# Comparative Analysis of Fashion Captioning for Multimodal Fashion Recommendation

Gwendolyn Rippberger, Julia Neidhardt

*CD Lab for Recommender Systems, TU Wien, Austria*

**Abstract**

Multimodal information provides new opportunities for recommender systems, especially in the fashion domain, where both visual and textual information can be utilized to provide a comprehensive understanding of the product. In this work, we focused on the task of fashion captioning, a specialized form of image captioning for fashion items. We fine-tuned pretrained vision-language models on two distinct fashion datasets to evaluate how effectively they capture dataset-specific ground truths. We were able to fine-tune the models successfully to a competitive result with specifically trained models. The resulting captioning models are applied in two key scenarios: (1) as components for generating richer multimodal embeddings in recommender systems, and (2) for modality imputation, where automatically generated descriptions are used to fill in missing textual data. We show that different modalities work better depending on the size of the dataset and the list length but none outperform the traditional item-based collaborative filtering technique using a real-life dataset with over 1M users and 31M transactions. Additionally, we present a detailed analysis of the two fashion datasets, highlighting critical aspects such as item presentation and textual style, which are often overlooked yet essential for effective modeling.

**Keywords**

Multimodal Recommendation, Fashion Captioning, NLP, Generative AI

## 1. Introduction and Motivation

The fashion domain poses challenges due to sparse purchase data for traditional recommender systems (RS) such as Collaborative Filtering (CF) [1, 2]. There are various approaches trying to leverage different modalities to overcome this problem e.g. using product images [1, 3–8], textual descriptions or customer reviews [9–11] and even video [12].

Fashion is inherently multimodal [13], combining visual information like product images with textual descriptions. Previous work with multimodal recommender systems [14–16] shows that leveraging high-level features from multimodal items outperforms CF recommendation [1, 17–21]. However, real-world data struggles with missing data modalities [22, 23]. Common solutions are dropping or imputation [24] with the risk of discarding valuable items or introducing noise. Previous work [23] shows that imputation using traditional methods, e.g., random, zeros, and global mean, can preserve the performance gap between multimodal and pure collaborative recommender systems. Still, those methods are rigid and might not capture the diversity of items. Another approach [22] uses feature propagation in the context of graph networks, though this is limited to graph networks.

We propose an alternative for the fashion domain: fine-tuning pretrained vision-language models for fashion captioning. Fashion captioning describes the domain-specific task of image captioning (generating text based on images) for fashion items that focuses on long captions with fine-grained attributes with an enchanting expression style [25] or tailor details [26]. Using the fine-tuned models to generate item descriptions we can augment missing text descriptions and the fine-tuned models can be used as feature extractors in order to experiment with multimodal embeddings. This method can be used as a preprocessing step for different recommendation algorithms and resulting text descriptions can be inspected as they are human-readable (compared to feature vectors). With the different modalities fashion has to offer, the question arises: which feature representations are most effective for

*DaQuaMRec, 1st International Workshop on Data Quality-Aware Multimodal Recommendation, RecSys 2025*
Corresponding author.
gwendolyn.rippberger@tuwien.ac.at (G. Rippberger)

recommending items? To address this, we analyze unimodal and multimodal feature spaces derived from pretrained models, as well as features extracted from our fine-tuned fashion captioning models.

We benchmark these feature sets using multiple recommendation algorithms, including content-based methods (e.g., k-NN), hybrid models (e.g., VBPR [1]), and unpersonalized algorithms (e.g. most popular), to understand how the choice of feature space impacts recommendation quality. By systematically comparing these approaches, we aim to identify the representation that best captures item semantics and user preferences, ultimately bridging the gap between sparse user interaction data and rich item content. To summarize, the research questions and related contributions of this work are:

**RQ1: To what extent can fine-tuning improve the performance of off-the-shelf image captioning models on domain-specific fashion datasets?** Firstly, we want to analyze if fine-tuning achieves good enough results in order to use the models for augmentation. We experiment with four different models (Section 2.1) using a dataset specifically curated for creating item descriptions of fashion items (FACAD) and a real-life dataset based on items from the fashion store H&M (more details in Section 3). This is done in order to better understand capabilities and limitations of fine-tuning. The evaluation is done quantitatively and qualitatively to get a full picture (Section 2.2). Based on seven different metrics, we show that fine-tuning achieves good results, especially when it comes to identifying item specific attributes. However, our qualitative analysis shows that the models struggle with abstract concepts and hidden details.

**RQ2: Which feature embeddings (textual, visual, or multimodal) provide the best recommendations?** We compare two setups: (1) filtered items (dropped missing modalities and sparsely represented items, (2) unfiltered items and missing text descriptions augmented. This is done with a real-life data set [26] and we use VBPR [1], a model that extends matrix factorization with feature vectors (originally visual), to compare the different feature spaces. Our results show that comparing the different feature spaces the textual features perform best except in terms of precision for the augmented dataset. Still, ItemKNN outperforms all setups.

Following our research questions, we contribute the following: (1) We conduct an in-depth analysis of the effectiveness of fine-tuning pretrained models. (2) We explore the use of previously unutilized query embeddings derived from image captioning models for fashion recommendation. (3) We present results on a real-world, underexplored fashion recommendation dataset to evaluate which feature spaces yield the best recommendation performance. Additionally, we make all models publicly available via our Hugging Face Space at https://huggingface.co/CDL-RecSys, and we release the code at https://github.com/omgwenxx/multimodal-fashion-analysis/.

Finally, this work is structured as follows: we first present the experiment setup, including the models (Section 2.1) used for fine-tuning, feature extraction (Section 2.3.1), and recommendation algorithms (Section 2.3.2). We then compare the datasets in detail (Section 3), present the results (Section 4), and conclude with discussion and future directions (Section 5).

## 2. Experiment Setup

### 2.1. Models for Image Captioning

For the image captioning task, we focused on open-source models to ensure transparency and reproducibility. We selected models available via Hugging Face [27]. We evaluated BLIP-2 [28] variants based on OPT [29] and LLaVA-1.5 [30], a multimodal conversational model, with different number of parameters. While BLIP-2 generates text from images alone, LLaVA relies on prompt-based inputs and is optimized for instruction-following. As a baseline on the FACAD dataset, we used results from Yang et al. [25], replicating their setup.

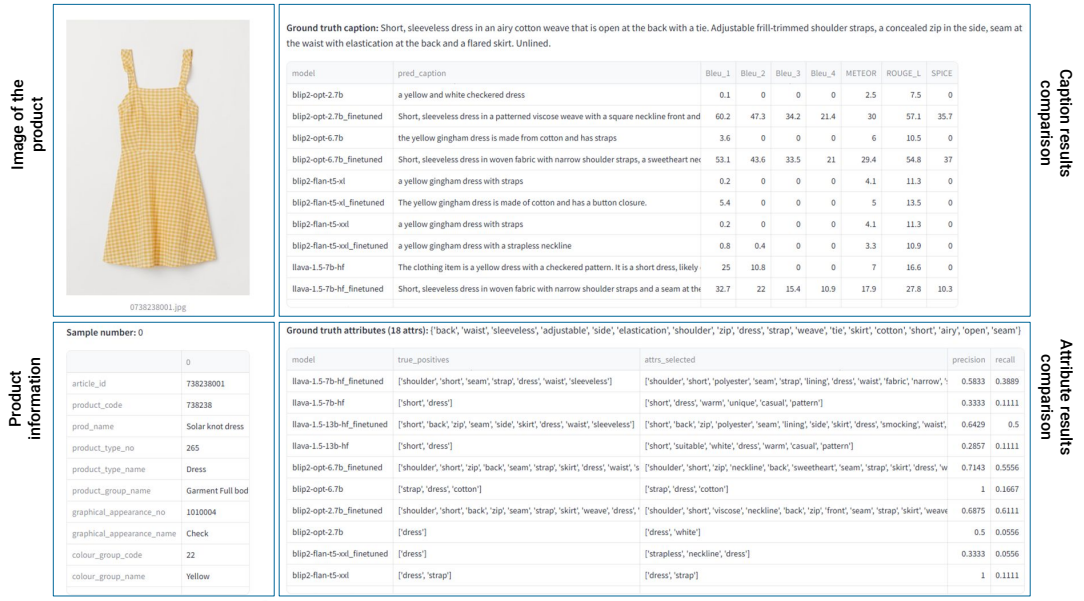


**Figure 1:** Example (using H&M item) of the information displayed for qualitative analysis. It includes captions, scores (except for CIDEr due to the zero vector caused by a single reference sentence), attributes, and the respective recall and precision. We created this interface using the Streamlit [38] library for Python.

## 2.2. Evaluation

We evaluated our results using well-established image captioning measures, namely BLEU [31], ROUGE-L [32], CIDEr [33], METEOR [34] and Spice [35]. As image captioning focuses on the task of generating a "correct" caption, the majority of those metrics measure word overlap either based on precision (BLEU), recall (ROUGE-L), or a combination of both (METEOR, focusing on word order). CIDEr incorporates term frequency-inverse document frequency (TF-IDF) weighting to emphasize informative words, while SPICE evaluates scene-graph-based semantic content.

Additionally, we re-implemented the measures introduced by Yang et al. [25] for category accuracy and "mean average precision". What the authors reported as mean average precision is the average precision over all captions. We keep the naming for comparison. Also, the authors originally pretrained a 3-layer text CNN [36] for category classification which was not provided. As a result, we used a pretrained BERT model [37] for text classification and fine-tuned it on each dataset, achieving a test accuracy of 90.3% on the FACAD dataset (78 classes) and 94.7% on the H&M dataset (89 classes).

For the qualitative comparison of the captions, we implemented a web application (see Figure 1), showing each product and its ground truth caption and attributes. We then proceeded to manually check the first 20 samples in the H&M test set and the first 3 distinct items in the FACAD test set (because the dataset includes the same caption for multiple images containing the same item).

## 2.3. Algorithms for Recommendations and Feature Extraction

### 2.3.1. Feature Extraction

We used the Ducho-meets-Elliot framework presented by Attimonelli et al. [14]. The framework integrates Ducho [14] for feature extraction. For our experiments, we explored six different setups: (1) visual features extracted from ResNet50 [39], (2) textual features from SentenceBERT [40], and (3) multimodal features from CLIP [41].

Additionally, we included (4) multimodal features obtained by concatenating ResNet50 and Sentence-BERT embeddings, (5) features extracted from the Q-Former component of the fine-tuned BLIP-2 model (32x768 values), and (6) visual features extracted from the same fine-tuned BLIP-2 model (257x1408 values). We used average global pooling to reduce the size to 1x1408.

**Table 1**
Statistics for preprocessed versions of the H&M dataset. We compared the original, filtered, and augmented datasets, including the number of users, items, transactions, and the resulting train/test split sizes. Items that were never bought are not included in the split. All items describe all items that appear in the dataset. Transaction items describe items that (based on the pool of items) appear in the transactions.

| Dataset | Users | Items (All/Transactions) | Transactions | Splits (Train/Test) |
|---|---|---|---|---|
| Original | 1,372,980 | 105,542 / 104,547 | 31,788,324 | - |
| Filtered | 1,360,178 | 104,232 / 103,251 | 31,400,864 | 24,516,873 / 6,883,991 |
| Unfiltered + Augmented | 1,361,823 | 105,100 / 104,106 | 31,651,678 | 24,717,300 / 6,934,378 |

#### 2.3.2. Recommendation Algorithms

We used algorithms supported by the elliot pipeline [42]. We chose to use Visual Bayesian Personalized Ranking [1] as it supports using different features for recommendation and implicit feedback. As a baseline, we used unpersonalized recommendation algorithms that suggest items without considering individual user preferences, namely Most Popular (MostPop) and Random. For additional comparison, we ran neighborhood-based approaches based on item similarity (Amazon's item-to-item collaborative filtering, ItemKNN) [43] and user similarity (algorithm used by GroupLens, UserKNN) [44].

**Data Split.** We applied an 80/20 temporal split to divide the transactions into training and test sets. For the first setup, we used the filtered articles (see Section 3.2) to inspect an isolated setup for the comparison of the features. For the second setup, we compared a "real-world scenario" where we augment the missing item descriptions using the generated captions from BLIP-2 (404 items, all fashion-related) and keep all items with images (105,100). We give a summary of the numbers in Table 1. 9699 customers do not appear in the transactions.

## 3. Datasets

We evaluated the fine-tuned models on FACAD (FAshion CAptioning Dataset) and the H&M dataset [26]. FACAD enables comparison with Yang et al. [25] and is, to our knowledge, the only available dataset focused on fashion captioning. H&M includes user-item transactions, allowing us to test generated captions and fine-tuned embeddings for recommendations.

### 3.1. FACAD

The dataset contains 993K images and 130K captions that were initially split into 794K (~80%) image-description pairs for training, 99K (~10%) for validation, and the remaining 100K (~10%) for testing. However, the dataset that is currently provided by the authors has a different distribution, with 888,293 pairs designated for training, 19,915 for validation, and 101,225 for testing (a total of 1,009,463 samples). As per the authors, this was done so that validation does not take as long[1].

The authors extracted 990 attributes from item metadata using the Stanford Parser [45]. It should be noted, that working with the dataset, we noticed that the test attribute file was incorrect (did not match the test items), we therefor manually created the correct attributes by checking caption overlap with the ground truth data and then using the attributes provided in the metadata file.

### 3.2. H&M

**Preprocessing.** The primary focus was on preparing the article dataset and associated images for creating an image captioning dataset. The first step was to clean the dataset by removing any articles that did not have a detailed description or a corresponding image file. We then decided to filter based

[1] https://github.com/xuewyang/Fashion_Captioning/issues/5, last accessed 13.03.2025, 15:40

**Table 2**
Comparison of both datasets. CAT: category, AT: attribute, CAP: caption, *The H&M dataset includes original images of different sizes, but a majority are around 1166x1750 pixels large. **The original images for FACAD are not provided by the authors; they are only preprocessed versions downsized to 256x256. They can however be extracted from the metadata file provided.

| Dataset | #img | img size | #img (per cap.) | vocab size | #CAP avg len | #CAT | #AT |
|---|---|---|---|---|---|---|---|
| H&M | 104K | *~1166x1750 | 1 | 7.6K | 23.9 | 89 | 1014 |
| FACAD | 993K | **256x256 | 7 ~ 8 | 15.8K | 21.0 | 78 | 990 |

on the number of articles per product type and kept only those categories with at least 7 items (because 25% of the product categories have less than 7 items, keeping 75% of the initial items). Non-fashion-related categories, e.g. *Dog Wear* and *Sleeping Sack*, were removed manually to retain only fashion and accessory items. The resulting dataset contains 89 categories after filtering. The filtered dataset was split into training, validation, and test set, maintaining a distribution of 80% training, 10% validation, and 10% test data. The final count of articles after preprocessing was 104,232, distributed as follows into 83,385 articles (train set), 10,423 (validation set), and 10,424 articles (test set).

**Attributes.** To compute the average precision, attributes within the product descriptions need to be extracted. The `detail_desc` column was used to extract nouns, adjectives, and proper nouns based on Universal POS tagging definitions [46, 47]. The extraction process used the Stanza NLP library [48] to identify and filter these attributes. Before tagging, the descriptions are lowercase, and hyphen-connected words, e.g., t-shirt, are split. Then, we extract the lemmatized attributes and filter, keeping only attributes appearing at least 10 times. This is done to ensure the significance and relevance of attributes kept. Using the train set items, we then collect all extracted attributes (1014 in total). These are then used as a pool to select attributes from generated captions and to generate the ground truth for the test set.

### 3.3. Comparison of Datasets

We use this section to emphasize the differences between both datasets to provide a better understanding of the results presented in Section 4.1. Both datasets are domain-specific fashion datasets but differ in many aspects (see Table 2). One of them is the size, with FACAD being almost 10 times larger than the H&M dataset. Providing more variety of item perspectives, including different angle shots, with and without a model and a material shot (see Figure 2a). The H&M dataset only offers single-item images without a model, sometimes only showing part of the item. Furthermore, the H&M dataset shows a 1-to-1 relationship between captions and images, whereas the FACAD dataset includes multiple items with the same description but differing in color or single items with many images.

Comparing the captions, the H&M dataset has more concise captions describing the item's "tailor" details, e.g., "frill-trimmed shoulder straps", whereas FACAD captions are more "enchanting" (examples can be seen in Figure 2). It includes expressions made for selling, e.g., "this neutral hued cotton sweater you'll wear everywhere". This can be also noticed in the size of the vocabulary (see Table 2).

## 4. Results

### 4.1. Fashion Captioning

**Quantitative Analysis.** We present the results of the pretrained models and the fine-tuned models side-by-side for the H&M dataset in Table 3 and for the FACAD dataset in Table 4.

*H&M.* All models' performance improved with fine-tuning. Among them, BLIP-2-6.7B achieves the best performance. We observed, and investigated during our qualitative analysis, the trade-off between the precision and recall of the attributes for the different models. Due to the longer captions produced by the LLaVA models, they achieve higher recall scores but then lack precision. The opposite for the

(a) this **shawl collar jersey** blazer **bloom** with **black** and **white floral** patterning inspired by the **work** of rising spanish photographer coco capit n

(b) **Short, sleeveless dress** in an **airy cotton weave** that is **open** at the **back** with a **tie**. **Adjustable** frill-trimmed **shoulder straps**, a concealed **zip** in the **side**, **seam** at the **waist** with **elastication** at the **back** and a flared **skirt**. Unlined.

**Figure 2:** Examples taken from the respective datasets (FACAD left, H&M right) with their image(s) and their ground truth captions with attributes highlighted bold.

**Table 3**
Fashion captioning results for dataset H&M in pretrained (PT) and fine-tuned (FT) scenarios. The best results highlighted in bold. *CIDEr shows results beyond 100 due to multiple scaling [33].

| H&M | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **BLEU-4** | | **METEOR** | | **ROUGE-L** | | **CIDEr** | | **SPICE** | | **MAP/MAR** | | **Acc** | | **% Impr.** |
| | **PT** | **FT** | **PT** | **FT** | **PT** | **FT** | **PT** | **FT** | **PT** | **FT** | **PT** | **FT** | **PT** | **FT** | |
| BLIP-2-2.7B | 0.3 | 40.8 | 5.3 | 33.5 | 14.8 | 63.4 | 7.0 | 275.2* | 8.0 | 44.7 | 37.2 / 11.4 | 70.8 / 65.4 | 53.0 | 83.5 | 2685.50 |
| BLIP-2-6.7B | 0.3 | **41.4** | 5.5 | **33.9** | 15.3 | **63.8** | 7.3 | **281.3*** | 8.3 | **45.4** | 38.3 / 11.6 | **70.8** / 66.0 | 53.8 | **83.8** | 2696.34 |
| LLaVA-1.5-7B | 0.5 | 23.4 | 7.4 | 32.8 | 14.5 | 46.0 | 0.8 | 24.2 | 5.3 | 34.6 | 15.2 / 13.0 | 53.1 / **71.3** | 32.0 | 82.9 | 1289.53 |
| LLaVA-1.5-13B | 0.4 | 22.9 | 7.4 | 32.3 | 14.2 | 45.5 | 1.2 | 22.0 | 5.8 | 34.9 | 17.8 / 13.8 | 54.0 / 70.8 | 31.0 | 83.5 | 1255.67 |

BLIP-2 models. They achieve better overall performance by having a balance between precision and recall due to their ability to adapt to the ground truth length.

*FACAD.* Comparing the accuracy reported by Yang et al. [25] for the category classification, we noticed that our models (fine-tuned BERT classifier) worked better than the reported CNN classifier. The accuracy metric generally describes how well the captioning models recognize the correct category from the image because the accuracy models will predict the category based on the caption input (with an accuracy of >90%, based on the performance on the testset). The zero-shot setup (using BLIP-2 results) achieves nearly 40% better accuracy than the results reported by Yang et al. We hypothesize that this improvement may be due to the BERT model outperforming the originally used CNN-based model or because the model by Yang et al. (Semantic Rewards guided Fashion Captioning, SRFC) struggles to accurately classify clothing items. The results show that the fine-tuned models do not achieve the same performance as SRFC in terms of image captioning metrics. However, results are competitive and models outperform when it comes to attributes precision and recall as well as recognizing the product category. We see the same behavior for BLIP-2 and LLaVA in terms of precision and recall.

**Qualitative Analysis.** *H&M.* Based on the setup explained in Section 2.2, we find that the fine-tuned LLaVA models retain the core content of the original H&M captions but tend to hallucinate details, such as materials or funding sources, not present in the training data, likely influenced by frequent patterns e.g. polyester. Due to their fixed setting, they often overgenerate beyond the caption's natural

**Table 4**
Fashion captioning results for the FACAD dataset in pretrained (PT) and fine-tuned (FT) scenarios. The last two models show the worst (CNN-C) and best (SRFC, by the authors) models reported by Yang et al. [25], with the best values for each metric highlighted in bold.

| FACAD | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **BLEU-4** | | **METEOR** | | **ROUGE-L** | | **CIDEr** | | **SPICE** | | **MAP/MAR** | | **Acc** | | **% Impr.** |
| | **PT** | **FT** | **PT** | **FT** | **PT** | **FT** | **PT** | **FT** | **PT** | **FT** | **PT** | **FT** | **PT** | **FT** | |
| BLIP-2-2.7B | 0.3 | 3.7 | 4.6 | 10.1 | 12.6 | 19.7 | 4.3 | 36.4 | 6.4 | 10.3 | 17.4 / 7.9 | **24.6** / 22.5 | 52.0 | **69.9** | 313.21 |
| BLIP-2-6.7B | 0.3 | 3.5 | 4.6 | 9.8 | 12.9 | 19.1 | 4.1 | 34.4 | 6.4 | 9.8 | 17.7 / 7.8 | 23.3 / 21.2 | 51.4 | 69.0 | 297.97 |
| LLaVA-1.5-7B | 0.2 | 1.5 | 6.8 | 11.2 | 12.0 | 15.4 | 0.1 | 1.2 | 4.1 | 7.4 | 9.8 / 8.7 | 14.7 / **27.9** | 45.1 | 65.5 | 288.39 |
| LLaVA-1.5-13B | 0.2 | 0.7 | 6.5 | 8.3 | 12.4 | 12.6 | 0.8 | 0.1 | 4.5 | 3.7 | 9.9 / 8.5 | 8.4 / 16.3 | 44.2 | 27.4 | 17.27 |
| CNN-C [49] | 2.1 | | 7.2 | | 16.3 | | 20.8 | | 6.5 | | 4.9 / - | | 10.8 | | - |
| SRFC [25] | **6.8** | | **13.2** | | **24.2** | | **42.1** | | **13.4** | | 9.5 / - | | 18.2 | | – |

end, unlike BLIP-2 models, which better learn caption lengths post fine-tuning. This overgeneration leads to higher attribute recall but lower precision, with the reverse observed for BLIP-2, especially in shorter captions. All models struggle with nuances in material (satin vs. velvet), product size (e.g. 33 cm), and disambiguation of visually similar items (crop top vs. sports bra).

*FACAD.* For the FACAD dataset sample in Figure 2a, we observed that certain caption parts, such as references to designers or inspirations, lack visual grounding, adding noise for models learning image-text alignment. The fine-tuned LLaVA models often adopt a "sales-like" tone which works well for marketing language but tends to overgenerate for straightforward item descriptions (FACAD vs. H&M). LLaVA models are prone to overgeneration, even reproducing prompt templates (e.g., starting responses with `ASSISTANT:`). Category predictions are sometimes off—e.g., mislabeling a blazer as a jacket or a tee/top, and misclassifications increase for partial views or when items are not fully visible (e.g., back details or side views). Some orignal attributes (e.g., "chrissy", "teigen") do not correspond to visual features, complicating evaluation. Overall, both datasets reveal that the models struggle with fine-grained visual distinctions in material, color (dark blue vs. black), or category (jacket vs. blazer), likely due to image resizing during preprocessing and the inherent visual limitations in the dataset.

### 4.2. Recommendations

With the results presented in Table 5 we are able to answer which feature embeddings provide the best recommendations (RQ2). Based on the H&M dataset the answer is textual embeddings, except for MAP in the augmented dataset. However, comparing the results, we see that all different feature spaces perform similarly. and not as good as the best overall result using the ItemKNN algorithm. This might also indicate that VBPR might not be the right algorithm for this dataset in general.

## 5. Discussion and Conclusion

To evaluate the impact of fine-tuning, we analyze how well it improved our pretrained models. We report the average percentage improvement in performance across six metrics, computed by first calculating the percentage change for each metric individually and then averaging these values per model, presented in Table 3 and Table 4. We observe that fine-tuning was more effective for the H&M dataset, which we attribute to the presence of a one-to-one relationship between images and captions as well as the smaller size. To further investigate this hypothesis, one could subsample the FACAD dataset to include only one-to-one samples or a more refined selection of images [50]. Compared to SRFC by Yang et al., our fine-tuned models, despite showing slightly lower image captioning metrics, generate more accurate captions in terms of attributes and categories. This demonstrates the strength of our approach, which is a simpler pipeline that still maintains competitive performance. Our qualitative

**Table 5**
Recommendation results for H&M dataset on the test set explained in Section 2.3.2. The best results highlighted in bold. We highlight the best results overall (using ItemKNN) and for the different feature spaces.

| Setting | Original | | | | Unfiltered & Augmented | | | |
|---|---|---|---|---|---|---|---|---|
| | NDCG | | MAP | | NDCG | | MAP | |
| | @5 | @12 | @5 | @12 | @5 | @12 | @5 | @12 |
| Random | 0.00005 | 0.00007 | 0.00005 | 0.00004 | 0.00006 | 0.00008 | 0.00005 | 0.00005 |
| MostPop | 0.00538 | 0.00640 | 0.00439 | 0.00373 | 0.00534 | 0.00634 | 0.00437 | 0.00371 |
| UserKNN (k=10) | 0.01198 | 0.01355 | 0.00558 | 0.00379 | 0.01214 | 0.01370 | 0.00564 | 0.00384 |
| ItemKNN (k=20) | **0.05760** | **0.06421** | **0.03557** | **0.02534** | **0.05762** | **0.06423** | **0.03560** | **0.02537** |
| Visual (ResNet50) | 0.02414 | 0.02882 | 0.01213 | 0.00917 | 0.02414 | 0.02881 | 0.01220 | 0.00921 |
| Visual (BLIP-2) | 0.02167 | 0.02631 | 0.01092 | 0.00840 | 0.02141 | 0.02606 | 0.01080 | 0.00834 |
| Textual (SentenceBERT) | **0.02509** | **0.02986** | **0.01270** | **0.00955** | **0.02471** | **0.02950** | 0.01246 | 0.00924 |
| Multimodal (ResBERT) | 0.02427 | 0.02892 | 0.01220 | 0.00922 | 0.02425 | 0.02887 | 0.01226 | 0.00923 |
| Multimodal (CLIP) | 0.02455 | 0.02917 | 0.01241 | 0.00934 | 0.02466 | 0.02925 | **0.01254** | **0.00943** |
| Queries (BLIP-2) | 0.02451 | 0.02912 | 0.01247 | 0.00939 | 0.02465 | 0.02914 | 0.01240 | 0.00933 |

analysis (Section 4.1) further reveals model limitations, especially when captions reference non-visible or abstract information and the trade-off between precision and recall for the LLaVA models. For future work, we recommend using captions that focus on visible item features and avoiding repeated captions across multiple images and experiments with image quality [51] to improve performance. Interestingly, our recommendation results (Section 4.2) show that (1) with increasing list length our NDCG increases but MAP decreases, meaning that the algorithms retrieve relevant items but not rank them optimally, (2) textual features (with our dataset) returned the best results (except for MAP on the augmented dataset) and (3) ItemKNN outperforms all multimodal approaches. ItemKNN likely performs well because it leverages repetitive user behavior and trends which is common in the fashion domain. The gap between VBPR and ItemKNN leads us to believe that our embeddings are either not representative enough to learn user preferences or VBPR may not be well-suited for this dataset. Future work could explore alternative fusion strategies and evaluate more multimodal recommendation algorithms, e.g. FREEDOM [52] or BM3 [53]. However, we show that despite fashion being a visually dominant domain, textual descriptions had the best results, highlighting their importance for future research.

## 6. Acknowledgments

The financial support by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

## References

[1] R. He, J. McAuley, VBPR: visual Bayesian Personalized Ranking from implicit feedback, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, AAAI Press, Phoenix, Arizona, 2016, pp. 144–150.

[2] T. M. A. U. Gunathilaka, P. D. Manage, J. Zhang, Y. Li, W. Kelly, Addressing sparse data challenges in recommendation systems: A systematic review of rating estimation using sparse rating data and profile enrichment techniques, Intelligent Systems with Applications 25 (2025) 200474. URL:

https://www.sciencedirect.com/science/article/pii/S2667305324001480. doi:https://doi.org/10.1016/j.iswa.2024.200474.

[3] V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, N. Sundaresan, Large Scale Visual Recommendations From Street Fashion Images, 2014. URL: http://arxiv.org/abs/1401.1778. doi:10.48550/arXiv.1401.1778, arXiv:1401.1778 [cs].

[4] Y. Deldjoo, T. Di Noia, D. Malitesta, F. A. Merra, Leveraging content-style item representation for visual recommendation, in: Advances in information retrieval: 44th european conference on IR research, ECIR 2022, stavanger, norway, april 10–14, 2022, proceedings, part II, Springer-Verlag, Berlin, Heidelberg, 2022, pp. 84–92. URL: https://doi.org/10.1007/978-3-030-99739-7_10. doi:10.1007/978-3-030-99739-7_10, number of pages: 9 Place: Stavanger, Norway.

[5] W. Chen, P. Huang, J. Xu, X. Guo, C. Guo, F. Sun, C. Li, A. Pfadler, H. Zhao, B. Zhao, POG: Personalized Outfit Generation for Fashion Recommendation at Alibaba iFashion, 2019. URL: https://arxiv.org/abs/1905.01866v3.

[6] W.-C. Kang, C. Fang, Z. Wang, J. McAuley, Visually-Aware Fashion Recommendation and Design with Generative Image Models, 2017. URL: https://arxiv.org/abs/1711.02231v1.

[7] R. He, J. McAuley, Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering, 2016. URL: http://arxiv.org/abs/1602.01585. doi:10.1145/2872427.2883037, arXiv:1602.01585 [cs].

[8] Q. Liu, S. Wu, L. Wang, DeepStyle: Learning User Preferences for Visual Recommendation, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 841–844. URL: https://doi.org/10.1145/3077136.3080658. doi:10.1145/3077136.3080658.

[9] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, 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, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 765–774. URL: https://dl.acm.org/doi/10.1145/3331184.3331254. doi:10.1145/3331184.3331254.

[10] K. Zhao, X. Hu, J. Bu, C. Wang, Deep Style Match for Complementary Recommendation, 2017. URL: https://arxiv.org/abs/1708.07938v1.

[11] C. Bracher, S. Heinz, R. Vollgraf, Fashion DNA: Merging Content and Sales Data for Recommendation and Article Mapping, 2016. URL: http://arxiv.org/abs/1609.02489. doi:10.48550/arXiv.1609.02489, arXiv:1609.02489 [cs].

[12] M. Yang, K. Yu, Real-time clothing recognition in surveillance videos, in: 2011 18th IEEE international conference on image processing, 2011, pp. 2937–2940. doi:10.1109/ICIP.2011.6116276.

[13] Y. Deldjoo, M. Schedl, B. Hidasi, Y. Wei, X. He, Multimedia recommender systems: Algorithms and challenges, 2022, pp. 973–. doi:10.1007/978-1-0716-2197-4_25.

[14] M. Attimonelli, D. Danese, A. D. Fazio, D. Malitesta, C. Pomo, T. D. Noia, Ducho meets Elliot: Large-scale Benchmarks for Multimodal Recommendation, 2024. URL: http://arxiv.org/abs/2409.15857. doi:10.48550/arXiv.2409.15857, arXiv:2409.15857 [cs].

[15] K. Laenen, M.-F. Moens, Attention-based Fusion for Outfit Recommendation, 2019. URL: http://arxiv.org/abs/1908.10585. doi:10.48550/arXiv.1908.10585, arXiv:1908.10585 [cs].

[16] X. Song, C. Wang, C. Sun, S. Feng, M. Zhou, L. Nie, MM-frec: Multi-modal enhanced fashion item recommendation, IEEE Transactions on Knowledge and Data Engineering 35 (2023) 10072–10084. doi:10.1109/TKDE.2023.3266423.

[17] W. Yinwei, W. Xiang, N. Liqiang, H. Xiangnan, C. Tat-Seng, GRCN: Graph-refined convolutional network for multimedia recommendation with implicit feedback, 2021. URL: https://arxiv.org/abs/2111.02036, arXiv: 2111.02036 [cs.IR].

[18] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, T.-S. Chua, MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video, in: Proceedings of the 27th ACM international conference on multimedia, Mm '19, Association for Computing Machinery, New

York, NY, USA, 2019, pp. 1437–1445. URL: https://doi.org/10.1145/3343031.3351034. doi:10.1145/3343031.3351034, number of pages: 9 Place: Nice, France.

[19] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, L. Wang, Mining latent structures for multimedia recommendation, in: Proceedings of the 29th ACM international conference on multimedia, Mm '21, ACM, 2021, pp. 3872–3880. URL: http://dx.doi.org/10.1145/3474085.3475259. doi:10.1145/3474085.3475259.

[20] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, F. Jiang, Bootstrap latent representations for multi-modal recommendation, in: Proceedings of the ACM web conference 2023, Www '23, ACM, 2023, pp. 845–854. URL: http://dx.doi.org/10.1145/3543507.3583251. doi:10.1145/3543507.3583251.

[21] X. Zhou, Z. Shen, A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation, in: Proceedings of the 31st ACM international conference on multimedia, Mm '23, ACM, 2023, pp. 935–943. URL: http://dx.doi.org/10.1145/3581783.3611943. doi:10.1145/3581783.3611943.

[22] D. Malitesta, E. Rossi, C. Pomo, F. D. Malliaros, T. D. Noia, Dealing with Missing Modalities in Multimodal Recommendation: a Feature Propagation-based Approach, 2024. URL: http://arxiv.org/abs/2403.19841. doi:10.48550/arXiv.2403.19841, arXiv:2403.19841 [cs].

[23] D. Malitesta, E. Rossi, C. Pomo, T. Di Noia, F. D. Malliaros, Do We Really Need to Drop Items with Missing Modalities in Multimodal Recommendation?, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 3943–3948. URL: https://dl.acm.org/doi/10.1145/3627673.3679898. doi:10.1145/3627673.3679898.

[24] C. Wang, M. Niepert, H. Li, LRMM: Learning to Recommend with Missing Modalities, 2018. URL: http://arxiv.org/abs/1808.06791. doi:10.48550/arXiv.1808.06791, arXiv:1808.06791 [cs].

[25] X. Yang, H. Zhang, D. Jin, Y. Liu, C.-H. Wu, J. Tan, D. Xie, J. Wang, X. Wang, Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards, 2022. URL: http://arxiv.org/abs/2008.02693. doi:10.48550/arXiv.2008.02693, arXiv:2008.02693 [cs].

[26] C. G. Ling, H&M Personalized Fashion Recommendations, 2022. URL: https://kaggle.com/competitions/h-and-m-personalized-fashion-recommendations.

[27] Hugging Face – The AI community building the future., 2025. URL: https://huggingface.co/.

[28] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, 2023. URL: http://arxiv.org/abs/2301.12597. doi:10.48550/arXiv.2301.12597, arXiv:2301.12597 [cs].

[29] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, OPT: Open Pre-trained Transformer Language Models, 2022. URL: http://arxiv.org/abs/2205.01068. doi:10.48550/arXiv.2205.01068, arXiv:2205.01068 [cs].

[30] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual Instruction Tuning, 2023. URL: http://arxiv.org/abs/2304.08485. doi:10.48550/arXiv.2304.08485, arXiv:2304.08485 [cs].

[31] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, pp. 311–318. URL: https://dl.acm.org/doi/10.3115/1073083.1073135. doi:10.3115/1073083.1073135.

[32] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of summaries, 2004, p. 10.

[33] R. Vedantam, C. L. Zitnick, D. Parikh, CIDEr: Consensus-based Image Description Evaluation, 2015. URL: http://arxiv.org/abs/1411.5726. doi:10.48550/arXiv.1411.5726, arXiv:1411.5726 [cs].

[34] M. Denkowski, A. Lavie, Meteor Universal: Language Specific Translation Evaluation for Any Target Language, volume 6, 2014, pp. 376–380. doi:10.3115/v1/W14-3348.

[35] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: Semantic Propositional Image Caption Evaluation, 2016. URL: http://arxiv.org/abs/1607.08822. doi:10.48550/arXiv.1607.08822, arXiv:1607.08822 [cs].

[36] Y. Kim, Convolutional Neural Networks for Sentence Classification, in: A. Moschitti, B. Pang,

W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751. URL: https://aclanthology.org/D14-1181. doi:10.3115/v1/D14-1181.

[37] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805, arXiv: 1810.04805 tex.bibsource: dblp computer science bibliography, https://dblp.org tex.timestamp: Tue, 30 Oct 2018 20:39:56 +0100.

[38] Streamlit • A faster way to build and share data apps, 2021. URL: https://streamlit.io/.

[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015). URL: http://arxiv.org/abs/1512.03385, arXiv: 1512.03385 tex.bibsource: dblp computer science bibliography, https://dblp.org tex.timestamp: Wed, 25 Jan 2023 11:01:16 +0100.

[40] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 conference on empirical methods in natural language processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, 2021. URL: http://arxiv.org/abs/2103.00020. doi:10.48550/arXiv.2103.00020, arXiv:2103.00020 [cs].

[42] V. W. Anelli, A. Bellogin, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, T. Di Noia, Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 2405–2414. URL: https://dl.acm.org/doi/10.1145/3404835.3463245. doi:10.1145/3404835.3463245.

[43] G. Linden, B. Smith, J. York, Amazon.com recommendations: item-to-item collaborative filtering, IEEE Internet Computing 7 (2003) 76–80. URL: https://ieeexplore.ieee.org/document/1167344. doi:10.1109/MIC.2003.1167344, conference Name: IEEE Internet Computing.

[44] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, in: Proceedings of the 1994 ACM conference on Computer supported cooperative work, CSCW '94, Association for Computing Machinery, New York, NY, USA, 1994, pp. 175–186. URL: https://dl.acm.org/doi/10.1145/192844.192905. doi:10.1145/192844.192905.

[45] R. Socher, J. Bauer, C. D. Manning, A. Y. Ng, Parsing with Compositional Vector Grammars, in: H. Schuetze, P. Fung, M. Poesio (Eds.), Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 455–465. URL: https://aclanthology.org/P13-1045.

[46] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal dependencies, Computational Linguistics 47 (2021) 255–308. URL: https://doi.org/10.1162/coli_a_00402. doi:10.1162/coli_a_00402, tex.eprint: https://direct.mit.edu/coli/article-pdf/47/2/255/1938138/coli_a_00402.pdf.

[47] J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, D. Zeman, Universal Dependencies v2: An evergrowing multilingual treebank collection, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the twelfth language resources and evaluation conference, European Language Resources Association, Marseille, France, 2020, pp. 4034–4043. URL: https://aclanthology.org/2020.lrec-1.497/.

[48] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020.

[49] J. Aneja, A. Deshpande, A. Schwing, Convolutional Image Captioning, 2017. URL: http://arxiv.org/abs/1711.09151. doi:10.48550/arXiv.1711.09151, arXiv:1711.09151 [cs].

[50] C. Cai, K.-H. Yap, S. Wang, Attribute Conditioned Fashion Image Captioning, in: 2022 IEEE International Conference on Image Processing (ICIP), 2022, pp. 1921–1925. doi:10.1109/ICIP46576.

2022.9897417, iSSN: 2381-8549.

[51] B. McKinzie, Z. Gan, J.-P. Fauconnier, S. Dodge, B. Zhang, P. Dufter, D. Shah, X. Du, F. Peng, F. Weers, A. Belyi, H. Zhang, K. Singh, D. Kang, A. Jain, H. Hè, M. Schwarzer, T. Gunter, X. Kong, A. Zhang, J. Wang, C. Wang, N. Du, T. Lei, S. Wiseman, M. Lee, Z. Wang, R. Pang, P. Grasch, A. Toshev, Y. Yang, MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training, 2024. URL: http://arxiv.org/abs/2403.09611. doi:10.48550/arXiv.2403.09611, arXiv:2403.09611 [cs].

[52] X. Zhou, Z. Shen, A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation, in: Proceedings of the 31st ACM International Conference on Multimedia, MM '23, ACM, 2023, pp. 935–943. URL: http://dx.doi.org/10.1145/3581783.3611943. doi:10.1145/3581783.3611943.

[53] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, F. Jiang, Bootstrap Latent Representations for Multi-modal Recommendation, in: Proceedings of the ACM Web Conference 2023, WWW '23, ACM, 2023, pp. 845–854. URL: http://dx.doi.org/10.1145/3543507.3583251. doi:10.1145/3543507.3583251.