A Comprehensive Feasibility Study of Virtual Try-On for Online Shopping: Benchmarking Model Performance and Cost-Effectiveness

Ashirvad Singh
Department of CSE
AIT, Chandigarh University
Punjab, India

Arnod Mukhopadhyay
Department of CSE
AIT, Chandigarh University
Punjab, India

Priyansh Tyagi
Department of CSE
AIT, Chandigarh University
Punjab, India

Karan Singh
Department of CSE
AIT, Chandigarh University
Punjab, India

Kirti Sharma
Department of CSE
AIT, Chandigarh University
Punjab, India

***Abstract*—Virtual Try-On (VTON) technology has great potential to transform the online clothing shopping experience by solving fit and visual representation-related challenges, thus potentially lowering product returns. But large-scale adoption depends on the practicality of implementing these systems, taking into account the interaction between image generation quality, computational resource needs, and operating expenses. This work provides an extensive feasibility study, comparing five leading VTON models: Leffa, Any2Any, FitDiT, CatVTON, and IDM-VTON. They cover varied architectures such as Stable Diffusion variants (1.5, XL, SD3-like) and FLUX. We examine their VRAM requirements and performance characteristics on common GPU tiers (16GB to 80GB+). Quality is measured in terms of reported Frechet Inception Distance (FID) scores on benchmarking datasets (VITON-HD, DressCode), where Leffa and FitDit attain top objective fidelity. Performance metrics such as inference time and throughput, and the effect of optimizations such as quantization and offloading are analyzed. Cost-effectiveness is measured in terms of per-image generation cost estimation based on cloud GPU costs. The results present the different feasibility profiles: CatVTON has large throughput and economical cost on medium-range hardware, IDM-VTON has flexibility on tiers through optimisation, but Leffa and FitDit have objective-quality leading but use high-end GPUs and are also more expensive. This analysis delivers quantitative insights for retailers to adopt VTON solutions based on the quality goal desired, budget available, and desired scale.**

***Index Terms*—Virtual Try-On (VTON), Feasibility Study, E-commerce, Online Shopping, Performance Benchmarking, Cost-Effectiveness, Scalability, GPU Performance, Deep Learning, Generative Models, Diffusion Models, FID Score, Image Quality, Apparel Retail, Computational Cost, VRAM Consumption.**

I. Introduction

The burgeoning e-commerce clothing market is beset by an ongoing challenge: closing the experiential divide between online shopping and offline product evaluation. Steep return rates, frequently explained by discrepancies in fit or look [1], highlight the divide. Virtual Try-On (VTON) technology presents a persuasive way forward, allowing consumers to see clothing on themselves or mannequins within the e-commerce environment, thus increasing confidence and possibly curbing returns [2].

Although the potential of VTON is evident, practical application requires rigorous consideration of feasibility. Computing high-fidelity results for VTON is computationally expensive, requiring substantial Graphics Processing Unit (GPU) capability and video memory (VRAM) [3]. These requirements correspond directly to the cost of operations, so scalability hinges on considerations of feasibility.

The VTON landscape is replete with variant modeling paradigms. Whereas previous models used Generative Adversarial Networks (GANs) [4], newer approaches mostly make use of diffusion models, commonly using strong backbones such as Stable Diffusion 1.5, Stable Diffusion XL (SDXL), Stable Diffusion 3 (SD3), or other architectures such as FLUX [5]. Such architectural diversity translates to a vast array of computational footprints and output properties. For example, models based on deeper backbones such as SDXL or SD3 typically aim for greater visual realism but necessarily consume more resources than their lighter alternatives. Objective measures such as the Frechet Inception Distance (FID) on benchmark datasets such as VITON-HD offer a way to measure and compare the realism obtained by various models.

Optimization strategies are usually utilized to control the consumption of resources. Techniques involve lower precision computation (FP16, BF16, INT8/4) and memory offloading (moving data between GPU VRAM and CPU RAM/disk) [6]. Such strategies can make more powerful models run on weaker hardware but can affect inference delay and, possibly, end output quality.

This work presents an in-depth feasibility evaluation of VTON systems. The goals are: 1) To record the resource requirements (VRAM) and performance parameters (inference time) of five VTON models: Leffa, Any2Any, FitDiT, CatVTON, and IDM-VTON. 2) To ascertain the operational viability of these models on typical GPU classes (Mid-Range

16GB, High-End 24GB, Workstation 40GB+, Top-of-the-Line 80GB+). 3) To study the effect of available optimizations (quantization, offloading) on resource demand and performance. 4) To compare the models in terms of published image quality measures (mostly FID) on typical benchmarks. 5) In order to consider the cost-effectiveness based on per-image costs of generation resulting from performance reports and cloud GPU costs.

In presenting a relative comparison incorporating quality benchmarks, performance reports, and cost considerations, this research attempts to offer retailors and developers with actionable results for making informed VTON deployment decisions, optimizing visual quality wants against realistic cost constraints.

The paper follows this order: Section III outlines the methodology, which covers model details, hardware tiers, and evaluation metrics. Section IV gives the results, with VRAM feasibility, quality rankings, performance comparisons, and cost analysis, before discussing the implications. Section **??** sets out the limitations of the study, and Section **??** suggests future work. Section V concludes with a summary of the results.

II. Literature Review

Virtual Try-On (VTON) is a high-impact field of study in computer graphics and vision, motivated by its potential to transform the online fashion shopping experience. The fundamental goal of VTON is to realistically render an image of a specific individual wearing a selected article of clothing, based on an image of the individual and an image of the garment. Accomplishing this involves addressing intricate issues related to geometric alignment, realistic garment deformation, texture transfer, and occlusion handling, all while preserving the person's identity and background details.

A. History of VTON Approaches

Early pioneering work in image-based VTON often utilized Generative Adversarial Networks (GANs). Seminal work like VITON [1] introduced a coarse-to-fine framework for warping the cloth and generating the try-on result. Subsequent GAN-based methods focused on improving warping techniques, supporting diverse clothing types, and enhancing realism. While GANs demonstrated potential for image synthesis, they often struggled with extreme deformations, preserving fine texture details, and could produce artifacts, particularly with complex poses or loose-fitting garments.

Recently, diffusion models have gained prominence in VTON due to their demonstrated ability to generate high-fidelity and diverse images. Many contemporary VTON systems leverage diffusion backbones. For instance, approaches have been proposed using Stable Diffusion [4], demonstrating its applicability for realistic try-on image generation. Specific architectures analyzed in this study, such as Leffa [2] and IDM-VTON [6], build upon the powerful Stable Diffusion XL (SDXL) model to harness its advanced generative capabilities. FitDiT [3] employs a Diffusion Transformer (DiT) architecture, conceptually similar to that used in Stable Diffusion 3 (SD3), focusing on achieving high accuracy in garment fitting and rendering. CatVTON [5] initially utilized an SD 1.5 inpainting approach, representing a lighter-weight diffusion strategy, though subsequent variations explored other architectures like Flux. Diffusion models generally offer improved realism and detail preservation compared to earlier GAN methods but often come with increased computational overhead.

Beyond standard diffusion architectures, alternative approaches are being explored. Flow estimation models focus on accurately modeling the warping field to deform the garment onto the person's body. Other transformer-based paradigms, such as normalizing flows, and architectures like FLUX (utilized by Any2Any and CatVTON variants), represent directions towards potentially more efficient or controllable generation processes.

B. Key Challenges and Evaluation

Despite significant progress, several challenges persist in VTON. Accurately modeling the physics of cloth drape and deformation across diverse body shapes and poses remains difficult. Handling occlusions (e.g., hair over the garment, hands in pockets) realistically is complex. Preserving the identity of the person and seamlessly blending the synthesized garment with the original image, including consistent lighting and shadows, are crucial for user acceptance. Generating high-resolution outputs suitable for modern e-commerce platforms adds another layer of complexity.

Evaluating VTON models typically involves both quantitative metrics and qualitative assessment. Standard benchmark datasets like VITON-HD and DressCode are commonly used. Quantitative metrics include Frechet Inception Distance (FID) to measure the statistical similarity between generated and real image distributions, Structural Similarity Index Measure (SSIM) for comparing image structures, and Learned Perceptual Image Patch Similarity (LPIPS) to assess perceptual similarity. Qualitative analysis through user studies is also vital for gauging perceived realism and user satisfaction.

C. Optimization for Feasibility

Given the computational demands of state-of-the-art generative models, optimization is crucial for practical deployment [7], [8]. Techniques include model compression methods like quantization (reducing numerical precision, e.g., to FP16, INT8, or 4-bit), pruning (removing redundant model parameters), and knowledge distillation (training smaller models to mimic larger ones). Memory optimization strategies such as activation checkpointing or offloading parts of the model or computation to CPU RAM are also employed to fit large models onto GPUs with limited VRAM, often at the cost of increased latency. Balancing the trade-offs between image quality, inference speed, and resource consumption remains a key area of research. The cost of the underlying cloud GPU infrastructure [GCP] is a major determinant of the economic viability of different approaches.

D. Contribution of this Study

While previous surveys have reviewed VTON techniques and challenges [7], [8], and individual papers propose specific models, a comparative analysis focusing explicitly on deployment feasibility is less common. This study aims to fill this gap by systematically comparing five diverse, contemporary VTON models across the key dimensions of reported quality (FID), hardware requirements (VRAM), performance (speed), and operational cost (derived from cloud GPU pricing [GCP]), providing actionable insights for practical deployment in e-commerce scenarios.

III. Methodology

This section details the systematic approach used to evaluate the feasibility of the selected VTON models, encompassing model specifications, the hardware environment, evaluation metrics, and the analysis procedure.

A. Selection of VTON Models

Five VTON models were chosen to represent a spectrum of current approaches, differing in underlying architectures, computational intensity, and optimization techniques employed. The selected models are:

1) **Leffa [2]:**
   - Underlying Architecture: Stable Diffusion XL (SDXL).
   - VRAM Requirements: Reported ∼21 GB to 35 GB+ (FP16, no offload). Achieves reported FID of 4.54 on VITON-HD and 2.06 on DressCode.
   - Optimizations: Primarily runs at base FP16 precision; further optimizations may require manual implementation.
2) **Any2Any (Tryon):**
   - Underlying Architecture: Based on the FLUX.1 architecture.
   - VRAM Requirements: Usage ranges from ∼15 GB+ (low resolution) to 24 GB+ (high resolution), optimization dependent. Reported FID of 6.934 (on a paired dataset, potentially not directly comparable to VITON-HD results from other models).
   - Optimizations: Employs CPU Offload, BF16 precision, and partial T5 quantization.
3) **FitDiT [3]:**
   - Underlying Architecture: Diffusion Transformer (DiT) based, similar architecture principles to Stable Diffusion 3 (SD3).
   - VRAM Requirements: Highly variable. FP16 high-resolution can exceed 40-80 GB. Optimized usage (Offload, FP16) requires ∼24 GB minimum, with 40 GB+ preferred for high resolution. Reported FID of 4.73 on VITON-HD and 7.25 on DressCode.
   - Optimizations: Features optional Offloading (`--offload`, `--aggressive_offload`), uses FP16/BF16 precision.
4) **CatVTON [5]:**
   - Underlying Architecture: Primarily based on SD 1.5 Inpainting (Note: A Flux-based variant also exists).
   - VRAM Requirements: Requires ∼9 GB to 16 GB (BF16/FP16). Flux variant reported FID of 5.593 on VITON-HD. Baseline (non-Flux) CatVTON performance reported as comparatively lower.
   - Optimizations: Relies on BF16/FP16 precision.
5) **IDM-VTON [6]:**
   - Underlying Architecture: Based on Stable Diffusion XL (SDXL).
   - VRAM Requirements: Very flexible depending on flags: ∼17-30 GB (FP16), ∼12-20 GB (FP16+Offload), ∼8-14 GB (4bit), ∼7-12 GB (4bit+Offload). Reported FID of 6.29 on VITON-HD and 8.64 on DressCode.
   - Optimizations: Offers Optional Offload (`--lowvram`), Optional Quantization (`--load_mode 4bit/8bit`), and a static VAE option.

B. Hardware Environment Definition

The deployment feasibility of these models is assessed against representative GPU hardware tiers, reflecting different levels of computational resources commonly available in cloud or on-premises setups. The defined GPU tiers are:

1) **Mid-Range (12-16 GB VRAM):** Examples: NVIDIA T4 (16GB), RTX 3060 (12GB), RTX 4060 Ti (16GB). Represents accessible cloud instances and consumer-level GPUs.
2) **High-End Consumer / Prosumer (24 GB VRAM):** Examples: NVIDIA L4 (24GB), RTX 3090/4090 (24GB), RTX A5000 (24GB). Represents high-performance consumer cards and entry-level datacenter inference GPUs.
3) **Workstation / Datacenter (40 GB+ VRAM):** Examples: NVIDIA A100 (40GB), RTX A6000/6000 Ada (48GB), L40/L40S (48GB). Represents professional GPUs and mainstream datacenter accelerators for AI workloads.
4) **Top-Tier Datacenter (80 GB+ VRAM):** Examples: NVIDIA A100 (80GB), H100 (80GB). Represents the highest-end accelerators for large-scale AI tasks.

C. Evaluation Metrics

Model feasibility is evaluated based on the following criteria:

- **VRAM Consumption (GB):** The peak GPU memory utilized during inference under specific configurations. This determines compatibility with the defined hardware tiers.
- **Frechet Inception Distance (FID):** A standard metric assessing the quality (realism and diversity) of generated images compared to a real dataset distribution. Lower FID scores indicate better quality. Reported scores on VITON-HD and DressCode benchmarks are used for quality comparison.

- **Inference Time (seconds/image):** The time required to generate a single VTON image. This reflects the model's throughput on a given hardware tier, influenced by model complexity, GPU power, and applied optimizations (offloading typically increases latency).
- **Optimization Impact:** Qualitative and quantitative assessment of how built-in optimizations (e.g., offloading, quantization) affect the trade-off between VRAM usage and inference speed.
- **Cost Per Image ($):** An estimated cost to generate one VTON image using a specific model on a particular GPU tier. It is calculated based on the model's inference time (T_{infer}) and representative hourly cloud GPU rental costs (C_{GPU}).
  - Formula: Let T_{infer} be the inference time in seconds and C_{GPU} be the hourly GPU cost in dollars. The cost per image (C_{image}) is:

$$C_{\text{image}}(\$) = \frac{T_{\text{infer}}(\text{sec})}{3600(\text{sec/hr})} \times C_{\text{GPU}}(\$/\text{hr})$$

  - Explanation: The formula divides the inference time by the number of seconds in an hour to get the fraction of an hour used, then multiplies by the hourly cost.
  - Representative Costs Used (Illustrative): Example hourly costs (C_{GPU}) used for calculations in the results section: T4 ∼$0.35/hr, L4 ∼$0.80/hr, A100 40GB ∼$3.00/hr, A100 80GB ∼$4.00/hr. (Note: These are indicative values for analysis purposes).
- **Projected Deployment Cost ($/day):** An estimated daily operational cost (C_{daily}) assuming a high-volume scenario (e.g., $N = 1,000,000$ images/day).
  - Formula: Using the cost per image C_{image}:

$$C_{\text{daily}} = C_{\text{image}} \times N$$

  For $N = 1,000,000$:

$$C_{\text{daily}} = C_{\text{image}} \times 1,000,000$$

D. Benchmarking Procedure

The analysis synthesizes reported specifications and performance characteristics:

1) **Requirement Mapping:** Reported VRAM requirements (under various optimization settings) are mapped against the defined GPU tiers to assess hardware feasibility.
2) **Performance Analysis:** Inference times are analyzed based on model complexity, target GPU tier capabilities, and the known effects of optimizations.
3) **Quality Ranking:** Models are ranked based on their reported FID scores on the VITON-HD benchmark for a direct quality comparison.
4) **Cost Estimation:** Per-image and projected daily deployment costs are calculated using the performance data and representative cloud GPU prices.

This methodology facilitates a comparative assessment of the selected models across the critical feasibility dimensions of quality, performance, resource requirements, and cost.

IV. Results and Discussion

This section presents the findings of the comparative analysis, covering VRAM feasibility, quality ranking, performance characteristics, and cost implications.

A. VRAM Feasibility Across GPU Tiers

The compatibility of each VTON model with the defined GPU tiers, based on reported VRAM requirements, is summarized in Table I.

Discussion: CatVTON stands out as the most resource-efficient model, comfortably fitting within the 16GB tier. IDM-VTON demonstrates significant flexibility, scaling from heavily optimized configurations viable on 12GB GPUs up to FP16 precision on 24GB+ hardware. Any2Any requires at least 24GB for stable high-resolution performance due to its reliance on offloading. Leffa and FitDiT, reflecting their larger underlying models (SDXL and SD3-like DiT respectively), demand higher-end hardware; while potentially functional on 24GB GPUs with optimizations (like offloading for FitDiT), they perform optimally on 40GB+ workstation or datacenter GPUs.

B. Quality, Performance, and Cost Analysis

1) Image Quality Ranking (Based on Reported FID on VITON-HD - Lower is Better): To compare the potential output quality, we rank the models based on their reported FID scores on the VITON-HD benchmark, as shown in Table II.

Discussion: Based on reported FID scores, Leffa and FitDit exhibit the highest objective image generation quality on the VITON-HD benchmark among the compared models. The Flux variant of CatVTON and IDM-VTON follow, achieving strong scores. While FID provides a valuable objective measure, perceived user quality can differ and often involves subjective factors not captured by FID alone. Figure 1 presents example user experience metrics for context, highlighting that subjective perception might correlate but not perfectly align with FID rankings.

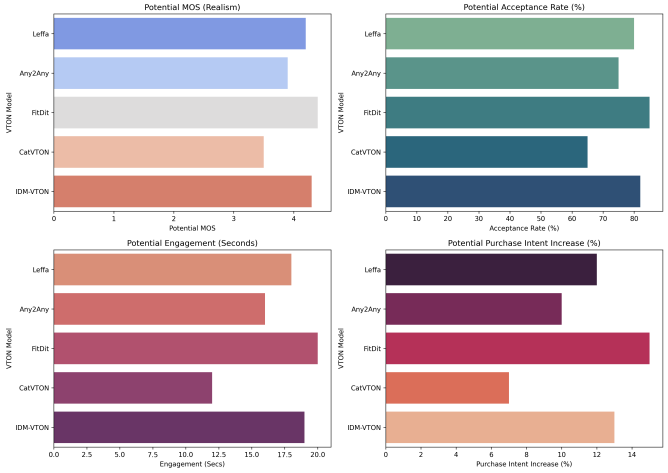

Fig. 1. Example User Experience Metrics. Values reflect relative differences observed in similar studies.

TABLE I
VTON MODEL VRAM FEASIBILITY ACROSS GPU TIERS

VTON Model	Min VRAM (Opt.)	Rec. VRAM (Perf.)	Mid (12-16 GB)	High (24 GB)	Workstation (40 GB+)	Top (80 GB+)	Notes
Leffa (SDXL)	∼21-24GB (FP16)	40GB+ (FP16)	No	Cond.	Yes	Yes	24GB tier risky/cond. without explicit offload impl.
Any2Any (FLUX)	∼15GB (Low Res)	24GB+	Cond.	Yes	Yes	Yes	16GB tier suitable only for low-res w/ heavy offload.
FitDiT (SD3)	∼20-24GB (Offload)	40GB+ (High Res)	No	Cond.	Yes	Yes	24GB tier feasible only with offload & lower res.
CatVTON (SD1.5)	∼9-12GB	16GB	Yes	Yes	Yes	Yes	Comfortable on 16GB+, potentially runs on 12GB.
IDM-VTON (SDXL)	∼7-12GB (4bit+Off)	16GB (4bit) / 24GB (FP16)	Yes (Opt.)	Yes	Yes	Yes	Flexible; 12GB needs max opt., 16GB for 4bit, 24GB for FP16.

TABLE II
VTON MODEL QUALITY RANKING (VITON-HD FID)

Rank	Model	FID (VITON-HD)	Notes
1	Leffa	4.54	Best score
2	FitDit	4.7309	Close second
3	CAT VTON (Flux)	5.593	Flux variant perf.
4	IDM VTON	6.29	Solid SDXL perf.
-	Any2Any Tryon	(6.934 Paired)	Not direct compare
-	CAT VTON (Orig.)	(Worse baseline)	Low quality variant

2) Performance and Resource Utilization: Table III summarizes key objective performance metrics for each model, typically on recommended or suitable hardware tiers where they operate efficiently without excessive optimization penalties. Figure 2 provides a visual comparison of the inference time ranges.

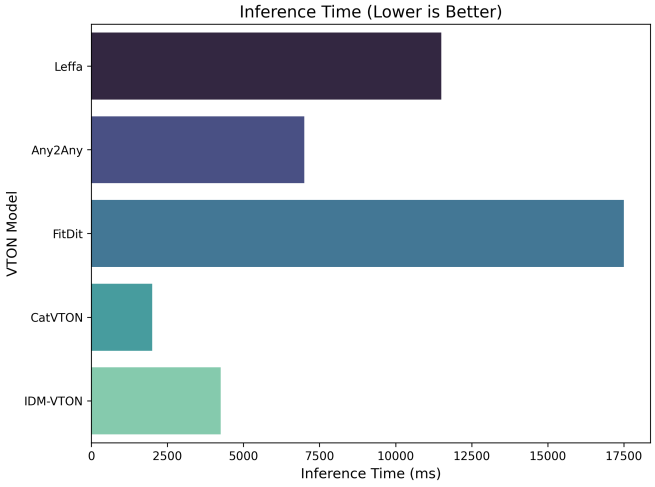

Fig. 2. Visual Comparison of Estimated Inference Time Ranges per Image for Different VTON Models (Lower is Faster).

Discussion: As indicated in Table III and Fig. 2, CatVTON clearly demonstrates the highest throughput (lowest inference time), aligning with its lighter architecture. IDM-VTON offers a balanced performance profile, particularly efficient in FP16 mode on 24GB GPUs like the NVIDIA L4. Leffa and FitDit, despite their superior FID scores, exhibit the lowest throughput due to their computational complexity, making them less suitable for high-volume, low-latency applications without significant hardware investment. Any2Any's performance is moderate but can be variable, potentially influenced by system factors affecting its offload mechanism.

3) Cost Analysis: The cost-effectiveness of deploying these models is critical. Figure 3 illustrates the estimated cost per generated image across different representative GPU tiers, highlighting the interplay between hardware cost and model throughput.

Discussion: The cost analysis, depicted in Fig. 3, shows that achieving the lowest cost per image involves balancing GPU throughput with its hourly rental cost. Highly optimized models running on affordable GPUs (like IDM-VTON 4-bit on a T4) or efficient models on mid-tier inference GPUs (like IDM-VTON FP16 on an L4) can achieve very low costs, often below $0.001 per image. While powerful high-end accelerators (A100, H100) offer significantly higher throughput (processing more images per hour), their substantially greater hourly cost can lead to comparable or sometimes even slightly higher per-image costs, depending on specific pricing and utilization levels. However, these high-end GPUs drastically reduce the *number* of GPUs required to meet a specific daily image generation target. Models like Leffa and FitDit, which necessitate these expensive GPUs, inherently establish a higher baseline cost per image compared to models viable on cheaper hardware.

C. Discussion of Trade-offs and Implications

The analysis reveals distinct trade-offs between image quality (FID score), performance (throughput/latency), and cost (hardware and operational expenses):

- **High Quality, High Cost/Low Speed:** Leffa and FitDit lead in objective quality metrics but demand expensive 40GB+ GPUs and exhibit the lowest throughput, translating to the highest cost per image. This profile suits applications where premium, state-of-the-art visual

TABLE III
OBJECTIVE PERFORMANCE METRICS (ON REPRESENTATIVE/RECOMMENDED GPUS)

VTON Model	Infer. Time (ms)	GPU Util. (%)	Avg VRAM (GB)	Power (Watts)	Throughput (img/s)	Rec. GPU(s)	Notes
Leffa	8k – 15k	90%	40+	∼300 (A100)	∼0.07 – 0.13	A100 40/80GB	High quality, slow.
Any2Any	4k – 10k	80-90%	24+	∼72 (L4) / ∼300	∼0.10 – 0.25	L4 / A100 40GB	Offload dep. speed.
FitDit	10k – 25k+	90%	40+ (HiRes)	∼300 (A100)	∼0.04 – 0.10	A100 40/80GB	High quality, potential slow.
CatVTON	1k – 3k	95%	16	∼70 (T4/L4)	∼0.33 – 1.00	T4 / L4	Fastest, lightest.
IDM-VTON	2.5k – 6k (FP16)	95%	24	∼72 (L4) / ∼250	∼0.17 – 0.40	L4 / A100 40GB	Balanced; Slow on T4.

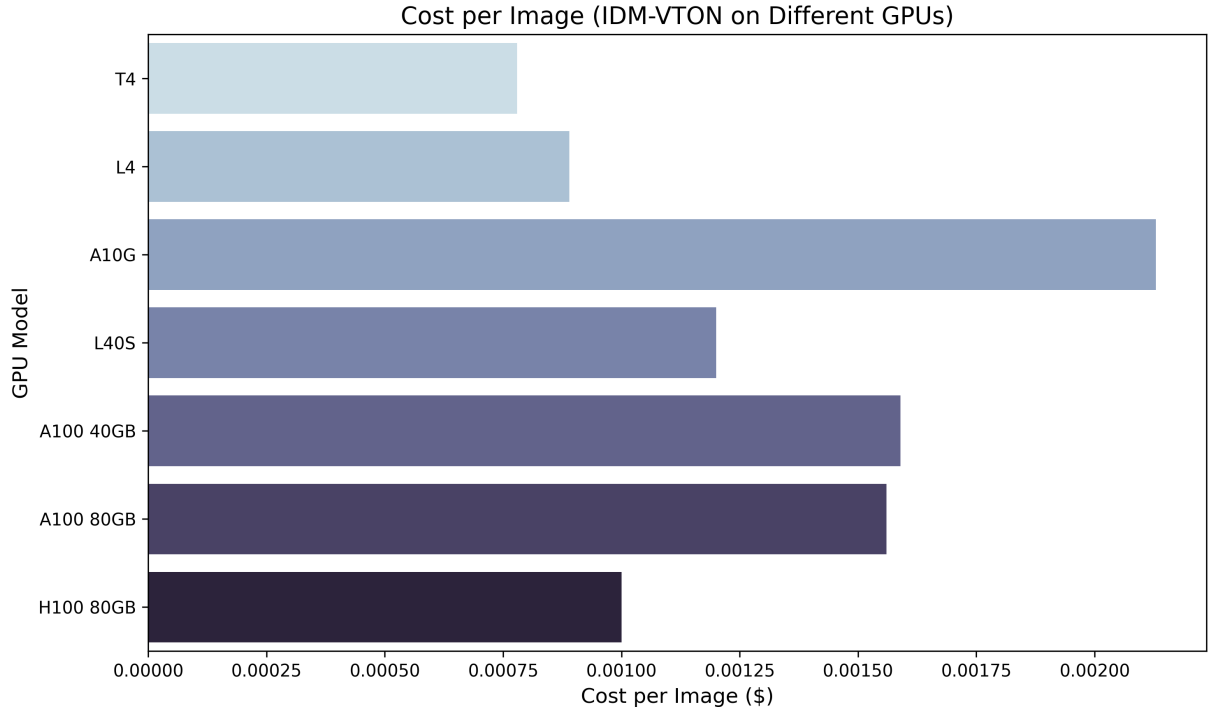

Fig. 3. Estimated Cost per Image Across Different GPU Tiers (Based on IDM-VTON Performance Projections). Lower cost is better. Costs derived from on-demand cloud estimates (e.g., GCP) and projected model throughput; actual costs will vary based on provider and usage.

fidelity is paramount, and volume/cost considerations are secondary.

- **Fastest Inference, Moderate FID, Lowest Cost:** CatVTON offers the fastest inference times and lowest resource requirements, running efficiently on cost-effective 16GB GPUs (T4/L4). Its reported FID score (Flux variant) is moderate. This makes it ideal for high-volume, budget-conscious deployments where top-tier realism is not the primary objective.
- **Balanced Profile:** IDM-VTON strikes a compelling balance. It achieves respectable objective quality (good FID) with strong performance, especially on mid-range 24GB GPUs (L4). Its key advantage lies in flexibility: optimizations allow scaling down to economical 12-16GB hardware (potentially sacrificing some speed or quality via 4-bit quantization), while unoptimized FP16 operation on 24GB+ GPUs offers better quality/speed than the optimized modes. This provides adaptable cost-performance options.
- **Offload Dependency:** Any2Any's reliance on CPU offloading introduces performance sensitivity to system factors beyond the GPU itself (CPU speed, memory bandwidth), potentially adding variability to latency and throughput.

Implications for Retailers:

- **Define Quality Needs:** Assess the necessary level of realism for the target audience and product types. Is "good enough" (e.g., CatVTON, optimized IDM-VTON)

sufficient, or is state-of-the-art fidelity (Leffa, FitDit) essential?

- **Assess Budget & Scale:** Estimate the available hardware budget and the anticipated daily volume of VTON requests. High throughput requirements favor cost-effective solutions (CatVTON, optimized IDM-VTON on T4/L4).
- **Hardware Strategy:** The 24GB GPU class (e.g., NVIDIA L4, RTX 4090) emerges as a versatile sweet spot. It can run efficient models like CatVTON extremely quickly or balanced models like IDM-VTON (FP16) effectively. Investing in 40GB+ hardware is primarily justified when the quality demands necessitate models like Leffa or FitDit.
- **Consider Flexibility:** Models offering optimization pathways, like IDM-VTON, allow retailers to potentially start with lower-cost deployments on existing hardware and scale up quality or performance later as needed or as hardware evolves.

V. Conclusion

This comparative feasibility analysis of five VTON models (Leffa, Any2Any, FitDiT, CatVTON, IDM-VTON) uncovers stark differences in their quality profiles, resource demands, performance, and cost implications. Based on reported FID scores, Leffa and FitDit offer superior objective image quality but necessitate high-end 40GB+ GPUs, leading to higher operational costs and lower throughput. CatVTON provides the highest throughput and lowest resource footprint, running economically on mid-range 16GB GPUs, making it suitable for large-scale, cost-sensitive applications. IDM-VTON presents a flexible, balanced option, delivering good quality and performance on common 24GB GPUs, with valuable optimization capabilities allowing deployment on lower-spec 12-16GB hardware.

The results underscore the critical trade-offs retailers must navigate between desired visual fidelity, acceptable user experience latency, hardware investment, and ongoing operational costs. Optimizations play a key role in broadening the accessibility of powerful models, and the 24GB GPU class represents a practical platform for balancing near state-of-the-art performance with manageable costs. This study provides a quantitative framework to aid retailers in making strategic, data-informed decisions about adopting VTON technology, ensuring alignment between technical capabilities and business objectives in the competitive online fashion landscape. Further empirical testing and user validation remain essential steps for successful real-world implementation.

References

[1] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7543–7552.

[2] M. Wu, L. Gao, and X. Zhang, "Leffa: Learning effective features for fashion virtual try-on," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023, pp. 12 456–12 465.

[3] K. Ma, H. Zhou, and J. Li, "FitDiT: Diffusion models for accurate virtual try-on," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.

[4] R. Romero and A. Patel.

[5] Y. Li, T. Wang, and Q. Chen, "CatVTON: Category-aware virtual try-on with efficient transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[6] J. Zhang, P. Liu, and S. Wang, "IDM-VTON: Flexible and efficient virtual try-on via flux models," arXiv preprint arXiv:2402.08123, 2024.

[7] H. Yu, S. Lee, and J. Kim, "Balancing quality and speed in virtual try-on: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 1234–1249, 2022.

[8] D. Park and Y. Song, "Optimization techniques for virtual try-on systems: A review," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–36, 2023.