Weekly Status Report

Week of 06.09.2025/07.09.2025 - 8 hrs

Report: Assess existing Virtual try-on models [both commercial and non-commercial]

1. IDM-VTON [https://huggingface.co/yisol/IDM-VTON] Or Improved Diffusion Models for Virtual Try-ON.
   It consists of two different components: 1) the image prompt adapter (IP-Adapter) that encodes the high-level semantics of the garment, and 2) the UNet encoder, which is GarmentNet, that extracts low-level features to preserve fine-grained details.

License: Creative Commons Attribution Non Commercial Share Alike 4.0 (CC-BY-NC-SA-4.0)
Components: IP-Adapter for base codes.
OOTDiffusion and DCI-VTON for masking generation.
SCHP for human segmentation.
Densepose for human densepose.

2. Virtual Try-On tool using IP-Adapter
   [https://huggingface.co/blog/tonyassi/virtual-try-on-ip-adapter, https://github.com/tencent-ailab/IP-Adapter]
   t IP-Adapter, an effective and lightweight adapter to achieve image prompt capability for the pretrained text-to-image diffusion models. The key design of our IP-Adapter is decoupled cross-attention mechanism that separates cross-attention layers for text features and image features. IP-Adapter is reusable and flexible. IP-Adapter trained on the base diffusion model can be generalized to other custom models fine-tuned from the same base diffusion model. Moreover, IP-Adapter is compatible with other controllable adapters such as ControlNet, allowing for an easy combination of image prompt with structure controls.

Components: utilize the open-source SD model as our example base model to implement the IP-Adapter. SD is a latent diffusion model conditioned on text features extracted from a frozen CLIP text encoder. The architecture of the diffusion model is based on a UNet with attention layers.
License: Apache License Version 2.0

https://sm4ll-vton.github.io/sm4llvton/

https://huggingface.co/martintomov/rayban-meta-glasses-v1

3. SM4LL-VTON [https://huggingface.co/spaces/sm4ll-VTON/sm4ll-VTON-Demo]

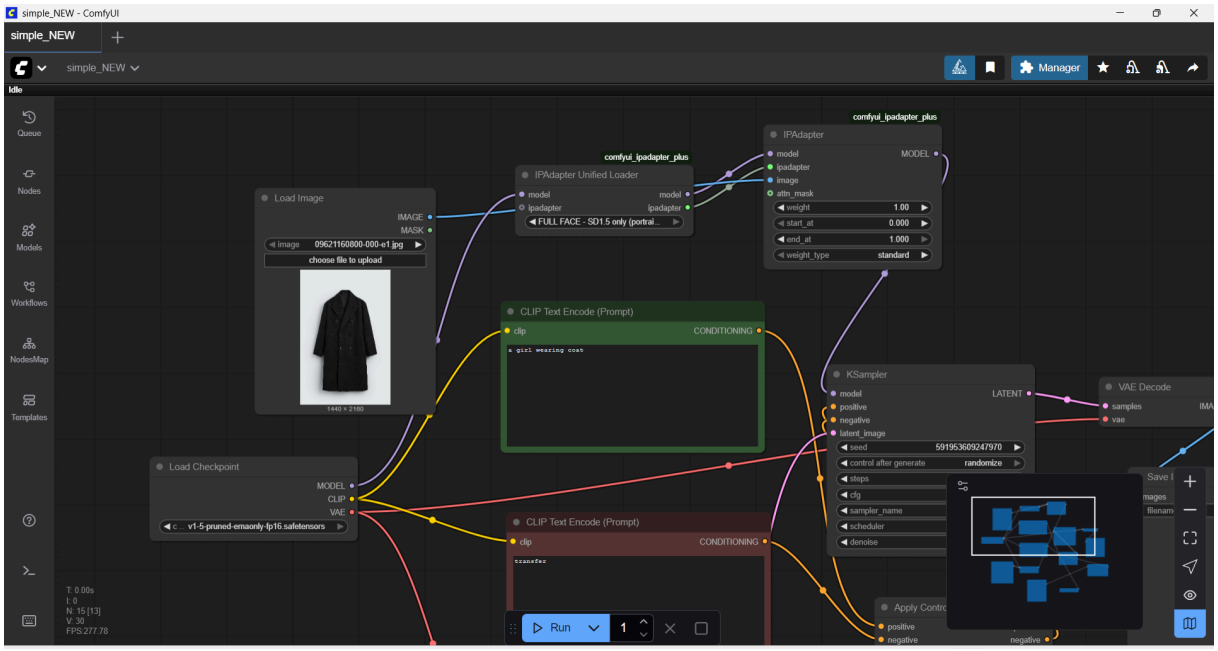
Release only demo. The sm4llVTONs family consists of several lightweight models, each an expert in a specific VTON domain. This specialization allows them to achieve state-of-the-art results on relatively small, targeted datasets.Inference is handled via ComfyUI, in an environment that mirrors the training conditions.

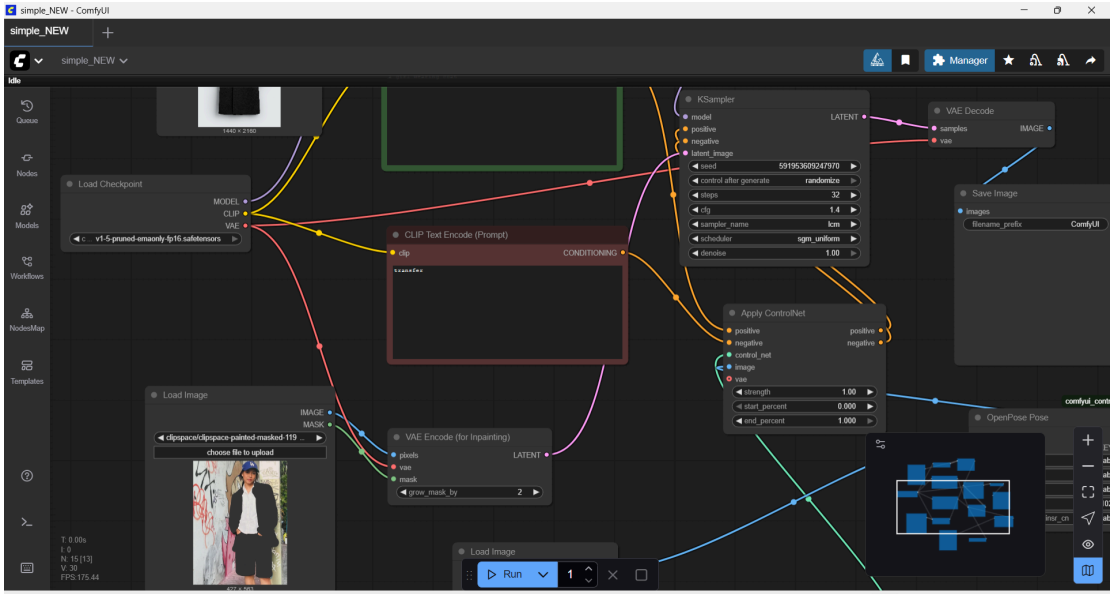
Components: They don't explicitly tell what base models have been used, no code or model released.
License: non-commercial license

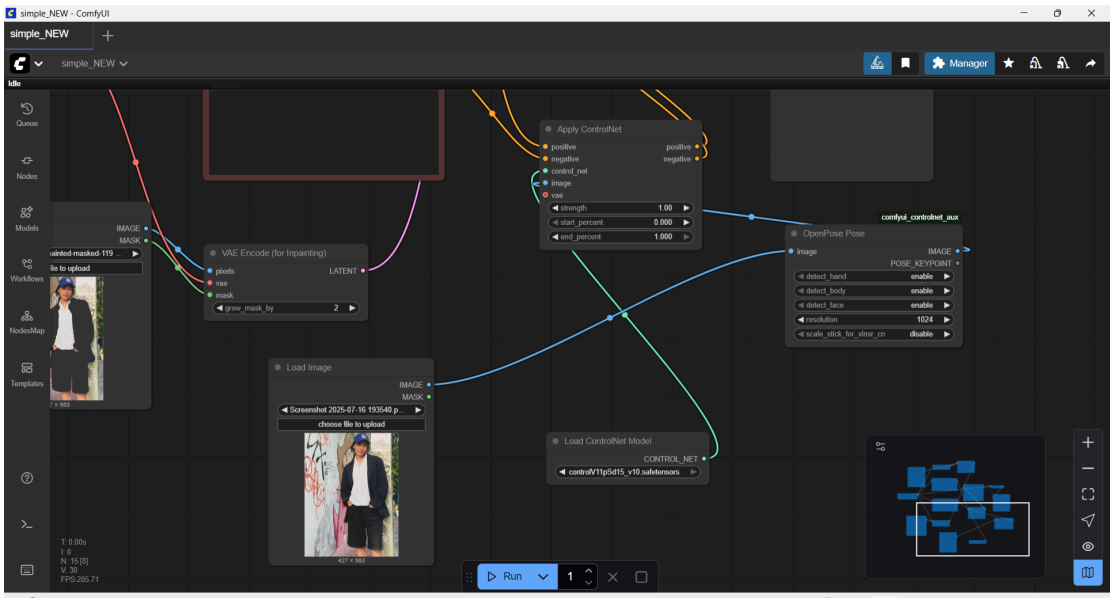
4. Comfy UI Virtual tryon

Base model: SD 1.5 prune
IPADAPTER: full face with CLIP VISION CLIP-ViT-bigG-14-laion2B-39B-b160k
CONTROLNET: controlnetv11sd15

Basic workflow/structure model connections [simplevt_v1 https://github.com/fashly-ai/fashly-research]







✅ Completed:
- Compile all existing VTON models on HuggingFace.
-successfully run the workflow on Comfy UI
-Config all required models components

🔄 In Progress:
-Explore feasibility on other open source models.
- Adjust and improve clothes quality on human body input

⚠️ Blockers:

- Most of the available models on HuggingFace are garment try on not sunglasses and they release on Non-commercial license.
- Virtual machine
- limitation on specific pretrained model on sunglasses

📅 Next Week:
- Try to config and explore/trial and error model components to improve hyper realistic quality Could be SDXL or FLUX instead of version 1.5

💡 Notes:
-Storage limitation on heavy vision models (base model usually 64 GB),
- Complicated parameters configuration
-the model seems to capture bodies better than sunglasses, but the quality needs to be improved more