VestiME: A RAG-based Multimodal Fashion Outfit Recommender System

Amanda Maduro
Facultad de Ciencias y Tecnologías (CyT)
Universidad Católica Nuestra Señora de la Asunción
Asunción, Paraguay
amandamaduro8@gmail.com

Carlos Rodríguez
Facultad de Ciencias y Tecnologías (CyT)
Universidad Católica Nuestra Señora de la Asunción
Asunción, Paraguay
carlos.rodriguez@uc.edu.py

Sebastián Ortiz Chamorro
Facultad de Ciencias y Tecnologías (CyT)
Universidad Católica Nuestra Señora de la Asunción
Asunción, Paraguay
sebastian.ortiz@uc.edu.py

***Abstract*—Fashion plays a significant role in cultures around the world and in the global economy. Similarly to what has happened with other consumer products, the Internet has transformed the way people shop for fashion, transitioning from an in-store experience to an increasingly online one thanks to ever-evolving e-commerce. While much of today's e-commerce experience for fashion still relies on a keyword-search-based approach for web and mobile catalogs, the advent of Generative Artificial Intelligence (GenAI) and multimodal Large Language Models (LLMs) has brought unprecedented opportunities for searching, exploring, and recommending fashion products online. This paper aims to leverage state-of-the-art GenAI and multimodal LLM technologies to provide a different customer experience based on conversational, intent-driven exploration of fashion catalogs and outfit recommendations through our proposed solution, *VestiME*. The results of our user study involving fashion experts show the potential of these highly interactive, conversational, and multimodal e-commerce experiences for fashion.**

***Index Terms*—Fashion outfit recommendation, RAG, LLM, AI agent**

I. Introduction

Fashion is an important cultural element that also plays a vital role in the global economy [1]. The way people shop for fashion has gone through a series of transformations over the past centuries. Starting in the pre-industrial era, clothing was handmade at home or produced by local artisans. This was a time-consuming and expensive process that made owning new clothing a luxury for most people. The 19th century witnessed the birth of department stores and ready-to-wear clothing thanks to the Industrial Revolution that enabled its production at scale, which made shopping for clothes more accessible. During the late 19th and entering the 20th century, mail-order catalogs (see Fig. 1) became popular and allowed customers to order clothing and benefit from delivery services directly to their door.

Toward the end of the 20th century and into the 21st, similarly to other consumer products, fashion has evolved through digital transformation driven by the Internet, web technology development, and e-commerce, moving its traditional in-store and printed-catalog-based experiences into an online one bringing convenience, innovation, and personalization to customers. In this context, online recommender systems play a crucial role helping customers discover products that are aligned with their preferences and tastes by leveraging user interactions with e-commerce platforms, including visits, product purchases, and reviews. This allows these platforms to better understand customer preferences that can be used to determine *what*, *how*, and *when* to recommend. At the same

Fig. 1. Fall and Winter, 1890-91 Fashion Catalogue - H. O'Neill and Co. (1890). Accessed from Wikimedia Commons [2] , licensed under CC0 (https://creativecommons.org/publicdomain/zero/1.0/).

This work has been partially funded by PRONII-CONACYT (Paraguay), Res. 90/2023.

Citation: A. Maduro, C. Rodríguez, S. Ortiz-Chamorro. "VestiME: A RAG-based Multimodal Fashion Outfit Recommender System." The 51st Latin American Computing Conference (CLEI 2025), October 2025, Valparaíso, Chile.

time, in today's e-commerce ecosystem, recommender systems have become essential: Approximately 35% of Amazon's sales and more than 80% of the content consumed by Netflix's users come from such systems [3].

Recommender systems face unique challenges when making recommendations in the context of fashion outfits. Deldjoo et al. [3] highlight the need to take into consideration the compatibility of products when recommending a complete outfit and the importance of personalization, which needs to account for the place where the outfit will be used, the season, occasion, and the social context of the customer. Additionally, the limitations of existing web- and mobile-app-based interactions (typically based on clicks, menu navigation, keyword searches, etc.) make it hard to provide a closer experience to that which clients are used to when making fashion decisions and purchases, that is, having interactions where clients ask for fashion advice and exchange opinions with family members, friends, and salespeople through conversational interactions. This paper focuses on the challenge of providing a fashion virtual assistant, *VestiME*, which offers outfit recommendations through a conversational interface. *VestiME* is powered by multimodal LLMs [4] that are grounded in a catalog of fashion products, on top of which the solution can provide outfit recommendations based on user intents and preferences expressed through a natural-language conversation. Our user study involving fashion experts demonstrates the usefulness and potential of the proposed solution, and, at the same time, sheds light on interesting open problems and challenges that are worth exploring further.

This paper is organized as follows. Section II presents related work. Section III discusses the major challenges faced by recommender systems for fashion. Next, Section IV introduces *VestiME*, our RAG-based multimodal outfit recommender agent. Section V presents the implementation and user study we carried out to validate our proposed solution. Finally, Section VI presents our concluding remarks and future work.

II. Related work

The fashion industry has used recommender systems for almost as long as e-commerce has existed. Such systems have been classified according to specific objectives, including the recommendation of fashion products, complete outfits, advice on sizes and measurements, and explanation capabilities for the recommendation [3]. Fashion recommendations typically focus on suggesting individual products that satisfy the customers' preferences, using techniques such as collaborative filtering and Bayesian personalized ranking, which optimize personalization through implicit feedback and user-product interaction probabilities. Other systems focus on recommending the clothing size that best fits customers' body characteristics. Systems that support explanations provide reasons for why they recommend a specific product using text or visual mechanisms to augment transparency and decision-making [5].

In the context of fashion recommender systems, *Fashion Outfit Recommendation* (FOR) [5] focuses specifically on recommending outfits where the individual product items harmonize with each other and as a whole based on product item compatibility. A variation of FOR is the personalized FOR (PFOR) [5], which adapts recommendations to the tastes of customers instead of focusing solely on the harmony of the product items that make up the outfit. In this context, Ding et al. [6] use a *fill-in-the-blank* approach to combine and score combinations of users, outfits, and product items. A model based on Conditional Random Fields (CRF) is trained to predict the compatibility of an outfit with a user. Other techniques have also been used in this space including Convolutional Neural Networks (CNN) [7] and Transformers [8] as alternative methods.

The advancements in Natural Language Processing (NLP) that characterized the 2010s boosted the interest in conversational user interfaces. Challenges such as fast-changing user preferences and natural language feedback are typically at the center of this line of research [9]. For example, Sapna et al. [10] proposed Athena, a chatbot based on artificial neural networks that offers recommendations based on graphs and trends. Zhang [11] proposed Outfit Helper, a chatbot that assists men in choosing outfits by combining NLP, computer vision, and dialog structuring. Husak et al. [12] designed a Telegram-based system to generate a list of clothing styles based on user needs. The main challenges faced in this work included issues in language interpretation due to the use of dialects and individual communication styles. Similar problems were common in research work published in this space before the massive launch and democratization of access to LLMs and GenAI technologies.

Recent progress made in the space of LLMs and GenAI technologies has spiked the interest in leveraging such technologies to improve the experience in conversational user interfaces. Several studies demonstrate their potential in this context. For example, Kolb et al. [13] proposed an architecture that combines both local and global knowledge through chat history rewriting using Generative Pre-trained Transformers (GPT) [8]. Forouzandehmehr et al. [14] addressed the challenge of recommending outfits based on famous people using LLMs to obtain style information from users and prompt engineering techniques to better understand their preferences. While the proposed solution showed promise and produced good results, issues were identified related to gender bias and lack of interactive, multi-turn conversations.

While recent advances in outfit recommendations have been promising, challenges persist when it comes to personalization and explainability of traditional recommender systems [15], many of which are based on generic user profiles with limited (or no) natural language interactions, making it difficult for such systems to recommend outfits that adapt to user taste, preference changes, exploration and discovery, and continuously changing fashion trends.

III. Major Challenges in Recommender Systems for Fashion

Fashion Recommender Systems have faced several challenges over the past decades, many of which are still out-

standing. In this section, we summarize major challenges that represent relevant open research problems within this domain as discussed by Deldjoo et al. [3].

The first such challenge is *item representation in fashion*. Here, most recommender systems based on collaborative and content-based filtering typically struggle with making good recommendations due to the limited information they work with, including a lack of consideration for the visual appearance of the products, which is arguably the most essential element in fashion. Recent work has started to incorporate additional elements such as text descriptions and reviews [16], [17], as well as product images and videos [18]–[20].

A second challenge is the *compatibility of items*. Questions such as "Does this neckless go well with my dress?" or "Is the color of my purse a good match for the color of my belt and shoes?" are very challenging to address with today's fashion recommender systems. Work that tries to address this challenge typically relies on co-purchase times [21], outfits obtained from social media images [22], and latent style types [23], with limited success.

Another challenge lies in *fit and personalization*. A good recommendation in the fashion realm typically accounts for aspects such as trends, season, occasion, and location [24]–[26]. This means that recommender systems need to go beyond just the product characteristics and take into account contextual information [27]. Part of this context includes also fit, which needs to consider body shapes and sizes, among other criteria [28].

Interpretation and explainability of the recommendations are key (but challenging) for customers to buy in. For the most part, traditional recommender systems do not offer explanations as to why they are making a recommendation. Current approaches that address this problem go as far as highlighting certain aspects of the image of the product or its description [29], [30], providing no clear rationale behind it.

Recent progress made in the context of multimodal LLMs and GenAI [4], Retrieval Augmented Generation (RAG) systems [31], and Agentic AI [32] bring new opportunities to address the core challenges listed above. For example, item representation can be helped by multimodal LLMs by combining text, images, and even videos; interpretation and explainability can be explored through the use of LLMs; while compatibility of items can be explored with RAG, reasoning LLM, and chain-of-though (CoT) [33].

IV. VestiME: A Fashion Virtual Assistant

The latest advancements in LLMs, GenAI, and agent technologies have brought unprecedented opportunities for e-commerce and online shopping, which enable the construction of recommender systems with conversational user interfaces with superior capabilities (w.r.t. traditional recommender systems) to interpret and generate natural language. In this section, we introduce *VestiME*, an agent-based virtual assistant that leverages multimodal LLM and RAG technologies to make personalized outfit recommendations that are grounded on a catalog of fashion product items.

A. An Architectural Perspective

Fig. 2 presents an architectural view of our proposed solution. The solution is divided into two main parts: (i) Data preparation, embedding, and indexing (left-hand side of the architecture), and (ii) retriever and generation agent (on the right-hand side). In the first part, we process the images from our catalog, extract attributes and descriptions with the help of a large language model $\mathcal{LLM}$-1 (we provide the details of the actual models we use in Section V-A), generate the embeddings out of them, and index them for faster search and retrieval. In the second part, we leverage a second large language model, $\mathcal{LLM}$-2, to manage intents and conversations, and make decisions based on *prompting* techniques. This second part, in turn, consists of two phases: (a) Gathering of user preferences, and (b) outfit generation based on the product items found in our catalog. We will delve into these two parts separately in the next sections.

B. Data Preparation, Embedding and Indexing

We start by first discussing the catalog of fashion product items we use in this work, which consists in the widely used Cleaned Maryland Dataset[1]. This dataset contains close to 127K product item images that are organized into 20 different product categories spanning from outwear and dresses to neckwear and bracelets. While the dataset includes some metadata about the product items, we opted to augment this metadata through the use of multimodal LLMs to include information such as occasion and style. After a manual inspection of the dataset, we curated it by removing non-essential categories such as brooches. We also removed around 20K duplicated images, resulting in a dataset of approximately 101K product items.

To perform this additional data curation and augmentation, we used $\mathcal{LLM}$-1, which is a multimodal LLM for summarization, chat applications, data extraction, and captioning. This allowed us to extract additional attributes such as the name of the product, description, category, style, materials, color, silhouette, occasion, and season. With the use of prompting and a few-shot [34] approach, we implemented the workflow presented in Fig. 2, where images were encoded using Base64 for image processing, and instructions for Pydantic (https://pydantic.dev/) were included in the prompt to guide $\mathcal{LLM}$-1 in the identification of attributes. We used LangChain (https://www.langchain.com/) for the orchestration of the interactions with $\mathcal{LLM}$-1, which allowed us to streamline the communication between $\mathcal{LLM}$-1 and the Pydantic Output Parser, which validates the outputs (in JSON format) generated by the LLM. As an optimization mechanism for this image curation and augmentation workflow, we leveraged LangChain's batch processing, with batches of sizes ranging from 50 to 100 images. This allowed us to improve image processing times

[1]https://github.com/AemikaChow/AiDLab-fAshIon-Data/blob/main/Datasets/cleaned-maryland.md

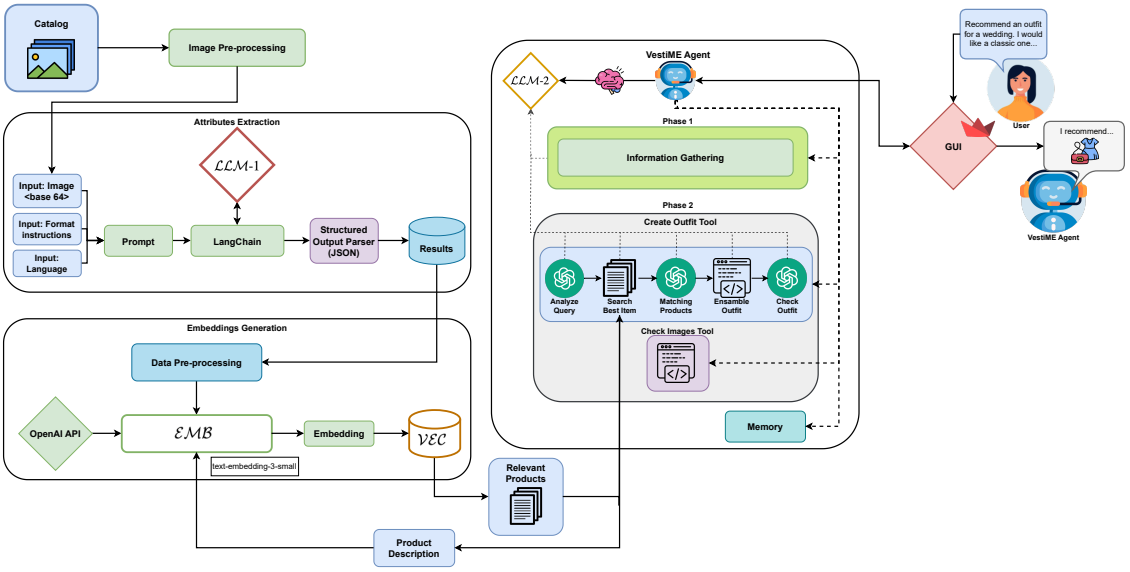

Fig. 2. Architecture of VestiME: The left-hand side of the architecture focuses on data preparation, embedding, and indexing, while the right-hand side focuses on the retriever and generation agent.

and the handling of formatting errors. This workflow allowed us to process all 101K images and generate the corresponding metadata, making this dataset ready for our next steps in this workflow: Embedding generation and indexing.

We leverage an embedding model, $\mathcal{EMB}$, to generate the embeddings from this enriched dataset. The embeddings for each of the product items in our dataset were generated from a paragraph that was created from the descriptions obtained from the attributes contained in the product's augmented metadata. As a result, we ended up with 101K vector representations (embeddings) of the products contained in our cleaned and enriched catalog. As a quick validation of the quality of the embeddings, we randomly sampled 1,000 embeddings and performed dimensionality reduction with t-Distributed Stochastic Neighbor Embedding (t-SNE) [35], which allowed us to visualize the embeddings in a two-dimensional space (see Fig. 3). We observed a clear separation among the identified categories, which allowed us to validate that the embeddings accurately capture the essence and semantic differences in the product items contained in our dataset.

As a final step in this curation and augmentation workflow, we stored and indexed the resulting embeddings in a vector database, $\mathcal{VEC}$, which allows for representing, storing, indexing, and retrieving data in the form of tensors that contain synthesized text, metadata, vector representations, and unique identifiers. With this step, we enabled our *VestiME* Agent (explained next) to search for and retrieve fashion products in order to provide outfit recommendations.

C. Retriever and Generation Agent

The right-hand side of Fig. 2 shows the flow that is followed by *VestiME* to process users' requests, starting from the

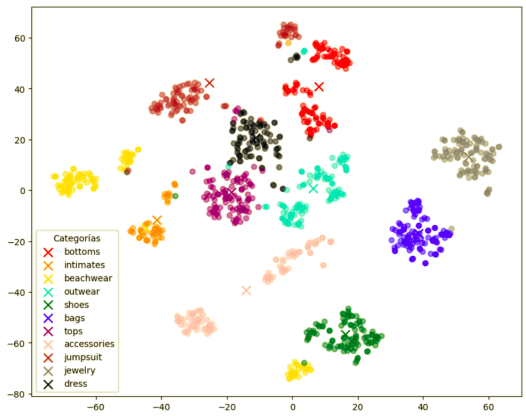

Fig. 3. Visualization of categories after applying dimensionality reduction with t-SNE.

initial intent down to the final recommendation for the outfits. *VestiME* is the core agent of our virtual assistant, managing the conversational interactions with end-users and assuming the role of a fashion expert to make outfit recommendations. This agent leverages GenAI, LLM and LLM-orchestration technologies to keep track of previous interactions with end-users ensuring dialog coherence.

As stated previously and shown in Fig. 2, the *VestiME* agent is implemented across two phases. The first phase focuses on the interaction with the end user to gather user preferences, which include aspects such as style, color, and occasion. These are fundamental elements needed for the agent to recommend an outfit. If such details are not provided by the user during the first interaction, the agent prompts the user to collect their

preferences.

During the second phase, the agent generates the outfit with the help of two key components: *CreateOutfit* and *CheckImages*. The first step in this flow involves the synthesis of the user's intent based on the information gathered during the first phase. This synthesized information is then fed into the *CreateOutfit* component, which is responsible for generating the outfit. This component also checks whether the generated outcome matches user's preferences, occasion, and style. Below, we summarize the core sequence of steps followed in the *CreateOutfit* component:

- **Analyze query**: This step uses $\mathcal{LLM}$-2 to generate the description of the main product item from an outfit, based on the synthesized user intent. This product item is considered the main element of the outfit, such as a dress, shirt, pants, or skirt. The underlying prompt employed in this step uses a one-shot approach [34], providing an example that can help guide the model. In this prompt, we instruct the model to analyze the user's intent and describe a key product item (e.g. a dress or shirt) that can help build an outfit. The selection of the remaining product items will then be based on the compatibility with this main product item. This step is also very useful when dealing with loosely defined preferences, ensuring that the recommendation process starts on a solid foundation for the outfit construction.
- **Search best item**: This step uses RAG with two retrievers (described next), which access the product catalog in our vector database $\mathcal{VEC}$ and uses $\mathcal{LLM}$-2 to identify the product that best matches the product description generated in the previous step. The first retriever performs a semantic search based on cosine similarity, retrieving the top 20 product items most similar to the main product identified previously. Next, using this list of 20 products as context, we issue a new prompt assigning the role of an expert in women's fashion, whose task is to select the product description that best matches the intent. With this resulting description, the second retriever searches for the product that most closely matches it, without losing valuable and relevant information.
- **Get matching products**: The main task for this step is to generate a JSON output with a list of product items that are compatible with the main element of the outfit, taking into account its image, description, and the user's synthesized intent. Using one-shot prompting again, we assign $\mathcal{LLM}$-2 the role of an expert in women's fashion, tasked with analyzing the image of the product item and generating a list of other items for the outfit, considering the user preferences for style and occasion. We also provide an example of the expected output in JSON. This is a crucial step for structuring the outfit and facilitating the search for product items in our database.
- **Ensemble outfit**: The function of this step is to assemble the outfit starting with the main product item and the complementary ones returned by the previous step. It leverages key attributes from the JSON representation of items, such as occasion, style, and gender, to generate a suitable description of the outfit. The outcome of this step is a list containing the ensemble (with the items and their attributes) and a description of the overall outfit.
- **Check outfit**: In this final step, the agent evaluates the generated outfit to ensure its suitability. It receives as input the images and descriptions of the outfit, as well as the synthesized user intent. Through prompting, we assign $\mathcal{LLM}$-2 the role of a *Fashion Critic* and task it with the responsibility of analyzing the outfit's harmony, style, and compatibility in light of the user's preferences expressed in the intent. We also instruct the Fashion Critic to reject items that are redundant or inappropriate for the occasion. The outcome of this step is the rejection or approval of the outfit, along with a short justification for the decision. In case of a rejection, we use the justification as context and rerun the entire flow up to three times. If, by the third attempt, we do not get an approved outfit, we notify the user that the agent could not find a suitable outfit in the catalog.

The *CheckImage* component focuses on sanitizing the output generated by the *CreateOutfit* component. Specifically, it verifies whether the file paths of images for product items actually exist in our database. This serves as an additional guardrail that helps us detect potential hallucinations. The component returns a boolean value and invokes the *CreateOutfit* component when the value is false.

Finally, the *VestiME* agent generates a JSON document with an explanation of the recommendation, a short description of each product item included in the outfit, the reason for including each item, and the URL of the corresponding images. All this information is used to provide an answer to the end-user, along with a visual representation of the full outfit. The graphical user interface (GUI) of VestiME provides a chatbot-like experience, offering end-users a familiar conversational interface resembling ChatGPT, currenlty the most popular AI-powered chatbot[2].

V. Implementation and Validation

A. Implementation

VestiME was implemented with a combination of technologies, tools, and libraries commonly used in building modern RAG systems. The actual model used for $\mathcal{LLM}$-1 in Fig. 2 is Gemini-1.5-flash[3] (a multimodal LLM suitable for image description, captioning, and answering image-related questions). For $\mathcal{LLM}$-2, the model is GPT-4o-mini[4] (a lightweight LLM well-suited for real-time conversations and virtual assistants), while for $\mathcal{EMB}$, we used text-embedding-3-small[5] (a lightweight embedding model well-suited for effi-

[2]https://www.demandsage.com/chatgpt-statistics/?utm_source=chatgpt.com

[3]https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/1-5-flash

[4]https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[5]https://platform.openai.com/docs/models/text-embedding-3-small

cient, cost-sensitive applications). The choice of these models provides a good combination of effectiveness, efficiency, and cost for the purposes and budget of this research work. Our implementation facilitates the replacement of these models with newer and more performant alternatives. For orchestrating these models in *VestiME*'s architecture and workflow, we used LangChain. The vector database $\mathcal{VEC}$, used to store the embeddings generated by $\mathcal{EMB}$, is DeepLake[6].

The GUI was implemented using Streamlit (https://streamlit.io/), which provides a framework for web-based chatbot development. The programming language used for the implementation is Python 3.13. Fig. 4 shows a screenshot of *VestiME*'s main GUI.

B. Validation

We validated *VestiME*'s usability and ability to generate personalized outfit recommendations through a user study that followed a *think-aloud protocol* [36]. This method allowed us to capture the perceptions, thoughts, and potential issues faced by end-users while they interacted with the agent, providing insights into its suitability and areas for improvement.

In our user study, we involved a total of four subject matter experts with seniority ranging from mid-level to senior Fashion Designer/Consultant. We decided to involve these participants in order to obtain rich expert feedback on *VestiME*. Participation was voluntary and non-remunerated. Participant P1 (female, 21 years old) is a 4th-year undergraduate student from a Fashion Design program with experience using ChatGPT; participant P2 (female, 27 years old) is a senior Fashion Designer with more than 11 years of professional experience in the domain with no experience with ChatGPT; participant P3 (female, 22 years old) is a 4th-year undergraduate student from a Fashion Design program with experience with ChatGPT; and participant P4 (female, over 40 years old) is a senior Fashion Consultant and former University Professor with no experience with ChatGPT. Participants were video-recorded, and the facilitator/observer took notes during task performance for later analysis in combination with the video recordings.

Two documents were prepared for this study. The first document is an information sheet that describes the user study in detail and seeks the consent of the participant to take part in the study. The second document is a task sheet describing the tasks to be performed by participants during the think-aloud sessions. The whole user study was conducted in Spanish (including the chat interactions with *VestiME*). Specifically, participants were tasked with interacting with *VestiME* to receive outfit recommendations for the following five occasions:

- OC1: Job interview (in person).
- OC2: Vacations at the beach.
- OC3: Christmas party.
- OC4: Wedding.
- OC5: Open occasion (freely chosen by the participant).

[6]https://www.deeplake.ai/

Fig. 4. *VestiME*'s main GUI. This is the Spanish version of the tool used for our user study.

C. Analysis of Results

We start by reporting on the average time (in minutes) it took to complete each of the above tasks (see Table I). Although we report on the average time per task accross different participants, it is worth mentioning that experts took considerably different amounts of time per task, which were influenced by the amount of comments made, the number of turns it took to accept a recommendation, and the number of times *VestiME* called the *CreateOutfit* function to generate a recommendation. We next analyze the results of the study across the following themes: (i) Prompts and multi-turn dialogue, (ii) suitability of recommendations, (iii) domain-specific terminology, (iv) hallucination and other limitations, and (v) overall satisfaction with the tool.

Prompts and multi-turn dialogue

We begin our analysis of the results by examining the way participants interacted with *VestiME* through prompting. One noticeable characteristic was that participants who reported previous experience with tools such as ChatGPT were consistently more specific and detailed in writing their prompts. As a result, the tool achieved a faster and higher success rate when making recommendations and required fewer turns to find an outfit that was accepted by the participant.

For example, in the case of OC1 (job interview), participant P1 prompted the tool:

"Recommend me an outfit for a job interview. I prefer it to be formal and in neutral colors [...]." (P1).

For OC2 (vacations at the beach), the same participant

TABLE I
AVERAGE TIME TO COMPLETE EACH TASK AND FINALLY ACCEPT THE RECOMMENDED OUTFIT. THE REPORTED TIME INCLUDES THE COMMENTS MADE DURING THE THINK ALOUD PROCESS.

Task	Occasion	Avg. time (minutes)
OC1	Job interview (in person)	6.06
OC2	Vacations at the beach	5.65
OC3	Chirstmas party	3.97
OC4	Wedding	3.02
OC5	Open occasion (freely chosen by the participant)	2.99
	Avg. time accross occasions	**4.34**

prompted:

"I would like it (the outfit) to include sandals, without a platform, and a loose dress with a hat. I think I would like the colors to be pastel tones." (P1).

In both cases, the participant not only provided context in terms of the occasion but also gave hints on the preferred style, explicitly stating the colors. Instead, participant P2 used the following prompt for the same occasion:

"Hi, I'm looking for recommendations for a formal job interview in banking." (P2).

In this case, while the participant provided context for the occasion, she did not give any indications regarding style preferences. As a result, the tool needed additional turns to prompt the user and inquire about her preferences for style, colors, etc.

As participants became comfortable with the tool, they began experimenting with the recommended outfits by changing some of the elements or styles. For example, for OC2 (vacations at the beach), participant P4 used multiple turns to try different elements of the initially recommended outfit, e.g. by making it explicit that she wanted:

"a crop top [...] with a thread skirt with bright colors [...]." (P4).

Multi-turn interactions like these were common throughout the study, demonstrated participant engagement, and helped facilitate product item discovery in the catalog. While there were cases where the recommended outfit was accepted on the first attempt, in other cases where participants were more exploratory or simply not fully satisfied with the recommendation, it took up to 7 turns before they finally accepted the recommendation.

Suitability of recommendations

We focus now on the suitability of the outfit recommendations made by *VestiME*. Except for a couple of cases where the tool showed hallucination (which we will discuss in more detail later in this section) or faced other issues (e.g. not displaying the outfit), the outfits recommended by *VestiME* were largely appreciated by the participants, in many cases on the first attempts, or after a number of turns where they "played" with the outfit options or provided more context. Regarding the latter, for OC3 (Christmas party), when participant P1 prompted the tool, she received an outfit recommendation made of warm clothes, to which she immediately responded:

"The outfit is very nice; I love it! But I forgot to mention that over here (The Southern Hemisphere) we have Christmas in summer!" (P1).

After an additional turn, where the participant further indicated that she needed a cool dress for summer, she finally received something more convincing and expressed:

"I find this very nice. It is an interesting one (the outfit) and it includes green colors, which gives the impression of a Christmas style." (P1).

We learned from this that *VestiME* shows the ability to understand and dynamically adapt to external conditions (weather, season, etc.), subject to the user providing enough contextual information.

In other cases, *VestiME* was able to provide outfit recommendations that were immediately accepted without further turns. For the most part, these were cases where the participant had a clear idea and preference for the outfit and intentionally expressed them in their prompt. For instance, for OC4 (wedding), participant P2 asked for a long, yellow dress for a hot weather, with a mix of golden and silvery accesories, and a black purse. When reviewing the recommendation (see Fig. 5), the participant expressed:

"I like it, absolutely yes! I love it, I swear I love it! I wouldn't change a thing." (P1).

The participant even went further and responded to *VestiME*:

"Congrats, well done! I love this outfit." (P1),

showing a great deal of satisfaction with the outcome. Furthermore, for occasion OC5 (open occasion), participant P3 opted for a romantic dinner with her fiancé and prompted:

"[...] a black dress that highlights my body shape." (P3).

When reviewing the recommended outfit, the participant expressed with noticeable satisfaction:

"Wow, Oh my God! I now only need to find a fiancé!" (P3).

"My God, this is beautiful! I love it. Where can I buy this?" (P3).

It is worth noting that fun, laughter, and engagement remained consistent throughout the study for all participants.

Fig. 5. Outfit recommended to participant P2 for OC4 (wedding). Participant's prompt: *"I have a formal wedding, during a hot summer, I'm looking for an elegant, yellow maxi dress, with golden and silver accesories, and black clutch"* (P2). The outfit description was translated from Spanish.

Participants also expressed satisfaction with *VestiME*'s ability to choose the combination of colors and tones, demonstrating a good grasp of fundamental principles of colorimetry. For example, for occasion OC5 (open occasion), participant P1 chose a date at a restaurant as the occasion and asked for a dress with a romantic style. Upon receiving the recommendation, P1 expressed her satisfaction with the outcome and emphasized that even without specifying the color, by simply prompting that she wanted a romantic style, the tool automatically adapted the outfit to colors aligned with the sytle, expressing:

"It (VestiME) knows about the colors that align with the idea of romantic." (P1).

Domain-specific terminology

Throughout the study, participants also noticed *VestiME*'s ability to describe the outfit in great detail using domain-specific terminology. For instance, participant P2 specifically pointed this out by expressing:

"I like that it (the description) is fairly specific, because it says here that it is a "maxi white dress". This is very good. Terminologies such as maxi, mini and midi is a specific (has a well-defined meaning) and technical term used in the fashion domain." (P2).

Similar impressions were shared regarding the tool's use of specific, technical terminology and jargon in the fashion domain related to colorometry (e.g. by using sophisticated colors from the color palette and their combinations), nature and texture of fabric (e.g. chiffon, crêpe, tulle, silk), footwear (e.g. stiletto, sandals, ankle strap heels), dress styles (e.g. v-neckline, bodycon dress, coral off-the-shoulder top), and more. Furthermore, participants praised not only the mere presence of such terminology in the description of the outfit but also the precision and coherence of its usage w.r.t. the prompted preferences and the visual representation of the outfit.

Hallucination and other limitations

As in the majority of LLM- and GenAI-based systems, *VestiME* is not completely free of potential hallucinations and errors in the provided responses [37]. For OC4 (wedding), participant P1 experienced problems with the tool, which generated a description of the outfit but was unable to show the items. We suspect that this is because the response (text) did not match any of the product descriptions in our catalog. In response, P1 prompted the tool:

"I cannot see the outfit, please make your suggestion once again with the actual images (of the outfit)." (P2).

This time, the tool was able to produce an outfit description with the visual representation, which participant P2 was satisfied with and accepted.

There were other situations where the tool was unable to fully satisfy the participant's preferences. For instance, for OC1 (job interview), participant P2 asked for pants (preferably beige), a white shirt, a blazer of the same color as the pants, and black stilettos. Yet, the tool did not fully follow the participant's preferences recommending a white shirt, beige blazer, black pants (which should also have been beige), and black stilettos. After a couple of additional attempts, the participant finally received an outfit that satisfied her preferences.

Another limitation that we observed in some cases was that the tool was not always successful in modifying a previously recommended outfit. For example, it made a recommendation with the specific requested modification but an overall outfit that was different. This happened when, for example, participants asked for a change in color (e.g. a white purse instead of black) or style (e.g. strapless dress instead of spaghetti straps) of a product item.

Overall experience and satisfaction with the tool

In order to capture the participants' perception of the overall experience and satisfaction with *VestiME*, at the end of the think-aloud session, we explicitly asked them for feedback in this regard. All participants responded positively when asked whether they would use this tool again and recommend it to others. Specifically, participant P1 expressed

"[...] I would recommend this tool to everybody." (P1),

arguing that anyone would need this type of tool and guidance at some point. Participant P4, on the other hand, shared that she would definitely use the tool with friends and others because

" [...] it would make it much easier for me to explain the outfit to others." (P4),

showing the potential of this tool not only as a product recommender system in the online store and retail business but also as an aid for fashion and image consulting.

Another pinpointed aspect was that, although the recommended outfits were widely accepted and liked by the participants, some of these items were not very updated or fresh. That is, the tool sometimes recommended outdated product items that are currently not in high demand in the *fast fashion* world but can be easily found in stores such as Marshalls (https://www.marshalls.com/) and Ross (https://www.rossstores.com/), which focus on past fashion seasons. In light of this feedback, we clarified to the participants that this limitation is due to the fact that we are relying on a catalog of products used in studies that date back to the year 2022, which leads to recommendations with product items that are not fully aligned with the most recent trends. In this sense, one of the participants pointed out (and we fully agree with the comment) that this would be easy to fix:

"[...] if a catalog from Forever 21[7] *or Zara*[8] *is used instead, because it will feed the database with product items that are*

[7]https://www.forever21.com/
[8]https://www.zara.com/

aligned with their fast fashion and weekly shifts." (P2).

When comparing this tool with others typically used to search for outfit ideas, one of the participants said that a tool like *VestiME* is much more focused and specific than the alternative option of looking for ideas and recommendations in applications such as Pinterest (https://www.pinterest.com/) where, although one can find many options, the experience can ultimately be very frustrating. In particular, participants appreciated the fact that *VestiME* is able to justify its recommendations:

"It (VestiME) provides an explanation of why it is recommending the outfit. This is very important." (P3).

We observed that having such explanations contributes to generating confidence and trust in both the recommended outfit and the tool.

Discussion

The user study provided expert validation of *VestiME*'s ability to recommend personalized outfits for different occasions. Throughout the study, we observed a high degree of accuracy in the recommendations, clearly reflected in the appropriate selection of colors, styles, combinations, and the overall satisfaction of the experts. In this section, we further discuss the results from the perspective of the proposed solution's effectiveness, efficiency, and ease of use.

Effectiveness. In the study, all the assigned tasks were completed successfully, indicating that the tool fulfilled the experts' expectations. We did not observe critical errors that prevented them to fulfill their tasks or caused the system to halt. However, we observed non-critical issues such as potential hallucinations and lack of visual representation of the outfit, which were easily overcome with further interactions (turns) with tool by the participants. In the latter case, participants highlighted that even when no visual representation of the outfit was provided (which happened in a couple of cases), the descriptions were still clear enough to provide a good understanding of the outfit and its components.

Efficiency. Time and the number of turns required to complete a task are two key aspect when building tools such as chatbots and virtual assistants. The average time for *VestiME* to provide a recommendation given a prompt was approximately 55 seconds. Compared to conventional chatbots, where answers are provided in about 10 seconds [10], [11], this delay can be considered high for end-users accustomed to faster response times. This is a relevant area of improvement if this tool is to be productionized and used in real settings. On the other hand, most of the outfit recommendations were accepted by the experts in less than three turns, with the average number of turns being two (rounded to the nearest whole number). This indicates a good performance by *VestiME* in making the adjustments needed based on the initial prompt or requirement from the experts.

Ease of use. We observed that during the study, the participants were able to navigate the tool intuitively and without extensive explanations or instructions, which suggests that the combination of a conversational interface, GUI, and standard chatbot *look and feel* contributed to the familiarity with the tool and its ease of use. As previously explained, in all cases, participants expressed their satisfaction with the tool, stating that they would both use it again and recommend it to others.

VI. Conclusion

This paper presented a RAG-based multimodal fashion recommender system for outfits based on a catalog of single product items. Our proposed solution aims to provide an alternative to traditional recommender systems, where interactions are typically limited to keyword searches, menu navigation, and clicks on recommended items, where the recommendations are typically based on past searches, item- and user-based collaborative filtering and other traditional methods. We leverage state-of-the-art multimodal LLMs, embedding models, prompting, LLM orchestration, and RAG to propose *VestiME*, our conversational agent for outfit recommendation, exploration, and discovery.

Our user studies revealed the viability, usability, and potential of *VestiME*. All four experts involved in the study evaluated the tool positively, expressing that they would use the tool again and recommend it to others. Some of the positive aspects explicitly highlighted by participants include the suitability of the recommended outfits based on their expressed intents and expectations; the correct and accurate use of terminology and jargon from the fashion domain when presenting outfit descriptions; a good notion of style, combination, and colorimetry based on the occasion, among other advantages. The observers of the study also identified strong engagement from participants during the interaction with the tool, and a sense of curiosity and fun when using *VestiME*.

This work comes with its own limitations. First, as commonly happens when using LLM and GenAI technologies, *VestiME* is not free from potential errors and hallucinations. This typically results in the tool being unable to find matching product items that fit the description of the recommended outfit. In very few cases, we also observed that the tool was unable to fully satisfy the user's requirements, for example, having pants of the same color as the blazer. At times, this is due to the fact that we simply cannot fulfill the requirement because we do not have such an item in the catalog (e.g. we do have blazers in beige, but not pants in that same color); ideally, the tool should be able to explain and clarify this to the user when such situations happen. The tool needs to improve its ability to modify specific items from the outfit (while keeping the rest of the items unchanged) as per the user request. Finally, data privacy should also be taken into account to prevent data leakage when used in production environments.

In future work, we plan to address each of the limitations listed above, prioritizing the errors identified in our user study and the ability to modify specific items, which were perceived as the most important issues. We are also interested in exploring the possibility of evolving *VestiME* into a truly Agentic AI solution that integrates with real-time APIs (for

catalogs) and domain-specific, multimodal LLMs by leveraging emerging technologies such as Model Context Protocol (MCP) [38]. Finally, we plan to expand our user validation through quantitative research methods.

REFERENCES

[1] J. Entwistle, "The cultural economy of fashion buying," *Current Sociology*, vol. 54, no. 5, pp. 704–724, 2006.

[2] "Wikimedia Commons: Illustrated fashion catalogue, summer, 1890," https://commons.wikimedia.org/wiki/File:Illustrated_fashion_catalogue_-_summer,_1890_%281890%29_%2814780822881%29.jpg, Accessed: 2025-08-10.

[3] Y. Deldjoo, F. Nazary, A. Ramisa, J. Mcauley, G. Pellegrini, A. Bellogin, and T. D. Noia, "A review of modern fashion recommender systems," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–37, 2023.

[4] M. Xu, W. Yin, D. Cai, R. Yi, D. Xu, Q. Wang, B. Wu, Y. Zhao, C. Yang, S. Wang *et al.*, "A survey of resource-efficient LLM and multimodal foundation models," *arXiv preprint arXiv:2401.08092*, 2024.

[5] Y. Ding, Z. Lai, P. Mok, and T.-S. Chua, "Computational technologies for fashion recommendation: A survey," *ACM Computing Surveys*, vol. 56, no. 5, pp. 1–45, 2023.

[6] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–38, 2024.

[7] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] Y. Wu, C. Macdonald, and I. Ounis, "Multimodal conversational fashion recommendation with positive and negative natural-language feedback," in *Proceedings of the 4th Conference on CUI*, 2022, pp. 1–10.

[10] Sapna, R. Chakraborty, K. Vats, K. Baradia, T. Khan, S. Sarkar, and S. Roychowdhury, "Recommendence and fashionsence: Online fashion advisor for offline experience," in *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, 2019, pp. 256–259.

[11] B. Zhang and Q. Wang, "Outfit helper: A dialogue-based system for solving the problem of outfit matching," *Journal of Computer and Communications*, vol. 7, no. 12, pp. 50–65, 2019.

[12] V. Husak, O. Lozynska, I. Karpov, I. Peleshchak, S. Chyrun, and A. Vysotskyi, "Information system for recommendation list formation of clothes style image selection according to user's needs based on NLP and chatbots." *COLINS*, vol. 2604, pp. 788–818, 2020.

[13] T. E. Kolb, A. Wagne, M. Sertkan, and J. Neidhardt, "Potentials of combining local knowledge and LLMs for recommender systems," in *Proceedings of the Fifth Knowledge-aware and Conversational Recommender Systems Workshop co-located with 17th ACM Conference on Recommender Systems (RecSys 2023)*, vol. 3560. CEUR-WS. org, 2023, pp. 61–64.

[14] N. Forouzandehmehr, Y. Cao, N. Thakurdesai, R. Giahi, L. Ma, N. Farrokhsiar, J. Xu, E. Korpeoglu, and K. Achan, "Character-based outfit generation with vision-augmented style extraction via LLMs," in *2023 IEEE International Conference on Big Data*. IEEE, 2023, pp. 1–7.

[15] W. Chen, P. Huang, J. Xu, X. Guo, C. Guo, F. Sun, C. Li, A. Pfadler, H. Zhao, and B. Zhao, "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.

[16] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha, "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 765–774.

[17] K. Zhao, X. Hu, J. Bu, and C. Wang, "Deep style match for complementary recommendation." in *AAAI Workshops*, vol. 17, 2017.

[18] R. He and J. McAuley, "VBPR: Visual bayesian personalized ranking from implicit feedback," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[19] V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan, "Large scale visual recommendations from street fashion images," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1925–1934.

[20] M. Yang and K. Yu, "Real-time clothing recognition in surveillance videos," in *2011 18th IEEE international conference on image processing*. IEEE, 2011, pp. 2937–2940.

[21] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 43–52.

[22] S. Jaradat, "Deep cross-domain fashion recommendation," in *Proceedings of the 11th ACM conference on RecSys*, 2017, pp. 407–410.

[23] C. Bracher, S. Heinz, and R. Vollgraf, "Fashion dna: merging content and sales data for recommendation and article mapping," *arXiv preprint arXiv:1609.02489*, 2016.

[24] Z. Al-Halah and K. Grauman, "From Paris to Berlin: Discovering fashion style influences around the world," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 136–10 145.

[25] Y. Liu, Y. Gao, S. Feng, and Z. Li, "Weather-to-garment: Weather-oriented clothing recommendation," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 181–186.

[26] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Hipster wars: Discovering elements of fashion styles," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 472–488.

[27] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, "Neuroaesthetics in fashion: Modeling the perception of fashionability," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 869–877.

[28] S. C. Hidayati, C.-C. Hsu, Y.-T. Chang, K.-L. Hua, J. Fu, and W.-H. Cheng, "What dress fits me best? Fashion recommendation on the clothing style for personal body shape," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 438–446.

[29] X. Han, X. Song, J. Yin, Y. Wang, and L. Nie, "Prototype-guided attribute-wise interpretable scheme for clothing matching," in *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, 2019, pp. 785–794.

[30] M. Hou, L. Wu, E. Chen, Z. Li, V. W. Zheng, and Q. Liu, "Explainable fashion recommendation: A semantic attribute region guided approach," *arXiv preprint arXiv:1905.12862*, 2019.

[31] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[32] D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous intelligence for complex goals – A comprehensive survey," *IEEE Access*, 2025.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[34] Y. Li, "A practical survey on zero-shot prompt design for in-context learning," *arXiv preprint arXiv:2309.13205*, 2023.

[35] G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger, "Efficient algorithms for t-distributed stochastic neighborhood embedding," *arXiv preprint arXiv:1712.09005*, 2017.

[36] M. E. Fonteyn, B. Kuipers, and S. J. Grobe, "A description of think aloud method and protocol analysis," *Qualitative health research*, vol. 3, no. 4, pp. 430–441, 1993.

[37] G. Perković, A. Drobnjak, and I. Botički, "Hallucinations in LLMs: Understanding and addressing challenges," in *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*. IEEE, 2024, pp. 2084–2088.

[38] X. Hou, Y. Zhao, S. Wang, and H. Wang, "Model Context Protocol (MCP): Landscape, security threats, and future research directions," *arXiv preprint arXiv:2503.23278*, 2025.