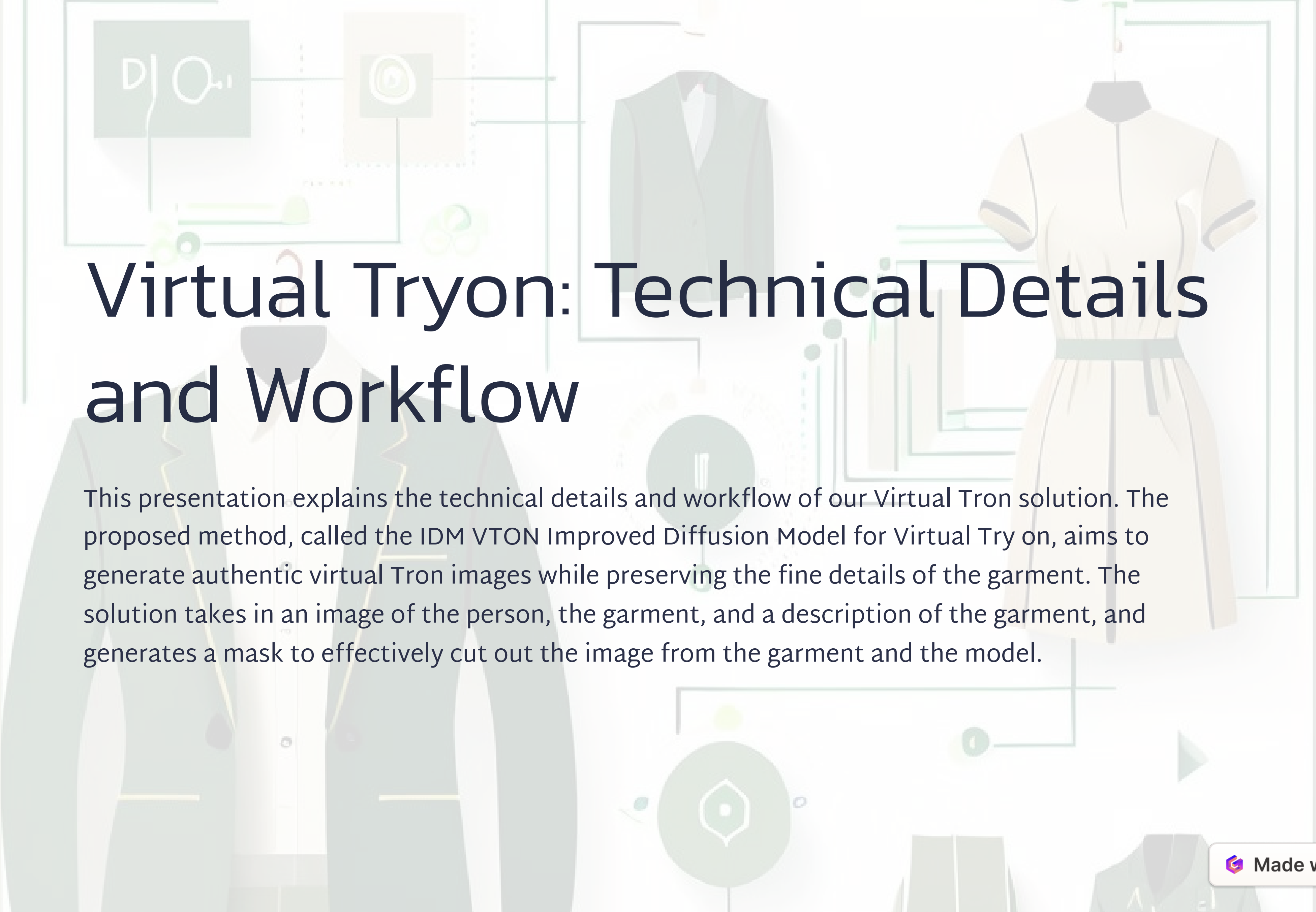
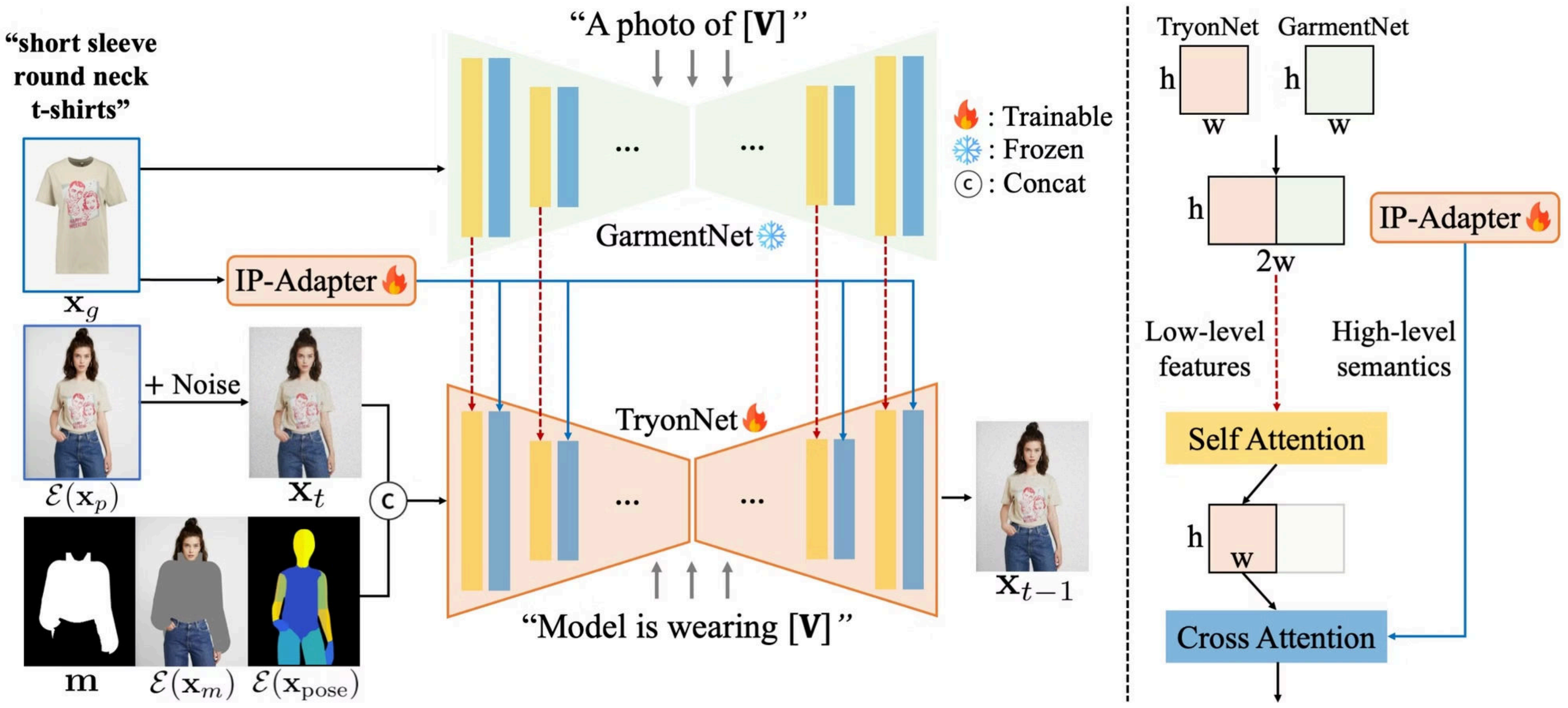
Virtual Tryon: Technical Details and Workflow

This presentation explains the technical details and workflow of our Virtual Tron solution. The proposed method, called the IDM VTON Improved Diffusion Model for Virtual Try on, aims to generate authentic virtual Tron images while preserving the fine details of the garment. The solution takes in an image of the person, the garment, and a description of the garment, and generates a mask to effectively cut out the image from the garment and the model.



Architecture of the Virtual Tryon Model



The architecture of the Virtual Tron model consists of several key components. The Tron Net is a base unit model that processes the person's image, taking in the latent representation of the person, segmentation mask, latent of the mask person, and the latent of the person's pose. The Image Prompt Adapter encodes the high-level semantics of the garment image using a frozen CLIP image encoder, and the encoded features are then fused with the Try on Net via a cross-attention mechanism. The Garment Net is a model unit encoder that extracts the low-level features from the garment image, and the extracted features are concatenated with the intermediate features from the Tron Net and passed through a self-attention layer to preserve the fine-grained details of the garment.

Leveraging Garment Descriptions

1

Garment Description

The description of the garment can be automatically generated using visual language models or provided by the user. This information is provided to both the Garment Net and the Try on Net to enhance the authenticity of the generated image.

2

Enhancing Authenticity

By incorporating the garment description, the model can leverage these details to generate a more authentic virtual Tron image, ensuring that the fine patterns and details of the garment are accurately represented.

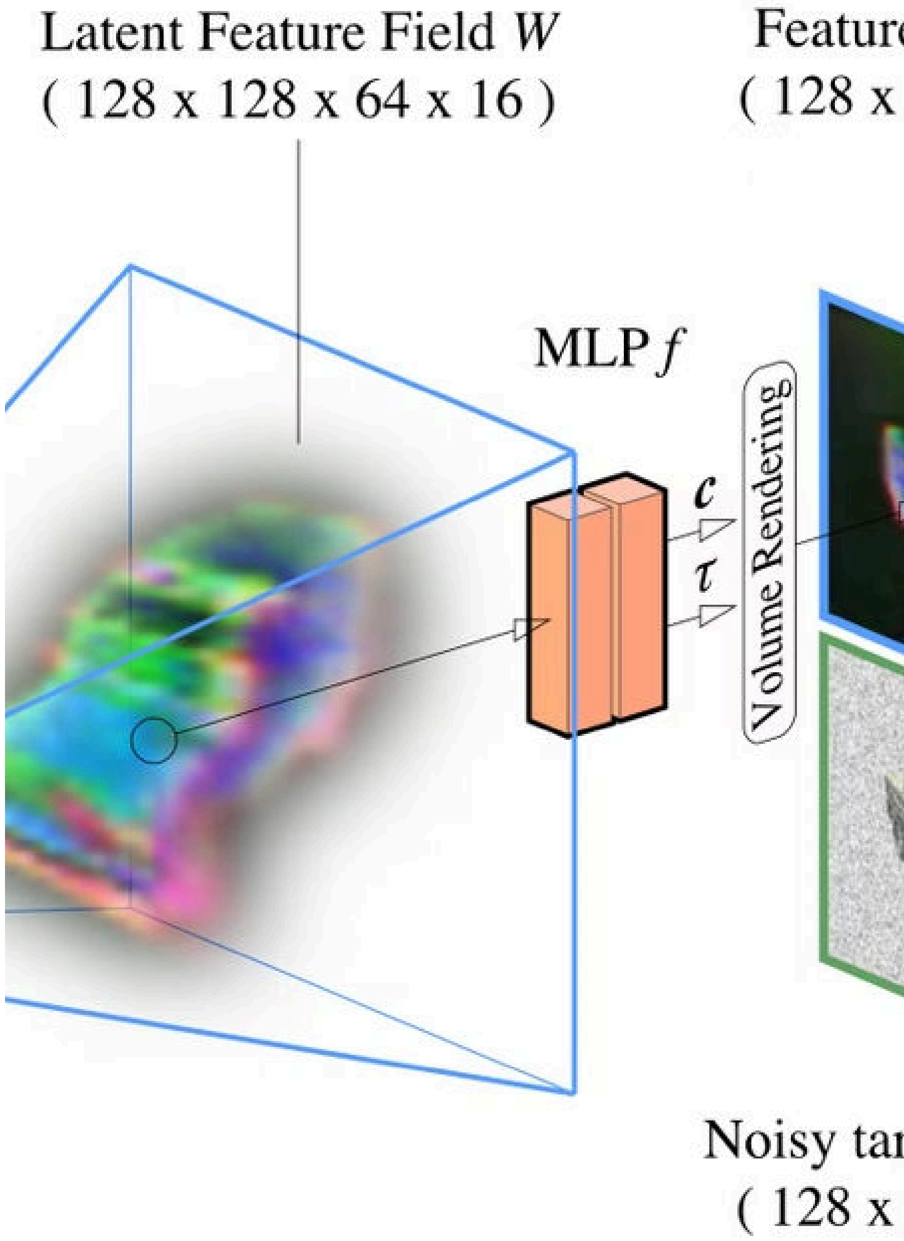
3

Combining Components

The complete Virtual Try on solution is a combination of the Tron Net, the Image Prompt Adapter, and the Garment Net, working together to produce the final result in about 15 seconds.

Novel View Synthesis

After generating the Virtual Try on image, the next step is to generate novel view synthesis in a GIF format. This is done using a fine-tuned version of the 0123++, a model used for generating novel view synthesis from a single image. The fine-tuned model is part of the Instant Mesh framework and has been trained to generate a 3x2 grid of six multi-view images from a single input image.



Fine-Tuning the 0123++ Model

1 Fine-Tuning Process

The fine-tuning process for the 0123++ model involved training it for 1,000 steps, using a linear noise schedule and V-prediction loss. The input images were randomly resized, and the target images were stitched into a 3x2 grid.

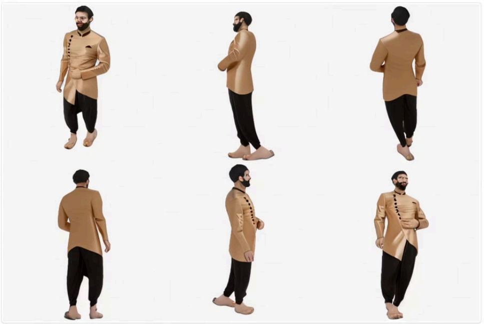
2 Background Optimization

The original model had an inconsistent gray background, which caused artifacts in the later reconstruction step. The fine-tuned model uses a white background instead, improving the overall quality of the generated images.

3 Limitations and Improvements

The resolution of the individual poses in the 6-view output is relatively low, so an XD-XL refiner is used to fix deformed facial features and increase the resolution to 4K. In the future, further fine-tuning of the 0123++ model could eliminate the need for this post-processing step.

Generating the Final GIF



Multi–View Synthesis

The fine-tuned 0123++ model generates a 3x2 grid of six multi-view images from the single input image. This provides a more comprehensive representation of the virtual Tron, allowing for better visualization and interaction.

GIF Conversion

The individual poses extracted from the 6-view output are then converted to a GIF format using a Python function. This final GIF representation provides a dynamic and engaging way to showcase the virtual Tron solution.

Improved Quality

The Virtual Tryon solution, with its fine-tuned 0123++ model and post-processing steps (Magic Image Refiner), produces a higher-quality GIF output compared to traditional Image to 3D methods, offering a more realistic and immersive experience for users.

Performance and Optimization

Total Time	Approximately 15 seconds for the Virtual Tron generation, and less than a minute for the final GIF output when hosted on dedicated servers with GPUs.
Limitations	The current setup uses Replicate cloud servers that shut down after few seconds of idle time due to budget and compute limitations. Hosting the solution on dedicated servers with GPUs can significantly improve the performance and reduce the overall processing time.
Future Improvements	Further fine-tuning of the 0123++ model to generate higher-quality multi-view images could eliminate the need for the Magic Image refiner, reducing the overall processing time and improving the efficiency of the Virtual Tryon solution.

VTON Tab

The VTON tab is responsible for virtual try-on. It takes a garment image and a model image as input and generates an output image of the model wearing the garment. The tab uses the Replicate API to run the IDM-VTON model, which is a diffusion-based model for virtual try-on.

Code Flow

1. The user uploads a garment image and a model image using the file uploader.
2. The uploaded images are saved to temporary files using `tempfile.NamedTemporaryFile`.
3. The Replicate API is used to run the IDM-VTON model with the uploaded images as input.
4. The output image is generated and displayed in the output area.

Custom Functions

- `enhance`: This function takes an image as input and enhances it using the Replicate API with the IDM-VTON model.
- `create_gif`: This function takes an image as input and generates a GIF using the enhanced image.

Garment Search Tab

The Garment Search tab is responsible for searching for garments based on a search query. It uses web scraping to extract garment images from various websites.

Code Flow

1. The user enters a search query and selects a website (Bewakoof, Amazon, or Ajio).
2. The search query is used to scrape garment images from the selected website using BeautifulSoup and Selenium.
3. The scraped images are displayed in a grid layout.

Custom Functions

- `scrape_from_bewakoof`, `scrape_from_amazon`, `scrape_from_ajio`: These functions are used to scrape garment images from the respective websites.

Multi-View Tab

The Multi-View tab is responsible for generating multiple views of a garment using a diffusion-based model.

Code Flow

1. The user uploads an image of a garment.
2. The uploaded image is used as input to the Zero123++ model, which generates six views of the garment.
3. The generated views are refined using the Magic Image Refiner model.
4. The refined views are used to generate a GIF using the `create_gif` function.
5. The GIF is displayed in the output area.

Custom Functions

- `gen_mul_six`: This function takes an image as input and generates six views of the garment using the Zero123++ model.
- `refine_images`: This function takes an image as input and refines it using the Magic Image Refiner model.
- `create_gif`: This function takes an image as input and generates a GIF using the refined views.

Models

- IDM-VTON: A diffusion-based model for virtual try-on.
- Zero123++: A diffusion-based model for generating multiple views of a garment.
- Magic Image Refiner: A model used to refine the output of the Zero123++ model.

Magic Image Refiner Model

The Magic Image Refiner model is used to refine the output of the Zero123++ model. It takes the generated views as input and enhances them to produce high-quality images. The model is based on a diffusion-based approach and is fine-tuned on a dataset of high-quality images.