multi-stage verification, we effectively ensured glyph accuracy and layout consistency. MojiTextCN comprised four complementary subsets: CleanChar (pure text, approximately 2.01 million images), PicTextLayout (image-text mixed layouts, approximately 419,000 images), SlideDoc (real-world presentation slides, approximately 80,000 images), and SceneTextSynth (synthetic scene text), scaling up to over 3 million images in total. To comprehensively validate the data quality, we proposed a multi-dimensional cross-evaluation framework. In terms of textual legibility, OCR-based evaluations demonstrate a Character Error Rate (CER) as low as 2.40% on the complex scene subset. Furthermore, models trained on this dataset achieve a remarkably low CER of 0.37% in image super-resolution tasks, reinforcing the necessity of constructing native Chinese corpora to bridge the domain distribution gap between Chinese and English. Additionally, zero-shot layout evaluations leveraging large vision models yield a high score of 4.27 (out of 5.0). Our benchmarking results also reveal that current mainstream models still suffer from character distortion and layout misalignment in long-text rendering and interleaved document generation tasks. Ultimately, MojiTextCN provides a robust data foundation and evaluation benchmark for enhancing cross-modal generation and understanding within complex Chinese contexts, offering significant value for both academic research and practical applications.

Key words: multimodal dataset; Chinese text image; text generation; dense text; cross-modal learning

0 引 言

近年来，多模态生成模型取得了显著进展。从早期的DALL[1]，Stable Diffusion[2]，到近期发布的SDXL[3]、GPT-Image-1[4]和GPT-4o[5]，文本驱动的图像生成能力持续增强。这些模型不仅能够理解更长的上下文输入，还在跨语言、多模态推理等方面取得显著的进展，使得基于自然语言指令的图像生成具备了更强的可控性与表现力，并在一定程度上缓解了过去在英文富文本图片生成上的局限。尤其是GPT-Image-1[4]与Qwen-Image[6]等新一代模型，将语言模型的长文本理解能力与视觉生成过程深度融合，为复杂布局建模、跨模态语义对齐及多语言应用带来了新的可能性。

尽管模型性能持续提升，但其在中文密集文本图像的生成与理解方面仍存在显著短板。深入剖析这一现象，其原因在于中文文本生成的内在机制极其复杂，且现有底层数据资源存在严重的领域错配(Domain Gap)。在形态学上，中文字符系统相较于拉丁字母具有高得多的拓扑复杂度与结构信息熵，要求生成模型必须具备极高的像素级细粒度控制力；在排版逻辑上，真实的中文应用场景往往呈现出高密度、多层级的图文混排特征。但是，现有大规模数据集如LAION-2B[7]、COCO Captions[8]等主要被英文数据主导，即便包含文字信息，也多为稀疏的短句或标签；另一方面，已有中文数据集大多聚焦于场景文字识别或简单的合成样本，缺乏对复杂排版、字体多样性与语境变化的系统覆盖。这种原生高质量数据的匮乏，直接导致多模态模型在处理复杂中文语境时，频繁出现字形扭曲、排版错乱与语义断裂等严重缺陷。

为了弥补当前研究的不足，本文提出了MojiTextCN，这是一个面向中文密集文本图像生成与理解任务的大规模多模态数据集。有别于将第三方生成模型视为功能组件进行简单"线性拼接"的传统黑盒构造范式，MojiTextCN在数据管线的设计中深度融合了针对中文特性的显式规则约束与领域先验知识。与现有资源相比(见图1与表1)，MojiTextCN在以下方面具有显著优势：1)在规模上，数据集包含从单词、短句到段落级文本的多粒度样本；2)在场景上，重点覆盖广

告、幻灯片、古诗、书籍封面等高密度排版场景，并设计了基于空间感知的自适应图文交错布局算法，高保真模拟真实排版；3）在构建方法上，并未局限于单一的 API 调用，而是提出并实施了“生成-解析-校验”的多级闭环反馈机制，结合了大模型生成、OCR 自动标注、《通用规范汉字表》的底层字形硬约束与人工校验，有效确保了数据的多样性、可读性与字形的一致性。

表 1　现有富文本图像生成数据集对比

Tab. 1　Comparison of existing rich-text image generation datasets

数据集名称	样本数	标注方式	数据集类型	标注方式	Token长度
AnyWords3M[9]	3×10^6	Caption+OCR	真实数据	机器标注	9.92
TextCaps[10]	28×10^3	Caption	真实数据	人工标注	26.36
SynthText[11]	0.8×10^6	OCR	合成数据	机器标注	13.75
Marion10M[12]	10×10^6	Caption+OCR	真实数据	机器标注	16.13
RenderedText[13]	12×10^6	Text	合成数据	机器标注	21.21
TextAtlas5M[14]	5×10^6	Caption/OCR/Text	真实+合成数据	人工+机器标注	**148.82**
MojiTextCN	3×10^6	Caption/OCR/Text	真实+合成数据	人工+机器标注	**119.24**

图 1　现有数据集对比

Fig. 1　Comparison of existing datasets

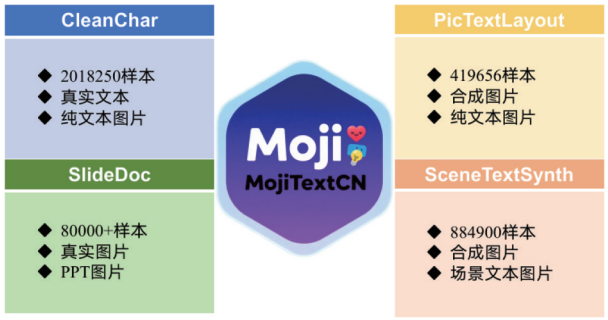

图 2　MojiTextCN数据集的构成

Fig. 2　Composition of the MojiTextCN dataset

MojiTextCN 在一定程度上扩充了中文密集文本图像的数据资源，并为文本驱动图像生成、文字超分辨率、检测识别及跨模态检索等任务提供了可供参考的实验支撑。同时，该数据集能够为多模态大模型在中文语境下的进一步研究提供有益的实验基础。

本文的主要贡献如下：

1）提出了一套融合生成式大模型与显式中文排版规则约束的数据构造流程，减少了传统黑盒API拼接带来的排版错乱与字形扭曲问题。

2）构建了 MojiTextCN，一个覆盖从词汇到段落级文本的高质量多粒度数据集，涵盖广告、幻灯片、真实交错文档等多种场景，切实反映复杂排版需求。

3）构建了涵盖文本基础可读性、跨模态语义对齐度，以及视觉排版合理性的多维交叉评估体系，从多个维度验证了数据集的质量与可靠性。MojiTextCN 可作为中文多模态智能技术预训练与下游评估的数据基础。

1　相关工作

1.1　文本条件图像生成

近年来，生成建模的发展主要集中在两大框架：扩散模型[15-17]与自回归模型[1,18-20]。扩散模型（如DALL·E[21]、Parti[22]）通过逐步去噪生成高保真图像，但推理速度相对较慢。相比之下，自回归方法（如VQGAN[18]、MAGVIT2[19]、PIXEL[23]）基于离散token进行序列建模，在生成效率与图像质量之间取得了更好的平衡。近期的 GPT-Image-1[4]与GPT-4o-0325[5]等模型在英文文本驱动的图像生成任务上展现出前所未有的能力，但在中文密集文本图像生成方面仍面临显著挑战。

1.2　文本密集图像生成与文字渲染

为增强生成模型的文字渲染能力，许多已有研究在扩散框架中引入了字符感知机制，例如：TextDiffuser[12,24]、AnyText[9]、GlyphDraw[25]与GlyphControl[26]等。这些方法在短文本或稀疏文本生成任务中表现良好，但在长文本生成、复杂排版及跨字体场景下仍存在显著不足，难以实现稳定的文字结构与语义一致性。

1.3　文本图像数据集

高质量的数据集对文本图像生成与识别的发展

至关重要。早期的数据集如SynthText[11]与COCO-Text[27]主要聚焦于自然场景中的文字检测与识别，而Marion10M[28]与AnyWords-3M[29]将研究扩展到多语言短文本生成。最近的TextAtlas-5M[14]首次制作了聚焦大规模密集英文文本的图像，为长文本驱动的图像生成提供了重要基准。然而，在中文语境下，仍缺乏覆盖密集文本、多样字体与复杂排版的大规模数据资源，且现有的模型往往难以解决中文特有的长尾字形扭曲与高密度排版错乱等问题。

为了缓解上述中文图片生成中的问题，本文提出了一个专注于中文密集文本图像生成与理解的大规模多模态数据集MojiTextCN，该数据集在数据构造流程中深度融合了针对中文特性的严格字形约束、基于空间感知的自适应排版算法以及环境感知渲染机制，从底层工程层面系统性保障了数据的排版整洁性与视觉一致性。该数据集可为多模态大模型在复杂中文场景下的研究与评测提供高质量、高可靠性的数据。

2 数据集构建方法

本文所构建的MojiTextCN数据集是一个覆盖多种密度中文文本的多场景数据集。该数据集在设计上兼顾真实应用的多样性及数据构建的可控性和质量，通过“合成子集＋真实子集”的两级策略实现平衡：一方面，以大规模、结构可控的合成样本支撑模型对密集文本与复杂排版结构的学习；另一方面，利用多源真实样本补充视觉多样性与域外泛化能力，从而提高模型在开放场景下的稳健性。数据集地址为https://modelscope.cn/datasets/shoohi/MojiTextCN。

MojiTextCN共包含4个子集：CleanChar、PicTextLayout、SlideDoc和SceneTextSynth。其中，CleanChar子集主要收录真实文本与纯文本图像，用于字符级识别与字体分析；PicTextLayout子集包含多样化的合成排版样例，聚焦文本区域的结构布局与语义对齐；SlideDoc子集由真实幻灯片及文档图像构成，用于模拟办公类密集文本场景；SceneTextSynth子集则通过多场景背景合成生成，反映真实环境中的复杂文字分布。

2.1 CleanChar子集

本文构建的纯文字图像子集CleanChar包含了单字、短词与完整句子的高质量清晰渲染，如图3所示。该子集的文本来源于《通用规范汉字表》一级常用字及其组成的常见词汇、语料与文学文本，确保了内容的广泛性与语言基础覆盖。为保证字符可读性与渲染一致性，所有生成内容均限定在一级常用字范围内，非一级常用字及生僻字样本在数据预处理阶段被系统性筛除，从而确保了字符分布的稳定性与兼容性。

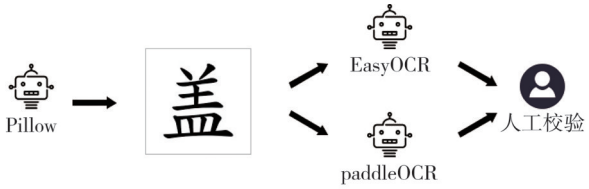

图3 CleanChar部分数据生成步骤

Fig. 3 The data generation process of the CleanChar subset

在数据生成过程中，基于Python图像处理库Pillow来实现文字绘制，以白底画布为背景，采用多种常用中文字体进行渲染，可以在保持规范性的同时提供多样性的字体风格。为进一步确保字体兼容性与字形一致性，从多个中文字体网站收集了约40种字体，涵盖宋体、黑体等多种风格字体，并针对《通用规范汉字表》一级常用字进行了系统筛选。然后分别在所有候选字体下进行渲染，并使用PaddleOCR与EasyOCR对生成结果进行识别。若任一模型无法正确识别该字符，则将该样本标记为“识别失败”，并进入人工复核环节。人工复核阶段由标注人员确认该字形是否与标准简体字一致。最终保留约20种兼容性良好的字体用于后续渲染。以上工作有效避免了“缺字”“方框字”等问题。最终，CleanChar子集共包含约201万张样本，覆盖从单字到句子的多粒度文本结构。

2.2 PicTextLayout子集

本文在CleanChar子集的基础上构建了图文混排子集PicTextLayout，用于模拟真实应用中文字与图片交织的复杂排版场景，如图4所示。该子集的设计理念源于交错数据的结构特征，旨在让模型学习文字和视觉内容在空间布局与语义层面的对应关系。

构建流程主要包括4个阶段：

1）预设了8个核心主题类别，分别为“西幻”“动漫”等，每个大类下进一步细分若干小类，覆盖多种真实语义场景。为确保内容多样性与语义可控性，每个大类均由GPT-4o生成不少于1 000条文本描述，随后每条prompt由Stable Diffusion 3.5生成20～25张图像样本。所有prompt在生成前被封装为JSON

格式，以统一模型接口与样本元数据管理。

2）针对生成的图像，利用多模态大模型Qwen2.5-VL-7B对图片内容生成相应的中文文本描述与英文caption。需要说明的是，为了最大限度地贴近真实世界交错文档（Interleaved Data）的客观物理分布，不强制所有图文样本具备严格的语义对齐关系。这一设计理念借鉴了大规模交错数据集（如MMC4、OBELICS等）的构建共识，即在真实的网页、海报等排版中，文本通常是对视觉元素的上下文延伸或背景补充，而非纯粹的图像描述。如果强行施加1∶1的绝对语义对齐，不仅会破坏图文排版的自然多样性，还会导致数据集退化为传统的人工合成图文对，容易诱导预训练模型产生过度拟合的偏置，从而难以适应真实应用中松散、复杂的图文结构。因此，保留适度的“弱关联”样本，是保障本数据集在多模态语义分布上具备高保真度的关键选择。

3）图文交错布局生成。为模拟interleaved data的空间混排特征，在白底画布上随机放置图像，随后在图片四周环绕生成对应文字段落。文本排版遵循从上到下、从左到右的阅读顺序，根据图片位置自动划分可排区域，实现图文交错的自然过渡。每个样本包含1张图片与1段中文文本，平均文本长度约为25～60字。

4）渲染与版式控制。本文采用Python图像处理库Pillow完成最终文本绘制。

PicTextLayout子集共生成约41.9万张样本，涵盖多领域、多语义层次的图文混排场景。该子集在保证图文语义一致性与排版合理性的同时，维持了与CleanChar子集一致的渲染规范与字体安全性，为中文密集文本的图文混排生成、视觉文本理解及跨模态布局学习提供了高质量的实验数据支撑。

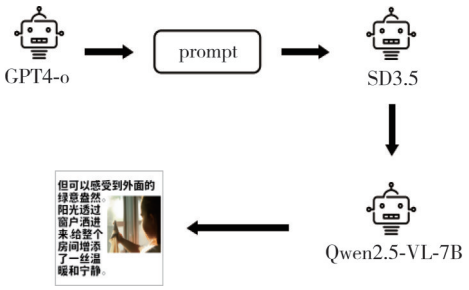

图4 PicTextLayout部分数据生成步骤

Fig. 4 The data generation process of the PicTextLayout subset

2.3 SlideDoc子集

为了覆盖真实办公与学术场景中的高密度文本图像，进一步构建了SlideDoc子集，其数据主要来源于约8万张幻灯片图片，如图5所示。这部分幻灯片的主题分布如图6所示，涵盖家庭教育等多个领域，能够较全面地反映教育、科普与公共文化类文档场景。

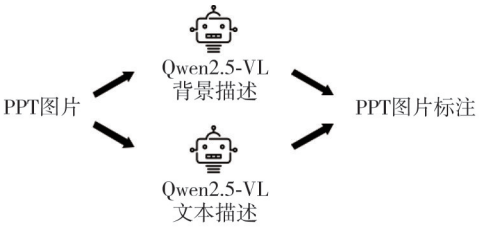

图5 SlideDoc子集数据生成步骤图

Fig. 5 The data generation process of the SlideDoc subset

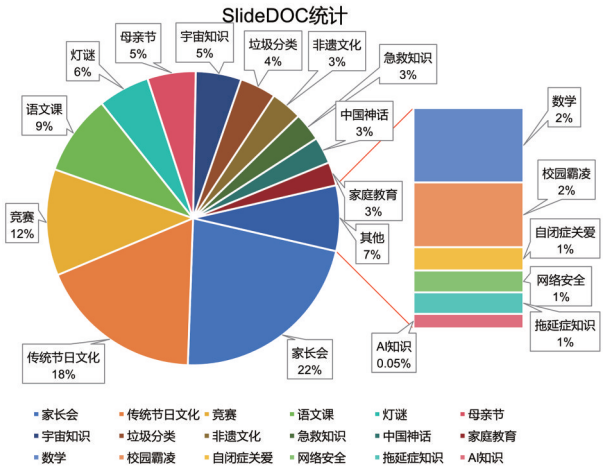

图6 SlideDoc子集主题分布

Fig. 6 Topic distribution of the SlideDoc subset

在数据清洗阶段，首先采用EasyOCR对所有幻灯片图像进行粗筛，过滤掉文字区域极少或无文本的样本，仅保留文字内容识别超过10个字的页面。随后，对剩余样本使用多模态大模型Qwen2.5-VL-7B进行语义解析与标注，如图5所示，模型一方面生成与幻灯片整体语义一致的页面级描述，另一方面执行文字区域的识别与抽取，以保证文本内容的完整性。本文选用视觉能力相对有限的7B模型，是为了平衡文生图与多模态大模型的视觉能力，即平衡两者对模糊、低对比度或复杂背景文字的理解能力。此类模糊区域往往连人类也难以辨识，因此所得结果对模型的鲁棒性具有重要参考价值。

最后，将背景描述与文字识别结果结合，生成结构化标注文件，以JSON格式存储，包括图像路径、页面级语义描述及文字区域标注三层信息。在整个筛选过程中，约有7%的幻灯片因文字不

足或识别置信度低而被剔除，最终保留约8万张高质量样本。该子集显著提升了MojiTextCN在真实办公类文档场景中的覆盖深度，为跨模态理解任务提供了坚实的数据支撑。

2.4 SceneTextSynth子集

不同于子数据集PicTextLayout的环绕式图文混排，SceneTextSynth子集旨在模拟真实场景下的文本图像生成，用于训练和评估模型在复杂背景条件下的文字生成与识别能力，如图7所示。

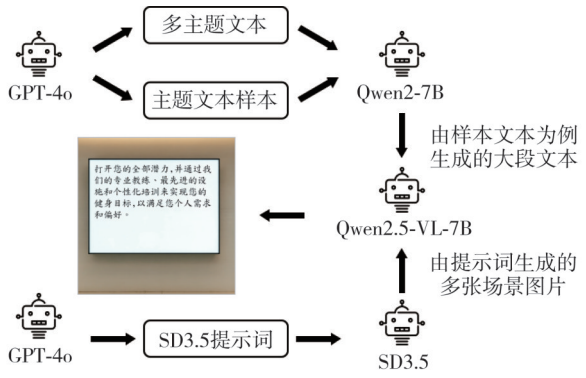

图7 SceneTextSynth子集数据生成步骤

Fig. 7 The data generation process of the SceneTextSynth subset

构建流程主要包括4个阶段：

1）背景图生成：使用Stable Diffusion 3.5按照预设主题生成多样化场景图片，共获得约88.5万张样本。部分背景图片来源于TextAtlas5M[14]中生成的开放式文本图像数据，以提升场景多样性与真实感。

2）候选区域标注：利用多模态大模型Qwen2.5-VL-7B对生成图片进行分析，自动标注适合渲染文字的背景区域。模型输出矩形边界框并附带背景亮度均匀度、纹理复杂度与区域可读性等特征，以指导后续文字嵌入位置的筛选。

3）文本生成：使用Qwen3-8B[30]模型生成与场景语义相匹配的中文文本描述(如广告语、标题、说明文字等)，通过模型自身的上下文一致性能力完成语义校正与格式化处理，确保文本内容自然且与场景主题匹配。

4）文本嵌入与渲染：在候选区域中执行文字嵌入，采用Python图像处理库Pillow完成最终渲染，复用CleanChar的20种安全字体与字号范围。在渲染过程中，根据候选区域的背景亮度与纹理特征自动判断颜色对比度，优先选择黑白两色以获得最高的可读性；同时结合亮度自适应、透视变换与轻量模糊校正，使文字在光照方向和视觉透视上与背景自然融合。

SceneTextSynth子集有效模拟了广告、街景与产品展示等真实应用中的高密度文本分布，为模型在复杂背景下的文本生成与识别提供了丰富的训练与评测样本。其场景主题分布如图8所示，涵盖公告栏、广告牌、学术报告、黑板、广告海报、显示器、数字显示屏与横幅等多种类别，整体分布多样且符合实际场景特征。

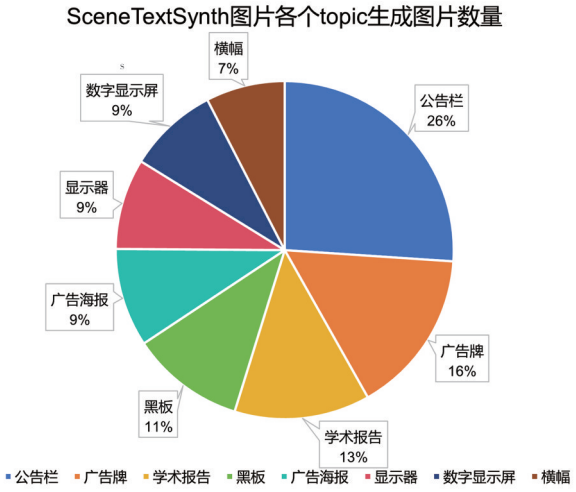

图8 SceneTextSynth子集主题分布

Fig. 8 Topic distribution of the SceneTextSynth subset

3 数据集分析

为了全面评估MojiTextCN数据集的特征及其在多模态任务中的潜在应用价值，本节将从统计学特征、分布规律以及质量评估等多个维度对其进行深入剖析。一个具备高训练价值的数据集，不仅需要具备足够的样本规模，更需要在语义长度、视觉风格及排版布局上呈现出良好的多样性与平衡性。因此，本节首先梳理了数据集的整体规模与构成，随后重点考察了各子集内部的文本密度分布与长尾特征，旨在揭示该数据集在模拟真实世界复杂中文场景方面的优势。同时，引入了基于OCR的质量评估与对比分析，以量化证明合成数据在清晰度与结构一致性上的表现。总体而言，MojiTextCN共汇集了约300万张图像及对应的高精度文本标注，为训练鲁棒的中文多模态生成与理解模型提供了坚实的数据基础。

3.1 标注长度分布

对MojiTextCN数据集个子集图像中文字标注

部分进行统计，结果如图 9 所示。

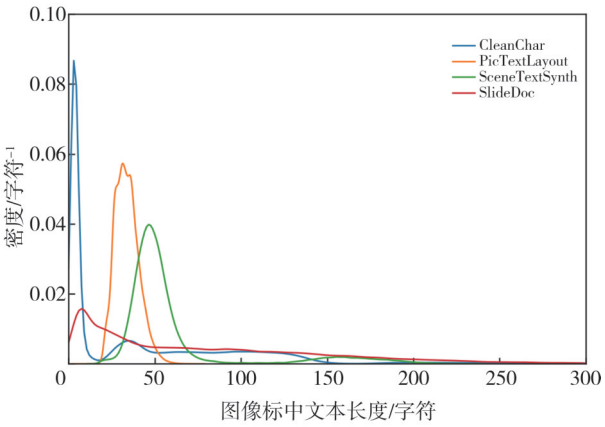

图 9　MojiTextCN各子集的标注文本长度分布

Fig. 9　Distribution of annotated text lengths in MojiTextCN subsets

由图 9 可以看出，MojiTextCN不同子集在文本长度分布上呈现出显著差异。SlideDoc子集以20字左右的短文本为主，反映其侧重于标题或要点的特征；CleanChar子集分布较为集中，大多落在10～40字之间，对应单词或短句；PicTextLayout与SceneTextSynth子集整体右移，平均文本更长，其中SceneTextSynth在50～150字区间形成长尾分布，涵盖了复杂场景中的段落级文本。整体而言，MojiTextCN各子集在短文本与长文本之间形成互补，为从字符级识别到长文本建模的多样化任务提供了支持。

3.2　长文本图像生成

长文本生成，尤其是涉及复杂汉字结构与密集排版的图像合成，一直是多模态生成领域的关键瓶颈。为了深入探究现有技术在该任务上的性能边界，本文设计了定性对比评测实验，图 10 为各模型在相同中文长文本输入时生成的样例，其中，GT表示数据集原始图像（Ground Truth）。

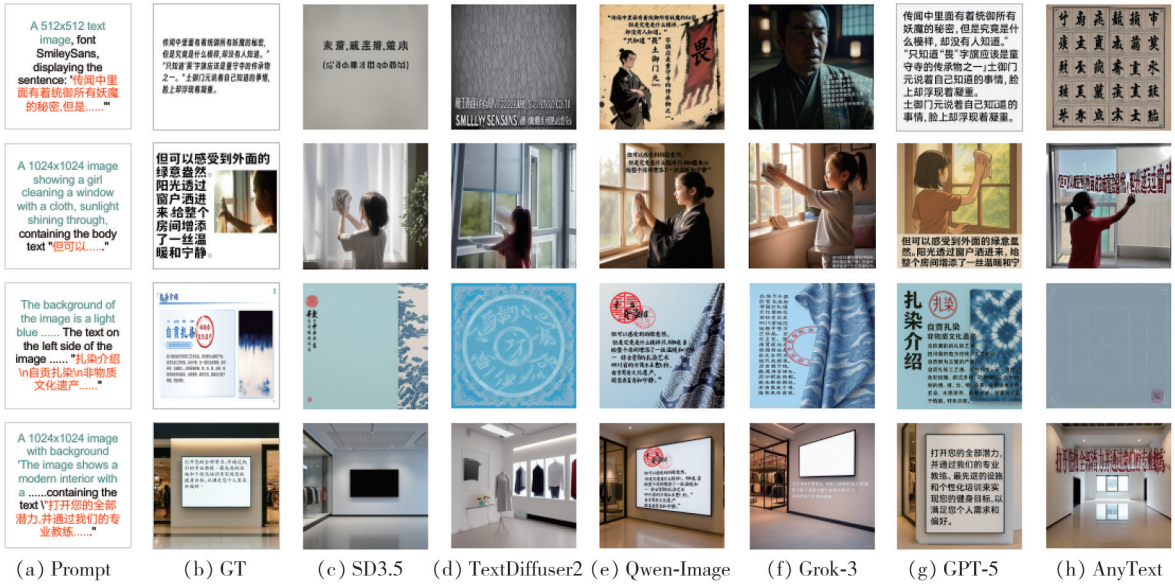

图 10　主流文生图模型的生成的图片

Fig. 10　Generated images from mainstream text-to-image models

本文数据集在基准模型的选择上力求覆盖广泛的技术路径，既包含了AnyText和TextDiffuser等针对文字渲染优化的开源模型，也包含了Qwen-Image、Grok-3等当前最先进的通用多模态大模型。本文基于实验结果分析了当前技术面临的主要挑战与局限性，主要结论如下：

1）短文本与长文本生成对比。开源模型在渲染短文本片段时表现良好，但随着输入长度的增加，其性能迅速下降，常出现文本被截断或字符严重扭曲的情况。

2）先进模型与早期的生成模型的对比。当前的先进模型在文本长度建模、上下文一致性以及鲁棒性方面取得了显著提升。尽管如此，这类模型在细粒度文字生成任务中仍存在一定局限，表现为部分字符的辨识度不足、重复或缺失现象以及字符间距不规则等问题。

3）交错文档生成问题。在各类文生图系统中，最具挑战性的任务依然是交错文档的生成。多数传统模型在版面规划与文本布局上缺乏全局一致性，难以实现文字与背景元素的精确协同。然而，最新的自回归多模态模型已在该方向取得了显著进展。借助统一的视觉-语言表示与层级式生成机制，该模

型能够在一定程度上完成背景语义与文本区域的联合建模，实现相对稳定的排版规划与语义对齐，缓解了过去模型在交错场景下的错位问题。

3.3 基于OCR的生成文字图像质量评估

本文采用基于OCR的定量评估方法来衡量生成文字图像的文字可读性与保真度。使用EasyOCR与InternVL3-8B[31]模型分别对生成图像中的文本进行识别，并以原始输入文本作为GT进行比对。通过计算字符错误率(CER)、字符准确率(ACC)以及归一化编辑距离(NED)，评估不同数据子集在字符级一致性上的差异。

1) 评估方法。

字符错误率(CER)：编辑距离/真实字符数，用于衡量识别文本与GT之间的平均差异程度。

字符准确率(ACC)：表示识别结果中正确字符的比例，$ACC=1-CER$。

归一化编辑距离(NED)：对Levenshtein距离进行长度归一化处理，反映整体字符序列的相似度。

2) 实验设置。

评估数据来自MojiTextCN的多个生成子集，包括CleanChar、PicTextLayout以及SceneTextSynth子集。每个子集随机抽取1000张样本图像进行评测。

3) 评估结果。

由表2的评估结果可以看出：CleanChar子集的识别效果最佳，CER仅为0.94%，说明文字清晰、结构规整；SceneTextSynth次之(CER 2.67%)，仍保持较高可读性；而PicTextLayout子集的CER相对较高(4.29%)，这表明在密集文本的图文交错样本下，OCR识别面临更大的挑战。总体而言，生成图像整体可读性良好，能够有效反映文本内容。

表2 基于OCR的生成文字图像可读性评估结果

Tab. 2 Readability evaluation results the generated text images based on OCR

子集	CER/%	ACC/%	NED/%
CleanChar	0.94	99.06	95.33
PicTextLayout	4.29	95.71	95.76
SceneTextSynth	2.67	97.33	98.28

3.4 文字图像超分任务中数据集的验证与分析

为了验证MojiTextCN数据集在文字图像超分辨率(STISR)任务中的有效性，本文基于STISR-TCDM框架，分别使用大规模英文基准数据集TextAtlas5M与本文构建的MojiTextCN进行训练，并在同一中文测试集上进行评估。实验中两者均保持一致的网络结构与训练超参数，唯一变量为训练数据来源。模型性能通过Qwen2.5-VL-7B视觉语言模型进行OCR识别，并计算字符错误率(CER)、字符准确率(ACC)及归一化编辑距离(NED)，以客观衡量模型在恢复文字清晰度与可识别性方面的表现。评估结果如表3和图11所示。

表3 基于Qwen2.5-VL的超分图像OCR性能评估结果

Tab. 3 Performance evaluation results of super-resolution image OCR based on Qwen2.5-VL

输入类型	CER/%	ACC/%	NED/%
高分辨率原图(HR)	0.37	99.63	99.67
低分辨率输入(LR)	11.03	88.97	90.53
超分结果(TextAtlas 5M)	33.09	66.91	65.67
超分结果(MojiTextCN)	**0.37**	**99.63**	**99.67**

图11 图像超分任务示例

Fig. 11 Example of the image super-resolution task

由表3可以看出，在严格的评估设置下，基于MojiTextCN训练的模型在中文测试集上取得了0.37%的极低字符错误率(CER)，该指标达到了与高分辨率原图(HR)相媲美的识别水平(0.37%)。该性能结果表明，MojiTextCN具备出色的像素级清晰度与字形结构准确性，能够为下游模型高水准地实现中文拓扑结构恢复提供可靠的数据支撑。

此外，鉴于当前开源社区尚缺乏大规模的中文密集文本基准数据集，本实验引入了当前最具代表性的大规模英文数据集TextAtlas5M作为参照基线。结果显示，尽管TextAtlas5M具有极高的英文数据质量，但基于其训练的模型在中文超分任务上表现出严重的泛化障碍，其CER不降反升，激增至33.09%。这一现象深刻揭示了中英文字符在形态拓扑与空间密度上存在显著的领域分布差异，印证了在多模态生成领域构建MojiTextCN这样专门针对复杂中文语境的大规模数据集具有重要的研究价值与现实意义。

综合上述多维度的数据集分析与实验验证可以看出，无论是长文本高保真渲染、交错文档布

局规划，还是像素级字形结构恢复，多模态生成模型在处理复杂中文场景时仍面临严峻挑战。MojiTextCN 数据集正是针对这些核心痛点而构建，通过提供大规模的中文密集文本图像与复杂排版样例，为后续相关研究提供了扎实的数据基础与客观的评测基准。

3.5 基于 Chinese-CLIP 的跨模态语义对齐评估

为了精准量化 MojiTextCN 数据集中视觉特征与文本标签之间的跨模态映射质量，本文引入了预训练的 Chinese-CLIP（ViT-B-16）模型[32]作为核心客观评价指标。相比传统的基于规则或单纯依赖文本 N-gram 的匹配指标，CLIP 的视觉-语言双编码器架构能够更深层次地捕获复杂的跨模态语义特征，从而有效克服中文字符形态多变带来的特征对齐难题。CLIP Score 通过计算图文对在共享的多模态嵌入空间中的余弦相似度，能够直观且客观地反映图像内容与描述文本的语义契合度。本文在MojiTextCN的4个子集中分别进行了无放回的均匀随机抽样（每组 N=1 000），以确保评估结果的统计显著性与分布的无偏性。

如表 4 所示，数据集在4个子集上均展现出了高度均衡且优异的语义对齐性能。其中，SceneTextSynth 子集获得了0.362 2的最高分。

表 4 MojiTextCN 与国际主流数据集的语义一致性对比（CLIP Score）

Tab. 4 Comparison of semantic consistency (CLIP Score) between MojiTextCN and international benchmarks

数据集类型	数据集名称	CLIP Score
本研究	CleanChar	0.346 7
	PicTextLayout	0.352 5
	SceneTextSynth	**0.362 2**
	SlideDoc	0.334 8
	MojiTextCN (Overall)	0.349 1
国际基准（baseline）	MS-COCO (Val)	0.312 5
	CC3M	0.281 2
	LAION-400M	0.278 4

这一表现证明，本文提出的数据合成引擎在处理复杂的自然背景时，不仅实现了物理视觉层面上的自然贴合，更在宏观语义层面精准匹配了对应的环境特征，有效规避了合成数据中常见的“语义断层”现象。此外，对于排版最为密集、信息复杂度极高的 SlideDoc 子集，尽管其图像中包含大量细粒度的长文本和多层级视觉元素（这通常会导致全局语义焦点发生一定程度的弥散），但其 CLIP Score 依然达到了0.334 8。该结果证明即使在极端复杂的图文排版场景下，本文数据集的文本内容与视觉主体的核心语义关联依然得到了有效保留。

综合来看，基于 Chinese-CLIP 的量化评估结果印证了 MojiTextCN 作为大规模多模态预训练数据的高纯净度与可靠性。其图文相似度表现证明，该数据集在宏观语义对齐上已经达到了工业级应用标准，能够为多模态大语言模型的训练提供极具价值的、低噪声的语料支撑，从而有效缓解下游模型在复杂文本图像生成与理解任务中容易出现的幻觉问题。

3.6 基于视觉大语言模型的布局评估

为了不拘泥于传统基于规则或纯语义特征的评价指标，进一步量化 MojiTextCN 数据集在复杂排版场景下的视觉自然度与布局合理性，本文引入了先进的视觉大语言模型InternVL3-8B[33]作为零样本评测器，采用VLM-as-a-Judge的前沿评估范式。评测器被要求从“布局合理性（即文字位置是否自然、是否存在严重的视觉重叠）”与“语义和谐度（即文字风格与背景主体是否匹配）”两个核心维度，对随机抽样的4 000个图文对进行1.0~5.0分的盲评打分，结果如表 5 所示。可以看出，CleanChar 与 PicTextLayout 子集均值均超过4.5，验证了本文提出的“生成-解析-校验”闭环机制在可控环境下能有效保障字形准确与排版一致。SceneTextSynth 与SlideDoc子集得分相对偏低（均值与中位数不高于4.0），反映出在真实复杂背景、高密度排版场景下，自动化渲染流程在光照自适应、透视校正与多层元素协同方面仍有优化空间。所有子集的中位数均高于均值，说明评分分布呈轻微左偏，即存在少量低分样本拉低整体均值，但绝大多数样本获得4.0以上高分，数据集主体质量稳定可靠。

表 5 基于视觉大模型InternVL3-8B的MojiTextCN视觉与布局质量量化评估

Tab. 5 Quantitative evaluation of visual and layout quality on MojiTextCN based on InternVL3-8B

数据子集	均值	中位数
CleanChar	4.60	4.80
PicTextLayout	4.58	4.70
SceneTextSynth	3.98	3.80
SlideDoc	3.92	3.80
Overall	4.27	4.50

4 LLM 生成示例与数据集示例

4.1 GPT 调用示例

本文通过GPT-5生成用于SD3.5图像生成的

结构化提示词。明确要求模型在输出中包含场景、主体、风格与环境等视觉要素，生成结果如图 12 所示。可以看出，GPT-5 生成结果符合预期，JSON 格式、UUID 格式以及 prompt 风格均正确，可以用于后续 SD3.5 生图。

GPT4o 输出：

```
[
 {"id": "b6a934e4-36c2-4ff5-a40a-9b1b312fbc4d",
"prompt": "A lone astronaut walks through a vast alien desert under twin suns, surrounded by towering crystal monoliths and drifting energy dust, cinematic lighting and detailed reflective armor."
 },...
]
```

图 12 GPT-4o 生成 SD3.5 提示词的示例

Fig. 12 Example of GPT-4o generating prompts for SD3.5

4.2 SD3.5 生成示例

给出一段提示词：一名孤独的宇航员行走在双日照耀下的广袤外星沙漠中，周围环绕着高耸的水晶巨石和漂浮的能量尘埃，以及电影般的灯光和精细的反光盔甲，利用 SD3.5 并根据 GPT-5 生成的结构化提示词完成图像的生成，如图 13 所示。可以看出，SD3.5 生成了与描述相符合的图片。

对 Stable Diffusion 3.5 输入：

A lone astronaut walks through a vast alien desert under twin suns, surrounded by towering crystal monoliths and drifting energy dust, cinematic lighting and detailed reflective armor.

Stable Diffusion 3.5 输出：

图 13 SD3.5 图像生成的示例

Fig. 13 Example of SD3.5 image generation

4.3 Qwen2.5-VL 文字渲染区域 bbox 标注

本文采用 Qwen2.5-VL 模型进行图像标注，结果如图 14 所示。可以看出，模型输出结果为[263，350，754，568]与目标区域高度一致，表明 Qwen2.5-VL 能够准确理解视觉场景中的指令语义，并完成精确的空间位置标注任务。

对 Qwen2.5-VL 输入：

图片中有一个 noticeboard 的区域，可以让我们渲染一些文字，能帮我标注一下。不要输出其他任何内容，仅仅输出[x1,y1,x2,y2]即可。

Qwen2.5-VL 输出：

[263,350,754,568]

图 14 Qwen2.5-VL 标注示例

Fig. 14 Qwen2.5-VL annotation example

4.4 Qwen2.5-VL 图像描述示例

本文在构造过程中采用了 Qwen2.5-VL-7B 对输入图片生成文字描述，结果如图 15 所示。可以看出，Qwen2.5-VL 输出了针对输入图片的描述(Caption)，同时没有输出关于图中文字的相关内容。

对 Qwen2.5-VL 输入：

Please describe this picture. Just describe the background of the picture. Do not describe the text in the picture. Do not output other content. Just output the description within 50 words.

Qwen2.5-VL 输出：

The background features a bright yellow central panel framed by blue and red geometric shapes....

图 15 Qwen2.5-VL 图像描述示例

Fig. 15 Qwen2.5-VL image description example

4.5 CleanChar 示例

本文生成的 CleanChar 样例如图 16 所示，可以看出样例涵盖了不同长度的纯文本图片。

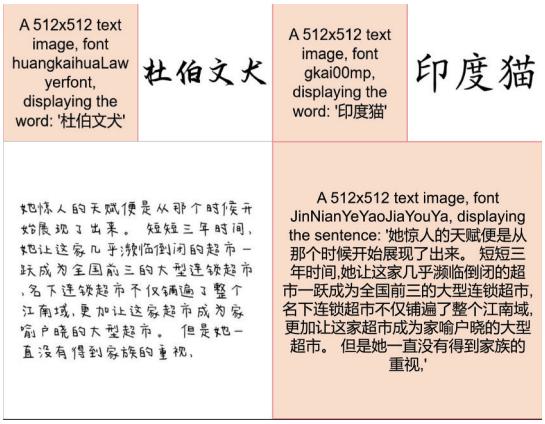

图 16　CleanChar 例图

Fig. 16　CleanChar example

4.6　PicTextLayout 示例

本文生成的 PicTextLayout 子集中具有代表性的图文混排样本如图 17 所示。这些样例广泛覆盖了从短句说明到长段落描述等不同长度的文本分布情况。

图 17　PicTextLayout 例图

Fig. 17　PicTextLayout example

4.7　SlideDoc 示例

SlideDoc 子集中具有代表性的幻灯片样本如图 18 所示，可以看出这些图像覆盖了从简短标题到密集正文等多种文本长度。

图 18　SlideDoc 例图

Fig. 18　SlideDoc example

4.8　SceneTextSynth 示例

SceneTextSynth 子集中具有代表性的合成场景样本如图 19 所示。可以看出，本文利用 Qwen2.5-VL 对图片中的可渲染字体区域进行了准确的标注，在此基础上对这一部分数据中的字符 bbox、字符字体进行了准确标注。

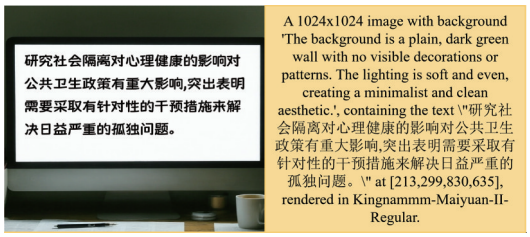

图 19　SceneTextSynth 例图

Fig. 19　SceneTextSynth example chart

5　结论与展望

本文构建并开源了一个面向中文多场景与密集文本图像生成的大规模多模态数据集 MojiTextCN。针对现有模型在复杂中文排版与多样化字体生成上的局限性，本研究采用“合成与真实相结合”的构建策略，数据涵盖了从字符、词句到复杂图文交错及真实场景的多种文本粒度与版式特征。跨模型评估结果表明，尽管当前主流多模态大模型在视觉-语言推理上取得了进步，但在处理中文长文本高保真渲染与复杂交错文档生成任务时仍存在一定局限，而本文提出的 MojiTextCN 丰富了中文密集文本图像领域的数据资源，为相关模型的训练与评测提供了高质量的基准。该数据集的构建有助于提升光学字符识别（OCR）、文本图像生成及跨模态对齐等基础任务的性能，同时也为探索多模态大模型在复杂视觉语境下的鲁棒性与泛化能力提供了数据基础。

未来计划开源 MojiTextCN 的合成方法与核心代码，以支持社区复现并提升轻量化模型的微调与部署效率。同时，本研究将把合成框架拓展至日、韩等非拉丁文字体系，以构建多语言密集文本数据集。此外，通过引入真实业务增量数据并深化版面结构与语义关联标注，将会进一步提升数据集在复杂环境下的表征能力，从而为文档理解与生成研究提供更精确的评测支撑。

利益冲突声明/Conflict of Interests

所有作者声明不存在利益冲突。

All authors declare no relevant conflict of interests.

参考文献:

[1] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation[C]//International Conference on Machine Learning, 2021: 8821-8831.

[2] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10674-10685.

[3] Esser P, Kulal S, Blattmann A, et al. Scaling rectified flow transformers for high-resolution image synthesis [C]//41st International Conference on Machine Learning, 2024: 12606-12633.

[4] Introducing our latest image generation model in the API [EB/OL]. (2025-04-23)[2026-03-25]. https://openai.com/index/image-generation-api/.

[5] Hello GPT-4o [EB/OL]. (2024-05-13)[2026-03-25]. https://openai.com/index/hello-gpt-4o.

[6] Wang P, Bai S, Tan S N, et al. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution[PP/OL]. V2. (2024-10-03)[2026-03-25]. https://arxiv.org/abs/2409.12191.

[7] Beaumont R, Cherti M, Coombes T, et al. Laion-5B: an open large-scale dataset for training next generation image-text models[J]. Advances in Neural Information Processing Systems, 2022, 35: 25278-25294.

[8] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context [C]//European Conference on Computer Vision, 2014: 740-755.

[9] Tuo Y X, Xiang W M, He J Y, et al. Anytext: Multilingual visual text generation and editing [PP/OL]. V4. (2023-12-15)[2026-03-25]. https://arxiv.org/abs/2311.03054v4.

[10] Sidorov O, Hu R, Rohrbach M, et al. Textcaps: a dataset for image captioning with reading comprehension[C]//European Conference on Computer Vision, 2020: 742-758.

[11] Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2315-2324.

[12] Chen J Y, Chen Q F, Cui L, et al. Textdiffuser: diffusion models as text painters[J]. Advances in Neural Information Processing Systems, 2023, 36: 9353-9387.

[13] Wendler C. RenderedText dataset [EB/OL]. (2024-11-02)[2026-03-25]. https://huggingface.co/datasets/wendlerc/RenderedText.

[14] Wang A J, Mao D, Zhang J, et al. Textatlas5m: a large-scale dataset for dense text image generation[PP/OL]. V1. (2025-02-11)[2026-03-25]. https://arxiv.org/abs/2502.07870.

[15] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C]//34th Conference on Neural Information Processing Systems, 2020: 1-12.

[16] Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution[C]//33rd Conference on Neural Information Processing Systems, 2019: 1-13.

[17] Ho J, Salimans T. Classifier-free diffusion guidance [PP/OL]. V1. (2022-07-26)[2026-03-25]. https://arxiv.org/abs/2207.12598.

[18] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 12866-12878.

[19] Yu L J, Lezama J, Gundavarapu N B, et al. Language model beats diffusion: tokenizer is key to visual generation [PP/OL]. V1. (2023-10-09)[2026-03-25]. https://arxiv.org/abs/2310.05737.

[20] Sun P Z, Jiang Y, Chen S F, et al. Autoregressive model beats diffusion: llama for scalable image generation [PP/OL]. V1. (2024-06-10) [2026-03-25]. https://arxiv.org/abs/2406.06525.

[21] Betker J, Goh G, Jing L, et al. Improving image generation with better captions[EB/OL]. [2026-03-25]. https://cdn.openai.com/papers/dall-e-3.pdf.

[22] Yu J H, Xu Y Z, Koh J Y, et al. Scaling autoregressive models for content-rich text-to-image generation [PP/OL]. V1. (2022-06-22)[2026-03-25]. https://arxiv.org/abs/2206.10789.

[23] Rust P, Lotz J F, Bugliarello E, et al. Language modelling with pixels[PP/OL]. V1. (2022-07-14)[2026-03-25]. https://arxiv.org/abs/2207.06991.

[24] Chen J Y, Huang Y P, Lü T C, et al. Textdiffuser-2: Unleashing the power of language models for text rendering [C]//European Conference on Computer Vision, 2024: 386-402.

[25] Ma J, Zhao M J, Chen C, et al. GlyphDraw: seamlessly rendering text with intricate spatial structures in text-to-image generation[PP/OL]. V1. (2023-03-31) [2026-03-25]. https://arxiv.org/abs/2303.17870.

[26] Yang Y, Gui D, Yuan Y, et al. GlyphControl: glyph conditional control for visual text generation [C]// Advances in Neural Information Processing Systems, 2023: 44050-44066.

[27] Veit A, Matera T, Neumann L, et al. Coco-text:

dataset and benchmark for text detection and recognition in natural images [PP/OL]. V1. (2016-01-26) [2026-03-25]. https://arxiv.org/abs/1601.07140.

[28] Hernandez E, Sharma A S, Haklay T, et al. Linearity of relation decoding in transformer language models [PP/OL]. V1. (2023-08-17)[2026-03-25]. https://arxiv.org/abs/2308.09124.

[29] Tuo Y X, Xiang W M, He J Y, et al. Anytext: multilingual visual text generation and editing [PP/OL]. V3. (2023-11-30)[2026-03-25]. https://arxiv.org/abs/2311.03054.

[30] Yang A, Li A F, Yang B S, et al. Qwen3 technical report [PP/OL]. V1. (2025-05-14) [2026-03-25]. https://arxiv.org/abs/2505.09388.

[31] Zhu J G, Wang W Y, Chen Z, et al. InternVL3: exploring advanced training and test-time recipes for open-source multimodal models[PP/OL]. V3. (2025-04-19) [2026-03-25]. https://arxiv.org/abs/2504.10479.

[32] Yang A, Pan J S, Lin J Y, et al. Chinese CLIP: contrastive vision-language pretraining in Chinese [PP/OL]. V2. (2022-11-03)[2026-03-25]. https://arxiv.org/abs/2211.01335.

[33] Chen Z, Wu J N, Wang W H, et al. Intern VL: scaling up vision foundation models and aligning for generic visual-linguistic tasks [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 24185-24198.