

AI VIRTUAL TRY-ON

AI虚拟换装技术 最新进展报告

从「填空题」到「改写题」：CtrlVTON与OOTDiffusion如何重新定义虚拟试衣间

97.2%

VIP-SAM分割精度

3-5秒

单次推理时间

AAAI 2025

OOTDiffusion

2026年7月31日 · AI虚拟换装技术资讯

网购衣服总「见光死」？AI虚拟试衣间正在彻底改变这一困境。从「把衣服抠掉再填进去」的填空思维，升级为「让AI理解穿衣逻辑」的改写范式——NXN Labs与KAIST联合发布的CtrlVTON，配合AAAI 2025录用的OOTDiffusion，一场虚拟换装技术革命正在悄然发生。

核心要点

- CtrlVTON的语义级+像素级双重控制机制，从「填空题」升级为「改写题」
- OOTDiffusion的开源实现与商业落地，AAAI 2025录用水准
- 阿里妈妈、抖音电商等平台的规模化应用，电商退货率有望大幅降低
- 技术路线从图像修补迈向图像编辑，分割精度达97.2%，推理速度进入秒级时代

一、传统虚拟试衣的三大困境

绝大多数现有AI试衣系统采用「图像修补」(inpainting)方法——先把人物身上的衣服「擦掉」，再用AI把新衣服填进去。这个看似直观的思路，实际操作中会产生一系列连锁问题。

⚠ 传统方案三大痛点

1 服装变形失真

擦除范围太小，旧衣服残留干扰；范围太大，人物身份信息被抹

2 姿态不匹配

挖空区域形状影响生成结果，裤子的遮罩可能生成裙子

3 光影割裂

新衣服与原图光照条件不一致，合成图像不自然

这些麻烦源自一个根本性的设计逻辑：把试衣当成「填空题」。AI在填充时会同时受到挖空区域形状和外部环境的双重束缚，无法专注于「如何把这件衣服穿好」。

二、从「填空题」到「改写题」：CtrlVTON的范式革命

NXN Labs与韩国科学技术院（KAIST）的联合研究团队决定彻底换一套玩法：不再做填空题，改做「改写题」。其框架不再是「把衣服抠掉再填上新的」，而是「给AI看同一人的不同穿衣照 + 参考服装图，让它直接生成换装效果」。

2.1 三阶段数据质量筛选

训练这样一个模型需要大量「同一个人穿不同衣服」的三元组数据。 研发团队开发了一套自动化数据生产流水线，配合三阶段质量筛选机制：

阶段	方法	目的
第一阶段	视觉语言模型四问审核	人物身份/姿势/背景保留，修改范围正确
第二阶段	轮廓匹配分数(CMF)	剔除新旧衣服轮廓高度重叠的样本
第三阶段	人工标注员审核	最终质量把关

2.2 VIP-SAM：精确到「这件衣服」的分割技术

团队还发布了专门解决服装精确分割问题的VIP-SAM。在自建时装数据集上达到**97.2% mIoU**，在COCO-20i和PASCAL-5i标准测试集上分别达到74.0%和79.1%。

三、两种控制维度：语义级 + 像素级

CtrlVTON-base · 语义级控制

full_swap

全部替换同类服装

partial_swap

部分替换指定服装

add

叠加新服装

CtrlVTON · 像素级控制

◆

衬衫扎不扎进裤子

◆

拉链拉多高

◆

袖子卷不卷

◆

衣服宽松还是修身

四、OOTDiffusion：AAAI 2025的开源突破

OOTDiffusion是AAAI 2025录用的开源项目，全称「Outfitting Fusion based Latent Diffusion for Controllable Virtual Try-on」，通过潜在扩散模型实现可控的虚拟试穿体验。

4.1 核心技术架构

OOTDiffusion的技术核心在于其独特的Outfitting Fusion机制，通过三个关键组件实现：

1

服装特征编码器

使用CLIP-ViT模型对服装图像进行深度特征提取，将服装的视觉信息转化为高维向量表示，完整保留纹理、颜色和图案信息。

2

人体解析与姿态估计

通过ONNX格式模型进行精确的人体语义分割，识别18个关键身体区域，包括上半身、下半身和细节区域（面部、手臂、腿部等）。

3

扩散模型融合

采用两阶段UNet架构：单步融合UNet快速整合服装特征，多步去噪UNet精细优化生成结果。

4.2 性能基准

3-5秒

单次推理时间

10张/分

批量处理速度

8-12GB

显存占用(RTX 4090)

五、VIP-SAM：精确到「这件衣服」的分割革命

虚拟试衣的一个核心难题是：如何让AI精确识别「这件特定的格纹衬衫」，而不是把牛仔外套或整个上半身都圈进去？NXN Labs发布的VIP-SAM专门解决这一问题。

5.1 传统分割的困境

现有视觉参考分割模型（如VRP-SAM）的问题在于：负责处理目标照片的图像编码器是「固定」的，它在处理照片时并不知道我们在找什么。当照片里有两件颜色相近的叠穿衬衫时，这个编码器根本没有能力区分哪件是哪件。

5.2 VIP-SAM的改进思路

VIP-SAM的改进可以这样理解：不要等到最后一步才告诉AI我们在找什么，而是从一开始就让AI带着「寻找目标」的意识来看照片。具体实现方式是在图像编码器的每个处理阶段都插入「适配器」模块，这些适配器会把参考服装的特征信息注入到编码器处理目标照片的中间过程里。

VIP-SAM 性能对比

模型配置	COCO-20i	PASCAL-5i	自建时装数据集
Hiera-L + DINOv3-H	74.0%	79.1%	97.2%
ViT-B + ResNet-50 (轻量)	—	—	显著超越VRP-SAM

六、CtrlVTON的技术细节

6.1 通道维度的遮罩注入

CtrlVTON在空间控制能力上，通过在冻结的base模型上面叠加一个轻量级的LoRA适配器实现。遮罩信息通过通道拼接的方式注入到模型的潜在表示空间中。

注入方式对比实验

令牌维度拼接

IoU 0.8925

速度: 0.38步/秒 | 显存: 62.65GB

通道维度拼接 

IoU 0.9610

速度: 0.91步/秒 | 显存: 43.47GB

 通道维度拼接利用遮罩与图像天然的空间对齐关系，效率效果双赢

6.2 多服装场景的颜色编码

CtrlVTON还引入了一套颜色编码方案：每件参考服装被分配一种独特的RGB颜色，这种颜色在目标遮罩和对应的服装遮罩中保持一致，让AI能够在生成图像时同时处理多件服装，并准确地把每件服装放到正确的位置上。

6.3 训练成本

CtrlVTON-base是对预训练的图像编辑扩散变换器（FLUX.2 Klein）进行全参数微调，训练成本约20个H200 GPU天。空间控制能力通过LoRA适配器实现，大幅降低了新增功能的训练成本。

七、电商平台的规模化落地

7.1 阿里妈妈「元宇宙魔盒」升级

2026年7月22日，阿里妈妈正式发布了「元宇宙魔盒」的升级版本， 淘宝天猫全面接入3D视觉AI虚拟试衣。这一升级的核心卖点在于： 从「猜尺码」到「看上身」，实现电商购物体验的历史性跨越。

7.2 抖音电商「虚拟上身」功能

抖音电商于2026年第一季度开始内测「虚拟上身」功能。用户上传一张全身照，AI自动识别体型，然后推荐合适版型的衣服， 还能看到不同场景下的穿着效果。这一功能直接针对女装电商的高退货率痛点。



女装电商退货率问题

传统电商购物中，女装退货率居高不下，核心原因之一是「买家秀」与「卖家秀」差距过大。AI虚拟试衣间让用户在购买前就能看到「真实自己」穿上衣服的效果， 从源头降低因「不合适」导致的退货。

7.3 腾讯云AI模特解决方案

基于服装素材图与模特姿态图双输入的AI虚拟试衣，是当前计算机视觉与扩散生成模型交叉领域的核心技术。 主流技术栈包括： 人体姿态检测 + SAM语义分割 + 单目深度估计 + SMPL三维人体建模 + 网格形变场预测 + 多条件扩散渲染。

八、未来展望：从技术突破到生活方式

OOTDiffusion代表了AI虚拟试穿技术的前沿方向，未来发展方向包括：



实时交互体验

通过WebGL和WebGPU技术实现在线实时虚拟试穿，用户可以通过摄像头直接试穿虚拟服装。



多模态融合

结合文本描述、语音指令等多种输入方式，实现更加智能的服装推荐和试穿体验。



社交化功能

用户可以将虚拟试穿结果分享到社交媒体，形成用户生成内容的时尚社区。



个性化定制

基于用户的身材数据、风格偏好，提供个性化的服装设计 and 定制服务。

AI虚拟换装技术正在重新定义时尚电商的未来。无论是个人开发者还是企业用户，都可以基于OOTDiffusion等开源项目，构建属于自己的虚拟试穿解决方案。从简单的试穿应用到复杂的时尚推荐系统，AI在时尚领域的应用正在打开一扇全新的大门。