VastrAI: Conversational Fashion Outfit Generator Enhanced with GenAI and CLIP-BM25 Models

Manikumari Illa*1, P.Divya Bhavana*2, G.Priyanka Devi*3, A.Bharath Chandu*4, A.Vivek*5, T.Eswar Hanok Kumar*6
[1-6]Department of AI&DS, Vignan's Institute of Information Technology, Visakhapatnam, India
{ [1]mani.illa95, [2]20l31a5450, [3]20l31a5427, [4]20l31a5401, [5]20l31a5405, [6]20l31a5456 }@vignaniit.edu.in

Abstract

In general, It is very challenging for people to choose the right Outfit for the respective occasion. People have a collection of dresses but sometimes make mistakes in wearing the outfit. We describe an approach for this task based on a transformer that auto-regressively models the text and generates image tokens as a single data stream. So our project is to create a generative AI model that can suggest the best outfit for the people based on the occasion given through the prompt in the model. It also describes the outfit and explains the impression that will be created after wearing the outfit. This idea has been originated from the text-to-image conversion model using OpenAI where it generates the images from the text given through the prompt. The outfits that are generated include top wear, bottom wear, footwear, earrings, watches, etc. With available data, our approach is competitive with previous domain-specific models.

Keywords: OpenAI, Prompt, Domain-specific, Generative AI.

I. INTRODUCTION

In today's world, making an impression with clothes is a priority for many people. It does not matter what their level of knowledge on fashion design is or what interest they have in the field. Poor or misguided fashion choices can range from displaying a lack of personal style and self-expression to appearing disinterested, unaware, or even oblivious to the latest trends. This has subsequently led to a surge in the fashion industry where they have received a high demand for clothes and accessories which saw the opening of outfit recommendation systems in fashion retail. Fashion and its essence of being sophisticated and distinctly presentable have been there for centuries. However, the high speed of fashion brands is the key thing as it is becoming more difficult for an ordinary buyer to wardrobe a trend and create an ensemble look. On the other hand, this completes the role of outfit recommendation systems, which are all about helping those who are not very flair in fashion yet would like to appear more stylish to look great.

This research aims at the development of a recommendation system that will assist users with only a basic understanding of fashion. A user could start a chat with the bot by typing, 'I need an outfit for a casual dinner with friends,' and then the virtual assistant would follow with questions about the preferred colours and styles. The GenAI model, based on the user's responses, would generate a list of outfits along with a description of each outfit that could fit the user's preferences, and the BM25 model, would rank them according to their importance to the user. The chatbot will next present the top trends to the user based on the user prompt.

Patterns of searching that involve traditional fashion often include keyword requests and surfing through the numerous catalogues which require a lot of time and ultimately man may not find the desired outcome. In opposition to the conventional fashion discovery which involves the users submersing in instructions with no real-time engagement, conversational fashion discovery offers a more interactive and natural way of engaging with the system that is similar to talking to a stylist. The machine learning algorithms that we have proposed add an extra dimension to search operations, and this goes far in trying to analyse and interpret the user preferences accurately. This Artificial Intelligence-fueled tool can focus on factors including style preference, as well as the occasion, concluding the social media insights into recent fashion trends. Subsequently, it will offer personalized suggestions that obey one's fashion personality. The system in addition gives a full-encompassing outfit plan complete with clothing items, accessories, footwear and additional matching pieces that blend well with the look. Our platform will combine the power of AI with deep fashion expertise, thus still simplifying the outfit selection and removing the need for the user to make choices that are not stylish, but confident.

Text-video models which generally are concerned with determining temporal attributes and finding how different frames of video content are related to each other, the whole text-to-outfit job is of another nature which is different in the very nature of the outfit data. At their core, ensembles indicate the outfits, which are the combinations of individual items of clothing. The model does not have to recall any formal relationships of the items since they are independent at all times. If the explanatory prose was the only source of information given to recognize casual wear, this model needs to pinpoint the border between styles and classify the attire as formal or casual by itself. In short, keyword or phrase matching is used to find those essential descriptive words in the text that will correspond to the required component.

II. LITERATURE SURVEY

Jang et al., 2023 [1] derived traditional stylists translate customer desires into complete outfits, considering colour, pattern, and material. Existing recommendation systems, though helpful, require some customer input. We propose a novel text-to-outfit retrieval system that generates full outfits based on user descriptions. Our model analyses data at three levels (item, style, outfit) to create a cohesive recommendation. Inspired by image-text pretraining models, we tackle the complexity of outfit styles. Our approach outperforms text-video retrieval models on benchmark datasets, proving its effectiveness for fashion recommendations. Besides introducing an alternative way of recommendation system, the research also reads an individual style through the short description texts.

Cucurull et al., 2019 [2] figured out if clothes go well together involve visual appeal and personal taste, influenced by trends and culture. We propose a new method to predict compatibility based on item features and their surroundings (compatible items we call context). Unlike other methods that just compare item features, ours uses a special neural network (graph neural network) to learn how features interact with context. We tested our model on two fashion datasets (Polyvore, Fashion-Gen) and a smaller Amazon dataset for two tasks: predicting missing items in an outfit and overall outfit compatibility. When we include context, our model performs better than previous methods, and its performance improves as we provide more context.

In his study, Lee et al., 2017 [3], has pointed out that in line with the online fashion industry's unprecedented growth, consumers' need for effective fashion recommendation systems is also feeling extremely high. In the fashion trend forecasting field, it is essential to put forth the idea of items befitting those few that are in existence when discussing style. Everything-changing array over time, the comparative accuracy of assessing those hidden style-definers by the ratings only is irrational. Thus, the efficient representation of fashion item styles, hence, imply the coming of the other new items that will align the style sets formed by the sub-groups of previously shopped-for items. However, the existing vector representation for fashion items is not enough, and hence, we introduce Style2Vec, a vector representation model that can be used for fashion. Similar to the distributional semantics in word embeddings, in its fashion context, Style2Vec learns item representations from similar items which also constitute matching outfits. We use two convolutional neural networks to build the item co-occurrence by raising the chance of the probability of co-occurrence. A fashion semantics test, in which our model captured styles that are represented by parameters such as shapes, colors, patterns and latent styles, demonstrated evaluation through a fashion analogy. Furthermore, Style2Vec features for style classification also gives superior better position than other baseline methods.

In this project, Sarkar et al., 2023 [4] Creating outfit recommendations requires understanding how clothes work together. Our OutfitTransformer model tackles this by learning how each item interacts with the entire outfit. It achieves this with two techniques: task-specific tokens and a self-attention mechanism. For outfit compatibility prediction, the model uses a special token to capture the overall outfit and a classification loss function during training. To recommend complementary items for a partial outfit, it uses a target item token that considers the desired item's category or description. Here, a special loss function ranks entire outfits based on their fit with the target item. Additionally, pre-training and curriculum learning boost retrieval performance. By analysing the whole outfit at once, OutfitTransformer captures complex relationships between items better than methods that compare them in pairs. Our model surpasses previous methods in compatibility prediction, outfit completion, and complementary item retrieval tasks. User studies further confirm the quality of our recommendations.

Su et al., 2020 [5] put forward an approach called VL-BERT (Visual-Linguistic BERT) for the tasks where language and visual contents are combined. VL-BERT utilises the Transformer model as its foundation, but the aim is to have it accept both image and sentence data at the input stage. An input unit can be a single word or a designated semantic region (ROI) within an image. This design facilitates the VL-BERT to produce the results of various visual-linguistic tasks. VL-BERT becomes more impactful because it was pre-trained on such a large dataset with lots of pictures as well as text. It synergises image and text, thereby improving visual tasks such as visual question answering and image-to-text understanding. Significantly, the VL-BERT model produced the best result on the visual commonsense reasoning evaluation.

Vasileva et al., 2018 [6] described online fashion as involving many clothing types (tops, bottoms, shoes) that need to look good together. To recommend outfits, we need a system that considers both similarity (e.g., interchangeable tops) and compatibility (different items that work together). This research proposes a method that learns image representations considering clothing types and learns both similarity and compatibility simultaneously. To test this, we analysed over 68,000 user-created outfits from Polyvore. Our method outperforms previous methods by 3-5 per cent on tasks like predicting outfit compatibility and completing partial outfits, using both our large dataset and a smaller established one. This approach allows for various useful outfit recommendations.

Veit et al., 2015 [7] showed a surge in smartphone use has led to an explosion of photos, including clothing and accessories. To recommend outfits based on a single item (e.g., shoes), we need to understand not just visual similarity, but also how different categories of clothes can work together. This paper proposes a new system to address this challenge. It uses a special neural network (Siamese CNN) to transform images into a compatibility-focused space. Training involves pairs of items labelled as compatible or incompatible, leveraging real-world co-purchase data. To ensure the system learns compatibility across categories (e.g., shoes and tops), training pairs prioritise items from different categories. While applicable to various domains, this research focuses on learning compatible clothing styles. The results show the system can effectively capture style information and generate outfits with items from different categories that complement each other.

Vittayakorn et al., 2015 [8] this paper explores how computers can analyse fashion, from high-fashion runways to everyday wear. The authors use computer vision techniques to study fashion at a large scale. They introduce a new dataset of runway images, design features to capture outfit appearances, and gather human opinions on outfit similarity. Finally, they develop algorithms that learn to mimic human judgments of outfit similarity based on these features. To assess their model's performance, the researchers evaluate how well it predicts outfit similarity, as well as attributes like season, year, and brand. One potential application is tracking how runway trends influence everyday street fashion.

Ding et al., 2023 [9] specified that personalized fashion matching is about calculating the potential of different style combinations regarding the available choices of customers. Conventional techniques typically fail to account for such hidden relations and hence deliver suboptimal results. For the response, this paper draws attention to CP-TransMatch, an innovative solution that takes into account the multi-relational connectivity based on personalized item matching. In contrast to the traditional multi-component translation processes that capture the complex relationships in fashion datasets in second order, CP-Transmatch adopts a single-operation translation algorithm that scrutinizes the third-order interactions of the user-item pairs. Moreover, it includes graph learning models to improve matching accuracy from both contextual and path sides. The results of the comparison of CP-TransMatch with three fashion datasets validate the supremacy of CP-TransMatch and set a new standard for personalized fashion matching, which is the best.

Yang et al., 2019 [10] ”pointed out the growing emphasis in the fashion industry on knowing mix-and-match patterns in garments”. The current approach depends on visual content for evaluating similarity but it does not explain why two items togetherlook nice. To address this gap, the study suggests the Attribute-based Interpretable Compatibility (AIC) method. AIC comprises three modules: a tree-based module for decision rule extraction, an embedding module for attribute semantics and a joint model module for integration of visual and rule embedding to predict matching scores. AIC has its validation on the Lookastic dataset, which contains fashion attributes. AIC gives more accurate results as compared with the state-of-the-art methods. It is also interpretable.

Zhan et al., 2021 [11] emphasizes that individualized recommender systems would play a major role in the Internet fashion industry to ensure that customer convenience and platform profitability are achieved. Unsurprisingly, fashion recommendation systems in place are oblivious to the user's overwhelming choice of outfits as well. This issue is tolerated through the proposal of the Attentive Attribute-Aware Fashion Knowledge Graph (A3-FKG) for personalized outfit preference prediction. A3-FKG establishes associations between outfits based on attributes and incorporates a two-level attention mechanism: a customized relation-aware layer for targeting specific user tastes and a multi-level target-aware layer with the ability to inspire wider user interests. Foremostly, large-scale fashion datasets are put under control that prove the model's efficiency in improving personalized fashion suggestions and ensuring maximum user satisfaction and platform revenue.

Baldrati et al., 2021 [12] showed that taking advantage of the latest achievements in multimodal zero-shot representation learning to test the possibility of features of the CLIP model to be used for content-based image search. Through the by-passing of a reference image and the addition of text-based user preferences, the Combiner network is utilized to correctly join both modalities. The role of the image is defined, text input is engaged, and the image is retrieved as a whole because of this network. The empirical comparison between a baseline of CLIP characteristics serves as a metric to rank the Combiner network performance, which turns out to be more accurate than the other methods of the state-of-the-art in the FashionIQ dataset test. The most essential, particularly, multi-modal approach is a more advanced method for image retrieval that enables the way of new perspectives of those who are interested in deeper perceptions in image searching.

III. METHODOLOGY

This foundational step focuses on creating an interactive and user-friendly interface where individuals can effortlessly express their style preferences and needs. Through a conversational interface, users articulate their requirements, such as occasions, preferred colours, styles, and any other relevant details, in natural language. This approach ensures that users feel comfortable and engaged throughout the recommendation process, leading to more accurate and personalized outfit suggestions. Leveraging state-of-the-art conversational AI techniques, the system employs sophisticated natural language processing (NLP) algorithms for understanding and interpreting the user prompts. By comprehending the nuances of human language, the system engages in meaningful dialogue with users, allowing for seamless communication and interaction. Furthermore, the framework enables the system to cope with different personal selections and conversational styles, besides providing a better user experience.

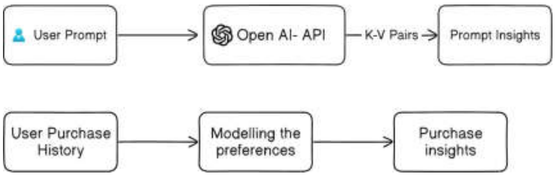


Fig.1: Schematic representation of VastrAI,for generating user-personalized fashion outfits, leveraging GenAI and CLIP-BM25 models.

The efficacy of the recommendation system relies heavily on the data quality and diversity or variety of the data it utilizes. Comprehensive data collection efforts are undertaken, encompassing fashion catalogues, trend reports, social media feeds, and user interactions. Subsequently, collected data undergoes rigorous preprocessing to extract relevant features, clean noise, and ensure uniformity across datasets. This meticulous process lays the groundwork for accurate and insightful outfit recommendations. To facilitate effective model training and recommendation generation, advanced feature engineering techniques are employed to extract meaningful representations of clothing items, styles, colours, and trends from the preprocessed data. This involves encoding textual descriptions, image features, and metadata into vector representations suitable for machine learning algorithms. By capturing the essence of fashion elements, this step enables the system to generate relevant and diverse outfit suggestions.

Machine learning models, including deep neural networks, are trained using curated datasets to learn patterns, correlations, and preferences inherent in fashion data. Model parameters are optimized through techniques such as gradient descent, regularization, and hyperparameter tuning to maximize predictive performance and recommendation accuracy. This iterative process ensures that the models are adept at capturing the complexities of fashion styling and can provide high-quality recommendations to users. Upon receiving user input, the recommendation system leverages the GenAI model, a sophisticated deep-learning architecture specifically designed for outfit generation. This model analyzes user responses, extracts style cues, and generates a diverse range of outfit options customized to the user's preferences. By incorporating advanced algorithms and fashion expertise, the GenAI model ensures that each outfit recommendation is stylish, relevant, and reflective of the user's unique taste.

The BM25 Ranking model is a key aspect of the fashion recommendation system. The model draws

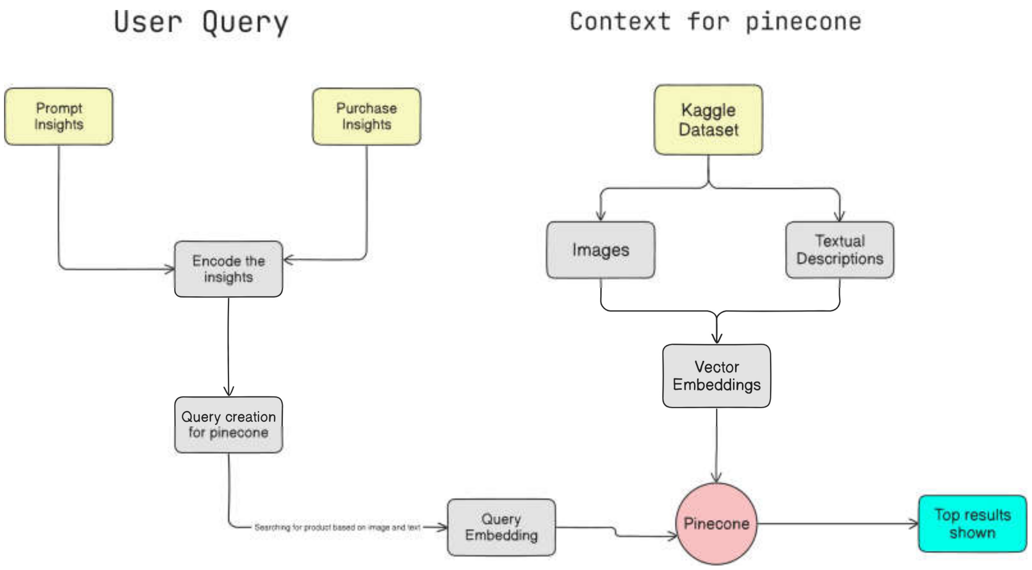


Fig.2: Architecture view of VastrAI

links between different sets of clothing thus enabling the user to get offered new options that are similar to their needs and preferences. To achieve this, the system incorporates the BM25 ranking algorithm, which evaluates the relevance of outfit recommendations based on the user's input. By assigning importance scores to each outfit suggestion, BM25 ensures that the most relevant and preferred options are presented to the user, enhancing the overall user experience.

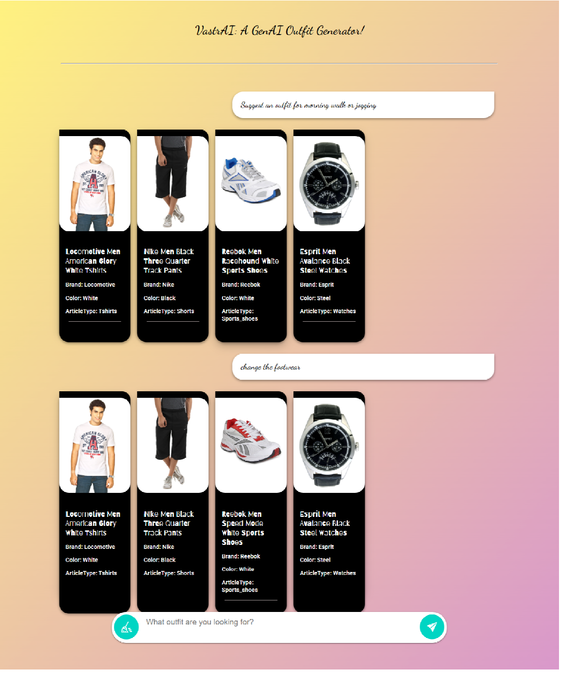
Integration of CLIP-BM25 Models to further enhance semantic understanding and relevance ranking of outfit suggestions. CLIP (Contrastive Language-Image Pretraining) leverages contrastive learning which enables the crossmodal retrieval of visually and semantically similar items by mapping images and text into a shared embedded space. By combining CLIP with the BM25 ranking algorithm, the system ensures that outfit recommendations not only match the user's preferences but also resonate with current fashion trends and styles. By integrating social media insights and trend forecasting algorithms, the system ensures that users receive outfit recommendations that are not only personalized but also aligned with the latest fashion trends, enhancing their overall fashion experience.

In addition to individual clothing items, the recommendation system provides comprehensive outfit plans that include accessories, footwear, and additional matching pieces. This holistic approach to outfit planning ensures that users receive cohesive and well-coordinated ensembles that reflect their unique style preferences. By considering every aspect of outfit composition, the system enables users to effortlessly create stylish and polished looks for any occasion.

Throughout the outfit selection process, the recommendation system serves as an AI-powered decision support tool, guiding users towards stylish and confident fashion choices. By leveraging deep fashion expertise and AI-driven insights, the system simplifies the selection process and empowers users to make informed decisions. Whether it's recommending the perfect dress for a formal event or suggesting the ideal accessories for a casual outing, the system provides valuable guidance and assistance every step of the way.

IV. RESULTS

The user provides outfit preferences through prompts. OpenAI API processes the prompt, generating KeyValue pairs to understand the user's specific requirements. User insights from OpenAI API are combined with purchase history analysis and social media trends. Insights from the latest fashion trends are extracted by scraping Instagram and Pinterest. A refined dataset is created, incorporating the Kaggle dataset and the top trending fashion posts stored in a CSV file. Insights are processed to create a search query for Pinecone, a vector database. Key-value pairs are transformed into search queries and filters, later converted to embeddings for efficient searching. Pinecone is used to search for matching outfits, employing multi-modal search techniques (CLIP for image embeddings, BM25 for text embeddings).



VastrAI: A GenAI Outfit Generator!

Suggest an outfit for morning walk or jogging

Locomotive Men American Glory White Tshirts
Brand: Locomotive
Color: White
ArticleType: Tshirts

Nike Men Black Three Quarter Track Pants
Brand: Nike
Color: Black
ArticleType: Shorts

Reebok Men Racehound White Sports Shoes
Brand: Reebok
Color: White
ArticleType: Sports_shoes

Esprit Men Avalance Black Steel Watches
Brand: Esprit
Color: Steel
ArticleType: Watches

change the footwear

Locomotive Men American Glory White Tshirts
Brand: Locomotive
Color: White
ArticleType: Tshirts

Nike Men Black Three Quarter Track Pants
Brand: Nike
Color: Black
ArticleType: Shorts

Reebok Men Speed Mode White Sports Shoes
Brand: Reebok
Color: White
ArticleType: Sports_shoes

Esprit Men Avalance Black Steel Watches
Brand: Esprit
Color: Steel
ArticleType: Watches

Fig.3: VastrAI Interface

V. CONCLUSIONS

This study introduces a method in fashion recommendation systems using advanced AI techniques that include the GenAI model and CLIP-BM25 model. Using appropriate data preprocessing and iterative model training, the system makes outfit selection easy and personalized for users and with the application of natural language processing and multimodal retrieval techniques, the system greatly helps in meeting different fashion tastes and needs of the users making it simple for them to make fashionable and apt fashion decisions with ease.

In the future, several areas for more research and progress can be found in order to make the proposed fashion recommendation system even more efficient and functional. These include further semantic comprehension through application of intricate technology of natural language processing techniques to be contextual and subject-specific in providing recommendations. Furthermore, human interface enhancement via the development of more user-friendly interfaces, adding voice interactions and virtual try-on abilities will increase the level of engagement and satisfaction of the users. On-going exploration of real-time trend analysis can bring the possibility of timely adjustment of suggestions taking into account current fashion trends, while setting up user responsive feedback mechanisms will facilitate the adaptation of suggestions to individual preferences over time. Additionally, the ethical consideration of data bias and algorithm fairness is also crucial to guaranteeing disparity less outfit recommendations to all users with different demographics. Through this way, the fashion recommendation system can become a more advanced and user-oriented one that plays a big role in personalized fashion styling and outfit choice.

VI. REFERENCES

[1] Jang, J., Hwang, E., & Park, S.-H. (2023). Lost Your Style? Navigating with Semantic-Level Approach for Text-to-Outfit Retrieval. arXiv e-prints, Article arXiv:2311.02122, arXiv:2311.02122. https://doi.org/10.48550/arXiv.2311.02122

[2] Cucurull, G., Taslakian, P., & Vazquez, D. (2019). Context-aware visual compatibility prediction. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 12617–12626.

[3] Lee, H., Seol, J., & Lee, S. (2017). Style2vec: Representation learning for fashion items from style sets. CoRR, abs/1708.04014. http://arxiv.org/abs/1708.04014

[4] Sarkar, R., Bodla, N., Vasileva, M. I., Lin, Y., Beniwal, A., Lu, A., & Medioni, G. (2023). Outfittransformer: Learning outfit representations for fashion recommendation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 3601–3609.

[5] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2020). VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In International Conference on Learning Representations. Retrieved from https://openreview.net/forum?id=SygXPaEYvH

[6] Vasileva, M., Plummer, B., Dusad, K., Rajpal, S., Kumar, R., & Forsyth, D. (2018). Learning type-aware embeddings for fashion compatibility [Publisher Copyright: © 2018, Springer Nature Switzerland AG.; 15th European Conference on Computer Vision, ECCV 2018 ; Conference date: 08-09-2018 Through 14-09-2018]. In Y. Weiss, V. Ferrari, C. Sminchisescu, & M. Hebert (Eds.), Computer vision – eccv 2018 - 15th european conference, 2018, proceedings (pp. 405–421). Springer. https://doi.org/10.1007/978-3-030-01270-0_24

[7] Veit, A., Kovacs, B., Bell, S., McAuley, J., Bala, K., & Belongie, S. (2015). Learning visual clothing style with heterogeneous dyadic co-occurrences.

[8] Vittayakorn, S., Yamaguchi, K., Berg, A., & Berg, T. (2015). Runway to realway: Visual analysis of fashion. Proceedings - 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015, 951–958. https://doi.org/10.1109/WACV.2015.131

[9] Yujuan Ding, PictureP.Y. Mok, PictureYi Bin, PictureXun Yang, PictureZhiyong Cheng. MM '23: Proceedings of the 31st ACM International Conference on MultimediaOctober 2023 Pages 7047–7055 https://doi.org/10.1145/3581783 .3612583

[10] Xun Yang, PictureXiangnan He, PictureXiang Wang, PictureYunshan Ma, PictureFuli Feng, PictureMeng Wang, PictureTat-Seng Chua. SIGIRi 19: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval July 2019 Pages 775–784 https://doi.org/10.1145/3331184.3331242.

[11] Zhan, H., Lin, J., Ak, K., Shi, B., Duan, L.-Y., & Kot, A. (2021). A3-fkg: Attentive attribute-aware fashion knowledge graph for outfit preference prediction. IEEE Transactions on Multimedia, PP, 1–1. https://doi.org/10.1109/TMM.2021.3059514

[12] Baldrati, A., Bertini, M., Uricchio, T., & Bimbo, A. (2021). Conditioned image retrieval for fashion using contrastive learning and clip-based features, 1–5. https : / / doi . org / 10 . 1145 / 3469877.3493593