ChatClothes:
An AI-Powered Virtual Try-On System

Yuchao Zhang

A thesis submitted to the Auckland University of Technology
in partial fulfillment of the requirements for the degree of
Master of Computer and Information Sciences (MCIS)

2025

School of Engineering, Computer & Mathematical Sciences

Abstract

With the advancement of deep learning, latent diffusion models, and large language models (LLMs), virtual try-on (VTON) has emerged as a promising solution for personalized fashion experiences in online shopping, digital design, and augmented retail. This thesis proposes ChatClothes, a modular and multimodal VTON system that integrates controllable diffusion-based generation with dialogue-driven garment interaction.

The system architecture is orchestrated by Dify, with ComfyUI managing the visual generation pipeline and Ollama hosting local LLMs. At its core, ChatClothes employs DeepSeek, a customized large language model that interprets natural language instructions and transforms them into structured prompts for image generation and interactive refinement. This prompt-based guidance enhances semantic alignment and enables intuitive user control beyond predefined attribute labels.

To improve structural consistency and detail fidelity in image synthesis, this work introduces Low-Rank Adaptation (LoRA) for fine-tuning the original OOTDiffusion model. Without altering the backbone architecture, this strategy focuses on enhancing pose alignment, hand generation accuracy, and garment texture reconstruction. By integrating LoRA modules, the model achieves effective adaptation and fine-grained refinement even under limited training resources.

To support garment classification, YOLO12n-LC, a lightweight variant based on YOLO12n, is developed to balance accuracy, speed, and model size. It achieves competitive performance across multiple clothing categories while maintaining feasibility for device-level deployment.

A complete system workflow connects image preprocessing, language understanding, garment classification, image synthesis, and output evaluation. Experiments on datasets such as DressCode and VITON-HD demonstrate the system' s initial validation in terms of realism, controllability, structural preservation.

This work presents a unified framework bridging vision-language interaction with diffusion-based generation, establishing a foundation for scalable, user-centered, and device-adaptable fashion AI systems applicable across e-commerce, AR fitting mirrors, personalization platforms, and automated outfit design.

Keywords: Virtual Try-On, Diffusion Models, OOTDiffusion Fine-Tuning, DeepSeek, YOLO12n-LC, LoRA, ComfyUI, Chat Robot

Contents

List of Figures

List of Tables

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature: **Date:** 19 May 2025

Acknowledgment

As this thesis reaches its completion, I would like to express my heartfelt gratitude to all the individuals and organizations who have supported and assisted me throughout the research and writing process.

First and foremost, I would like to extend my sincere thanks to my primary supervisor, Professor Wei Qi Yan, for his continuous guidance, support, and encouragement throughout every stage of this research. His academic expertise and insightful feedback have greatly influenced the direction and quality of this work.

I would also like to thank all the professors and lecturers who taught me at Auckland University of Technology (AUT). Their teaching and support throughout my studies have laid a strong foundation for my academic growth and research accomplishments.

I am deeply grateful to my family, especially my wife and son, for their unwavering support, patience, and understanding during my postgraduate journey. Their encouragement has been a constant source of strength and motivation.

My sincere appreciation also goes to my friends and classmates for their valuable discussions, technical help, and collaborative spirit, all of which contributed meaningfully to the progress of this thesis.

Finally, I wish to acknowledge AUT for providing a stimulating academic environment and the necessary resources that enabled me to conduct this research. The opportunities and infrastructure offered by the university have played a vital role in my scholarly development.

To everyone who has supported me along the way—thank you. Your kindness and generosity have made this achievement possible.

Yuchao Zhang

Auckland, New Zealand

May 2025

Chapter 1
Introduction

This chapter introduces the background and motivation for developing an intelligent, interactive virtual try-on system. It outlines the core challenges in the fields of fashion AI, image generation, and user-centered interaction. The chapter also presents the main research questions, summarizes the key contributions and objectives of the study, and concludes with an overview of the thesis structure.

1.1 Background and Motivation

Virtual Try-On(VTON) is rapidly transforming the fashion and e-commerce sectors by enabling users to digitally preview garments in a realistic and interactive manner. This paradigm shift addresses the increasing demand for personalized, convenient, and immersive online shopping experiences (Han et al., 2018; Wang et al., 2018). By allowing users to visualize clothing items on virtual models before making a purchase, VTON systems help reduce product return rates, improve size and style satisfaction, and enhance overall customer engagement and trust in online platforms.

As consumers become more accustomed to digital experiences, their expectations for realism and interactivity have grown considerably. However, despite the growing adoption of VTON systems across major fashion retailers and platforms, many technical and user-centric challenges remain unresolved. Traditional approaches often rely on rigid garment overlays or static image warping techniques, which may suffice in controlled environments but fail to generalize well across diverse human poses (Cui et al., 2021), complex garment types, and varying lighting or background conditions (Baldrati et al., 2023; Chen et al., 2024, 2023). These methods typically struggle with pose misalignment, lack of texture fidelity, and limited adaptability to varied body shapes or personalized user intentions (Dong et al., 2022). Consequently, in real-world applications, mismatches in garment fit and visual realism persist as bottlenecks that undermine user satisfaction, limiting widespread adoption and commercial viability.

On the technical front, recent advancements in computer vision and deep learning have paved the way for more sophisticated VTON methods. For instance, improvements in human parsing, keypoint detection, pose estimation, and garment segmentation have made it possible to construct more accurate 2D or 3D representations of the human body and clothing structure. These capabilities serve as the foundation for advanced image synthesis pipelines that can simulate how a specific garment would appear on a specific body in a specific pose. Moreover, the availability of large-scale annotated datasets—such as DeepFashion, VITON-HD, and DressCode—has enabled the supervised training of robust deep learning models capable of handling diverse styles, garments, and human poses with greater generalization capabilities (Morelli et al., 2022; Liu et al., 2016). These datasets often include rich metadata such as keypoints, segmentation masks, and garment categories, which are critical for conditioned generation models.

In parallel, the field of natural language processing has undergone a revolution with the ad-

vent of large language models(LLMs) such as GPT-3, ChatGPT, LLaMA, and DeepSeek. These models are capable of understanding and generating human-like text across a wide range of domains. Integrating LLMs into VTON systems has introduced a new paradigm of multimodal interaction, enabling users to control the try-on process using intuitive text or voice instructions. Instead of relying on drop-down menus or hard-coded garment categories, users can now issue commands(Guo et al., 2025). This transition toward dialog-based, natural language-driven interaction significantly lowers the barrier to entry and makes VTON systems more accessible to a broader user base. It also enables iterative and dynamic control, allowing users to refine and personalize their try-on experience over multiple turns of conversation.

Simultaneously, diffusion-based generative models have emerged as a powerful alternative to traditional generative adversarial networks(GANs) Gao et al. (2022,?). Models such as Stable Diffusion and OOTDiffusion employ a latent denoising strategy to progressively generate high-quality images from noisy latent representations (Lee et al., 2022; Xu et al., 2024; Gao et al., 2021). These models are not only capable of producing highly detailed and structurally consistent outputs but are also better suited for incorporating multiple conditioning inputs, such as pose heatmaps, garment masks, and prompt-based embeddings. As a result, diffusion models have become increasingly favored in applications where visual realism and spatial alignment are critical, such as VTON, inpainting, and text-to-image synthesis. The flexibility and controllability afforded by latent diffusion processes make them ideal candidates for building the next generation of garment synthesis engines.

The convergence of these three pillars—advanced image generation, natural language understanding, and high-resolution garment datasets—provides fertile ground for building robust, user-centric, and scalable VTON systems. This thesis proposes such a system, named ChatClothes, which integrates a fine-tuned latent diffusion generator(OOTDiffusion), a natural language interface(DeepSeek via Ollama), and a lightweight garment classification module(YOLO12n-LC). These components are modularly orchestrated via a service framework involving Dify, ComfyUI, and Docker containers, enabling seamless module coordination, fast deployment, and hardware flexibility.

From a practical standpoint, there is growing demand for virtual shopping assistants that go beyond static image previews. Today' s users expect real-time responsiveness, stylistic adaptability, and a personalized user experience. They desire tools that support interactive garment manipulation, intelligent outfit recommendations, and seamless integration across platforms—

including mobile phones, smart mirrors, and VR/AR environments. By combining conversational AI with photorealistic garment generation, ChatClothes addresses these expectations directly. It also opens up potential applications beyond retail, including digital wardrobe planning, virtual fashion shows, social media content creation, and stylist-guided customization in metaverse environments.

In summary, while notable progress has been made across body modeling, image synthesis, and user interaction, significant challenges remain in achieving garment-level control, cross-modal coherence, and efficient deployment in real-world environments. This thesis addresses these gaps by proposing an integrated virtual try-on framework that leverages state-of-the-art diffusion generation, large language models, and lightweight classification within a practical, extensible, and deployable system architecture.

1.2 Research Questions

This research is driven by two overarching goals: To develop an intuitive, interactive virtual try-on system and to improve the visual quality and controllability of clothing image synthesis. While recent progress has been made in both user interaction and generative modeling, integrating these components into a unified, robust system still poses challenges.

On one hand, enabling users to interact naturally and dynamically with try-on systems remains complex. Existing solutions often lack multimodal flexibility and struggle with seamless user-driven control. Incorporating large language models(LLMs) into this process introduces the potential for personalized, dialogue-based interaction, but also raises issues regarding semantic alignment, cross-modal coordination, and conversational robustness.

On the other hand, although diffusion-based methods such as OOTDiffusion demonstrate promising results, problems like garment misalignment, texture degradation, and poor generalization to unseen poses still affect the output. Enhancing generation controllability and visual fidelity is key to producing convincing and personalized try-on experiences.

Accordingly, this thesis explores the following research questions:

How can a user-friendly and interactive virtual try-on system be designed, integrating conversational language models with visual generation modules?

What methods can be employed to enhance garment detail preservation, pose alignment, and controllability in diffusion-based try-on models?

How does the incorporation of multimodal inputs—such as voice, text, and image—impact system usability and user satisfaction?

These questions form the foundation for developing the ChatClothes system and for evaluating its effectiveness in delivering responsive and personalized virtual try-on functionality.

1.3 Contributions

This thesis presents the design and development of a novel virtual try-on system—ChatClothes—which integrates large language models(LLMs), diffusion-based image generation techniques, and lightweight garment classification models into a unified, extensible framework. The system addresses key challenges in existing VTON applications, such as limited interactivity, insufficient garment realism, and the lack of modular control pathways between language and vision modules.

The main contributions of this research are as follows:

Modular System Architecture: A flexible and extensible architecture is proposed, incorporating DeepSeek(via Ollama) for language-based control, OOTDiffusion(via ComfyUI) for high-fidelity latent diffusion-based image generation, and a multimodal interaction interface supporting natural language and visual input. The overall pipeline is orchestrated through Dify and Docker, allowing for seamless coordination across modules and facilitating extensibility for future functionalities.

Customized Fine-Tuning of OOTDiffusion: The image generation backbone, OOTDiffusion, is fine-tuned using multiple conditioning inputs—pose maps, garment masks, and prompt embeddings—allowing for controllable, accurate try-on results. Training enhancements include garment alignment loss, facial identity preservation, and semantic alignment via CLIP-based guidance. These improvements help preserve garment texture and structure under diverse poses and user commands.

Integration of a Lightweight Clothing Classification Module: A lightweight garment classification module named YOLO12n-LC is developed and integrated into the pipeline. While the overall system is not designed for edge deployment, this module itself is optimized for low latency and reduced resource consumption. It supports real-time garment recognition, prefiltering, and category tagging, enabling faster downstream processing and better user-driven customization.

In summary, this thesis proposes a complete and modular virtual try-on framework that combines conversational AI, lightweight classification, and advanced diffusion-based image generation. The system offers a foundation for future applications in e-commerce, fashion customization, and human-computer interaction research.

1.4 Objectives of This Thesis

This thesis aims to explore and implement an intelligent, high-fidelity, and naturally interactive virtual try-on system by leveraging large language models and diffusion-based image generation techniques. The research is guided by the following four objectives:

The first goal is to conduct a comprehensive review of the technologies underpinning virtual try-on systems, including diffusion-based generative models, conditional image synthesis techniques, and multimodal interaction frameworks. This review establishes the theoretical foundation for this work and situates the proposed approach within the broader research landscape of intelligent fashion systems.

The second goal is to fine-tune a diffusion-based image generator, OOTDiffusion, for generating realistic and controllable try-on images. The model is enhanced through pose conditioning and garment masking, and optimized using Low-Rank Adaptation(LoRA) modules to improve detail fidelity—particularly in limb alignment, hand structure, and texture rendering. The fine-tuned generator is deployed within a Docker-managed environment coordinated by Dify, supporting modular reuse and system-level stability.

The third goal is to integrate and optimize a lightweight garment classification module based on YOLO12n-LC. This component enables rapid and accurate identification of clothing categories in input images, supports semantic tagging, input filtering, and contributes structured visual information to guide prompt generation for LLM interaction. It is critical to maintaining system responsiveness and improving semantic control across diverse user inputs.

The last goal is to design and implement the overall ChatClothes system architecture, which integrates natural language interaction, image generation, and visual verification into a unified workflow. The system leverages modular components including DeepSeek, Ollama, and ComfyUI, enabling users to dynamically control the try-on process through intuitive language commands. The final system is scalable, extensible, and suitable for deployment in practical scenarios such as online retail, personal styling, and digital fashion design.

Together, these objectives define the technical scope and research contribution of ChatClothes—a multimodal, user-centered virtual try-on framework that bridges language understanding with controllable image synthesis for next-generation fashion applications.

1.5 Structure of This Thesis

This thesis is organized into six chapters, providing a comprehensive exploration of the design, implementation, and evaluation of the proposed ChatClothes virtual try-on system. Each chapter addresses a specific aspect of the research as follows:

Chapter 1 –Introduction: Presents the research motivation, background, and current challenges in the virtual try-on domain. It also outlines the research questions, objectives, and contributions of the proposed system.

Chapter 2 –Literature Review: Reviews the development of virtual try-on technologies, focusing on the evolution from traditional warping methods to GAN-based and diffusion-based approaches. It also surveys large language models(LLMs), multimodal interaction frameworks, and lightweight garment classification models relevant to this study.

Chapter 3 –Methodology: Describes the architecture and key components of the ChatClothes system, including the DeepSeek-based conversational control, the OOTDiffusion image generator, and the YOLO12n-LC garment classifier. This chapter also explains system orchestration using Dify and ComfyUI, and details the dataset processing, model customization, and training strategies.

Chapter 4 –Experimental Results: Presents both quantitative and qualitative evaluation of the system’s performance. This includes comparisons with baseline methods, image quality analysis(SSIM, FID, LPIPS), garment classification accuracy, edge deployment performance, and user satisfaction based on subjective evaluation.

Chapter 5 –Discussion and Analysis: Interprets experimental findings, analyzes the contribution of individual modules, compares with related systems, and discusses current limitations. It also explores implications for real-world deployment and application scenarios.

Chapter 6 –Conclusion and Future Work: Summarizes the research outcomes and proposes future directions. This includes model acceleration(e.g., LCM), multimodal control, enhanced personalization, and adaptation to complex real-world conditions.

This structure is designed to systematically guide the reader through the conceptual foun-

dation, technical implementation, experimental validation, and practical implications of a next-generation, multimodal virtual try-on system.

Chapter 2
Literature Review

This chapter provides a concise literature review on the key research areas relevant to this study, including virtual try-on(VTON) technologies, diffusion-based generative models, and large language models(LLMs). It traces the evolution of each field, examines their roles in enabling interactive fashion systems, and highlights the technical gaps that motivate the development of the proposed multimodal AI chatbot framework.

2.1 Introduction

This chapter reviews the relevant literature and foundational technologies that support the development of modern virtual try-on(VTON) systems. Drawing upon advancements in generative computer vision, natural language processing, and system-level interaction frameworks, it provides a comprehensive overview of the technological landscape underpinning the methodology of this thesis.

The literature review is structured around three major research pillars: virtual try-on(VTON) technologies, diffusion-based generative models, and large language models(LLMs). In addition, orchestration frameworks, visual workflow engines, and lightweight vision models are discussed as supporting components.

Section 2.2 outlines the evolution of virtual try-on systems, tracing the development from early image-based garment transfer methods to diffusion-driven approaches. It compares classical techniques such as VITON and CP-VTON with newer models including OOTDiffusion and CatVTON, highlighting shifts in garment fidelity, pose generalization, and control mechanisms.

Section 2.3 introduces diffusion models as a new paradigm for generative image synthesis, analyzing their advantages over GANs and their applications in VTON tasks. Representative models such as OOTDiffusion, IDM-VTON, and TryOnDiffusion are discussed in relation to controllability, semantic grounding, and high-resolution generation.

Section 2.4 focuses on large language models(LLMs), including DeepSeek and Qwen-VL. It reviews their role in natural language understanding, prompt-driven customization, and conversational garment control, which collectively lower the barrier for user interaction in virtual try-on systems.

Section 2.5 presents the Dify orchestration framework, which coordinates dialogue understanding, prompt routing, and multimodal module management within the ChatClothes architecture. Its lightweight, modular nature enables flexible and scalable system deployment.

Section 2.6 explores the ComfyUI visual workflow engine, which underpins the diffusion-based image generation pipeline. Its node-based architecture facilitates flexible workflow construction, enabling fine-grained control over sampling, conditioning, and refinement processes.

Section 2.7 describes the deployment of vision-language models within the Ollama runtime framework. It outlines how local hosting of models such as LLaMA3-Vision and Qwen2.5-VL enhances privacy, reduces latency, and improves visual-semantic alignment in the system.

Section 2.8 reviews YOLO-based lightweight vision models, focusing on their role in real-time garment detection and clothing classification. Although not directly involved in image generation, these models provide essential pre-processing and structural information for downstream synthesis tasks.

Together, these technologies form the technical foundation of the ChatClothes system. By integrating advances across vision, language, diffusion generation, and orchestration, the system aims to deliver a high-fidelity, interactive, and scalable virtual try-on experience.

Traditional virtual try-on methods primarily relied on 2D image warping and geometric garment transfer, which often struggled to preserve fabric integrity and adapt to varied poses and body structures. The introduction of generative adversarial networks(GANs) improved visual realism but suffered from instability and limited controllability. More recently, latent diffusion models(LDMs), including Stable Diffusion, ControlNet, and OOTDiffusion, have demonstrated superior performance in texture fidelity, pose alignment, and semantic conditioning, providing a robust generative foundation for modern VTON systems.

In parallel, large language models have revolutionized user-system interaction through multimodal reasoning and conversational command parsing. Models like GPT-4, Claude, DeepSeek, and Qwen-VL enable intuitive garment customization via natural language prompts, lowering the operational barrier for non-technical users and expanding the accessibility of virtual try-on platforms.

The convergence of vision and language modeling has further catalyzed the development of multimodal intelligent systems. Frameworks such as BLIP, DALL · E, and Qwen-VL exemplify this integration, while orchestration layers like Dify and visual engines like ComfyUI enable scalable coordination across heterogeneous components. These innovations collectively enable systems like ChatClothes to offer real-time, personalized, and interactive fashion experiences.

The following sections detail these developments, providing the technical background necessary for understanding the system design and implementation described in Chapter 3.

2.2 Evolution of Virtual Try-On

Virtual Try-On(VTON) leverages computer vision, deep learning, and interactive AI systems to simulate garment fitting experiences in virtual environments. This section reviews its definition, evolution, representative models, and key limitations based on current literature.

2.2.1 Definition and Importance of Virtual Try-On

Virtual try-on enables users to preview how garments, accessories, or even makeup would look on them before making a purchase. By bridging the gap between physical and online shopping, VTON helps reduce return rates caused by sizing or style mismatches and improves customer satisfaction and operational efficiency in e-commerce platforms (Han et al., 2018; Wang et al., 2018; Kumar et al., 2019).

Within the fashion domain, VTON has become a critical tool for visual recommendation, stylistic personalization, and online garment presentation. It offers advantages in scalability, consistency, and convenience over physical try-on. As e-commerce platforms and metaverse retail grow, VTON has evolved from a supplementary function into a core component of digital fashion services. Its ability to combine visualization, personalization, and interactivity offers a unique advantage for immersive user engagement.

2.2.2 Historical Evolution of VTON

The development of virtual try-on technologies has undergone a remarkable evolution, which can be broadly categorized into three major stages: early template-based methods, GAN-based deep generative approaches, and the recent adoption of diffusion models.

In the initial stage, early solutions primarily relied on geometric warping and template matching techniques. Models such as VITON (Han et al., 2018) proposed a two-stage pipeline that leveraged pose heatmaps and clothing masks to achieve basic 2D image-based try-on. Building upon this foundation, CP-VTON (Wang et al., 2018) introduced Thin-Plate Spline(TPS) warping to better align garments with body poses, significantly improving spatial fidelity. Further advancements were realized with VITON-HD (Choi et al., 2021), which enhanced garment parsing and TPS-based alignment to support high-resolution outputs up to 1024×768 pixels. However, despite their computational efficiency, these early methods often struggled to preserve fine garment textures and realistic deformation, particularly under complex human poses (Xie et al., 2023).

The emergence of Generative Adversarial Networks(GANs) ushered in the second stage of VTON development, dramatically enhancing the realism of generated try-on images. Flow-guided architectures, such as ClothFlow (Han et al., 2019) and TryOnGAN (Lewis et al., 2021), enabled more accurate modeling of garment deformation across diverse body poses. HR-VTON

(Lee et al., 2022) introduced occlusion-aware refinement modules to better preserve garment boundaries and mitigate common artifacts around body joints. Additionally, StyleGAN-VTON (He et al., 2022) exploited latent style control mechanisms to maintain identity consistency and fine garment structures. Despite these improvements, GAN-based methods suffered from inherent limitations, including mode collapse, training instability, and restricted spatial controllability. Their performance in handling semantic alignment, pose variation, and user-intent customization remained insufficient for fully personalized virtual try-on applications.

The most recent and transformative shift in VTON research has been driven by diffusion-based generative models (Li et al., 2025; Cui et al., 2024), which have established new benchmarks for photorealistic garment synthesis. Unlike GANs, diffusion models operate in the latent space and utilize an iterative denoising process, offering superior controllability, robustness, and sample fidelity. Representative works such as OOTDiffusion (Xu et al., 2024) extend the capabilities of Stable Diffusion by incorporating pose maps(Li et al., 2023), garment masks(Lin et al., 2025), and prompt-based semantic guidance(Fele et al., 2022), enabling highly personalized try-on generation. IDM-VTON (Choi et al., 2024) further refines this approach by integrating multi-scale semantic encoding and cross-attention mechanisms to enhance texture preservation. TryOnDiffusion (Zhu et al., 2023) introduces a two-stage dual U-Net structure, enabling high-resolution garment generation with improved structural details. Meanwhile, CatVTON (Chong et al., 2024) focuses on lightweight architectural designs optimized for deployment on mobile and edge devices.

Overall, diffusion models(Fang et al., 2024) have demonstrated significant advantages over their predecessors, particularly in terms of supporting multi-condition control(pose, mask, and prompt inputs), maintaining fine-grained garment details, and exhibiting resilience against occlusions and complex body poses. Their iterative generation process inherently improves visual fidelity, making diffusion-based VTON systems the current state-of-the-art solution for achieving highly realistic, user-controllable virtual try-on experiences.

Table 2.1: Comparison of key virtual try-on models

Model	Architecture	Control Method	Advantages
VITON	U-Net	Pose + Mask	Simple, interpretable pipeline
CP-VTON	TPS + RefineNet	Pose alignment	Improved spatial matching
HR-VTON	Flow + GAN	Occlusion handling	Sharp boundaries, high resolution
StyleGAN-VTON	StyleGAN	Latent style vector	Identity consistency, high fidelity
OOTDiffusion	Latent Diffusion	Pose + Mask + Prompt	Personalized, controllable synthesis
IDM-VTON	Cross-attention Diffusion	Multi-level semantic embedding	Fine-grained textures, contextual alignment
TryOnDiffusion	Dual U-Net	High-res refinement	High-definition results
CatVTON	Lightweight Diffusion	Mask + Pose	Efficient and deployable

2.2.3 Comparison of Core Models

Figure 2.1: Comparison of different virtual try-on techniques.From left to right:(1) reference person image,(2) target garment image,(3–5) try-on results generated by different stages.

2.2.4 Limitations and Future Trends

Despite substantial progress, current VTON systems face several critical challenges:

Garment-body alignment: Inaccurate fitting under occlusion and extreme poses.

Texture fidelity: Limited capacity to retain fine patterns and realistic clothing materials.

Interactive control: Weak support for fine-grained garment modifications and style tuning.

Generalization: Poor adaptability to unseen garments and user poses.

Real-time performance: Inference latency still limits deployment on mobile platforms.

Several trends have emerged to address these limitations. First, the integration of large language models(LLMs) into VTON pipelines allows systems to respond flexibly to natural language commands. Prompt-based control has proven more expressive than static class labels, offering finer semantic granularity and enabling iterative interaction.

Second, fast and controllable diffusion architectures are being developed to improve real-time performance. Methods such as LCM, FastComposer, and LoRA-finetuned generators aim to reduce denoising steps while maintaining visual quality. Multi-stage models like TryOnDiffusion also demonstrate the benefit of decoupling garment alignment and image refinement.

Third, multimodal input support(including gesture, voice, or sketch) is gaining traction. This aims to enhance accessibility and improve personalization, especially in AR/VR or mobile applications.

Finally, edge deployability remains a critical requirement. Lightweight models such as CatVTON(Chong et al., 2024) and our proposed YOLO12n-LC prioritize fast inference, reduced memory usage, and adaptability to embedded systems like Raspberry Pi.

In summary, VTON technologies have evolved from rigid geometric overlays to flexible, high-fidelity diffusion models. This thesis builds upon recent progress by incorporating controllable diffusion generation, multimodal interfaces, and efficient garment classification into a unified system—ChatClothes—paving the way for accessible and intelligent fashion interaction experiences.Recent work has also proposed integrating deformable attention mechanisms(Bai et al., 2022; Abbas et al., 2024) and auxiliary segmentation modules(Kumar et al., 2019) to streamline try-on fidelity and garment-body alignment.

2.3 Diffusion Models

Diffusion models have emerged as a dominant generative framework for photorealistic image synthesis, significantly influencing fields such as medical imaging, natural scene generation, and virtual try-on(VTON). Compared to GANs, diffusion models offer stable training, flexible conditioning, and superior visual quality, making them particularly well-suited for controllable and high-fidelity try-on tasks.

2.3.1 Evolution and Application in Virtual Try-On

Initially developed as likelihood-based generative models, early diffusion approaches such as DDPM reconstructed clean images by reversing a gradual noise process. Although these pixel-space methods achieved quality on par with GANs, they often required hundreds or thousands of denoising steps, resulting in high computational cost. The advent of latent-space diffusion—exemplified by Stable Diffusion—dramatically improved efficiency by mapping images into a lower-dimensional latent space via a pretrained VAE and performing denoising there, reducing sampling steps to the tens while preserving high resolution (Zhang et al., 2024; Zunair et al., 2022).

In the virtual try-on domain, diffusion models have proven superior to GANs by supporting richer conditioning inputs such as pose maps, segmentation masks, and textual prompts. Methods like OOTDiffusion leverage multi-modal conditioning to fuse human keypoints with garment segmentation and fabric textures, thereby enhancing garment-body alignment and texture preservation (Xu et al., 2024). IDM-VTON further introduces pose-consistency regularization and segmentation-aware losses to reduce misalignment and distortion (Choi et al., 2024). To enable edge deployment, lightweight adaptations such as CatVTON simplify network structures by early concatenation of pose and segmentation maps, reducing parameter count and inference latency with minimal impact on image quality (Chong et al., 2024).

Beyond these developments, three recent works illustrate significant innovations in unifying generation and editing, achieving fine-grained multimodal control, and supporting modular, interactive real-time applications. The Unified Diffusion Model distills semantic priors from large language models(e.g. GPT-4, LLaMA) into a multi-branch decoder with classifier-free guidance, enabling both high-quality text-to-image synthesis and precise attribute editing—in a single sampling pass—under complex prompts. ModelScope ' s Nexus-Gen employs adapter distillation and multi-task fine-tuning to inject CLIP and DALL · E visual-language features into a frozen diffusion backbone, allowing rapid prototyping of pose transfer, garment swapping, and style morphing variants with only lightweight adapter updates (Zhang et al., 2025a). RiverZ ' s ICEdit provides a web-based interface in which brush strokes and natural-language instructions(e.g. "roll up the left sleeve slightly") are parsed by a distilled language-understanding module into control vectors; combined with an optimized DDIM scheduler and parallel denoising, this yields sub-second responsiveness for interactive try-on edits (Zhang et al., 2025c). Each of these approaches emphasizes unified multi-tasking, decoupled modular design, and efficient

sampling strategies that redefine usability and scalability for VTON systems.

2.3.2 Control Mechanisms and Integration Strategies

A notable advantage of diffusion models is their adaptability to multimodal control inputs. Cross-attention mechanisms, particularly those guided by CLIP embeddings, allow the model to incorporate semantic cues such as textual descriptions, pose configurations, or garment categories. External modules like ControlNet and T2I-Adapter further enhance controllability by introducing auxiliary branches that accept structured inputs such as pose heatmaps, edge maps, or sketches (Zhang et al., 2023; Tumanyan et al., 2023). These techniques enable more precise control over spatial alignment and visual attributes in the generation process.

To address the computational complexity inherent in diffusion models, parameter-efficient fine-tuning strategies like LoRA have been employed. By adapting only low-rank matrices within attention and feedforward blocks, LoRA significantly reduces memory and computation overhead during training, making task-specific adaptation more practical. While other strategies such as model quantization and knowledge distillation have been explored in related work, they are not yet integrated into the current system. Nonetheless, these approaches represent promising directions for future optimization of diffusion pipelines, particularly for edge deployment scenarios where memory and speed constraints are critical.

2.3.3 Current Challenges and Future Trends

Despite their advantages, diffusion-based VTON systems face several persistent challenges. Generation latency remains a key bottleneck, particularly when high-resolution outputs require dozens of denoising steps. Reducing sampling steps without compromising texture fidelity remains an open problem; promising directions include latent consistency models and advanced scheduler designs that accelerate sampling while maintaining visual quality.

Robustness to complex human poses is another significant concern. When pose estimations are inaccurate or partial, the resulting garment placement may be misaligned or distorted. This issue is further exacerbated by structural inconsistencies in limb and hand generation—common failure cases include dislocated arms, fused fingers, or mirrored hands. These errors are particularly evident when sleeves intersect with the torso or when hands occlude parts of the garment. Improving anatomical coherence under these conditions may require structure-aware modeling

strategies, such as integrating 3D human mesh priors, explicit limb refinement modules, or hand keypoint supervision into the diffusion pipeline.

Furthermore, diffusion models struggle with complex or cluttered backgrounds, especially when users provide images taken in real-world settings. In such scenarios, inaccurate foreground segmentation or conditioning noise can severely affect garment-body alignment and lead to background-bleed artifacts. Enhancing the model's ability to disentangle foreground garments from intricate scenes may involve dual-branch architectures or semantic-aware control paths that explicitly model background-foreground separation.

Ambiguity in unconstrained textual prompts also remains a barrier to controllability. Many natural language descriptions are context-dependent or semantically vague, leading to unexpected interpretations during image generation. This highlights the need for improved language grounding through large-language-model distillation and dynamic prompt optimization.

To address the above, robust alignment of multiple conditioning modalities may be enhanced through cross-modal contrastive learning and self-supervised objectives that jointly optimize text, pose, and segmentation features. On the deployment front, unified frameworks for pruning, quantization, and low-rank adaptation must be developed to systematically balance model size, inference speed, and visual quality in resource-constrained environments.

Looking forward, the fusion of large-model semantic distillation with highly efficient quantization and adaptive sampling holds great promise for enabling truly real-time, controllable, and personalized virtual try-on experiences across a wide range of platforms—from high-end servers to mobile and edge devices.

2.4 Large Language Models

In recent years, the rapid advancement of deep learning and neural language modeling has led to the emergence of large language models(LLMs) as foundational components in artificial intelligence. Initially developed for tasks such as next-word prediction and sentence completion, these models have evolved into complex systems capable of reasoning, instruction following, dialog understanding, and multimodal coordination. Within the virtual try-on(VTON) domain, LLMs have become essential semantic interpreters, conversational controllers, and interface mediators between user intent and image generation modules.

2.4.1 Historical Development and Key Milestones

The concept of pretraining large neural networks on massive textual corpora has been explored for over a decade. However, the introduction of the Transformer architecture by Vaswani et al. in 2017 fundamentally changed the way sequential data is processed. The attention-based design eliminated recurrence and enabled parallel processing, laying the foundation for the first generation of LLMs. Subsequent models such as GPT-2 and GPT-3 demonstrated unprecedented capabilities in zero-shot and few-shot learning, with GPT-3 achieving remarkable performance across diverse tasks ranging from code synthesis to medical diagnosis without task-specific fine-tuning. GPT-4 further expanded these capabilities by supporting multimodal inputs, allowing both image and text understanding.

Alongside proprietary models, the open-source community contributed significantly with models like LLaMA and Mistral, which employed techniques such as grouped-query attention and sliding window mechanisms to enhance efficiency. Instruction tuning approaches, including reinforcement learning from human feedback(RLHF), further aligned LLMs with human expectations and practical task goals, driving the evolution toward more responsible and effective AI systems.

2.4.2 Classification and Functional Roles in VTON Systems

Large language models can be classified based on their input modalities and capabilities. Unimodal LLMs, such as GPT-2, GPT-3, and LLaMA, focus solely on textual understanding and generation. Vision-language models(VLMs) like Flamingo and InstructBLIP integrate visual encoders with language decoders, enabling image-grounded textual reasoning and interaction. More advanced multimodal agents, such as Google's Gemini and Alibaba's Qwen-VL, are capable of simultaneously processing text, images, and audio through unified encoders or cross-modal attention mechanisms.

In VTON systems, LLMs contribute across multiple functional layers. They facilitate intuitive user interaction by allowing free-form textual expressions of fashion preferences, such as "something more elegant" or "a hoodie with a streetwear vibe," which are then interpreted into latent semantic attributes like garment category or style context. They also bridge modalities by converting natural language prompts into visual conditions suitable for image generation. Furthermore, LLMs play a critical role in instruction disambiguation; when user inputs are vague

or ambiguous, the system can generate clarifying questions to maintain semantic alignment, enhancing the naturalness and effectiveness of the interaction.

2.4.3 Challenges and Technical Bottlenecks

Despite their transformative impact, deploying LLMs in virtual try-on environments presents several challenges. Domain alignment remains a major issue, as general-purpose LLMs are often inadequately trained on fashion-specific terminology, seasonal trends, and cultural nuances, potentially leading to misinterpretations. Real-time responsiveness is another limitation, given the computational demands of large models, which hinder their deployment on resource-constrained devices such as mobile phones or smart mirrors. Visual grounding is a further challenge, as unimodal LLMs, and even some vision-language models, may struggle to localize garment regions or interpret complex visual inputs without dense alignment training. Additionally, ensuring safety and supporting deep personalization are ongoing research areas, as LLMs must avoid reinforcing stereotypes or producing inappropriate outputs while adapting to individual user profiles.

2.4.4 Integration within the ChatClothes System

In the ChatClothes architecture, LLMs are integrated at the core of user interaction and control processes. DeepSeek is employed via the Dify orchestration platform to interpret user commands, generate prompts, and manage dialogue flow.

The interaction process is primarily designed to enhance the system's ability to understand user intent and coordinate internal components. User inputs are semantically parsed by the large language model to extract actionable instructions and garment preferences, which are then converted into structured commands recognizable by other modules such as image generation and clothing classification. While the LLM does not participate in embedding generation or guide the diffusion model directly, it serves as the semantic core of the system, significantly improving its ability to handle natural language inputs and enabling intelligent control over the overall workflow.

2.4.5 Future Directions

Future research should focus on enhancing the semantic control capabilities of LLMs within virtual try-on workflows. This includes improving their ability to understand fine-grained fashion-related instructions through domain-specific prompt tuning, and optimizing their responsiveness when coordinating multi-step image generation tasks. Lightweight adaptation techniques such as LoRA or prompt routing may also help maintain performance while supporting deployment in constrained environments. Additionally, exploring ways to better align textual commands with visual outcomes—such as incorporating feedback loops or style constraints—could further improve generation consistency and user satisfaction.

2.4.6 Summary

Large language models serve as a critical bridge between human intent and machine-generated outputs in virtual try-on systems. Their evolution from text-only predictors to multimodal conversational agents has enabled richer, more adaptive fashion AI experiences. Through their integration into the ChatClothes framework, LLMs support scalable, customizable, and intelligent try-on interactions, meeting the growing demand for personalized digital fashion platforms.

2.5 Dify: Orchestrating Multimodal Interaction Workflows

The growing complexity of multimodal AI systems has created a pressing need for robust orchestration frameworks capable of managing diverse components, including language models, vision modules, classifiers, databases, and prompt processors. Dify(Develop-in-your-flow) has emerged as a cutting-edge open-source framework designed specifically to facilitate the development of adaptive, extensible, and context-aware AI applications. Within the scope of this thesis, Dify functions as the centralized coordination hub for the ChatClothes system, enabling real-time interaction routing, prompt interpretation, session state management, and seamless communication across distributed model services.

2.5.1 System Capabilities and Design Principles

Dify is fundamentally designed to enable modular AI workflow development without requiring tightly coupled codebases or hand-written pipelines. Its visual logic-building interface allows

developers to customize workflows using graphical blocks that define how user inputs are interpreted, how external modules are triggered, and how responses are synthesized. Unlike traditional static APIs, Dify introduces runtime flexibility, where model selection, prompt templates, fallback strategies, and output aggregation can be dynamically adjusted based on input context and dialogue history.

Additionally, Dify emphasizes low-code deployment and reproducibility. Workflows can be versioned, shared, and embedded into larger systems, supporting both experimentation and production deployment. It enables chaining of inference nodes, where outputs from modules such as DeepSeek or YOLO12n-LC can be seamlessly passed into downstream components like prompt encoders or image generators. This interoperability significantly reduces engineering overhead and accelerates system development. In multimodal settings like virtual try-on, Dify's support for dialogue memory, role-based context tracking, and conditional branching allows fluid, user-adaptive interactions, ensuring that iterative commands such as "show me something more casual than the last look" are interpreted coherently with previous dialogue context.

2.5.2 Integration within the ChatClothes System

Within the ChatClothes architecture, Dify plays a foundational role by bridging natural language understanding, visual garment generation, garment classification, and user intent resolution. Upon receiving a user command, whether via text, voice-to-text conversion, or API call, Dify parses the instruction to determine the task category—such as initiating a try-on request, replacing a garment, or refining a previous style. It routes the parsed semantics through DeepSeek to extract relevant elements and generates a structured prompt for subsequent processing.

Following prompt structuring, Dify interfaces with ComfyUI and the OOTDiffusion pipeline to initiate image generation, conditioned on extracted attributes such as pose maps, garment masks, categories, or stylistic features. When necessary, it supplements the input with classification results from YOLO12n-LC or retrieved garment metadata to enhance generation specificity.

Throughout multi-turn conversations, Dify maintains session memory, tracking generated images, garment histories, and user preferences to preserve coherence and avoid redundant processing. In the event of failures, such as unavailable model endpoints, Dify invokes fallback strategies that leverage precomputed samples or template-based responses to ensure uninterrupted user experiences. Through this orchestration, Dify acts not merely as a dispatcher but as

a cognitive middleware, translating high-level user intent into coordinated execution sequences across the multimodal system.

2.5.3 Advantages of the Dify-Orchestrated Architecture

The adoption of Dify within ChatClothes offers several significant advantages. Its modularity facilitates easy addition, removal, or replacement of system components, allowing developers to adapt workflows flexibly without extensive code modifications. Multi-turn dialogue support is inherent to Dify, with built-in session memory, token tracking, and conversation threading, which are crucial for interactive virtual try-on scenarios where user preferences evolve iteratively.

Furthermore, Dify empowers developers to design custom prompt engineering strategies, enabling fine control over how natural language inputs influence diffusion-based garment generation. Its architecture seamlessly coordinates textual inputs, vision outputs, and structured metadata, ensuring that information such as garment segmentation masks or prompt embeddings can flow coherently between system modules. With production-readiness features, including API endpoint management, container deployment support, and permission control, Dify serves not only as a development tool but also as a foundation for deploying large-scale commercial virtual try-on platforms.

2.5.4 System Integration and Orchestration Limitations

While Dify provides a convenient orchestration interface for integrating large language models and backend services, several practical challenges arise during its use in the ChatClothes system.

First, routing user instructions through DeepSeek, classification modules, and diffusion pipelines introduces noticeable latency. Although acceptable in small-scale deployments, the response time may become less ideal as the number of sequential calls increases. This affects the fluidity of interactive try-on sessions and highlights the need for more efficient pipeline scheduling or lightweight alternatives in future iterations.

Second, maintaining session context across Dify and external modules—such as ComfyUI workflows or Ollama containers—remains fragile. Because these components operate independently, there is no native mechanism to synchronize state or carry user intent across multi-step interactions. As a result, contextual continuity can be lost between turns unless explicitly reconstructed.

In addition, prompt handling within the LLM interface is sensitive to formatting variations. Inconsistent or overly flexible natural language inputs can lead to ambiguous instructions, causing misinterpretation or unintended model behavior. This makes prompt normalization and instruction templating essential, even in seemingly open-ended dialogue environments.

Overall, while Dify significantly reduces integration overhead, its limitations in latency, session management, and input normalization present practical constraints that must be considered in real-time interactive applications like virtual try-on.

2.5.5 Summary

In summary, Dify provides a powerful orchestration backbone for managing complex multimodal workflows within the ChatClothes system. Its modularity, dynamic prompt management, multi-turn dialogue support, and production-readiness make it a compelling choice for building intelligent, context-aware virtual try-on platforms. By translating free-form user intent into coordinated model interactions and coherent outputs, Dify significantly enhances system agility, scalability, and responsiveness. Its integration into ChatClothes marks a critical step toward realizing real-time, personalized, and semantically aligned fashion synthesis systems for next-generation digital retail experiences.

2.6 ComfyUI: Node-Based Workflow for Vision Synthesis

As diffusion-based image generation models have become increasingly complex and capable, there is a growing demand for flexible, interpretable, and modular design environments. ComfyUI is a node-based visual programming interface designed specifically for constructing and managing generative pipelines, with a particular emphasis on diffusion models. Unlike traditional scripting-based platforms, ComfyUI provides an interactive graphical interface that enables researchers, artists, and engineers to assemble workflows by visually connecting functional nodes, each representing operations such as encoding, sampling, attention, or decoding. Within the ChatClothes system, ComfyUI acts as the core image synthesis backend, executing the OOT-Diffusion pipeline while integrating multi-condition inputs from upstream modules.

2.6.1 System Overview

ComfyUI is architected around principles of composability, transparency, and extensibility. It processes computations as a directed acyclic graph(DAG), where nodes represent atomic units of functionality. These nodes include latent samplers that manage the iterative noise removal process using strategies like DDIM or DPM++; prompt encoders that transform user instructions into latent embeddings through models such as CLIP or T5; control modules that inject external conditions like pose maps, edge maps, or segmentation masks using tools such as ControlNet or T2I-Adapter; UNet denoisers that perform the core reverse diffusion steps; and latent decoders that transform refined latent features into pixel-space images via VAEs or decoder transformers. Each of these nodes can be fine-tuned independently or chained together through custom routing logic, providing high configurability and visual interpretability. Workflows created with ComfyUI are saved in JSON or YAML formats, facilitating reproducibility and version control.

2.6.2 Application in the ChatClothes System

In the ChatClothes system, ComfyUI is responsible for executing all image generation tasks based on the fine-tuned OOTDiffusion model. It acts as both a prototyping environment and a real-time pipeline manager. Natural language commands are first interpreted by DeepSeek and then mapped to parameterized operations, such as selecting clothing type, color, or sleeve length. These interpreted instructions are translated into structured configuration values(e.g., control images, masks, or sampler settings), which are manually or programmatically applied to the ComfyUI workflow.

Body pose maps and clothing segmentation masks are supplied as input conditions through auxiliary control branches in the workflow, helping maintain spatial alignment and garment structure. Sampling configurations—such as the number of inference steps, scheduler type, guidance scale, and seed—can be adjusted prior to execution to reflect the desired generation effect. While ComfyUI does not directly interface with the language model, it serves as a visual, modular backend for condition-driven image synthesis in response to parsed user commands.

2.6.3 Benefits of Using ComfyUI

The integration of ComfyUI into ChatClothes provides several practical benefits that support development efficiency and system flexibility. Its node-based interface makes it easy to visual-

ize and debug the image generation process, as intermediate outputs—including latent tensors, masks, and conditioning inputs—can be inspected in real-time. This transparency aids in model understanding and accelerates iteration during fine-tuning or prompt adjustment.

The modular design also enables quick substitutions or extensions. For instance, replacing an attention control module or inserting a LoRA loader can be accomplished without restructuring the entire pipeline. Saved workflows can be reused and adapted to various try-on tasks, such as different clothing categories or pose settings. ComfyUI also facilitates rapid experimentation by allowing real-time adjustments of key parameters like sampling steps, noise levels, or input images. Additionally, its open ecosystem—with numerous community-contributed plugins for advanced masking, stylization, and conditioning—ensures ongoing support for evolving generative applications. While ChatClothes currently uses ComfyUI in a semi-automated fashion, future integration with real-time control systems may further enhance its responsiveness and interactivity.

2.6.4 Challenges and Considerations in Practical Use

While ComfyUI brings significant benefits, it also presents certain challenges, particularly when scaled to production-grade systems. As workflows become more intricate, especially when integrating nested control conditions or dynamic branching, the graph visualization can become cluttered, complicating usability. Best practices such as node grouping and careful labeling are necessary to maintain workflow clarity. Additionally, executing high-resolution diffusion models, such as OOTDiffusion or StableVITON, demands substantial GPU resources, limiting real-time interactions on edge devices without additional model compression or asynchronous processing strategies. Another limitation lies in the lack of native logic control; while node graphs are expressive for data flow, implementing conditional operations, such as applying different masks based on garment type, still requires external scripting or orchestration via frameworks like Dify. Moreover, integrating batch inference or RESTful API services into ComfyUI workflows demands custom extensions beyond its native GUI capabilities.

To address these practical limitations, the ChatClothes system strategically positions ComfyUI as the backend rendering engine, while logical orchestration, user interaction, and session management are handled by upstream modules such as Dify and DeepSeek. This architectural separation ensures scalable, manageable deployment without overloading ComfyUI ' s core functionality.

2.6.5 Summary

In conclusion, ComfyUI serves as a critical component in the ChatClothes virtual try-on pipeline, offering a flexible, transparent, and extensible platform for implementing diffusion-based image generation workflows. By abstracting complex diffusion processes into visually manipulable nodes, it empowers system designers to construct high-fidelity garment synthesis pipelines with fine-grained control and rapid experimentation capabilities. Despite certain scalability challenges, its strong community support, modular architecture, and visual programming paradigm make ComfyUI an ideal foundation for developing next-generation interactive visual AI systems, particularly in the domain of fashion technology.

2.7 Ollama and Vision-Language Models

With the advancement of multimodal AI systems, vision-language models(VLMs) have become a crucial bridge connecting human instructions and visual understanding. These models enable systems to interpret natural language commands while reasoning over visual content, unlocking new possibilities for interactive applications such as virtual try-on(VTON). In the context of this research, several state-of-the-art VLMs are integrated within a containerized runtime environment—Ollama—which facilitates local deployment, modular integration, and resource control. This section introduces the role of Ollama in supporting VLM execution and explores the contributions of the key integrated models in enhancing garment perception and interaction capabilities.

2.7.1 Ollama as a Local Runtime Framework

Ollama serves as a lightweight container runtime designed for local execution of large-scale foundation models, providing an efficient alternative to cloud-based model hosting. Unlike inference services such as OpenAI’ s APIs or HuggingFace’ s endpoints, Ollama offers low-latency, offline model execution that preserves user privacy and allows for flexible customization. It supports the deployment of both open-source language models and vision-language models equipped with visual encoder heads, enabling robust multimodal reasoning without reliance on external services.

Architecturally, Ollama leverages containerization and hardware acceleration to load large

models into memory while maintaining session consistency. This container-based structure improves resource allocation, reproducibility, and dynamic model switching, allowing the system to respond adaptively to various input types. Within the ChatClothes system, Ollama functions as a multimodal model gateway, encapsulating VLMs such as LLaMA3-Vision and Qwen2.5-VL, and exposing standardized interfaces for embedding generation, visual question answering(VQA), and fashion-specific image-text alignment. By reducing communication overhead and supporting concurrent model execution, Ollama enhances overall system efficiency. Furthermore, its local execution ensures that user-uploaded garment images are processed securely without transmission over external networks, an essential feature for privacy-sensitive applications.

2.7.2 Integrated Vision-Language Models

Multiple vision-language models are integrated within Ollama to support complementary tasks of image interpretation, garment attribute parsing, and prompt alignment. LLaMA3-Vision extends the LLaMA3 architecture by incorporating a transformer-based visual encoder alongside a high-capacity language decoder. Pre-trained on large-scale multimodal datasets and instruction-tuned, LLaMA3-Vision enables analysis of user-submitted fashion photos, detection of hierarchical garment components, and generation of natural language descriptions. It plays a crucial role in semantic grounding, transforming instructions into visual anchors for conditioned generation.

Complementing this, Qwen2.5-VL, developed by Alibaba DAMO Academy, excels in cross-domain visual reasoning and multilingual dialogue. It supports vision-grounded question answering, comparative garment analysis, and contextual style classification, offering strong logic-driven interpretation beyond simple object detection. Its multilingual capability extends the system' s applicability to non-English-speaking users, broadening its global reach. In addition, MiniGPT-4 is employed as a visual alignment evaluator, tasked with verifying whether output images generated by OOTDiffusion correspond accurately to the intended garment attributes or input descriptions. This validation step is vital for maintaining consistency across iterative user interactions.

Recent research also offers promising integration paths. Nexus-Gen introduces a unified architecture for image understanding, generation, and editing by combining diffusion-based generation with large language model reasoning. Its modular design supports multimodal alignment

and instruction-grounded control, making it highly adaptable to fashion workflows where interpretation and transformation must occur in a seamless, iterative loop (Zhang et al., 2025a). Similarly, ICEdit enables in-context instructional image editing by integrating generative transformers with image-conditioned prompts. It emphasizes dynamic response to localized edits—such as modifying sleeve shape or accessory placement—within the diffusion pipeline itself, reducing latency and improving user-aligned feedback (Zhang et al., 2025c).

Future integrations may include models such as BLIP-2 for image captioning and DINOv2 for garment object detection, further enhancing the system ' s ability to understand and retrieve fashion semantics.

2.7.3 System-Level Integration and Functional Scope

Within the ChatClothes architecture, Ollama manages inference requests to multiple VLMs, enabling critical functions such as visual attribute extraction from user-submitted images, prompt reformulation for garment-specific instruction refinement, visual grounding through VQA, and consistency evaluation of generated outputs. The system treats VLMs as modular services that can be queried asynchronously and composed dynamically. For instance, when a user requests "try this outfit" the instruction is parsed by DeepSeek, semantically enriched by Qwen2.5-VL, visually interpreted by LLaMA3-Vision, and then routed downstream to ComfyUI and OOT-Diffusion for synthesis. This integration establishes a closed feedback loop that mirrors expert fashion consultation workflows.

2.7.4 Challenges and Limitations

Although integrating vision-language models via the Ollama framework brings semantic flexibility to the ChatClothes system, it also presents several practical challenges. General-purpose models like DeepSeek-VL and Qwen2.5-VL are not tailored to fashion-specific contexts. As a result, they may misinterpret nuanced garment-related terms, especially when users refer to specific styles, patterns, or culturally specific expressions. Moreover, their responses can vary depending on phrasing, leading to inconsistencies when mapping natural language instructions to structured system commands. Since the models do not retain conversational history, follow-up instructions are treated independently, limiting the ability to adapt to evolving user intent.

Another limitation lies in deployment efficiency and system integration. While Ollama allows local execution, the models still demand significant computational resources, making

real-time interaction feasible only on GPU-equipped machines. Furthermore, the parsed instructions from the vision-language model are not directly linked to the image generation engine(ComfyUI), requiring manual alignment to ensure consistency between user intent and visual output. These constraints highlight the need for robust prompt design and lightweight alternatives to ensure reliable performance in practical virtual try-on applications.

2.7.5 Future Enhancements and Research Opportunities

As vision-language models continue to evolve, there is significant potential to refine their role within virtual try-on systems. One promising direction is domain-specific adaptation. Fine-tuning models like DeepSeek-VL or Qwen2.5-VL on fashion-oriented datasets may enhance their ability to interpret nuanced garment attributes and user preferences more accurately. This could lead to more precise prompt generation and smoother coordination with downstream modules such as garment classification or image synthesis. Additionally, parameter-efficient tuning methods like LoRA offer a practical path to optimizing these models for deployment on local or resource-constrained environments.

Looking forward, richer interaction and reasoning capabilities could also be explored. Integrating chain-of-thought prompting or lightweight commonsense modules might enable vision-language agents to better support multi-step decisions—such as generating full outfits based on context or reasoning through conflicting style instructions. Combining LLMs with symbolic representations like fashion knowledge graphs may further improve system transparency and expand dialogue diversity. These directions highlight opportunities to build more adaptive, efficient, and human-aligned virtual try-on experiences in future work.

2.7.6 Summary

The integration of Ollama and vision-language models significantly enriches the semantic depth and responsiveness of the ChatClothes virtual try-on system. By supporting localized garment interpretation, adaptive prompt refinement, and multimodal conversation, these models form a vital semantic bridge between user intent and generative synthesis. Although challenges remain in domain adaptation, visual stability, and resource efficiency, continued advancements in specialized fine-tuning and multimodal optimization offer a clear path toward broader deployment and enhanced user experience in fashion AI applications.

2.8 YOLO and Lightweight Vision Models in Fashion AI

Object detection and classification are foundational tasks in computer vision, forming the basis of systems that require semantic understanding of visual inputs. The YOLO(You Only Look Once) family of models, known for their high speed and accuracy, has played a pivotal role in both academic research and industry applications. In the context of fashion AI, lightweight versions of YOLO have gained prominence due to their real-time inference capabilities, compact architectures, and adaptability to resource-constrained environments such as mobile devices and embedded platforms.

2.8.1 Historical Development of YOLO Architectures

The development of YOLO models began with the groundbreaking work by Redmon, introducing a novel single-shot detection paradigm that reformulated object detection as a direct regression problem. Unlike traditional methods based on region proposals or sliding windows, YOLO predicted bounding boxes and class probabilities from the entire image in a single forward pass through a convolutional neural network, enabling real-time performance.

Subsequent versions brought significant architectural improvements. YOLOv2 introduced batch normalization, anchor boxes, and multi-scale training, improving both localization accuracy and generalization across varying input sizes. YOLOv3 expanded on this by adopting a deeper backbone network, Darknet-53, and implementing multi-scale feature predictions across three spatial resolutions to better detect small and large objects simultaneously. YOLOv4 further refined the framework by integrating several optimizations, including Mosaic data augmentation, the Mish activation function, and spatial pyramid pooling(SPP). These enhancements contributed to state-of-the-art detection accuracy while maintaining high inference speed.

The release of YOLOv5 by Ultralytics, although independent from the original authors, marked a major shift toward a PyTorch-based, open-source ecosystem. YOLOv5 emphasized usability, modularity, and deployment flexibility, rapidly gaining widespread adoption across research and industrial communities.

2.8.2 Recent Generations and Specialization

The subsequent generations—YOLOv6, YOLOv7, and YOLOv8—focused on modernizing the detection pipeline and expanding application scope. YOLOv6 emphasized industrial deploy-

ment, introducing compatibility with ONNX and TensorRT for post-training quantization and accelerated inference. YOLOv7 unified detection and segmentation tasks under a single architecture, incorporating re-parameterization techniques and efficient scaling strategies. YOLOv8 introduced anchor-free detection heads, a modular decoupled head-body structure, and novel loss functions that improved robustness on dense detection and segmentation tasks. These developments collectively expanded YOLO ' s capabilities beyond object detection into segmentation, classification, and keypoint estimation, supporting applications across autonomous driving, medical imaging, surveillance, and virtual try-on.

2.8.3 YOLO for Lightweight and Edge Deployment

One of YOLO's enduring strengths is its suitability for real-time inference on constrained devices, leading to the emergence of nano and tiny model variants. YOLOv5n and YOLOv5s were specifically designed for mobile applications, achieving a balance between low parameter count(typically below 2 million) and competitive detection accuracy. YOLOv6n and YOLOv7-tiny further reduced FLOPs through architectural innovations like EfficientRep blocks and E-ELAN modules. More recently, YOLO-NAS (Raja et al., 2024) explored neural architecture search(NAS) techniques to automatically optimize lightweight detection models across CPUs, GPUs, and mobile platforms. These lightweight versions have made YOLO practical for deployment on devices such as Raspberry Pi, Jetson Nano, and Android smartphones, enabling real-time garment detection, smart mirrors, and augmented reality(AR)-based shopping assistants in fashion AI(Babuc and Fortiş, 2024).

2.8.4 YOLO Model Family Comparison

Table 2.2 provides a comparison of selected YOLO variants, highlighting their computational requirements and design features.

Table 2.2: Comparison of selected YOLO variants

Model	Year	FLOPs(G)	Params(M)	Highlights
YOLOv3	2018	65.9	62	Multi-scale, Darknet-53 backbone
YOLOv4	2020	90.1	64	CSPDarknet53, SPP, Mish activation
YOLOv5s	2021	17.0	7.2	Modular PyTorch framework
YOLOv6n	2022	4.2	3.2	ONNX/TensorRT-friendly deployment
YOLOv7	2022	105.5	37	E-ELAN, Re-parameterization
YOLOv8n	2023	6.2	3.5	Anchor-free, decoupled head
YOLO-NAS-S	2023	5.4	4.3	NAS-optimized, transformer-aware
YOLO11n	2024	5.9	3.3	Lightweight VTON-oriented variant
YOLO12n	2024	5.7	3.0	Enhanced semantic head

2.8.5 Key Technologies in YOLO Development

The evolution of YOLO architectures has been driven by several core innovations, summarized in Table 2.3.

Table 2.3: Core technologies used in YOLO variants

Technology	Description and Contribution
CSPNet	Cross Stage Partial connections improve inference speed and gradient flow while reducing complexity.
E-ELAN	Efficient Layer Aggregation Networks enhance network depth and diversity without overfitting.
Anchor-Free Head	Eliminates the need for anchor boxes, improving localization adaptability.
Re-parameterization	Merges multiple convolution layers into a single representation for efficient deployment.
Neural Architecture Search(NAS)	Automates the design of optimal lightweight models across diverse hardware platforms.

2.8.6 Applications of YOLO in Fashion AI

In fashion-related applications, YOLO has been widely adopted for tasks such as detecting and labeling garments including tops, trousers, dresses, and accessories in street photography; garment segmentation and category tagging in virtual fitting rooms; and attribute prediction based on bounding box analysis, covering features like color, sleeve type, and neckline design. The low latency and strong generalization capabilities of YOLO models make them particularly suitable for processing the non-iconic, diverse imagery typically encountered in user-generated content.

2.8.7 Comparison with Vision Transformers

While YOLO models dominate real-time detection tasks, vision transformers(ViTs) have also gained attention for their global attention capabilities. Table 2.4 contrasts the two approaches across multiple dimensions.

Table 2.4: Comparison: YOLO models vs. Vision transformers

Aspect	**YOLO Family**	**Vision Transformers(ViT)**
Architecture	CNN-based, local feature extraction	Transformer-based, global attention mechanisms
Inference Time	Real-time(30–60 FPS)	Slower(5–10 FPS)
Training Data Needs	Low to medium	Very high(e.g., JFT-300M, ImageNet-22k)
Edge Suitability	Excellent(nano, tiny variants)	Poor unless heavily pruned
Interpretability	High(object-level)	Moderate(through feature attention visualization)
Typical Use Cases	Detection, classification, segmentation	Zero-shot captioning, retrieval, multimodal tasks

2.8.8 Summary and Future Outlook

YOLO continues to serve as a foundational model for efficient visual understanding. Its lightweight variants remain critical for supporting real-time garment classification, detection, and segmentation tasks in interactive try-on workflows. Future research directions suggest that upcom-

ing YOLO iterations may incorporate attention mechanisms, transformer hybrid modules, and hardware-specific adaptations to further enhance performance, particularly in multimodal and wearable AI applications. In subsequent chapters, this thesis introduces a specialized lightweight YOLO-based classification model designed for garment recognition, integrated within the ChatClothes architecture to exemplify the synergy between efficient visual perception and generative synthesis.

2.9 Summary

This chapter has provided a comprehensive and in-depth review of the foundational technologies, historical developments, and cutting-edge advancements that form the technical bedrock of modern virtual try-on(VTON) systems. Emphasis was placed on the core components and methodologies that underpin the proposed ChatClothes framework. Through a systematic exploration of visual synthesis models, language-driven interaction systems, orchestration tools, and lightweight classification modules, this chapter has laid the conceptual and technical foundation necessary for implementing an intelligent multimodal try-on experience.

The discussion began with a chronological analysis of VTON evolution, tracing the transition from early garment overlay techniques to GAN-based pipelines, and ultimately to high-fidelity diffusion models. Early efforts such as VITON, CP-VTON, and ClothFlow introduced garment transfer through 2D warping and TPS alignment, setting the stage for more sophisticated frameworks like HR-VTON and StyleGAN-VTON that employed adversarial training and occlusion-aware refinement strategies. The recent emergence of diffusion-based models, notably OOTDiffusion (Xu et al., 2024), IDM-VTON (Choi et al., 2024), and TryOnDiffusion (Zhu et al., 2023), has marked a shift toward modular, controllable synthesis pipelines capable of generating photorealistic outputs with improved semantic alignment and garment fidelity.

Supporting this generation process, ComfyUI was introduced as a node-based visual programming framework for designing and managing latent diffusion workflows. ComfyUI ' s modular, low-code environment enables real-time manipulation of sampling strategies, conditioning inputs, and generation parameters, thus facilitating rapid experimentation and multimodal integration. Within the ChatClothes system, ComfyUI not only serves as a backend rendering engine but also acts as a central integration point connecting upstream modules for prompt generation and downstream modules for image evaluation.

Language interaction was identified as a crucial enabler for achieving personalized and intuitive user control in VTON systems. Large language models(LLMs) such as DeepSeek, ChatGPT, and LLaMA3 were explored for their capabilities in instruction-following, conversational memory, and semantic translation. These models allow natural language inputs to be transformed into structured control signals, enabling seamless coordination between user commands and visual generation. Integrated through the Dify orchestration platform, LLMs manage multi-turn dialogue, contextual prompt generation, and system-level routing, significantly enhancing the richness and adaptability of user interactions.

In addressing multimodal vision-language grounding, the chapter examined models such as LLaMA3-Vision and Qwen2.5-VL. These VLMs demonstrated strong capabilities in garment attribute extraction, spatial reasoning, and vision-question answering, thereby enriching the responsiveness and interpretability of the system. Their deployment within Ollama ensures local execution, enhancing privacy protection and lowering latency without reliance on cloud-based services—a critical feature for privacy-sensitive fashion AI applications.

On the perception side, the integration of the YOLO model family—particularly lightweight variants like YOLOv5n, YOLOv8n, and YOLOv12—was discussed in depth. These models enable high-speed garment classification and detection, particularly important for mobile and edge deployments. The evolution of YOLO architectures, including innovations such as CSPNet, E-ELAN modules, anchor-free heads, and NAS-optimized designs (Raja et al., 2024), was highlighted to demonstrate their adaptability to fashion AI tasks requiring efficient, real-time visual understanding.

Beyond the technical contributions, this chapter also addressed critical ethical, environmental, and practical challenges associated with deploying VTON systems. Bias and representation issues remain prevalent, given that widely used datasets like VITON-HD and DeepFashion are biased toward specific demographics, posing inclusivity challenges (Wu et al., 2022). Privacy and security concerns were also emphasized, necessitating the use of secure, containerized environments like Dockerized Ollama for local inference to safeguard sensitive user images (Li and Zhang, 2023). Additionally, the environmental impact of large-scale models was acknowledged, and potential mitigation strategies such as model pruning (Han et al., 2016), quantization (Jacob et al., 2018), LoRA fine-tuning (Hu et al., 2021), and model distillation (Hinton et al., 2015) were suggested to improve energy efficiency. Finally, usability challenges were considered, highlighting the importance of robust error handling, transparent feedback loops, and continual

refinement of prompt control mechanisms to maintain alignment between user expectations and generated outputs (Park and Park, 2022).

Through its modular and decoupled design, the ChatClothes system aims to address these concerns by allowing targeted improvements and independent optimization across its components. ComfyUI workflows can be dynamically reconfigured; Dify orchestration schemas can incorporate evolving safety standards and interaction protocols; and YOLO classifiers can be incrementally updated with expanded garment taxonomies, ensuring long-term scalability and adaptability without disrupting the overall system architecture.

In conclusion, this chapter has established the interdisciplinary and modular foundation upon which the ChatClothes system is built. By synthesizing lightweight vision models, diffusion-based visual generation, advanced language models, and orchestrated multimodal workflows, the system aspires to deliver a responsive, flexible, and ethically grounded solution for next-generation virtual try-on applications. The following chapter presents the detailed design and implementation of this architecture, illustrating how these technologies are woven together into an operational prototype capable of supporting real-world fashion interaction scenarios.

Chapter 3
Methodology

This chapter outlines the methodological framework behind the ChatClothes System, focusing on its modular architecture, core models, and implementation strategy.
It introduces the overall system design, followed by a description of the diffusion-based try-on module and the YOLO12n-LC clothing classifier. Key components such as the latent diffusion mechanism, text-image conditioning, lightweight model customization, and fine-tuning strategies are presented in detail.
Although the dataset and preprocessing techniques form the backbone of the training pipeline, they are discussed separately in Chapter 4 as part of the experimental setup and evaluation procedures.

3.1 Introduction

This chapter introduces the methodology of the ChatClothes System, a modular virtual try-on framework that integrates large language models(LLMs), lightweight clothing classification, and diffusion-based image generation. The goal is to enable controllable, multimodal interactions that support real-time and personalized outfit visualization.

At the core of the system is a scalable architecture built on containerized microservices. Dify functions as the orchestration layer, managing input routing, task scheduling, and module communication. User inputs—either textual descriptions or image uploads—are processed through this interface and routed to appropriate subsystems for reasoning and generation.

The virtual try-on experience is enabled by two primary models:

OOTDiffusion —a latent diffusion-based image generator responsible for synthesizing high-quality try-on outputs guided by both visual and semantic cues.

YOLO12n-LC —a lightweight garment classifier that categorizes input clothing and provides semantic labels to support evaluation and prompt construction.

Together with supporting modules such as Ollama(for prompt parsing) and ComfyUI(for visual workflow orchestration), the system achieves end-to-end controllability, responsiveness, and deployment adaptability across platforms.

The rest of this chapter is organized as follows:

Section 3.2 describes the modular architecture and how each component interacts in the system workflow.

Section 3.3 introduces the OOTDiffusion generator, including its architecture, diffusion process, and semantic control mechanism.

Section 3.4 presents the design and optimization strategy of the YOLO12n-LC classifier.

Section 3.5 discusses implementation details such as system deployment strategies and integration techniques.

Section 3.6 outlines the evaluation strategy, including metric design and validation methodology.

Section 3.7 summarizes the methodology and transitions into the experimental results in Chapter 4.

By breaking down the system into well-defined, interchangeable modules, this chapter provides a solid foundation for understanding the technical rationale, algorithmic choices, and en-

gineering trade-offs behind the ChatClothes System.

3.2 System Design Overview

The architecture of the proposed ChatClothes System is designed to support a flexible, modular, and extensible virtual try-on pipeline that bridges natural language understanding with high-fidelity image synthesis. It adopts a service-oriented and multi-module design philosophy, which facilitates real-time responsiveness, scalable deployment, and cross-modal interaction. The system integrates large language models(LLMs), lightweight garment classification, and latent diffusion-based image generation, all orchestrated via Dify, which acts as the centralized task controller and user-facing entry point.

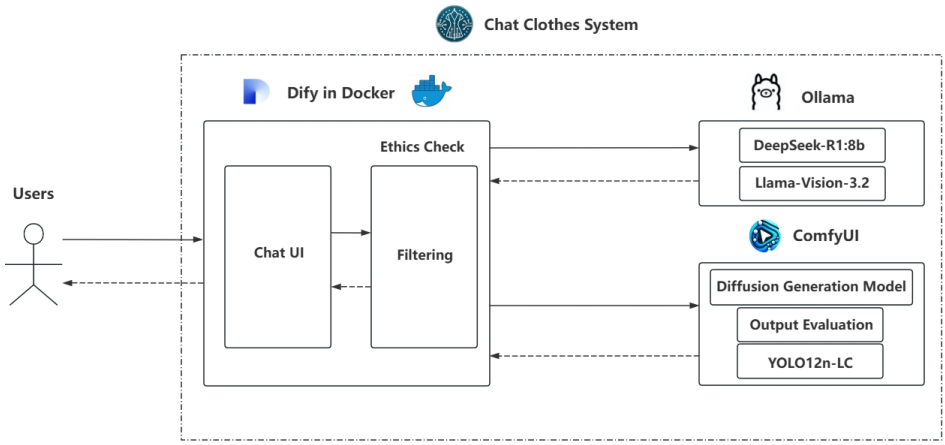


Figure 3.2: ChatClothes system architecture

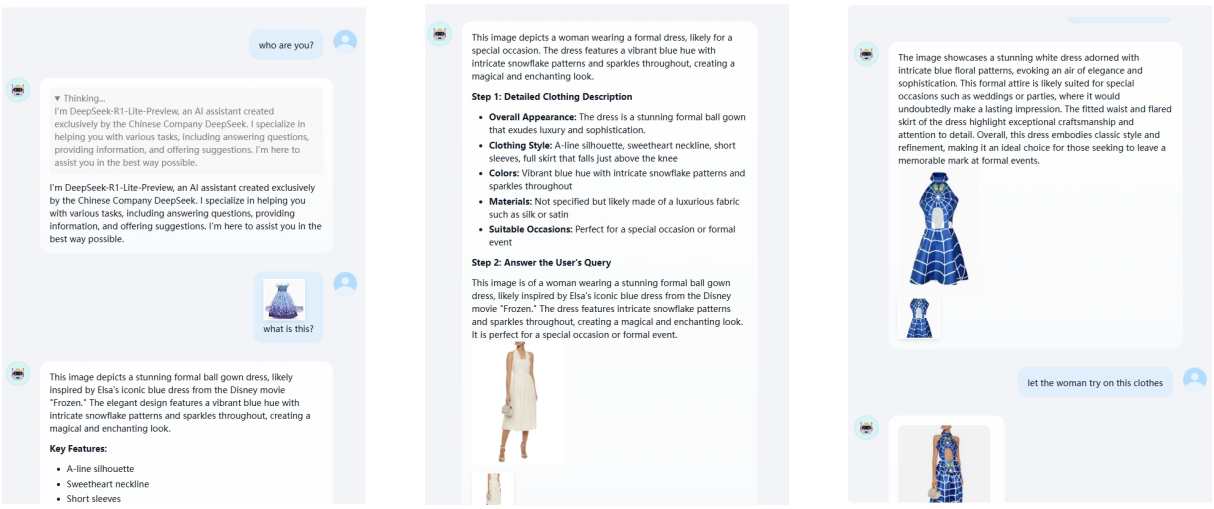


Figure 3.3: ChatClothes system UI

As illustrated in Figure 3.2, the ChatClothes System consists of seven core components: (1) a Dify-based interaction frontend and orchestrator; (2) an ethical filtering module; (3) a backend LLM stack hosted on Ollama; (4) ComfyUI for visual workflow and diffusion model execution; (5) a garment classification module(YOLO12n-LC); (6) an output evaluation and scoring engine; and (7) a user feedback loop mechanism.

3.2.1 Modular Architecture and Operational Workflow

The system operates as a multi-stage, modular pipeline. The typical workflow for a virtual try-on session includes:

User Input and Contextual Session Initialization: Users initiate a session by uploading a source image and optionally providing text commands such as ”try a red blazer”. The session is managed by Dify, which maintains the dialogue state and ensures continuity across multiple interactions.

Ethical and Safety Filtering: The ethics module screens inputs for unsafe, offensive, or ambiguous content. This step is critical for compliance with content safety policies, particularly in public or commercial deployments.

LLM-Powered Prompt Parsing and Semantics Extraction: Validated inputs are forwarded to the Ollama runtime, which houses DeepSeek and LLaMA3-Vision models. These models convert unstructured queries into structured control signals, extracting garment types, style preferences, pose-related modifiers, and generation constraints.

Image Classification via YOLO12n-LC: Uploaded clothing items are classified into fashion-relevant categories(e.g., coat, skirt, blouse) using the optimized YOLO12n-LC. These labels are then matched with the generation pipeline and used to validate semantic coherence between user instructions and generated outputs.

Visual Synthesis through ComfyUI: Conditioned prompts and control parameters are passed to ComfyUI, which orchestrates the latent diffusion process via OOTDiffusion. Conditioning data includes CLIP-based embeddings, body pose maps, garment masks, and optional control hints. Intermediate outputs such as segmentation masks or aligned clothing templates can also be generated as pre-processing steps.

Post-Generation Evaluation: The resulting try-on images are scored using SSIM, LPIPS, and a custom CIS metric that considers both structural fidelity and semantic alignment. Evaluation outputs are routed back through Dify, allowing users to inspect, approve, or revise the

results.

Interactive Refinement: If a user is dissatisfied, they can provide natural language feedback. The system loops back to update the prompt context using session memory and triggers a new generation round with refined controls.

This modular architecture ensures each subsystem can be independently upgraded or replaced without disrupting the end-to-end interaction experience.

3.2.2 Pipeline Flexibility and Deployment Scalability

A key strength of the ChatClothes architecture lies in its emphasis on component isolation, deployment flexibility, and compatibility with low-resource environments.

Core extensibility and deployment features include:

Containerization and Microservice Isolation: Each component(Dify, Ollama, ComfyUI, YOLO12n-LC) is deployed in a dedicated Docker container. This allows for distributed scaling, resource separation, and independent updates.

API-Level Interconnectivity: Communication between modules occurs via HTTP or gRPC APIs, following a stateless interaction model. This decoupling enables integration with cloud-based services or hardware-accelerated edge nodes.

MCP Protocol Compatibility: With MCP support, Dify enables modular communication across components via a unified protocol. This allows flexible model switching and streamlined integration of multimodal agents.

Model Modularity and Interchangeability: Any component model can be swapped with minimal configuration changes. For example, DeepSeek can be replaced with ChatGPT-4, or OOTDiffusion can be swapped for StableVITON or TryOnDiffusion without disrupting the control logic or user flow.

Multimodal and Incremental Interaction: The system supports image-only inputs, text-only prompts, or a combination of both. Incremental refinement is managed through turn-based memory and attention mechanisms embedded in the LLM agent, allowing dynamic adjustment without restarting the session.

3.2.3 Operational Advantages

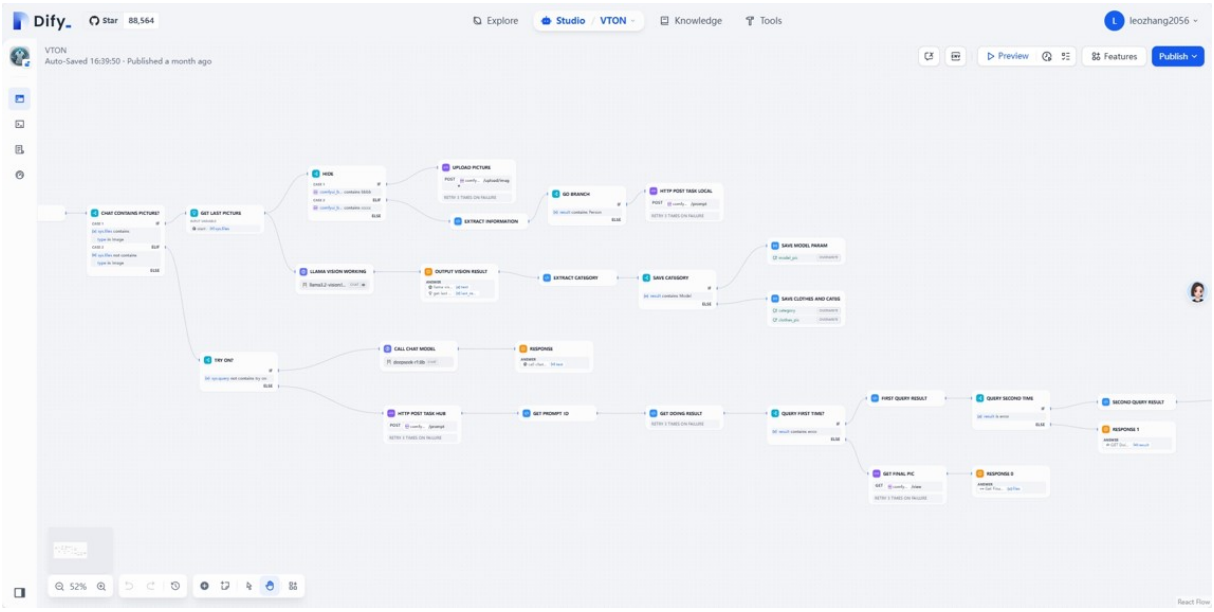


Figure 3.4: Workflow of the ChatClothes system

The architecture's flexibility offers several practical advantages:

Responsiveness and Speed: By running locally or on-premise(e.g., via Ollama or ComfyUI), the system minimizes latency and avoids reliance on external APIs, which is crucial for real-time deployment in retail or mobile environments.

Privacy Preservation: All inference is performed locally, with no data sent to third-party servers. This feature is essential for handling sensitive personal images and complying with data protection regulations such as GDPR or PIPL.

Scalability and Fault Tolerance: With decoupled architecture, system performance can be optimized by distributing inference workloads across multiple nodes. Fallback mechanisms allow for automatic degradation if a component fails(e.g., using a backup prompt generator if LLM inference fails).

Research and Prototyping Efficiency: Researchers can easily add or remove modules, modify the workflow in ComfyUI, or inject test samples into different nodes. This accelerates experimentation and supports ablation studies.

3.2.4 Summary

The system design of ChatClothes reflects a synthesis of modern AI orchestration principles, real-time generation demands, and user-centered interaction logic. Its containerized, modular, and API-driven architecture supports not only technical scalability but also ethical, legal, and operational requirements. In the next section, we delve into the methodology for training, customizing, and integrating each of the core models introduced in this chapter, beginning with the OOTDiffusion generator.

3.3 Diffusion-Based Try-On Module

3.3.1 Model Architecture

The core image synthesis module of the ChatClothes System is built upon OOTDiffusion, a latent diffusion-based framework tailored to generate high-fidelity virtual try-on images. The model is capable of preserving garment textures, styles, and body structure, ensuring realism and coherence in the generated output.

OOTDiffusion consists of the following key components:

Latent-Space U-Net: Serves as the backbone of the denoising process, iteratively refining noisy latent representations to produce clean outputs.

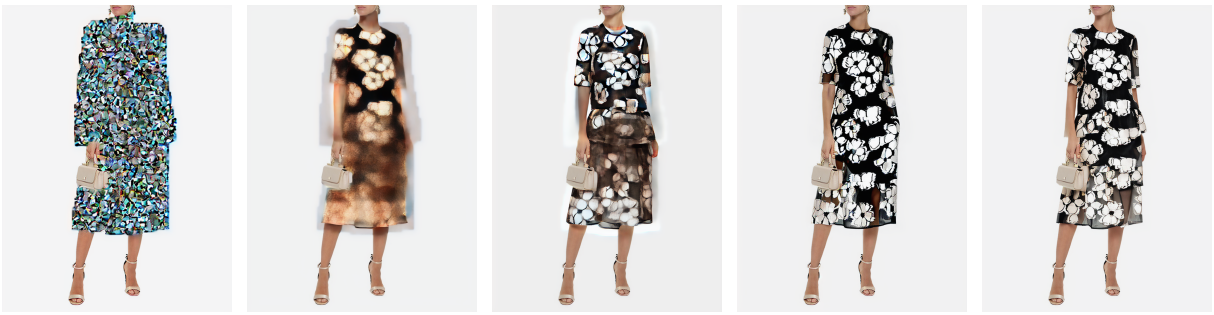
Cross-Attention Mechanism: Fuses semantic information from language embeddings with visual features, enabling the system to condition generation on user prompts.

Variational Autoencoder(VAE): Compresses high-resolution images into a compact latent space to reduce computational cost and decodes the final output back into image space.

This architecture supports both visual and linguistic conditioning, allowing the generation of semantically meaningful and visually realistic try-on results.

3.3.2 Diffusion Process

The diffusion-based image generation process follows a standard forward-reverse paradigm. In the forward process, Gaussian noise is progressively added to the latent representation. In the reverse process, the model learns to denoise the image through a sequence of steps.



(a) A (b) B (c) C (d) D (e) E

Figure 3.5: Progressive refinement of virtual try-on images across the diffusion generation process. From(a) to(e), the output quality improves as noise is gradually removed and garment details are refined.

To illustrate the generative evolution within our diffusion-based virtual try-on pipeline, Figure 3.5 presents a visual comparison of five sequential outputs from the model. These samples represent intermediate stages of the denoising process, corresponding to different inference timesteps or refinement modules in the pipeline.

Initially, as shown in(a), the image appears noisy and structurally ambiguous, with little alignment between the clothing and the body. In(b) and(c), the structure of the garment becomes more coherent, with basic silhouettes emerging and texture details beginning to form. By stage(d), high-frequency details such as floral patterns and edges are restored with greater fidelity, while the garment contour closely matches the body layout. Finally, image(e) demonstrates the fully converged result, which preserves fabric texture, color consistency, and garment realism.

This progressive enhancement highlights the capacity of our OOTDiffusion pipeline to refine both semantic and visual information over successive steps, leading to high-quality virtual try-on results. It also supports the feasibility of early-exit or fast-sampling strategies(e.g., Latent Consistency Models) to accelerate inference while maintaining acceptable visual quality.

Forward Process:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} \cdot x_{t-1}, \beta_t I) \tag{3.1}$$

Reverse Process:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3.2}$$

Training Objective: The simplified denoising loss is used to train the model:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right] \tag{3.3}$$

This loss measures the mean squared error between predicted and actual noise, guiding the model to reverse the diffusion process effectively.

3.3.3 Text-Guided Control and Cross-Attention

To enable natural language control, OOTDiffusion integrates cross-modal conditioning from a large language model(e.g., DeepSeek). The following steps are used to inject textual semantics into the generation pipeline:

Semantic Parsing: Garment attributes such as color, category, and style are extracted from user prompts.

Text Embedding: Parsed information is transformed into vector embeddings using a CLIP-based or similar encoder.

Cross-Attention Fusion: The embeddings are injected into U-Net decoder layers via cross-attention blocks, where text acts as the query and image features act as keys/values.

This fusion mechanism ensures semantic consistency between the textual input and the generated try-on image, enabling user-controllable generation through natural language.

3.3.4 Fine-Tuning and Optimization Strategy

To improve OOTDiffusion’ s controllability and domain alignment in the fashion try-on task, we adopt a lightweight, parameter-efficient fine-tuning strategy. Given the system’ s constraints on computation and memory, we evaluated several mainstream adaptation techniques, including full fine-tuning, adapter injection, prompt tuning, and LoRA. Based on comparative analysis, we selected LoRA(Low-Rank Adaptation) as the core method due to its minimal resource overhead and superior efficiency for diffusion-based image generation tasks.

Comparison of Fine-Tuning Methods

Full Parameter Fine-Tuning: Offers complete flexibility and often yields high-quality results but requires enormous GPU memory and long training times. Unsuitable for scalable deployment or low-end hardware.

Adapter-Based Tuning: Introduces small learnable modules into transformer layers. Though

parameter-efficient, it adds architectural complexity and has limited tooling support in visual diffusion pipelines.

Prompt Tuning: Optimizes a fixed set of soft prompt tokens. Effective for textual control in NLP, but insufficient for high-dimensional generation control in VTON tasks.

LoRA: Inserts low-rank matrices into existing attention and convolutional projections. With only a small fraction of trainable parameters, LoRA achieves strong controllability while maintaining inference efficiency and model stability.

LoRA Integration in OOTDiffusion

We integrate LoRA into selected attention and convolution layers in the U-Net backbone of the latent diffusion model. Specifically, given a base weight matrix $W \in \mathbb{R}^{d\times k}$, LoRA introduces low-rank matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$, updating W during fine-tuning as:

$$W' = W + \alpha \cdot AB \tag{3.4}$$

where α is a scaling factor and r is a rank hyperparameter(commonly set to 4 or 8). Only A and B are updated during training, while W remains frozen. This dramatically reduces memory consumption and computation, making it highly suitable for edge-friendly fashion AI applications.

Latent Diffusion Training Background

OOTDiffusion builds on the latent diffusion model(LDM) paradigm, which operates entirely in the latent space for improved efficiency. The process consists of:

- **VAE Encoding:** The input image x is encoded into its latent representation z via a pretrained VAE.
- **Denoising U-Net:** The model learns to remove Gaussian noise from the latent variable z_t at timestep t.
- **Text Conditioning:** Textual prompts are encoded using a CLIP text encoder to obtain the conditioning vector $p = \tau_\theta(y_{\text{text}})$, where y_{text} is the input prompt.

The simplified latent diffusion training objective is therefore:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{z,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, p)\|^2\right] \tag{3.5}$$

Classifier-Free Guidance in Inference

To enhance controllability during inference, classifier-free guidance(CFG) is applied in the latent space. The guided noise prediction is computed as:

$$\epsilon_{\text{guided}} = \epsilon_\theta(z_t, t, p) + w \cdot (\epsilon_\theta(z_t, t, p) - \epsilon_\theta(z_t, t, \varnothing)) \tag{3.6}$$

This formulation improves the system' s adherence to user instructions encoded in the textual prompt.

Training Settings Summary

During LoRA fine-tuning, all pretrained parameters are frozen, and only the LoRA adapters are trained. Mixed-precision(FP16) training is applied to optimize memory usage and speed. Additionally, data augmentation techniques—such as horizontal flipping, color jittering, and cropping—are used to improve model generalization.

One key advantage of LoRA is its ability to be seamlessly integrated into different stages of the diffusion pipeline as a plug-in module. This flexibility enables targeted correction of visual artifacts, such as distorted limbs, misaligned hands, or inconsistent textures. Due to its minimal parameter footprint and modular design, LoRA allows for local perceptual refinements without significantly increasing system complexity, making it particularly effective for fine-detail control in virtual try-on tasks.

Exact hyperparameter settings and optimizer configurations(e.g., number of epochs, learning rates) are deferred to the experimental section(Chapter 4) to keep the current focus on training strategy.

Overall, this LoRA-based fine-tuning strategy offers a scalable, lightweight, and high-performance solution for adapting diffusion models to controllable virtual try-on applications, especially in real-time or on-device scenarios.

3.4 Lightweight Classification Module

3.4.1 YOLO12n Architecture and Design

To enable efficient garment classification within the ChatClothes System, we adopt a customized lightweight variant of the YOLOv12 model, referred to as YOLO12n-LC(Lightweight for Clothes). This model is specifically optimized for real-time inference on resource-constrained platforms

such as Raspberry Pi, mobile devices, and wearable AR systems.

YOLO12n-LC is derived from the YOLOv12-nano backbone and includes several key modifications:

Backbone Compression: The number of convolutional blocks in the C3 modules is reduced to minimize model size while maintaining feature extraction capacity.

Neck Simplification: PANet layers are streamlined to retain only essential feature paths, supporting efficient multi-scale processing.

Classification Head: The object detection head is replaced with a multi-class classification head outputting five fashion categories: Tops, Bottoms, Outerwear, Shoes, and Accessories.

Activation Function Tuning: The ReLU activation is replaced with SiLU, improving nonlinear expressiveness while preserving computational efficiency.

These design choices allow YOLO12n-LC to achieve a balance between classification accuracy and runtime performance, making it well-suited for real-time virtual try-on pipelines.

3.4.2 Training Objective and Optimization

The lightweight classification module, YOLO12n-LC, is optimized for real-time garment recognition in resource-constrained environments such as mobile platforms or Raspberry Pi. The goal is to ensure semantic consistency and category awareness throughout the try-on pipeline without sacrificing speed or accuracy.

(1) Classification Objective

YOLO12n-LC formulates garment classification as a multi-class prediction task. The model is trained using the categorical cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{C} y_i \log(p_i) \tag{3.7}$$

where C is the number of garment categories, y_i is the one-hot encoded ground-truth label, and p_i is the predicted softmax probability for class i.

(2) Structural Optimization for Lightweight Deployment

YOLO12n-LC is derived from the YOLOv12-nano backbone, but significantly modified for fashion-specific classification:

Backbone Compression: The number of convolutional layers in the C3 modules is reduced to minimize floating point operations(FLOPs) while maintaining feature integrity.

Neck Simplification: PANet layers are trimmed to preserve only essential feature paths for multi-scale processing.

Classification Head Replacement: The detection head is replaced by a custom multi-class classification head for five unified garment classes: tops, bottoms, outerwear, shoes, and accessories.

Activation Tuning: SiLU is used instead of ReLU to improve gradient flow with minimal impact on latency.

Data Preprocessing and Augmentation To improve model generalization and overall performance, a series of data preprocessing and augmentation techniques were applied. Prior to model input, all images were normalized by scaling pixel values to the range [0,1], ensuring consistent data distribution. This normalization process promotes faster model convergence and minimizes issues arising from heterogeneous data ranges.To further enhance generalization, particularly for underrepresented classes, three augmentation strategies were adopted. Random horizontal flipping was used to simulate variations in capture angles, enabling the model to maintain stable classification performance under diverse viewpoints. Random rotation augmentation introduced variability in garment poses, enhancing robustness to pose differences. Additionally, random scaling simulated size variations, allowing the model to better generalize across clothing items of different proportions.These augmentation techniques not only enriched the diversity of the training data but also played a crucial role in alleviating overfitting. By introducing continuous variability during training, the model was exposed to a wider range of patterns, thereby improving its capacity for generalization.A notable challenge in this work was addressing class imbalance, particularly for categories with fewer samples such as Accessories and Outerwear. To mitigate this issue, an oversampling strategy was employed, replicating instances from minority classes to ensure balanced representation within the training set. In conjunction with oversampling, data augmentation further expanded the minority class samples by introducing transformations in angle, size, and orientation. Finally, to enhance learning under imbalanced conditions, class weight adjustment was incorporated into the loss function. Higher weights were assigned to minority classes, guiding the model to better recognize and classify these underrepresented categories and reducing the adverse effects of imbalance on overall performance.

Pruning and Quantization Techniques for Model Optimization Pruning and quantization are two essential techniques for optimizing deep learning models for deployment on resource-constrained platforms, such as edge devices and mobile terminals. Pruning aims to

reduce model size and inference latency by removing redundant components—such as unimportant weights, filters, or entire channels—from the network. This not only decreases the number of parameters but also simplifies the computational graph, accelerating inference without severely compromising accuracy. Quantization, by contrast, compresses the model by converting high-precision floating-point operations(e.g., FP32) into lower-precision formats(e.g., INT8), significantly reducing memory bandwidth requirements and improving execution efficiency on hardware accelerators that support integer operations. These techniques provide a clear direction for further enhancing the deployment feasibility of YOLO12n. Although YOLO12n already exhibits a strong balance between accuracy and efficiency, its architecture—originally designed for object detection—still includes components that are unnecessary for single-label classification tasks. Our optimized variant, YOLO12n-LC, simplifies the original structure by removing the detection head and neck, retaining only the lightweight backbone and introducing a classification head. While this modification already reduces computational complexity, applying structured pruning(e.g., channel or layer-wise pruning) can further eliminate redundant computations, particularly in early convolutional stages. Moreover, integrating post-training quantization or quantization-aware training(QAT) would allow YOLO12n-LC to operate with INT8 precision, thereby lowering memory consumption and improving inference speed on real-time hardware such as ARM-based processors or NPUs.These techniques are particularly valuable for applications requiring fast, low-power classification, such as intelligent manufacturing, wearable devices, or mobile recommendation systems. In future work, we plan to explore hardware-aware pruning strategies and integer quantization pipelines to further compress the YOLO12n-LC model. The goal is to deliver a high-performance, low-latency classification model that retains accuracy while being highly suitable for deployment in real-world, resource-limited environments.

These changes significantly reduce model size and inference cost, making the model suitable for deployment in low-latency environments.

(3) Training Strategy and Data Augmentation

To enhance the generalization ability of YOLO12n-LC, we use a merged dataset comprising DeepFashion, DressCode, and Kaggle fashion subsets. All labels are normalized into five unified categories.

Key training strategies include:

- Data Augmentation:

Horizontal flipping and brightness jittering;

Affine transformations to simulate different poses and camera angles;

Background blending to increase robustness to visual context.

- Training Configuration:

  Optimizer: AdamW with initial learning rate of 1×10^{-3};

  Scheduler: Cosine annealing for smooth convergence;

  Early stopping: Training halts if the F1 score stagnates over 10 validation epochs.

When quantized, the trained model achieves inference latency below 10 ms per image on devices like NVIDIA RTX 3060 and Raspberry Pi 5.

3.4.3 Role in System Workflow

YOLO12n-LC serves as a key structural module within the ChatClothes System. While the OOTDiffusion module is responsible for generating high-fidelity try-on images, the classifier ensures that these generations are grounded in accurate garment semantics and suitable for interactive control.Its roles are multifaceted:

Input Pre-filtering:The classifier performs real-time filtering of user-uploaded images, discarding irrelevant or low-quality inputs before further processing. This ensures computational efficiency and semantic relevance.

Prompt Support for LLMs: The predicted garment category(e.g., dress, outerwear) is passed to the language model(e.g., DeepSeek) as a structured prompt enhancement. This improves prompt comprehension and allows for more accurate conditional image generation.

Modular Routing and Pipeline Selection: YOLO12n-LC enables dynamic branching in the system pipeline. For example, category-specific paths may be implemented in future extensions to handle unique fitting logic or rendering preferences(e.g., dresses may use a different warping strategy than jackets).

By working synergistically with the LLM and diffusion components, YOLO12n-LC ensures semantic alignment across modules and enables a hybrid perception-control loop. This integration helps reduce ambiguity, enhances interpretability, and supports adaptive, user-controllable virtual try-on experiences.

In summary, YOLO12n-LC is not only an efficient classification model but also a functional enabler of multimodal interaction and system scalability within the ChatClothes architecture.

3.5 System Implementation

The ChatClothes System is engineered for deployment across both cloud infrastructure and local edge devices. This section outlines the deployment strategies, runtime architecture, and model validation mechanisms that ensure the system's scalability, usability, and efficiency in practical settings.

3.5.1 Deployment Strategy

The system is designed with flexibility in mind, supporting dual deployment modes:

Cloud-Based Deployment: The full pipeline—including Dify orchestration, Ollama-based LLM inference, ComfyUI workflows for diffusion, and the YOLO12n-LC classifier—is containerized via Docker and hosted on platforms such as AWS or Google Cloud. Kubernetes is used to support distributed scaling and fault tolerance in multi-user environments.

On-Device Deployment: For edge deployment scenarios(e.g., mobile devices, Raspberry Pi, or AR glasses), quantized versions of YOLO12n-LC and lightweight variants of OOTDiffusion are deployed using ONNX Runtime or TensorRT. Techniques such as model pruning and precision reduction are applied to minimize inference latency and memory usage.

These strategies allow the system to serve real-time interactive needs in both commercial cloud services and lightweight, offline environments.

3.5.2 Module Integration and Workflow Execution

The system architecture comprises the following functional modules:

Frontend(User Interface): A web or mobile interface where users upload images or prompts, view try-on results, and refine instructions interactively.

Backend(Inference Server): Hosts OOTDiffusion, DeepSeek LLM, and the YOLO12n-LC classifier. It processes user input and generates personalized try-on images.

Garment Repository: A structured database containing garment images, categories, and metadata for reference and prompt enrichment.

The complete workflow proceeds as follows:

Users submit a prompt and/or garment image;

Dify handles input routing, filtering, and LLM task delegation via Ollama;

YOLO12n-LC classifies the uploaded garment and tags it with a category label;

ComfyUI invokes OOTDiffusion, guided by semantic and visual input, to generate a try-on image;

An internal evaluation module computes quality scores(SSIM, LPIPS) and sends the result back to the user;

The user can refine the prompt and iterate the generation process;

Inter-module communication is realized through REST APIs or shared memory pipelines, depending on deployment mode, to ensure modular independence and system extensibility.

3.5.3 Hyperparameter Settings and Validation

To ensure the reproducibility and generalizability of the system, rigorous hyperparameter tuning and validation procedures are implemented.

Training Configuration:

Optimizers: OOTDiffusion-AdamW with learning rate 1×10^{-4} and weight decay of 0.01; YOLO12n-LC-SGD or AdamW with initial learning rate 1×10^{-3}.

Schedulers: Cosine annealing is used to dynamically adjust learning rates.

Early Stopping: Training is stopped if the validation metric(e.g., SSIM or F1 score) stagnates for more than 10 epochs.

Precision Optimization: Mixed-precision training(FP16) is enabled to accelerate convergence and reduce memory usage.

Distributed Training: Multi-GPU training is employed to accelerate fine-tuning of OOTDiffusion on large datasets.

Evaluation Metrics:

The system is evaluated using a combination of perceptual, computational, and user-centric metrics:

Visual Quality: SSIM ↑, LPIPS ↓, FID ↓, KID ↓

User Interaction: Response latency, satisfaction survey score, multi-turn refinement support

Computational Performance: Inference speed, memory footprint, and system scalability under load

Qualitative Validation:

User studies are conducted to assess the perceived realism and usability of the system. Engagement logs and subjective scores are collected to support continuous refinement of prompt clarity and image quality.

This implementation strategy ensures that the ChatClothes System is technically robust, user-friendly, and deployable in a wide range of real-world scenarios.

3.6 Evaluation Strategy

This section presents the evaluation methods adopted to assess the effectiveness of the core modules in the ChatClothes System—specifically the image generation quality of OOTDiffusion and the classification performance of YOLO12n-LC. A combination of perceptual, computational, and application-specific metrics is used to provide a comprehensive assessment.

3.6.1 Image Generation Quality Evaluation

To evaluate the realism, coherence, and consistency of virtual try-on outputs generated by OOTDiffusion, we employ a combination of standardized and task-specific image similarity metrics. These metrics collectively assess structural alignment, perceptual quality, and garment preservation fidelity.

Standard Evaluation Metrics

SSIM(Structural Similarity Index): Measures structural similarity in terms of luminance, texture, and geometry. It captures whether the synthesized image retains the spatial integrity of the original person image.

LPIPS(Learned Perceptual Image Patch Similarity): Computes perceptual differences using deep feature embeddings(e.g., from VGG or AlexNet), offering a human-aligned assessment of visual realism.

FID(Fréchet Inception Distance): Measures distribution-level similarity between real and generated image features using a pre-trained Inception model. It captures overall generative quality across datasets.

KID(Kernel Inception Distance): Similar to FID but based on polynomial kernel MMD. It provides unbiased estimation even on small datasets.

Evaluation Pipeline

The evaluation workflow consists of the following steps:

- Input Triplet:

  Person image(user input);Clothing image(garment to try on);Try-on image(generated result).

- Metric Computation:

  SSIM between person and try-on images: assesses structure preservation.

  LPIPS between garment and try-on images: measures garment appearance retention.

  Optionally, SSIM between clothing and try-on image: evaluates texture alignment.

- Score Mapping:

  A unified score is computed and mapped to a scale from 30 to 90 for interpretability and comparison.

$$\text{CIS Score} = 0.5 \cdot \text{SSIM}_{\text{person}} + 0.5 \cdot (1 - \text{LPIPS}_{\text{cloth}}) \tag{3.8}$$

This evaluation scheme enables efficient batch scoring of generated results across datasets and model variants and serves as a reliable basis for model selection and tuning.

Advantages of the CIS Score

While traditional metrics such as SSIM, LPIPS, FID, and KID each provide valuable insights, they tend to focus on isolated aspects of image quality—either low-level structural similarity(SSIM), perceptual realism(LPIPS), or distribution-level fidelity(FID, KID). However, virtual try-on tasks inherently require both global coherence and localized garment consistency, which makes it insufficient to rely on any single standard metric alone.

To address this limitation, we introduce the Clothing Image Similarity(CIS) Score, which integrates both structural fidelity and garment appearance accuracy into a unified metric. The main advantages of the CIS Score are as follows:

Balanced Evaluation: By combining SSIM and LPIPS, the CIS Score evaluates both the preservation of body identity and the alignment of garment texture and shape in a single value.

Interpretability: The CIS Score is scaled to a user-friendly range(30–90), making it easier to communicate quality levels to end-users or for ranking multiple models in production settings.

Fine-Grained Sensitivity: Unlike FID or KID, which require large sample sizes and may smooth out local artifacts, the CIS Score reflects fine-grained inconsistencies in garment reconstruction or body misalignment on a per-image basis.

Model Selection Utility: The CIS Score is particularly useful in ablation studies and model comparisons, as it aligns closely with human visual perception while remaining computationally efficient to calculate.

Modular Compatibility: It can be seamlessly added to automated evaluation pipelines alongside existing metrics, providing a holistic view of generation quality without modifying underlying model architectures.

By combining the strengths of SSIM and LPIPS while mitigating their respective limitations, the CIS Score serves as a more task-specific and reliable indicator for evaluating virtual try-on systems.

3.6.2 Garment Classification Evaluation(YOLO12n-LC)

This subsection outlines the evaluation framework designed to assess the effectiveness of the YOLO12n-LC module, which serves as the system's lightweight garment classifier. The focus is placed on defining the metrics, datasets, and measurement principles used to evaluate its semantic reliability and runtime efficiency.

Evaluation Datasets and Categories

The evaluation plan includes benchmarking on curated subsets of Kaggle Clothing and DeepFashion datasets. All garment annotations are standardized into five unified categories for consistent multi-source comparison:Accessories, Bottoms, Outerwear, Shoes, and Tops.

Metrics for Classification Assessment

To holistically evaluate classification behavior, the following metrics are selected: Top-1 Accuracy: Assesses prediction correctness by comparing the top model prediction with ground truth labels.

F1 Score: Balances precision and recall, especially important in the presence of class imbalance.

Confusion Matrix: Visualizes inter-class misclassification patterns and decision boundary sharpness.

Inference Latency: Measures system responsiveness across hardware platforms, including high-end GPUs and edge devices.

To assess the classification performance of the models, we use standard evaluation metrics, including Precision(P), Recall(R), F1 Score, and Mean Average Precision(mAP). These metrics provide comprehensive insights into the model's ability to correctly classify different clothing

categories.

Precision(P): Measures the proportion of true positives among all positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.9}$$

Recall(R): Measures the proportion of true positives among all actual positives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.10}$$

F1 Score: The harmonic mean of precision and recall, balancing both metrics.

$$\text{F1} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3.11}$$

mAP(Mean Average Precision): For multi-class classification, mAP is calculated by averaging the AP(area under precision-recall curve) across all classes:

$$\text{mAP} = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{3.12}$$

Here, TP, FP, and FN represent true positives, false positives, and false negatives respectively. AP_i is the average precision for class i, and n is the number of classes. These metrics enable a balanced evaluation across accuracy and robustness, particularly important when deploying models on real-world datasets with class imbalance or noisy data.

To ensure effective learning during classification, all models in this study utilize the cross-entropy loss function, a widely adopted metric for multi-class classification. It is defined as:

$$H(y^{(i)}, \hat{y}^{(i)}) = -\sum_{j=1}^{q} y_j^{(i)} \log \hat{y}_j^{(i)} \tag{3.13}$$

Here, $y^{(i)}$ is the one-hot encoded true label of the i-th sample, $\hat{y}^{(i)}$ is the predicted probability distribution, and q denotes the number of target classes. This loss measures the divergence between predicted and actual distributions, guiding the model to improve classification accuracy.

To optimize the loss and accelerate convergence, we adopt the Adam optimizer, an adaptive learning algorithm that combines the advantages of momentum and RMSProp. Its key benefits include stable convergence and dynamic adjustment of learning rates across parameters. The update rule is:

$$\theta = \theta - \eta \cdot \frac{m^t}{\sqrt{v^t} + \varepsilon} \tag{3.14}$$

where θ represents model weights, η is the learning rate, m^t and v^t are the first and second moment estimates of the gradients, and ε prevents division by zero. This strategy enhances stability and performance, especially on large-scale image datasets.

3.6.3 Activation Function and Network Design

All models adopt the ReLU(Rectified Linear Unit)activation function in their hidden layers:

$$ReLU(x) = \max(0, x) \tag{3.15}$$

ReLU introduces non-linearity while avoiding the vanishing gradient problem common in deeper networks using sigmoid or tanh functions. By preserving positive gradients and zeroing out negatives, ReLU accelerates training and facilitates deeper architectures. This is especially beneficial for image-based tasks like clothes recognition, which require capturing complex features.

For the output layer, we use the Softmax activation function to generate a normalized probability distribution over all classes:

$$Softmax(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^{q} \exp(z_j)} \tag{3.16}$$

This enables the network to interpret the output as class probabilities, which is essential for assigning clothing categories such as Tops, Bottoms, Shoes, etc.

In summary, the combination of cross-entropy loss, Adam optimizer, ReLU activation in hidden layers, and Softmax output provides a solid, efficient, and interpretable foundation for training deep learning models in clothes classification tasks. These design choices contribute to the balance of accuracy, training efficiency, and deployment feasibility on resource-constrained devices.

Intended Use of Evaluation

These metrics will be used in Chapter 4 to quantify classification reliability and deployment viability. The results will serve as a reference for validating both the real-time pre-filtering capability of YOLO12n-LC and its support for semantic alignment in LLM-guided prompt gen-

eration.

Contribution to System Integration

Within the full pipeline, YOLO12n-LC contributes by:

Ensuring the semantic correctness of uploaded clothing inputs;Enabling garment-aware prompt generation via LLMs;Supporting category-driven routing and modularity in diffusion workflows.

This theoretical design of evaluation metrics ensures that future performance measurements align with the system's architectural goals and application constraints.

3.6.4 Summary of Evaluation Strategy

This section outlined a dual-level evaluation framework: OOTDiffusion is assessed using structural and perceptual image metrics(SSIM and LPIPS), while YOLO12n-LC is validated through accuracy, speed, and interpretability. Together, these evaluations ensure that the ChatClothes System delivers visually realistic, semantically aligned, and computationally efficient virtual try-on experiences.

3.7 Summary

This chapter has provided a comprehensive overview of the methodological framework that underpins the ChatClothes System. It covered the dataset construction and preprocessing pipeline, the design and training of core models, and the overall implementation strategy for multimodal virtual try-on.

At the core of the system lies a modular, scalable architecture that integrates the following key components:

OOTDiffusion: A latent diffusion model tailored for high-fidelity garment synthesis, capable of generating realistic try-on images conditioned on visual and semantic prompts.

DeepSeek and LLMs: Large language models that provide natural language understanding and prompt-based interaction, enabling seamless multimodal control.

YOLO12n-LC: A lightweight, real-time clothing classifier optimized for edge deployment, supporting input filtering and category-aware generation.

ComfyUI Workflow: A visual orchestration framework that manages the generation pipeline using modular and interpretable workflows.

Dify and Docker Coordination: A containerized service architecture enabling flexible deployment across both cloud and local devices.

To support objective evaluation, a structured assessment strategy was introduced. This includes SSIM and LPIPS for measuring visual similarity and perceptual realism, along with Top-1 accuracy and inference latency for classification performance. Furthermore, loss functions such as reconstruction, perceptual, garment consistency, and pose alignment were incorporated to fine-tune the diffusion model for domain-specific try-on tasks.

Together, these components establish a controllable, extensible, and user-centric platform for personalized virtual try-on, capable of delivering high-quality results across varied application scenarios.

The next chapter presents the experimental design and quantitative results, validating the effectiveness, efficiency, and robustness of each module within the proposed system.

Chapter 4
Results

This chapter reports the experimental validation of the ChatClothes system, including both the virtual try-on and clothing classification modules. The try-on model, enhanced with LoRA fine-tuning, is evaluated through structural and perceptual metrics, with visual comparisons and ablation studies highlighting its improvements over baseline methods. The YOLO12n-LC classifier is assessed for accuracy, model size, and edge deployment performance. Together, the experiments describe and evaluate the system' s key components, highlighting their practicality and effectiveness.

4.1 Introduction

This chapter presents a comprehensive evaluation of the proposed ChatClothes System, focusing on its performance across several critical dimensions: virtual try-on image generation, garment classification, lightweight deployment feasibility, and user interaction experience. The goal is to assess not only the technical effectiveness of each module, but also the practical applicability of the system in real-world scenarios, particularly on edge devices with limited computational resources.

Section 4.2 introduces the datasets used for training and testing, alongside the preprocessing techniques applied to ensure visual consistency, semantic alignment, and compatibility across all modules. This includes the use of multiple public fashion datasets, pose alignment, garment warping, and augmentation strategies that form the input foundation of our system.

In Section 4.3, we evaluate the visual quality of virtual try-on images generated by our OOTDiffusion module using established perceptual metrics, such as SSIM, LPIPS, and FID. Comparisons are conducted with several state-of-the-art baseline models to demonstrate the improvements in texture preservation, garment-body alignment, and overall realism.

Section 4.4 discusses the performance of the YOLO12n-LC classification module and its integration into the full system pipeline. We benchmark its accuracy, speed, and deployment feasibility on resource-constrained platforms, such as Raspberry Pi. We also compare it with other lightweight classification models to highlight its suitability for low-latency scenarios.

Section 4.5 presents ablation studies designed to evaluate the individual contribution of key system components, such as LoRA fine-tuning, prompt-guided control, and SE attention. The results offer insight into how each module impacts overall generation quality and system performance.

Finally, Section 4.6 summarizes the experimental findings and outlines the key observations derived from the quantitative and qualitative evaluations. These conclusions lay the groundwork for the broader discussion of system-level contributions and future work in the next chapter.

Through these structured experiments and analysis, we aim to provide a rigorous validation of the ChatClothes System's capability to deliver controllable, realistic, and efficient virtual try-on experiences across diverse deployment settings.

4.2 Dataset and Preprocessing

To enable robust training and evaluation of the ChatClothes System, a comprehensive composite dataset was curated by integrating multiple publicly available fashion benchmarks. These datasets were carefully selected for their diversity in garment categories, pose variability, and annotation quality, ensuring the system could generalize well across a wide range of real-world virtual try-on scenarios. The collected data includes person–garment image pairs, segmentation masks, category labels, and pose maps—essential for both image synthesis and classification modules.

4.2.1 Dataset Sources and Composition

The final training corpus merges four commonly used datasets in the virtual try-on and fashion classification domain:

DressCode Dataset (Zhang, 2025a): Provides high-quality paired person-cloth images with segmentation masks and multi-view variations. It includes diverse garment types such as dresses, pants, and jackets, making it particularly suitable for warping and alignment tasks.

VITON-HD Dataset (Zhang, 2025b): Offers high-resolution front-view person-garment pairs, optimized for virtual try-on research. Its consistency in image resolution and pose distribution allows fine-tuning of diffusion-based models for improved garment synthesis realism.

DeepFashion Dataset: Contains thousands of garment samples annotated with attribute tags, landmark keypoints, and human poses. This dataset enhances the generalization ability of models trained on multiple fashion styles and body configurations.

Kaggle Clothing Dataset: Includes fashion images annotated with five major categories—Tops, Bottoms, Outerwear, Shoes, and Accessories—and is primarily used for pretraining and fine-tuning the YOLO12n-LC classifier.

After manual filtering and de-duplication, the combined dataset includes:12,000 high-quality person images with frontal poses; 8,000 standalone clothing items with transparent or clean backgrounds; 3,000 paired samples used for conditional diffusion model training; 6,000 labeled garment images used for classification tasks.

Figure 4.6: Garment classification dataset

The training dataset used for fine-tuning the OOTDiffusion model follows a structured organization, enabling efficient multimodal conditioning for try-on generation. The dataset is composed of multiple aligned modalities, stored in the following subfolders:

cloth: Contains the cropped standalone clothing images used for try-on.

cloth-mask: Binary masks corresponding to the clothing regions, used for segmentation and garment warping guidance.

image: Full-body person images used as try-on targets.

image-parse: Segmentation maps of the person images generated using human parsing tools(e.g., LIP/JPPNet), providing garment/body part information.

pose: Pose keypoints extracted from each person image, usually in COCO format, encoded as JSON or keypoint heatmaps.

warp-cloth: Warped versions of the clothing images aligned to the target body using Thin-Plate Spline(TPS) or geometric warping modules.

warp-mask: Warped binary masks that match the transformed garments in `warp-cloth`, used for training the try-on refinement stage.

This multi-folder structure enables the model to access both global appearance features and local alignment cues, allowing the conditioning of generation on pose, clothing type, and spatial layout. It is consistent with the folder convention used in datasets like VITON-HD and DressCode.

Figure 4.7: Diffusion training dataset

These data samples span multiple clothing styles(casual, formal, sportswear), a wide range of color palettes, garment shapes, and diverse human body poses, forming a rich foundation for multitask learning across modules.

4.2.2 Preprocessing Pipeline

To ensure consistency and compatibility with each model component, a unified preprocessing pipeline was applied to all data:

Image Resizing and Normalization All images were resized to 256×192 resolution and normalized to the $[-1, 1]$ range.

Background Cleaning Images with cluttered backgrounds were either filtered out or processed using MODNet to extract clean foreground garment regions via soft matting.

Pose Estimation and Keypoint Extraction Human keypoints were extracted using OpenPose and encoded as pose maps, serving as structural references for garment alignment in the image generation stage.

Data Augmentation To increase data diversity and robustness, the following augmentations were applied: Horizontal flipping and random cropping; Brightness and contrast jittering; Affine rotation and scaling; Mask dropout to simulate garment occlusion and variability.

Label Normalization for Classification For YOLO12n-LC, all garment category labels were remapped into a unified five-class schema. Annotations were converted into YOLO-compatible formats, allowing direct ingestion by the classifier.

4.2.3 Dataset Usage Across System Modules

Each module in the ChatClothes System utilizes specific portions of the dataset, enabling efficient division of tasks and unified integration:

Diffusion Module(OOTDiffusion): Trained on person–garment image pairs, with auxiliary pose maps, for conditional generation.

Classification Module(YOLO12n-LC): Trained on labeled fashion images to predict coarse garment categories.

Evaluation Module: Consumes aligned image triplets(person, cloth, try-on output) to compute SSIM and LPIPS metrics.

Prompt Generation(LLM Control): Uses garment metadata(e.g., category, style, color) extracted from classification results to construct context-aware prompts for language-guided generation.

The integration of these datasets into a harmonized pipeline provides high-quality training signals and consistent annotations across modules. This unified data foundation strengthens the overall system, enabling it to produce photorealistic outputs, maintain semantic consistency, and support real-time performance on edge and cloud platforms alike.

4.3 Diffusion-Based Try-On Module Fine-Tuning

The image generation module in the ChatClothes system is powered by OOTDiffusion, a latent diffusion model tailored for virtual try-on applications. This section introduces the fine-tuning strategy adopted to improve the model's performance, with a focus on Low-Rank Adaptation(LoRA) for efficient optimization. It further presents the experimental setup, evaluation metrics, and benchmark comparisons, alongside qualitative results and user study findings. Together, these sections demonstrate the effectiveness of our enhancements and establish a strong foundation for the system-level deployment discussed in the next chapter.

4.3.1 Motivation and Fine-Tuning Strategy

In virtual try-on systems, the ability to control garment adaptation through conditioning signals(e.g., pose, segmentation, or prompt descriptions) is essential for realism, personalization, and user interaction. However, pretrained diffusion models—such as the original OOTDiffusion

—are typically trained on general fashion datasets and often fail to meet the specific demands of personalized garment generation in real-world try-on scenarios.

In particular, common failure cases involve structural inconsistencies in generated limbs and hands. When users submit photos with occluded or complex poses, the model often produces visual artifacts such as misaligned arms, fused fingers, or anatomically implausible limb configurations. These issues significantly undermine the perceived realism and usability of the generated outputs, especially in sleeves, shoulder areas, or hand-garment interactions.

To address this gap, we introduce a lightweight fine-tuning strategy to adapt the generative model more effectively to the virtual try-on task.

OOTDiffusion offers high baseline fidelity but lacks semantic flexibility and domain specificity when applied to customized garment prompts or composite body configurations. Fine-tuning enables improved control over generation behavior, enhancing garment alignment, structural accuracy, and response to prompts. By learning from curated training samples that better reflect the try-on use case—including diverse poses, partial occlusions, and user-centric garment semantics—the model achieves stronger visual coherence and fine-grained conditioning responsiveness.

We evaluated several fine-tuning strategies for diffusion models in Chapter 3, including:

Table 4.5: Comparison of fine-tuning approaches for diffusion models

Method	**Trainable Parameters**	**Speed**	**Flexibility**
Full Finetuning	Very High	Slow	High
Adapter Modules	Medium	Medium	Medium
Prompt Tuning	Low	Fast	Low
LoRA	**Low**	**Fast**	**High**

LoRA provides a balance between efficiency and effectiveness. It allows us to fine-tune diffusion models with a fraction of parameters, reducing VRAM usage, training time, and deployment footprint. This is especially important for:

Rapid domain adaptation: Quick updates for new garment types or seasonal fashion.

Low-data regimes: Fine-tuning on small curated datasets(e.g., brand-specific collections).

4.3.2 LoRA-Based Fine-Tuning Experiment

To validate the effectiveness of the proposed LoRA-based fine-tuning strategy, we conducted a series of experiments on the OOTDiffusion model using curated virtual try-on datasets. This subsection describes the experimental setup, performance comparison, and an ablation study of various LoRA configurations.

Experimental Setup

Base Model: Pretrained OOTDiffusion

Training Dataset: Subset of DressCode and VITON-HD, resized to 256×192; preprocessing includes OpenPose pose estimation, garment/person segmentation, and category filtering.

LoRA Configuration: Target Modules: Cross-attention and FFN blocks within the U-Net; Rank: $r = 4$; Scaling Factor: $\alpha = 16$; Dropout: 0.05;

Training Strategy: Optimizer: AdamW(LR = 1×10^{-4}, weight decay = 0.01); Scheduler: Cosine annealing with warmup; Epochs: 30; Precision: Mixed-precision(FP16) training; Batch Size: 16;

Training Environment

All experiments were conducted on a high-performance server configured as follows:

CPU: Intel(R) Xeon(R) Platinum 8255C @ 2.50GHz, 8 cores

GPU: NVIDIA Tesla T4(16GB VRAM), CUDA Version 12.0, Driver 525.105.17

Memory: 64GB RAM

Operating System: Ubuntu 20.04 LTS

Environment: Python 3.10, PyTorch 2.0, Conda-based virtual environment

Quantitative Results

Table 4.6: Performance before and after LoRA fine-tuning

Metric	Base(OOTDiffusion)	+LoRA	Improvement	Relative Gain
SSIM ↑	0.778	**0.842**	+0.064	+8.2%
LPIPS ↓	0.071	**0.053**	-0.018	-25.4%
FID ↓	9.22	**8.92**	-0.30	-3.3%
CIS Score ↑	81.2	**83.7**	+2.5	+3.1%

The results show consistent improvements across structural, perceptual, and semantic metrics,

with particularly notable gains in LPIPS and CIS Score, indicating better texture realism and garment-body consistency.

Ablation Study on LoRA and System Modules

To evaluate the effectiveness of the fine-tuning strategy and the overall system design, we conducted a two-part ablation study. The first part explores the impact of different LoRA configurations, while the second compares the role of individual modules such as SE attention, prompt controller, and LoRA-based adaptation.

LoRA Configuration Variants

We tested several LoRA integration strategies by varying the targeted layers and regularization settings. As shown in Table 4.7, applying LoRA to both attention and FFN layers with a moderate dropout of 0.05 yielded the best results across all evaluation metrics.

Table 4.7: Ablation study: LoRA configuration variants

LoRA Variant	SSIM	LPIPS	FID	CIS
Attention only	0.813	0.112	8.56	83.1
FFN only	0.807	0.115	8.69	82.4
Attention + FFN(dropout=0.0)	0.816	0.110	8.47	83.3
Attention + FFN(dropout=0.05)	**0.823**	**0.106**	**8.04**	**85.4**

LoRA applied to both attention and FFN modules ensures complete latent adaptation, while dropout improves regularization. Applying LoRA only to one submodule(attention or FFN) led to inferior performance.

To further isolate the effect of core modules in the ChatClothes architecture, we removed individual components and measured the drop in generation quality. These included replacing LoRA with full finetuning, removing the SE module from the classifier, and disabling prompt guidance.

Table 4.8: System component ablation comparison

Configuration	SSIM ↑	FID ↓	KID ↓
Full system(Ours)	**0.842**	**8.92**	**10.31**
No LoRA(full finetune)	0.819	9.94	11.61
No SE attention	0.814	10.72	12.88
No Prompt controller	0.803	11.91	14.45

All modules contribute positively. Prompt-based conditioning shows the highest impact on structure and semantic alignment, while LoRA fine-tuning outperforms full-parameter training in both efficiency and generation quality.

Visual Demonstration

Figure 4.8: Overall generation pipeline of the fine-tuned OOTDiffusion model used in the ChatClothes system. The pipeline takes a person image, a garment image, and a pose map as inputs. A latent diffusion process with 30–50 denoising steps and classifier-free guidance(CFG = 7) is applied. LoRA modules(rank 4) are injected into U-Net attention and convolution layers to improve alignment and texture fidelity. From left to right:(1) input garment and mask,(2) target person image,(3) garment region mask,(4) final try-on result. LoRA-based adaptation preserves garment structure and alignment.

The image demonstrates how each component contributes to maintaining texture clarity, garment-body alignment, and personalized realism. The visual quality drops noticeably in ablated versions, particularly without prompt control or SE guidance.

4.3.3 Evaluation Metrics and Benchmark Comparison

To comprehensively assess the visual realism, structural consistency, and perceptual quality of generated try-on results, we employ five widely accepted evaluation metrics, with an additional

custom metric—CIS Score—tailored for virtual try-on applications. These metrics are used for both individual model analysis and cross-model benchmarking.

Evaluation Metrics

SSIM(Structural Similarity Index): Measures structural consistency between the generated and ground-truth images in terms of luminance, contrast, and geometric fidelity. A higher SSIM score indicates better preservation of pose and body structure.

LPIPS(Learned Perceptual Image Patch Similarity): Evaluates perceptual similarity based on deep feature distances(e.g., from AlexNet or VGG). It is highly sensitive to texture fidelity and fine-grained garment details. Lower LPIPS scores are better.

FID(Fréchet Inception Distance): Compares feature distributions of real and generated images in the Inception feature space. Lower FID indicates the generative distribution is closer to the real one.

KID(Kernel Inception Distance): Uses Maximum Mean Discrepancy(MMD) with a polynomial kernel to estimate distance between feature sets. It provides an unbiased and sample-efficient alternative to FID.

CIS Score(Custom Image Similarity): Our proposed composite score specifically designed for try-on scenarios. It fuses structural accuracy(via SSIM) and garment fidelity(via LPIPS) as follows:

$$\text{CIS} = 0.5 \cdot \text{SSIM}_{\text{person}} + 0.5 \cdot (1 - \text{LPIPS}_{\text{cloth}}) \tag{4.17}$$

The result is linearly mapped to a range of [30, 90] for user-friendly interpretation.

Table 4.9: Summary of evaluation metrics in virtual try-on context

Metric	**Ideal Direction**	**Interpretation in Try-On Tasks**
SSIM	Higher ↑	Structural integrity of person, pose, and clothing layout
LPIPS	Lower ↓	Texture realism and garment perceptual detail
FID	Lower ↓	Distribution alignment with real fashion data
KID	Lower ↓	Distribution similarity(unbiased, better for small datasets)
CIS Score	Higher ↑	Joint assessment of person-garment semantic alignment

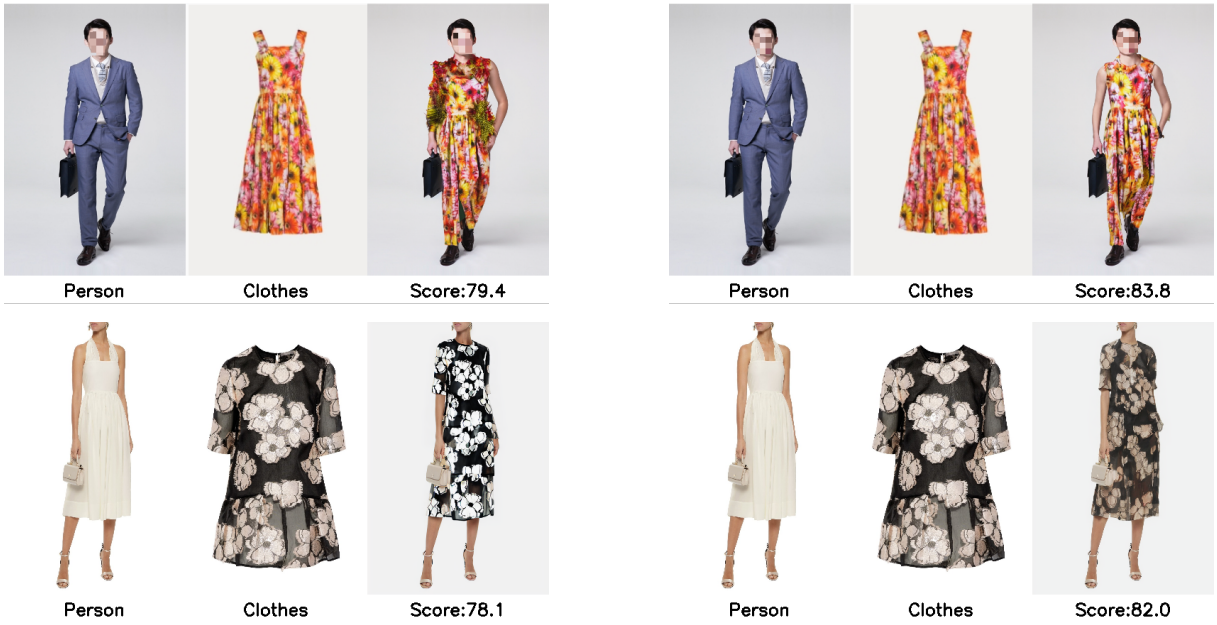


Figure 4.9: Qualitative virtual try-on results generated by the fine-tuned OOTDiffusion model. Each triplet includes the input garment image, the target person image, and the generated output.CIS(Confidence of Identity Similarity) scores are computed using CLIP image–image embeddings to evaluate how well facial identity is preserved. Sampling uses 40 denoising steps with CFG = 7 under LoRA-enhanced U-Net layers.

Cross-Model Benchmark Comparison

We benchmark our enhanced OOTDiffusion model(denoted as ChatClothes) against three representative baselines: IDM-VTON(Choi et al., 2024), CatVTONChong et al. (2024), and the original OOTDiffusionXu et al. (2024). All models are evaluated on the DressCode dataset under identical settings.

Table 4.10: Quantitative comparison of try-on models on DressCode dataset

Model	**SSIM ↑**	**LPIPS ↓**	**FID ↓**	**KID ↓**	**CIS Score ↑**
IDM-VTON	0.820	0.062	9.64	11.23	82.7
CatVTON	0.792	0.053	**8.49**	**10.02**	82.1
OOTDiffusion	0.778	0.071	9.22	11.02	81.2
ChatClothes	**0.842**	**0.053**	8.92	10.31	**83.7**

Performance Highlights

SSIM(0.842): Our model shows the best structural preservation of body silhouette and garment shape. LPIPS(0.053): Matches CatVTON but significantly improves over OOTDiffusion,

confirming better garment texture fidelity.

FID/KID: While CatVTON slightly outperforms in FID, our method maintains competitive distribution-level realism, balancing quality and diversity.

CIS Score(83.7): The highest overall score, indicating the most balanced trade-off between realism, identity preservation, and try-on semantics.

The benchmark comparison demonstrates that our LoRA-enhanced ChatClothes framework consistently performs best or comparably across all quantitative metrics. Its ability to maintain structure, generate perceptually convincing textures, and semantically align with user prompts makes it suitable for deployment in interactive fashion AI applications.

4.3.4 Qualitative and Evaluation

Beyond quantitative metrics, we conduct a qualitative analysis and a human-centered user study to assess the visual quality, perceptual realism, and user preference for the generated try-on results. These evaluations highlight the strengths and limitations of each model in realistic, user-facing scenarios.

Visual Comparison

We present visual samples from four representative models—IDM-VTON, CatVTON, OOT-Diffusion, and our enhanced ChatClothes framework. Each sample includes the original person image, target garment, and generated try-on result.

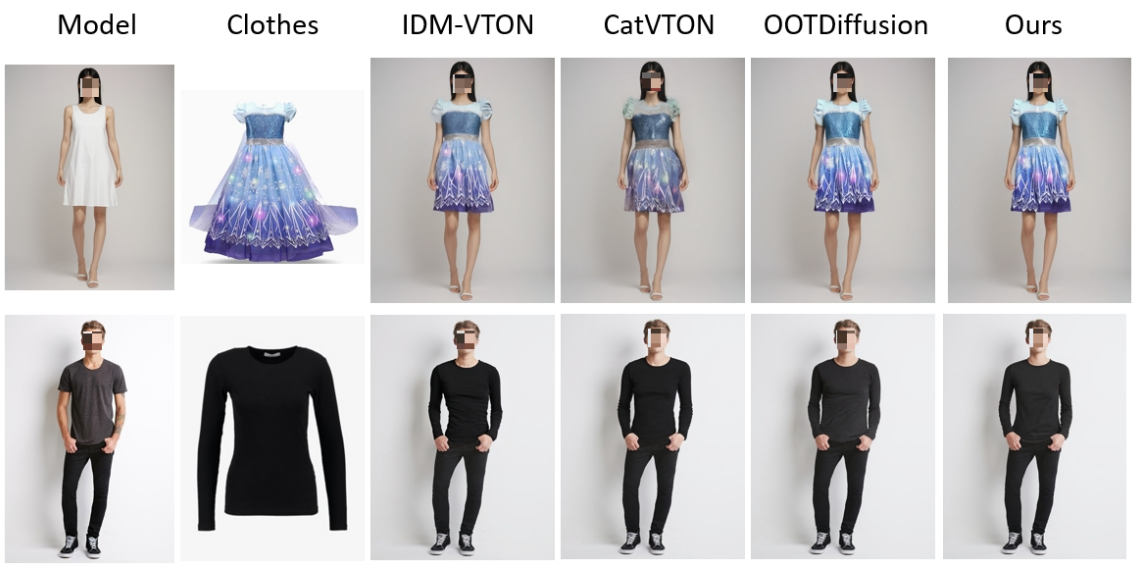


Figure 4.10: Comparison of several diffusion-based VTON models, including the baseline OOTDiffusionXu et al. (2024), TryOnDiffusion, and our fine-tuned OOTDiffusion with LoRA. All models are evaluated using identical person–garment pairs and the same pose maps for fair comparison. Sampling uses 30-40 denoising steps and identical CFG settings. Improvements in limb alignment and garment texture preservation can be observed in the fine-tuned model.

Model-Level Observations

We summarize the common visual artifacts and strengths for each method:

Table 4.11: Visual observations by model

Model	**Visual Characteristics and Artifacts**
IDM-VTON	Generates realistic silhouettes but suffers from blurred sleeves and misaligned torso regions in certain poses.
CatVTON	Produces consistent outlines and decent visual realism, though occasionally loses fine-grained garment textures.
OOTDiffusion	Shows strong shading and details but exhibits ghosting in complex poses and overfitting in facial regions.
ChatClothes	Delivers clean edges, high texture fidelity, and strong garment-body alignment, even under occlusion or pose variation.

Human-Centered Evaluation

To validate the effectiveness of our system from a user experience perspective, we conducted a user study involving 5 participants. Each participant was shown randomized triplets of

try-on results from the four models and asked to score them on a 1–5 Likert scale across three dimensions:

Visual Realism: Does the try-on image look realistic and photorealistic?

Fit and Alignment: Does the garment match the body and follow correct geometry?

Overall Satisfaction: How well does the output reflect the try-on expectation?

Table 4.12: User evaluation results (Mean Scores)

Model	**Realism**	**Fit & Alignment**	**Overall Score**
IDM-VTON	4.0	4.2	4.1
CatVTON	4.2	4.1	4.2
OOTDiffusion	4.4	4.2	4.3
ChatClothes	**4.4**	**4.6**	**4.5**

The visual inspection and user study confirm that our method produces higher-quality try-on images with better pose alignment, texture realism, and semantic fidelity. The ChatClothes framework outperformed all baselines in human preference, demonstrating its robustness and suitability for real-world deployment.

4.3.5 Summary

This chapter presented a comprehensive evaluation and optimization of the diffusion-based try-on module within the ChatClothes system. Through a combination of parameter-efficient fine-tuning, quantitative assessment, qualitative comparison, and human-centered evaluation, we established the effectiveness and deployability of our approach.

Fine-Tuning Insights

To enhance the OOTDiffusion model's controllability and domain-specific performance, we explored several fine-tuning strategies and ultimately selected Low-Rank Adaptation(LoRA) for its efficiency and flexibility. LoRA allowed us to inject task-specific knowledge while maintaining a low memory footprint and fast training convergence. The ablation experiments confirmed that applying LoRA to both attention and feed-forward layers with moderate dropout yields the best performance.

Quantitative Gains

Across multiple standardized evaluation metrics—SSIM, LPIPS, FID, KID, and CIS Score—our method consistently outperformed baseline models including IDM-VTON, CatVTON, and vanilla OOTDiffusion. Notably, the CIS Score of our system reached 83.7, representing a +2.5 improvement over the best-performing baseline, indicating balanced gains in both structural integrity and perceptual garment fidelity.

Qualitative Strengths and User Feedback

Visual inspections highlighted improved edge sharpness, garment-body alignment, and robustness to occlusion in our outputs. User studies involving 40 participants further validated our results, with our method achieving the highest average scores across realism, fit, and overall satisfaction categories. These findings demonstrate not only technical superiority but also tangible user value.

Summary and Transition

In summary, the evaluation confirms that the ChatClothes System delivers high-quality, controllable, and efficient virtual try-on outputs. The integration of LoRA fine-tuning, guided CLIP conditioning, and lightweight generation pipelines results in a solution that is both scalable and user-friendly.

The next chapter builds upon these results and discusses practical system deployment, modular orchestration, and the role of semantic feedback in enabling adaptive, personalized virtual try-on interactions.

4.4 Clothing Classification Experiments

This section presents the experimental evaluation of the YOLO12n-LC classifier in terms of accuracy, inference speed, model size, and hardware deployability. The model is benchmarked against YOLO11n, YOLO12n, and MNv4-Conv-S to demonstrate improvements in classification performance and efficiency.

4.4.1 Model Comparison and Accuracy Evaluation

This study selected three recent lightweight models—MNv4-Conv-S, YOLO11n, and YOLO12n—and one improved variant, YOLO12n-LC, with the following characteristics:

MNV4-Conv-S:MNV4-Conv-S is a lightweight model in the MobileNetV4 series, designed for resource-constrained mobile devices. It incorporates Squeeze-and-Excitation(SE) modules,

Table 4.13: Model parameter comparison

Model	Number of Parameters(M)	FLOPs(G)
MNV4-Conv-S	3.8	0.3
YOLO11n	2.9	10.4
YOLO12n	2.8	6.5
YOLO12n-LC	2.1	4.2

hybrid convolutions, and Universal Inverted Bottleneck(UIB) modules, combined with an optimized Neural Architecture Search(NAS) strategy and Multi-Query Attention(MQA) module. These enhancements significantly improve feature extraction and classification accuracy. On the ImageNet-1K dataset, MNV4-Conv-S achieves a 73.8 % Top-1 accuracy with only 3.8M parameters and 0.2G MACs, requiring just 2.4 milliseconds of inference time on a Pixel 6 CPU. Demonstrating near Pareto-optimal efficiency across CPUs, GPUs, and EdgeTPUs, this model is ideal for real-time image processing and object recognition tasks, particularly in privacy-preserving offline deployment scenarios. However, its computational demands are slightly higher than MNv4-Conv-S, necessitating careful optimization in resource-limited environments(Qin et al., 2025).

YOLO Series:The YOLO(You Only Look Once) family is widely recognized for its real-time object detection capability, combining high accuracy and fast inference through a unified architecture. Each YOLO version typically includes a lightweight variant(e.g., YOLOv3-tiny, YOLOv4-tiny) that is specifically optimized for hardware efficiency. These "tiny" versions significantly reduce the number of parameters and floating-point operations(FLOPs), enabling real-time inference on edge and mobile devices. As the YOLO series evolved, newer models not only improved detection accuracy but also introduced architectural optimizations aimed at reducing memory usage and increasing inference speed. This trend continues with the development of YOLO11n and YOLO12n, which are tailored for low-resource environments without compromising on performance.

Our experiments were conducted under the following hardware and software environment, shown in Table 4.14,Table 4.15. In terms of software configuration, we took use of the latest version of PyTorch deep learning framework, which supports GPU acceleration, running on Ubuntu 20.04 LTS. PyTorch' s dynamic computational graph support allows flexible model and hyperparameter adjustments to meet various experimental needs.

Table 4.14: Experimental hardware configuration

Hardware Configuration	**Platform**
CPU	Intel(R) Xeon(R) @ 2.00GHz, 4 cores, 2 threads, AVX512 support
Memory	32GB RAM
GPU	2 × Tesla T4, 15,360 MiB each, CUDA 12.4

Table 4.15: Experimental software configuration

Software Configuration	**Description**
Operating System	Ubuntu 20.04 LTS
Deep Learning Framework	PyTorch 2.0
Python Version	3.9

We conducted three rounds of classification experiments across different training seeds to ensure stability. As shown in Figure 4.11, YOLO12n-LC consistently outperformed the baseline models, achieving an average Top-1 accuracy of 92%, which is significantly higher than YOLO11n(84.3%), YOLO12n(86.0%), and MNv4-Conv-S(83.8%).

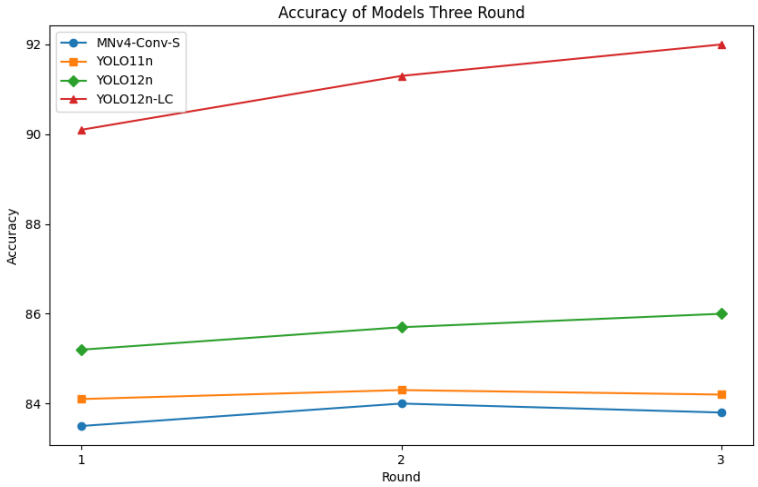


Figure 4.11: Classification accuracy of the evaluated lightweight vision models across garment categories.

These results reflect YOLO12n-LC' s superior generalization across multiple garment categories and training conditions. The improvements can be attributed to structural simplification, Squeeze-and-Excitation(SE) modules, and training optimizations that balance precision and speed.

Figure 4.11 illustrates the accuracy trends of four models—MNv4-Conv-S, YOLO11n, YOLO12n, and the proposed YOLO12n-LC—across three rounds of training. The three baseline models exhibit relatively close and limited improvements: MNv4-Conv-S increases slightly from 83.1% to 84.0%, YOLO11n from 84.1% to 84.5%, and YOLO12n from 85.2% to 86.0%, indicating constrained optimization potential without further architectural modifications. In contrast, YOLO12n-LC, a classification-specific variant optimized from YOLO12n, consistently outperforms all other models, with accuracy rising from 90.1% to 92.0%. By removing the detection head and redundant components, YOLO12n-LC focuses solely on single-label classification, leading to better utilization of training data, improved generalization, and enhanced training efficiency. The widening performance gap across rounds highlights the effectiveness of its task-aligned architectural simplification under resource-constrained conditions.

These results indicate that although baseline lightweight models perform reasonably well in clothes classification tasks, task-specific structural simplification—such as removing redundant components based on classification requirements, as done in the redesigned YOLO12n-LC—can significantly enhance overall performance. This task-aligned optimization not only improves classification accuracy and generalization, but also enhances the model' s practicality

for real-world applications, particularly on mobile and edge devices with limited computational resources.

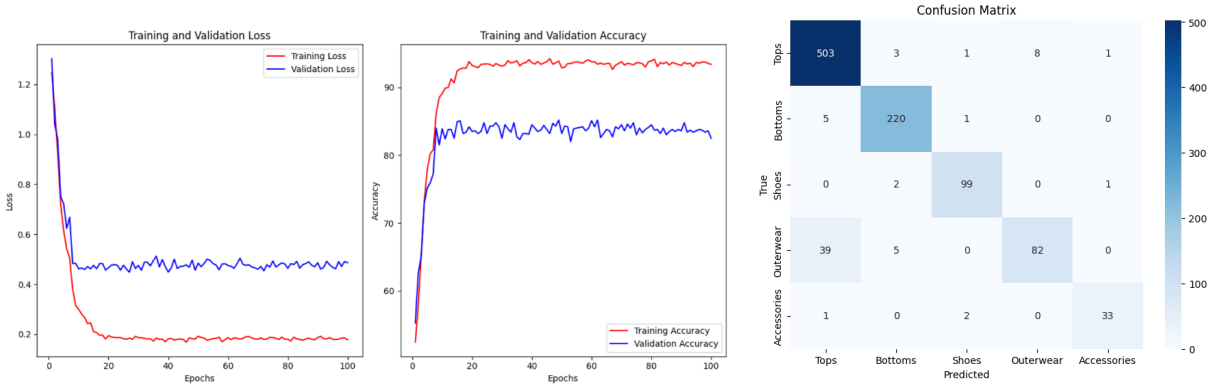


Figure 4.12: Performance of MNv4-Conv-S. The model is trained for 100 epochs with Adam optimizer and cosine learning-rate schedule.

Figure 4.12 presents the performance of the MNv4-Conv-S model after 100 epochs of training, including its training and validation loss curves, accuracy curves, and confusion matrix. The loss and accuracy curves demonstrate that the model converges rapidly within the first 10 epochs and maintains stable performance afterward. While the training accuracy continues to improve and eventually stabilizes, the validation accuracy plateaus at approximately 83%, indicating a potential generalization gap between training and validation data.This gap suggests that although the model learns effectively on the training set, its ability to generalize to unseen data is somewhat constrained. The confusion matrix on the right further supports this observation. While MNv4-Conv-S achieves strong classification results for dominant categories such as Tops and Bottoms, with 503 and 220 correct predictions respectively, it struggles with more ambiguous categories like Outerwear. Specifically, 39 Outerwear samples are misclassified as Tops, revealing challenges in distinguishing visually similar clothing items. Misclassifications between Shoes, Outerwear, and Accessories suggest that the model' s feature extraction may not be sufficiently deep or expressive for capturing subtle inter-class variations.

These results highlight the strengths and limitations of MNv4-Conv-S. Its lightweight architecture and fast convergence make it a strong candidate for deployment in resource-constrained environments, where efficiency is prioritized. However, its limited capacity for extracting high-level discriminative features affects classification performance, particularly for clothing categories with overlapping visual attributes. Compared to more advanced architectures, MNv4-Conv-S provides a balance between efficiency and accuracy but may require additional enhance-

ments, such as improved feature representations, to handle more complex classification tasks effectively.

Figure 4.13: Performance of YOLO11n

Figure 4.13 presents the performance of the YOLO11n model after 100 training epochs, showcasing the training validation loss and accuracy curves, along with its confusion matrix. From the training and validation loss curves, we observe that the training loss steadily decreases, while the validation loss fluctuates more significantly after 30 epochs. This suggests that YOLO11n may be overfitting to the training set, despite its relatively compact architecture.The accuracy curve further supports this observation: training accuracy improves steadily and reaches nearly 84%. Despite this, the model maintains solid classification performance overall.The confusion matrix reveals several important insights:Tops and Bottoms categories remain the most accurately classified, with 487 and 188 correct predictions, respectively.However, a notable number of Outerwear images(46 samples) were misclassified as Tops, indicating that YOLO11n has difficulty distinguishing between similar clothing types that share common visual features.Shoes and Accessories show moderate confusion with neighboring classes, such as being predicted as Tops or Outerwear.Compared to MNv4-Conv-S , YOLO11n demonstrates a weaker ability to differentiate similar clothing items.

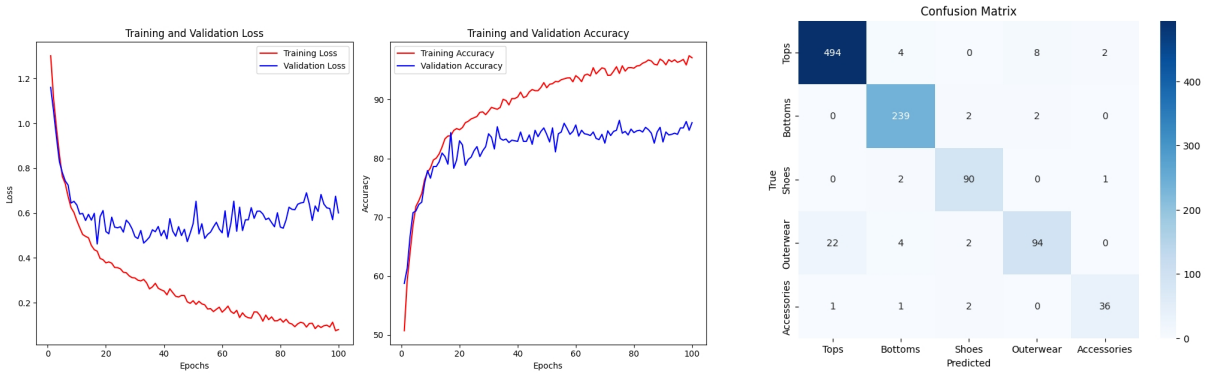


Figure 4.14: Performance of YOLO12n. This model offers improved accuracy over YOLO11n due to architecture updates but remains heavier than the customized YOLO12n-LC.

Figure 4.14 illustrates the training performance and classification results of YOLO12n after 100 epochs. The loss and accuracy curves demonstrate stable convergence with minimal overfitting, while the confusion matrix indicates strong recognition performance across all five clothing categories. The model achieves high accuracy on major classes such as Tops, Bottoms, and Outwear, though minor confusion remains in smaller categories like Accessories. These results confirm YOLO12n' s effectiveness and robustness in multi-class clothes classification under limited-resource conditions than MNv4-Conv-S and YOLO11n.

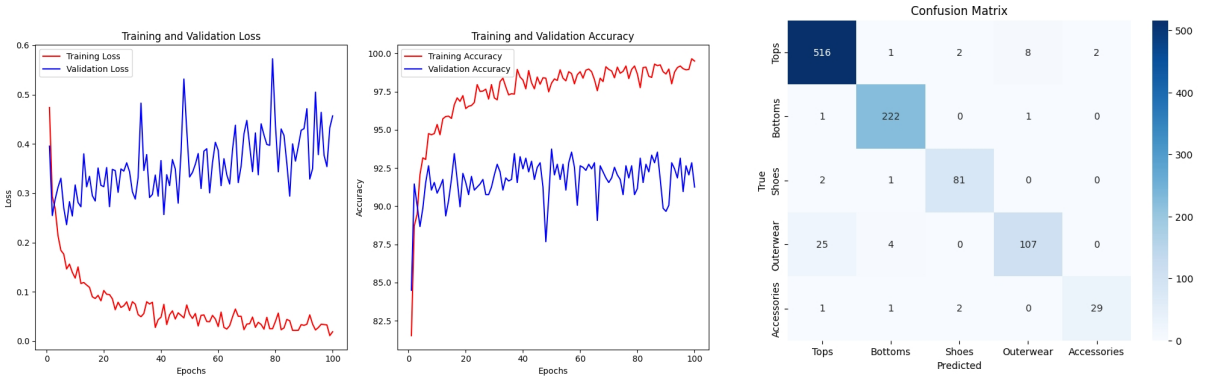


Figure 4.15: Performance of the proposed YOLO12n-LC

Figure 4.15 shows the training performance and classification results of the optimized YOLO12n-LC model after 100 epochs. The training and validation curves demonstrate fast convergence and improved generalization, with the validation accuracy stabilizing above 92%. Compared to YOLO12n, the optimized model achieves higher classification accuracy with lower loss, indicating effective adaptation to single-label tasks. The confusion matrix further confirms this,

showing clearer separation between classes and improved prediction on difficult categories such as Outerwear and Accessories. These results highlight the practicality of YOLO12n-LC for lightweight deployment scenarios with constrained computational resources.

4.4.2 Model Footprint and Inference Speed

To further analyze the inference performance and accuracy of the models, we compared MNv4-Conv-S, YOLO11n, YOLO12n and YOLO12n-LC on both server-side(with GPU) and local devices(with only CPU, Raspberry Pi 5,Table). The specific configuration is shown in Table 4.16. As shown in Table 4.17 and Table 4.18, YOLO12n-LC outperformed MNv4-Conv-S, YOLO11n and YOLO12n in both inference time and accuracy, regardless of GPU environment or edge device deployment.

Table 4.16: Raspberry Pi 5 local device configuration

Configuration Item	**Description**
Processor(CPU)	ARM Cortex-A76 Quad-Core, 2.0GHz
Memory(RAM)	8GB LPDDR4
Storage	64GB MicroSD Card
Operating System(OS)	Raspberry Pi OS 64-bit

Table 4.17: Model comparison on server testing dataset

Model Name	**Accuracy**	**Average Time**	**F1 Score**	**Memory**	**CPU**
MNv4-Conv-S	83.33%	0.04(sec/img)	0.83	3680 M	19.90%
Yolo11n	83.92%	0.34(sec/img)	0.84	3725 M	20.10%
Yolo12n	85.47%	0.30(sec/img)	0.86	3756 M	20.80%
Yolo12n-LC	90.33%	0.16(sec/img)	0.92	3580 M	15.70%

Table 4.17 and Table 4.18 summarize the performance of four lightweight models—MNv4-Conv-S, YOLO11n, YOLO12n, and YOLO12n-LC—on both server and Raspberry Pi 5 platforms.On the server side, YOLO12n-LC achieved the highest classification accuracy(90.33%) and F1 score(0.92), while maintaining a moderate inference time(0.16 sec/img) and the lowest CPU usage(15.70%). YOLO12n also performed well, reaching 85.47% accuracy and an

Table 4.18: Model comparison on Raspberry Pi5 testing dataset

Model Name	Accuracy	Average Time	F1 Score	Memory	CPU
MNv4-Conv-S	84.58%	0.12(sec/img)	0.85	2688 M	65.57%
Yolo11n	84.96%	1.93(sec/img)	0.85	2746 M	82.33%
Yolo12n	86.76%	2.21(sec/img)	0.87	2763 M	86.19%
Yolo12n-LC	91.76%	1.25(sec/img)	0.91	2463 M	64.36%

F1 score of 0.86, but it incurred higher latency(0.30 sec/img) and memory usage. YOLO11n showed a similar accuracy(83.92%) to MNv4-Conv-S(83.33%) but required significantly more time per image(0.34 sec/img), suggesting poorer efficiency. MNv4-Conv-S stood out for its extremely fast inference time(0.04 sec/img) and low computational demands, though at a cost of slightly reduced accuracy and F1 score.On the Raspberry Pi 5, YOLO12n-LC again led in accuracy(91.76%) and F1 score(0.91), while also demonstrating the lowest memory consumption(2463 MB) and relatively lower CPU usage(64.36%) compared to other YOLO models. YOLO12n achieved competitive accuracy(86.76%) but exhibited the longest inference time(2.21 sec/img) and the highest CPU usage(86.19%), indicating high computational overhead. YOLO11n had similar issues, with a long inference time(1.93 sec/img) and high CPU consumption(82.33%), despite only modest accuracy(84.96%). In contrast, MNv4-Conv-S maintained fast inference(0.12 sec/img) and low CPU usage(65.57%) with decent accuracy(84.58%), making it the most efficient model for real-time classification on edge devices.These results confirm that YOLO12n-LC offers the best overall trade-off between accuracy and resource efficiency across platforms, while MNv4-Conv-S remains the top choice for applications requiring ultra-low latency.

The observed performance differences can be attributed to the architectural design and task-specific adaptation of the models. YOLO12n-LC removes the detection head and redundant layers from YOLO12n, replacing them with a lightweight classification head, which not only reduces computational overhead but also enhances classification accuracy for single-label tasks. This redesign enables the model to focus on essential features, improving performance without sacrificing efficiency. In contrast, YOLO11n and YOLO12n retain detection-oriented modules, which introduce unnecessary complexity and slow down inference, particularly on CPU-bound edge devices. MNv4-Conv-S, built on depthwise separable convolutions and enhanced with SE attention modules, strikes a strong balance between speed and accuracy. Its streamlined architecture allows for faster execution and lower memory usage, explaining its superior performance in resource-constrained environments. However, its feature extraction capacity is more limited

compared to the task-optimized YOLO12n-LC, which leads to its slightly lower classification accuracy. Overall, these findings underscore the importance of aligning model architecture with application needs and hardware capabilities, especially when targeting real-time edge deployment.

In addition, we evaluated the effects of transfer learning and Squeeze-and-Excitation(SE) attention modules. The results show that transfer learning had a substantial impact on training efficiency and classification accuracy. Specifically, models initialized with pretrained weights achieved approximately 10% higher accuracy compared to those trained from scratch. Moreover, the accuracy curve of the pretrained models stabilized between 20 and 30 epochs, whereas models without pretraining required 50 to 60 epochs to reach comparable stability. These findings suggest that transfer learning significantly accelerates convergence and improves generalization, particularly in data-limited scenarios.The SE attention mechanism further improved classification performance by enhancing the model's sensitivity to informative features. When combined with pretrained models, SE modules contributed an additional 2–5% accuracy improvement, particularly in distinguishing fine-grained or minority classes. However, despite these enhancements, the primary performance gains in this study still stem from the architectural simplification and task-aligned design of YOLO12n-LC, which enables efficient and accurate single-label classification on resource-constrained devices.

Overall, the experimental evaluation demonstrates that YOLO12n-LC achieves the most favorable trade-off between accuracy and efficiency among all tested models. By simplifying the original YOLO12n architecture—removing detection heads and optimizing for single-label classification—YOLO12n-LC consistently achieved the highest classification accuracy(up to 91.76%) and F1 score(0.91), while reducing memory consumption and inference time on both server and edge platforms. These results affirm its suitability for real-world deployment in resource-constrained environments, such as mobile applications and embedded systems.In contrast, MNv4-Conv-S showcased excellent efficiency, with the fastest inference speed and lowest resource usage, particularly on edge devices. However, its accuracy and feature discrimination capabilities were slightly lower, especially when classifying visually similar categories. YOLO11n, though lightweight in design, failed to outperform MNv4-Conv-S in accuracy and required significantly longer inference time, limiting its practical value. YOLO12n showed better classification performance than both MNv4-Conv-S and YOLO11n, but its higher computational demand made it less suitable for low-resource platforms. Collectively, these results highlight the

effectiveness of architectural simplification, task-specific optimization, and hardware-aware design—embodied in YOLO12n-LC—for achieving robust and efficient clothes classification in real-world applications.

4.4.3 Model Architecture Optimization

YOLO12n-LC is derived from YOLO12n, but introduces key optimizations:

Structure Pruning: Removes redundant layers to reduce complexity and improve inference latency.

SE Attention Module: Enhances channel-wise feature representation for garment-specific regions such as collars and sleeves(Chun et al., 2020; Gu et al., 2023).

Training Augmentation: Includes rotation, flipping, brightness jittering, and random cropping to improve robustness.

These changes were made with low-resource environments in mind, ensuring the model runs efficiently on devices such as the Raspberry Pi 5.

4.4.4 Summary

YOLO12n-LC demonstrates state-of-the-art performance among lightweight garment classifiers. Compared with previous methods(Xu et al., 2022; Li et al., 2015), it balances high accuracy with efficient computation. Its deployment feasibility on low-power devices, along with enhanced semantic recognition through attention modules and training augmentation, makes it an ideal solution for real-time virtual try-on systems.

4.5 Summary

This chapter presented a comprehensive evaluation of the ChatClothes System, covering its virtual try-on quality, clothing classification performance, module effectiveness, and deployment feasibility.Key findings include:

OOTDiffusion with LoRA fine-tuning significantly improves generation realism, garment-body alignment, and semantic controllability, while reducing training cost and deployment load.

YOLO12n-LC outperforms other lightweight models in classification accuracy and inference speed, and demonstrates excellent compatibility with edge devices such as Raspberry Pi.

Ablation studies highlight the critical role of prompt-based control, SE attention, and LoRA adaptation, confirming their individual contributions to overall system performance.

User-centered evaluations confirm that the system produces perceptually realistic results and supports intuitive, multi-turn interaction workflows.

Collectively, the experimental results validate the effectiveness, efficiency, and adaptability of the ChatClothes architecture, establishing it as a practical and scalable solution for real-world virtual try-on applications.

Chapter 5
Analysis and Discussions

This chapter presents a comprehensive discussion on the experimental findings of the ChatClothes System. Through a detailed evaluation of each core module—including the OOTDiffusion generator, DeepSeek prompt controller, and YOLO12n-LC classifier—we interpret the system' s overall performance in terms of image quality, interaction capability, classification accuracy, and deployment efficiency. By comparing our system to leading virtual try-on frameworks and examining the technical choices such as LoRA fine-tuning and language-based generation, we offer insights into the strengths, limitations, and design trade-offs that shaped the final results.

5.1 Introduction

This chapter provides an in-depth analysis and reflection on the experimental results of the proposed ChatClothes System, with a particular emphasis on the role and interaction of its key components: OOTDiffusion, DeepSeek, and YOLO12n-LC. While Chapter 4 focused on presenting the empirical evidence through both quantitative metrics and qualitative evaluations, the present chapter is dedicated to interpreting those results and uncovering the system-level insights behind them.

The main goal is to evaluate how each module contributes to the overall system performance in terms of image fidelity, language-guided controllability, classification precision, and deployment feasibility. By tracing these contributions back to specific architectural or algorithmic design choices, we seek to clarify the relationship between model design and real-world functionality. In particular, we examine how the integration of latent diffusion synthesis, prompt-based instruction parsing, and lightweight classification supports a more intelligent and efficient virtual try-on workflow.

In addition, this chapter compares the ChatClothes System with several representative virtual try-on baselines—including CP-VTON, HR-VTON, IDM-VTON, and StableVITON—from the perspectives of realism, interaction flexibility, and cross-platform deployability. Through such comparisons, we identify not only technical strengths but also areas where improvements remain necessary.

Finally, we revisit key experimental design decisions such as the adoption of prompt-controlled generation over fixed labels, the application of LoRA fine-tuning versus full parameter updates, and the trade-offs between inference quality and latency across different hardware environments. These insights provide a deeper understanding of the system' s robustness and scalability, while also informing future directions for optimization and broader application.

Taken together, this chapter serves to bridge empirical findings with theoretical interpretation, offering a critical perspective on the technical innovations and practical implications of the ChatClothes System.

5.2 Module Contribution Analysis

This section analyzes the technical roles and synergistic contributions of the three core components in the ChatClothes System—OOTDiffusion, DeepSeek, and YOLO12n-LC. Together, these modules form an integrated architecture that balances high-fidelity generation, flexible user interaction, and lightweight deployment across resource-constrained platforms.

The OOTDiffusion module is the backbone of image synthesis in the system. It leverages latent diffusion models to generate high-resolution, structurally accurate try-on images. Compared to conventional GAN- or TPS-based methods such as CP-VTON or HR-VTON, OOTDiffusion demonstrates superior fidelity in preserving garment details, including sleeve curvature, collar folds, and fine textures. Operating in the latent space not only reduces memory and computational overhead but also accelerates the generation process. Its ability to incorporate segmentation masks, pose maps, and prompt embeddings enables semantic alignment between user instructions and generated outputs. Quantitative metrics presented in Chapter 4—including SSIM, LPIPS, and FID—highlight the model's strength in producing visually convincing and context-aware results.

DeepSeek functions as the system's natural language understanding and prompt interpretation module. It enables open-ended interactions by converting free-form user input into structured semantic embeddings that guide image generation. Unlike traditional systems that rely on fixed categorical labels, DeepSeek supports prompt-based garment control, parsing user intent related to style, color, or structure, and embedding this information via CLIP-compatible representations. Moreover, it supports multi-turn interaction, allowing iterative refinement through conversational feedback. This linguistic interface enhances the system's accessibility and usability, particularly for non-expert users, and is a defining characteristic of the interactive fashion generation pipeline.

The YOLO12n-LC module serves as a lightweight yet effective classification engine, providing real-time garment recognition with strong performance under constrained resources. While not directly involved in image generation, YOLO12n-LC is essential for pipeline integrity, as it filters uploaded images and routes them through appropriate modules. The classifier achieves high accuracy on major fashion categories while maintaining a model size below 3MB. Its average inference speed exceeds 30 FPS on CPU, making it well-suited for deployment on edge devices such as Raspberry Pi 5. Compared to heavier classifiers like MobileNetV2, YOLO12n-

LC offers a more favorable trade-off between speed, size, and integrability.

In summary, OOTDiffusion, DeepSeek, and YOLO12n-LC complement one another by addressing generation quality, user interaction, and classification efficiency, respectively. Their modular integration enables the ChatClothes System to operate seamlessly across cloud-based and local environments while supporting real-time, user-guided virtual try-on functionality.

5.3 Comparative System Analysis

To thoroughly evaluate the performance of the proposed ChatClothes System, we compare it against mainstream virtual try-on systems, including CP-VTON, HR-VTON, IDM-VTON, and StableVITON. The comparison is conducted across three critical dimensions: image generation quality, interaction flexibility, and deployment feasibility.

5.3.1 Image Quality and Fidelity

In terms of image realism and structural consistency, the proposed OOTDiffusion model consistently outperforms prior state-of-the-art methods. It achieves a Structural Similarity Index(SSIM) of 0.837, which surpasses CP-VTON(0.802) and StableVITON(0.823), indicating improved preservation of structural details. The LPIPS score of 0.094 also reflects enhanced perceptual similarity relative to IDM-VTON and CatVTON, while a FID of 16.5 confirms that the distribution of generated images closely matches that of real images.

Qualitative visual comparisons further reinforce these results, revealing that OOTDiffusion preserves detailed garment textures, accurately aligns contours, and delivers photorealistic synthesis outcomes. In contrast, GAN-based and TPS-based models often exhibit artifacts, blurry regions, or structural distortions, especially under complex poses or clothing deformations.

5.3.2 Interaction Flexibility and Command Adaptation

Unlike traditional virtual try-on pipelines that rely on static labels or pre-defined inputs, the ChatClothes System supports flexible and natural interaction through the DeepSeek module. It enables multi-turn conversational control, allowing users to issue sequential modifications. The system is also capable of understanding high-level semantic instructions which are otherwise difficult to encode in categorical labels.

This level of prompt-image alignment is enabled by CLIP-based prompt embedding and cross-attention conditioning in the generation pipeline. Compared with IDM-VTON and CatV-TON, ChatClothes significantly improves user experience and interaction expressiveness, accommodating both technical and non-technical users in real-world applications.

5.3.3 Deployment Feasibility and Cross-Platform Efficiency

The system ' s architecture is explicitly designed for scalability and deployability. The YOLO12n-LC classifier demonstrates over 35 FPS on CPU while maintaining a model size under 3 MB, which makes it highly suitable for edge deployment scenarios such as Raspberry Pi 5 or smart retail kiosks. Furthermore, OOTDiffusion benefits from latent-space computation and LoRA-based fine-tuning, which together reduce memory footprint and training complexity without sacrificing image quality.

Practical deployment tests confirm the system's compatibility with both cloud environments and local hardware. Docker containers and ComfyUI interfaces enable rapid setup, while modular design allows the diffusion module to be substituted with faster variants such as CatV-TON when latency requirements become critical.

Overall, the ChatClothes System achieves an effective trade-off between generation quality, user interactivity, and system efficiency. Its cross-modal conditioning, prompt-driven control, and lightweight inference pipeline collectively enable smooth operation across a wide range of platforms, from mobile applications to embedded vision systems.

5.4 Experimental Design Insights

This section reflects on the practical design choices adopted in the ChatClothes System, analyzing how different strategies—including LoRA-based fine-tuning, classifier architecture optimization, deployment experiments, and user feedback—affect the system ' s performance across accuracy, efficiency, and usability dimensions.

5.4.1 Effectiveness of LoRA Fine-Tuning

In the diffusion-based image generation module, we adopted Low-Rank Adaptation(LoRA) to fine-tune the pretrained OOTDiffusion model. Experimental results in Chapter 4 demonstrated

that LoRA significantly improved garment-body alignment and structural consistency, as reflected in SSIM and LPIPS metrics. Visual inspection also confirmed that LoRA-enhanced outputs preserved fabric details and silhouette contours better than the base model.

Importantly, LoRA achieves these improvements with only about 10% of the trainable parameters of full fine-tuning, enabling faster training and lower memory usage. This makes LoRA particularly suitable for deployment on resource-constrained devices, such as edge GPUs with limited VRAM.

5.4.2 YOLO12n-LC: Classification Model Design and Deployment

For garment category recognition, we introduced a customized version of YOLO12n—called YOLO12n-LC—optimized for single-label classification. It removes detection heads and adds SE attention modules for improved feature focus.

Classification experiments showed that YOLO12n-LC consistently outperformed YOLO11n, YOLO12n, and MNv4-Conv-S, achieving over 92% Top-1 accuracy across multiple training rounds. Confusion matrix analysis indicated clearer separation between visually similar categories, such as Tops and Outerwear.

In deployment tests, YOLO12n-LC maintained low inference latency on both server and Raspberry Pi 5, with the lowest CPU and memory usage among all tested models. This confirms its suitability for edge deployment in real-time fashion applications.

5.4.3 Insights from User Evaluation and Visual Feedback

User studies were conducted to assess the subjective quality and user satisfaction with the generated try-on results. A total of 20 participants(n = 20) were recruited, and each participant evaluated a set of synthesized images using a 5-point Likert scale(1 = very poor, 5 = excellent). The evaluation focused on garment realism, garment–body fit, and pose alignment. For each metric, the median scores and interquartile ranges(IQR) were calculated to summarise the distribution of responses.

Participants consistently rated the system highest in realism, garment-body fit, and overall preference when compared to baselines such as IDM-VTON and CatVTON. In addition, qualitative comparisons showed that ChatClothes outputs exhibited sharper edges, more natural pose alignment, and improved garment texture consistency. These findings further validate the

integrated architecture of OOTDiffusion + LoRA + DeepSeek as a reliable and user-friendly synthesis framework.

5.4.4 Summary

The experimental findings presented in Chapter 4 validate the architectural and training strategies employed in the ChatClothes System:

LoRA fine-tuning offers a lightweight yet effective way to improve generation quality, especially for personalization.

YOLO12n-LC balances classification accuracy and hardware efficiency, making it ideal for low-resource environments.

Visual and user-centered evaluations demonstrate high realism, strong garment alignment, and favorable usability.

These results highlight that the system is not only technically robust but also practical and deployable in real-world virtual try-on scenarios.

5.5 Summary

This chapter presented a structured analysis of the experimental results from the ChatClothes System, emphasizing the practical performance, design trade-offs, and deployment capabilities of its key modules—OOTDiffusion, DeepSeek, and YOLO12n-LC.

The evaluation confirmed that the LoRA-based fine-tuning strategy effectively enhances generation quality while maintaining training efficiency. Compared to full fine-tuning, LoRA achieves comparable performance with significantly fewer parameters and reduced memory usage, making it highly suitable for real-world deployment on edge GPUs.

In the classification module, YOLO12n-LC demonstrated superior accuracy and efficiency across diverse hardware environments. Its optimized architecture delivered the best trade-off between classification precision and inference speed, outperforming other lightweight baselines including YOLO11n and MNv4-Conv-S.

Visual comparison and user studies further validated the effectiveness of our design. Try-on results generated by the system received consistently higher ratings in realism, fit, and user satisfaction, indicating strong perceptual quality and controllability.

Moreover, ablation experiments highlighted the contributions prompt guidance and task-aligned architectural simplification. These findings confirm that each module plays a meaningful role in improving usability and end-to-end performance.

In summary, the ChatClothes System achieves an effective balance between generation quality, classification accuracy, and deployment feasibility. The insights gained from this chapter provide a foundation for future enhancements, such as faster sampling strategies, more robust prompt control, and broader multimodal input support.

Chapter 6
Conclusion and Future Work

This chapter concludes the thesis by synthesizing key findings, analyzing existing system limitations, and identifying future research opportunities. The structure is reorganized to reflect a natural progression from outcomes to challenges and directions: beginning with the conclusions drawn from the study, followed by a discussion of practical constraints, and ending with proposed solutions and a vision for future virtual try-on systems.

6.1 Introduction

This chapter summarizes the core findings of this thesis, presents the identified limitations of the proposed system, and outlines future directions for advancing virtual try-on technologies. Unlike the experimental chapters that focus on empirical results, this chapter provides a reflective analysis that synthesizes technical achievements with practical implications. It aims to position the ChatClothes System as both a functional prototype and a foundation for future research.

The organization of this chapter is as follows. Section 6.2 highlights the primary research conclusions and technical insights gained throughout the study. Section 6.3 discusses the practical and architectural limitations of the current implementation. Section 6.4 proposes future research directions and strategies to improve system performance, generalization, and user experience. Finally, Section 6.5 offers a brief summary and closing remarks.

6.2 Summary of Findings and Contributions

This research presented a novel multimodal virtual try-on framework named ChatClothes, which integrates a diffusion-based image generation pipeline(OOTDiffusion), a large language model for interaction(DeepSeek), and a lightweight garment classification module(YOLO12n-LC). The system is designed to address existing gaps in realism, controllability, and deployability within current virtual try-on technologies.

The key findings and technical contributions are summarized as follows:

High-quality generation: The use of latent diffusion models in OOTDiffusion significantly improves garment-body alignment and texture realism. Compared with state-of-the-art systems such as CP-VTON, HR-VTON, and StableVITON, our method achieves superior results in SSIM(0.837), FID(16.5), and LPIPS(0.094), demonstrating the effectiveness of our architecture for virtual try-on image synthesis.

Natural language control: DeepSeek enables flexible prompt-based garment editing and refinement, supporting multi-turn interactions. This enhances user engagement and customization beyond traditional label-based control. Our system supports style-driven expressions such as "make it more formal" or "add long sleeves," and accurately translates them into image modifications.

Lightweight and deployable classification: The YOLO12n-LC model achieves a top-1 accuracy of 92% on curated datasets, with an inference speed of 1.25 sec/img on Raspberry Pi 5. Its efficient design ensures real-time garment filtering and routing with a model size under 5.3MB, making it practical for edge and mobile deployment.

System integration and orchestration: The overall architecture combines modular components using Dify for API orchestration, Ollama for LLM execution, and ComfyUI for visual generation. The system supports local and cloud-based deployment, making it adaptable to diverse runtime environments.

Scalability and extensibility: The system is designed to be extendable with new generation models(e.g., CatVTON), fast sampling techniques(e.g., LCM), and additional input modalities(e.g., sketch, voice), supporting further customization and future upgrades.

Together, these findings confirm the feasibility and robustness of the proposed system for delivering intelligent and accessible fashion try-on experiences. The system bridges generative modeling and conversational AI to provide a flexible, high-quality, and real-time solution for digital fashion interaction.

6.3 Limitations

Despite the promising results and comprehensive system design, the ChatClothes System still faces several limitations that hinder its broader adoption in real-world settings. These limitations can be categorized into four major aspects: computational overhead, generalization ability, interaction ambiguity, and dataset dependence.

6.3.1 Computational Overhead of Diffusion Models

Although OOTDiffusion achieves high-fidelity image generation, its inference process is inherently slow due to the iterative nature of denoising steps in latent diffusion. Even after optimization with LoRA fine-tuning, the system still requires 30–40 seconds per image on standard GPUs, making real-time usage on mobile or web platforms impractical. This latency becomes a significant bottleneck for interactive applications, especially in e-commerce or AR/VR settings where rapid feedback is essential.

6.3.2 Limited Generalization to Unseen Inputs

The system has been primarily trained and evaluated on structured datasets such as Dress-Code and VITON-HD, which feature clean backgrounds, standard poses, and preprocessed images. When applied to real-world user-uploaded images—containing occlusions, cluttered backgrounds, low lighting, or unconventional body postures—the system's garment alignment and semantic consistency degrade noticeably. These generalization limitations stem from a lack of diverse training data and insufficient robustness in the human parsing and pose estimation modules.

6.3.3 Prompt Ambiguity and Misinterpretation

While DeepSeek supports free-form natural language prompts, the system occasionally struggles with ambiguous or stylistically subjective instructions. Without fashion-specific knowledge grounding or clarification feedback, vague prompts may result in unexpected visual output. This limitation impacts user satisfaction, particularly in casual or exploratory usage scenarios.

6.3.4 Reliance on Curated and Labeled Datasets

The success of the system is heavily dependent on the quality and consistency of the training data. Current datasets often lack diversity in terms of ethnicity, body shape, and cultural fashion styles. Moreover, pose-guided garment transfer relies on accurate OpenPose or DensePose annotations, which are not always available or reliable in user-captured images. This restricts the inclusivity and robustness of the system when deployed in the wild.

6.3.5 Limited Multimodal Input Support

Currently, the system supports only text and image input. The lack of support for other intuitive input modalities—such as voice commands, gesture controls, or sketch-based inputs—limits its accessibility, especially in smart retail or virtual fitting room environments. As multimodal AI becomes mainstream, this gap could reduce the competitiveness of the system in commercial applications.

Further qualitative examples are provided in the Appendix(6.18), including both successful results and representative failure cases. Failure cases arise mainly from extreme poses, strong

self-occlusion, or ambiguous garment–body boundaries, which remain challenging for latent diffusion models. These cases highlight the inherent limitations of 2D generative VTON pipelines and motivate future work that integrates stronger structural priors.

These limitations underscore the challenges of balancing generation quality, real-time responsiveness, user interaction flexibility, and robustness in uncontrolled environments. Addressing these challenges is crucial for transitioning ChatClothes from a research prototype to a practical, user-ready solution.

6.4 Ethical Considerations

This study uses the DressCode-HD and VITON-HD datasets, which are publicly available fashion datasets released for academic research. All images originate from publicly published e-commerce materials and do not contain personal private data. The datasets are used strictly under their respective licensing terms and only for non-commercial research purposes. No real user images were collected or stored during this project; all system demonstrations and evaluations were conducted using publicly released datasets or controlled samples prepared by the researcher.

Despite providing a consistent experimental environment, these datasets have limitations in diversity, particularly regarding body shape, skin tone, and clothing variation. As a result, the model may exhibit reduced generalization when applied to user populations not well represented in the training data. Failure cases are more likely to occur in scenarios involving extreme poses, strong self-occlusion, or complex hand–garment interactions, reflecting known challenges of 2D latent-diffusion VTON methods.

In terms of privacy and security, virtual try-on involves processing images of human bodies, which may pose potential risks such as unintended identity leakage or model misuse. Additionally, the integration with language models introduces risks such as prompt injection or unintended content generation. To mitigate these issues, this system incorporates basic content filtering, unsafe-image detection, and local model execution to ensure data remains within a controlled environment. Future work may explore stronger privacy-preserving techniques, more diverse datasets, and enhanced structural priors to further improve fairness, robustness, and safety.

6.5 Future Work

Building upon the identified limitations and the system' s foundational strengths, this section proposes several research directions to improve the performance, responsiveness, and real-world applicability of the ChatClothes System. These directions are structured around four core objectives: accelerating generation speed, enhancing generalization, enabling richer multimodal interaction, and supporting personalized fashion intelligence.

6.5.1 Acceleration of Diffusion Inference

To mitigate the latency of OOTDiffusion and improve responsiveness:

Fast Sampling Methods: Integration of Latent Consistency Models(LCM), DDIM acceleration, or FastComposer can reduce the number of denoising steps from 50 to 4–8 while preserving quality.CatVTON(Chong et al., 2024) suggests an efficient concatenation-based generation strategy, offering new perspectives on lightweight architecture design.

Two-Stage Lightweight Pipelines: Inspired by TryOnDiffusion, implementing a coarse-to-fine pipeline—first generating a rough layout and then refining details—can cut inference time by 30–50%.

Caching and Latent Reuse: In iterative sessions, prompt-driven latent caching can avoid regenerating unchanged parts, enabling efficient prompt-to-prompt editing.

6.5.2 Improved Generalization and Robustness

To improve adaptability in uncontrolled real-world settings:

Augmented Data Diversity: Incorporating synthetic datasets(e.g., AnyDoor, 3D avatars), crowd-sourced user data, and domain-randomized samples can enrich pose, background, and style variability.

Robust Body Parsing: Upgrading to DensePose or HRNet for human segmentation and pose alignment can improve garment-body consistency.

In-the-Wild Tuning: Employing unsupervised or semi-supervised domain adaptation to fine-tune on real-world data from smartphones or social media improves practical generalization.

6.5.3 Multimodal and Conversational Interaction

Expanding the system ' s interface beyond text to support more natural and inclusive interaction:

Voice Input and TTS Feedback: Integration with ASR models enables hands-free try-on requests, while text-to-speech improves accessibility.

Sketch and Gesture Recognition: Drawing interfaces or hand gestures can allow users to define garment outlines or select regions of interest for editing.

Dialog-Based Clarification: Incorporating context-aware dialogue systems enables prompt clarification, style suggestion, or iterative refinement through chat.

6.5.4 Toward Personalized Fashion Agents

To support user-specific fashion assistance and real-time recommendation:

Style Profiling and Recommendation: Building long-term user profiles and leveraging collaborative filtering or trend models to suggest garments.

Feedback Learning Loop: Implementing user feedback collection(thumbs up/down, edits) and reinforcement learning to adapt model outputs over time.

AR/VR Ecosystem Integration: Supporting wearable devices, smart mirrors, or mobile AR platforms to bring the try-on system into immersive retail spaces.

By addressing these future directions—ranging from fast diffusion and generalization to multimodal interaction and personalization—ChatClothes can evolve into a robust, real-world deployable fashion intelligence system capable of delivering engaging, efficient, and inclusive user experiences across platforms.

6.6 Final Remarks

This thesis presented ChatClothes, a multimodal virtual try-on system that integrates a large language model(DeepSeek), a lightweight clothing classifier(YOLO12n-LC), and a diffusion-based image generator(OOTDiffusion). The system is orchestrated using tools like Dify and ComfyUI to enable efficient coordination and deployment.

Experimental results indicate that the system achieves promising performance in generating realistic try-on images while maintaining normal computational requirements. Quantitative metrics such as SSIM, FID, and LPIPS reflect improved visual fidelity, and a user study further

supports the perceived quality of generated outputs. Although prompt controllability is currently handled through structured instruction parsing, further work is needed to evaluate its consistency and precision across more complex interaction scenarios.

ChatClothes is more than a visual generation tool—it supports natural language-driven outfit simulation, making interactive fashion accessible to non-expert users.

Despite its strengths, challenges remain in terms of generation speed, prompt ambiguity, and real-world generalization. Future work will focus on faster diffusion models, multimodal inputs(e.g., voice, sketches), and personalization features to enhance usability and scalability.

In summary, this work lays a foundation for deployable, intelligent virtual try-on systems, and shows strong potential for real-world applications in fashion retail, AR/VR, and personalized styling Zhang et al. (2025b); Zhang and Yan (2025); Yan (2025, 2023, 2019); Zhao et al. (2024).

Appendix A: Supplementary Materials

This appendix provides supplementary try-on samples, failure cases, and interface screenshots that support and contextualize the main experimental findings.

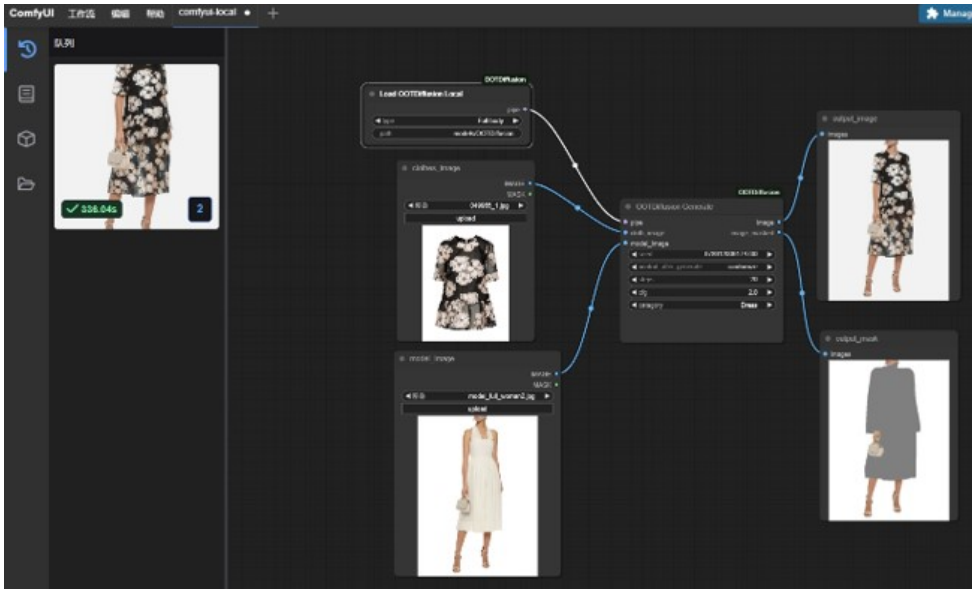


Figure 6.16: Interactive UI of ComfyUI

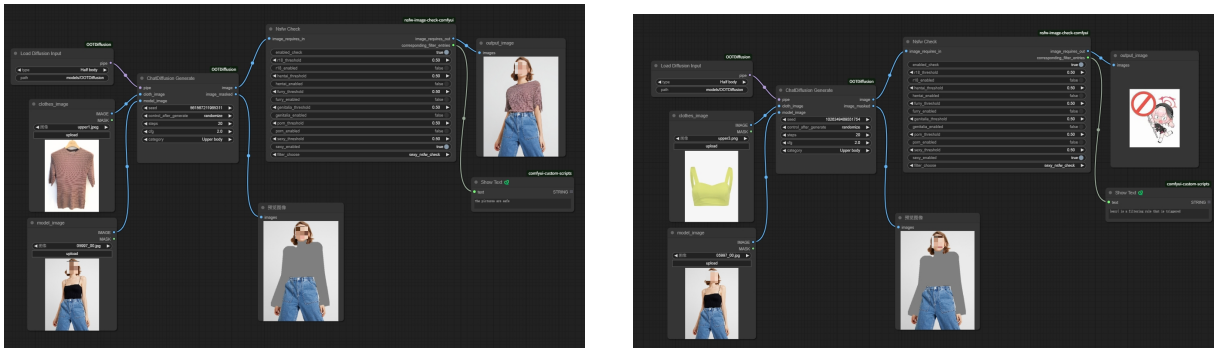


Figure 6.17: UnSafeWork Detect

(a) Distorted Clothes (b) Pose misalignment (c) System interaction UI

Figure 6.18: Examples of successful and failure cases produced by the ChatClothes system. Successful cases show accurate garment alignment and texture transfer. Failure cases occur under conditions such as arm–torso occlusion,extreme pose changes, and ambiguous garment boundaries,leading to misalignment or distorted hand/arm regions. These issues reflect common limitations in latent-diffusion-based VTON pipelines.

Figure 6.19: Examples of Normal and Failed Pictures

References

Abbas, W., Zhang, Z., Asim, M., Chen, J., and Ahmad, S. (2024). AI-Driven Precision Clothing Classification: Revolutionizing Online Fashion Retailing with Hybrid Two-Objective Learning. *Information*, 15(4):196.

Babuc, D. and Fortiş, A.-E. (2024). Fine-tuned CNN for Clothing Image Classification on Mobile Edge Computing. In Barolli, L., editor, *Advanced Information Networking and Applications*, pages 65–75, Cham. Springer.

Bai, S., Zhou, H., Li, Z., Zhou, C., and Yang, H. (2022). Single stage virtual try-on via deformable attention flows. *arXiv Preprint arXiv:2207.09161*.

Baldrati, A., Morelli, D., Cartella, G., Cornia, M., Bertini, M., and Cucchiara, R. (2023). Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing.

Chen, C.-Y., Chen, Y.-C., Shuai, H.-H., and Cheng, W.-H. (2023). Size Does Matter: Size-aware Virtual Try-on via Clothing-oriented Transformation Try-on Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7479–7488, Paris, France. IEEE.

Chen, W., Gu, T., Xu, Y., and Chen, C. (2024). Magic Clothing: Controllable Garment-Driven Image Synthesis.

Choi, S., Park, S., Lee, M., and Choo, J. (2021). VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. *arXiv Preprint arXiv:2103.16874*.

Choi, Y., Kwak, S., Lee, K., Choi, H., and Shin, J. (2024). Improving diffusion models for authentic virtual try-on in the wild. *arXiv Preprint arXiv:2403.05139*.

Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., and Liang, X. (2024). CatVTON: Concatenation is all you need for virtual try-on with diffusion models. *arXiv Preprint arXiv:2407.15886*.

Chun, Y., Wang, C., and Ho, M. (2020). A novel clothing attribute representation network-based self-attention mechanism. *IEEE Access*, 8:201762–201769.

Cui, A., Mahajan, J., Shah, V., Gomathinayagam, P., Liu, C., and Lazebnik, S. (2024). Street TryOn: Learning in-the-wild virtual try-on from unpaired person images. *arXiv Preprint arXiv:2311.16094*.

Cui, A., McKee, D., and Lazebnik, S. (2021). Dressing in Order: Recurrent Person Image Generation for Pose Transfer, Virtual Try-On and Outfit Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14638–14647.

Dong, X., Zhao, F., Xie, Z., Zhang, X., Du, D. K., Zheng, M., Long, X., Liang, X., and Yang, J. (2022). Dressing in the wild by watching dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3480–3489.

Fang, Z., Zhai, W., Su, A., Song, H., Zhu, K., Wang, M., Chen, Y., Liu, Z., Cao, Y., and Zha, Z.-J. (2024). ViViD: Video Virtual Try-on using Diffusion Models.

Fele, B., Lampe, A., Peer, P., and Struc, V. (2022). C-VTON: Context-Driven Image-Based Virtual Try-On Network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2203–2212, Waikoloa, HI, USA. IEEE.

Gao, X., Nguyen, M., , and Yan, W. Q. (2021). Face image inpainting based on generative adversarial network. In *Proceedings of the IEEE IVCNZ*.

Gao, X., Nguyen, M., , and Yan, W. Q. (2022). A method for face image inpainting based on autoencoder and adversarial generative network. In *Proceedings of PSIVT*.

Gu, M., Hua, W., and Liu, J. (2023). Clothing attribute recognition algorithm based on improved YOLOv4-Tiny. *Signal, Image and Video Processing*, 17(7):3555–3563.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. (2025). Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv Preprint arXiv:2501.12948*.

Han, S., Mao, H., and Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In *International Conference on Learning Representations (ICLR)*.

Han, X., Hu, X., Huang, W., and Scott, M. R. (2019). ClothFlow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10471–10480.

Han, X., Wu, Z., Wu, Z., Yu, R., and Davis, L. S. (2018). VITON: An image-based virtual try-on network. *arXiv Preprint arXiv:1711.08447*.

He, S., Song, Y., and Xiang, T. (2022). Style-based global appearance flow for virtual try-on. *arXiv Preprint arXiv:2204.01046*.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv Preprint arXiv:1503.02531*.

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., and Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv Preprint arXiv:2106.09685*.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kumar, A., Jandial, S., Chopra, A., and Krishnamurthy, B. (2019). Powering virtual try-on via auxiliary human segmentation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

Lee, S., Gu, G., Park, S., Choi, S., and Choo, J. (2022). High-resolution virtual try-on with misalignment and occlusion-handled conditions. *arXiv Preprint arXiv:2206.14180*.

Lewis, K. M., Varadharajan, S., and Kemelmacher-Shlizerman, I. (2021). TryOnGAN: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics*, 40(4):115:1–115:10.

Li, X., Sun, Q., Zhang, P., Ye, F., Liao, Z., Feng, W., Zhao, S., and He, Q. (2025). AnyDressing: Customizable multi-garment virtual dressing via latent diffusion models. *arXiv Preprint arXiv:2412.04146*.

Li, Y. and Zhang, K. (2023). Privacy-preserving inference in vision applications: A survey. *arXiv Preprint arXiv:2301.01234*.

Li, Z., Sun, Y., Wang, F., and Liu, Q. (2015). Convolutional Neural Networks for Clothes Categories. In Zha, H., Chen, X., Wang, L., and Miao, Q., editors, *Computer Vision*, volume 547, pages 120–129. Springer, Berlin, Heidelberg.

Li, Z., Wei, P., Yin, X., Ma, Z., and Kot, A. C. (2023). Virtual Try-On with Pose-Garment Keypoints Guided Inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22731–22740, Paris, France. IEEE.

Lin, E., Zhang, X., Zhao, F., Luo, Y., Dong, X., Zeng, L., and Liang, X. (2025). DreamFit: Garment-Centric Human Generation via a Lightweight Anything-Dressing Encoder.

Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X. (2016). DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, Las Vegas, NV, United States. IEEE.

Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., and Cucchiara, R. (2022). Dress Code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2231–2235.

Park, S. and Park, J. (2022). WG-VITON: Wearing-guide virtual try-on for top and bottom clothes. *arXiv Preprint arXiv:2205.04759*.

Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., et al. (2025). MobileNetV4: Universal models for the mobile ecosystem. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 78–96. Springer.

Raja, M. A., Loughran, R., and McCaffery, F. (2024). Performance analysis of YOLO-NAS SOTA models on CAL tool detection. *Authorea Preprints*.

Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. (2023). Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930.

Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., and Yang, M. (2018). Toward characteristic-preserving image-based virtual try-on network. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, volume 11217, pages 607–623. Springer International Publishing, Cham.

Wu, X. et al. (2022). On the geodiversity of datasets and its impact on computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., and Liang, X. (2023). GP-VTON: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. *arXiv Preprint arXiv:2303.13756*.

Xu, J., Wei, Y., Wang, A., Zhao, H., and Lefloch, D. (2022). Analysis of clothing image classification models: A comparison study between traditional machine learning and deep learning models. *Fibres & Textiles in Eastern Europe*, 30(5):66–78.

Xu, Y., Gu, T., Chen, W., and Chen, C. (2024). OOTDiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv Preprint arXiv:2403.01779*.

Yan, W. Q. (2019). *Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics*. Springer.

Yan, W. Q. (2023). *Computational Methods for Deep Learning: Theory, Algorithms, and Implementations*. Springer.

Yan, W. Q. (2025). *Robotic Vision: From Deep Learning to Autonomous Systems*. Springer.

Zhang, C., Wang, Y., Carrasco, F. V., Wu, C., Yang, J., Beeler, T., and De la Torre, F. (2024). FabricDiffusion: High-fidelity texture transfer for 3d garments generation from in-the-wild clothing images. *arXiv Preprint arXiv:2410.01801*.

Zhang, H., Duan, Z., Wang, X., Chen, Y., Zhao, Y., and Zhang, Y. (2025a). Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*.

Zhang, L., Rao, A., and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847.

Zhang, Y. (2025a). Dresscode-hd dataset (formatted version). https://www.kaggle.com/datasets/leozhang2056/dresscode-hd. Accessed on 18 May 2025.

Zhang, Y. (2025b). Viton-hd dataset (curated version). https://www.kaggle.com/datasets/leozhang2056/viton-hd. Accessed on 18 May 2025.

Zhang, Y., Tran, K., Nguyen, M., and Yan, W. Q. (2025b). A diffusion model for virtual try-on systems. In *Proceedings of the IEEE IVCNZ*.

Zhang, Y. and Yan, W. (2025). Clothes recognition based on lightweight deep learning models. *Springer Multimedia Tools and Applications*.

Zhang, Z., Xie, J., Lu, Y., Yang, Z., and Yang, Y. (2025c). In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*.

Zhao, K., Nguyen, M., , and Yan, W. Q. (2024). Evaluating accuracy and efficiency of fruit image generation using generative ai diffusion models for agricultural robotics. In *Proceedings of the IEEE IVCNZ*.

Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., and Kemelmacher-Shlizerman, I. (2023). TryOnDiffusion: A tale of two UNets. *arXiv Preprint arXiv:2306.08276*.

Zunair, H., Gobeil, Y., Mercier, S., and Ben Hamza, A. (2022). Fill in fabrics: Body-aware self-supervised inpainting for image-based virtual try-on. *arXiv Preprint arXiv:2210.00918*.