Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(3), 731-742
Published Online May 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.153069

Hans汉斯

融合大模型微调的工业产品可控生成设计研究

滕　龙

浙江师范大学工学院，浙江　金华

收稿日期：2026年4月16日；录用日期：2026年4月30日；发布日期：2026年5月14日

摘　要

针对传统工业设计迭代低效，以及现有AIGC图像模型(如U-Net架构)在处理严苛机械几何约束时易发生透视失真的技术痛点，本文以工具车为验证载体，提出一种高精度、强约束的工业产品可控生成设计方法。研究确立具备全局空间感知能力的Flux大模型为基座，通过构建融合单图多视角重构技术的领域数据集，并应用LoRA微调技术进行深度微调，成功内化了品牌家族化特征。为突破单一文本提示的控制局限，本文进一步构建了融合显式几何结构锁定与隐式视觉风格迁移的双向多模态控制矩阵。实验表明，该生成范式在高保真地维持产品三维几何结构的前提下，实现了广域的材质与美学泛化，为装备制造企业的数字化研发与敏捷创新提供了切实可行的系统工程方案。

关键词

工业设计，AIGC，Flux模型，LoRA微调，可控生成

Research on Controllable Generation Design of Industrial Products Based on Fine-Tuning of Large Models

Long Teng

College of Engineering, Zhejiang Normal University, Jinhua Zhejiang

Received: April 16, 2026; accepted: April 30, 2026; published: May 14, 2026

Abstract

In response to the inefficiency of traditional industrial design iterations and the technical challenges of existing AIGC image models (such as the U-Net architecture) when dealing with strict mechanical geometric constraints, which often lead to perspective distortion, this paper uses a tool vehicle as a verification platform and proposes a high-precision and strongly constrained controllable generation design

文章引用: 滕龙. 融合大模型微调的工业产品可控生成设计研究[J]. 人工智能与机器人研究, 2026, 15(3): 731-742.
DOI: 10.12677/airr.2026.153069

method for industrial products. The study establishes the Flux large model with global spatial perception capabilities as the base. By constructing a domain dataset that integrates single-image multi-view reconstruction technology and applying LoRA fine-tuning technology for deep fine-tuning, the brand family characteristics have been successfully internalized. To break through the control limitations of a single text prompt, this paper further constructs a bidirectional multimodal control matrix that integrates explicit geometric structure locking and implicit visual style transfer. Experiments show that this generation paradigm can maintain the three-dimensional geometric structure of the product with high fidelity, and achieve wide-area material and aesthetic generalization, providing a practical and feasible system engineering solution for the digital R&D and agile innovation of equipment manufacturing enterprises.

Keywords

Industrial Design, AIGC, Flux Model, LoRA Fine-Tuning, Controllable Generation

1. 引言

在制造业向“服务型制造”与品牌化转型的宏观语境下，工业装备的外观造型已不再是性能指标的附属品，而是重塑企业核心竞争力的关键维度[1]。然而，在实际的设计推演中，业界广泛采用的“双钻模型”正面临严重的“效率悖论”[2]：受限于设计师手工绘图速度与三维建模的高昂时间成本，概念发散阶段往往难以穷尽有效的设计空间，导致产出的方案往往是为了迁就工艺而妥协的产物。同时，二维草图向三维工程转化的技术鸿沟，极易导致最终交付出现“形技分离”的断层。这种极度依赖个体经验且迭代低效的传统模式，已成为制约装备制造业品牌化跃升的显著瓶颈。

人工智能生成内容(AIGC)技术的爆发式增长，为化解上述设计效能矛盾提供了契机。从早期的生成对抗网络(GANs) [3]到如今的潜空间扩散模型[4]，机器视觉生成已跨越了简单的特征映射阶段。但在要求严苛的工业级设计场景中，现有主流生成模型仍暴露出诸多底层架构带来的局限。早期的 GANs 模型在处理工业产品的刚性矩形框架或正圆脚轮时，频繁出现直线弯曲与透视错乱。随后崛起的基于 U-Net 架构的扩散模型(如 Stable Diffusion)，虽在材质渲染上表现惊艳，但受制于卷积神经网络的“局部感受野”偏置，难以建立跨区域的长程依赖关系，导致其在面对具有严格几何约束的机械结构时，极易产生空间逻辑悖论。

为突破这一技术桎梏，本文提出一种融合大模型微调与多模态协同生成的新型设计范式。本研究首先通过多维度的对比，确立了具备全局空间感知能力的 Flux 大模型作为底层基座；随后通过构建高质量垂直领域数据集并引入 LoRA 低秩自适应微调技术[5]，将企业的品牌基因深度内化为模型底层参数。在此基础上，本文引入 ControlNet 几何约束[6]与 IP-Adapter 视觉风格迁移技术[7]，构建双向控制矩阵，彻底打通了从创意发散到工程精准收敛的关键路径。

2. AIGC 图像生成模型比较研究

在工业装备设计场景中，生成模型必须同时兼顾“高保真视觉表现”与“严苛的空间物理约束”。本节将从架构演进与工程应用的角度，深入探讨主流算法在工业生成任务中的表现差异，以论证本系统基座选型的科学性。

2.1. 传统架构的工程局限性

生成对抗网络(GANs)曾是图像生成领域的开拓者，其核心思想在于生成器与判别器的零和博弈。然而在工业设计实践中，GANs 难以克服“模式崩塌”缺陷，导致生成的方案往往千篇一律，无法提供双钻模型所需的广域创意探索。其在处理工具车这类包含矩形柜体、正圆脚轮等刚性结构的任务时，由于缺乏对全局几何逻辑的理解，极易出现严重的失真现象。

随后占据主流地位的扩散模型(如 Stable Diffusion)将生成过程重构为迭代式去噪。尽管其在审美表现力上实现了质的飞跃，但核心的 U-Net 架构在处理复杂工业结构时显露出疲态。由于卷积核倾向于捕捉局部特征，模型难以维持严谨的远近透视关系，常出现“远端轮子大于近端轮子”的空间逻辑悖论。测试数据表明，SDXL 等模型在文字渲染准确率上不足 40%，且 FID 指标[8]徘徊在 23.8 左右，难以满足高保真工业设计交付的标准。

2.2. 基于 DiT 架构的 Flux 模型优势剖析

针对传统架构的短板，本研究确立了以基于 DiT (Diffusion Transformer)架构的 Flux 模型作为核心基座[9]。Flux 彻底摒弃了 U-Net 的卷积堆叠模式，转而利用 Transformer 的多头自注意力机制(Multi-Head Self-Attention)处理图像特征块[10]。这种底层逻辑的革新赋予了模型类语言逻辑的全局空间感知力，使得图像中的任意一个微小图块都能与全图产生跨区域的信息交互，如图 1 所示。在生成具有严苛透视要求的工业渲染图时，这种架构能够有效维持产品硬边线条的平直性与三维结构的严谨性。

此外，Flux 引入了流匹配(Flow Matching) [11]与校正流(Rectified Flow) [12]技术，将高度弯曲的生成路径“拉直”为理想的线性插值轨迹。这一改进大幅缩短了推理距离，使得模型在极少采样步数下即可完成高质量生成，且显著增强了对金属拉丝、高光塑料等复杂工业材质物理光影的还原能力。量化对比显示，Flux 在未微调状态下的 FID 指标即达到 18.2 的最优水平，文字渲染准确率超过 90%，在结构严谨性与图像保真度上实现了跨代超越。

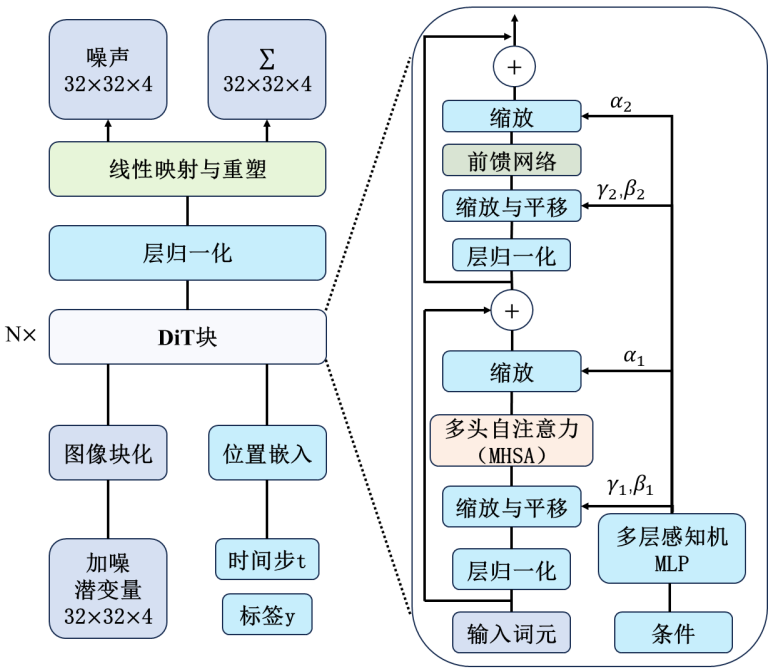

Figure 1. DiT structure diagram
图 1. DiT 结构图

3. 面向工业产品外观设计的生成模型微调与应用

LoRA 是一种通过对预训练大模型进行微调的技术，其核心思想是通过低秩调整对大模型中的部分参数进行优化，从而提升模型在特定任务上的表现。LoRA 方法相较于传统的全量训练方法，具有显著的优势：它可以在保持大模型强大生成能力的同时，大幅度减少训练成本和计算资源的消耗。尤其在工业设计这样的任务中，LoRA 的训练方式使得模型能够更加高效、精准地生成所需的设计图像。为了让模型能够适应工具车外观设计这一特定任务，本研究进行了大量的实验与调整，确保训练过程高效、稳步地推进。

3.1. 数据集预处理

即使引入了具备强大空间感知能力的通用生成模型，若在实际应用中仅依赖自然语言提示词进行引导，模型依然难以精准复刻特定企业独有的家族化设计特征。因此，构建高质量的垂直领域数据集并执行深度的底层权重微调，便成为了实现“品牌 DNA 内化”的必由之路。本研究以金华市精工工具制造有限公司的旗舰工具车为工程标的，采取了现场高精度拍摄与 Solidworks 三维建模并行的双轨数据采集策略，共计构建了 1500 张高分辨率图像的垂直领域数据集。在获取原始影像后，由于复杂的真实车间背景会严重干扰神经网络对主体特征的学习，并未采取粗暴的裁剪处理，而是依托基于 Flux Kontext [13] 的图像分割工作流，对工具车前景进行了像素级的精准剥离与人工边缘修补，最终将所有样本统一规范至 1024 × 1024 像素的高分辨率标准。更为关键的是，单一的视角分布极易导致模型在推理阶段陷入“视角过拟合”，使其将三维结构与特定的二维渲染角度强行绑定。为破解这一数据维度的死局，本研究创新性地部署了自动化多视角重构机制。该机制基于 Qwen 模型强大的跨模态理解与图像编辑能力，并针对性地引入了 Next-scene LoRA 模型以增强空间一致性与视角泛化表现。通过该组合算法，模型能够从单张工具车单视角图像中，精准逆向推演出符合透视逻辑的正视、背视及俯视图。这种数据层面的空间拓维，强行迫使模型在二维像素阵列中建立起零部件在三维空间下的连续拓扑认知，极大地增强了其空间泛化能力。完成清洗后，研究进一步构建了递进式的语义标签体系，将全局触发词“jinggong tool cart”与“heavy duty caster”、“aluminum drawer handle”等专业特征标签深度耦合，从而构筑了极其严密的图文映射网络。

3.2. 训练参数配置

考虑到 Flux 模型高达 120 亿(12 B)的庞大参数规模，其推理与微调所需的显存开销远超传统的 Stable Diffusion 模型。为了满足这一苛刻的算力需求，本实验依托高性能计算工作站进行部署，硬件核心选用单张配备 24 GB GDDR6X 显存的 NVIDIA GeForce RTX 4090 GPU。在软件生态方面，底层基于 PyTorch 2.1.2 深度学习框架，并深度集成了 xformers 加速库以优化注意力机制的计算效率。整个训练流程依托对 Flux 架构具有完善支持的 Kohya_ss GUI 框架展开。

首先，在模型优化的核心环节，标准的 AdamW 优化器由于需要同时保存一阶矩和二阶矩，会吞噬海量显存。本研究因此转而采用 8-bit 量化版本的 AdamW8bit，在几乎不牺牲梯度更新精度的前提下，极大地卸载了优化器状态的显存重担。省下的显存空间被用于支撑更稳定的数值计算——本研究摒弃了传统的 fp16，选用了 bf16 混合精度进行训练。由于 bf16 拥有与全精度 fp32 完全相同的 8 位指数位宽，其动态范围更为宽广。这对于 Flux 这类深层 Transformer 架构而言至关重要，它能有效遏制训练中极易出现的梯度溢出(Gradient Overflow)与数值震荡，保障了整个微调过程的平稳收敛。

其次，在梯度更新策略上，单卡 24 GB 的极限迫使单次物理批量大小(Batch Size)只能被设定为 1。众所周知，极小批量的训练极易引发剧烈的梯度震荡。为化解这一矛盾，本研究引入了梯度累积(Gradient

Accumulation)机制，将累积步数设定为 4。这意味着模型在连续前向传播并计算 4 个样本的误差后，才集中进行一次反向传播与权重更新。这种处理巧妙地实现了等效批量大小(Effective Batch Size)为 4 的效果，不仅大幅平滑了梯度噪声，更显著提升了模型对全局特征捕捉的稳定性。

最后，针对 LoRA 网络的拓扑结构，本研究经过反复权衡，将网络维度(Network Rank)精准锚定在 16。对于工业产品外观的特征提取而言，这是一个绝佳的平衡点：若维度过低(如 4 或 8)，模型将无力承载倒角、螺丝、接缝等精密机械细节的重构；若维度过高(如 64)，不仅会徒增显存负荷，还会加剧模型对训练集背景噪声的过拟合风险。与此同时，用于控制权重缩放的 Alpha 值也被同步设定为 16。这种 Alpha 与 Rank 比例为 1:1 的“全强度”更新策略，赋予了模型更强的学习动能，使其能够以最快速度跨越通用图像域，深度适应并拟合全新的工业设计垂直领域数据。

3.3. LoRA 模型训练

模型的训练绝非将参数进行简单的线性输入，本质是一个需要不断监控与动态博弈的寻优过程。在初始探索阶段，本研究采取了相对激进的训练策略，即在未引入任何正则化数据的情况下，将恒定学习率设定为 1e−4 并计划执行 5000 步的总训练量。然而，通过 TensorBoard 实时监控发现，当训练推进至 3500 步左右时，虽然损失函数值呈现出持续走低的假象，但模型在验证集上的泛化能力却发生了严重退化。测试生成的图像暴露出典型的过拟合症候：模型陷入了语义僵化，丧失了对提示词的响应与解耦能力。例如，当输入“红色木质材质”的指令时，模型依然强行输出原本的蓝色金属工具车。这种“灾难性遗忘”现象在于，过高的初始学习率导致模型权重在损失函数的局部极小值中深陷，加之数据集样本的高度同质化，使得模型选择了“死记硬背”的捷径来降低 Loss 值，而非真正学习目标物体的数据分布特征。

针对上述过拟合与概念混淆的痛点，本研究在后续的优化阶段引入了“先验保护(Prior Preservation)”机制，并对学习率调度策略进行了彻底重构。具体而言，首先利用 Flux 底模预先生成了 50 张基于白色背景、风格各异的通用工具车图像，将其作为正则化约束集。在损失函数的计算环节，算法被强制要求同时衡量“特定精工工具车”与“通用工具车”之间的差异。这一机制相当于在底层逻辑上引导模型剥离通用属性，仅聚焦于精工品牌的独特设计语言。与此同时，本研究将学习率调度器更替为带有重启机制的余弦退火策略(Cosine with Restarts)，并将初始学习率下调至 5e−5。这种动态策略使得模型在训练初期能够凭借较大的步幅快速逃离初始的平坦区，而在训练后期则随着余弦曲线的平滑衰减，以极小的步长在全局最优解附近进行精细化搜索，从而有效避免了在极值点附近的震荡跳出。

在完善了上述约束机制后，本研究重新设定了训练周期，将总轮次(Epochs)控制在 10 轮，折合总步数约 2800 步。结合 TensorBoard 记录的 Loss 曲线轨迹，瞬时损失函数值如图 2 所示，整个训练过程呈现出三个特征鲜明的阶段。

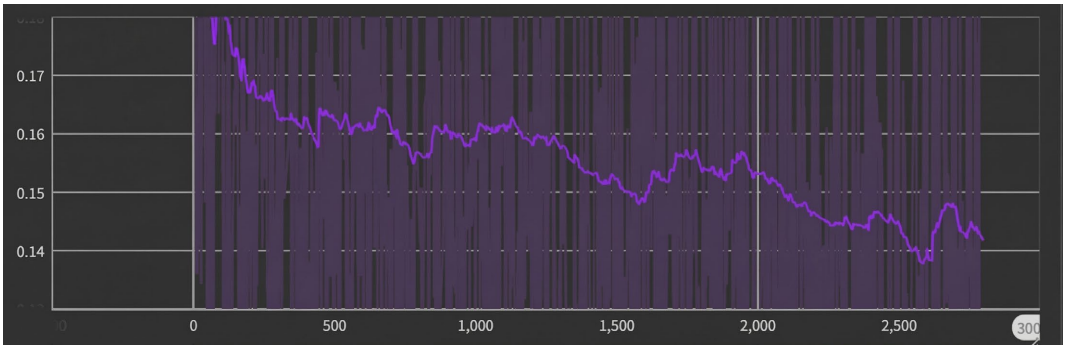

Figure 2. Instantaneous loss function value
图 2. 瞬时损失函数值

在前 8 轮的震荡下降期(约 0~2240 步)，Loss 值总体呈下降趋势，模型逐步建立起对工具车宏观结构的认知，但生成的图像在把手接缝等微观纹理上仍显粗糙，偶有结构扭曲发生。当训练推进至第 9 轮(约 2520 步)时，模型迎来了最佳收敛期。此时的 Loss 曲线触底并趋于平稳，抽取该节点的检查点(Checkpoint)进行测试发现，模型不仅完美重构了铝合金把手等复杂的工业细节，还展现出了极高的提示词响应度，能够精准执行颜色与材质的跨域修改。然而，一旦跨入第 10 轮(2520 步之后)，Loss 值开始出现反弹上涨的危险信号，生成的图像伴随产生高频噪点与伪影，且对背景的依赖性再次抬头。这明确表征模型已跨过最优拟合点，开始强行记忆训练集中的非关键光斑与噪点。基于对这一动态博弈过程的严密观察，本研究果断触发了早停(Early Stopping)机制，舍弃了出现过拟合端倪的最终轮次，将第 9 轮(约 2520 步)的存档确立为最终发布的最佳模型版本。

3.4. 模型性能评估与生成效果分析

为了深入探究 LoRA 模型是否真正将精工工具车的外观特征“内化”于底层参数，而非单纯依赖文本提示词的表层引导，本研究精心设计了一项严苛的“零提示词干扰”盲测评估实验。该实验的核心控制变量在于严格锁定全局随机种子(Seed = 42)。这一举措确保了每次生成时的初始高斯噪声分布绝对一致，从而彻底排除了随机性波动的干扰，使得最终图像的任何形态差异均能唯一归因于模型权重的演进。在具体的生成环节，仅输入唤醒模型的基础触发词“jinggong tool cart”，其余无任何提示词输入，并提取了从第 1 轮至第 10 轮的全部存档点进行纵向的演化比对，如图 3、图 4 所示。

Figure 3. 1~7 rounds of LoRA sampling with 15, 20, and 25 steps effect diagrams
图 3. 1~7 轮 LoRA 采样 15、20、25 步效果图

观察这十个轮次的演变轨迹，可以清晰地捕捉到模型“内隐知识”积累的四个动态阶段。在最初的第 1 至第 3 轮语义发散期，由于 LoRA 权重尚未形成有效约束，生成过程完全由 Flux 底模的随机性主导。此时的模型倾向于输出带有木质纹理的复古工作台或结构松散的通用铁柜，材质与形态均不具备精工品牌的识别度。随着训练推进至第 4 到第 7 轮，模型进入了特征震荡与结构重组阶段。此时柜体材质已开始初步具有数据集的结构和材质，表明 LoRA 开始接管材质控制权；然而其空间逻辑依然极不稳定，生成的工具车顶部甚至出现了非典型的开放式层架，抽屉排布混乱，模数化特征尚未成型。

真正的质变发生在了第 8 至第 9 轮的最佳平衡期。在这一阶段，模型不仅精准复刻了顶部气撑盖板

与高光黑色粉末涂层，还完美重现了铝合金把手的银色光泽与稳固的脚轮结构，实现了品牌 DNA 的深度内化与“形神兼备”。然而，当训练继续跨入第 10 轮(约 2800 步)时，模型不可避免地陷入了过拟合漂移。生成的图像开始丧失原有的质感，结构趋于简单化。这表明模型已越过拟合临界点，开始过度关注训练集中特定光照与角度下的局部反光，反而破坏了产品的整体物理形态。

Figure 4. 8~10 rounds of LoRA sampling with 15, 20, and 25 steps effect diagrams
图 4. 8~10 轮 LoRA 采样 15、20、25 步效果图

在进行纵向轮次比对的同时，本研究亦对 15、20 以及 25 步三个梯度的采样步数进行了横向的敏感性分析。结果显示，15 步的欠采样虽然能勾勒出基本轮廓，但金属表面的光泽感尤为平淡，暗部细节对比度孱弱，整体呈现出一种“灰蒙蒙”的未完成态。当步数提升至 20 步时，图像质量迎来了跃升，黑色烤漆的通透感与把手的锐利度均达到了极佳的状态，确立了最佳的效率平衡点。而继续增加至 25 步时，除了顶部阴影过渡略显柔和外，整体视觉增益已呈现出明显的边际效应递减。综合纵横两轴的实验数据，本研究最终确立了最优的部署策略：选用第 9 轮模型存档作为最终发布版本，并推荐在 20 至 25 步的采样区间内，配合 3.5 的提示词相关性(CFG Scale)进行推理，以求在最高画质与生成耗时之间取得最优解。

在确立最终模型后，为进一步验证其是否具备强大的特征解耦能力——即能否在死守核心物理结构的前提下，对外观的色彩、材质与图案进行自由编辑，本研究实施了一项极端风格化的压力测试。测试中，刻意向系统输入了包含“色彩鲜艳丰富、图案繁复、充满想象力”等极易诱发结构崩溃的极端渲染指令。面对如此密集的艺术化诉求，第 9 轮模型展现出了令人惊叹的结构与纹理正交控制力，如图 5 所示。首先，画面中的金属框架、脚轮以及拉手等核心功能部件依然稳如泰山，完全遵循了精工产品的严谨工程逻辑，未因“充满想象力”的修饰词而出现任何扭曲或穿模。其次，模型展现出了极高的空间语义理解力，将所有复杂的艺术图案精准地限制在了抽屉面板与侧板等非功能性平面区域，成功划定了“装饰区”与“功能区”的界限。这项压力测试确凿地证明，优化后的 LoRA 模型已彻底跨越了简单的特征记忆阶段，真正实现了“工业结构特征”与“艺术风格特征”的深度解耦，能够作为一块高度可控的“数字画布”，稳稳承接设计师天马行空的创意推演。

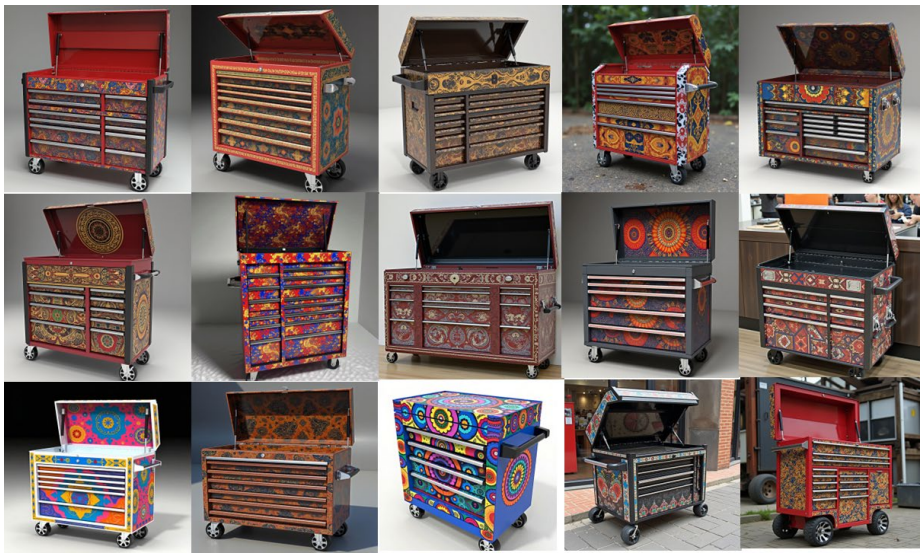

Figure 5. Color prompt words and LoRA effect images
图 5. 色彩提示词加 LoRA 效果图

3.5. 基于 FID 指标的生成质量量化评估

为进一步从严密的数学统计学视角衡量微调模型的生成质量与领域自适应(Domain Adaptation)水平，本节引入图像生成领域的标准化量化指标——弗雷歇起始距离(Fréchet Inception Distance, FID)。该指标并非简单对比图像的像素级差异，而是通过预训练的深层神经网络(如 Inception-V3)提取真实图像与生成图像的高阶语义特征向量。在假设这些高维特征服从多元高斯分布的前提下，通过计算真实数据分布与生成数据分布之间的弗雷歇距离来量化两者的相似度。

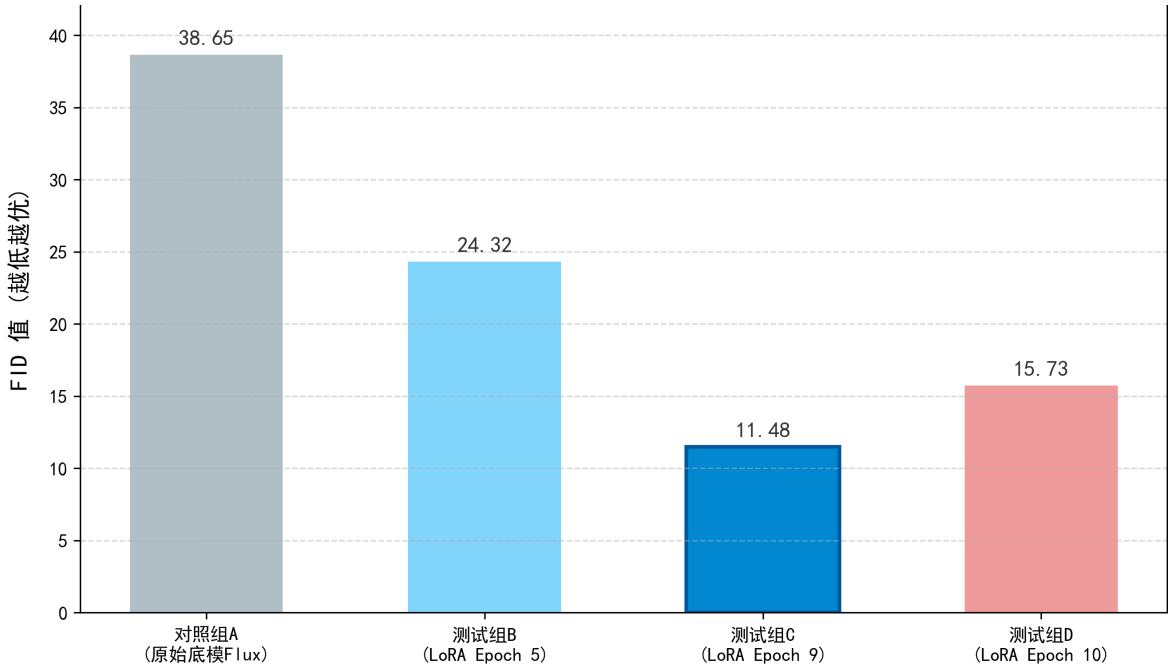

Figure 6. FID indicators of the Flux LoRA model at different training stages of the vehicle
图 6. Flux LoRA 模型不同训练阶段工具车 FID 指标

为确保量化评估结果具备统计学意义，本研究构建了严密的对比实验组。首先，从精工工具车数据集中随机等距抽取 500 张高分辨率图像作为真实基准组。随后，设定统一的工业提示词，分别利用原始 Flux-Dev 底模以及不同训练阶段的 LoRA 权重严格独立生成 500 张测试图像进行对照计算，结果如图 6 所示。

经过量化计算，不同模型配置与训练阶段的 FID 核心指标呈现出了极具规律性的演进轨迹。对于未经微调的对照组(原始 Flux 底模)，其计算得出的 FID 值高达 38.65。在视觉特征上，底模生成的图像虽然具备工具车的基础框架，但严重缺乏精工品牌特征，工业材质表现偏向通用化。这一高初始值从数学层面印证了“通用大模型”与“特定工业设计”之间存在着显著的领域鸿沟，同时也构成了本研究实施深度微调的核心必要性前提。

随着 LoRA 特定领域数据的注入，模型的生成保真度开始大幅跃升。在训练进行到 Epoch 5 时，FID 值迅速下降至 24.32。此时模型的特征生成处于震荡期，虽然品牌轮廓初现，但微观物理逻辑仍存错乱。当微调稳步推进至 Epoch 9 时，FID 值大幅收敛并达到全局最低点 11.48。这一极其优异的核心数据确凿地证实：经过深度微调的生成流形(Generative Manifold)已经高度贴合了真实的工业产品分布，模型在此时达到了最佳数据拟合，完美重现了重型脚轮与金属拉丝等高频物理细节。

然而，当训练继续跨入第 10 轮(Epoch 10)时，指标并未进一步优化，FID 值反而反弹退化至 15.73。这种数值的回升本质上是模型陷入过拟合状态的量化表征。过拟合导致模型丧失了泛化能力，在生成的图像中产生了非自然的光斑、噪点与高频伪影，这些细微的失真被 Inception 网络的深层特征提取迅速捕捉并放大。

综上所述，这一基于数学统计的量化结果，与前文基于 Loss 曲线及视觉评估的结论实现了理论与数据的三维闭环。严密的数据推演不仅证实了微调策略在跨越工业设计“领域鸿沟”中的决定性作用，更在数学层面上无可辩驳地确立了 Epoch 9 为最佳的模型归档权重。

3.6. 基于条件控制的生成模型应用研究

尽管经过微调的 Flux LoRA 模型已具备极强的品牌特征还原能力，但在面对工业设计中严苛的“工程约束”与“定向风格迭代”需求时，单纯依靠自然语言提示词仍存在随机性过大、空间控制精度不足的问题。为实现从“概率生成”向“确定性设计”的跨越，本研究进一步引入了 ControlNet 与 IP-Adapter 技术，构建了“几何结构锁定 + 视觉风格迁移”的双重控制流矩阵。

在视觉风格迁移方面，本研究引入了 IP-Adapter 这一轻量级视觉提示适配模块。不同于传统的将图像特征与文本特征简单拼接的做法，本研究对主干模型(DiT)的注意力机制进行了重构，采用了解耦交叉注意力(Decoupled Cross-Attention)机制。设网络某层的查询向量为 Q，由文本提示词提取的键值对为 K_{text}、V_{text}，由参考图像提取的风格键值对为 K_{IP}、V_{IP}，融合后的多模态交叉注意力输出 Z 可表示为：

$$Z = \text{Softmax}\left(\frac{QK_{\text{text}}^{\text{T}}}{\sqrt{d}}\right)V_{\text{text}} + \lambda_{\text{IP}} \cdot \text{Softmax}\left(\frac{QK_{\text{IP}}^{\text{T}}}{\sqrt{d}}\right)V_{\text{IP}} \tag{1}$$

其中，λ_{IP} 为风格强度系数。这种解耦设计确保了视觉风格特征能够在不干扰文本语义逻辑的前提下，精准地注入到扩散模型的生成链路中，有效解决了复杂美学特征在文本描述中的语义瓶颈。

与此同时，为了弥补扩散模型在处理机械结构时易产生的透视失真与逻辑悖论，本研究引入了 ControlNet 架构作为高度显式的空间结构引导模块。针对 Flux 的 DiT 架构特性，控制信号不再是简单的前置叠加，而是作为多尺度残差信号逐层注入到 Transformer Block 中。本研究利用 Canny 边缘检测预处理器从参考图中提取物理空间特征，并转化为控制图 c_{control}。在生成过程中 ControlNet 提取的各层特征通

过零卷积处理后注入主干网络。设第 i 层 DiT Block 的隐藏状态为 h_i，则该层的更新逻辑为：

$$\hat{h}_i = \mathrm{DiT\_Block}\left(h_i, Z\right) + \lambda_{\mathrm{ctrl}} \cdot \mathrm{ZeroConv}\left(c^i_{\mathrm{control}}\right) \tag{2}$$

λ_{ctrl} 为几何约束权重。这种深度干预机制确保了生成的工具车在宏观轮廓与微观零件(如脚轮位置、抽屉间隙)上均能严格遵循预设的工程轨迹，从根本上杜绝了工业设计中的“形技分离”现象。

在实际应用中，IP-Adapter 与 ControlNet 的协同需要精细的联调以达成风格与结构的平衡，如图 7 所示。本研究通过实验，将 IP-Adapter 与 ControlNet 的控制权重从 0.0 至 1.0 进行等步长(步长为 0.1)的二维矩阵式递增，以系统观测模型在不同干预强度下的特征耦合表现。确立了最佳权重组合：当 $\lambda_{\mathrm{ctrl}} \in [0.8,1.0]$ 且 $\lambda_{\mathrm{IP}} \in [0.8,1.0]$ 时，模型既能稳固地锁定工具车的机械框架，又能灵活地完成涂装风格的跨域迁移。为提升设计效能，本研究搭建了基于 Web 端的辅助设计界面，将上述复杂参数封装为可视化的滑块控件。设计师只需通过“参考图(结构) + 参考图(风格) + 简单提示词”的组合，即可在秒级时间内获取高保真的产品效果图。

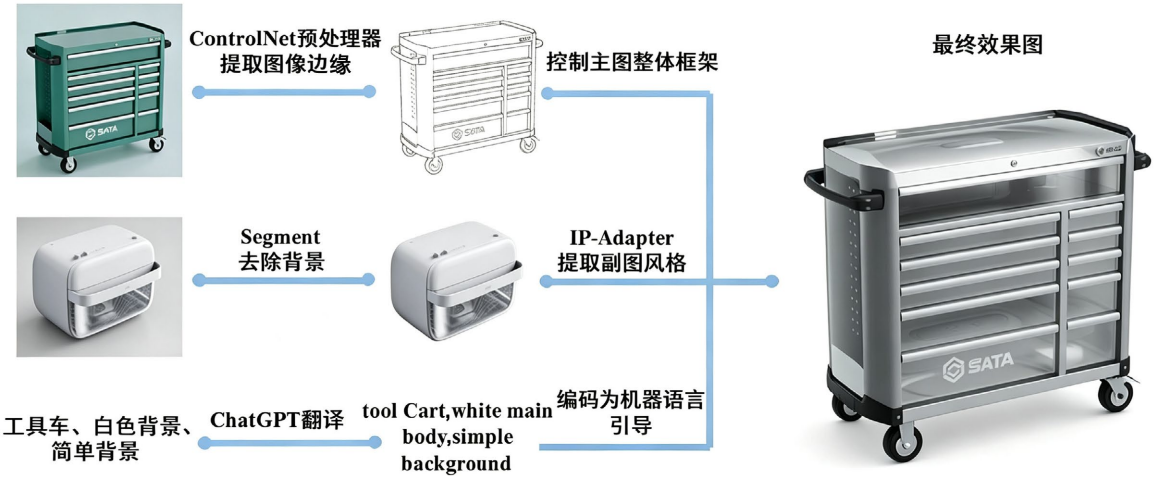

Figure 7. The implementation route of IP-Adapter combined with ControlNet
图 7. IP-Adapter 加 ControlNet 实现路线

4. 局限性与未来工作

尽管本研究提出的基于 Flux 与 LoRA 微调的双向多模态控制矩阵在工具车的可控生成中表现出显著的工程实用性，但将其泛化至更广泛的工业设计领域时，仍存在若干亟待解决的挑战。

首先，当前方法在处理具有复杂拓扑结构的非正交曲面产品时面临建模精度的瓶颈。本研究选取的验证载体“工具车”具有典型的刚性矩形特征，模型能够较好地捕捉其物理边界。然而，对于汽车外饰、流线型小家电等涉及高阶非均匀有理 B 样条(NURBS)曲面的产品，扩散模型在潜空间中的生成逻辑往往难以精确维持曲率的连续性。在实际测试中发现，当设计指令涉及大面积流线型曲面时，模型生成的反光影调常出现断裂或扭曲，这表明模型对精密曲面拓扑的理解尚未达到工业生产级的 A 级曲面标准。

数据集的构建成本与质量控制亦是制约该范式大规模落地的关键因素。虽然引入 Qwen 模型与 Next-scene LoRA 辅助生成多视角数据，但为了实现“品牌 DNA”的深度内化，依然需要对原始素材进行像素级的背景剥离与边缘修补。这种对高质量、结构化数据的依赖，在面对产品品类众多的企业时，会导致前期研发的时间成本陡增。此外，自动化重构机制在处理遮挡关系复杂的零部件时，偶尔会产生虚假的几何信息，这些数据进入训练环节后，会显著降低最终模型的收敛质量。

从生成结果的鲁棒性来看，本研究也观察到了若干具有代表性的失败案例。当外部参考图提供的视觉风格与目标产品的原始几何结构存在严重的透视冲突或语义排斥时，模型常会出现“特征漂移”现象。例如，在尝试将一种具有高度复杂肌理的艺术风格迁移至具有细长把手的结构时，由于空间权重分配的竞争，把手接缝处往往会出现非自然的噪点堆叠或结构断层。这些案例表明，显式结构控制与隐式风格嵌入之间的协同平衡仍需更精细的动态权重调度算法支撑。

未来的研究工作将聚焦于引入更高维度的三维先验(如深度图与法线贴图的联合约束)，以强化模型对复杂曲面的物理感知识别。同时，探索基于少样本学习(Few-Shot Learning)的微调策略，旨在降低对海量标注数据的依赖，从而为装备制造企业提供更加敏捷、低成本的数字化设计解决方案。

5. 结论

本文针对传统工业设计迭代低效，以及现有 AIGC 模型在严苛几何约束下易发生透视失真的痛点，提出了一种高精度、强约束的工业产品可控生成设计方法。研究确立了具备全局空间感知能力的 Flux 大模型作为底层基座，有效克服了传统 U-Net 架构在处理复杂机械结构时频发的空间逻辑悖论。为实现企业品牌基因的深度内化，本文构建了融合多视角重构技术的垂直领域数据集，并结合先验保护机制对模型进行了 LoRA 深度微调。量化评估证实，微调后的模型在第 9 轮达到最佳收敛，其 FID 指标降至 11.48 的全局最低点，能够极其精准地重构出符合真实工业分布的物理细节。

在此基础上，为突破单一文本提示带来的随机性与控制盲区，本研究进一步引入了 ControlNet 与 IP-Adapter 技术，构建起“显式几何结构锁定”与“隐式视觉风格迁移”的双向控制矩阵该系统不仅彻底实现了工业结构特征与艺术风格特征的正交解耦，还能在高保真地维持产品三维几何约束的前提下，赋予设计过程广域的美学泛化能力。这一全新的生成范式搭配前端可视化操作界面，大幅降低了设计的转化门槛，为装备制造企业化解“形技分离”断层、实现敏捷研发提供了切实可行的系统级解决方案。

参考文献

[1] 殷艳娜, 徐剑. 面向服务型制造的区域物流体系要素关系与特征识别——基于多案例的扎根理论分析[J]. 沈阳工业大学学报(社会科学版), 2020, 13(4): 332-339.

[2] Tang, X., Windham, J. and Bush, B. (2024) Pre-AI and Post-AI Design: Balancing Human Creativity and AI Tools in the Industrial Design Process. *Proceeding of the* 2024 *International Conference on Artificial Intelligence and Future Education*, Shanghai, 1-2 November 2024, 100-108. https://doi.org/10.1145/3708394.3708413

[3] 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**, 139-144. https://doi.org/10.1145/3422622

[4] Ho, J., Jain, A. and Abbeel, P. (2020) Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, **33**, 6840-6851.

[5] Hu, E.J., Shen, Y., Wallis, P., *et al*. (2022) LoRA: Low-Rank Adaptation of Large Language Models. 2022 *International Conference on Learning Representations*, Online, 25-29 April 2022, 1-20.

[6] Zhang, L., Rao, A. and Agrawala, M. (2023) Adding Conditional Control to Text-to-Image Diffusion Models. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 3836-3847. https://doi.org/10.1109/iccv51070.2023.00355

[7] Ye, H., Zhang, J., Liu, S., *et al*. (2023) IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. https://doi.org/10.48550/arXiv.2308.06721

[8] Heusel, M., Ramsauer, H., Unterthiner, T., *et al*. (2017) GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems*, **30**, 6629-6640.

[9] Peebles, W. and Xie, S. (2023) Scalable Diffusion Models with Transformers. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 4199-4209. https://doi.org/10.1109/iccv51070.2023.00387

[10] Jun, W., Tianliang, Z., Jiahui, Z., Tianyi, L. and Chunzhi, W. (2023) Hierarchical Multiples Self-Attention Mechanism for Multi-Modal Analysis. *Multimedia Systems*, **29**, 3599-3608. https://doi.org/10.1007/s00530-023-01133-7

[11] Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., *et al*. (2023) Flow Network Matching: Generating Infinite Resolution Continuous-Normalizing Flows. *2023 International Conference on Learning Representations*, Kigali, 1-5 May 2023, 1-28.

[12] Liu, X., Gong, C. and Qi, L. (2023) Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. 2023 *International Conference on Learning Representations*, Kigali, 1-5 May 2023, 1-41.

[13] Labs, B.F., Batifol, S., Blattmann, A., *et al*. (2025) FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space.