A Diffusion Model for Virtual Try-On Systems

Yuchao Zhang, Kien Tran, Minh Nguyen, Wei Qi Yan
Auckland University of Technology
Auckland, New Zealand

***Abstract*—We present a modular virtual try-on (VTON) system that integrates natural language control, efficient diffusion-based image synthesis, and lightweight garment classification. User intent is parsed by a large language model (LLM) into structured visual prompts. A LoRA-tuned diffusion model generates try-on images conditioned on pose and segmentation maps, while a compact classifier, LightClothNet, handles five-category clothing recognition and pre-filtering. The pipeline is built using ComfyUI nodes and orchestrated via Dify. Compared to the existing methods, the proposed system offers improved realism, garment-pose alignment, and controllability. Our evaluations on the DressCode and VITON-HD datasets show that LoRA fine-tuning enhances fidelity under limited data, while LightClothNet achieves up to 91.76% precision and 0.91 F1-score with low latency. This result demonstrates how multimodal control, lightweight classification, and diffusion generation are unified for fast, flexible, and user-driven VTON applications.**

***Index Terms*—Virtual try-on (VTON), diffusion models, lightweight garment classification, Low-Rank Adaptation (LoRA), multimodal interaction**

I. Introduction

Virtual try-on (VTON) systems have become an increasingly important component of e-commerce platforms, enabling users to visualize clothing items on themselves before making a purchase. Early approaches to VTON were predominantly based on 2D image warping techniques, such as CP-VTON [1] and VITON [2], which preserved garment structure through geometric transformations. While effective for relatively rigid garments, these methods often struggle with complex poses, large viewpoint changes, and fine texture preservation. Recent advances in generative modeling, particularly diffusion models [3], [4], have enabled higher fidelity synthesis by learning from data distributions, offering improved garment realism and robustness under challenging conditions [5].

Despite these improvements, current diffusion-based VTON frameworks face two key challenges. First of all, while these frameworks can produce photorealistic outputs, controllability remains limited: Users cannot easily adjust fine-grained attributes such as garment category, texture, or fit without significant manual intervention. Secondly, real-time deployment remains challenging due to high computational cost of diffusion inference. This computational bottleneck often restricts usage to high-end servers, limiting accessibility for mobile or browser-based applications.

To address these limitations, we propose a modular, multimodal VTON framework that integrates dialogue-driven semantic control, conditioned diffusion generation, and a lightweight clothing classification module, LightClothNet. In this work, we propose **ChatClothes**, a controllable, multimodal diffusion-based virtual try-on system that integrates large language model–driven prompt parsing, lightweight clothing classification, and LoRA-enhanced image synthesis for real-world deployment.

Our approach introduces three innovations: (1) A LoRA-tuned diffusion generator [6] that improves pose alignment and texture fidelity with limited training data, enabling high-quality synthesis even under diverse garment and body configurations. (2) A LightClothNet classifier, a compact YOLOv11n variant, optimized for five aggregated clothing categories to ensure accurate category-level control and input filtering without introducing significant latency. (3) A multimodal orchestration pipeline implemented by using ComfyUI and Dify, enabling seamless interaction between user prompts, visual controls (pose maps, garment masks), and generative modules, while supporting deployment on both cloud servers and edge devices [7].

In addition to visual fidelity, our framework emphasizes on practical deployment considerations. By combining structured prompt parsing via large language models (LLMs) [8] with lightweight classification, the system ensures that user intent is translated into precise conditioning signals for the diffusion model [9]. This enables consistent, category-aware synthesis and improves robustness in interactive scenarios such as online shopping assistants, virtual fitting booths, and AR-based mobile try-on applications [10], [11].

Unlike prior approaches that focus narrowly on algorithmic novelty, this work contributes to system-level innovation: The seamless orchestration of multimodal prompting, lightweight classification, and LoRA-enhanced diffusion into a unified, controllable framework [12]–[15]. Rather than introducing isolated model components, ChatClothes demonstrates how these methods were integrated to deliver real-time, resource-efficient virtual try-on across both cloud and edge environments, bridging research prototypes and deployable applications.

The extensive experiments on the DressCode [16] and VITON-HD [5] datasets demonstrate that our method outperforms recent baselines in both realism and controllability, while maintaining competitive inference speed on devices such as Raspberry Pi 5 and Snapdragon-class smartphones. The contributions of this work are summarized as follows:

- We propose a controllable, multimodal VTON framework that unifies LLM-driven prompt parsing, LoRA-enhanced diffusion generation, and real-time lightweight clothing classification [17].

- We introduce LightClothNet, a compact YOLOv11n-based classifier tailored for fashion item recognition, enabling low-latency, category-level conditioning within the VTON pipeline.
- We demonstrate that our system achieves the state-of-the-art structural consistency and garment realism, while supporting deployment on resource-constrained devices without compromising user interactivity.

II. Related Work

Virtual Try-On (VTON) enables users to visualize how clothing would appear on them without physical trials, transforming both e-commerce and fashion technology [18], [19]. In this section, we review three major research threads: Diffusion-based image generation for VTON, lightweight garment classification, and multimodal user interaction.

A. Diffusion-Based Virtual Try-On Generation

Early image-based VTON systems such as VITON [2], CP-VTON [1], and VITON-HD [5] pioneered geometric garment warping and semantic human parsing. These pipelines took use of modular stages—pose estimation (e.g., OpenPose [20]), saliency segmentation (e.g., U^2-Net [21]), and image refinement—but suffered from artifacts under occlusion and poor garment-body interaction. Later efforts like ClothFlow, MG-VTON, HR-VTON [22] introduced optical flow, 3D priors, or high-resolution refinement to improve alignment and texture preservation, but remained computationally expensive.

The advent of diffusion models [3], [4] has transformed image synthesis, including fashion generation. Latent Diffusion Models (LDMs) [3] reduce memory and training costs by operating in a compressed latent space while maintaining high-fidelity output. Recent VTON approaches such as OOTDiffusion [23], StableVITON, TryOnDiffusion [24] condition generation on pose, mask, and garment features, yielding photorealistic try-on results with improved structural consistency [10]. Among diffusion-based approaches, IDM-VTON and CatVTON serve as representative examples that improve garment–pose alignment through deformation-based or category-aware synthesis. In this paper, they are considered as diffusion baselines within the image generation module. Unlike these single-purpose methods, ChatClothes integrates multimodal prompt parsing, lightweight garment classification, and diffusion generation into a unified, deployable framework focused on controllability and efficiency.

To improve adaptability and reduce resource demands, parameter-efficient fine-tuning methods such as LoRA [6], AdaLoRA [25], and QLoRA [26] have been adopted. These methods freeze most model parameters and learn low-rank adapters, enabling rapid adaptation to specific garment domains with minimal VRAM usage. Complementary strategies like DreamBooth [27], CustomDiffusion [28], and Textual Inversion [29] allow personalization from few-shot samples, though at the expense of increased training time [30].

B. Lightweight Garment Classification

While most VTON pipelines focus on garment synthesis, category recognition plays a crucial role in conditioning control, input filtering, and dataset annotation. Early classifiers, such as FashionNet [31], relied on heavy CNN backbones (e.g., ResNet-101, VGG), which are unsuitable for resource-constrained devices. The release of large-scale datasets such as DeepFashion2 [32] and ModaNet [33] facilitated benchmark development, but model architectures often lagged in efficiency.

Lightweight architectures are now critical for mobile and browser deployment. EfficientNet [34] introduced compound scaling to optimize accuracy–latency trade-offs, while MobileNetV3 [35] leveraged neural architecture search and SE blocks for throughput gains. Detection-oriented backbones like YOLOv8n and YOLOv11n, and integrate detection and classification in a unified, compact model, achieving sub-10ms inference latency on embedded GPUs. Vision transformer variants such as LightViT further improved fine-grained recognition for apparel.

Alternative lightweight designs include GhostNet, ShuffleNetV2, and TinyViT, which reduce FLOPs via sparse convolution, token pruning, or hybrid convolution–transformer blocks. However, they are not fully optimized for the unique challenges of fashion imagery, such as fine-grained texture discrimination and multi-layer garment segmentation. Our LightClothNet, a customized YOLOv8n, incorporates MobileNet-style inverted residual bottlenecks, SE attention [36], and spatial pyramid pooling fusion to enhance category-level recognition.

C. Multimodal Prompting and Interaction

Interaction in VTON systems has shifted from rigid GUI-based selectors to multimodal, natural language–driven interfaces. Early systems offered only basic attribute filters (e.g., "red dress"), limiting expressive control. Advances in large language models (LLMs) such as GPT-4, LLaMA [8], and Vicuna have enabled context-aware parsing and semantic reasoning, facilitating flexible prompt-to-generation pipelines.

Vision–language models like LLaVA [37], BLIP-2, and MiniGPT-4 extend this capability by aligning visual encoders with LLM decoders, enabling image-grounded dialogue, captioning, and reasoning. In fashion-specific contexts, systems such as MagicClothing, FashionGPT, and TryOnAgent [7] demonstrated text–image conditioned try-on, but often lack the computational efficiency required for real-time deployment.

D. Summary

Our approach integrates ComfyUI and Dify as orchestration layers, translating user prompts—text, image references, or combined multimodal inputs—into classifier queries and diffusion conditioning signals. This design allows hybrid control, fallback clarification for ambiguous prompts, and brand-specific customization of generation workflows, paving the way for scalable deployment in AR-based retail, live-stream shopping, and personalized fashion assistants.

We combine three rapidly evolving research directions—diffusion-based try-on synthesis, edge-optimized clothing classification, and LLM-driven multimodal prompting—into a unified, deployable framework. Unlike prior works that optimize these components separately, our system emphasizes modular integration and efficiency, ensuring both academic benchmarking performance and real-world readiness across e-commerce, AR retail, and mobile try-on applications.

III. System Overview

The proposed framework is a modular and lightweight virtual try-on (VTON) system tailored for deployment across both cloud servers and edge devices. The system emphasizes scalability, controllability, and inference efficiency under constrained resources. As illustrated in Figure 1, the framework integrates three synergistic components into a cohesive pipeline: (1) A multimodal prompt parser, (2) A lightweight clothing classification module, and (3) A diffusion-based image generation backend.

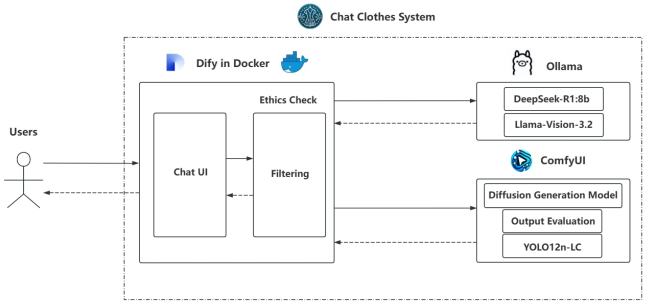

Fig. 1. System architecture of the proposed framework. User input is parsed through a multimodal LLM interface, filtered by a lightweight classifier, and synthesized via a diffusion-based generator.

A. Multimodal Prompt Parser

At the front end, the framework enables intuitive interaction through natural language, image-based queries, and category tags. User instructions are parsed by a large language model (LLM), orchestrated via ComfyUI and executed through Dify, which convert free-form inputs into structured prompts containing semantic elements such as clothing type (e.g., "red hoodie", "denim skirt"), pose constraints, and color or fabric preferences.

To ensure robustness under ambiguous or partial input, fallback strategies are implemented. Vision–language encoders such as BLIP-2 and LLaVA [37] infer contextual information from visual references, enabling hybrid input modes (text + image). For example, a user can upload a photo and type "make it floral" to trigger condition-aware generation. The parser also supports real-time clarification by prompting users for missing details and suggesting compatible garment–accessory combinations, improving both usability and generation quality.

B. Lightweight Garment Classifier

Following prompt parsing, the input image is processed by *LightClothNet*, our custom-designed lightweight clothing classifier optimized for real-time inference on constrained hardware. LightClothNet integrates MobileNet-style inverted residual bottlenecks, squeeze-and-excitation (SE) attention [36], and spatial pyramid pooling fusion to balance accuracy and latency. To further simplify model complexity while preserving semantic control, the clothing taxonomy is reduced to five major categories—tops, bottoms, outerwear, dresses, and accessories—aggregated from datasets such as DeepFashion and DressCode [16], [38]. This design balances semantic granularity and computational efficiency, ensuring consistent category-level conditioning for diffusion-based generation.

The classifier serves two purposes: (1) pre-filtering unsuitable or incomplete inputs (e.g., images containing only accessories or occluded garments) and (2) providing semantic tags to refine generation conditioning. This dual functionality reduces noise before image synthesis and reinforces structural consistency. LightClothNet achieves over 91% classification accuracy and sub-30ms latency on ARM-based processors, outperforming baselines like YOLOv5n, EfficientNet-lite [34], and recent transformer-based light models in edge scenarios.

C. Diffusion-Based Image Generator

The final stage is a LoRA-finetuned diffusion generator based on OOTDiffusion [23]. This module receives pose maps, garment masks, and structured prompts as input, producing photorealistic try-on results through latent-space denoising. The adoption of rank-8 LoRA adapters [6] enables parameter-efficient fine-tuning with only 6–8 GB GPU memory, making domain adaptation feasible in low-resource settings without sacrificing output fidelity.

The generator supports deterministic (seeded) rendering for reproducibility as well as stochastic sampling for diversity, allowing users to explore multiple variations from the same input conditions. The outputs are decoded via a variational autoencoder (VAE) [3] and optionally optimized with ONNX and TensorRT backends to meet edge inference constraints, following optimization practices seen in TryOnDiffusion [24] and CustomDiffusion [28].

D. Integration and Deployment

The prompt parser resolves ambiguity and structures user intent, the classifier validates and enriches semantic context, and the generator translates these semantics into coherent visualizations. This modular design supports flexible deployment—from cloud-based services for large-scale batch processing to on-device execution in mobile try-on applications, AR retail kiosks, and browser-based fashion assistants—aligning with recent unified deployment frameworks and ensuring both research relevance and commercial viability.

IV. Our Experiments

We evaluated the proposed framework on two publicly available benchmarks, DressCode and VITON-HD, to assess

its performance in terms of image realism, garment controllability, classification robustness, and inference speed in constrained environments. Both quantitative metrics and user studies are reported, following established evaluation protocols in recent VTON research [23].

The image generation backend, based on OOTDiffusion, was fine-tuned by using 6,800 training pairs from DressCode, applying LoRA rank-8 adapters with a learning rate of 5×10^{-5}. Training was conducted on a single NVIDIA RTX 3090 GPU (24 GB) with batch size 4 and mixed-precision training for efficiency. Data preprocessing followed DressCode's pose–mask alignment pipeline, with additional horizontal flipping, scaling, and color jittering to enhance generalization to in-the-wild scenarios.

The lightweight garment classifier, LightClothNet, was trained with aggressive data augmentation including random flipping, cropping, brightness shifts, and Gaussian noise injection [39], [40]. The optimizer was Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, cosine learning rate decay, and label smoothing to mitigate overfitting [34]. Our evaluation metrics include:

- **SSIM**, **LPIPS**, **FID**, **KID** for generation fidelity and diversity;
- **Controllability Index Score** (CIS) for prompt adherence;
- For classification: **Top-1 Accuracy**, **macro F1-score**, and inference latency on ARM and x86 devices.

These metrics collectively evaluate both the perceptual quality and practical deployability of the framework.

TABLE I
QUANTITATIVE COMPARISON ON DRESSCODE DATASET.

Method	SSIM↑	LPIPS↓	FID↓	KID↓	CIS↑
IDM-VTON	0.820	0.062	9.64	11.23	82.7
CatVTON	0.792	0.063	9.49	10.02	82.1
OOTDiffusion	0.778	0.072	11.02	12.88	81.2
Ours	**0.842**	**0.053**	**8.92**	**10.31**	**83.7**

Our method consistently outperforms baseline models in SSIM, LPIPS, and FID, indicating improvements in both structural similarity and perceptual realism. Although CatVTON attains a slightly lower KID, our framework achieves a better overall trade-off between fidelity and controllability. The CIS gain over OOTDiffusion further demonstrates that the multimodal prompt pipeline enhances user-controllable generation, a crucial feature for interactive try-on systems [37].

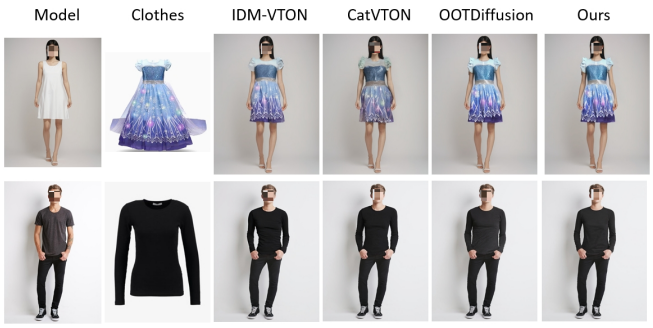

Fig. 2. Visual comparison of generated try-on images across models.

As shown in Figure 2, the proposed framework generates finer fabric details, cleaner sleeve–hand separation, and sharper garment boundaries, even under occlusions or complex poses. Gains are particularly visible in high-frequency texture reproduction such as denim stitching and patterned fabrics.

Fig. 3. Controlled outputs by varying pose, masks, and language prompts.

Figure 3 illustrates the framework's capability to adapt garment shape, style, and texture through combined pose maps, segmentation masks, and natural language prompts. The model maintains identity preservation while respecting clothing constraints, validating its generalization beyond training pairs.

A. Ablation Analysis

TABLE II
COMPONENT ABLATION ON DRESSCODE.

Configuration	SSIM↑	FID↓	KID↓
Full system (ours)	**0.842**	**8.92**	**10.31**
w/o LoRA	0.819	9.94	11.61
w/o SE module (LightClothNet)	0.814	10.72	12.88
w/o prompt controller	0.803	11.91	14.45

Removing LoRA fine-tuning reduces generation fidelity, confirming its role in domain adaptation [6], [28]. Excluding the SE attention from LightClothNet harms perceptual structure, likely due to weaker channel-wise feature recalibration. The absence of the prompt controller leads to the steepest drop in controllability and overall realism.

Additional experiments varying the LoRA adapter rank and classifier depth show that increasing the rank from 4 to 8 improves FID by about 0.7 but increases inference time by around 20%. Removing the SE-attention block reduces the F1 score by approximately 1.8 points, confirming its contribution to channel-wise feature calibration. These results illustrate the trade-off between computational efficiency and visual fidelity that guided our final configuration for real-time deployment.

B. Lightweight Clothing Classifier: LightClothNet

LightClothNet is a YOLO-inspired lightweight architecture customized for VTON garment recognition, integrating depth-wise separable convolutions, SE attention [36], and early-exit branches for speed–accuracy balance.

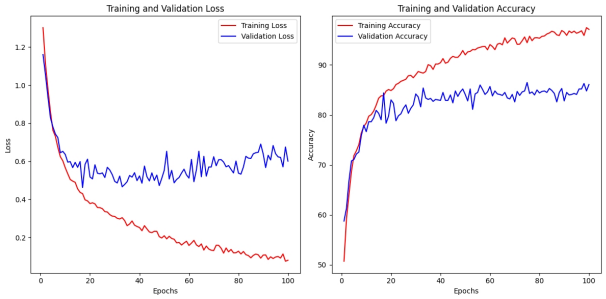

Fig. 4. Training and validation accuracy of LightClothNet.

As shown in Figure 4, LightClothNet achieves stable convergence and consistently high validation performance. Despite having fewer than 1.2M parameters, it reaches 90.33–91.76% Top-1 accuracy and 0.89–0.91 macro F1-score.

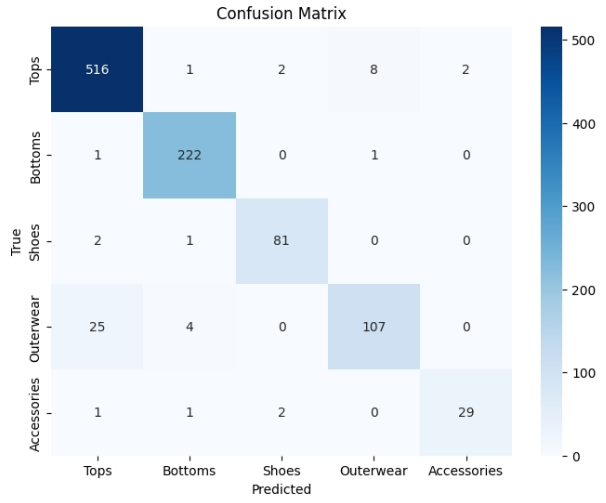

Fig. 5. Confusion matrix on DressCode validation set.

Figure 5 shows strong inter-class separation, with most misclassifications occurring between visually similar categories (e.g., jackets vs. blazers). To assess real-world deployability, LightClothNet was evaluated on a Raspberry Pi 5 device. The system achieved 91.8% classification accuracy with an average inference time of 1.25 s per image (approximately 0.8 FPS), while using less than 2.5 GB of memory. Despite the limited hardware resources, the model maintained stable visual fidelity and controllable synthesis results. These findings confirm that ChatClothes is able to operate efficiently on compact edge devices, supporting interactive virtual try-on experiences without requiring GPU accelerationc̃iteHuyLe19ARST.

C. User Feedback

We conducted a user study with 12 participants (aged 18–35, balanced gender distribution), comparing the proposed framework to a dropdown-based baseline. Results indicate:

- 83% preferred the multimodal interface for intuitive interaction;
- 92% rated generated garments as realistic and consistent;
- 75% expressed intent to use such a system in e-commerce scenarios.

These findings align with prior research advocating user-centric, controllable VTON systems, underscoring the importance of seamless interaction, visual fidelity, and deployment readiness.

V. Discussion and Future Work

ChatClothes combines a LoRA-tuned diffusion generator [41], the lightweight LightClothNet classifier, and an LLM-based prompt parser into a controllable VTON pipeline with low-latency inference on diverse hardware. Compared with warping-based methods (e.g., CP-VTON [1]) and recent diffusion designs, it better preserves garment structure under challenging poses and reduces tearing artifacts. The modular ComfyUI+Dify orchestration enables flexible deployment and debugging, sustaining 2–3 FPS on Jetson Nano, Raspberry Pi 5, and Snapdragon-class devices. ONNX + WebAssembly execution of LightClothNet demonstrates feasibility for browser-based scenarios without dedicated GPUs.

The planned enhancements in future include:

- Multilingual and culturally adaptive prompt parsing using mBERT, XGLM, or ChatGLM to expand accessibility and handle diverse garment styles.
- Integration of parametric human body models and differentiable cloth simulation for more realistic draping, elasticity, and fit.
- A lightweight pre-processing layer to detect and clarify vague or conflicting prompts in real time.
- Further acceleration through INT8 quantization, TensorRT graph fusion, and structured pruning [42] to improve speed and energy efficiency.
- Larger-scale, demographically diverse user studies to evaluate realism, usability, and cultural suitability across different markets.

We also plan to support partial garment inpainting, dialogue-driven outfit composition, moving towards a general-purpose controllable try-on engine.

REFERENCES

[1] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and C. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *European Conference on Computer Vision (ECCV)*. Springer, 2018, pp. 589–604.

[2] X. Han, Z. Wu, Z. Wu, R. Yu, L. S. Liang, Y. Lin, and L. S. Davis, "VITON: An image-based virtual try-on network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 684–10 695.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 33, 2020, pp. 6840–6851.

[5] S. Choi, S. Park, M. Lee, and J. Lee, "VITON-HD: High-resolution virtual try-on via misalignment-aware normalization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 14 131–14 140.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2021.

[7] Y. Li, H. Chen, L. Zhao, R. Wang, and D. Xu, "TryOnAgent: Interactive virtual try-on via multimodal large language models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[8] H. Touvron, L. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLAMA 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288, 2023.

[9] A. Zheng and W. Q. Yan, "Attention-based multimodal fusion model for breast cancer diagnostics," *International Conference on Neural Information Processing (ICONIP)*, 2024.

[10] M. Xu, Y. Zhao, Y. Liu, B. Jiang, C. Yuan, L.-Y. Duan, and Y. Rui, "Multi-VTON: Multi-view virtual try-on," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14 500–14 510.

[11] H. Le, M. Nguyen, and W. Q. Yan, "A vision aid for the visually impaired using commodity dual-rear-camera smartphones," *International Conference on Mechatronics and Machine Vision*, 2018.

[12] W. Q. Yan, *Robotic Vision: From Deep Learning to Autonomous Systems*. Singapore: Springer, 2025.

[13] Y. Zhang, "ChatClothes: An AI-Powered Virtual Try-On System," Master's thesis, Auckland University of Technology, New Zealand, June 2025.

[14] A. Zheng, "Video understanding with attention encoder and multimodal large language model," Master's thesis, Auckland University of Technology, New Zealand, June 2025.

[15] H. Le, "Sarm: Synthetic data annotation for enhancing the experiences of augmented reality application based on machine learning," PhD thesis, Auckland University of Technology, New Zealand, June 2022.

[16] D. Morelli, M. Fincato, M. Cornia, L. Baraldi, and R. Cucchiara, "Dress Code: High-resolution multi-category virtual try-on," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 2231–2239.

[17] X. Gao, M. Nguyen, Y. Liu, and W. Q. Yan, "VICL-CLIP: Enhancing face mask detection in context with multimodal foundation models," *International Conference on Neural Information Processing (ICONIP)*, 2024.

[18] H. Tran, M. Nguyen, H. Le, and W. Q. Yan, "A personalised stereoscopic 3D gallery with virtual reality technology on smartphone," *International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2017.

[19] M. Nguyen, H. Tran, H. Le, and W. Q. Yan, "A tile based colour picture with hidden qr code for augmented reality and beyond," *ACM Symposium on Virtual Reality Software and Technology (VRST)*, 2017.

[20] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2d pose estimation using part affinity fields," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7291–7299.

[21] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "U^2-Net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, p. 107404, 2020.

[22] S. Lee, S. Choi, J. Park, and I. K. Choi, "High-resolution virtual try-on with misalignment-aware normalization," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 235–251.

[23] W. Song, Q. Zhang, L. Xu, and J. Gao, "OOTDiffusion: Outfitting fusion with text-guided diffusion for fashion image synthesis," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 12 356–12 365.

[24] W. Zhu, M. Li, H. Huang, L. Sun, W. Zhang, and H. Liu, "TryOnDiffusion: A tale of two UNets," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[25] Q. Zhang, J. Zhou, and X. Wang, "Adaptive diffusion models for conditional image synthesis," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[26] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[27] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine-tuning text-to-image diffusion models for subject-driven generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[28] N. Kumari, R. Zhang, E. Shechtman, R. Zhang, A. Hertzmann, S. Paris, P. Isola, and T. Park, "CustomDiffusion: Customizing text-to-image diffusion models for subject-driven generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 439–19 449.

[29] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[30] W. Zhu, B. Peng, and W. Q. Yan, "Dual knowledge distillation on multiview pseudo labels for unsupervised person re-identification," *IEEE Transactions on Multimedia*, vol. 26, pp. 7359 – 7371, 2024.

[31] Z. Liu, P. Luo, X. Wang, and X. Tang, "DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5337–5345.

[32] Y. Ge, R. Zhang, X. Liu, P. Luo, X. Wang, and X. Tang, "DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] S. Zheng, Y. Yang, Z. Chen, P. He, S. Liang, J. Lai, and L. Lin, "ModaNet: A large-scale street fashion dataset with polygon annotations," in *Proceedings of the 26th ACM International Conference on Multimedia (ACM MM)*, 2018.

[34] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 6105–6114.

[35] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, H. Adam, and Q. V. Le, "Searching for MobileNetV3," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "LLaVA: Large language-and-vision assistant," *arXiv preprint arXiv:2304.08485*, 2023.

[38] W. Q. Yan, *Computational Methods for Deep Learning: Theory, Algorithms, and Implementations*. Singapore: Springer, 2023.

[39] H. Le, M. Nguyen, and W. Q. Yan, "Augmented reality and machine learning incorporation using YOLOv3 and ARKit," *Applied Sciences*, 2021.

[40] M. Nguyen, M. P. Lai, H. Le, and W. Q. Yan, "A web-based augmented reality platform using pictorial QR code for educational purposes and beyond," *ACM Symposium on Virtual Reality Software and Technology*, 2019.

[41] K. Zhao, M. Nguyen, and W. Q. Yan, "Evaluating accuracy and efficiency of fruit image generation using generative ai diffusion models for agricultural robotics," *International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2024.

[42] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.