FashionAlign: Multimodal Search System for Fashion Exploration

Eunhye Kim, Taehyun Yang, Byeolyi Yoon

Seoul National University

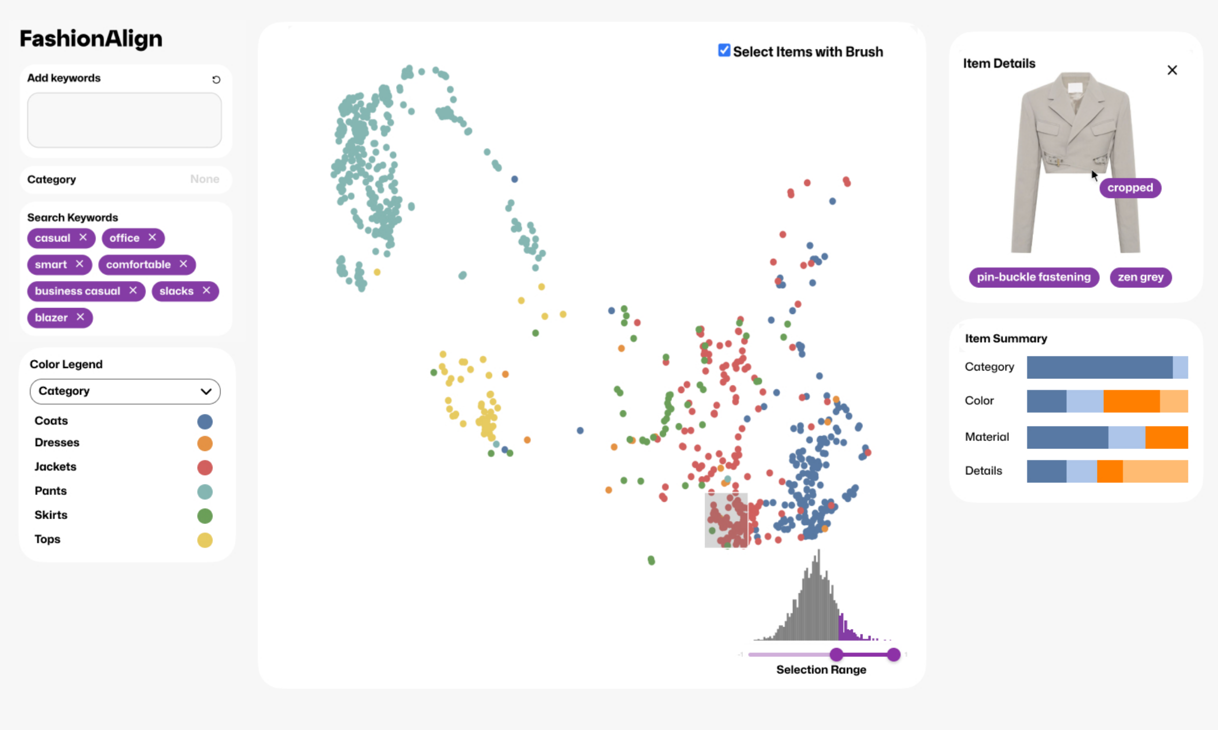

Figure 1: The user interface of *FashionAlign* consists of the left sidebar, center map view and right sidebar. The left side bar allows the user to enter a query and control the map colors. The map view allows navigation of the fashion space based on visual features and similarity range. The right sidebar allows the user to hover over item(s) to find the textual descriptions for visual features.

Abstract

In the field of searching for fashion items, novice users face the challenge of query formulation and articulating their style preferences into searchable terms, due to the lack of domain knowledge. This disconnect between user intentions and search capabilities results in unsatisfactory experiences, where the search result may not capture the users intentions. Addressing this issue, we introduce *FashionAlign* as our proposed solution, a system designed to bridge this gap by facilitating natural language-based searches and semantically aligning textual descriptions and visual features. Employing parameter optimized UMAP visualization, *FashionAlign* transforms abstract style descriptions into a curated visual display of fashion items. Through a user study, we prove that this approach demonstrates marked improvements in user experience, notably in matching user queries with relevant fashion items, and accurate refinement of queries. Our full system can be found on our GitHub page: https://github.com/gracekim027/FashionAlign.

Index Terms: Human-centered computing—Fashion Technology—Interactive Search Systems;

1 Introduction

The integration of large language models and advanced vision technologies has reshaped the landscape of fashion technology, introducing new possibilities in the fashion domain [8, 10]. This change has introduced new systems like FashionIQ [17], StyleFinder [4], and Deepstyle [15], each harnessing these technologies to enhance online fashion search and recommendations.

Moreover, AI-powered image recognition and processing technologies have opened up new doors for the fashion item cataloging. [16]. These systems can accurately identify and tag various fashion elements from images, making it easier for users to find specific styles. AI has also played a crucial role in trend forecasting, analyzing emerging patterns in fashion data to predict future trends, aidng designers and retailers to predict user's personal style interests [14]. AI has also opened up creative avenues in fashion design and customization. Algorithms can now generate unique fashion keywords based on their visual inputs, offering a scalable method to query human-designed fashion items [12]. This synergy of AI with human designers is creating new opportunities in AI-powered fashion systems.

However, these systems and technology often cater to a more technical audience and are not readily accessible or intuitive for the average end-user [13]. Their complexity often necessitates a significant level of domain specific knowledge, and they typically do not support the kind of exploratory, user-driven browsing experience that many shoppers seek. This creates a gap between the capabilities of these advanced systems and the practical needs of everyday fashion consumers.

In response to these limitations, we introduce *FashionAlign*, a system that stands apart in its user-centric approach. Unlike previous sytems, *FashionAlign* is designed with the end-user in mind, focusing on ease of use without compromising on the underlying technology [18]. It offers a solution that bridges the gap between complex fashion search algorithms and user-friendly interfaces. It introduces a new solution for previous systems: the need for technical expertise and the lack of support for intuitive browsing and query refinement. At its core, *FashionAlign* utilizes a Large Language Model (LLM) powered natural language processing system, coupled with a parameter-optimized Uniform Manifold Approximation and Projection (UMAP) for visualizing fashion items [12]. This unique combination allows users to input abstract style descriptions, which the system then translates into a visually curated display of fashion items, facilitating an intuitive and dynamic exploration of fashion options.

In summary, our major contributions through **FashionAlign** include:

1. The development of an natural language-based search interface that simplifies the need for accurate domain specific knowledge for novice users.

2. The integration of UMAP visualization with image segmentation and labeling.

3. A user study that demonstrate the effectiveness and usability of our system.

2 Related Works

Our work aligns closely with three primary areas: (i) query (re)formulation, (ii) clothing retrieval and datasets, and (iii) multimodal search models.

Query (Re)formulation In the realm of query (re)formulation, the Berrypicking Model [2] suggests that search queries are dynamic and evolve over time. Real-life searches involve users starting with a single feature or relevant reference and gradually exploring various sources. Instead of presenting information in a comprehensive set, the search system should deliver it in smaller increments, such as color, design, or shape in the domain of fashion. Previous research, including that by Kelly and Fu [9], emphasizes the importance of additional information, such as domain knowledge, information needs, and search motivation in enhancing retrieval performance. Notably, current fashion systems often overlook the user's domain knowledge, given the varied and abstract nature of fashion descriptions. PromptMagician [5] addresses this issue by helping users refine their search prompts, bridging the gap in user knowledge in the domain of image generation.

Clothing Retrieval and Datasets Concerning clothing retrieval and data sets, various works have contributed to data sets like DeepFashion [11], which includes labeled images with detailed textual information. FashionNet [7] predicts landmark locations, while FashionIQ [17] provides human-generated captions and real-world product descriptions. DeepStyle [15] proposes a search engine that combines visual cues for aesthetically similar multimedia database retrieval. Previous work emphasizes the importance of detailed textual information in the fashion domain.

Multimodal Search Models In the domain of multimodal search models, Chia et al. [3] introduced FashionCLIP, a model adapted for the fashion industry. They underscore the benefits of domain-specific fine-tuning for real-world product search. Additionally, Baldrati et al. [1] demonstrated that CLIP-based features can effectively support conditioned image retrieval systems, incorporating user feedback as natural language input. While previous research has delved into the technical aspects of multimodal search systems, there is a noticeable gap in understanding the interface design for these systems.

3 System Overview

FashionAlign is designed to interpret user search queries in natural language, and refinement the query using a large language model. The most distinctive feature is the clustering of fashion items in a grouped format based on their similarity of each visual features. With additional components that provide textual alignment of visual features, users can navigate, compare, and select from the fashion space.

3.1 Design Requirements

The system is structured to accomplish three major tasks: visual browsing, semantic alignment, and query refinement, each addressing a specific aspect of the fashion item search process. These tasks are supported by four key requirements that guide the system's functionality.

R1: Visual Overview Map: The system should provide a map-based overview of fashion items, organized according to text and image parameters. This feature allows users to get a comprehensive view of the available fashion space.

R2: Keyword Recommendations: To further aid in refining searches, the system should offer keyword recommendations. These suggestions are tailored to align with the visual aspects of items that users show interest in during their search process.

R3: Textual-Visual Description: Following a search, the system provides a detailed textual-visual description of the results. This should allow users to gain a better understanding of the the terminology of fashion items, and help them refine their textual query to describe their initial intentions best.

R4: Breakdown of User Queries: The system should assist users in deconstructing their initial, often abstract, queries into specific fashion attribute keywords. This functionality is particularly useful for users who may not have precise knowledge of how to articulate their style preferences.

3.2 Components

This section details the various components that make up the system.

Query Input View serves as the starting point of the search session. Users can input their search queries in natural language, accommodating queries that range from specific item requests to more abstract style descriptions.

Query Breakdown View shows the keyword extractions of the user query. Utilizing the GPT-4 model, the system breaks down the natural language query into meaningful keywords. The breakdown view also includes keywords that are recommended by the GPT model to help make an abstract query be specific with accurately capturing the user's intent.

Map View is the primary window for users to explore the fashion space. In the map view, fashion items are placed based on their image embeddings generated by FashionCLIP, which have been dimension-reduced using Uniform Manifold Approximation and Projection (UMAP). This method effectively groups items that share similar categories or visual attributes, resulting in intuitive clusters for easy browsing. The most optimal parameters have been picked using the parameters (**mindist** was set to 0.0, **nnneighbor** was set to 15, and **spread** was set to 1.0). This parameters have been picked by reiterating through every possible parameter and asking five users which cluster seems to have the best visibility.

Color Control View allows users to customize the color scheme of the map view, which aids in visually distinguishing between different categories or attributes. When a category is defined in the query, users can select from category-specific attributes.

Map Control View enables users to filter items based on their overall similarity to the query. This similarity, which is the cosine similarity of the item image embedding and the text embedding of the query, is recalculated dynamically as the keywords change.

Item Detail View shows a one-to-one alignment of visual features and keywords. As a user hovers over an item in the map view, the system displays the item image along with relevant keywords. When a user hovers over a specific part of an item's image, the corresponding keyword is highlighted. The user can then choose to add this keyword to their search. The keywords were generated using the GPT-4 vision model, while the local attributes in the images were labeled through fine-tuning a SAM model. Subsequently, both the keywords and bounding boxes underwent supervision by a domain expert for refinement.

Item Summary View provides a diagram-based summary of multiple items selected by the user. When the user brushes over items in the map view, the system displays the a stacked bar chart of the features. When the brushed items are of different category, the bar chart's axis is set to global attributes. When the the brushed items are of same category, the bar chart's axis is set to category specific attributes. To provide an overview of the selected items, the summary was generated using a KMeans clustering technique, with the maximum number of clusters set to four.

4 User Study

We conducted a user study to evaluate the effectiveness and usability of our system in natural language fashion search and query refinement. Specifically, we aim to evaluate (1) the helpfulness of the query refinement model, (2) the usability of the overall visual browsing flow, and (3) the image to text alignment support.

4.1 Participants

We recruited five participants (four females and one male, aged 22-30) from a local university. The participants are mainly undergraduate and master's students from various disciplines. We recruited mostly female users for the test since our system currently supports only female fashion data. They had a keen interest in fashion, but lacked specific domain knowledge on the terminology.

4.2 Procedure and tasks

The user study was conducted in four stages: introduction, training, visual exploration, and questionnaire. During the training and visual exploration tasks, the users were instructed to speak out their thoughts and feelings (Think-aloud) while manipulating the system.

Introduction: We first provided an introduction to our research background, including the research motivation and the study protocol. We then introduced the components of our system and demonstrated their usage with examples.

Training task: To help users become familiar with the system, we set a training task where the participant was given a random image of an item, and the participant was required to find it using the system components.

Visual exploration and search: In this stage, the participants were required to think of an item that they would like to buy. The initial search goal could be both abstract, such as a specific style, or detailed, such as an item of a specific category and design. The participants were required to utilize all components at least once regardless of order.

Post-study Questionnaire: For the heuristic evaluation, we referred to the questionnaire made by Granollers [6]. We edited a few questions that were not applicable to our system. The resulting principle set has 11 principles, with a total of 22 questions. For each question, the scale was given from 1 (strongly agree) to 5 (strongly agree).

4.3 Results Analysis

4.3.1 Effectiveness of the Query Refinement Model

Participants found the keywords recommended by the system meaningful and easy to understand. For example, during the exploration task, when Participant 2 (P2), searched for *"casual yet professional workwear,"* the system updated the query to less abstract keywords such as *"smart casual," "blazers,"* and *"loafers."* P2 remarked, The keywords perfectly encapsulated the style I had in mind but couldn't articulate. Participant 4 (P4) sought a *"bohemian summer look"* and was introduced to keywords like *"maxi dress"* and *"gladiator sandals,"* items they hadn't originally considered. P4 expressed delight in discovering new styles that fit their desired aesthetic. The refinement model significantly enhanced the aesthetic appeal and specificity of searches. Participant 5 (P5) wanted to find an *"elegant evening gown"* and was presented with keywords like *"silk," "A-line,"* and *"off-shoulder,"* which helped them refine their search to find exactly what they envisioned. P5 appreciated how these keywords *"transformed a broad concept into a precise fashion statement."*

4.3.2 Effectiveness of the Visual Browsing System

The *Map View*, which color-codes clusters and items based on the selected attribute, was appreciated by participants for easy understanding of the fashion space. Participant 4 (P4) expressed, *"The color-coded map legend made it easy to identify different materials at a glance."* However, some people also thought that it made the cluster overcrowded as in a larger view, participants were unable to see any important detail from the map legend. The system's ability to cluster clothing features based on similarity scores was mostly appreciated for its intuitive use. Users found that the zoom feature, which allows for a closer inspection of individual clothing items, significantly enhanced their browsing experience. For instance, Participant 1 (P1) noted, *"Zooming into clusters was very natural and it gave me a clear view of each item, making it easier to find what I was looking for."*

4.3.3 Image and Text Alignment Support

Most participants agreed that the *Image Detail View* and *Image Summary View* helped them understand the relationship between visual features and textual description. Participant 3 (P3) remarked, *"The hover feature that reveals clothing details is interesting as we can search more keywords based on the indicated features."* Participant 1 (P1) expressed their appreciation for how easily they could align their mental image of a desired outfit with the actual items presented by the system. They noted, *"The alignment between what I had in mind and what the system showed me was impressively accurate."* Participant 2 (P2) highlighted the usefulness of individual keyword displays, saying, *"Being able to see each keyword linked to an item helped me understand why it was suggested and explore similar styles more efficiently."* This feedback underscores the system's effectiveness in bridging the gap between visual and textual fashion queries.

4.3.4 Usability

All participants agreed that the links of the components were clearly defined within our system, and that the information appears in a logical order. The participants thought that the workflow of the system was intuitive, and the information was displayed consistently. Many participants pointed out that there was some latency when creating the item summary, and that it was not easy to understand the system status when there was latency.

5 Discussion

In this study, we found that *FashionAlign* helps users to easily generate and refine text prompts.

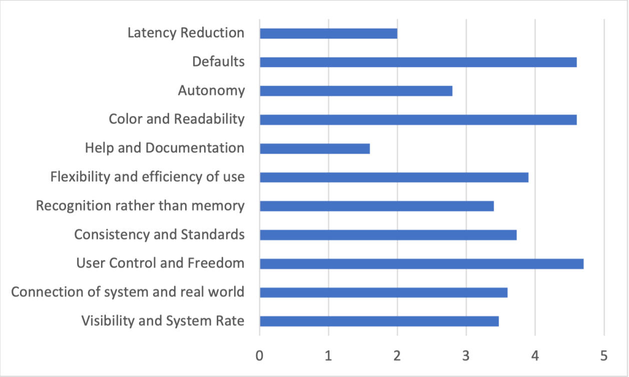

Figure 2: The results of the usability questionnaire. The questionnaire was comprised of 11 principles and a total of 22 questions.

Abstract to non-abstract workflow *FashionAlign* employs a depth-first search approach to navigate the fashion space. This approach targets the problem scenario that novice users who lack domain knowledge find it hard to articulate features that can be visually imagined. With the user study, we noticed that many novice users started the search with an abstract natural language query, and refined the query to be more non-abstract with domain specific keywords. Users found the one-to-one mapping of visual features and textual descriptions help navigate this depth-first search flow. However, some commented that a backwards approach, where the summary elements show not only specific keywords but also an abstract keyword that captures the style or aesthetic, might help them understand the fashion space as well.

Quick view of the fashion space By presenting items in clusters of visual features, *FashionAlign* helps users navigate the fashion space, compared to the traditional method of layering in a grid format. By allowing users to change the range of similarity scores, the items that appeared on the map changed dynamically, and this feature helped users match textual descriptions and visual attributes. However, some commented that the map does not show hierarchical information, such as which item has the highest score within the range, and wished that the rank could be shown through dot size or a supplementary list.

Latency and system status Due to the use a large language model API during the keyword breakdown, feature clustering and summary generation process, the users commented on the latency of the system. Initially, *FashionAlign* did not incorporate descriptions of why the feedback took longer than expected, and many users found it hard to understand the system status. This proved a need to include faster feedback and an indicator of the system status in future development.

5.1 Limitations and Future Work

FashionAlign currently allows users to configure textual queries. Other hyper-parameters such as determining the level of abstraction in the recommended keywords or summary can improve the user gain textual knowledge of the domain. The system also currently only provides data for six categories for women's clothing items. In the future, we plan to expand the database to include more categories and items.

6 Conclusion

Our study presents *FashionAlign*, an innovative system that combines Large Language Models (LLM) and visualization methods, providing a dynamic approach to fashion search and recommendations. We address the challenges related to formulating queries that demand domain-specific knowledge by offering a depth-first search solution. Key contributions of our system encompass the creation of a natural language-based search interface and the alignment of textual and visual features. Through a user test involving five inexperienced users, the findings indicate that *FashionAlign* effectively suggests relevant keywords, thereby improving the overall discovery experience.

References

[1] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo. Conditioned image retrieval for fashion using contrastive learning and clip-based features. In *ACM Multimedia Asia*, pp. 1–5. 2021.

[2] M. J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online review*, 13(5):407–424, 1989.

[3] P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Magalhães, D. Goncalves, C. Greco, and J. Tagliabue. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12(1):18958, 2022.

[4] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *Proceedings of the IEEE Conference on computer vision and pattern recognition workshops*, pp. 8–13, 2013.

[5] Y. Feng, X. Wang, K. K. Wong, S. Wang, Y. Lu, M. Zhu, B. Wang, and W. Chen. Promptmagician: Interactive prompt engineering for text-to-image creation. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[6] T. Granollers. Usability evaluation with heuristics, beyond nielsen's list. In *The Eleventh International Conference on Advances in Computer-Human Interactions (ACHI 2018)*, 2018.

[7] T. He and Y. Hu. Fashionnet: Personalized outfit recommendation with deep neural network. *arXiv preprint arXiv:1810.02443*, 2018.

[8] I. Kang, S. Ruan, T. Ho, J.-C. Lin, F. Mohsin, O. Seneviratne, and L. Xia. Llm-augmented preference learning from natural language. *arXiv preprint arXiv:2310.08523*, 2023.

[9] D. Kelly and X. Fu. Eliciting better information need descriptions from users of information search systems. *Information Processing & Management*, 43(1):30–46, 2007.

[10] C. Li, X. Li, M. Chen, and X. Sun. Deep learning and image recognition. In *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, pp. 557–562. IEEE, 2023.

[11] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104, 2016.

[12] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[13] R. Miñón, L. Moreno, P. Martínez, and J. Abascal. An approach to the integration of accessibility requirements into a user interface development method. *Science of Computer Programming*, 86:58–73, 2014.

[14] S. Seymour. *Fashionable technology: The intersection of design, fashion, science, and technology*. Springer, 2008.

[15] I. Tautkute, T. Trzciński, A. P. Skorupa, Ł. Brocki, and K. Marasek. Deepstyle: Multimodal search engine for fashion and interior design. *IEEE Access*, 7:84613–84628, 2019.

[16] Y. Tian. Artificial intelligence image recognition method based on convolutional neural network algorithm. *IEEE Access*, 8:125731–125744, 2020.

[17] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11307–11317, 2021.

[18] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12647–12657, 2021.