Ensino que te conecta

BACHARELADO EM

CIÊNCIA DA COMPUTAÇÃO

A MULTI-AGENT ARCHITECTURE FOR MULTIMODAL FASHION RECOMMENDATION

JOSÉ GUILHERME DOS SANTOS VASCONCELOS

Brasília - DF, 2025

JOSÉ GUILHERME DOS SANTOS VASCONCELOS

A MULTI-AGENT ARCHITECTURE FOR MULTIMODAL FASHION RECOMMENDATION

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção de grau de Bacharel em Ciência da Computação, pelo Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP).

Orientador
Me. Klayton Rodrigues de Castro

Brasília - DF, 2025

Código de catalogação na publicação – CIP

V331m Vasconcelos, José Guilherme dos Santos

A Multi-Agent Architecture for Multimodal Fashion Recommendation / José Guilherme dos Santos Vasconcelos. — Brasília: Instituto Brasileiro Ensino, Desenvolvimento e Pesquisa, 2025.

49 f. : il.

Orientador: Prof. Me. Klayton Rodrigues de Castro

Monografia (Graduação em Ciência da Computação) - Instituto Brasileiro Ensino, Desenvolvimento e Pesquisa – IDP, 2025.

1. Moda. 2. Inteligência artificial. 3. Ambiente digital. I. Título

CDD 006.3

Elaborada por Biblioteca Ministro Moreira Alves

JOSÉ GUILHERME DOS SANTOS VASCONCELOS

A MULTI-AGENT ARCHITECTURE FOR MULTIMODAL FASHION RECOMMENDATION

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção de grau de Bacharel em Ciência da Computação, pelo Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP).

Aprovado em 12/12/2025

Banca Examinadora

gov.br Documento assinado digitalmente
KLAYTON RODRIGUES DE CASTRO
Data: 19/12/2025 16:34:17-0300
Verifique em https://validar.iti.gov.br

Me. Klayton Rodrigues de Castro - Orientador

gov.br Documento assinado digitalmente
LUCAS MAURICIO CASTRO E MARTINS
Data: 19/12/2025 16:52:13-0300
Verifique em https://validar.iti.gov.br

Me. Lucas Maurício Martins e Castro - Examinador interno

gov.br Documento assinado digitalmente
LORENA DE SOUZA BEZERRA BORGES
Data: 19/12/2025 17:01:30-0300
Verifique em https://validar.iti.gov.br

Me. Lorena Bezerra de Souza Borges - Examinadora interna

ABSTRACT

This work investigates the challenges of personalized fashion recommendation in digital environments, with emphasis on semantic consistency, interpretability, and structured representation of user style. Traditional approaches based on monolithic language model architectures often suffer from context dilution and limited transparency when handling multimodal inputs and complex interaction histories. As an alternative, this study proposes a hierarchical multi-agent architecture in which specialized agents independently perform visual analysis, user profile modeling, and recommendation synthesis. Agent orchestration is inspired by principles of the Model Context Protocol (MCP), aiming to provide explicit control over context flow and reasoning processes. The experimental evaluation was conducted through a direct comparison with a monolithic approach using the FashionRec dataset. Results were analyzed using semantic similarity metrics based on Sentence-BERT and a qualitative assessment grounded in the LLM-as-a-Judge paradigm. The experiments indicate that the hierarchical architecture achieves higher personalization consistency, greater semantic stability, and improved explanatory clarity in recommendations. Although reasoning decomposition incurs higher computational cost, the results demonstrate that functional separation among agents mitigates context loss and supports style-sensitive recommendations, constituting a promising approach for the development of more interpretable, coherent, and adaptive fashion recommendation systems.

Keywords: Recommendation systems. Fashion. Multi-agent architecture. Multimodality. Interpretability..

RESUMO

Este trabalho investiga os desafios da recomendação personalizada de moda em ambientes digitais, com foco na consistência semântica, interpretabilidade e representação estruturada do estilo do usuário. Abordagens tradicionais baseadas em arquiteturas monolíticas de modelos de linguagem tendem a sofrer com diluição de contexto e baixa transparência ao processar informações multimodais e históricos complexos de interação. Como alternativa, é proposta uma arquitetura hierárquica multiagente, na qual agentes especializados realizam, de forma desacoplada, a análise visual, a modelagem do perfil do usuário e a síntese da recomendação. A orquestração entre os agentes é inspirada em princípios do Model Context Protocol (MCP), visando maior controle sobre o fluxo de contexto e o processo de raciocínio. A avaliação experimental foi conduzida por meio de uma comparação direta com uma abordagem monolítica, utilizando o conjunto de dados FashionRec. Os resultados foram analisados com base em métricas de similaridade semântica (Sentence-BERT) e em uma avaliação qualitativa baseada no paradigma LLM-as-a-Judge. Os experimentos indicam que a arquitetura hierárquica apresenta maior consistência de personalização, estabilidade semântica e clareza explicativa nas recomendações. Embora a decomposição do raciocínio implique maior custo computacional, os resultados demonstram que a separação funcional entre agentes reduz efeitos de perda de contexto e favorece recomendações mais sensíveis ao estilo em uma abordagem promissora para obtenção de sistemas de recomendação de moda mais interpretáveis, coerentes e adaptativos.

Palavras-chave: Sistemas de recomendação. Moda. Arquitetura multiagente. Multimodalidade. Interpretabilidade..

LIST OF FIGURES

LIST OF TABLES

CONTENTS


1

1 INTRODUCTION

1.1 BACKGROUND AND CONTEXT

Fashion e-commerce is one of the most relevant segments of global digital retail. The sector generated more than US$820 billion in 2023, with projections that exceed the US$1 trillion mark by 2027 [1]. In the Brazilian context, the fashion category remains among the leaders in national e-commerce, driven by the growing adoption of consumption via mobile devices, the integration between digital platforms, social networks and marketplaces, as well as the search for practicality in shopping experiences. The Brazilian fashion e-commerce market reached a volume of US$10.38 billion in 2024, with expected annual growth of between 10% and 15% in the coming years [2]. This scenario of accelerated expansion highlights the need for innovative technological solutions that provide not only transactional efficiency, but also personalized experiences that meet consumers' stylistic preferences, in the face of an increasingly broad, dynamic and competitive digital catalog.

Despite the wide range of products on offer, this abundance can generate counterproductive effects, such as the so-called paradox of choice, in which an excess of alternatives leads to indecision, frustration and reduced consumer satisfaction [3]. In the context of digital fashion, this translates into the difficulty of finding pieces that reflect the user's personal style among thousands of available options. About this challenge, solutions based on artificial intelligence (AI) have established themselves as essential tools for personalizing the online shopping experience. Intelligent recommendation systems, capable of interpreting aesthetic preferences and suggesting relevant items, have become a competitive differentiator. However, many of these systems still operate based on historical purchase data, generic filters or rigid categories, without capturing subjective nuances such as style, occasion or cultural context.

This work is situated within this context and proposes the use of multimodal agents for clothing recommendation based on personal style, addressing the demands of contemporary fashion e-commerce for advanced, responsive, and meaningful personalization. Consequently, the need for more sophisticated models capable of jointly understanding language and visual information becomes evident. Despite the widespread adoption of recommender systems in fashion e-commerce, most existing solutions re-

main primarily optimized for transactional efficiency rather than for capturing the subjective and contextual nature of personal style. While current systems can effectively suggest items based on popularity, similarity, or past interactions, they often fail to model style as a coherent, interpretable, and evolving concept. As a result, recommendations may appear relevant at a surface level yet lack deeper alignment with the user's aesthetic identity, cultural background, or situational preferences. This gap between operational recommendation performance and meaningful stylistic understanding constitutes a central challenge for contemporary fashion e-commerce platforms and motivates the problem addressed in this research.

1.2 PROBLEM STATEMENT AND RESEARCH HYPOTHESES

In the context of fashion e-commerce, recommender systems have become essential tools for managing the overwhelming variety of options and delivering personalized shopping experiences. However, despite recent advances in multimodal agents powered by language and vision models, such as FashionM3 [4], a significant limitation remains: the challenge of capturing and representing the concept of style in an explicit, structured, and adaptive manner.

Current approaches often rely on implicit inferences derived from user interaction histories and visual or semantic similarity metrics. While effective for basic personalization, these methods fall short of modeling fashion style as an abstract, culturally informed, and ever-evolving construct. The intelligence of these systems is typically embedded within a single, monolithic model, making it difficult to isolate, update, or explain its understanding of a specific aesthetic.

Moreover, the stylistic knowledge in such systems tends to be static and prone to 'context dilution' when processing complex multimodal queries. Monolithic models often struggle to balance immediate visual inputs with long-term user history, frequently defaulting to generic recommendations that fail to capture the unique, long-tail stylistic preferences of individual users. This limitation hinders the system's ability to adapt to the specific, evolving tastes of each user, resulting in a 'one-size-fits-all' experience.

This research departs from monolithic recommendation models and explores a framework based on a system of specialized style agents. Each agent is responsible for representing and reasoning over specific fashion concepts, such as style attributes, visual patterns, or contextual preferences. From an implementation perspective, the proposed architecture is informed by general principles of modular orchestration and controlled context management commonly discussed in recent agent-based AI. In particular, it draws conceptual inspiration from ideas associated with the Model Context Protocol (MCP), which emphasizes structured interaction boundaries between models and external resources [5]. Its high-level principles, such as role separation, explicit context ownership and mediated coordination are used as design guidelines to support

modularity and reduce uncontrolled context accumulation.

Within this architecture, individual style agents maintain localized operational contexts and can be independently refined or extended using multimodal data. This design choice aims to enhance adaptability and interpretability, as recommendations can be traced back to specific agents and fashion concepts, without relying on opaque internal representations of a single monolithic model.

This research is justified by its potential to generate significant contributions across multiple domains. From a theoretical perspective, it investigates an alternative paradigm for modeling subjective and evolving concepts, such as fashion style, by exploring distributed reasoning through specialized agents rather than relying on monolithic representations. Methodologically, the work examines how general principles of modular orchestration and explicit context management—commonly discussed in recent agent-based AI systems can be applied to multimodal recommendation tasks.

From a practical standpoint, the proposed approach has the potential to enhance personalization in fashion e-commerce by providing recommendations that are more aligned with users' stylistic identities and easier to explain. Technically, the design contributes to system robustness and transparency, mitigating common issues of context dilution and unpredictability in large generative models.

The proposed architecture follows general orchestration principles inspired by recent agent-based AI frameworks, including task decomposition, role specialization, and coordinator-mediated interaction. Rather than implementing the full Model Context Protocol (MCP), these principles are adopted at an abstract level to investigate whether modular, agent-oriented designs yield measurable improvements in recommendation quality when compared to monolithic baselines.

Based on the identified limitations of monolithic architectures, this study evaluates the following research hypotheses:

- **H1 (Personalization)**: A hierarchical multi-agent architecture improves recommendation personalization compared to monolithic baselines by explicitly modeling user preferences and distributing reasoning across specialized agents.
- **H2 (Semantic Robustness)**: Decomposing the recommendation workflow into specialized agents reduces semantic inconsistencies and context dilution commonly observed in monolithic multimodal models.
- **H3 (Efficiency Trade-off)**: Under increasing multimodal contextual complexity, a hierarchical multi-agent architecture achieves higher inference stability and recommendation quality than a monolithic architecture by reallocating computational effort toward structured reasoning.

1.3 OBJECTIVES AND STRUCTURE OF THE WORK

The main goal of this research is to design and experimentally evaluate a style-aware clothing recommendation framework based on a hierarchical multi-agent architecture. The specific goals of this work are as follows:

- Investigate the limitations of monolithic multimodal recommender systems in capturing the dynamic, contextual, and subjective nature of fashion style.
- Design and analyze a hierarchical multi-agent architecture in which specialized agents explicitly encapsulate distinct reasoning capabilities, including visual analysis, style profiling, and final item recommendation.
- Implement and validate a functional prototype that operationalizes MCP-inspired orchestration principles using the OpenAI Agents SDK, enabling controlled coordination and communication among agents without relying on the full MCP protocol.
- Evaluate the proposed framework in terms of personalization quality, interpretability, semantic consistency, and robustness, comparing its behavior against a monolithic baseline under the same generative backbone.

This chapter presents the contextualization of the topic, the definition of the research problem, the justification, the objectives, and the guiding hypotheses. The remainder of this work is organized as follows:

- **Chapter 2 – Literature Review**: Reviews the relevant literature on fashion recommender systems, multimodal learning, multi-agent architectures, and representative state-of-the-art approaches in fashion recommendation.
- **Chapter 3 – Methodology**: Details the experimental design, the proposed hierarchical multi-agent architecture, the implementation of agents inspired by MCP principles, and the evaluation metrics.
- **Chapter 4 – Results and Discussion**: Presents the comparative analysis between the proposed multi-agent system and a monolithic baseline, validating the hypotheses through quantitative and qualitative evidence.
- **Chapter 5 – Conclusion**: Summarizes the research findings, highlights the achievement of objectives, discusses limitations related to computational cost and sample size, and proposes directions for future work.

2

2 LITERATURE REVIEW

2.1 THEORETICAL FOUNDATIONS

This section presents the theoretical foundations that support the proposed architecture, outlining the principles of recommendation systems, multimodal learning, generative models, and autonomous agents as applied to personalized fashion recommendation.

2.1.1 Recommendation Systems

Recommendation systems are based on algorithms capable of suggesting items according to user preferences. Traditional approaches include collaborative filtering, content-based filtering, and hybrid systems [6]. These systems function as an extension of the social recommendation process, aggregating individual opinions and delivering them in a personalized manner.

Collaborative filtering, one of the most widely used methods, can be based on user similarity (user-based) or item similarity (item-based), employing metrics such as cosine similarity, correlation, and adjusted cosine. Predictions can be made through weighted summation or regression. Although originally designed in a centralized manner, such systems also show potential for exploration under a multi-agent perspective, in which autonomous agents could dynamically adapt to user preferences in distributed environments [7].

Over the past decade, advances in Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) have significantly expanded the capabilities of collaborative filtering [8]. These approaches enable models to capture complex user–item interaction patterns, learn high-dimensional latent representations, and integrate contextual and temporal signals [9].

A particularly impactful shift is the rise of Generative AI, which introduced models capable not only of classifying or retrieving information but also of generating coherent and context-aware content—ranging from product descriptions to stylistic suggestions. Large Language Models (LLMs), such as GPT and LLaMA, exemplify this new generation of systems that can understand and produce fluent natural language. When coupled with Retrieval-Augmented Generation (RAG) mechanisms [10], these mod-

els achieve greater factual accuracy by incorporating external knowledge into their responses.

The integration of Large Language Models (LLMs) with multimodal capabilities—such as vision-language models like CLIP [11] and BLIP [12]—further expands their potential in domains like fashion, where both textual and visual understanding are essential for personalized recommendations. These models enable the alignment of image and text representations in a shared embedding space, facilitating contextual interpretations of style, color, and composition.

Building upon these advances, this work proposes a multi-agent architecture that leverages generative reasoning and multimodal embeddings to enhance the relevance, diversity, and transparency of clothing recommendations.

The evolution of recommendation systems reflects a progressive increase in representational complexity, moving from explicit preference modeling toward learned semantic and stylistic representations. Figure 1 summarizes this conceptual progression.

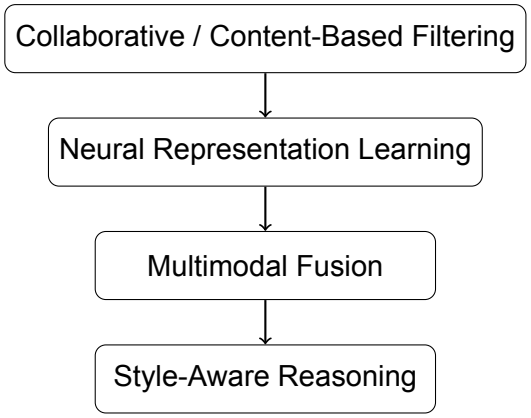

Figure 1: Conceptual evolution of recommendation systems from classical filtering approaches to style-aware multimodal reasoning.

This progression highlights how contemporary fashion recommendation systems increasingly rely on latent representations and multimodal signals to capture subjective notions such as style and compatibility.

2.1.2 Neural Architectures for Representation Learning

Modern recommendation systems increasingly rely on the ability to learn abstract, high-dimensional representations of user preferences and item characteristics. This capability, known as representation learning, is essential for modeling latent features that go beyond explicit ratings or categories—such as aesthetic style, visual coherence, and contextual fit. Among the most effective tools for this task are Artificial Neural Networks (ANNs), whose evolution has profoundly influenced the design of intelligent, data-driven recommendation frameworks [13].

Inspired by biological neurons, the Perceptron was originally conceived to investigate how the brain encodes information. In this connectionist model, knowledge is not stored as "copies" of stimuli, but rather emerges from the connections formed between computational units. Although simple, the Perceptron modeling introduced the use of neural networks in recognition and classification tasks [14]. The Perceptron's convergence procedure required manually predefined inputs ("feature analyzers"), which made it impractical for learning complex internal representations. As a result, linearly non-separable problems remained beyond its reach [14].

The publication by Rumelhart, Hinton, and Williams introduced the backpropagation algorithm [15], enabling the training of networks with hidden layers by iteratively adjusting weights to minimize output error. With this advancement, internal units began to extract abstract features, enabling solutions to more sophisticated tasks—such as symmetry detection. This evolution laid the foundation for deep neural architectures that now support multimodal recommendation systems, including those in the fashion domain.

With advances in computational resources and the growth of available data, more sophisticated architectures emerged, such as Transformers, introduced by Vaswani et al. [16]. These models replaced recurrence mechanisms with self-attention, allowing for parallel sequence processing and the modeling of long-range dependencies with high efficiency. Initially designed for natural language tasks, Transformers quickly expanded into multimodal applications, exemplified by models like CLIP, Florence, and BLIP, which project text and images into a unified embedding space. These models are particularly relevant in fashion recommendation scenarios, as they can accurately capture both visual context and stylistic semantics [11].

Building on this foundation, several neural architectures have been developed to address specific tasks such as image analysis, sequential modeling, and multimodal integration. The most prominent types include:

- **CNNs (Convolutional Neural Networks)**: they have become essential for visual representation in recommendation systems, particularly after the breakthrough of AlexNet by Krizhevsky et al. [17]. This deep architecture, composed of five convolutional layers and three fully connected layers, achieved unprecedented performance in the ImageNet ILSVRC-2012 competition. Techniques such as ReLU activation, local response normalization, dropout, and data augmentation contributed to faster training and reduced overfitting. The network was trained on multiple GPUs with millions of parameters, solidifying the role of CNNs in complex computer vision tasks.

- **RNNs (Recurrent Neural Networks) and Long Short-Term Memory networks (LSTMs)**: These architectures specialize in sequential data processing and are

used to model user behavior over time, such as click histories, views, or purchases. LSTMs, introduced by Hochreiter and Schmidhuber [18], overcome the limitations of basic RNNs by preserving long-term dependencies and are widely applied in sequential prediction tasks. In recommendation contexts, models like GRU4Rec use such architectures to predict users' next interactions with high accuracy [19].

- **Transformers**: As used in models like CLIP, Florence, and BLIP, Transformers enable multimodal recommendation systems by integrating text and image representations. Their ability to capture visual and stylistic semantics makes them particularly suitable for context-aware fashion recommendation [11].

- **GANs (Generative Adversarial Networks)**: Proposed by Goodfellow et al. [20], GANs consist of two networks trained competitively: a generator that creates synthetic samples, and a discriminator that evaluates whether those samples are real or generated. This minimax game leads the generator to produce increasingly realistic outputs, and GANs have become widely used in image synthesis tasks. In fashion, GANs have been employed to generate visually compatible outfits, such as in the OutfitGAN model. More recently, FashionDPO demonstrated how Direct Preference Optimization can fine-tune generative models based on explicit human preferences, improving both coherence and personalization in recommendations [21, 22].

- **KANs (Kolmogorov–Arnold Networks)**: Proposed by Liu et al. [23], KANs replace fixed activation functions in nodes (as in traditional MLPs) with learnable univariate functions on edges, typically parameterized as splines. Inspired by the Kolmogorov–Arnold representation theorem, KANs combine local precision with model interpretability, achieving high representational power while scaling better than MLPs in several tasks. Although still experimental and not yet widely adopted in production-grade recommender systems, KANs are considered a promising alternative to standard multilayer perceptrons, particularly in scenarios requiring compact and interpretable models.

2.1.3 Multimodal Learning and Style-Sensitive Recommendation

Multimodal learning models aim to create unified and enriched representations from multiple data sources, such as images, text, and audio. The central premise is that combining different modalities enables a deeper and more robust understanding than what would be possible using a single source of information. This section traces the evolution of such models—from early foundational works to the state-of-the-art architectures in vision and language—focusing on their applicability to complex tasks such as fashion recommendation.

A seminal contribution in this field is the work by Ngiam et al. [24], which explored how deep learning architectures could be extended to discover correlated representations across modalities. Inspired by the human ability to integrate sensory stimuli—as in audiovisual speech recognition, where lip movements complement sound—the authors aimed to overcome the limitations of manual feature engineering.

The initial approaches investigated by Ngiam et al. (2011) employed Restricted Boltzmann Machines (RBMs), evolving from separately trained RBMs for each modality to bimodal Deep Belief Networks (DBNs) trained over concatenated data. However, these architectures showed limitations in modeling complex, nonlinear correlations.

To address these shortcomings, the authors proposed the use of Deep Autoencoders, introducing two main variations:

- **Video-Only Deep Autoencoder**: Designed for cross-modal learning, this model is trained to reconstruct both audio and video features using only the video as input. This forces the network to learn a visual representation enriched by intermodal correlations.
- **Bimodal Deep Autoencoder**: A more robust architecture trained with corrupted inputs (where one modality may be missing), pushing the model to learn to reconstruct both modalities from partial data. This strategy enables the learning of not only modality-specific features but also a **shared representation** that is invariant to input and capable of inferring one modality from another.

As demonstrated by Ngiam et al. [24], these models improved performance in classification tasks and even replicated perceptual phenomena such as the McGurk effect—thus validating the creation of a truly multimodal internal representation. Building upon the foundations of multimodal learning, the field has expanded into several application domains—one of the most prominent being Multimodal Recommendation Systems (MRS). A comprehensive survey by Deldjoo et al. [25] maps this landscape, where visual, textual, and auditory features play a critical role in improving recommendation quality and mitigating classical challenges such as data sparsity. To organize this complex domain, Deldjoo et al. categorize the wide range of MRS models based on four core technical challenges they aim to address:

- **Modality Encoder**: The initial challenge of extracting meaningful features from raw data. Models such as Vision Transformers (ViT) for images and BERT for text are standard at this stage.
- **Feature Interaction**: Perhaps the most complex challenge, focused on effectively merging features from different modalities. Strategies include bridging (using graphs to transfer information), fusion (employing attention mechanisms to combine representations), and filtering (removing noise).

- **Feature Enhancement**: After fusion, the goal is to refine the quality of the joint representation. Two notable techniques are Disentangled Representation Learning (DRL), which isolates latent factors, and Contrastive Learning (CL), which semantically aligns the multiple modal views of a given item.
- **Model Optimization**: A practical challenge related to the high computational complexity of MRS, with strategies ranging from end-to-end training to two-stage learning pipelines.

This taxonomy not only systematizes existing approaches but also highlights the need for future solutions that are more unified, interpretable, and computationally efficient [25].

2.1.4 Language-Supervised Multimodal Learning

The technique of contrastive learning has been elevated to a new level and serves as the foundation of one of the most influential multimodal models to date: CLIP (Contrastive Language–Image Pre-training), proposed by Radford et al. [11]. Rather than relying on curated and labeled datasets with fixed categories, CLIP learns robust visual representations directly from natural language supervision.

This methodology is based on massive-scale data (a dataset of 400 million image-text pairs collected from the internet) and a contrastive pre-training objective. An image encoder and a text encoder are trained jointly to align their embeddings, maximizing the similarity between correct (image, text) pairs and minimizing it for incorrect pairs.

CLIP's most impactful innovation lies in its zero-shot transfer capability. Once trained, the model can perform classification tasks on entirely new datasets without any additional training—simply by comparing the embedding of an image to the embeddings of textual prompts that describe the candidate classes (e.g., "a photo of a dog").

CLIP not only established a new state of the art for transferable visual models but also sparked an important debate about its limitations. Radford et al. highlight its shortcomings [11] in highly specialized tasks and, more critically, its tendency to inherit and amplify the social biases embedded in unfiltered web data—underscoring the profound ethical challenges associated with large-scale web-trained models.

The latest generation of multimodal models, known as Vision-Language Models (VLMs), represents a direct evolution of CLIP's foundational ideas [11]. While CLIP introduced a new paradigm by aligning image and text representations for semantic search and zero-shot classification, subsequent models such as BLIP-2 [26] build upon this foundation by incorporating generative capabilities.

This new architecture, typically based on encoder-decoder structures, enables VLMs not only to understand the relationship between text and image but also to perform more flexible and powerful tasks such as image captioning, visual question answering, and

complex instruction following in natural language. For example, BLIP-2 introduces a mechanism that allows a pre-trained Large Language Model (LLM) to "see," learning to extract and translate visual information into representations the LLM can process.

This ability for semantic alignment and conditional generation makes VLMs particularly effective in fashion recommendation scenarios. By projecting images and texts into a shared, interpretable embedding space, these models enable systems that not only recognize visually similar items but also understand nuances of style, context, and preferences described in natural language—resulting in recommendations that are more coherent, personalized, and explainable.

2.1.5 Style-Aware Recommendation

Recommendation in creative domains such as fashion presents challenges that transcend traditional recommendation systems based on purchase history or content similarity. The central question shifts from "What other items are similar to this one?" to "What goes well with this piece?" Answering this requires understanding a complex and subjective concept: style. This need becomes increasingly pressing with the proliferation of e-commerce platforms and the vast volume of visual clothing data available online.

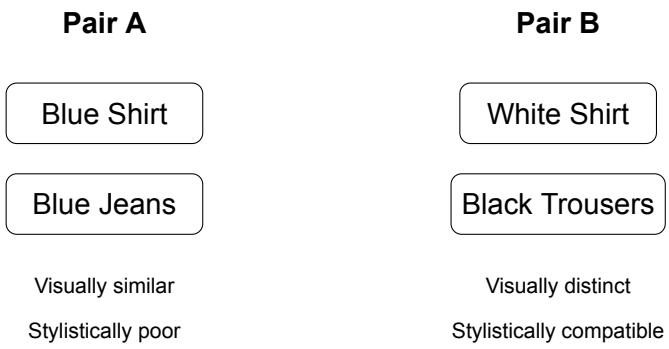

Figure 2: Illustration of the distinction between visual similarity and stylistic compatibility in fashion recommendation.

At the core of this challenge lies the fundamental distinction between visual similarity and stylistic compatibility [27, 28]. While two blue shirts may be visually similar, a white shirt and black trousers may be stylistically compatible despite their visual dissimilarity. Early approaches that relied on manually annotated fine-grained attributes (e.g., "dark slim formal trousers") proved limited: they fail to generalize to emerging trends, require expert knowledge, and demand costly labeled datasets to be created and maintained [27].

To overcome these limitations, modern research has focused on deep learning approaches that aim to model "style" as a latent embedding space. The central idea is to learn a transformation where compatible items—even across different categories such

as shoes and shirts—are mapped close together, and incompatible ones are mapped far apart. Supervision for this learning rarely comes from explicit “compatibility” labels but rather from large-scale implicit signals such as item co-occurrence in shopping carts or full outfit compositions created by users [27, 28].

However, even this approach introduces new challenges. Mapping all item types into a single latent space may result in undesired compression of stylistic variation and lead to violations of compatibility transitivity (invalid triangles), where compatibility becomes an incorrect transitive property [28].

2.1.6 Datasets for Fashion Recommendation

The development of intelligent and style-aware fashion recommendation systems relies heavily on datasets that combine visual data with structured annotations and semantic relationships between clothing items. These datasets are essential not only for training multimodal architectures but also for enabling supervised learning tasks such as visual embedding, compatibility modeling, and outfit generation. Among the most widely adopted datasets in the field are:

- **Polyvore** [29]: A benchmark dataset containing user-created outfits, along with item images and textual descriptions. It enables compatibility modeling and has been central to early research in outfit recommendation.
- **iFashion (Alibaba)** [30]: A large-scale dataset collected from e-commerce platforms, annotated with fine-grained categories, attributes, and user metadata. It is useful for both item classification and style analysis in realistic commercial scenarios.
- **Fashion32** [31]: A dataset focused on style diversity, including 32 distinct fashion styles labeled by experts. It supports tasks such as multi-style classification and personalized recommendation based on aesthetic attributes.

These datasets provide rich multimodal signals—images, style labels, item co-occurrences, and curated outfit compositions—which are crucial for training and evaluating the effectiveness of fashion recommendation models, particularly those that aim to capture complex relationships between style, context, and user preference.

2.1.7 Agent-Based Architectures and Interaction Protocols

Autonomous agents are computational systems capable of perceiving their environment, making decisions, and acting to achieve specific goals. As defined by Wooldridge and Jennings [32], the concept of agency may be classified into two perspectives: the *weak notion*, which emphasizes autonomy, proactivity, reactivity, and social ability; and the *strong notion*, which attributes mentalistic constructs such as beliefs, desires, and intentions.

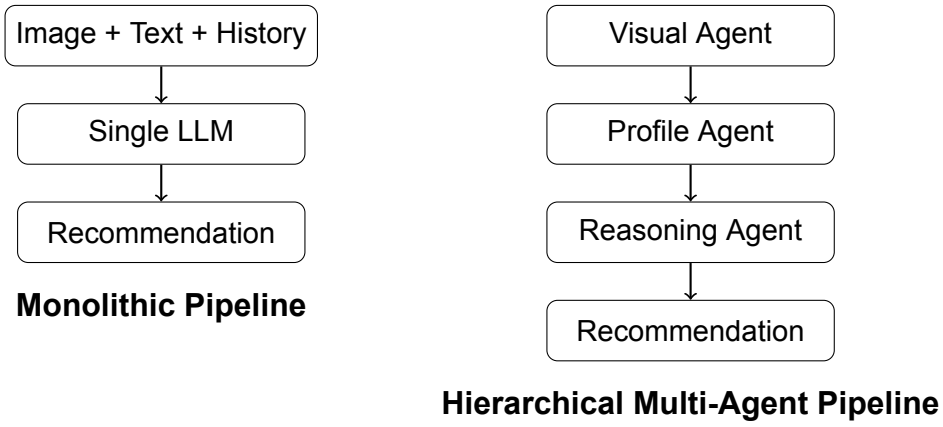

Figure 3: Comparison between monolithic reasoning pipelines and hierarchical multi-agent architectures for recommendation tasks.

To clarify the architectural motivation behind agent-based approaches, Figure 3 contrasts a monolithic reasoning pipeline with a hierarchical multi-agent organization. While monolithic architectures concentrate all reasoning into a single model, hierarchical designs explicitly separate perception, profiling, and decision-making, reducing context dilution and improving interpretability.

Originally applied in distributed computing, cooperative environments, and intelligent retrieval systems, agents have gained renewed importance with their integration into architectures driven by large language models (LLMs). These models greatly enhance the reasoning capabilities of agents, allowing them to interpret, explain, and act based on rich contextual inputs.

A particularly relevant advancement in this direction is the ReAct paradigm, introduced by Yao et al. [33], which enables agents to interleave symbolic reasoning (”thoughts”) with environment-driven actions (”acts”). Rather than relying solely on static decision trees or predefined pipelines, ReAct agents dynamically generate hypotheses, plan internally, and interact with external sources—such as APIs, search engines, or data platforms—before formulating a final response.

This model has proven effective in tasks like question answering, fact verification, and simulated navigation, often outperforming both pure reasoning (e.g., Chain-of-Thought) and pure tool-use baselines. Importantly, ReAct-based agents also improve transparency, as their thought-action sequences can be audited or guided by human operators—an essential feature in high-stakes or subjective domains.

In the context of fashion recommendation, this evolution supports the development of multimodal agents capable of understanding language, processing images, and adapting to user preferences. One illustrative example is the Neural Outfit Recommendation (NOR) framework [34], which incorporates a natural language generation module that explains recommendations through abstract comments, enabling agents to justify outfit

choices based on visual compatibility, inferred user intent, and contextual alignment.

As agent-based architectures evolve toward greater modularity and tool-oriented design, recent initiatives have proposed standardized interfaces to coordinate interactions between language models and external resources. One such initiative is the Model Context Protocol (MCP) [5], an open protocol designed to support secure and structured communication between AI agents and external tools. Rather than introducing new cognitive or reasoning mechanisms, MCP operates at the infrastructural coordination layer, reinforcing established principles in distributed systems and agent-based AI, such as explicit context boundaries, mediated tool interaction, and modular composition of autonomous agents.

Likewise, its reliance on controlled tool invocation reflects prior work on tool-mediated reasoning in language models [33, 35]. It is consistent with foundational principles of multi-agent systems and rational agent architectures, in which complex behaviors emerge from the coordination of specialized and loosely coupled agents [36–38].

By combining reasoning, action, and multimodal interaction, agent-based architectures form the foundation for next-generation fashion recommendation systems—capable of delivering not only accurate results but also meaningful, transparent, and user-aligned experiences.

2.1.8 Semantic Representation and Evaluation Foundations

This section outlines the theoretical foundations underlying the evaluation strategy adopted in this work, focusing on Transformer-based sentence embeddings and the *LLM-as-a-Judge* paradigm [39]. Figure 4 summarizes two complementary assessment perspectives commonly employed in contemporary multimodal and generative systems: quantitative semantic similarity analysis and qualitative, judgment-based evaluation.

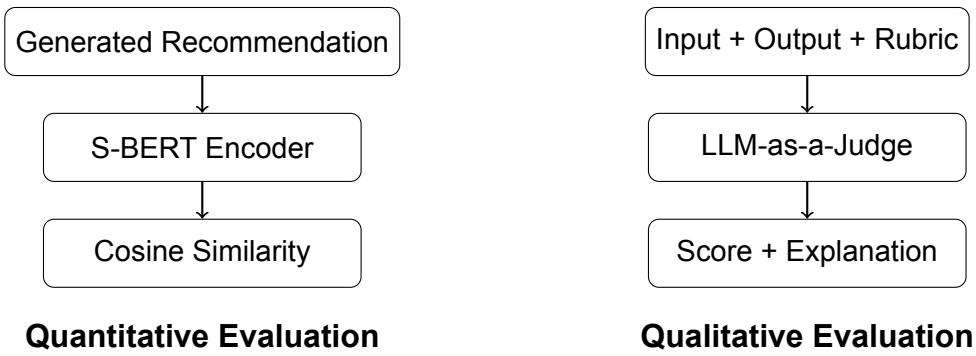

Figure 4: Overview of the semantic and qualitative evaluation strategies adopted in the literature.

The quantitative branch represents evaluation based on sentence-level semantic representations and similarity metrics, while the qualitative branch captures evaluation

paradigms grounded in large language models acting as interpretable judges. Together, these perspectives reflect a shift from surface-level textual matching toward semantically and contextually informed assessment strategies.

The introduction of BERT (Bidirectional Encoder Representations from Transformers) fundamentally altered Natural Language Processing by enabling models to learn deep, context-dependent representations of text [40]. However, standard Transformer-based models such as BERT produce sentence embeddings that are highly anisotropic and poorly suited for direct comparison using cosine similarity [41].

To address this limitation, Reimers and Gurevych proposed Sentence-BERT (S-BERT), a framework that fine-tunes Transformer models using siamese and triplet network architectures to generate semantically meaningful sentence embeddings [42]. This approach enables efficient computation of vector representations optimized for semantic similarity comparisons via cosine distance. In the literature, cosine similarity over S-BERT embeddings is widely adopted as a proxy for semantic relatedness between textual artifacts, particularly in evaluation scenarios where lexical overlap is insufficient.

While embedding-based similarity captures semantic proximity between texts, it does not assess higher-level properties such as stylistic coherence, personalization intent, or contextual appropriateness. To complement quantitative similarity metrics, recent research has proposed the *LLM-as-a-Judge* paradigm [39], in which a strong Large Language Model is positioned as a surrogate for human judgment.

In this framework, the model receives a structured prompt containing the input, the system output, and an explicit evaluation rubric (e.g., personalization or stylistic coherence), and produces both a numerical score and a natural language justification. Empirical evidence suggests that LLM-based judges can achieve agreement levels comparable to those observed between human annotators, enabling scalable, semantically grounded, and explainable evaluation of generative outputs.

2.2 RELATED WORKS

Several recent studies have addressed the challenges of fashion recommendation. Below is a selection of representative and state-of-the-art contributions:

- **FashionM3** [4]: FashionM3 is a multimodal assistant for fashion recommendation that integrates textual and visual input with multi-turn dialogue. Based on a vision-language model fine-tuned from Show-O and trained on the FashionRec dataset, it supports personalized suggestions, virtual try-on, and style variation generation with high precision. It outperforms baselines such as GPT-4o and LLaMA-3.2 in metrics like S-BERT, while maintaining lower computational complexity. User studies demonstrated high satisfaction and practical usability. However, its architecture relies on a centralized reasoning model, which may increase the risk

of context dilution when complex user histories and multimodal inputs must be processed simultaneously.

- **OutfitGAN** [21]: OutfitGAN introduces a generative approach to completing partial outfits with visually compatible items. It combines Generative Adversarial Networks (GANs) with a compatibility model trained on iFashion and Polyvore datasets. The generation is staged progressively to ensure visual coherence. It outperforms previous approaches in both compatibility and diversity, achieving visual quality comparable to real images—even with incomplete outfits.
- **FLLM** [43]: The Fashion Large Language Model (FLLM) is a domain-adapted LLM enriched with fashion-specific knowledge. It leverages automatic prompt generation and retrieval-augmented generation (RAG) to deliver personalized and context-aware recommendations. Evaluated on the Polyvore dataset, it achieved 67.21% accuracy on the Fill-in-the-Blank task and showed robustness with limited data. FLLM demonstrated strong performance in adapting to various styles and use-case contexts.
- **DiFashion** [44]: DiFashion is a diffusion-based model for generative outfit recommendation, capable of synthesizing compatible and personalized clothing items. It conditions generation on item categories, internal outfit coherence, and user preference history. Evaluated on the iFashion and Polyvore-U datasets, it outperformed Stable Diffusion and OutfitGAN in visual fidelity, compatibility, and personalization—being preferred by human judges in most qualitative tests.
- **AI-Yo** [45]: AI-Yo explores the psychosocial aspects and interaction design of chatbots in the context of fashion recommendation, with a focus on user trust, perception, and behavioral influence.

Approach	Multimodal	Generative	Dialog-Based	Style-Aware	Personalized	Multi-Agent Arch.
FashionM3	✓	✓	✓	✓	✓	—
OutfitGAN	—	✓	—	✓	—	—
FLLM	✓	✓	—	✓	✓	—
DiFashion	✓	✓	—	✓	✓	—
AI-Yo	—	—	✓	—	✓	—

Table 1: Comparison of related fashion recommendation approaches. Note that while FashionM3 utilizes a single-agent orchestration, none of the baseline approaches explicitly employ a hierarchical multi-agent architecture.

These studies highlight the shift toward deeply personalized and context-aware fashion recommendation systems that combine visual generation, multimodal interaction, and adaptive user modeling. Table 1 summarizes key characteristics of representative approaches, indicating the presence or absence of multimodal capabilities, generative reasoning, dialog support, explicit style modeling, personalization mechanisms, and agent-based interaction protocols.

Although these approaches differ in modality coverage, generative capability, dialog support, and personalization strategies, most rely on centralized or monolithic reasoning pipelines, in which perception, user modeling, and recommendation synthesis are tightly coupled within a single inference flow. This architectural concentration limits its transparency and flexibility, reinforcing the relevance of exploring explicitly modular and agent-based designs for fashion recommendation.

FashionM3 shares the same underlying dataset and multimodal scope as the present study but adopts a centralized architecture in which reasoning remains monolithic. In contrast, OutfitGAN and DiFashion primarily target visual generation and outfit compatibility, emphasizing image synthesis rather than contextual reasoning or decision decomposition. FLLM advances domain adaptation by specializing a large language model for fashion, yet still maintains a single-model reasoning pipeline without explicit task separation. AI-Yo, in turn, focuses on conversational interaction and user trust, addressing the social dimension of fashion chatbots rather than the architectural organization of the recommendation process.

In contrast, the proposed architecture departs from model-level extensions and instead investigates the impact of explicitly decomposing the recommendation process into specialized, autonomous agents. Rather than being defined by multimodality or generative capacity alone, this distinction emerges from how contextual information is structured, propagated, and combined during inference. Accordingly, the contribution of this work is orthogonal to existing solutions: it does not aim to replace their underlying models, but to examine how architectural decomposition influences interpretability, semantic stability, and personalization behavior in fashion recommendation.


3

3 METHODOLOGY

This section describes the experimental procedures adopted to validate the hypothesis that an architecture based on autonomous agents offers advantages over monolithic models in the task of fashion recommendation. The study is characterized as applied and experimental research, structured as a comparative analysis between two distinct architectural approaches.

3.1 DATA SELECTION AND SAMPLING

The experiments were conducted using data from the FashionRec dataset, specifically the Personalized Recommendation subtask. This dataset provides user interaction histories, item images, and recommendation dialogues, serving as a reference baseline for comparative evaluation.

A random sample of 30 test scenarios was extracted from the validation/test split of the dataset. The sample size was intentionally limited, as the objective of this study is not statistical generalization but controlled architectural comparison under identical experimental conditions.

Each sample consists of the following elements:

1. **User History:** Used to simulate the user's wardrobe and infer stylistic preferences.
2. **User Query:** The recommendation request expressed in natural language.
3. **Context Images:** A composite image or set of images representing the current clothing items.
4. **Ground Truth Response:** The original recommendation provided by the dataset, used as a reference baseline.

For each scenario, the same user history was provided to both architectures, and the ground-truth recommendation was not included in the input context, ensuring that no information leakage occurred during inference.

3.2 EXPERIMENTAL DESIGN

In this study, fashion recommendation serves as an experimental domain rather than a domain-specific contribution, enabling controlled evaluation of architectural design choices in multimodal, style-sensitive recommendation systems.

To ensure a fair comparison and isolate architectural design (Monolithic versus Multi-Agent) as the primary experimental variable, both pipelines employed the same generative backbone, GPT-4o, across all stages. This model was selected because it was used to generate the original recommendation dialogues in the FashionRec dataset.

The OpenAI GPT-4o model was selected as the generative backbone for both architectures to ensure consistency with the data generation process of the FashionRec dataset, whose original recommendation dialogues were produced using OpenAI-based models. Using an identical model configuration across both pipelines minimizes confounding factors related to model behavior, capacity, alignment, or training objectives, allowing the experimental analysis to focus exclusively on the effects of architectural orchestration.

The experiment therefore consisted of executing the same 30 scenarios through two distinct pipelines and comparing their outputs in terms of quality, robustness, and computational cost.

3.2.1 Architecture A: Monolithic Approach (Baseline)

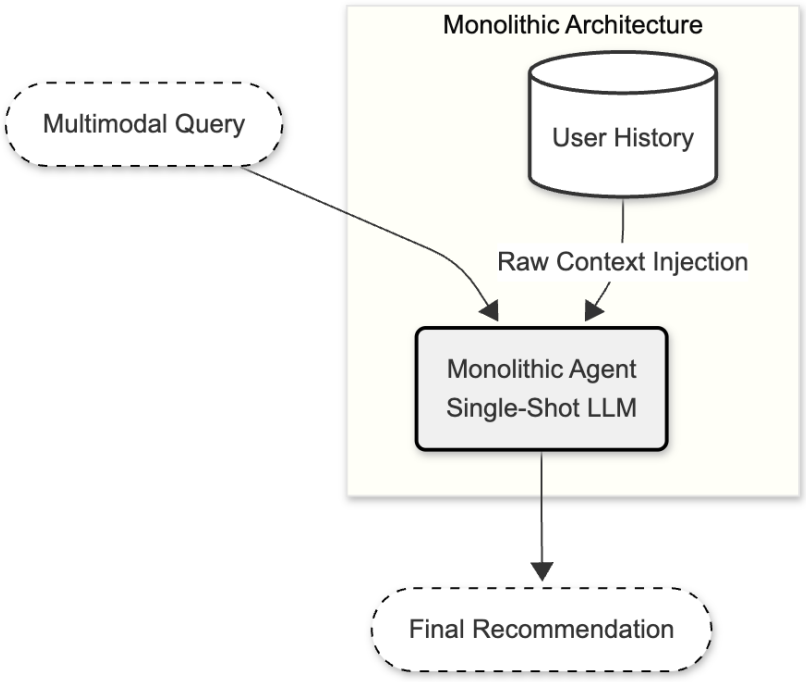

Figure 5: **Architecture A: Monolithic Baseline.** A single-shot pipeline in which user history, visual input, and instructions are jointly injected into the LLM context window.

This architecture simulates the standard operation of contemporary LLM-based recommendation systems without explicit task orchestration.

- **Operation:** A single-shot request is issued to the language model.
- **Input:** User history (textual and metadata format), user query, and item images are injected simultaneously into a single context window.
- **Objective:** To generate a recommendation through holistic interpretation of the entire input context.

3.2.2 Architecture B: Multi-Agent System (Proposed)

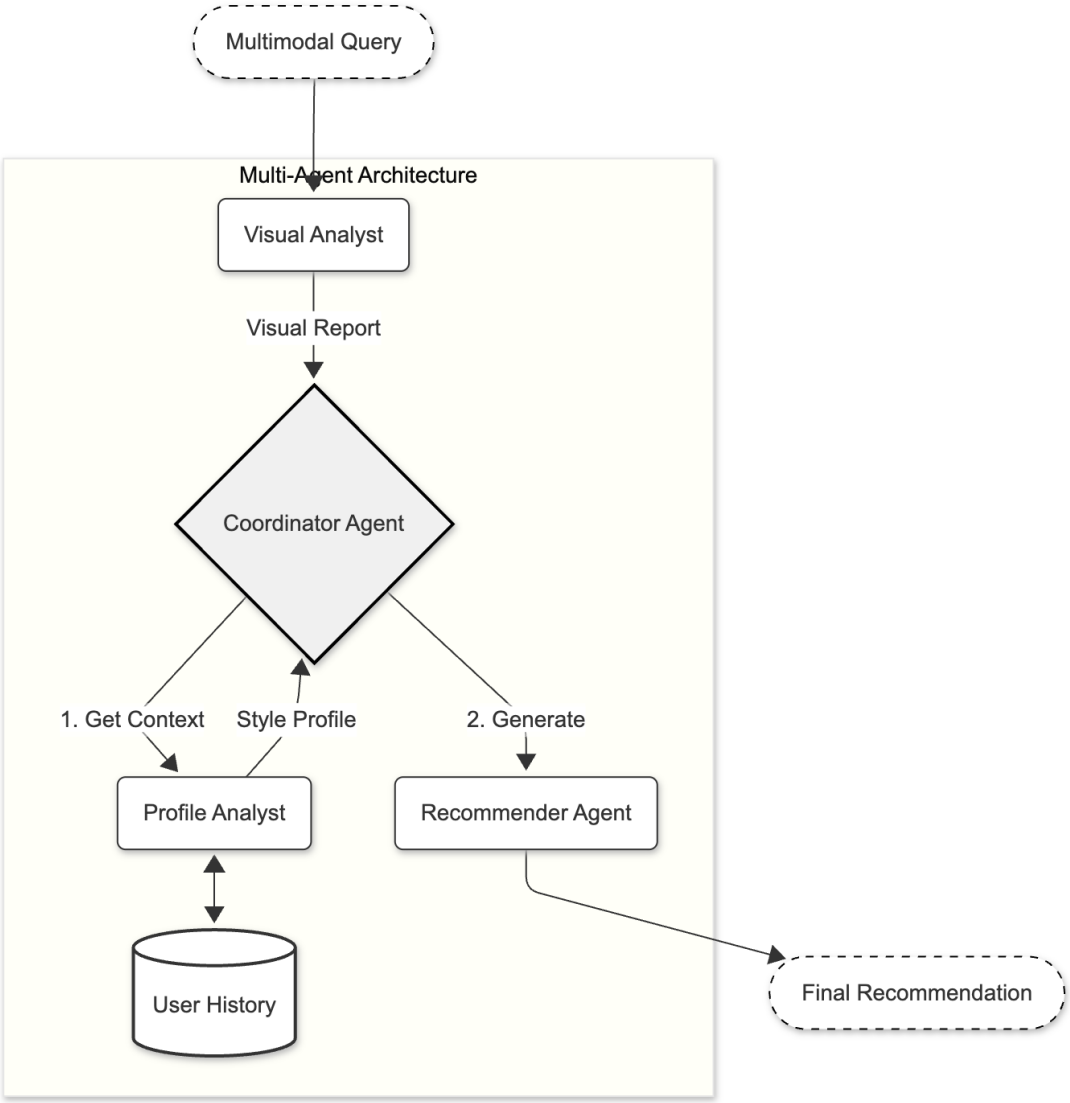

Figure 6: **Architecture B: Proposed Hierarchical Multi-Agent System.** Sequential pipeline illustrating the data flow among specialized agents.

The proposed architecture adopts task specialization and coordinated reasoning, inspired by architectural principles associated with the Model Context Protocol (MCP). Rather than implementing MCP directly, its core ideas are applied at an abstract architectural level to structure agent interaction and context flow. The system is composed of the following agents:

1. **Visual Fashion Analyst:** Responsible exclusively for visual interpretation, generating a technical descriptive report of the input items.
2. **Coordinator Agent:** Acts as the central orchestrator, invoking other agents as tools and managing information flow.
3. **Profile Analyst:** Processes the user's historical interactions to extract a structured stylistic profile.
4. **Recommender Agent:** Produces the final recommendation by integrating the visual report and stylistic profile.

3.2.3 Technical Stack and Implementation Environment

Both the proposed architecture and the monolithic baseline were implemented in a dedicated Python environment. The following tools and libraries were used:

- **Programming Language:** Python (3.13.x)
- **Agent Orchestration and LLM API:** OpenAI Agents SDK
- **Vector Similarity:** SentenceTransformers (all-mpnet-base-v2) and NumPy
- **Data Handling:** Pandas

All experiments were executed on the same local machine under identical software and hardware conditions. Inference was performed using a fixed model version and default API parameters to ensure consistent execution across all runs.

3.3 METRICS AND EVALUATION PROCEDURES

The evaluation procedures operationalize the semantic and judgment-based assessment framework discussed in Section X, combining embedding-level similarity metrics with rubric-driven qualitative evaluation. Two complementary evaluation strategies were applied to compare the outputs of the Monolithic and Multi-Agent architectures.

A fixed and standardized evaluation prompt was used across all scenarios and architectures to ensure consistency and avoid evaluation bias; the prompt structure, evaluation criteria, and scoring scale were kept constant, with variation occurring only in the system-generated output being assessed.

3.3.1 Semantic Similarity (Embedding Similarity)

Quantitative evaluation was performed using S-BERT similarity [42], computed as the cosine similarity between sentence embeddings of generated recommendations and their corresponding ground-truth responses. Embeddings were obtained using the pre-trained *all-mpnet-base-v2* model.

- **Procedure:** Outputs from both architectures were embedded and compared against the ground-truth response embeddings.
- **Objective:** To assess which architecture maintains higher semantic adherence to the reference responses.
- **Justification:** S-BERT serves as a proxy for **semantic adherence**, indicating whether factual constraints such as item category, attributes, and contextual alignment are preserved.

Semantic similarity, as measured by S-BERT, does not capture subjective notions of fashion quality or stylistic preference. Instead, it reflects semantic consistency and factual alignment with the reference response.

3.3.2 Evaluation via LLM-as-a-Judge (LLM Rubric Evaluation)

To complement embedding-based metrics, a qualitative evaluation was conducted using an advanced Large Language Model acting as a judge. This approach addresses the limitations of purely mathematical similarity measures, which do not capture reasoning depth or explanatory clarity.

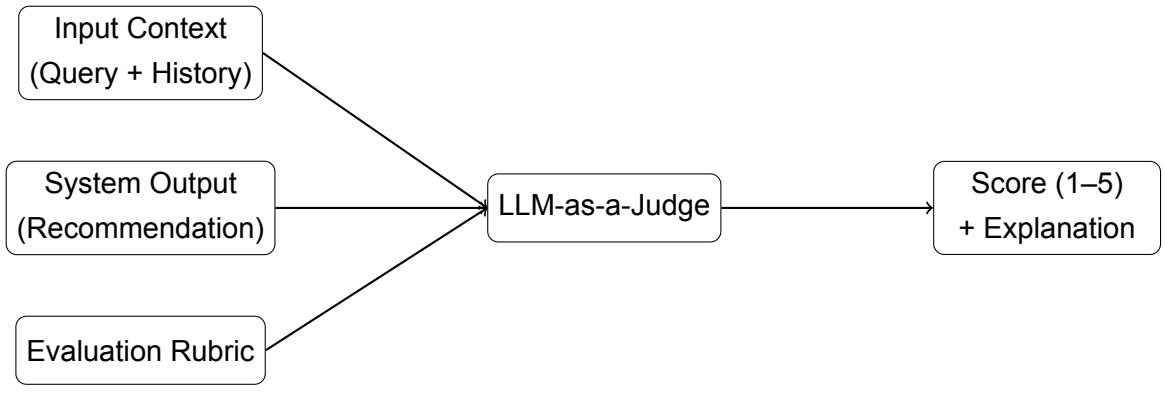

Figure 7: **LLM-as-a-Judge Evaluation Pipeline.** Rubric-driven qualitative assessment under blinded conditions.

Figure 7 illustrates the evaluation pipeline adopted in this study, highlighting how system outputs are assessed through a rubric-driven LLM-as-a-Judge process under blinded conditions.

As depicted, the judge model does not evaluate recommendations in isolation. Instead, it jointly considers the user context, the system-generated output, and an explicit evaluation rubric to produce both a numerical score and a natural-language justification.

Based on this setup, the judge evaluated blinded input–output pairs according to a structured rubric on a 5-point scale (1, 3, and 5 anchors), defined as follows:

- **Interpretability (Process Transparency):** Assesses the clarity and depth of the justification provided.
- **Personalization (Contextual Preferences):** Evaluates explicit use of the user's historical preferences.
- **Relevance (Constraint Adherence):** Measures compliance with prompt requirements.

The evaluation strategy adopted in this study aims to assess the effectiveness of the proposed recommendation architecture through controlled experimental analysis. Rather than comparing the system with commercial or large-scale industrial solutions, the evaluation focuses on intra-architectural comparison, analyzing how different architectural organizations influence semantic consistency, interpretability, and personalization stability. This approach enables a clear examination of the system's ability to represent stylistic preferences and produce coherent, explainable recommendations under well-defined conditions.

The LLM-as-a-Judge paradigm is employed as an evaluation mechanism aligned with the characteristics of style-sensitive fashion recommendation. Since stylistic coherence and semantic alignment are inherently subjective, the assessment relies on language-based judgments rather than explicit ground-truth labels. In this context, the qualitative evaluation functions as a comparative diagnostic tool to analyze architectural behavior, rather than as an absolute performance metric.

By constraining both the monolithic and hierarchical approaches to the same models, prompts, datasets, and execution conditions, the analysis isolates architectural structure as the primary variable under investigation. The results therefore reflect how reasoning decomposition and contextual separation influence recommendation behavior, rather than serving as a claim of market-level accuracy or general-purpose recommendation performance.


4

4 RESULTS AND DISCUSSION

This chapter presents the findings from the comparative analysis between the proposed Hierarchical Multi-Agent System and the Monolithic Baseline. The data is analyzed across three dimensions: (1) Quantitative Semantic Consistency using S-BERT embeddings, (2) Qualitative Performance using an LLM-as-a-Judge framework, and (3) Computational Efficiency.

The overall performance summary of both architectures is presented in Table 2. As observed, the Hierarchical System outperforms or matches the baseline across all quality metrics, with a notable trade-off in computational cost.

Metric	Hierarchical	Monolithic	Difference	Interpretation
Quality Metrics				
Interpretability	5.00	4.87	+0.13	Improved
Personalization	5.00	3.60	+1.40	Improved
Relevance	4.87	4.87	0.00	Equivalent
Overall Score	4.96	4.44	+0.51	Improved
Semantic Similarity	0.815	0.799	+0.016	Improved
Computational Cost				
Tokens Used	4447	782	+3665	Higher cost

Table 2: Comparison of Hierarchical vs. Monolithic Architectures Across Quality Metrics and Computational Cost.

The increased token usage observed in the hierarchical architecture reflects the expected computational overhead of reasoning decomposition and agent coordination. This cost trade-off is intrinsic to modular architectures and should be interpreted alongside the observed gains in interpretability and personalization, rather than as an efficiency regression.

4.1 QUANTITATIVE ANALYSIS: SEMANTIC SIMILARITY AND INFERENCE STABILITY

The semantic alignment of the generated recommendations with the ground truth (FashionRec dataset) was measured using cosine similarity on S-BERT embeddings. While the mean similarity showed a modest improvement (+2.0%), the distribution of scores reveals a critical operational advantage regarding system stability.

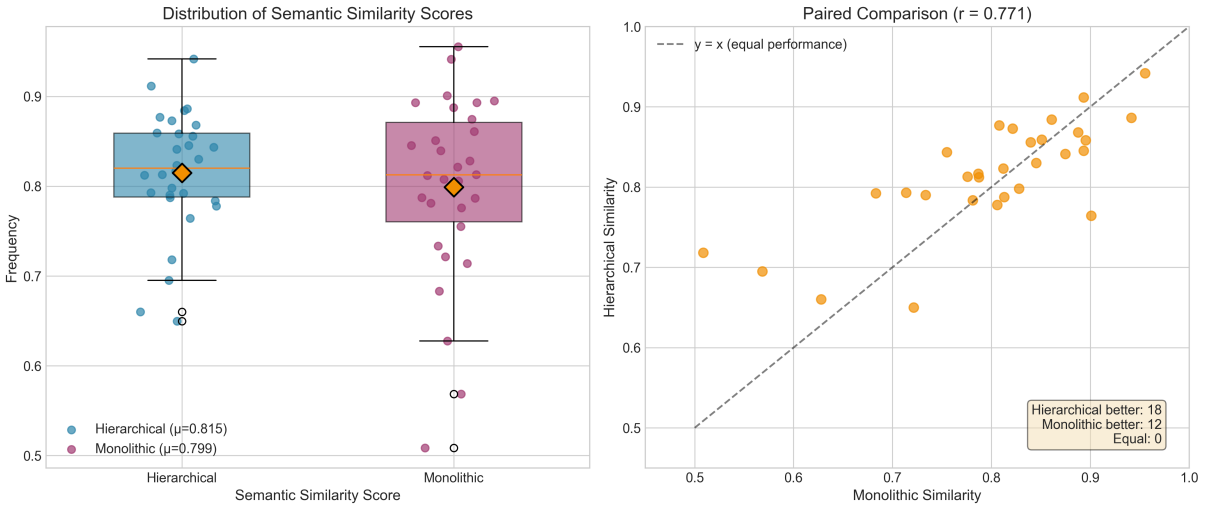

Figure 8: Distribution of Semantic Similarity Scores (Left) and Paired Comparison (Right)

As visualized in the boxplot in Figure 8, the Hierarchical System exhibits a significantly more compact distribution. The interquartile range is narrower than that of the Monolithic system, indicating higher consistency.

Furthermore, the scatter plot highlights the "performance floor." The data points falling below the $y = x$ diagonal line represent cases where the Monolithic system outperformed the Agents, but these are clustered near the top (high similarity). Conversely, the points where the Hierarchical system wins are those where the Monolithic system suffered degradation (scores below 0.7).

This visual evidence is consistent with **Hypothesis H2 (Semantic Robustness)**, suggesting that the decomposition of the recommendation workflow reduces semantic inconsistencies and mitigates context dilution commonly observed in monolithic multimodal models. By decomposing the task, the system prevents semantic drift frequently observed when monolithic models attempt to process visual and textual context simultaneously.

Beyond semantic robustness, the same distributional behavior provides indirect evidence for **Hypothesis H3 (Efficiency Trade-off)**. Although contextual complexity was not explicitly parameterized, the variability across user histories, visual inputs, and nat-

ural language queries introduces heterogeneous multimodal context conditions. Under these conditions, the Monolithic architecture exhibits sharper degradation in semantic similarity for lower-performing cases, while the Hierarchical system maintains a higher performance floor. This indicates greater inference stability under increasing contextual integration demands, achieved at the expense of higher computational cost.

4.2 QUALITATIVE ANALYSIS: STYLE AWARENESS AND PERSONALIZATION

To capture nuances beyond vector similarity, the systems were evaluated by an LLM Judge on a scale of 1 to 5. The comparative results are visualized in Figure 9.

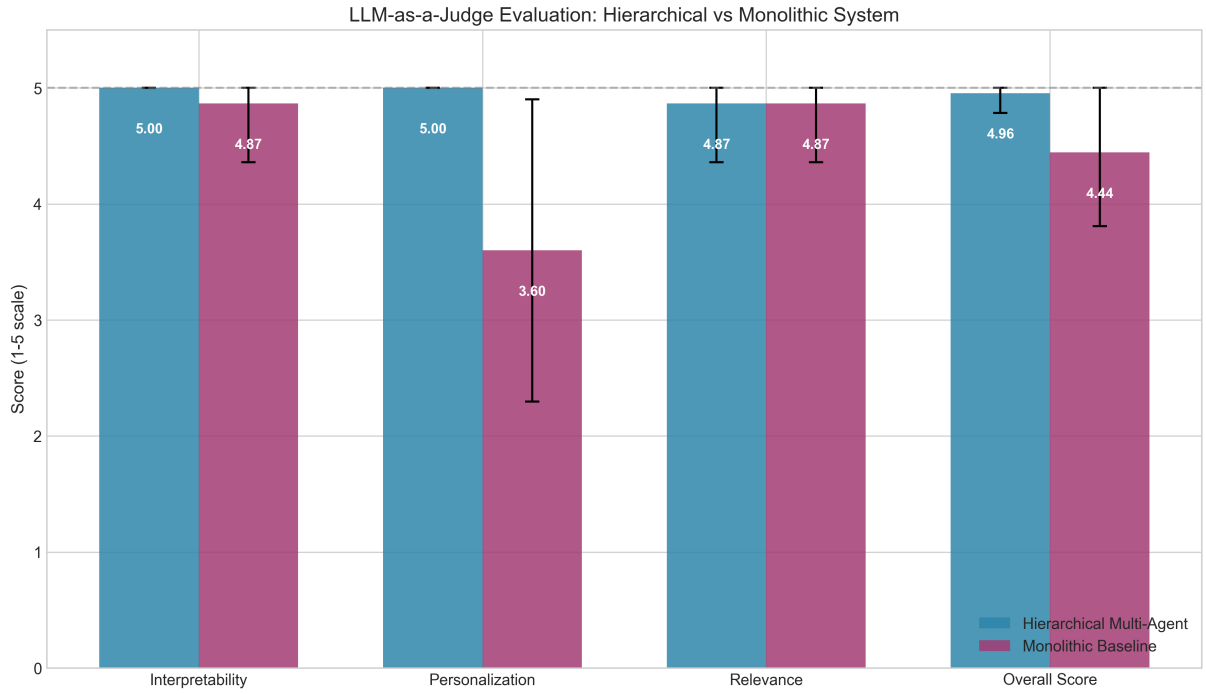

Figure 9: LLM-as-a-Judge Evaluation Results. The bar chart compares the mean scores (1-5 scale) across three criteria, with error bars indicating standard deviation. Note the significant gap in the Personalization metric.

The most striking finding is the disparity in the Personalization metric. The Hierarchical System achieved a perfect mean score of 5.00, whereas the Monolithic baseline dropped to 3.60. The error bars indicate that the Monolithic system had high variability in personalization (sometimes using the history, sometimes ignoring it), while the Agent system was perfectly consistent.

This confirms **Hypothesis H1 (Personalization)**. The Monolithic model, despite having the user history in its context window, often suffers from *context dilution*. By dedicating a specific Profile Analyst agent to process the history before the recom-

mendation is generated, the proposed architecture ensures that user preferences are treated as a hard constraint rather than optional context.

Although the LLM-as-a-Judge paradigm provides a scalable and structured mechanism for qualitative evaluation, it does not replace human judgment. Its role in this study is to approximate expert-level stylistic assessment in a consistent and reproducible manner, enabling comparative analysis under controlled conditions.

It is also acknowledged that the judge model shares architectural and training characteristics with the generative backbone used in both pipelines, which may introduce alignment bias. This risk is mitigated by blind evaluation, explicit rubrics, and the comparative nature of the analysis, but it remains a limitation to be addressed by future studies involving human evaluators. Therefore, the qualitative results should be interpreted as indicative rather than definitive measures of stylistic quality.

In terms of Relevance, both systems performed identically (4.87), suggesting that the underlying model (GPT-4o) is inherently capable of understanding instructions. The value of the Agent architecture lies specifically in deep integration of context (Style), not just content (Category).

4.3 COMPUTATIONAL EFFICIENCY ANALYSIS

A critical trade-off for the improved robustness and personalization is the computational cost, measured in token usage per request. This relationship is visualized in Figure 10.

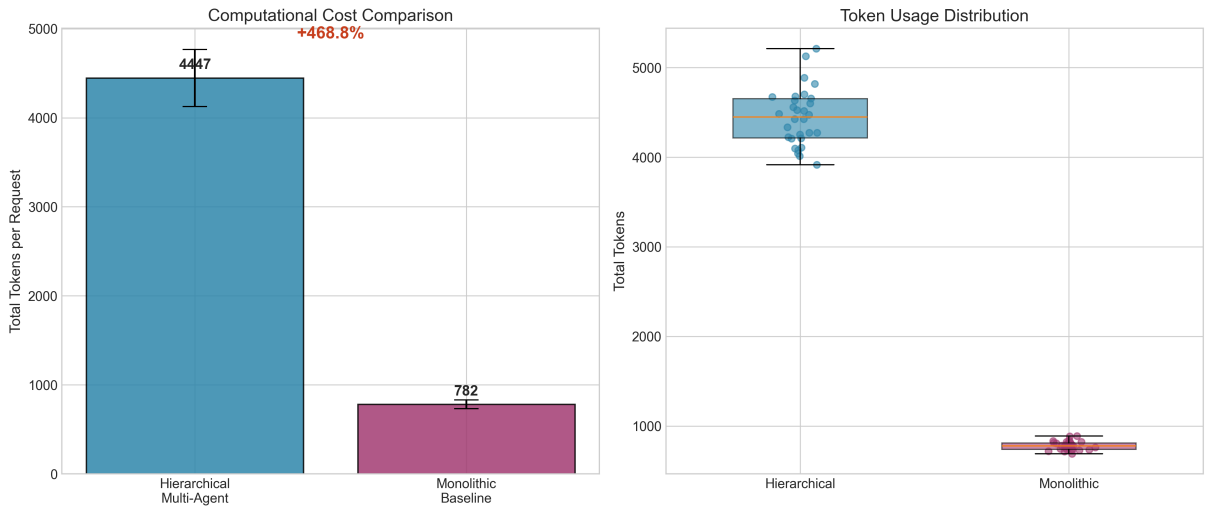

Figure 10: Computational Efficiency Analysis. The chart compares the average token usage per request between the Hierarchical Multi-Agent System and the Monolithic Baseline.

The Hierarchical System consumed an average of 4,447 tokens per request, compared to 782 tokens for the Monolithic baseline, corresponding to an approximate **4.7× increase** in token usage. As shown in the distribution plot, token consumption in the Hierarchical system is not only higher but also more variable. This overhead is primarily driven by explicit intermediate reasoning artifacts—such as visual analysis reports and profile summaries—exchanged between specialized agents. While this design incurs additional computational cost, it directly reflects the architectural decision to externalize and modularize reasoning processes rather than compressing them into a single context window.

This result reinforces the trade-off described in **Hypothesis H3**, where improved inference stability under complex multimodal context is obtained through increased token consumption and agent coordination overhead.

4.4 DISCUSSION OF FINDINGS

The visual and statistical evidence supports the following conclusions regarding the research objectives:

1. **Personalization as an Architectural Challenge:** The perfect score in personalization combined with the high token cost suggests that “Style-Awareness” requires explicit cognitive steps. The Monolith failed to personalize not because it lacked knowledge, but because it lacked the “attention span” to process multimodal and historical context simultaneously.

2. **Inference Stability via Specialization:** The tighter similarity distribution confirms that decomposing the task reduces the search space for each individual agent, minimizing abrupt inference degradation under complex context integration

3. **The Efficiency Trade-off:** The results confirm that high-quality, personalized reasoning comes at a cost in computational resources. While the proposed architecture delivers superior user experience and interpretability, future work must focus on optimization strategies—such as summarization of intermediate reasoning steps or use of smaller specialized agents—to improve efficiency without sacrificing gains.

5

5 CONCLUSION

Fashion recommendation systems pose challenges that go beyond item similarity or transaction optimization, particularly when they rely on the combined interpretation of visual cues, user history, and subjective aesthetic preferences.

To address these challenges, this work presents the design, implementation, and evaluation of a hierarchical multi-agent recommendation architecture inspired by principles of the Model Context Protocol (MCP). By decomposing the recommendation workflow into specialized agents responsible for visual analysis, user profile modelling, and recommendation synthesis, the proposed architecture restructures contextual processing and integration, and was evaluated on the FashionRec dataset in direct comparison with monolithic language-model-based recommendation approaches.

The hierarchical architecture exhibited greater personalization consistency and interpretability, along with improved semantic stability under heterogeneous multimodal inputs. By separating visual analysis from user profile modelling, the approach mitigates semantic drift commonly observed in monolithic pipelines, suggesting that architectural organization influences how large language models are applied in style-sensitive recommendation settings.

Efficiency is not solely a property of the model itself, but a design variable, particularly in multimodal recommendation systems where interpretability and personalization are primary design priorities. The increase in token consumption is a direct consequence of the modularization of reasoning, highlighting an architectural trade-off between efficiency and interpretability in multimodal systems.

Future work may explore the use of compact or fine-tuned language models, including retrieval-augmented generation (RAG), as well as adaptive summarization strategies to reduce coordination overhead.

Overall, hierarchical multi-agent architectures exhibit greater contextual stability, interpretability, and stylistic coherence than monolithic pipelines in fashion recommendation tasks.

REFERENCES

References

[1] Statista Research Department, "Global fashion e-commerce market size from 2018 to 2027," https://www.statista.com/topics/9288/fashion-e-commerce-worldwide/, 2024, accessed: 2025-09-15.

[2] ECDB Research, "Fashion e-commerce market in brazil," https://ecdb.com/resources/sample-data/market/br/fashion, 2025, accessed: 2025-09-15.

[3] B. Schwartz, *The Paradox of Choice: Why More Is Less*. Harper Perennial, 2005.

[4] K. Pang, X. Zou, and W. Wong, "Fashionm3: Multimodal, multitask, and multiround fashion assistant based on unified vision-language model," *IEEE Transactions on Neural Networks and Learning Systems*, 2025, inclui descrição do Model Context Protocol (MCP). [Online]. Available: https://arxiv.org/abs/2504.17826

[5] X. Hou, Y. Zhao, S. Wang, and H. Wang, "Model context protocol (mcp): Landscape, security threats, and future research directions," 2025. [Online]. Available: https://arxiv.org/abs/2503.23278

[6] P. Resnick and H. R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56–58, 1997.

[7] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th International Conference on World Wide Web (WWW '01)*, 2001, pp. 285–295.

[8] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th International World Wide Web Conference (WWW)*, 2017, pp. 173–182.

[9] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Computing Surveys*, vol. 52, no. 1, pp. 1–38, 2019.

[10] P. Lewis, E. Perez, A. Piktus *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. [Online]. Available: https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable

visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[12] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022. [Online]. Available: https://arxiv.org/abs/2201.12086

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org

[14] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain." *Psychological Review*, vol. 65, no. 6, p. 386–408, 1958. [Online]. Available: http://dx.doi.org/10.1037/h0042519

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS 2012)*. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in *Proceedings of the International Conference on Learning Representations (ICLR 2016), Workshop Track*, 2016, arXiv:1511.06939. [Online]. Available: https://arxiv.org/abs/1511.06939

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[21] M. Moosaei, Y. Lin, and A. e. a. Akhazhanov, "Outfitgan: Learning compatible items for generative fashion outfits," in *CVPR*, 2022, pp. 2273–2277.

[22] M. Yu, Y. Ma, L. Wu, C. Wang, X. Li, and L. Meng, “Fashiondpo: Fine-tune fashion outfit generation model using direct preference optimization,” *arXiv preprint arXiv:2504.12900*, 2025. [Online]. Available: https://arxiv.org/abs/2504.12900

[23] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, “Kan: Kolmogorov-arnold networks,” *arXiv preprint arXiv:2404.19756*, 2 2025. [Online]. Available: http://arxiv.org/abs/2404.19756

[24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011, pp. 689–696. [Online]. Available: https://proceedings.mlr.press/v15/ngiam11a.html

[25] Y. Deldjoo, F. Nazary, and A. e. a. Ramisa, “A review of modern fashion recommender systems,” *ACM Computing Surveys*, vol. 56, no. 4, 2023.

[26] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research. PMLR, 2023. [Online]. Available: https://proceedings.mlr.press/v202/li23q.html

[27] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie, “Learning visual clothing style with heterogeneous dyadic co-occurrences,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4642–4650.

[28] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, and D. Forsyth, “Learning type-aware embeddings for fashion compatibility,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4511–4520.

[29] Z. Lu, Y. Hu, Y. Jiang, Y. Chen, and B. Zeng, “Learning binary code for personalized fashion recommendation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 562–10 570.

[30] W. e. a. Chen, “Pog: Personalized outfit generation for fashion recommendation at alibaba ifashion,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2662–2670.

[31] J.-H. e. a. Lai, “Theme-matters: Fashion compatibility learning via theme attention,” *arXiv preprint arXiv:1912.06227*, 2019. [Online]. Available: https://arxiv.org/abs/1912.06227

[32] M. Wooldridge and N. R. Jennings, “Intelligent agents: Theory and practice,” *The Knowledge Engineering Review*, vol. 10, no. 2, pp. 115–152, 1995.

[33] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, “React: Synergizing reasoning and acting in language models,” *arXiv preprint arXiv:2210.03629*, 2023.

[34] Z. Lin, X. Song, L. Nie, Z.-J. Zha, Q. Tian, and T.-S. Chua, “Explainable outfit recommendation with joint outfit matching and comment generation,” in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 2020, pp. 1663–1672.

[35] T. Schick, J. Dwivedi-Yu, R. Dessì *et al.*, “Toolformer: Language models can teach themselves to use tools,” *arXiv preprint arXiv:2302.04761*, 2023.

[36] M. Wooldridge, *An Introduction to MultiAgent Systems*. John Wiley & Sons, 2009.

[37] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2020.

[38] Z. Xi, W. Xie *et al.*, “The rise and potential of large language model based agents,” *arXiv preprint arXiv:2309.07864*, 2023.

[39] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging llm-as-a-judge with mt-bench and chatbot arena,” *arXiv preprint arXiv:2306.05685*, 2023. [Online]. Available: https://arxiv.org/abs/2306.05685

[40] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[41] K. Ethayarajh, “How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings,” *Proceedings of EMNLP-IJCNLP*, pp. 55–65, 2019.

[42] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019.

[43] Z. Shi and S. Yang, “Integrating domain knowledge into large language models for enhanced fashion recommendations,” *arXiv preprint arXiv:2502.15696*, 2025. [Online]. Available: https://arxiv.org/abs/2502.15696

[44] Y. Xu, W. Wang, F. Feng, Y. Ma, J. Zhang, and X. He, “Diffusion models for generative outfit recommendation,” in *Proceedings of the 47th International ACM SIGIR Conference*, 2024, pp. 1350–1359.

[45] Z. Ye, M. Guo, J. Han, and J. Ma, “Ai-yo: Embedding psychosocial aspects in the fashion stylist chatbot design,” in *Proceedings of Creativity and Cognition*, 2024, pp. 520–526.

idp

Ensino que
te **conecta**