智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

AI图像生成的精细化可控:结构化与非结构化提示词及其轻量微调的对比研究

张伟,罗亚远

引用本文:
张伟, 罗亚远. AI图像生成的精细化可控:结构化与非结构化提示词及其轻量微调的对比研究[J]. 智能系统学报, 2026, 21(4): 908-918.
ZHANG Wei, LUO Yayuan. Fine-tuned control of AI image generation: a comparative study of structured and unstructured prompts and lightweight model fine-tuning[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(4): 908-918.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202509016

DOI: 10.11992/tis.202509016
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20260422.1627.004

AI 图像生成的精细化可控：结构化与非结构化提示词及其轻量微调的对比研究

张伟[1]，罗亚远[2]
（1. 卡罗琳斯卡学院 临床科学系，斯德哥尔摩 胡丁厄 14152；2. 辽宁工程技术大学 软件学院，辽宁 葫芦岛 125000）

摘　要：针对文本到图像（T2I）生成中结果不稳定、难以控图的痛点，结合生成式人工智能（artificial intelligence generated content, AIGC）技术中的提示词和微调模型，通过分析不同格式的提示词及其微调模型对控图能力的影响，建立“结构化提示词及 LoRA 微调”可复现生图模型以解决当前包装设计过程中 LLM(large language model) 绘图对生成结果的精细化可控等问题。以 FLUX-dev 为基座，围绕 100 个主题构建提示词结构化与否、LoRA 微调与否的四象限数据集；以 CLIPScore 与自动化美学评分为基准，配合线性混合效应与配对统计，并辅以同构图的可视化对照与全局—细节双重检验。在此基础上，增加 DALL·E3 跨模态模型以验证结构化语义一致性的普适性。结果表明：结构化主要稳定全局版式与层级关系，LoRA 打磨边缘与材质细节；针对基础文本生图模型，结构化提示词能稳健提升图文符合度；LoRA 微调能显著提升美学评分。此外，不同大模型侧重不同，新版本并不必然优于旧版本，匹配符合版本的高质量提示词微调组合尤为关键。本结论可为未来的提示词工程和低成本模型定制提供理论基础和实践参考。

关键词：生成式人工智能；提示词工程；结构化提示；LoRA 模型；CLIPScore；可控生成；美学评分；智能包装设计
中图分类号：TP391.41; TB482　　**文献标志码**：A　　**文章编号**：1673−4785(2026)04−0908−11

中文引用格式：张伟，罗亚远. AI 图像生成的精细化可控：结构化与非结构化提示词及其轻量微调的对比研究 [J]. 智能系统学报，2026, 21(4): 908–918.
英文引用格式：ZHANG Wei, LUO Yayuan. Fine-tuned control of AI image generation: a comparative study of structured and unstructured prompts and lightweight model fine-tuning[J]. CAAI transactions on intelligent systems, 2026, 21(4): 908–918.

Fine-tuned control of AI image generation: a comparative study of structured and unstructured prompts and lightweight model fine-tuning

ZHANG Wei[1], LUO Yayuan[2]
(1. Department of Clinical Science, Karolinska Institutet, Huddinge 14152, Sweden; 2. School of Software Engineering, Liaoning Technical University, Huludao 125000, China)

Abstract: To mitigate instability and poor controllability in text-to-image(T2I) generation, this study integrates prompt engineering and lightweight fine-tuning in artificial-intelligence-generated content(AIGC) systems. We examine how prompt formats and fine-tuned models jointly affect controllability of image generation and propose a reproducible pipeline built on “structured prompts + LoRA fine-tuning.” The method addresses the challenge of achieving precise control in LLM-assisted image generation for packaging design workflows. Building on FLUX-dev, we create a four-quadrant, 100-theme dataset that systematically varies prompt structure and LoRA configurations. The proposed method is evaluated using CLIPScore and automated aesthetic ratings. Statistical significance is examined through linear mixed-effects models and paired tests, while visual quality is assessed through isomorphic renderings and dual global–detail inspections. To assess the generalizability of the structured prompt strategy, we also evaluate the dataset with the cross-modal DALL·E3 model. Results show that structured prompts mainly stabilize overall layout and hierarchy, whereas LoRA chiefly sharpens edges and material details. For base T2I models, structured prompts consistently improve image–text alignment, while LoRA substantially elevates aesthetic scores. Different large models exhibit distinct strengths, and newer versions do not inherently surpass older ones; the key is pairing high-quality prompts with LoRA settings suited to each version. These findings provide both theoretical insights and practical guidance for prompt engineering and cost-effective model adaptation.

Keywords: artificial intelligence generated content; prompt engineering; structured prompts; LoRA model; CLIPScore; controlled generation; aesthetic scoring; intelligent packaging design

随着人工智能（artificial intelligence, AI）技术的不断迭代，社会进入新阶段，艺术创作能力被有效下沉，普通个体亦可凭人工智能将其艺术构

收稿日期：2025−09−09.　**网络出版日期**：2026−04−23.
基金项目：Shenzhen Science and Technology Innovation Program（JCYJ20240813143102004）.
通信作者：张伟. E-mail: wei.zhang.2@ki.se.

想实现为成品[1]。人工智能不仅能提高从业者的创造力和工作效率[2-3], 还在 AI 工具的使用中培养用户与这些工具互动的技能[4]。其中, 生成式人工智能(artificial intelligence generated content, AIGC)领域的文本生成图像技术迅速发展, 成为智能设计的重要研究方向。该技术以“提示词工程”[5-6]为基础, 通过输入自然语言(Prompt), 由大语言模型(large language model, LLM)生成对应的图像。LLM 文本–图像模型(DALL· E3、Stable Diffusion、Midjourney)降低了视觉内容生成门槛, 但输出的随机性与可控性不足限制其工程应用化。

为了使生成图像更符合预期并具备稳定性, 提示词工程与参数高效微调成为两条互补的关键路径。提示词的结构化设计与表达范式决定语义约束的显式性, 直接影响生成结果的质量与匹配精度; 低秩适配微调(low-rank adaptation, LoRA)作为对预训练基座的定向优化, 可在特定任务场景中增强模型的控图能力与稳健性, 使模型在不改写提示词的条件下更准确地理解并执行指令。然而, 现有研究仍缺乏对结构化与非结构化提示的系统量化对比、与评测指标相匹配的可复现实验模板以及资源受限环境下提示–微调协同的实践范式。

为此, 本文基于 FLUX-dev 搭建统一的生图与评测流水线, 围绕“结构化提示能否显著提升语义匹配度、两类提示在视觉质量上的差异机理以及轻量微调能否在不改写提示词时带来统计显著的改进”开展系列对比实验。采用“四象限”设置(是否结构化提示词、是否使用 LoRA 微调), 在固定采样策略与分辨率的前提下控制变量, 使用 CLIPScore 与审美评分等指标进行量化评估, 并给出可复现模板, 推动生成式人工智能在资源受限场景中的稳健应用, 同时为设计实务(包装或平面设计)提供方法论依据与效率提升路径。

1　文本到图像生成可控性的相关研究

1.1　文生图模型(Text-to-image)

不同阶段的代表性文生图模型沿时间线呈现出范式更迭与能力跃迁。早期以 AlignDRAW 等为开端, 随后生成对抗网络(generative adversarial network, GAN)体系奠定了生成学习的基础: Goodfellow 等[7] 提出对抗训练框架, 使生成器在判别器约束下逼近真实分布, 并衍生出在网络结构与训练策略上的改进[8-10]。然受梯度消失、模式坍塌与语义对齐不足等制约, 即便引入 Wasserstein GAN 缓解不稳定性[11], 早期方法在高质量成像与图文一致性上存在上限, 生成对抗网络及条件生成对抗网络(conditional generative adversarial network, CGAN)为文生图技术奠基, 但难以支撑细粒度语义与稳定可控的生成[12-14]。

进入近年, 跨模态对齐+扩散式生成成为主流路线。代表性系统包括 OpenAI 的 DALL·E2[15]、Google Brain 的 Imagen[16] 及商业化的 Midjourney 等, 它们将自然语言映射为语义一致的图像, 并在分辨率与质感上显著提升。对比语言–图像预训练模型(contrastive language-image pre-training, CLIP)的提出为文本–图像构建共享语义空间, 成为条件建模与评测的里程碑[17]; 结合扩散模型的稳定训练特性[18] 与潜空间表征的高效推理[19], 新一代方法进一步兼顾生成效率与视觉质量, 推动文生图从“可生成”走向“可用、可控”。

在文本驱动编辑方面, 研究形成两条互补路径: 其一是基于矢量量化生成对抗网络(vector quantized generative adversarial network, VQGAN)的全局重生成, 通过“噪声注入–再生成”等策略在保持主体语义的同时完成风格内容迁移[20-21]; 其二是局部受限编辑, 借助用户遮罩或结构条件对指定区域进行精细控制, 如 ControlNet 在不改动主干的前提下注入边缘/深度/姿态等条件图, 实现几何与布局的显式约束[22]。与此同时, 仅依赖文本修改亦可触发可观的编辑效果, 用于考察提示词变体对生成的边际影响[23]。大量证据表明, 提示词的词项与序列会显著影响稳定性与质量, 因而提示工程成为重要研究方向; 这也为后续关于结构化与非结构化提示及其与轻量微调协同机制的系统量化研究奠定了问题背景。

1.2　提示词工程

以文本指令引导生成式模型产出的实践通常被称为“提示工程(prompt engineering)”[24]。该术语最初由 OpenAI 在开发 GPT-3 语言模型的创意写作功能时引入[25]。现有研究表明, 提示词的组织与表达会显著影响语义对齐与感知质量, 掌握有效的提示策略可显著提升生成结果的可预期性与美学吸引力[5-6], 并推动更广泛的创作者群体参与到 AI 艺术实践中[26]。学术与工业界已涌现多类辅助工具: 一类利用预训练模型进行提示扩展(如 magicPrompt)以丰富输入语义; 一类基于模板/推荐为用户提供候选提示(如 novelAI prompt tool、Midjourney prompt helper); 另一类通过交互式重写与链式增改支持人机协同优化(如 opal、promptify、promptMagician)。尽管如此, 面向“逐步迭代–可解释推荐–前后对比评估”的闭环仍未完备, 提示工程在不同模型与任务间的可复现规

范亦有待统一，这正构成本文工作的出发点。

2　文本到图像生成可控化研究方法

2.1　CLIP 模型

2021 年，OpenAI 提出的 CLIP 模型成为文本生成图像研究的重要里程碑[27]。本文以 CLIP 作为核心测评器，其在大模型视觉–语言语料上的对比预训练建立了统一语义空间，具备跨模态联合理解与表征能力[28]。在图像美学评估（image aesthetic assessment, IAA）中对美学与情感维度亦有良好敏感性[29]，并在弱或无标注设定下较基于 ImageNet 预训练的模型展现更高的准确性与适用性[30]。具体实现上，将评测拆分为语义一致性与美学质量两条互补分量。其一，语义一致性采用标准 CLIPScore；其二，美学质量在 CLIP 表征上通过加权融合形成综合审美分。该组合能够稳定区分高低美学品质样本，作为对语义一致性的有效补充。

基于 CLIPScore 等无参考指标刻画图文语义一致性，采用 CLIP ViT-L/14(336) 计算 CLIPScore。

$$\mathrm{CLIPScore}(I,T)=\frac{<f_I(I),f_T(T)>}{\|f_I(I)\|\cdot\|f_T(T)\|}$$

式中 f_I、f_T 为图像/文本编码器，用以计算得分。

2.2　基于大语言模型的结构化提示词

提示词的质量一定程度决定了生成图片的质量。为研究提示词能否提高 AIGC 的稳定性、图片质量以及增强对应图片的可复现性，本文采用基于大语言模型的结构化提示词。先定义受控槽位（schema），再以大语言模型在约束解码下完成字段填充与统一回归，最后编译为可读或机读两路指令并绑定超参。该流程将自由文本分层、逻辑化，显著降低语义歧义，提升复杂场景下的构图稳定与风格一致性，为后续测评保持输入模式一致[31]。具体来说，结构化提示词分为背景设置、排版布局以及细节优化 3 部分。相较平铺提示，结构化提示词在多主体与强约束上能显著减少“版式飘逸、细节逸散”等问题，提高输出质量与统计可比性。

2.3　研究设计对照协议

本研究采用 2×2 因子成对设计：因素一为提示词结构化 $S\in\{\mathrm{struct,plain}\}$，因素二为 LoRA 微调 $L\in\{\mathrm{base,lora}\}$。其中，struct 为结构化提示词，plain 为非结构化提示词，base 为基础模型，lora 为使用 LoRA 微调。对每一主题 $t\in\{1,2,\cdots,T\}$（本研究 T=100）生成四象限图像。其定义范围为 $I_{t,s,l}$，$s\in S,l\in L$。

配对策略以“按主题配对”消除题材差异。为提高外部可证伪性，提出跨模态外部验证，仅改变“是否结构化”，其他保持一致。

2.4　数据与清洗

主题集合跨物体、场景、光照、人物与构图复杂度，覆盖 100 个主题，每个主题在四象限下各生成 1 张图像。以生成失败、严重遮挡、敏感或不适宜内容（not safe for work, NSFW）、文本越界（位置、字形、乱码）、主题缺失或主从关系颠倒、分辨率或纵横比例异常、语义越题作为失效判据。

采用双评机制，Cohen’s k 一致性为

$$k=\frac{p_0-p_e}{1-p_e}$$

当 k<0.75 时，引入第 3 方评审仲裁。对配对差异值 ΔC、ΔA 采用中位数绝对偏差（median absolute deviation, MAD）规模做文件异常检测。

$$\frac{|x-\mathrm{median}(x)|}{1.482\,6\times\mathrm{MAD}(x)}$$

若值大于 3，标记为异常，并开展敏感性分析。

2.5　结构化提示的 schema 与编译

为消除结构化、非结构化的信息混杂，先以固定 schema 将提示词拆解为槽位并 JSON(JavaScript object notation) 化，再编译为可读提示。核心槽位包括主体/关系、场景/时间、构图/镜头、风格/色盘和负面项。

槽位指示向量为

$$\boldsymbol{Z}=[z_1\ z_2\ \cdots\ z_k],\ z_k\in\{0,1\}$$

编译函数为 $P_{\mathrm{struct}}=\pi(z,\mathrm{values})$，为避免“信息量不等”带来的混杂，引入信息量对齐指数：

$$\mathrm{PI}=\alpha\frac{\sum_{k=1}^{K}z_k}{K}+(1-\alpha)\frac{\min(T_{\mathrm{struct}},T_{\mathrm{plain}})}{\max(T_{\mathrm{struct}},T_{\mathrm{plain}})}$$

式中：α=0.5；T 为两类提示词的 token 数；当 PI≥τ（经验阈值 τ=0.8）判定对齐通过，并以关键词集合的 Jaccard 指示附属核验。

$$J=\frac{|K_{\mathrm{struct}}\cap K_{\mathrm{plain}}|}{|K_{\mathrm{struct}}\cup K_{\mathrm{plain}}|}$$

2.6　模型与 LoRA 微调

在统一推理栈中仅改变 S 与 L 的水平，其余超参（分辨率、步数、CFG/指导尺寸、负面项）在四象限中保持一致。与传统基于 U-Net 的扩散模型不同，FLUX-dev 采用了基于流匹配的多模态架构。该范式通过建立噪声与数据分布之间的最优直线传输路径，在百亿参数规模下实现了更高的计算效率与跨模态语义对齐能力。模式表达为

$$\boldsymbol{W}'=\boldsymbol{W}+\alpha\boldsymbol{B}\boldsymbol{A},\boldsymbol{A}\in\mathbf{R}^{\mathrm{dout}\times r},\boldsymbol{B}\in\mathbf{R}^{\mathrm{dout}\times r},r\ll\min(d_{\mathrm{in}},d_{\mathrm{out}})$$

为了进一步提高基于大语言模型的图像生成精度和可控性, 本文引入 LoRA 微调模型来探究对其生成结果的影响。LoRA 通过低秩矩阵的分解对生成基座进行参数高效微调, 冻结大部分权重, 仅仅在多模态 Transformer 架构(MMDiT)的注意力机制等关键线性层中插入低秩增量, 以极少可训练参数完成风格、领域的对齐并降低过拟合与算力成本。训练后于推理阶段通过缩放系数调整 LoRA 的影响力, 实现场景任务匹配的“软融合”。在统一采样与分辨率设置下, LoRA 与结构化提示词分工明确, 前者补强局部表征(边缘连贯、材质质感、细部一致), 后者约束全局语义与分布。基于大语言模型的 LoRA 微调模型构建过程如图 1 所示。

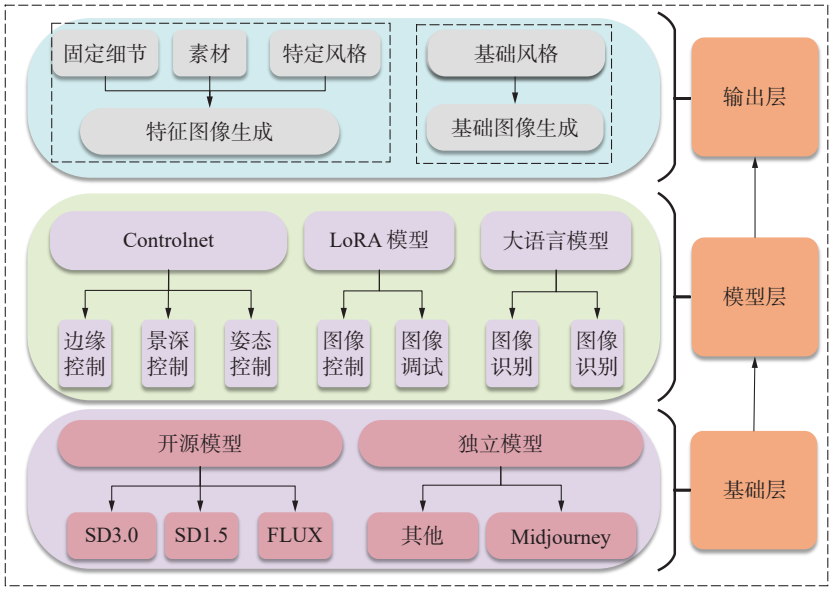

图 1　模型架构

Fig. 1　Model architecture

在 ComfyUI 与 FLUX-dev 框架下, 先选取覆盖物体、场景、光照、人物与复杂构图的 50 个主题的 400 张图片, 将 400 张样本按 320/80 划分为训练集与测试集, 基于 RTX 5090(12 GB)训练。LoRA 配置: 仅在 *L*=LoRA 条件下注入; CFG 为 1, LoRA 强度为 1。训练设定为 rank=4、epoch=8、batch size=1、AdamW、学习率 3×10^{-5}、LoRA dropout=0.05; 文本编码器与负面提示与 base 保持一致; 输出分辨率为 1 024 像素×1 024 像素。随后对生成结果实施严格的视觉质量审查, 系统性地剔除了严重遮挡或存在文本溢出的劣质测试集样本, 直到满足总数。为确立实验结论的鲁棒性与高度可复现性, 采用两组独立种子完成了全流程的交叉复现与量化评估。

3　结构化提示词与 LoRA 微调实验分析

3.1　研究设计与实现

本研究面向文本到图像生成, 采用 2×2 因子成对设计检验提示词结构化与 LoRA 微调对图文符合度 CLIP 以及美学评分的主效应和交互作用。以 FLUX-dev 为基座, 同一主题在四象限{*P-B,P-L,S-B,S-L*}下各生成一张图, 并按主题内配对统计, 以配对差抑制混乱。采用两组独立随机种子(651443027168890, 652542538796679), 除 *S*、*L* 外, 分辨率、采样步数、文本截断和负面项设定恒定条件下, 每主题四象限中各生成一张, 在两种子下重复。

为评估结论的外部效应, 在保持同一主题集与等语义内容前提下引入跨模态验证。将生成器切换为 DALL·E3, 仅改变提示词是否结构化, 其余评估流程与参数保持一致, 以此检验提示词结构化的跨模态稳健性。为研究结构化提示 schema 按槽位分解中的核心槽位, 引入消融实验, 估计核心槽位的边际贡献。

3.2　数据与预处理

本研究以物体、景物、人物、光照、构图复杂度等 100 个主题为基本单位, 在两组独立随机种子(652542538796679、651443027168890)下按照 2×2 被试内因子完全复刻。同一主题在四象限下各生成一张图像, 1 024 像素×1 024 像素输出, 推理保持 LoRA 强度为 1, CLIP 引导为 1, 共计 800 张图(每个象限 200 张)。为评估结论的外部效应, 在同一主题集与等语义下引入跨模态模型 DALL·E3, 生成 100 张图像; 为刻画结构化 schema 的关键贡

献，机理消融子集在 seed=651443027168890 下对结构化提示词进行构图槽位删除，生成 400 张图像。

针对上述的 1 300 张图像，依照 NSFW、构图越界、重复破图等进行筛查，实施双评审加仲裁，计算得 Cohen's k 介于 0.82~0.89，符合标准。经四象限中的任一缺失以“主题+种子”整组剔除，数值异常按 MAD 判据在配对内最小化处理后，评分实现与依赖版本均版本化与锁定。基于清洗后的部分样本集如图 2 所示，提示词参见表 1。

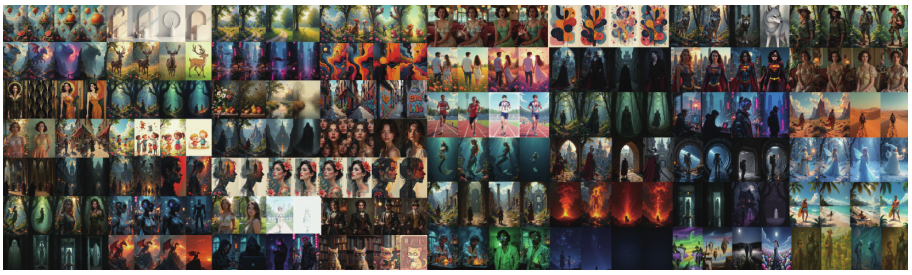

图 2　样本图集（部分）

Fig. 2　Sample gallery (partial)

表 1　四象限提示词

Table 1　Four quadrant prompt

组	结构化提示词	非结构化提示词
1	“主题关联: “漂浮岛屿、巨型水果” 场景时间: “” 构图视角: “均衡构图; 中景镜头, 微低角度; 50 mm镜头” 风格色调: “超现实主义; 鲜活明艳、梦幻朦胧、空灵辉光”	一幅浮岛场景: 超大水果, 采用超现实主义风格/色调; 鲜活、梦幻、空灵光晕; 构图均衡; 中景镜头、微低角度; 50 mm镜头。
2	主题关联: “未来主义建筑; 悬浮载具” 场景时间: “夜晚” 构图视角: “均衡构图; 中景镜头, 微低角度; 50 mm镜头” 风格色调: “赛博朋克; 霓虹、明亮; 未来主义、霓虹灯”	一幅展现未来主义建筑的场景: 悬浮车辆, 夜幕下, 采用赛博朋克风格/色调; 霓虹灯, 明亮; 未来感, 霓虹灯饰; 构图均衡; 中景镜头、微低角度; 50 mm镜头。
3	主题关联: “简单几何形状” 场景时间: “白天” 构图视角: “均衡构图; 中景镜头, 微低角度; 50mm镜头” 风格色调: “极简主义; 单色调、宁静; 柔和自然光”	一幅以简约几何图形为元素的场景: 白昼时分, 采用极简主义风格/色调; 单色调、宁静氛围; 柔和自然光线; 构图均衡; 中景镜头、微低角度; 50 mm镜头。

3.3　评价指标

本研究采用 CLIP 语义一致性评分 C 与美学评分 A 作为主要评价指标。为进行条件效应量度量，定义主题内成对差异：

$$\Delta C_t^{(s)} = C_{t,s,\text{lora}} - C_{t,s,\text{base}}$$

$$\Delta A_t^{(s)} = A_{t,s,\text{lora}} - A_{t,s,\text{base}}$$

结构化主效应为

$$\Delta C_t^{(l)} = C_{t,\text{struct},l} - C_{t,\text{plain},l}$$

$$\Delta A_t^{(l)} = A_{t,\text{struct},l} - A_{t,\text{plain},l}$$

相互作用采用差中之差：

$$\Delta_C^{\text{int}} = \overline{\Delta C}^{(\text{struct})} - \overline{\Delta C}^{(\text{plain})}$$

$$\Delta_A^{\text{int}} = \overline{\Delta A}^{(\text{struct})} - \overline{\Delta A}^{(\text{plain})}$$

为解释性汇总而不替代主分析，定义综合指标（本实验权重取 2）：

$$M_t^{(\cdot)} = \Delta C_t^{(\cdot)} + \lambda \Delta A_t^{(\cdot)}$$

采用主题为随机截距的线性混合模型：

$$y_{t,s,l} = \beta_0 + \beta_S[s=\text{struct}] + \beta_L[l=\text{lora}] + \beta_{SL}[s\times l] + \mu_t + \xi_{t,s,l}$$

式中：$y \in \{C, A\}, \mu \sim N \in (0, \sigma_\mu^2), \xi \sim N \in (0, \sigma^2)$。

3.4　主实验对比

根据两组随机种子以及四象限中生成图像的 CLIP 评分以及美学评分，得到 Aesthetic Score 2×2 交互图、CLIPScore2×2 交互图如图 3、图 4 所示。

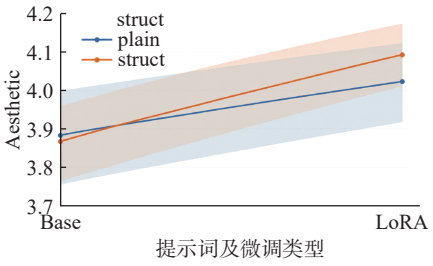

图 3　美学评分交互图

Fig. 3　Aesthetic score interaction chart

图 3、图 4 以主题聚合的交互图给出全局的走向，结构化提示与 LoRA 微调在美学评分上呈现一致的主效应，并表现出增益相加的趋势，而针对图文一致性则出现轻度下调。图 3 美学评分中两条曲线（struct、plain）从 Base 到 LoRA 同向上

升, 表明 LoRA 对审美质量有稳定的正向主效应; 且 struct 的斜率更陡, 在 base 端 struct 略低于 plain, 但随即 struct 渐高, 到 LoRA 端 struct 明显高于 plain, 呈现增益相加并伴随弱正交互的趋势。尽管置信区间部分重叠, 但方向一致、分离稳定, 表明结构化更能在 LoRA 条件下转化为可感知的审美增益。

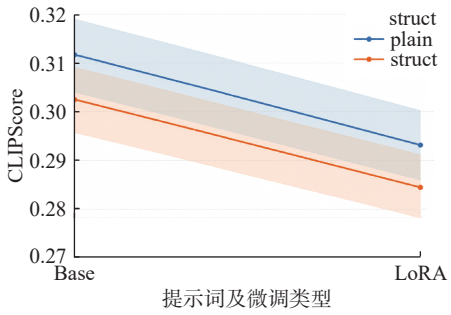

图 4　CLIPScore 交互图

Fig. 4　CLIPScore interaction diagram

图 4 中 CLIPScore 两条曲线从 base 到 LoRA 同向下降, LoRA 对 CLIP 一致性存在小幅负向主效应; 同时, struct 曲线整体低于 plain, 结构化提示词亦带来轻度负向主效应。两线近似平行、置信区间宽度相当且多出重叠, 指向交互效应弱、方差结构相近的稳健格局。

为将聚合趋势映射到全分布视角并对结构化与 LoRA 主效应作成对核验, 特辅以 Aesthetic Score 四象限箱线图、CLIPScore 四象限箱线图(图 5、图 6), 以分布证据补充交互图的点估计证据。

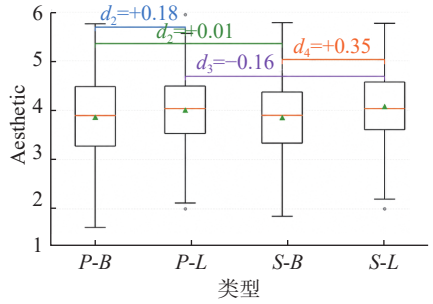

图 5　四象限美学评分箱线图

Fig. 5　Four-quadrant aesthetic rating box plot

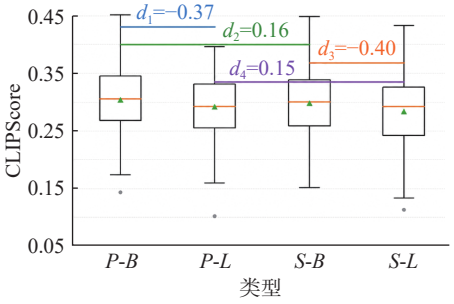

图 6　四象限 CLIPScore 箱线图

Fig. 6　Four-quadrant CLIPScore box plot

如图 5 所示, 在非结构化(plain)条件下从基础模型到 LoRA 微调 (*P-B* 到 *P-L*) 的配对效应量 d_z=+0.18, 在结构化条件下经 LoRA 微调(*S-B* 到 *S-L*)的 d_z=+0.35, 均值/中位数整体右移而 IQR 未异常增宽, 提示为小–中等级量且非极值驱动的提升, LoRA 对审美质量具有稳定正向主效应。而结构化的主效应随模型而异: 在基础模型下, 非结构化提示词过渡到结构化提示词(*P-B* 到 *S-B*)几乎为零效应(d_z=−0.01), 在 LoRA 优化下(*P-L* 到 *S-L*), 结构化提示词呈轻度正效应(d_z=0.16)。综合箱线图与交互图, 结构化×LoRA 微调的增益主要呈相加而交互弱, *S-L* 在均值与中位数上均为四象限最优。

如图 6 所示, LoRA 在无结构化提示词下(*P-B* 到 *P-L*)的 d_z=−0.37, 在结构化提示词下(*S-B* 到 *S-L*)的 d_z=−0.40; 结构化提示词在无 LoRA 微调时(*P-B* 到 *S-B*)的 d_z=−0.16, 经 LoRA 微调后(*P-L* 到 *S-L*)的 d_z=−0.15, 均属于轻–中等负向。结构化提示把文本嵌入推向“细而满”的方向, 但生成端只能部分兑现这些细节, 导致文本–图像嵌入的夹角变大, CLIPScore 下降。

表 2、表 3 为基于 FLUX-dev 的四象限 CLIPScore 和美学评分的分析结果。进行 *T* 检验分析后, 本文得到以下的结果: *P-B* 与 *P-L*、*S-B* 与 *S-L* 进行成对比较, 发现其在 CLIPScore 上具有显著差异(t=3.2944, p=0.0011; t=2.9909, p=0.0030); 美学评分上, 同样具有显著性差异 (t=−1.7422, p=0.0082; t=−2.8914, p=0.0041), 结果表明使用 LoRA 微调能显著提高其在语义一致性和美学质量上的效益。将 *P-B* 与 *S-B*、*P-L* 与 *S-L* 进行成对比较, 在 CLIPScore 上差异性并不显著(t=1.6395, p=0.1019; t=1.4871, p=0.1378), 但在美学质量方面差异性显著(t=0.1876, p=0.0513; t=−0.9565, p=0.0394), 结果表明, 结构化提示词能在美学质量方面带来提升效益, 但语义一致性上的提升并不充分, 支持下文的跨模态验证结构化提示词对其影响。

表 2　基于 FLUX-dev 的四象限 CLIPScore 和美学评分分析

Table 2　Analysis of quadrant-based CLIPScore and aesthetic scores using FLUX-dev

类型	CLIPScore		美学评分	
	平均值	标准差	平均值	标准差
P-L	0.2927	0.0460	4.0132	0.7240
P-B	0.3078	0.0453	3.8806	0.7961
S-L	0.2855	0.0512	4.0822	0.7180
S-B	0.3002	0.0470	3.8660	0.7769

表 3　基于 FLUX-dev 的四象限 CLIPScore 和美学评分 *t* 值、*p* 值分析

Table 3　Analysis of *t*-values and *p*-values for the four-quadrant CLIPScore and aesthetic scores based on FLUX-dev

类型	CLIPScore		美学评分	
	t	*p*	*t*	*p*
P-B vs *P-L*	3.2944	0.0011	−1.7422	0.0082
S-B vs *S-L*	2.9909	0.0030	−2.8914	0.0041
P-B vs *S-B*	1.6395	0.1019	0.1876	0.0513
P-L vs *S-L*	1.4871	0.1378	−0.9565	0.0394

鉴于上述的评分完全依赖 OpenAI 训练模型的自动打分，缺乏人类主观评估或外部验证，本文引入一轮人工评分实验，选取 30 名来自艺术创作一线的专家学者进行评估，参与者均拥有丰富的艺术实践经验，同时参与过 AIGC 的研究。为减轻批量评分的压力，该人工评分采用抽样方法选取图像，抽取 4 组提示词，每组提示词在两组随机种子四象限模式下生成 8 张图像，共 32 张图像用于专家评估，得到评估结果如下。

表 4、表 5 为人工评分的四象限 CLIPScore 和美学评分的分析结果。进行 *T* 检验分析后，本文得到以下的结果：首先比较 LoRA 微调的影响，将 *P-B* 与 *P-L*、*S-B* 与 *S-L* 进行成对比较，发现其在 CLIPScore 上差异显著（t=−3.5690, p=0.0005; t=−3.3317, p=0.0012）；美学评分上，亦呈显著差异（t=−3.4875, p=0.0007; t=−7.8785, p=0.0121），方向与自动评分结论一致，表明使用 LoRA 微调有助于提升图文一致性和美学质量。其次比较结构化的影响，将 *P-B* 与 *S-B*、*P-L* 与 *S-L* 进行成对比较，在 CLIPScore 上，*P-B* 与 *S-B* 有显著差异（t=−3.0205, p=0.0030），*P-L* 与 *S-L* 为边界不显著（t=−1.6501, p=0.1015）；在美学评分上一组显著（t=−3.4894, p=0.0017），一组不显著（t=−1.5852, p=0.1155）。人工评分结果基本表明使用 LoRA 微调能提高美学质量和图文一致性，结构化提示词也具有同等效益，但人工评分也存在一定的主观性。

表 4　人工评分的四象限 CLIPScore 和美学评分分析

Table 4　Four-quadrant analysis of CLIPScore and aesthetic scoring in manual evaluation

类型	CLIPScore		美学评分	
	平均值	标准差	平均值	标准差
P-L	0.3012	0.0281	4.1362	0.5741
P-B	0.2850	0.0230	3.7966	0.5265
S-L	0.3100	0.0321	4.4985	0.6003
S-B	0.2953	0.0147	3.9021	0.0798

表 5　人工评分的四象限 CLIPScore 和美学评分 *t* 值、*p* 值分析

Table 5　Four-quadrant CLIPScore and aesthetic score *t*-value and *p*-value analysis for manual scoring

类型	CLIPScore		美学评分	
	t	*p*	*t*	*p*
P-B vs *P-L*	−3.5690	0.0005	−3.4875	0.0007
S-B vs *S-L*	−3.3317	0.0012	−7.8785	0.0121
P-B vs *S-B*	−3.0205	0.0030	−1.5852	0.1155
P-L vs *S-L*	−1.6501	0.1015	−3.4894	0.0017

综合判断，模型自动评分和人工评分在总体趋势上相互印证：LoRA 稳定提升美学质量与图文一致性；结构化提示在 CLIPScore 上的效应依赖 LoRA 与基座设定，在美学评分上总体表现为中性至轻度正效应，从而支持后文开展跨模态验证以进一步检验结构化提示对语义一致性的影响。

3.5　机理消融与稳健性

以结构化提示词 schema 槽位中的“构图”作为待检因子，采用配对消融评估其边际贡献。在其余槽位与超参完全一致的前提下，对每一对“主题种子”成对生成全构图和去构图的两幅图，计算 Cohen’s dz 并给出 95%CI，结果见表 6。

表 6　结构化槽位构图消融对比

Table 6　Structured slot layout ablation comparison

评分类别	结构化及微调	消融前后差值	95% CI	p_holm
Aesthetic	*S-B*	0.309	0.49 [0.29, 0.69]	<0.001
Aesthetic	*S-L*	0.105	0.19 [−0.01, 0.39]	0.116
CLIPScore	*S-B*	0.003	0.06 [−0.13, 0.26]	0.518
CLIPScore	*S-L*	0.013	0.32 [0.12, 0.51]	0.006

表 6 所示的“构图”槽位的配对消融中呈现出清晰而稳健的分化效应：对美学评分，在结构化但未微调的设置（*S-B*）去除构图带来显著降幅（Δ=0.309），对应中等效应量（d_z=0.49, Holm p<0.001），表明显式的布局约束（主体定位、取景于层次）可系统性提升感知质量；而在结构化并经 LoRA 适配的设置（*S-L*）作用被明显稀释（Δ=0.105, d_z=0.19, Holm p=0.116）提示 LoRA 已部分吸收构图的先验。相较之下，对语义一致性（CLIPScore），基础模型下几乎零效应（*S-B*, Δ=0.003, d_z=0.06, Holm p=0.518），仅在与 LoRA 协同出现显著的正向效应（Δ=0.013, d_z=0.32, Holm p=0.006）。结构化提示词中的构图槽位主要改善视觉的感知维度，能显著提升图片的整体美观，而对文本–图像相符度的直接影响有限。

3.6　跨模型验证结构化提示对 CLIP 的影响

前文在 FLUX-dev 基座下采用四象限进行成对对比, 发现 CLIPScore 的组间差异均不显著, 随着大语言模型版本的提升, 不同的模型针对相同风格的提示词语义有不同程度的理解, 模型架构的差异化造就了模型的“百花齐放”或“一枝独秀”[31]。若 FLUX-dev 模型本身就是具有极强的自然语言解析与跨模态对齐能力, 当内容表述清楚时, 其语义一致性已经接近“天花板”, 达到模型的语义饱和。

本文中语义饱和侧重于模型的语义饱和, 指在固定采样与超参数下, 若同主题同象限内的评分配对差(Δ)的 95%CI 跨 0 且$|d_z|$<0.1, 则判定进入语义饱和区间。

在 FLUX-dev 基座上采用四象限设计(*P-B*、*P-L*、*S-B*、*S-L*), 按主题与种子共生成 1 600 张基线图像, 用于比较结构化提示与 LoRA 的主效应及其交互效应。在此基础上开展两条单变量敏感性复线: 1) 仅增量修改提示语义(token 或槽位填充质量); 2) 仅修改一个超参数(seed、步数、CFG 或负面提示), 其余条件不变, 额外生成 600 张敏感性图像, 合计 2 200 张。对每个主题/象限计算“修改前后”的配对差 Δ, 结果如表 7 所示。CLIPScore 与美学评分的均值差接近 0, 95%CI 跨 0 且$|d_z|$<0.1, 满足语义饱和的判定阈值, 表明在 FLUX-dev 设置下出现语义饱和。

表 7　语义饱和结果分析

Table 7　Semantic saturation result analysis

类型	组	CLIPScore			美学评分		
		均值差	95%CI	d_z	均值差	95%CI	d_z
P-L	组1	0.000 04	[−0.024 2~0.024 3]	0.000 7	−0.000 21	[−0.382 4~0.382 0]	−0.000 1
	组2	0.000 12	[−0.024 1~0.024 4]	0.001 9	0.000 03	[−0.381 7~0.382 2]	<0.000 1
P-B	组1	−0.000 16	[−0.024 0~0.023 7]	−0.002 5	0.007 82	[−0.412 5~0.428 2]	0.006 9
	组2	0.001 88	[−0.020 2~0.025 8]	0.029 4	0.004 10	[−0.416 2~0.424 4]	0.003 6
S-L	组1	0.000 13	[−0.026 9~0.027 1]	0.001 8	0.000 09	[−0.378 9~0.379 1]	<0.000 1
	组2	−0.000 62	[−0.027 6~0.026 4]	−0.008 5	−0.000 26	[−0.037 9~0.037 8]	−0.000 3
S-B	组1	−0.003 40	[−0.025 2~0.024 5]	−0.005 1	0.002 57	[−0.407 5~0.412 8]	0.002 4
	组2	0.000 15	[−0.024 7~0.024 9]	0.002 2	0.004 84	[−0.037 9~0.037 8]	0.004 4

上文已得到 FLUX-dev 模型出现语义饱和, 基于此引入 DALL·E3 模型跨模态验证图文符合度。除提示词方式不同外, 其余条件均与上模型一致。由于上文已得出 LoRA 微调的影响效应, 本次测试仅测评 CLIPScore(不含美学评分), 得到图 7 所示的结构化与否的 CLIP 评分图。

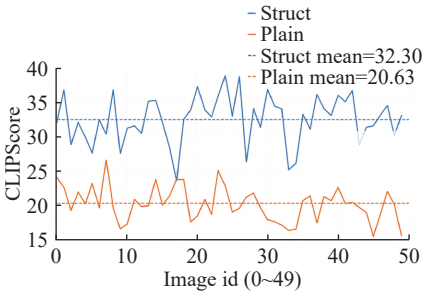

图 7　结构化与非结构化 CLIP 评分对比

Fig. 7　Comparison of structured and unstructured CLIP scores

如图 7 所示, 在 DALL·E3 模型上, 结构化提示对 CLIP 的提升强且稳健, 均值差异提升近 50%, 且成对效应量达到极大水平, 置信区间完全位于正区间。这说明当模型的文本理解与关系解析能力相对不那么“自带强约束”时, 结构化语法通过显式槽位(对象–关系–布局–限制/负面项)减少语义歧义、提高指令可执行性, 从而显著改善文本–图像的一致性。对比 FLUX-dev 的“持平或不显著”, 两者差异可解释为模型语言理解强度不同导致的“天花板效应”: 当基座已能在非结构化提示下充分抓住关键语义, 进一步的结构化约束难以在 CLIP 上产生可测增益[32]; 而在多数主流模型上, 这一约束仍然必不可少。

3.7　鲁棒性分析

为检验主结论对随机性的稳定性, 在两组独立随机种子(seed 为 651443027168890、65254253879-6679)和 4 个象限条件(*P-B*、*P-L*、*S-B*、*S-L*)上进行配对复测, 并对每个配对在“seed1~seed2”空间作散点拟合, 见图 8、9。图 8、9 中灰色虚线表示理想一致性($y=x$), 黑色实线为跨象限合并后的 OLS 拟合; 图例以颜色区分四象限。角标给出配

对样本量 n, Pearson 相关 r 与单项随机效应的组内相关系数 ICC(3, 1)。

如图 8 所示, 美学评分在两组种子间呈显著正相关(r=0.649、ICC(3, 1)=0.649), 对应中等偏上的重测信度。黑色 OLS 斜率明显小于对角线, 跨种子存在幅度收缩或再标定现象, 高分样本在另一 seed 下略有回落, 低分样本略有回升。这种“向均值回归”的形态在四象限中均可观察到(不同颜色散点交织而非分群), 说明结构化/LoRA 条件并不会放大 seed 噪声。

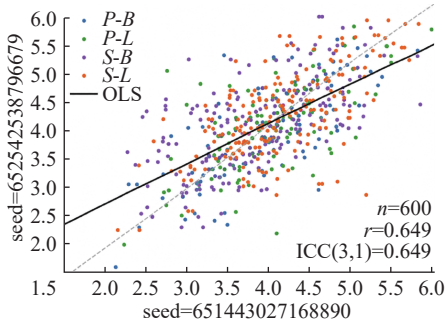

图 8　美学评分的跨 seed 一致性散点图

Fig. 8　Scatter plot of cross-seed consistency in aesthetic scoring

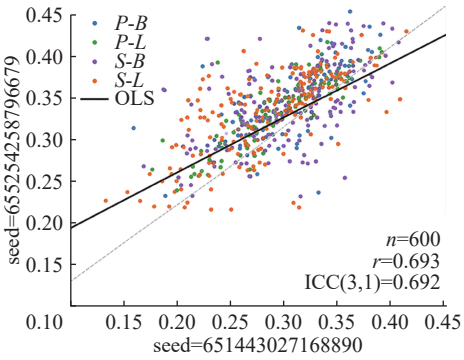

图 9　CLIPScore 的跨 seed 一致性散点图

Fig. 9　CLIPScore cross-seed consistency scatter plot

如图 9 所示, CLIPScore 的跨 seed 一致性更高(r=0.693、ICC(3, 1)=0.692), 显示中–高水平的重测信度, 且 OLS 线更贴近对角线, 表明该指标对随机种子的扰动更不敏感, 重复性优于 Aesthetic Score。同时, 不同象限的散点并未出现系统性偏移或异方差扩张, 提示结构化与 LoRA 因子不改变 seed-level 方差结构。

3.8　构图基础上的对比

CLIP 主要衡量“文本–图像语义一致性”, 对版式结构、相对位置、遮挡关系与轮廓连续性不够敏感, 而这些恰是设计场景的关键约束。为避免仅以 CLIPScore 得出片面的结论, 在 FLUX-dev 上差异不显著的同时, 转向构图层面的对比分析, 以配对样例观察主体锚点稳定性、尺度与姿态漂移、轮廓闭合与伪影等几何信号, 作为对 CLIP 与美学评分的补充证据, 从而更全面地刻画提示词结构化与 LoRA 对可控生成的真实影响。

轮廓对比主要体现视觉上带来的主观感受, 为进一步比较构图细节对图的影响, 进行图片的精细化对比, 得到图 10 所示的四象限对比结果。如图 10 所示, 在“四象限”对照中, 以红框标注结构化效应、蓝框标注 LoRA 效应。结构化主要决定全局版式、稳定主体与装饰的相对位置、文本与图元的基线对齐以及遮挡、层级关系, 显著抑制标题越界与元素漂移; LoRA 主要改善局部呈现, 画更连贯、边缘更干净、材质更一致, 但在少数主题上出现过锐与风格过拟合。二者呈明确分工且交互弱: 在已结构化前提下, LoRA 的收益集中于细节而非语义、构图; 在未结构化时, LoRA 难以弥补缺失的几何约束, 位置与层级错误仍频发。该质性观察与量化结果相吻合, FLUX-dev 上 CLIP 对结构化 LoRA 不敏感, 而跨模型验证显示结构化显著提升 CLIP, 美学更受 LoRA 强度与题材匹配影响。

图 10　精细化对比

Fig. 10　Detailed comparison

4 结束语

本文提出面向“结构化+LoRA 微调”的可复现生图模式, 并补充美学维度; 给出可复用的控图范式——“结构化定全局、LoRA 磨细节”; 揭示“模型语义饱和”会导致结构化效应失显这一关键边界条件。在 FLUX-dev 上, 结构化及 LoRA 微调对 CLIP 无显著增益, 呈“模型语义饱和”, 但 LoRA 在多数主题稳定提升美学分(细节与风格一致性)。跨模型验证(DALL·E3)中, 结构化显著提高 CLIP 语义一致性, 不同基座对提示结构的敏感点差异明显, 面向特定版本沉淀高质量提示往往优于盲目换模, 局限在于美学分以自动化打分为主、缺少大规模人评。之后将扩展至多基座/多语种, 引入人评与偏好学习, 探索“弱结构+自适应 LoRA 强度”闭环控制, 以兼顾提示简洁性、语义一致性与美学提升。

参考文献:

[1] XU Junping, ZHANG Xiaolin, LI Hui, et al. Is everyone an artist? a study on user experience of AI-based painting system[J]. Applied sciences, 2023, 13(11): 6496.

[2] WU Zhuohao, JI Danwen, YU Kaiwen, et al. AI creativity and the human-AI co-creation model[C]//Human-Computer Interaction. Theory, Methods and Tools. Online: HCII, 2021.

[3] WANG Changsheng. Art innovation or plagiarism? Chinese students' attitudes toward AI painting technology and influencing factors[J]. IEEE access, 2024, 12: 85795−85805.

[4] REDDY A. Artificial everyday creativity: creative leaps with AI through critical making[J]. Digital creativity, 2022, 33(4): 295−313.

[5] LIU Jiachang, SHEN Dinghan, ZHANG Yizhe, et al. What makes good in-context examples for GPT-3?[C]// Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. Dublin: ACL, 2022.

[6] WITTEVEEN S, ANDREWS M. Investigating prompt engineering in diffusion models[EB/OL]. (2022−11−21)[2025−08−20]. https://arxiv.org/abs/2211.15462.

[7] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27: 2672−2680.

[8] DENTON E L, CHINTALA S, SZLAM A. Deep generative image models using a laplacian pyramid of adversarial networks[J]. Advances in neural information processing systems, 2015, 28: 1486−1494.

[9] 胡铭菲, 左信, 刘建伟. 深度生成模型综述[J]. 自动化学报, 2022, 48(1): 40−74.
HU Mingfei, ZUO Xin, LIU Jianwei. Survey on deep generative model[J]. Acta automatica sinica, 2022, 48(1): 40−74.

[10] 谢天圻, 吴媛媛, 敬超, 等. GAN 模型生成图像检测方法综述[J]. 计算机工程与应用, 2024, 60(22): 74−86.
XIE Tianqi, WU Yuanyuan, JING Chao, et al. Survey of image detection methods generated by GAN models[J]. Computer engineering and applications, 2024, 60(22): 74−86.

[11] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein gans[J]. Advances in neural information processing systems, 2017, 30: 5767−5777.

[12] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[EB/OL]. (2015−11−19)[2025−08−20]. https://arxiv.org/abs/1511.06434.

[13] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis[C]//International Conference on Machine Learning. New York: PMLR, 2016.

[14] ZHANG Han, XU Tao, LI Hongsheng, et al. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017.

[15] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with clip latents[EB/OL]. (2022−04−13)[2025−08−16]. https://arxiv.org/abs/2204.06125.

[16] CHAN W, DENTON E, FLEET D, et al. Photorealistic text-to-image diffusion models with deep language understanding[C]//Advances in Neural Information Processing Systems 35. New Orleans: NeurIPS, 2022.

[17] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. Online: PmLR, 2021.

[18] NICHOL A, DHARIWAL P, RAMESH A, et al. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models[EB/OL]. (2021−12−20)[2025−08−16]. https://arxiv.org/abs/2112.10741.

[19] SHAVLOKHOVA V, VOLLMER A, ZOUBOULIS C C, et al. Finetuning of GLIDE stable diffusion model for AI-based text-conditional image synthesis of dermoscopic images[J]. Frontiers in medicine, 2023, 10: 1231436.

[20] AVRAHAMI O, LISCHINSKI D, FRIED O. Blended diffusion for text-driven editing of natural images[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022.

[21] KIM G, KWON T, YE J C. DiffusionCLIP: text-guided diffusion models for robust image manipulation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022

[22] AVRAHAMI O, FRIED O, LISCHINSKI D. Blended latent diffusion[J]. ACM transactions on graphics, 2023, 42(4): 1−11.

[23] HERTZ A, MOKADY R, TENENBAUM J, et al. Prompt-to-prompt image editing with cross attention control[EB/OL]. (2022−10−02)[2025−08−16]. https://arxiv.org/abs/2208.01626.

[24] OPPENLAENDER J, LINDER R, SILVENNOINEN J. Prompting AI art: an investigation into the creative skill of prompt engineering[J]. International journal of human–computer interaction, 2025, 41(16): 10207−10229.

[25] OPPENLAENDER J. The creativity of text-to-image generation[C]//Proceedings of the 25th International Academic Mindtrek Conference. Tampere: ACM, 2022.

[26] BROWNE K. Who (or what) is an AI artist?[J]. Leonardo, 2022, 55(2): 130−134.

[27] 王常圣. 面向大模型艺术图像生成的提示词工程研究[J]. 图学学报, 2024, 45(6): 1243−1255.
WANG Changsheng. Research on prompt engineering for generating artistic images using large models[J]. Journal of graphic sciences, 2024, 45(6): 1243−1255.

[28] DAI Kanyuan, SHAO Ji, GONG Bo, et al. CLIP-FSSC: a transferable visual model for fish and shrimp species classification based on natural language supervision[J]. Aquacultural engineering, 2024, 107: 102460.

[29] WANG Jianyi, CHAN K C K, LOY C C. Exploring CLIP for assessing the look and feel of images[J]. Proceedings of the AAAI conference on artificial intelligence, 2023, 37(2): 2555−2563.

[30] HENTSCHEL S, KOBS K, HOTHO A. CLIP knows image aesthetics[J]. Frontiers in artificial intelligence, 2022, 5: 976235.

[31] 郑明琪, 陈晓慧, 刘冰, 等. 提示学习中思维链生成和增强方法综述[J]. 计算机科学, 2025, 52(1): 56−64.
ZHENG Mingqi, CHEN Xiaohui, LIU Bing, et al. A review of methods for generating and enhancing thought chains in prompted learning[J]. Computer science, 2025, 52(1): 56−64.

[32] ZHANG Wei, JIN Wei, RHO S, et al. A federated learning framework for brain tumor segmentation without sharing patient data[J]. International journal of imaging systems and technology, 2024, 34(4): e23147.

作者简介：

张伟，博士后，主要研究方向为计算机视觉、深度学习。发表学术论文 30 余篇。E-mail：wei.zhang.2@ki.se。

罗亚远，本科生，主要研究方向为计算机视觉。E-mail：luoyayuan.pl@qq.com。