VTEdit-Bench: A Comprehensive Benchmark for Multi-Reference Image Editing Models in Virtual Try-On

Xiaoye Liang[1,2], Zhiyuan Qu[1], Mingye Zou[2,3], Jiaxin Liu[1], Lai Jiang[1,4], Mai Xu[1,4,5], and Yiheng Zhu[2⋆]

[1] School of Electronic Information Engineering, Beihang University, Beijing, China
[2] Zhongguancun Academy, Beijing, China
[3] Faculty of Computing, Harbin Institute of Technology, Harbin, China
[4] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[5] Department of Nephrology, First Medical Center of Chinese PLA General Hospital; State Key Laboratory of Kidney Diseases; and National Clinical Research Center for Kidney Diseases, Beijing, China
xiaoyeliang@buaa.edu.cn, zhuyiheng@zgci.ac.cn
Project Page: https://github.com/Hiuyee124/VTEdit-Bench

Abstract. As virtual try-on (VTON) advances, a growing number of real-world scenarios have emerged, pushing beyond the ability of the existing specialized VTON models. Meanwhile, universal multi-reference image editing models have progressed rapidly and exhibit strong generalization in visual editing, suggesting a promising route toward more flexible VTON systems. However, despite their strong capabilities, the strengths and limitations of universal editors for VTON remain insufficiently explored due to the lack of systematic evaluation benchmarks. To address this gap, we introduce VTEdit-Bench, a comprehensive benchmark designed to evaluate universal multi-reference image editing models across various realistic VTON scenarios. VTEdit-Bench contains 24,220 test image pairs spanning five representative VTON tasks with progressively increasing complexity, enabling systematic analysis of robustness and generalization. We further propose VTEdit-QA, a reference-aware VLM-based evaluator that assesses VTON performance from three key aspects: model consistency, cloth consistency, and overall image quality. Through this framework, we systematically evaluate 8 universal editing models and compare them with 7 specialized VTON models. Results show that top universal editors are competitive on conventional tasks and generalize more stably to harder scenarios, but remain challenged by complex reference configurations, especially multi-cloth conditioning.

Keywords: Virtual Try-on · Benchmark · Multi-Reference Image Editing Models

⋆ Corresponding author

Table 1: Summary of supported tasks for specialized VTON models and universal multi-reference image editing models.

Task	Specialized VTON Model												Universal Editing Model
	2024				2025								
	SD. [23]	Stable. [12]	IDM. [4]	TPD [34]	FastFit [6]	MV. [26]	Omni. [35]	Any2any. [8]	Cat. [5]	OOTDiff. [33]	ITA. [10]	IMAG. [22]	
S2M	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
S2MM	×	×	×	×	×	×	×	×	×	×	×	×	✓
S2MV	×	×	×	×	×	×	×	×	×	×	×	×	✓
M2M	×	×	×	×	×	×	✓	×	✓	×	×	×	✓
MS2M	×	×	×	×	✓	×	×	×	×	×	×	×	✓

Task abbreviations: S2M (shop-to-model), S2MV (shop-to-multi-view), S2MM (shop-to-multi-model), M2M (model-to-model), MS2M (multi-shop-to-model).

Fig. 1: Overview of VTEdit-Bench, illustrating VTON tasks and dataset composition, the comparison between universal and specialized models, and the proposed unified evaluation metrics.

1 Introduction

In recent years, virtual try-on (VTON) has increasingly shaped online shopping and digital fashion. For consumers, it provides intuitive visualization of garments on themselves or virtual avatars, improving engagement and decision-making. For retailers, VTON enables interactive product experiences that can boost conversion and reduce returns. As e-commerce continues to expand, practical VTON systems are expected to support a broad range of real-world scenarios, as illustrated in Fig. 1(a), from Shop2Model to Model2Model and beyond. Developing robust and adaptable VTON models is therefore crucial.

Recent progress has been driven by specialized VTON methods [4–6,8,10,12, 22,23,25,26,33–35,37], which largely adapt text-to-image backbones with mask-based generation or inpainting. While they achieve strong results in canonical settings such as Shop2Model, extending them to more complex VTON scenarios remains difficult. Their task-specific designs, fixed input assumptions, and

reliance on auxiliary conditions often limit flexibility and scenario coverage, as summarized in Tab. 1. Meanwhile, universal multi-reference image editing models [13, 15, 18, 27–29, 31] have advanced rapidly. Following a text-conditioned image-to-image paradigm, they directly edit reference images according to instructions and naturally support multi-reference conditioning. As depicted in Fig. 1(b), this offers a simpler and more flexible pipeline than specialized VTON methods, making universal editors well aligned with diverse VTON scenarios. This points to a potential paradigm shift, where universal multi-reference editors may evolve into a unified VTON solution and reduce reliance on task-specific pipelines. However, their effectiveness and limitations for VTON remain underexplored due to the lack of a dedicated benchmark covering realistic and diverse VTON settings. Existing VTON benchmarks [14, 32] often focus on the most canonical and simplified try-on scenario, which is insufficient to reveal both the strengths and weaknesses of universal editors. Moreover, standard VTON evaluation still relies heavily on coarse distributional metrics (e.g., FID [9], KID [1]), which overlook key factors such as model consistency, cloth consistency, and overall image quality.

To address these gaps, we introduce VTEdit-Bench, the first comprehensive benchmark for evaluating universal multi-reference editing models across diverse VTON tasks. As shown in Fig. 1(a), VTEdit-Bench contains 24,220 test image pairs spanning five representative VTON settings, with task complexity progressively increasing from canonical VTON to more challenging scenarios involving viewpoint variation, identity changes, and multi-item composition. To support fine-grained evaluation, we further propose VTEdit-QA, a reference-aware VLM-based metric that assesses VTON quality along three essential dimensions: model consistency, cloth consistency, and image quality, complementing coarse metrics and enabling more informative diagnosis. We benchmark eight universal editors against seven specialized VTON models, showing that top universal models are competitive on conventional tasks and often generalize more stably to harder settings, while challenges remain in strict identity preservation and compositional multi-item control. Overall, VTEdit-Bench provides a unified foundation for evaluating VTON systems in the era of fast-evolving universal editing models. The main contributions are as follows:

- We introduce the first comprehensive benchmark tailored for evaluating universal multi-reference image editing models under VTON scenarios with progressively increasing complexity.
- We introduce VTEdit-QA, a reference-aware VLM-based metric for fine-grained VTON evaluation along three key dimensions: model consistency, cloth consistency, and image quality.
- We benchmark eight universal editing models against seven specialized VTON models, showing comparable performance on conventional tasks and stronger generalization to novel scenarios, while leaving room for further improvement.

2 Related Work

2.1 Image-based VTON

Image-based VTON aims to transfer a garment from a source image onto a target person image, and is widely used in online shopping. Most specialized VTON [4–6, 8, 10, 12, 22, 23, 26, 33–35] models build on text-to-image backbones, treating the person, garment, and other references as conditioning signals, and follow task-specific pipelines that are broadly mask-based or mask-free. Mask-based methods [4, 6, 10, 12, 22, 23, 26, 33–35] estimate auxiliary conditions (e.g., masks/parsing/pose) and perform region-aware synthesis or inpainting, offering explicit correspondence but depending on the quality of these conditions. Mask-free methods [8, 25, 37] remove explicit masks and learn try-on directly from garment and person images, improving input flexibility. As VTON settings expand beyond Shop2Model to multi-view, multi-identity, and multi-item scenarios, most specialized models remain limited to a subset of tasks due to fixed input assumptions and task-specific designs.

2.2 Instruction-based Image Editing Model

Instruction-based image editing models [13, 27] perform text-conditioned image-to-image editing, directly modifying a reference image according to a natural-language instruction. A representative early method, InstructPix2Pix [3], distills supervision from large text-to-image models and trains an editor conditioned on the input image and instruction to generate the edited output. Recent models extend this paradigm from single-image editing to multi-reference conditioning, where multiple reference images (e.g., identity, garment, or style) are jointly used with text for more controllable edits. For example, Qwen-Image-Edit-2511 [27] supports multi-image inputs for instruction-guided editing, and FLUX.2-based editors [13] similarly enable flexible composition of multiple visual cues within a unified generative framework. This flexibility makes universal multi-reference editors a promising alternative to specialized VTON pipelines.

2.3 Image Editing Benchmark

Benchmarks play a key role in assessing model capabilities. As image editing models advance, a growing number of editing benchmarks have been proposed [11, 17, 20, 21, 30, 36], including general-purpose suites that span tasks from simple to complex (e.g., I2EBench [17], Ice-Bench [20], Imgedit [36]) and task-specific benchmarks such as KRIS-Bench [30] for knowledge-based reasoning edits. VTON benchmarks have also recently emerged [7, 14, 32], but most focus on a single VTON setting (e.g., StreetVTON [7] for in-the-wild try-on). In contrast, we propose VTEdit-Bench to comprehensively evaluate VTON models across diverse task types, including Shop2Model, Shop2MultiModel, Shop2MultiView, Model2Model, and MultiShop2Model.

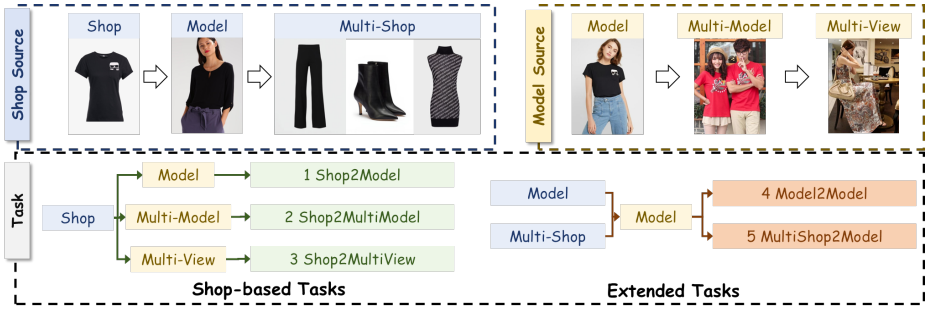

Fig. 2: Illustration of the five tasks in VTEdit-Bench, showing the diversity of shop sources and model sources in the dataset, as well as how different tasks are constructed and connected through shared input compositions.

3 VTEdit-Bench

To evaluate universal multi-reference editors for multi-scene VTON, we propose VTEdit-Bench, which evaluates multi-scene tasks with progressively increasing reference complexity. We further introduce VTEdit-QA, a reference-aware VLM-based evaluator for fine-grained assessment of multi-reference consistency and image quality.

3.1 Task Description

As shown in Fig. 2, VTEdit-Bench is organized into five representative tasks: Shop2Model, Shop2MultiModel, Shop2MultiView, Model2Model, and MultiShop-2Model. These tasks are designed to cover representative real-world VTON scenarios, including identity variation, viewpoint changes, multi-person try-on, and multi-item composition. The first three tasks are shop-based settings, where the clothing source is fixed as shop product images while the model source becomes progressively more complex, ranging from a single frontal model to multi-view and multi-person cases. The remaining two tasks extend conventional VTON beyond shop-to-person transfer by considering model-to-model garment transfer and multi-shop item composition. Overall, VTEdit-Bench comprises 13,521 human sources and 10,926 clothing sources, yielding 24,220 test pairs.

We further analyze the overall attribute distribution of VTEdit-Bench in Fig. 3. The benchmark covers diverse garment categories, including upper-body clothes, lower-body clothes, dresses, bags, and shoes. It also spans different scene types, viewpoints, and person-number settings. Indoor scenes and frontal views dominate conventional settings, while outdoor scenes, non-frontal views, and multi-person cases are introduced in more challenging tasks. These distributions show that VTEdit-Bench not only covers standard VTON scenarios but also incorporates realistic variations in clothing type, scene, viewpoint, and reference composition. Details of each task are provided below.

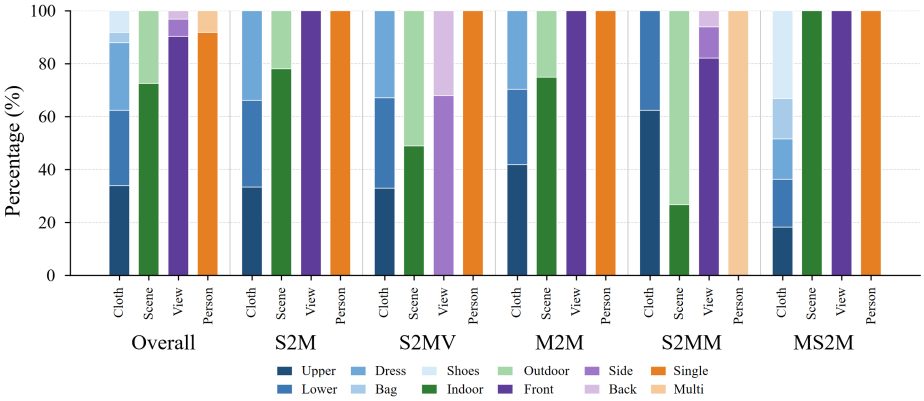

Fig. 3: Attribute distribution of VTEdit-Bench across the overall benchmark and the five tasks. Each group shows the percentage distribution of four key attributes: cloth category, scene type, view direction, and person number.

Shop2Model This task transfers common garments (tops, bottoms, dresses) from shop product images to model photos, evaluating realistic garment integration across indoor and outdoor scenes. Shop images and indoor models are taken from the DressCode and VITON-HD test sets, while outdoor models are collected from StreetVTON. In total, it contains 7,432 indoor models, and 2,089 outdoor models, yielding 9,521 test pairs.

Shop2MultiModel This task extends VTON to multi-person images, evaluating consistency and coherence across individuals. We collect 400 multi-person model images from DeepFashion [16] and an additional 1,600 images from the web, and pair them with upper-body garments from the DressCode test set, following the Shop2Model setup. In total, this yields 2,000 multi-person model images and 2,000 test pairs.

Shop2MultiView This task evaluates VTON under non-frontal viewpoints, focusing on clothing fidelity and structural coherence across perspectives. We collect 1,000 multi-view model images from DeepFashion [16] and an additional 1,000 from the web, and pair them with shop images (tops, bottoms, dresses) from the DressCode test set, following the Shop2Model setup. In total, it includes 2,000 model images and 5,400 product images, yielding 2,000 test pairs.

Model2Model This task transfers apparel from a source person to a target person, evaluating clothing extraction and faithful re-rendering under complex appearance and background. We consider four splits: indoor $\rightarrow$ indoor, indoor $\rightarrow$ outdoor, outdoor $\rightarrow$ indoor, and outdoor $\rightarrow$ outdoor. Indoor images are from VITON-HD and outdoor images are from StreetVTON. In total, it includes 2,032 clothing sources and 4,121 model pairs, yielding 8,299 test pairs.

MultiShop2Model This task composes multiple shop items (tops, bottoms, dresses, bags, shoes) onto a single person, stressing multi-item interaction, occlusion, and global coherence. Shop images are from the DressCodeMR test set. In total, it contains 8,894 product sets and 5,400 model images, yielding 2,400 test pairs.

Auxiliary conditions. To satisfy the input requirements of specialized VTON models, we provide auxiliary conditions for each model image, including OpenPose, human parsing, DensePose, and agnostic masks; for each clothing image, we additionally provide a cloth mask. Moreover, to ensure broad compatibility with these pipelines, we incorporate the extractability of such auxiliary signals as a constraint during data curation, filtering candidate samples to retain those for which the required conditions can be reliably computed.

3.2 Unified Evaluation Metrics

To enable systematic and fine-grained assessment across diverse VTON tasks, we design a unified evaluation protocol. Considering the common aspects across different tasks, we design a concise, universal metric that assesses outputs along three key dimensions: model consistency, cloth consistency, and image quality.

Model Consistency This dimension evaluates whether the synthesized VTON results faithfully preserve the identity and structural attributes of the source person. Key aspects include facial identity retention, body shape consistency, and pose alignment. These criteria collectively measure the model's ability to maintain subject-specific characteristics across diverse VTON conditions.

Cloth Consistency This dimension assesses the fidelity of garment transfer, focusing on preserving the semantic and visual attributes of the source clothing. Evaluation considers fine-grained texture details, color accuracy, and the alignment of garment boundaries with the body. Together, these factors reflect how well the model reproduces clothing characteristics across varied VTON tasks.

Image Quality This dimension evaluates the overall realism of the generated image, focusing on visual artifacts and distortions affecting the person or the clothing. Evaluation considers issues such as unnatural body proportions or deformations, distorted facial or limb details, and implausible clothing fit or fabric behavior. It also accounts for rendering artifacts (e.g., blurring or pixelation) and physical inconsistencies such as mismatched lighting or shadows.

By integrating these three dimensions, VTEdit-Bench establishes a standardized evaluation framework that jointly captures both fine-grained consistency and holistic realism. This unified design enables rigorous, reproducible, and comprehensive comparisons between specialized and universal multi-reference image editing models across a wide range of VTON tasks.

Fig. 4: The annotation pipeline for capturing human preferences.

3.3 Annotation Pipeline for Human Preferences

To capture human preferences for VTON results, we design an annotation pipeline based on pairwise comparisons. As shown in Fig. 4, given a pair of model and cloth samples, we generate multiple outputs (output 1 to 7) using different models. The outputs are then paired, and annotators are asked to select the output with the best VTON quality. If annotators find it difficult to distinguish between the outputs, we allow them to skip the annotation. This process is repeated for all image pairs. The pairwise comparison results are then aggregated using the Bradley-Terry model [2] to obtain the final ranking. Let P_{ij} represent the probability that image i is preferred over image j, and let r_i denote the latent "quality" or "ability" score of image i. The Bradley-Terry model computes P_{ij} as:

$$P_{ij} = \frac{e^{r_i}}{e^{r_i} + e^{r_j}}, \tag{1}$$

where r_i is the latent ranking score of image i, and P_{ij} is the probability that image i is preferred over image j. The pairwise comparison results are used to estimate the values of r_i for each image using Maximum Likelihood Estimation (MLE). The final rank for each image is determined by the estimated r_i values, with higher values corresponding to higher ranks. To ensure robust results, three annotators performed over 15k annotations, providing a comprehensive dataset of human preferences for further analysis.

3.4 VTEdit-QA

To enable scalable evaluation, we introduce VTEdit-QA, a reference-aware quality assessment framework leveraging the large vision-language model, GPT-4o [19], for automated scoring across three defined dimensions. Given a source model image S_m, a source cloth image S_c, and a generated output I_{out}, GPT-4o is provided with either S_m or S_c along with I_{out} to assess model consistency and cloth consistency, respectively. For image quality, GPT-4o receives only the

Fig. 5: Examples of VTEdit-QA for evaluating VTON outputs along model consistency, cloth consistency, and image quality.

generated output image I_{out}. For each dimension, we provide a specific prompt that details the evaluation criteria, which are further elaborated in the Supp. 1.2. Following the prompt, GPT-4o produces a numerical score between 0 and 5, reflecting the degree of consistency with the source images or the quality of the generated image. Additionally, GPT-4o generates detailed reasoning to justify its evaluation, enhancing transparency in the scoring process. Fig. 5 shows two examples where GPT-4o accurately detects inconsistencies in the model and cloth, demonstrating its ability to distinguish between high-quality and low-quality outputs based on the defined evaluation criteria and providing coherent justifications for its ratings. Since a successful virtual try-on result requires simultaneously preserving the model identity, accurately transferring the garment, and maintaining overall image quality, a failure in any dimension can significantly degrade the final result. Therefore, we adopt a conservative aggregation strategy and take the minimum of the three scores as the overall score.

3.5 Validating VTEdit-QA via Human Correlation

We validate VTEdit-QA by measuring how well its rankings align with human judgment. Specifically, we compare (i) the ranking consistency among human annotators and (ii) the ranking correlation between VTEdit-QA (GPT-4o) and each annotator. All correlations are computed using Spearman's Rank Correlation Coefficient (SRCC) [24], which measures the agreement between two ranked

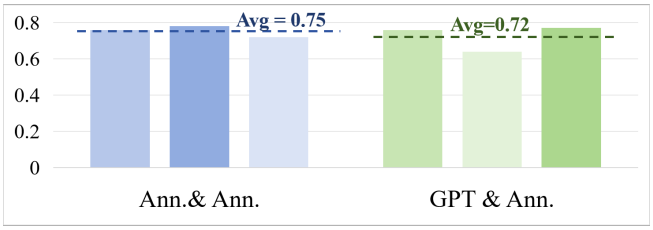

Fig. 6: SRCC between human annotators and GPT-4o under different scoring settings.

lists. Given two ranking sequences L_m and L_n, SRCC is defined as:

$$\mathrm{SRCC}(L_m, L_n) = 1 - \frac{6\sum_{j=1}^{N} d_j^2}{N(N^2-1)}, \tag{2}$$

where d_j denotes the rank difference of the j-th item and N is the total number of items. As shown in Fig. 6, human annotators exhibit strong agreement, with an average Annotator–Annotator SRCC of 0.75, indicating consistent human preference rankings. Under our proposed VTEdit-QA, GPT-4o achieves an average SRCC of 0.72 with human annotator rankings, approaching the human inter-annotator agreement level. This close gap suggests that VTEdit-QA provides rankings that are largely consistent with human judgment, making GPT-4o a reliable proxy annotator for scalable evaluation. More analysis and results on variant aggregation strategies and VLMs are in Supp. 1.3.

4 Experiment

4.1 Experimental Setup

We benchmark universal multi-reference image editing models and compare them against specialized VTON models as strong task-specific baselines on VTEdit-Bench. Specifically, we benchmark eight open-source editing models with native multi-reference conditioning, including Qwen-Image-Edit-2511 [27], Uniworld [15], Flux.2 [13], DreamOmni2 [31], OmniGen2 [28], UNO [29], DreamO [18], and Flux.2-klein [13]. For comparison, we select seven recent open-source specialized VTON methods whose required auxiliary conditions (e.g., OpenPose, human parsing, DensePose) are supported by VTEdit-Bench, including ITA-MDT [10], OOTD-Diffusion [33], CatVTON [5], Any2AnyTryon [8], FastFit [6], IDM-VTON [4], and StableVITON [12]. Universal multi-reference image editing models are evaluated zero-shot across all tasks using a single unified prompt (see Supp. 2.2), without any task-specific fine-tuning. Specialized VTON models are evaluated using their official checkpoints. For extended tasks (e.g., Shop2MultiModel and Shop2MultiView), we provide the required auxiliary conditions (e.g., OpenPose, human parsing) to enable zero-shot evaluation when applicable. However, most specialized VTON models are not compatible

Table 2: FID and KID results on VTEdit-Bench across multiple VTON tasks. The last column reports the average of task-wise rankings for each model.

Method	S2M		S2MV		S2MM		M2M		MS2M		Avg.
	FID	KID	FID	KID	FID	KID	FID	KID	FID	KID	RANK
Specialized VTON Model											
IDM-VTON	7.40	2.61	42.90	8.37	73.90	14.65	-	-	-	-	**5.4**
StableVITON	16.86	7.24	37.95	5.01	111.44	53.58	-	-	-	-	8.6
ITA-MDT	12.23	5.60	69.36	31.72	143.61	90.38	-	-	-	-	10.4
OOTD-Diffusion	7.49	2.36	67.94	22.70	90.57	26.52	-	-	-	-	7.6
CatVTON	10.82	6.95	43.28	10.54	76.30	20.03	14.98	8.79	-	-	**5.4**
Any2AnyTryon	11.37	7.19	48.99	19.88	86.12	35.68	-	-	-	-	9.2
FastFit	9.56	3.56	49.42	12.30	81.67	17.67	-	-	11.11	1.53	**4.8**
Universal Multi-reference Image Editing Model											
Qwen-Image-Edit-2511	9.53	4.17	43.48	7.88	71.46	13.14	13.36	4.43	46.81	21.60	**3.0**
Uniworld	38.73	24.63	64.38	29.54	80.83	24.91	27.68	10.08	-	-	9.2
Flux.2	10.88	6.54	36.51	7.26	65.81	13.68	14.51	7.62	15.85	5.36	**2.2**
DreamOmni2	27.41	14.36	91.79	45.58	115.67	43.89	32.36	12.84	18.77	6.10	9.8
OmniGen2	34.04	25.42	81.32	54.24	83.23	32.47	37.59	26.68	35.78	18.61	10
UNO	32.77	19.87	70.28	29.97	93.43	36.24	29.37	15.63	53.38	28.71	10.4
DreamO	14.91	7.74	60.54	22.89	91.66	27.31	24.62	11.38	20.53	8.94	**7.4**
Flux.2-klein	11.92	8.17	44.03	12.55	79.23	24.79	17.57	11.60	16.46	7.33	**5.6**

Task abbreviations: S2M (Shop2Model), S2MV (Shop2MultiView), S2MM (Shop2MultiModel), M2M (Model2Model), MS2M (MultiShop2model).

with the MultiShop2Model and Model2Model tasks due to their fixed input assumptions and task-specific requirements. All experiments are conducted on a computing cluster with eight NVIDIA A100 GPUs to ensure consistent runtime and reproducibility.

4.2 Preliminary Analysis on VTEdit-Bench

As shown in Tab. 2, we perform a preliminary analysis of universal multi-reference image editing models on VTEdit-Bench using FID and KID, where lower values indicate better realism and closer alignment to the target distribution. Here, the target distribution refers to the model images in the corresponding task-specific dataset, which serve as the reference distribution for each task.

Universal Models Achieve Competitive Performance on Shop-Based Tasks. On the Shop2Model task, specialized VTON models achieve strong performance, led by IDM-VTON (7.40/2.61) and OOTD-Diffusion (7.49/2.36). Notably, universal models such as Flux.2 and Qwen-Image-Edit-2511 remain highly competitive, indicating that universal multi-reference image editing models can already match strong specialized baselines in the standard VTON setting. As task difficulty increases on Shop2MultiView and Shop2MultiModel, most methods degrade, while Flux.2 remains the best on Shop2MultiView (36.51/7.26) and strong on Shop2MultiModel (65.81/13.68), with Qwen-Image-Edit-2511 also performing competitively (43.48/7.88 and 71.46/13.14). In contrast, some specialized baselines degrade sharply, e.g., ITA-MDT rises from 12.23/5.60 to 143.61/

90.38. These observations suggest that universal multi-reference editing models maintain more stable performance as scenario complexity increases.

Universal Models Demonstrate Stronger Cross-Task Generalization. The extended tasks, Model2Model and MultiShop2Model, are supported by only a subset of methods, further exposing differences in model flexibility. On Model2Model, universal models such as Qwen-Image-Edit-2511 and Flux.2 achieve competitive results comparable to CatVTON. On MultiShop2Model, FastFit attains the best scores (11.11/1.53), whereas universal models show larger degradation, suggesting that fine-grained multi-item and multi-source control remains challenging for current universal editing frameworks.

In addition, some methods fail on specific tasks and produce outputs that deviate significantly from the target distribution, resulting in extremely high FID/KID values. Representative failure cases are provided in the Supp. 2.3.

4.3 Fine-grained Analysis with VTEdit-QA

While FID/KID measure distributional similarity, they can be misleading for VTON when clothing transfer fails (e.g., copying the input). We therefore conduct fine-grained evaluation with VTEdit-QA, which evaluates model consistency, cloth consistency, and image quality, and aggregates them into an overall score via the minimum rule. Since some methods do not support certain tasks, we rank the methods by their FID/KID scores on supported tasks, as shown in Tab. 2, and analyze representative strong models here, including IDM, CatVTON, FastFit, DreamO, Qwen-Image-Edit-2511, Flux.2, and Flux.2-klein.

Universal Models Show More Stable Performance Across Shop-based Tasks. As shown in Tab. 3, we report VTEdit-QA results on the shop-based tasks supported by all evaluated methods. Across these tasks, specialized VTON models generally achieve high model and cloth consistency, but their overall scores are often limited by image quality as the settings become more challenging. In contrast, universal multi-reference editors maintain more stable overall performance across the three tasks, indicating stronger robustness to view changes and increased identity diversity. In particular, Flux.2-klein attains the highest overall scores across all three tasks (3.96/3.99/3.56), outperforming the best specialized baseline IDM (3.60/2.73/1.65).

Extended Tasks Expose Performance Gaps Among Universal Models. As shown in Tab. 4, we analyze two extended tasks, Model2Model and MultiShop2Model, which are only supported by a subset of methods. Across these tasks, universal editors exhibit clear performance divergence, with some models approaching specialized baselines while others fail under the same setting. On Model2Model, CatVTON achieves the best overall score (2.25). Among universal models, Flux.2 attains a comparable overall score (2.06), whereas Qwen-Image-Edit-2511 performs notably worse (1.17) due to low model-consistency (Mod.

Table 3: VTEdit-QA Results on shop-based tasks, including Shop2Model, Shop2MultiView, and Shop2MultiModel.

Method	Shop2Model				Shop2MultiView				Shop2MultiModel			
	Mod.	Clo.	Qual.	Overall	Mod.	Clo.	Qual.	Overall	Mod.	Clo.	Qual.	Overall
Specialized VTON Models												
IDM	4.68	4.55	4.02	3.60	4.66	4.54	3.03	2.73	4.51	4.73	1.72	1.65
CatVTON	4.67	4.49	3.79	3.39	4.75	4.12	2.49	2.14	4.45	4.18	1.39	1.26
FastFit	4.64	4.47	3.42	3.08	4.70	4.31	1.94	1.74	4.45	4.63	1.22	1.18
Univeral Multi-reference Image Editing Model												
DreamO	2.78	3.22	4.19	1.62	1.82	3.18	3.42	0.96	1.62	2.75	3.07	0.74
Qwen-Image-Edit-2511	4.23	4.77	4.34	3.64	3.94	4.20	3.30	2.46	4.37	3.74	2.90	2.14
Flux.2	4.62	4.09	4.30	3.36	4.79	4.25	3.95	3.38	4.82	4.36	3.40	3.02
Flux.2-klein	4.61	4.68	4.48	3.96	4.61	4.80	4.40	3.99	4.81	4.73	3.79	3.56

Metric abbreviations: Mod. (model consistency), Clo. (cloth consistency), Qual. (image quality)

Table 4: VTEdit-QA Results on Model2Model and MultiShop2Model Tasks.

Method	Model2Model				MultiShop2Model							
	Mod.	Clo.	Qual.	Overall	Mod.	Dresses	Upper	Lower	Shoe	Bag	Qual.	Overall
Specialized VTON Models												
CatVTON	4.44	3.69	2.96	2.25	-	-	-	-	-	-	-	-
FastFit	-	-	-	-	3.57	4.96	4.99	4.95	4.82	4.23	4.14	2.90
Universal Multi-reference Image Editing Model												
DreamO	1.53	3.68	4.15	0.90	2.77	2.57	3.38	2.17	0.92	1.00	4.32	0.01
Qwen-Image-Edit-2511	1.60	4.32	4.00	1.17	2.45	2.29	1.98	1.58	2.17	2.30	1.70	0.34
Flux.2	2.77	3.80	4.36	2.06	3.18	4.21	4.82	4.54	4.20	4.56	4.20	2.25
Flux.2-klein	1.55	2.17	4.63	1.03	3.52	4.97	4.65	4.05	4.56	4.49	4.45	2.68

Metric abbreviations: Mod. (model consistency), Clo. (cloth consistency), Qual. (image quality)

1.60), suggesting frequent identity drift despite visually plausible outputs. On MultiShop2Model, FastFit performs best (2.90 overall). Among universal editors, Flux.2-klein remains close (2.68 overall), while DreamO collapses to near-zero overall score (0.01), indicating failure to handle multi-item conditioning and composition.

4.4 Qualitative Analysis of Universal Editing Models.

To further understand universal multi-reference editing models, we present representative qualitative cases of Flux.2 and Flux.2-klein in Fig. 7. Fig. 7 (a,b) provide examples of successful outputs on Shop2MultiView and Shop2MultiModel, illustrating realistic synthesis and accurate garment transfer in some instances. Fig. 7 (c,d) show challenging cases where the models may confuse the person and clothing references in Model2Model, or exhibit inconsistent multi-subject control in multi-human settings, e.g., dressing only one person correctly while the other deviates from the cloth reference. Fig. 7 (e,f) further shows typical failure patterns, including unsuccessful garment replacement and incorrect garment understanding, often influenced by the source outfit. These examples qualitatively illustrate both the promise and remaining limitations of universal edi-

Fig. 7: Qualitative examples of universal editing models (Flux.2 and Flux.2-klein) across multiple VTON tasks. Scores are reported as "model consistency, cloth consistency, image quality | overall".

tors in reference understanding, identity-conditioned transfer, and robust multi-person/multi-item control.

5 Insights and Discussions

5.1 Potential and Limitations of Universal Models

Based on VTEdit-Bench, universal multi-reference image editing models show clear potential for VTON. A key reason is their simpler, more general editing paradigm: instead of relying on task-specific modules, they directly perform instruction-guided editing with flexible conditioning. Together with large-scale and diverse training data, this design yields stronger generalization and more stable performance across harder variants such as Shop2MultiView and Shop2MultiModel settings, suggesting a promising path toward a unified VTON solution. However, two bottlenecks persist: (i) model and cloth consistency degrades as reference complexity increases, and (ii) limited VTON-specific visual understanding and instruction following weaken fine-grained compositional control (e.g., attribute binding, localization, and occlusion reasoning). These findings suggest promising directions: adapting universal editors with high-quality VTON data to strengthen identity and garment constraints; improving multi-reference editing via spatially grounded reference binding and joint image–text understanding; and enabling reasoning-driven editing that iteratively analyzes failures and refines outputs toward constraint-satisfying results.

5.2 Benchmark Scope and Future Extensions

The current design of VTEdit-Bench uses specialized VTON models as baselines and thus primarily constructs tasks via different input image configura-

tions to ensure fair and consistent comparison. This enables a systematic evaluation of the differences and boundaries between universal and specialized models under realistic settings. Meanwhile, universal editing models support flexible instruction-following and prompt-driven control, opening a broader design space for VTON. Future extensions could incorporate prompt-guided settings (e.g., fine-grained constraints, multi-step instructions, localized garment replacement, and style control) to more comprehensively assess controllability and interaction capabilities.

5.3 Evaluation Insights with VTEdit-QA

VTEdit-QA suggests that VTON evaluation is better framed as constraint satisfaction rather than pure distribution matching. It produces rankings that are consistent with human preference and approach the level of inter-annotator agreement, indicating that the reference-aware VLM in VTEdit-QA can serve as a practical proxy for scalable assessment. More importantly, VTEdit-QA exposes semantic failure modes that are often invisible to distributional metrics such as FID/KID. For example, IDM achieves strong FID/KID scores on VTEdit-Bench, yet receives a lower overall VTEdit-QA score than Flux.2-klein, implying that distributional similarity can remain high even when identity or garment constraints are violated. This highlights the need for VTON benchmarks to report not only realism, but also explicit reference-consistency signals (e.g., identity preservation and garment adherence) to reflect real try-on reliability.

6 Conclusion

We introduced VTEdit-Bench, the first benchmark for evaluating universal multi-reference image editing models under VTON scenarios with progressively increasing complexity, and benchmarked them against strong specialized baselines. Results showed that top universal editors were competitive in canonical try-on and often generalized more stably to harder multi-view and multi-model settings, while still facing bottlenecks in model/cloth consistency and compositional control. We also introduced VTEdit-QA, a reference-aware VLM-based evaluator measuring model consistency, cloth consistency, and image quality. VTEdit-QA aligned well with human preferences and revealed semantic failures overlooked by coarse distributional metrics. Together, VTEdit-Bench and VTEdit-QA provide a unified foundation for diagnosing, comparing, and advancing VTON systems.

Acknowledgements

This work was supported by the Zhongguancun Academy under Grants C20250402 and XTS0051, and by the National Natural Science Foundation of China under Grants 62401027 and 62231002.

References

1. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
2. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
4. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for authentic virtual try-on in the wild. In: European Conference on Computer Vision. pp. 206–235. Springer (2024)
5. Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., Jiang, D., Liang, X.: Catvton: Concatenation is all you need for virtual try-on with diffusion models. arXiv preprint arXiv:2407.15886 (2024)
6. Chong, Z., Lei, Y., Zhang, S., He, Z., Wang, Z., Zhang, X., Dong, X., Wu, Y., Jiang, D., Liang, X.: Fastfit: Accelerating multi-reference virtual try-on via cacheable diffusion models. arXiv preprint arXiv:2508.20586 (2025)
7. Cui, A., Mahajan, J., Shah, V., Gomathinayagam, P., Liu, C., Lazebnik, S.: Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 1414–1423 (2025)
8. Guo, H., Zeng, B., Song, Y., Zhang, W., Liu, J., Zhang, C.: Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19085–19096 (2025)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
10. Hong, J.W., Ton, T., Pham, T.X., Koo, G., Yoon, S., Yoo, C.D.: Ita-mdt: Image-timestep-adaptive masked diffusion transformer framework for image-based virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28284–28294 (2025)
11. Jia, B., Huang, W., Tang, Y., Qiao, J., Liao, J., Cao, S., Zhao, F., Feng, Z., Gu, Z., Yin, Z., et al.: Compbench: Benchmarking complex instruction-guided image editing. arXiv preprint arXiv:2505.12200 (2025)
12. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8176–8185 (2024)
13. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025)
14. Li, J., Chen, T., Jiang, S., Wang, W., Luo, J., Wu, C.: Openvton-bench: A large-scale high-resolution benchmark for controllable virtual try-on evaluation. arXiv preprint arXiv:2601.22725 (2026)
15. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025)
16. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1096–1104 (2016)

17. Ma, Y., Ji, J., Ye, K., Lin, W., Wang, Z., Zheng, Y., Zhou, Q., Sun, X., Ji, R.: I2ebench: A comprehensive benchmark for instruction-based image editing. Advances in Neural Information Processing Systems **37**, 41494–41516 (2024)
18. Mou, C., Wu, Y., Wu, W., Guo, Z., Zhang, P., Cheng, Y., Luo, Y., Ding, F., Zhang, S., Li, X., et al.: Dreamo: A unified framework for image customization. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)
19. OpenAI: Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024
20. Pan, Y., He, X., Mao, C., Han, Z., Jiang, Z., Zhang, J., Liu, Y.: Ice-bench: A unified and comprehensive benchmark for image creating and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16586–16596 (2025)
21. Pathiraja, B., Patel, M., Singh, S., Yang, Y., Baral, C.: Refedit: A benchmark and method for improving instruction-based image editing model on referring expressions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15646–15656 (2025)
22. Shen, F., Jiang, X., He, X., Ye, H., Wang, C., Du, X., Li, Z., Tang, J.: Imagdressing-v1: Customizable virtual dressing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6795–6804 (2025)
23. Shim, S.H., Chung, J., Heo, J.P.: Towards squeezing-averse virtual try-on via sequential deformation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4856–4863 (2024)
24. Spearman, C.: The proof and measurement of association between two things. (1961)
25. Wan, Z., Xu, Y., Hu, D., Cheng, W., Chen, T., Wang, Z., Liu, F., Liu, T., Gong, M.: Mft-viton: High-fidelity virtual try-on with minimal input via a mask-free transformer-diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1985–1994 (2025)
26. Wang, H., Zhang, Z., Di, D., Zhang, S., Zuo, W.: Mv-vton: Multi-view virtual try-on with diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7682–7690 (2025)
27. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
28. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
29. Wu, S., Huang, M., Wu, W., Cheng, Y., Ding, F., He, Q.: Less-to-more generalization: Unlocking more controllability by in-context generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18682–18692 (2025)
30. Wu, Y., Li, Z., Hu, X., Ye, X., Zeng, X., Yu, G., Zhu, W., Schiele, B., Yang, M.H., Yang, X.: Kris-bench: Benchmarking next-level intelligent image editing models. arXiv preprint arXiv:2505.16707 (2025)
31. Xia, B., Peng, B., Zhang, Y., Huang, J., Liu, J., Li, J., Tan, H., Wu, S., Wang, C., Wang, Y., et al.: Dreamomni2: Multimodal instruction-based editing and generation. arXiv preprint arXiv:2510.06679 (2025)
32. Xiaobin, H., Yujie, L., Donghao, L., Xu, P., Jiangning, Z., Junwei, Z., Chengjie, W., Yanwei, F.: Vtbench: Comprehensive benchmark suite towards real-world virtual try-on models. arXiv preprint arXiv:2505.19571 (2025)
33. Xu, Y., Gu, T., Chen, W., Chen, A.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8996–9004 (2025)

34. Yang, X., Ding, C., Hong, Z., Huang, J., Tao, J., Xu, X.: Texture-preserving diffusion models for high-fidelity virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7017–7026 (June 2024)
35. Yang, Z., Li, Y., He, S., Li, X., Xu, Y., Dong, J., Du, Y.: Omnivton: Training-free universal virtual try-on. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16702–16711 (2025)
36. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)
37. Zhang, X., Song, D., Zhan, P., Chang, T., Zeng, J., Chen, Q., Luo, W., Liu, A.A.: Boow-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26399–26408 (2025)