中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-15

论文引用格式:Han Han, Yang Chuang, Jing Wei, Wang Qi. A method for text localization and translation in scene images: SceneTrans[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.250621. (韩涵, 杨创, 荆纬, 王琦. 面向自然场景的图像文本寻译方法:SceneTrans[J/OL]. 中国图象图形学报,XXXX:1-15. DOI: 10.11834/jig.250621.)[DOI:10.11834/jig.250621]

面向自然场景的图像文本寻译方法:SceneTrans

韩涵,杨创,荆纬,王琦*

西北工业大学光电与智能研究院,西安　710072

摘　要:**目的**　现有自然场景文本编辑方法聚焦在仅包含单一文本内容图像块的同语编辑,无法实现跨语言文本编辑。针对此问题,本文提出一种端到端的英文到中文自然场景图像文本寻译框架SceneTrans,实现自然场景图像文本的精准定位、英文到中文的跨语言翻译及视觉风格一致的图像文本编辑。**方法**　以MiniCPM-V 2.6多模态大模型为基线,设计基于位置增强提示模板的微调策略实现文本定位与翻译的联合学习;设计中文字形结构编码器实现中文字形先验的精准建模,引入中文字形识别监督机制保障生成图像文本的字形准确性;为解决领域内文本寻译数据匮乏的难题,构建英中配对的SynthTrans专用数据集,收集多种中英兼容字体,合成了20万对包含弯曲、倾斜等复杂字形变换的英中匹配自然场景文本图像。**结果**　实验在文本检测数据集及自建SynthTrans数据集上与其他方法进行了比较。在图像文本定位任务上,所提方法的F1值接近主流模型的检测效果;在图像文本编辑任务上,所提方法的平均结构相似性(structural similarity index measure,SSIM)达0.622、峰值信噪比(peak signal-to-noise ratio,PSNR)达18.51,文本渲染准确率(accuracy,ACC)值提升至71.13%;整体框架具备高效的自然场景图像文本寻译能力。**结论**　本文所提出的端到端寻译框架,实现了自然场景图像文本"定位-翻译-编辑"的一体化流程,可生成字形精确、与原始图像文本风格一致的中文图像文本,突破了传统方法的无法定位及同语编辑的局限,为自然场景图像文本寻译提供了新的解决方案与数据集支撑。

关键词:自然场景图像文本检测;自然场景文本编辑;图像文本寻译;多模态大模型微调;扩散生成

A method for text localization and translation in scene images: SceneTrans

Han Han, Yang Chuang, Jing Wei, Wang Qi*

School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China

Abstract: **Objective** Existing scene text editing methods mainly focus on monolingual editing within image patches containing single textual contents, and thus fail to support cross-lingual text editing. To address this limitation, this paper proposes SceneTrans, an end-to-end English-to-Chinese scene text localization-and-translation framework that enables accurate text localization in natural images, cross-lingual translation from English to Chinese, and visually consistent scene text editing. **Method** Using the multimodal large model MiniCPM-V 2.6 as the baseline, an fine-tuning strategy based on position-enhanced prompt templates is designed to jointly learn text localization and cross-lingual translation. A Chinese glyph structure encoder is introduced to explicitly model Chinese character shape priors, together with a glyph recognition supervision mechanism to ensure the structural accuracy of generated image text. To address the scarcity of cross-lingual scene text translation data, a dedicated English-Chinese paired dataset, SynthTrans, is constructed by collecting diverse

收稿日期:2025-12-09;**修回日期**:2026-08-10

* **通信作者**:王琦 crabwq@nwpu.edu.cn

基金项目:国家自然科学基金项目(62471394, U21B2041)

Supported by: National Natural Science Foundation of China (62471394, U21B2041)

English-Chinese compatible fonts and synthesizing 200,000 paired natural scene text images with complex glyph transformations such as curvature and inclination. **Result** Experiments are conducted on public scene text detection benchmarks and the constructed SynthTrans dataset for comparison with existing methods. In the image text localization task, the proposed method achieves an F1 score comparable to mainstream detection models. In the image text editing task, an SSIM of 0.622 and a PSNR of 18.51 are obtained, while the text rendering accuracy (ACC) is improved to 71.13%. Overall, the framework demonstrates efficient and effective performance for cross-lingual scene image text translation. **Conclusion** The proposed end-to-end seek-and-translation framework accomplishes an integrated "localization-translation-editing" pipeline for natural scene text images, generating Chinese scene text with accurate glyph structures and style consistency with the original text. It overcomes the limitations of previous methods that cannot perform localization or cross-lingual editing, offering a new technical solution and dataset foundation for cross-lingual scene text seek-and-translation in natural images.

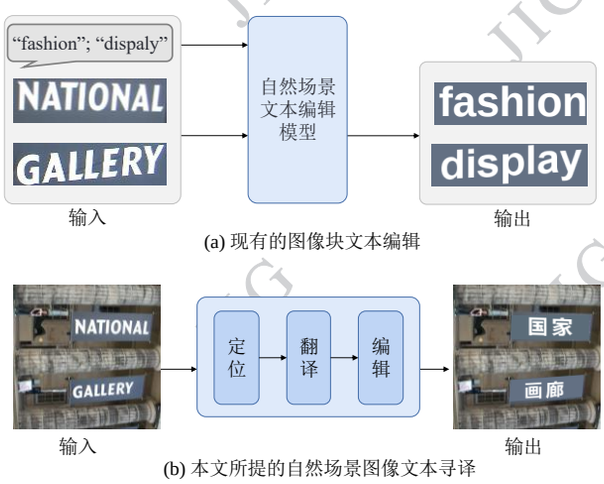
Key words: Scene Text Detection; Scene Text Editing; Text Image Localization and Translation; Large Multimodal Model Fine-tuning; Diffusion Generation

论文引用格式: Han Han, Yang Chuang, Jing Wei, Wang Qi. A method for text localization and translation in scene images: SceneTrans [J/OL]. Journal of Image and Graphics. DOI: 10.118343/jig.250621. (韩涵, 杨创, 荆纬, 王琦. 面向自然场景的图像文本寻译方法:SceneTrans [J/OL]. 中国图象图形学报. DOI: 10.118343/jig.250621.)

0 引 言

随着图像处理技术的飞速发展与智能化需求的持续升级,图像文本作为信息传递的重要载体,已深度融入日常生活中的多个场景,为人们的信息获取与交互带来了极大便利。当前图像文本处理技术已逐步从传统的人工操作走向智能化、自动化处理,用户不再满足于简单的文本提取、识别等基础功能,而是迫切需要能够与图像文本进行交互式处理的解决方案。

自然场景图像文本处理主要针对自然场景图像中所包含的文本信息进行处理。早期研究主要集中于文本检测与识别,并未涉及对图像中文本内容的直接修改。在实际应用过程中,当图像文本存在错误或需更新时,通常依赖人工修图方式,效率低下。为此,研究者提出自然场景文本编辑任务(scene text editing,STE),旨在对图像中指定区域的文本内容进行修改,并保持文本字体、风格及背景纹理的一致性,从而实现视觉层面一致的文本编辑效果,如图1(a)所示。然而,现有的自然场景文本编辑方法仍存在以下两方面局限:其一,当前方法仅支持同语文本替换,例如直接在图像上对某一英文单词修改为另一英文单词(Zeng等,2024),尚未实现跨语言文本编辑;其二,自然场景文本编辑方法通常不具备自动文本定位能力,仅局限于对文本图像块进行编辑,对自然场景图像编辑前需依赖人工预先裁剪待编辑文本区域。



((a) current image patch text editing; (b) proposed scene text image localization and translation)

图1 自然场景图像文本处理任务示意

Fig. 1 Illustration of scene text image processing task

针对上述问题,本文提出自然场景图像文本寻译任务。其目标是在自然场景图像中定位原语言文本区域并将其翻译为目标语言,最终在对应位置生成与原文本风格纹理一致的目标语言文本图像,如图1(b)所示。该任务以端到端的方式实现,输入待处理图像即可直接输出翻译结果,在统一模型中完成文本定位、翻译与图像编辑的融合流程。

现有方法在相关任务上仍然存在不足。在图像文本定位和翻译方面,传统文本检测模型(Liu等,

2020；Zhang等，2023）只能完成文本位置检测，无法直接提供翻译结果；另一方面，多模态大模型（Hurst等，2024；Yao等，2024）虽然能够执行跨语言翻译任务，但无法显式输出文本位置信息。这些方法均难以满足自然场景图像文本寻译的整体需求。在图像文本编辑方面，早期研究主要基于生成对抗网络（generative adversarial network，GAN）（Goodfellow等，2020）。STEFANN（Roy等，2020）提出了一种双模型架构，将保持结构完整性的字体自适应模型与专用风格迁移网络相结合。MOSTEL（Qu等，2023）通过引入显式笔画引导图精确划定编辑区域，并采用半监督学习策略结合合成数据与真实数据来提升模型泛化能力。其他基于生成对抗网络的方法（Krishnan等，2021）也通过架构优化提升文本生成质量。然而，这些方法在处理复杂文本结构和多样化风格时仍面临重大挑战，由于GAN模型容量受限以及准确分解文本样式的挑战，这些方法的泛化能力不可避免地受到限制，常导致生成文本出现不真实特征且视觉效果欠佳。

近年来，许多研究都专注于调整扩散模型进行自然场景文本编辑。DiffSTE（Ji等，2023）使用双编码器设计改进了预训练的扩散模型，其使用了用于渲染精度的字符编码器和用于样式控制的指令编码器。DiffUTE（Chen等，2024）进一步利用基于光学字符识别（optical character recognition，，OCR）的图像编码器作为CLIP（Radford等，2023）文本编码器的替代方案。TextCtrl（Zeng等，2024）则设计基于字形适应注意力机制的推理控制实现对生成字体形状的高质量生成。TextDiffuser（Chen等，2024）和UDiffText（Zhao等，2024）分别利用字符分割掩码进行条件输入和监督标签。然而上述方法在训练过程中只局限于英文字符的字形生成，导致推理时用于中文等其他语言字符的编辑时生成图片均出现字形扭曲，无法生成高质量的中文图像。在其他相关研究领域，已有学者提出面向少样本表盘缺陷图像生成的稳定扩散模型（Zhang等，2025）实现对裂纹形状和位置的精准控制。HCNet（Yang等，2024）提出一种生成高质量遥感图像文本描述方法，其对图像纹理特征提取机制对于中文字形的精细化生成任务具备一定的借鉴与适配价值。除上述研究方法外，GlyphControl（Yang等，2024）和AnyText（Tuo等，2024）等方法通过特征拼接引入间接潜在条件，虽然在单一语言文本生成方面表现良好，但其在跨语言场景下的风格一致性仍存在明显局限。实验结果表明，生成文本的风格往往受限于训练数据中所用字体的类型，难以灵活适应多样化的视觉语境。GlyphMastero（Wang等，2025）等研究尝试结合交叉注意力机制与图像修复技术，实现了对多语言字形特征的细粒度控制。RS-STE（Fang等，2025）在统一框架下整合文本识别与编辑功能，采用基于Transformer（Vaswani等，2017）架构的多模态并行解码器，可同时预测文本内容和风格化图像。但上述模型的编辑过程仍局限于语内替换（如英文到英文、中文到中文），尚未突破跨语言文本编辑的技术瓶颈，无法实现从原语言到目标语言的语义与视觉层面的有效跨越。

综上，为实现端到端的跨语言自然场景图像文本转换，本文提出了一种面向英文到中文的自然场景图像文本寻译框架SceneTrans。该框架包括以下两个组成部分：1）在保留多模态大模型原生翻译能力的基础上，设计基于位置增强提示模板的指令微调策略，显式地输出文本位置信息，为后续目标语言文本图像的编辑提供先验信息支持；2）在传统自然场景文本编辑模型的基础上，设计了中文字形结构编码器和中文识别监督机制，通过汉字字形先验的精准建模，实现中文图像文本的高质量编辑。此外，由于缺乏公开的自然场景图像文本寻译可用数据集，本文构建英文到中文的文本编辑数据集SynthTrans，该数据集基于真实街道场景背景与多样化中英兼容字体，合成了20万对包含弯曲、倾斜等复杂字形变换的英中匹配自然场景文本图像，具有较强的任务挑战性，能够有效验证所提方法的鲁棒性，并为相关领域建立基准。实验表明，所提框架在文本定位任务中实现了定位精度与翻译准确性的协同提升，在文本编辑任务上显著优化了中文字形生成质量与风格一致性，完成了从原语言图像文本定位到目标语言文本图像编辑的端到端闭环，实现了跨语言场景图像文本寻译任务。

1 方　法

1.1 整体结构

本文所提框架的整体流程如图2所示，包含两个核心阶段。首先，对输入的原始英文图像文本进行定位翻译，该阶段显式地输出文本区域的位置及

其对应的中文翻译结果;继而,这些位置信息与翻译文本字符序列被送入文本编辑模块,以在指定区域内生成视觉风格与原始图像一致的中文图像文本;最后,新生成的图像文本块被嵌入原始图像的对应位置,完成全图像文本寻译。

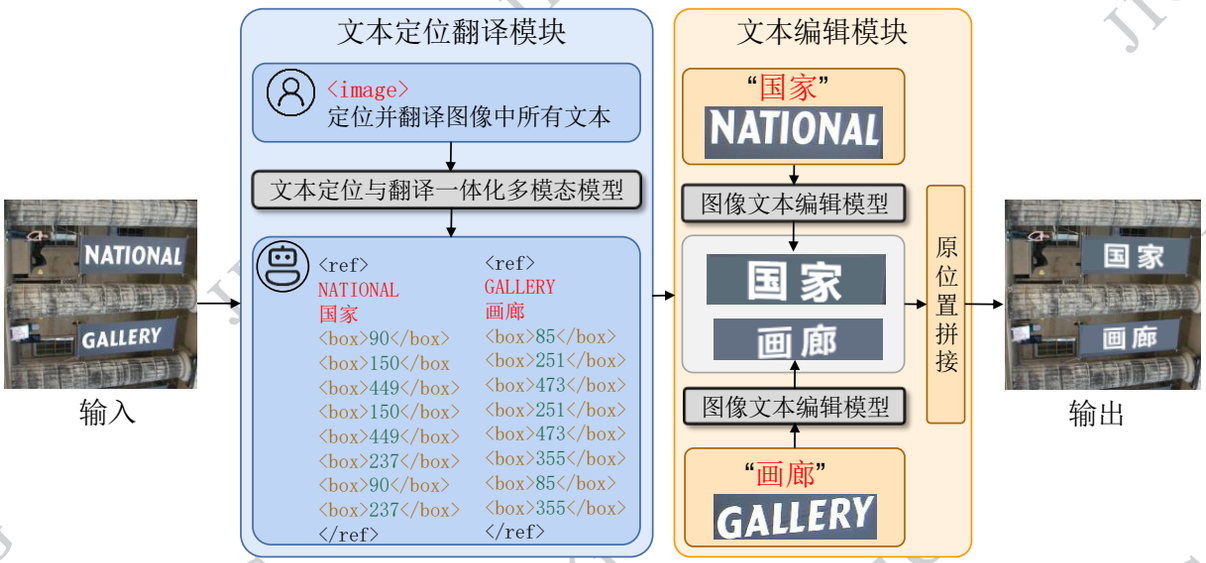


图2　本文所提框架整体流程图

Fig. 2　The work flow of the proposed framework

1.2　文本定位与翻译一体化多模态模型

1.2.1　基线模型

对于图像文本寻译任务,若直接采用传统的专用文本检测与识别模型体系,通常需要构建“文本检测-文本翻译-图像文本编辑”的三阶段级联框架。该框架不仅系统结构复杂、推理链路冗长,而且各阶段之间相互独立,误差易在模块间逐级传递与放大,最终对图像生成质量产生显著影响。在数据依赖与泛化能力上,传统专用文本检测与识别模型通常需要依赖大规模精细标注数据进行训练,往往在数十万级样本规模下才能获得较好的泛化性能。而多模态大模型在大规模跨模态与跨语言语料上进行预训练,已具备较强的通用视觉感知能力与跨语言语义建模能力。

基于上述考虑,本文以多模态大模型为基础进行微调,选用MiniCPM-V 2.6(Yao等,2024)作为基线模型。多模态大模型是一类能同时理解和生成多种模态信息(如图像、文本、语音等)的模型。其输入可为任意模态组合,输出也可为任意模态,具备跨模态理解、生成与推理能力,广泛应用于视觉问答、图文生成、视频理解等任务。MiniCPM-V 2.6是面壁智能发布的开源多模态大模型。在单图理解上,取得了优于GPT-4o mini(Hurst等,2024)、Gemini 1.5 Pro(Team等,2024)等商用闭源模型的表现。其具备强大的OCR识别与文档理解能力,支持超过30种语言的多模态翻译与推理任务,性能优于同类开源模型。尽管MiniCPM-V 2.6在上述任务中表现突出,但其完全不具备显式的文本定位能力,无法直接输出图像中文本实例的精确空间位置坐标。然而,该模型以图像理解为核心的整体架构为引入文本定位机制提供了良好的结构基础与改造空间,使其能够通过适当的结构扩展与任务约束,适配图像文本寻译任务的需求。因此,本文选取其作为基线模型,进一步设计针对文本定位与翻译的一体化方法。

1.2.2　基于位置增强提示模板的微调策略

本文设计基于位置增强提示模板对模型进行微调,模板如图3所示。

模板中用户的提示词说明任务,即定位和翻译图中文本,期望的模型回答要包含文本位置坐标和翻译结果。模板中输入图像的编码表示为

$$T_{img} = < img\_start > + < unk\_token > * num + < img\_end > \tag{1}$$

式中,num表示每张子图对应的视觉标记(token)数量,$< unk_token >$为占位符,后续将被实际图像特

征向量替换，< *img_start* > 与< *img_end* > 分别作为图像嵌入的起始和结束边界标记。视觉编码模块基于ViT(Dosovitskiy等，2020)提取图像特征，替换过程即将其映射为语言编码模块可理解的视觉语义表示。将用户的提示词和期望回答拼接为如下格式

T_{text} =< *用户* > *定位并翻译图像中所有文本* < AI >
< ref > NATIONAL *国家* < box > 90 < /box > …
< ref > GALLERY *画廊*… < /ref >

(2)

式中，< ref >< /ref >为保留标志符，用于区分每一个定框。< box >< /box >也为保留标志符，用于区分每一个坐标值，共8组，分别代表图中文本定位框的左上角坐标(x，y)，右上角坐标(x，y)、右下角坐标(x，y)和左下角坐标(x，y)。与传统多模态大模型仅预测语义结果不同，本文在期望输出中引入了显式的位置标签< box >< /box >，用于精确描述文本在图像中的空间坐标信息，从而实现“文本定位+文本翻译”的联合学习目标。

图3 位置增强提示模板

Fig. 3 Location enhancement prompt template

考虑到语言编码模块存在输入长度限制，大量冗余几何参数将显著增加token占用，不利于保持原有文本理解能力；并且视觉编码器的学习难度随几何参数维度上升而增加。因此，模板中的模型回答部分简化为8组坐标值表达有助于视觉-语言空间的稳定对齐。此外，不同文本实例在方向、倾斜度、比例等上存在差异，统一采用矩形框可保证表达格式一致性，避免复杂四边形或多边形拟合带来的不确定性与训练收敛困难。综合而言，8点坐标在任务表达能力、模型输入规模与训练稳定性之间取得了合理平衡。将位置信息编码为有序标签序列，与对应的翻译结果共同构成统一的输出模板，使语言编码模块在生成目标内容的同时具备空间理解能力。这种位置增强式提示-回答机制有效弥补了原模型无法显式输出文本框位置的局限。最终得到语言编码模块的输入为

$$T_{llm} = bos\_id + T_{img} * num\_slice + \text{tokenizer.encode}\left(T_{text}\right) + eos\_id \tag{3}$$

式中，*num_slice* 表示对图像进行切分后一张子图占用多少个标记。*bos_id* 和 *eos_id* 分别代表文本起始符和结束符。基于LLaMa3(Grattafiori等，2024)的嵌入层构建的语言编码器对该序列进行处理，得到对应的文本向量表示。

值得注意的是，< box >< /box >标签中的坐标与对应图像均来自场景图像文本检测数据集。对于原有数据集中缺失的中文或英文字符标注内容，本文进行了手动补全。

为最大化保持原模型的语言理解与翻译能力，本文在训练策略上使用参数高效微调方案。具体而言，采用低秩自适应(low-rank adaptation，LoRA)(Hu等，2022)的方式。在本节的微调策略中，将LoRA适配器嵌入原生基线模型的视觉编码器部分，训练过程中仅优化LoRA适配器参数及视觉编码模块的顶层参数，同时完全冻结预训练语言模型的全部权重。该策略具有双重优势：其一，大幅降低训练开销与显存占用；其二，有效规避了对语言模型固有语义知识的干扰，在保留原生多语言理解能力的同时，引导模型聚焦于视觉特征的定位感知与结构化输出能力习得，从而实现文本定位与翻译的端到端联合优化。使用所提模板进行微调流程如图4所示。依据预设的位置映射规则，将经适配后的视觉特征向量替换文本嵌入序列中的对应占位符。替换后多模态特征表示被直接输入到大语言模型进行解码。本节选用LLaMa3(Grattafiori等，2024)作为大语言模型，最终输出同时包含位置信息与目标语言译文的结构化向量表示。

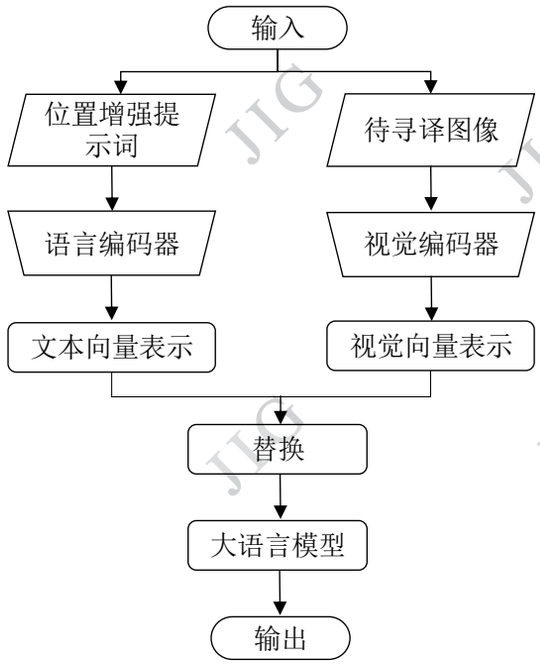


图4　微调流程

Fig. 4　Fine-tuning process

1.3　图像文本编辑模型

1.3.1　基线模型

$$\boldsymbol{I}_{\text{edit}} = G\left(\boldsymbol{C}_{\text{struct}}, \boldsymbol{C}_{\text{style}}\right) \tag{4}$$

式中，$\boldsymbol{C}_{\text{struct}}$和$\boldsymbol{C}_{\text{style}}$分别是文本字形结构特征和文本样式特征。$\boldsymbol{C}_{\text{struct}}$从中文字形结构编码器中获得，具体介绍见1.3.3小节。$\boldsymbol{C}_{\text{style}}$源自文本样式编码器$S$，二者分别实现字形结构和文本样式的细粒度解耦。模型主干采用潜在扩散模型(latent diffusion model，LDM)。LDM利用自编码器ε将输入图像转换为潜在空间表示$\boldsymbol{z}_{\text{edit}}^{0}$。然后训练基于时间条件U-Net(Ronneberger等，2015)的去噪网络ε_θ，在t时间步长估计添加的噪声ε。通过扩散生成器G对先验特征进行积分；对于字形结构特征$\boldsymbol{C}_{\text{struct}}$，由于潜在扩散模型中的U-Net包含自注意力和交叉注意力，且交叉注意力专注于潜在条件与外部条件之间的关联，因此将交叉注意力模块中的键值对替换为换为$\boldsymbol{C}_{\text{struct}}$的线性投影，为文本渲染提供准确的字形指导；对于样式特征$\boldsymbol{C}_{\text{style}}$，将其注入U-Net解码器中，实现对生成结果的高保真样式控制。构建此部分损失函数如下

$$L_{dn} = \mathbf{E}_{\varepsilon(\boldsymbol{x}),c,\varepsilon\sim N(0,1),t}\left[\left\| \varepsilon - \varepsilon_\theta\left(t, \boldsymbol{z}_{\text{edit}}^{t}, c\right)\right\|_2^2\right] \tag{5}$$

式中，$\boldsymbol{x}$为输入图像。

1.3.2　整体网络结构

本文所提图像文本编辑模型整体网络结构如图5所示。目标文本字符输入到中文字形结构编码器中，将目标文本特征与其多种视觉字体对齐，得到目标文本的字形结构特征。原始英文图像经由文本样式编码器解析，通过前景文本纹理迁移与背景提取两个任务共同学习跨语言文本的通用样式表示。文本样式编码器的主干网络S用于提取文本样式特征。该特征经线性投影后分别输入到两个子网络F^t和F^r中。其中，F^t由轻量级编码器-解码器构成，实现原始英文文本纹理的迁移。F^r由基于空间注意力机制的残差卷积块构建，实现原始图像背景的提取。本文使用所构建的SynthTrans数据集对上述编码器进行单独预训练，使得训练后的两个编码器能够提取具备较强泛化能力的文本字形结构特征$\boldsymbol{C}_{\text{struct}}$和文本样式特征$\boldsymbol{C}_{\text{style}}$，实现对文本多维信息的细粒度解耦。这两类特征共同输入至扩散生成网络中，扩散生成过程如图5(a)所示。

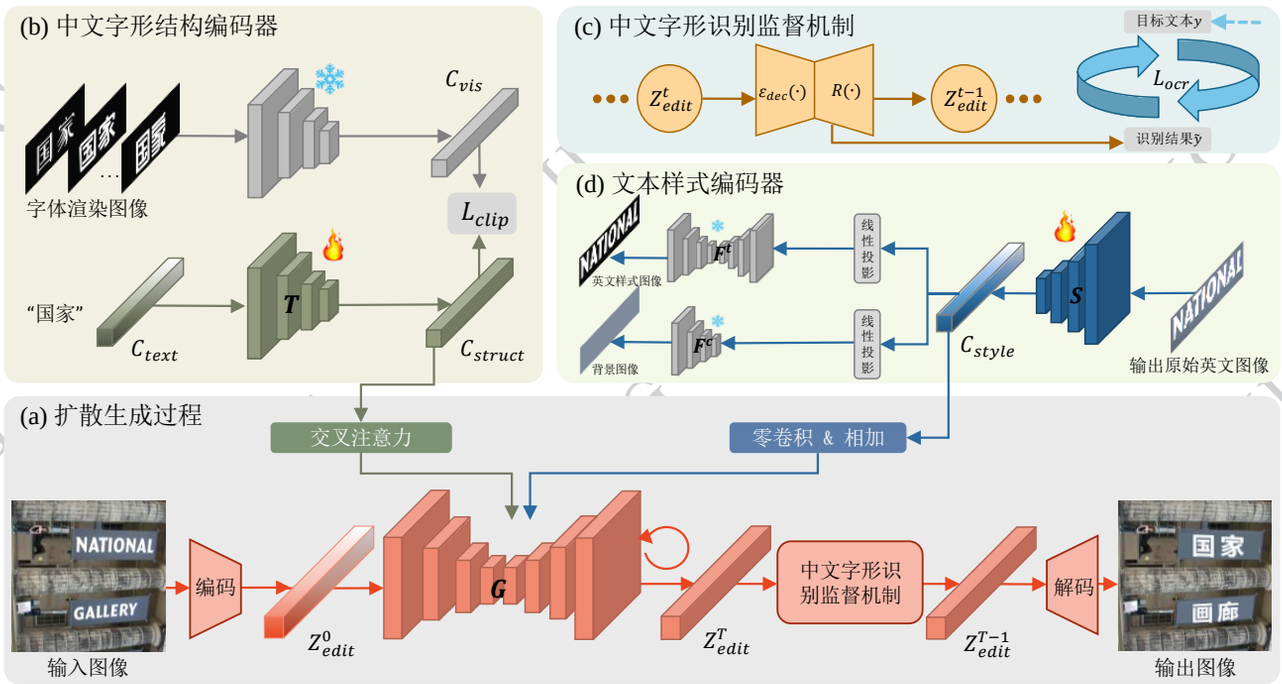
1.3.3　中文字形结构编码器

与自然对象不同的是，场景图像文本具有复杂的字形结构，中文文本由多个笔画构成，较小的笔画差异可以显著改变视觉感知并导致误解。在扩散生成过程中，如果前期没有较好的中文字形先验知识，会导致生成图像中文字结构的扭曲，对于汉字字符，即使最小的变化也可能造成汉字本身含义的改变，因此对编辑精度提出了独特的挑战。对于中文图像文本编辑，文本字形结构编码器需能够对有关字形结构而不是语义信息或者单一图像模板的目标文本进行编码。

为此，本文采用字符级文本编码器将目标文本特征与其视觉字形结构对齐。如图5(b)所示，字符级的目标文本嵌入使用Transformer(Vaswani等，2017)编码器T进行处理，定义为

$$\boldsymbol{C}_{\text{struct}} = T\left(\boldsymbol{C}_{\text{text}}\right) \tag{6}$$

式中，对于文本$\boldsymbol{C}_{\text{text}}$="*国家*"，编码器$T$分别对“国”、“家”进行特征嵌入，生成字形结构特征$\boldsymbol{C}_{\text{struct}}$。同时，将输入的中文文本按照不同字体渲染形成二值化的掩码文本图像。使用冻结的预训练中文场景图像文本识别器提取的视觉特征$\boldsymbol{C}_{\text{vis}}$。字形结构特征和视觉特征的对齐使用CLIP损失。多种字体类型的参与将丰富模型的视觉特征空间，使视觉特征



((a) backbone network; (b) chinese glyph structure encoder; (c) chinese glyph recognize-supervise mechanism; (d) text style encoder)

图5　图像文本编辑模型结构

Fig. 5　Text Image editing model structure

能够更充分地表达字形差异，从而实现更有效的跨模态特征对齐。中文字形结构编码器的训练独立于扩散生成器G的训练，在后续的扩散生成过程，预训练好的中文字形结构编码器直接提取字形结构特征$\boldsymbol{C}_{\text{struct}}$输入到$G$中。

1.3.4　中文字形识别监督机制

在扩散模型的生成过程中，会出现不期望的字形纹理退化。为了控制生成中文字符形状的准确性，本文进一步提出中文字形识别监督机制，如图5(c)所示。具体而言，扩散生成器G生成的中间潜变量$\boldsymbol{z}^{t}_{\text{edit}}$经过解码后，输入到一个预训练中文场景图像文本识别器R，对解码后的图像进行文本识别，得到中文文本识别结果$\tilde{\boldsymbol{y}}$。并与文本标签$\boldsymbol{y}$计算交叉熵损失，用于控制当前中间图像与目标图像文本之间的字形相似性。识别结果$\tilde{\boldsymbol{y}}$为

$$\tilde{\boldsymbol{y}} = R\left(\varepsilon_{dec}\left(\boldsymbol{z}^{t}_{\text{edit}}\right)\right) \tag{7}$$

式中，ε_{dec}为潜变量解码器。

1.3.5　损失函数

对扩散生成器设计三重引导损失，用于其训练监督，总损失为去噪损失L_{dn}、重构损失L_{cons}以及文字识别损失L_{ocr}的总和，表达式为

$$L = L_{dn} + L_{cons} + L_{ocr} \tag{8}$$

式中，去噪损失L_{dn}与1.3.1节所提内容相同。重构损失L_{cons}用于控制视觉效果的一致性，为生成图像$\tilde{\boldsymbol{I}}_{e}$与目标图像$\boldsymbol{I}_{e}$的像素损失

$$L_{cons} = S\left(\boldsymbol{I}_{e},\tilde{\boldsymbol{I}}_{e}\right) \tag{9}$$

文字识别损失为识别结果文本$\tilde{\boldsymbol{y}}$与目标文本$\boldsymbol{y}$的交叉熵损失

$$L_{ocr} = CE\left(\boldsymbol{y},\tilde{\boldsymbol{y}}\right) \tag{10}$$

该损失函数实现去噪生成、视觉重建与语义对齐的协同优化。

2　实　验

2.1　文本定位与翻译一体化多模态模型

鉴于经过微调后的模型相较于基线模型增强了对自然图像中文本的定位能力，本节对该能力的有效性进行验证。

2.1.1　数据集和评估指标

为了验证本文算法对自然场景图像文本的定位能力，使用以下数据集进行实验并对所提方法进行分析：

1）Total-Text数据集（Ch′ng等，2017）共有1255幅训练图像和300幅测试图像，包含了任意形状的文本实例，如水平文本、多方向文本和曲型文本等。数据集的标注为单词级别。

2）MSRA-TD500数据集（Yao等，2012）共有300幅训练图像和200幅测试图像，包含了多种语言、多方向的文本实例，由文本行级标注组成。

3）ICDAR 2015数据集（（Karatzas等，2015）共有1000幅训练图像和500幅测试图像，图像中包含如复杂光照、模糊、低分辨率等极端情况。内含多方向文本的检测，标注为单词级别，由文本4个角点的坐标组成。

评估指标作为量化所提方法的有效性与鲁棒性性能的关键，目标检测领域中常用的精确率（Precision）、召回率（Recall）与F1分数（F1-score）将在本文中使用，计算公式如下

$$\text{Precision} = \left(\frac{TP}{TP + FP}\right) \times 100\% \tag{13}$$

$$\text{Recall} = \left(\frac{TP}{TP + FN}\right) \times 100\% \tag{14}$$

$$\text{F1 - score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

式中，TP、FP、FN分别表示预测结果中的真阳性、假阳性和假阴性样本数。

2.1.2 实验设置

模型微调采用LoRA方法，将其适配模块插入语言编码模块中的自注意力投影层，并仅对视觉编码模块进行参数更新，保持语言编码模块冻结。初始预训练权重均从官方开源加载。最大训练步数为2000，每200步进行一次验证；训练批大小设置为每卡2，验证批大小为每卡1，梯度累积步数为1；优化器使用AdamW，学习率设定为1e-6，并启用梯度检查点以降低显存开销；设置输入图像的最大切片数设为9；采用DeepSpeed Zero-3策略在节点上并行，峰值显存约35GB；梯度累积步数设为32。算法训练在PyTorch框架下运行，所有实验均在2张NVIDIA A100 GPU上完成。

2.1.3 实验结果

本节的核心任务是实现自然图像中的文本定位，即为下游图像文本编辑任务提供所需文本区域的位置信息，因此实验中对模型输出的矩形框计算精确率（P）、召回率（R）及F1分数（F1）三项核心指标，用以量化评估文本定位性能。为全面验证所提方法的有效性，实验选取传统专用文本检测模型进行对比，定量对比结果如表1所示。

实验结果表明，传统文本检测模型经过针对性优化，在规则形状文本的定位任务中已展现出极高的性能水平。其中，ABCNet在MSRA-TD500数据集上实现了85.2%的F1分数，在ICDAR2015数据集上更是达到88.1%的F1分数，成为两类数据集中综合性能最优的传统模型；DBNet与VTD也表现出较高的指标。相较于这些技术成熟的传统模型，本文所提的SceneTrans方法在规则形状文本定位任务中，已能达到与之相近的性能水平。具体而言，在MSRA-TD500数据集上，SceneTrans的精确率为88.1%，仅比最优模型ABCNet低约1.3个百分点，召回率为78.3%，F1分数为82.8%，与主流传统模型的性能差距控制在3个百分点以内；在ICDAR2015数据集上，SceneTrans的F1分数为81.3%，虽略低于ABCNet、VTD、CIC等顶尖传统模型，但与EAST基本持平，且精确率与召回率均保持在80%左右，展现出良好的文本定位稳定性。需要指出的是，本文的目标并非在文本定位精度上全面超越专用的文本检测模型，而是致力于在统一框架下实现“定位-翻译-编辑”一体化的端到端能力。尽管在定位精度上与部分专用模型仍存在一定差距，但在性能相近的前提下，传统方法通常仅针对文本定位单一任务进行优化，而本文所提出的方法对MiniCPM-V 2.6此类原本不能显式输出坐标的多模态模型新增文本定位能力。在完成文本定位的同时，进一步实现了跨语言图像文本翻译。依托端到端的整体架构，所提方法能够同步满足自然场景图像中文本定位与跨语言图像文本寻译的双重需求。

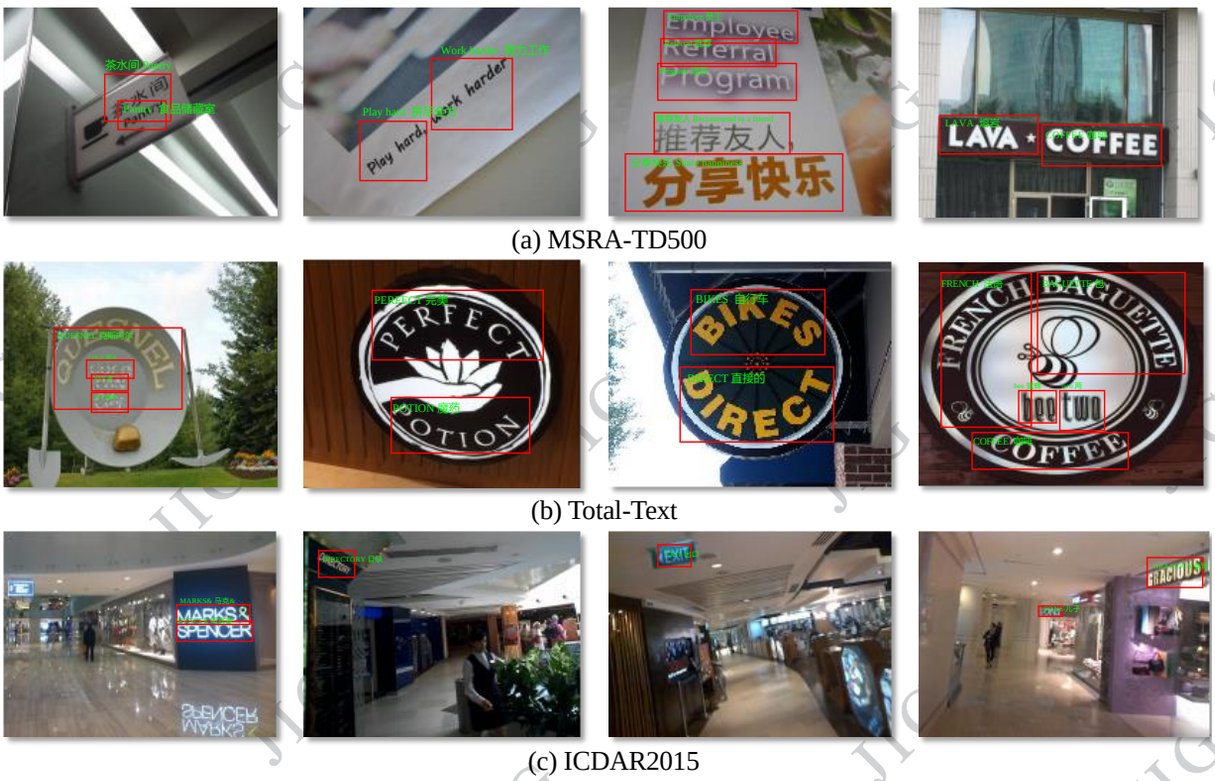
所提方法SceneTrans在弯曲文本等不规则形状的文本定位任务中表现相对有限，其整体性能低于在规则形状文本场景下的定位效果，与顶尖传统文本定位模型相比存在一定差距。在Total-Text数据集上，所提方法的F1分数为72.9%，低于ABCNet、DBNet、EATD等模型约14个百分点，虽显著优于EAST、TextSnake等仅适用于水平或倾斜文本的模型，但与当前先进的弯曲文本定位模型相比，其性能仍存在一定差距，总体处于中等水平。主要原因在于，本文所提方法的视觉编码模块基于通用多模态理解设计，未针对弯曲文本的形态特征进行专项优

表1　在多个公共数据集上进行性能比较，包括MSRA-TD500、ICDAR2015和Total-Text

Table 1　Performance comparisons on multiple public datasets, including MSRA-TD500, ICDAR2015 and Total-Text

方法	MSRA-TD500			ICDAR2015			Total-Text		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
EAST(Zhou等,2017)	87.3	67.4	76.1	83.3	78.3	80.7	49.0	43.1	45.9
TextSnake(Long等,2018)	83.2	73.9	78.3	89.9	80.4	82.6	61.5	67.9	64.6
DBNet(Liao等,2020)	91.5	79.2	84.9	82.1	82.7	85.4	87.1	82.5	84.7
ABCNet(Liu等,2020)	89.4	81.3	85.2	90.4	86.0	88.1	90.2	84.1	87.0
VTD(Zhang等,2023)	89.2	81.5	85.2	90.5	85.5	87.1	–	–	–
EATD(Yang等,2025)	93.3	87.1	90.1	89.8	83.6	86.6	89.4	85.9	87.6
CIC(Han等,2025)	95.1	87.1	90.9	–	–	–	90.2	85.5	87.8
SceneTrans	88.1	78.3	82.8	83.6	79.0	81.3	75.7	70.5	72.9

化，而传统模型通过引入弯曲文本拟合算法、不规则区域分割等专用机制，在该任务中具备天然优势。尽管如此，SceneTrans在弯曲文本定位任务中仍展现出一定的实用价值，由于后续图像文本寻译模块仅需明确文本所在的包含背景信息的矩形位置区域即可进行语义替换、内容修改等操作，无需精确到像素级的弯曲边界拟合，其F1分数稳定在72%以上，所以能够较为准确地定位弯曲文本的外接矩形区域，完全满足下游文本编辑任务的输入要求。



(a) MSRA-TD500

(b) Total-Text

(c) ICDAR2015

((a) results of MSRA-TD500; (b) results of Total-Text; (c) rsesults of ICDAR2015)

图6　本文所提方法对自然场景图像中文本定位结果的可视化

Fig. 6　Visualization of text localization results in scene text images using the proposed method

为直观呈现所提方法在不同数据集上的文本定位性能与适用场景，本节选取多组评估数据集推理结果进行可视化，实验结果如图6所示。在规则文本场景中能够给出稳定且边界清晰的预测；在弯曲文本或极端形变的场景中，模型虽倾向生成相对粗粒度的矩形区域，但仍能覆盖主要文本区域，满足后续文本编辑任务的需求。

2.2 文本编辑模型

编辑后的中文图像文本质量是寻译任务中的关键。因此，本节将验证所提方法对风格一致性与字形可读性的建模能力。

2.2.1 数据集和评估指标

数据集：本文基于SRNet（Wu等，2019）的工作，构建了一个用于英文到中文自然场景图像文本寻译任务的大规模合成数据集SynthTrans，数据集示例如图7所示。该数据集包含训练集SynthTrans-train与验证集SynthTrans-eval。训练集包含20万组中英配对图像文本，用于原语言与目标语言图像文本的风格解耦预训练。每组样本由风格一致的英文图像和其中文翻译文本图像组成，并提供对应的背景图像、目标文本前景图像、原始文本分割图像和目标文本字体图像。为充分覆盖现实场景中的视觉多样性，设置了字体、大小、颜色、空间变换和背景以最大化图像在风格上的不同。在字形上，采用100种同时兼容中英文字符的字体进行渲染，覆盖常见衬线、非衬线、印刷体、展示字体等多种风格，从而捕捉中文字符丰富的笔画结构和英文字符较强的一致性特征；在颜色上，覆盖60种常见色系；在空间变换上，引入了透视、旋转、倾斜、阴影及下划线等多种几何扰动。设置弯曲文本与非弯曲文本的比例为1∶1。所有图像文本尺寸介于256×256至512×512之间，字体大小根据中英文字符自适应调整。背景图像来源于开源网络收集的1050张4K及以上分辨率的实拍街道图片，合成时对其进行了随机裁剪以增强多样性。验证集包含1000组配对图像，在风格控制策略上与训练集保持一致，但在旋转角度、透视比例等空间变换的幅度上设置更为极端，以更好地评估模型在复杂场景下的鲁棒性。

图7 SynthTrans数据集示例

Fig. 7 Example of the SynthTrans dataset

评估指标：在编辑结果的视觉质量评估方面，采用以下指标衡量编辑结果图像和原始图像文本的风格一致性，（1）平均结构相似性（structural similarity index measure，SSIM），综合衡量图像亮度、对比度、结构三个维度；（2）峰值信噪比（peak signal-to-noise ratio，PSNR），计算生成图像与真实图像的像素级绝对误差；（3）特征分布差异（fréchet inception distance，FID），评估生成图像与真实图像的分布相似性。在文本渲染精度评估方面，采用中文文本识别器计算：（1）单词识别准确率（accuracy，ACC）；（2）归一化编辑距离（normalized edit distance，NED）。分别用于衡量生成图像与目标图像在结构一致性、视觉质量和纹理统计一致性方面的差异，并从字符识别角度评估生成文本的可读性与语义准确性。

2.2.2　实验设置

与TextCtrl(Zeng等,2024)的配置类似,采用以ViT(Dosovitskiy等,2020)为主干的风格编码器提取风格特征文本样式,中文字形结构编码器和中文字形识别监督机制中采用的预训练场景文本识别器均为加载预训练中文权重的PaddleOCR(Cui等,2025)。对于扩散生成主干,使用Stable Diffusion(Rombach等,2022) V1-52的预训练权重。模型输入图像调整为256×256,最大目标提示长度设置为24。训练过程使用256的批量大小,学习率为1e−5,总训练轮数为100。本文所提算法在PyTorch框架下运行,所有实验均在2张NVIDIA A100 GPU上完成。

2.2.3　实验结果

将所提方法与以下自然场景文本编辑方法在SynthTans-eval数据集上进行比较:TextCtrl(Zeng等,2024)、DiffSTE(Ji等,2023)、TextDiffuser(Chen等,2024)、AnyText(Tuo等,2024)、Flux-text(Lan等,2025)和TextFlux(Xie等,2025)。定量实验结果如表2所示,定性实验结果如图8所示。

表2　不同方法的文本风格一致性和文本渲染精度评估

Table 2　Text style consistency and text rendering accuracy assessment of different methods

方法	文本风格一致性				文本渲染精度
	SSIM ↑	PSNR (dB) ↑	FID ↓	ACC (%) ↑	NED ↑
TextCtrl	0.572	15.66	**49.78**	17.80	0.5084
DiffSTE	0.418	14.32	116.45	3.62	0.1262
TextDiffuser	0.391	14.69	65.11	5.15	0.1818
AnyText	0.457	14.78	61.73	62.69	**0.8013**
SceneTrans	**0.622**	**18.51**	59.88	**71.13**	0.7837

注:加粗字体为最优值。

图8　不同方法图像文本编辑效果的定性比较

Fig. 8　Qualitative comparison of text image editing results using different methods

在文本风格一致性上,本文所提方法在平均结构相似性(SSIM)、峰值信噪比(PSNR)指标上相较于其他方法取得最优,在特征分布差异(FID)指标方面,本文方法得分为59.88,虽略高于专注于风格迁移的TextCtrl方法的49.78,但显著优于DiffSTE的116.45和TextDiffuse的65.11,并与AnyText方法的61.73处于相当水平。扩散模型在样式控制上具备更高自由度,在某些情况下可能出现不期望的字体形态偏移,使生成结果在文本外观上与原图差异较大,如图8中DiffSTE和TextDiffuser列所示,导致相关评价指标得分偏低。Flux-text则在部分测试样本中出现生成字体颜色与原始英文不一致的问题。相比之下,本文方法通过引入专门的中文字形结构编码器,使模型获得丰富的中文字形先验,并结合原始文本样式特征,在扩散生成过程中实现双重约束,从而显著提升中文文本生成的结构准确性与风格一致性。此外,不同于基于整图区域修复(Inpainting)的AnyText,本文方法仅对裁剪后的文本区域背景进行重建,有效减少全图非文本区域颜色及纹理信息的干扰。

在文本渲染精度上,本文所提方法生成的编辑后的中文图像具有稳健的字形结构表示,在所有方法中取得了优异的拼写准确率。与现有先进方法相比,目标文本的渲染准确率(ACC)至少提升约

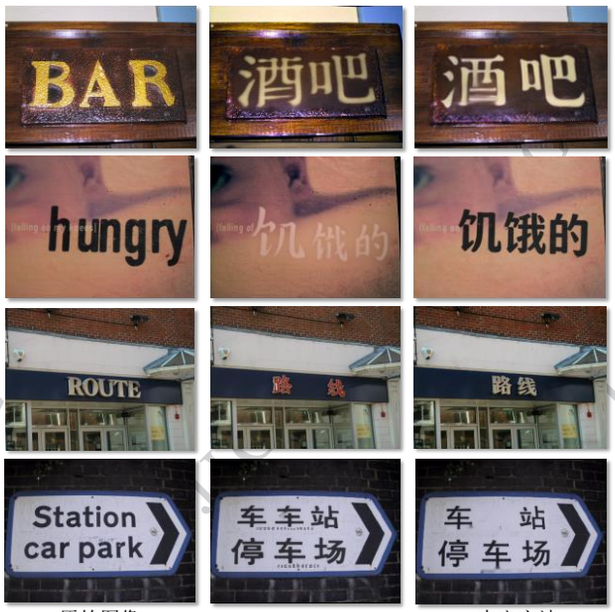
8.84%。在验证数据集中包含复杂背景、多样中文字体形态及颜色、透视等风格变化的任务难度下，仍实现了更高的整体可读性与结构一致性。与此同时，本文方法在归一化编辑距离（NED）指标上取得0.7837的分数，虽然略低于AnyText的0.8013，但已达到可比水平，差距较小。值得注意的是，在NED表现接近的情况下，本文方法在拼写准确率方面领先，表明所提方法在整体字形精确性与语义对齐方面更具稳定性。作为比较，图8中AnyText模型缺乏对中文字形结构的显式建模与控制，导致生成汉字笔画扭曲、结构塌陷等问题，难以适应复杂中文字形变化场景。在定性结果方面，尽管TextFlux和Flux-text能够生成可读的中文字符，但其生成的中文字体与原始英文字体在类型一致性上存在明显差异。本文提出的方法通过多种中英匹配字体的方式对中文字形结构编码器进行预训练，实现了对字体结构先验的显式建模，从而在生成结果的字形结构准确性上表现更优。

为进一步验证所提方法在真实应用场景下的有效性，本节从自然场景文本检测数据集中额外采集了小规模真实场景测试集，与Flux-text进行了可视化对比，部分定性结果如图9所示。结果显示，Flux-text在生成过程中易出现文本内容重复、颜色与原始背景不协调、中文字符语义不一致等问题。相较之下，本文方法在颜色还原度与字形结构准确性等方面表现更为稳定。原因在于本文所构建的Synth-Trans合成数据集中，引入了多种字体并设置弯曲、透视、光照、阴影等复杂几何变换，使模型在训练阶段学习到更丰富的场景变化表示，从而提升了在真实场景下的鲁棒性。

2.2.4 消融实验

为验证每个模块对实验结果的影响，本文进行消融实验分别验证所提模块的有效性。图像文本编辑模型由两部分组成：1）中文字形结构编码器；2）中文字形识别监督机制。在消融设置中，本文首先以仅包含扩散生成主干、不引入任何中文字形相关约束的模型作为基础对照，用以刻画缺乏结构与语义先验时的生成性能下限。在此基础上，分别对仅启用1）或仅启用2）两种情形进行实验，并与完整模型进行对比，结果如表3所示。

实验结果表明，仅采用扩散生成主干时，模型在文本风格一致性与渲染精度上均表现较差，尤其在



原始图像　Flux-text　本文方法

图9　不同方法在真实场景下寻译效果的定性比较

Fig. 9　Qualitative comparison of text image localization and translation results using different methods in real-world scenarios

文本可读性指标上几乎失效，表明在缺乏中文字形结构与语义约束的情况下，扩散模型难以稳定生成结构正确的中文文本。当只引入中文字形识别监督机制时，模型在文本风格一致性与渲染精度方面同样表现较差，说明缺乏对字形局部结构、笔画布局及轮廓几何特征的先验建模，使得模型难以在视觉层面精准还原字体风格，从而导致生成结果出现结构性失真。单独使用中文字形结构编码器时，模型的文本风格一致性和渲染精度均显著提高，从视觉结构层面对生成器提供稳定的先验约束，使得模型能够更准确地复现中文字符的几何形态与风格特征，表明字形结构先验对中文文本生成质量具有关键作用。当同时引入中文字形结构编码器与中文字形识别监督机制时，模型在保持风格一致性基本稳定的同时，文本生成准确率（ACC）和编辑可读性（NED）进一步提升。此现象表明中文形结构编码器提供强结构先验，增强生成器对视觉形态的表达能力；而识别监督则从语义层面校正字符类别，确保生成结果在可读性和语义一致性上进一步收敛。在视觉与语义双重约束的共同作用下，模型能够同时实现高保真的字形渲染和高准确度的字符生成。实验结果显示两模块具有互补优势，验证了本方法在英文到中文自然场景图像文本寻译任务中具备的中文字形编

码能力。

表3　消融实验

Table 3　Ablation study

方法	文本风格一致性			文本渲染精度	
	SSIM ↑	PSNR(dB) ↑	FID ↓	ACC(%) ↑	NED ↑
仅扩散主干	0.273	10.59	91.86	2.99	0.1174
中文字形结构编码器	0.531	17.19	65.96	68.13	0.6856
中文字形识别监督机制	0.372	14.66	77.78	4.70	0.2684
SceneTrans	**0.622**	**18.51**	**59.88**	**71.13**	**0.7837**

注：加粗字体为最优值。

图10　本文方法在真实场景下的失败案例

Fig. 10　Failure cases in real-world scenarios

2.2.5　失败案例分析

本文方法在少量真实场景下可能出现生成失败的情况，如图10所示。主要体现在以下两类场景：1）当原始英文文本为高度艺术化字体时，其字形结构偏离常规分布，模型难以准确解析，进而影响中文生成质量，如“Disney”；2）对于笔画密集、结构复杂的中文字符，生成过程中可能出现笔画粘连，导致语义可读性下降，如“英语餐厅”。

尽管在上述复杂场景下仍有改进空间，本文方法在大多数常见自然场景中能够稳定完成文本定位、跨语言翻译与视觉一致性编辑，可有效覆盖日常生活中的主要应用场景，具备一定的实用价值。

3　结　论

本文提出了端到端英文到中文自然场景图像文本寻译方法，旨在同时解决自然场景文本的定位、翻译与视觉一致性编辑问题。所提出框架包含两大核心设计：1）基于位置增强提示模板的指令微调策略，通过引入显式坐标标注增强多模态大模型的文本定位能力；2）构建面向中文的字形结构编码与识别监督机制，实现高保真中文生成与风格迁移。此外，本文构建了英文-中文匹配的SynthTrans数据集，为跨语言场景文本编辑提供支持。实验结果表明，所提方法在寻译任务中取得优异表现，可生成字形结构精确、风格一致的中文图像文本，显著提升跨语言自然场景图像文本处理能力。

但是，本文提出的模型仍然存在一些需要改进的问题。本文提出的文本位置定位翻译提取模块目前仅能输出矩形框，对于多方向文本或复杂布局的拟合效果不佳，后续拟通过设计多方向锚边框机制进一步提升大模型进行的定位精度。其次，文本编辑模型在生成复杂结构的中文字符时，对笔画细节和空间布局的还原仍有提升空间，尤其在生僻字或艺术字体上的表现不够稳定，未来计划引入更细粒度的字形先验知识库和结构增强的生成策略。

参考文献（References）

Chen H, Xu Z, Gu Z, Lan J, Zheng X, Li Y, et al. 2023. DiffUTE: universal text editing diffusion model//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: NIPS: 63062 - 63074 [DOI: 10.5555/

3666122.3668875]

Chen J, Huang Y, Lv T, Cui L, Chen Q and Wei F. 2023. TextDiffuser: diffusion models as text painters////Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: NIPS: 9353 - 9387 [DOI: 10.5555/3666122.3666532]

Ch'ng C K and Chan C S. 2017. Total-text: a comprehensive dataset for scene text detection and recognition//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Kyoto, Japan: IEEE: 935-942 [DOI: 10.1109/ICDAR.2017.157]

Cui C, Sun T, Lin M, Gao T, Zhang Y, Liu J, et al. 2025. PaddleOCR 3.0 technical report [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2507.05595

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2020. An image is worth 16×16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. Vienna, Austria: ICLR: 1 - 21 [DOI: 10.48550/arXiv.2010.11929]

Fang Z, Lyu P, Wu J, Zhang C, Yu J, Lu G, et al. 2025. Recognition-synergistic scene text editing//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13104 - 13113. [DOI: 10.1109/CVPR52734.2025.01223]

Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et. al. 2024. The Llama 3 herd of models [EB/OL]. [2025-02-14]. https://doi.org/10.48550/arXiv.2407.21783

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2020. Generative adversarial networks. Communications of the ACM, 63(11): 139-144. [DOI:10.1145/3422622]

Han X and Wang Q. 2025. Compensating for the incomplete with the complete: an efficient scene text detector. IEEE Transactions on Circuits and Systems for Video Technology, 35(12): 12096-12108 [DOI: 10.1109/TCSVT.2025.3588711]

Hu E J, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang L, et al. 2021. LoRA: low-rank adaptation of large language models//Proceedings of the International Conference on Learning Representations. Virtual: ICLR: 1 - 15 [DOI: 10.48550/arXiv.2106.09685]

Hurst A, Lerer A, Goucher A P, Perelman A, Ramesh A, Clark A, et al. 2024. GPT-4o system card [EB/OL]. [2025-02-14]. https://arxiv.org/pdf/2410.21276

Ji J, Zhang G, Wang Z, Hou B, Zhang Z, Price B, et al. 2023. Improving diffusion models for scene text editing with dual encoders [EB/OL]. [2023-04-12] https://arxiv.org/pdf/2304.05568

Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, et al. 2015. ICDAR 2015 competition on robust reading//Proceedings of the 13th International Conference on Document Analysis and Recognition. Tunis, Tunisia: IEEE: 1156-1160 [DOI: 10.1109/ICDAR.2015.7333942]

Krishnan P, Kovvuri R, Pang G, Vassilev B and Hassner T. 2023. TextStyleBrush: transfer of text aesthetics from a single example. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(7): 9122-9134 [DOI:10.1109/TPAMI.2023.3239736]

Lan R, Bai Y C, Duan X, Li M X, Jin D Y, Xu R, et al. 2025. Flux-text: a simple and advanced diffusion transformer baseline for scene text editing [EB/OL]. [2025-11-20]. https://arxiv.org/abs/2505.03329

Liao M H, Wan Z Y, Yao C, Chen K and Bai X. 2020. Real-time scene text detection with differentiable binarization//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press: 11474-11481 [DOI: 10.1609/aaai.v34i07.6812]

Liu Y L, Chen H, Shen C H, He T, Jin L W and Wang L W. 2020. ABCNet: real-time scene text spotting with adaptive bezier-curve network//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9806-9815 [DOI: 10.1109/CVPR42600.2020.00983]

Long S, Ruan J, Zhang W, He X, Wu W and Yao C. 2018. TextSnake: a flexible representation for detecting text of arbitrary shapes//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 19 - 35 [DOI: 10.1007/978-3-030-01216-8_2]

Qu Y, Tan Q, Xie H, Xu J, Wang Y and Zhang Y. 2023. Exploring stroke-level modifications for scene text editing//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 2119-2127 [DOI: 10.1609/aaai.v37i2.25305]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual: ACM: 8748 - 8763 [DOI: 10.48550/arXiv.2103.00020]

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10684 - 10695 [DOI: 10.1109/CVPR52688.2022.01042]

Ronneberger O, Fischer P and Brox T. 2015. U-net: Convolutional networks for biomedical image segmentation//International Conference on Medical Image Computing and Computer-Assisted Intervention. Daejeon, South Korea: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Roy P, Bhattacharya S, Ghosh S and Pal U. 2020. STEFANN: scene text editor using font adaptive neural network//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13225-13234 [DOI: 10.1109/CVPR42600.2020.01324]

Team G, Georgiev P, Lei V I, Burnell R, Bai L, Gulati A, et al. 2024. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context [EB/OL]. [2025-02-14].

https://arxiv.org/pdf/2403.05530

Tuo Y，Xiang W，He J Y，Geng Y，Xie X. 2023. AnyText：multilingual visual text generation and editing［EB/OL］［2023-11-06］ https://arxiv.org/pdf/2311.03054

Vaswani A，Shazeer N，Parmar N，Uszkoreit J，Jones L，Gomez A N，et al. 2017. Attention is all you need//Advances in Neural Information Processing Systems. Montreal，Canada：NIPS：5998 - 6008［DOI：10.48550/arXiv.1706.03762］

Wang T，Liu T，Qu X，Wu C，Liu L and Hu X. 2025. GlyphMastero：a glyph encoder for high-fidelity scene text editing//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：28523 - 28532［DOI：10.1109/CVPR52734.2025.02656］

Wu L，Zhang C，Liu J，Han J，Liu J，Ding E，et al. 2019. Editing text in the wild//Proceedings of the 27th ACM International Conference on Multimedia. Nice，France：ACM：1500-1508［DOI：10.1145/3343031.3350929］

Xie Y，Zhang J L，Chen P Y，Wang Z Y，Wang W H，Gao L W，et al. 2025. TextFlux：an ocr-free dit model for high-fidelity multilingual scene text synthesis［EB/OL］.［2025-05-23］. https://arxiv.org/abs/2505.17778

Yang C，Han X，Han T，Han H，Zhao B X and Wang Q. 2025. Edge approximation text detector. IEEE Transactions on Circuits and Systems for Video Technology，35(9)：9234-9245［DOI：10.1109/TCSVT.2025.3558634］

Yang Y，Gui D，Yuan Y，Liang W，Ding H，Hu H，et al. 2023. GlyphControl：glyph conditional control for visual text generation//Proceedings of the 37th Conference on Neural Information Processing Systems. New Orleans，USA：NIPS：［DOI：doi. org/10.48550/arXiv.2305.18259］

Yang Z G，Li Q，Yuan Y and Wang Q. 2024. HCNet：hierarchical feature aggregation and cross-modal feature alignment for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing，62：1-11，［DOI：10.1109/TGRS.2024.3401576］

Yao Y，Yu T，Zhang A，Wang C，Cui J，Zhu H，et al. 2024. MiniCPM-V：a gpt-4v level mllm on your phone［EB/OL］.［2025-02-14］. https://arxiv.org/pdf/2408.01800

Yao C，Bai X，Liu W Y，Ma Y and Tu Z W. 2012. Detecting texts of arbitrary orientations in natural images//Proceedings of 2012 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Providence，USA：IEEE：1083-1090［DOI：10.1109/CVPR.2012.6247787］

Zeng W，Shu Y，Li Z，Yang D and Zhou Y. 2024. TextCtrl：diffusion-based scene text editing with prior guidance control. /Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver，Canada：NIPS：138569-138594［DOI：10.52202/079017-4396］

Zhang J B，Feng W，Zhao M B，Yin F，Zhang X Y and Liu C L. 2023. Video text detection with robust feature representation. IEEE Transactions on Circuits and Systems for Video Technology，34(6)：4407-4420［DOI：10.1109/TCSVT.2023.3337247］

Zhang K，Sheng X，Xiao Y J，Yang J Y，Chen M J and Ren Z H. 2025. Stable diffusion model for few-shot generation of meter defect images. Journal of Image and Graphics，30(11)：3451-3464（张珂，盛鑫，肖扬杰，杨济远，陈美娟，任泽华. 2025. 面向少样本表计缺陷图像生成的稳定扩散模型. 中国图象图形学报，30(11)：3451-3464）［DOI：10.11834/jig.240777］

Zhao Y and Lian Z. 2024. UDiffText：a unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models//Proceedings of the European Conference on Computer Vision. Milan，Italy：Springer：217-233［DOI：10.1007/978-3-031-72751-1_13］

Zhou X Y，Yao C，Wen H，Wang Y Z，Zhou S C，He W R，et al. 2017. EAST：an efficient and accurate scene text detector//Proceedings of 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu，USA：IEEE：2642-2651［DOI：10.1109/CVPR.2017.283］

作者简介

韩涵，男，硕士研究生，主要研究方向为计算机视觉和自然场景图像文本处理。E-mail：han_han@mail.nwpu.edu.cn

王琦，通信作者，男，教授，主要研究方向为计算机视觉、模式识别和遥感图像解译。E-mail：crabwq@nwpu.edu.cn

杨创，男，助理研究员，主要研究方向为计算机视觉和自然场景图像文本处理。E-mail：omtcyang@gmail.com

荆纬，男，博士研究生，主要研究方向为计算机视觉和遥感图像解译。E-mail：wei_adam@mail.nwpu.edu.cn