Do Recommender Systems Really Leverage Multimodal Content? A Comprehensive Analysis on Multimodal Representations for Recommendation

Claudio Pomo*
claudio.pomo@poliba.it
Politecnico Di Bari
Bari, Italy

Matteo Attimonelli*
matteo.attimonelli@poliba.it
Politecnico Di Bari
Bari, Italy
Sapienza University of Rome
Rome, Italy

Danilo Danese*
danilo.danese@poliba.it
Politecnico Di Bari
Bari, Italy

Fedelucio Narducci
fedelucio.narducci@poliba.it
Politecnico Di Bari
Bari, Italy

Tommaso Di Noia
tommaso.dinoia@poliba.it
Politecnico Di Bari
Bari, Italy

Abstract

Multimodal Recommender Systems aim to improve recommendation accuracy by integrating heterogeneous content, such as images and textual metadata. While effective, it remains unclear whether their gains stem from true multimodal understanding or increased model complexity. This work investigates the role of multimodal item embeddings, emphasizing the semantic informativeness of the representations. Initial experiments reveal that embeddings from standard extractors (e.g., ResNet50, Sentence-Bert) enhance performance, but rely on modality-specific encoders and ad hoc fusion strategies that lack control over cross-modal alignment. To overcome these limitations, we leverage Large Vision–Language Models (LVLMs) to generate *multimodal-by-design* embeddings via structured prompts. This approach yields semantically aligned representations without requiring any fusion. Experiments across multiple settings show notable performance improvements. Furthermore, LVLMs embeddings offer a distinctive advantage: they can be decoded into structured textual descriptions, enabling direct assessment of their multimodal comprehension. When such descriptions are incorporated as side content into recommender systems, they improve recommendation performance, empirically validating the semantic alignment encoded in LVLMs outputs. Our study highlights the importance of semantically rich representations and positions LVLMs as a compelling foundation to build robust and meaningful multimodal representations in recommendation tasks[1].

CCS Concepts

• **Computing methodologies** → **Natural language processing**; **Information extraction**; *Learning latent representations*; • **Information systems** → **Recommender systems**; **Information extraction**.

Keywords

Multimodal Recommender Systems, Large Vision–Language Models, Multimodal Representation Learning

ACM Reference Format:
Claudio Pomo, Matteo Attimonelli, Danilo Danese, Fedelucio Narducci, and Tommaso Di Noia. 2025. Do Recommender Systems Really Leverage Multimodal Content? A Comprehensive Analysis on Multimodal Representations for Recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3746252.3761398

1 Introduction

Recommender Systems (RSs) have emerged as one of the most extensively studied and widely applied technologies, facilitating personalized and on-demand information services. Despite the success of collaborative filtering techniques in large-scale platforms, RSs continue to encounter inherent limitations, including data sparsity [22, 41], cold-start [39, 42] problems, scalability constraints, which ultimately hinder recommendation accuracy. In response to these challenges, *Multimodal Recommender Systems* (MMRSs) [18, 53, 55, 56] have become a cutting-edge solution for enhancing user satisfaction and engagement. By integrating multiple content modalities, such as images and textual descriptions of items, MMRSs improve the recommendation pipeline, addressing data sparsity issues and enabling more precise and tailored recommendations. While MMRSs have demonstrated empirical success in leveraging multimodal content to boost accuracy [34], a critical question remains insufficiently addressed: *Do MMRSs truly benefit from the semantic content of input features, or are observed performance gains mainly related to increased model complexity?*

This work addresses the aforementioned question through a comprehensive investigation of multimodal embeddings in recommender systems. We begin by examining whether the performance gains reported in MMRSs arise from genuine cross-modal

*These authors contributed equally to this work.
[1]Our code is available at https://split.to/mmrs.

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761398

understanding or are instead attributable to incidental factors such as increased model capacity (e.g., number of parameters). To this end, we compare classical collaborative filtering models with various multimodal alternatives through a series of controlled experiments, including the use of (a) Gaussian noise, (b) multivariate structured noise, and (c) multimodal item embeddings. Our findings confirm that multimodal content can indeed enhance recommendation performance relative to classical baselines. Furthermore, we note critical limitations in how multimodal features are typically integrated into MMRSs: most existing systems rely on pre-trained encoders, such as ResNet50 [16] for visual content and Sentence-Bert [40] for textual metadata, as they represent the most widely adopted feature extractors in the field. The embeddings from these encoders are combined using methods like concatenation, summation, or element-wise multiplication based on dimension compatibility [34, 53, 55, 56]. This straightforward fusion approach has two main issues. First, it lacks a systematic way to manage the weighting and preservation of information from each modality. Second, it offers limited visibility into how recommender systems use the resulting composite representations, reducing control over the distribution of modality-specific information. Consequently, multimodal signals integration may lack transparency and robustness, undermining the semantic richness intended by multimodal data.

Motivated by these findings, we explore a more principled alternative: the use of *multimodal-by-design* representations, which are embeddings generated by models trained to natively process and align multiple modalities. In particular, we focus on Large Vision–Language Models (LVLMs) [9, 28], which extend Large Language Models (LLMs) [6, 44] by incorporating advanced computer vision capabilities. These models are trained to unify visual and textual information within a shared semantic space, effectively tackling tasks such as image captioning and visual question answering (VQA) [3, 46]. By prompting LVLMs to describe the visual content of item images, we facilitate embedding extraction through a VQA task, effectively leveraging their in-context learning (ICL) capability (e.g., the ability to adapt to new tasks without explicit fine-tuning). The resulting embeddings capture cross-modal semantics by coherently integrating visual and linguistic information within a unified latent space. Unlike traditional late-fusion methods that combine separately encoded unimodal features, our experiments demonstrate that this approach yields more effective representations, leading to improved MMRS performance.

Furthermore, differently from traditional late-fusion approaches, embeddings extracted from LVLMs can be decoded into the original structured textual descriptions. This allows the designer to directly inspect the multimodal understanding of the LVLMs through analysis of the generated outputs. To evaluate the semantic informativeness of these textual representations and the associated multimodal-by-design embeddings, we incorporate the generated descriptions as additional content signals in collaborative recommendation models. The resulting improvements in recommendation performance, measured against both simple collaborative filtering baselines and more advanced MMRSs, serve as indicators of how well these descriptions capture the relevant item semantics and their practical utility in recommendation. Experimental results show that leveraging such descriptions not only enhances recommendation quality relative to classical models but, in some cases, matches or surpasses the performance of more complex MMRS architectures. This confirms the strong multimodal comprehension of LVLMs and the rich semantic content embedded within their latent representations.

To summarize, our contributions are the following:

- We demonstrate that the performance of MMRSs is primarily influenced by the quality and informativeness of the input multimodal features, rather than merely by the model capacity (e.g., number of parameters).
- We investigate the potential of state-of-the-art LVLMs as a principled alternative for generating multimodal item embeddings, given their ability to natively integrate visual and textual modalities without requiring explicit fusion of separately encoded features, thus preserving transparency and robustness of the recommendation results.
- We provide empirical evidence that embeddings extracted from LVLMs capture item semantics by analyzing the textual descriptions they generate and directly associate with item embeddings. Moreover, incorporating these descriptions as side content into classical recommender models enhances performance, achieving results comparable to MMRSs.

The remainder of the paper reviews relevant literature on multimodal recommendation and vision–language modeling, details our methodology and experimental setup, presents the results, and concludes with directions for future work.

2 Related Work

This section reviews the key areas supporting our study: MMRSs and LVLMs. We first examine MMRSs, highlighting challenges in achieving true cross-modal understanding, particularly with ad hoc fusion of unimodal features. We then turn to LVLMs as a promising paradigm for unified multimodal representation learning.

2.1 Multimodal Recommendation

MMRSs aim to enrich the semantics of user and item representations by integrating item-side multimodal content, such as textual descriptions and visual illustrations, with traditional collaborative filtering signals [32]. This integration is particularly valuable for addressing common issues like data sparsity and the cold-start problem [30, 34], leading to improved recommendation accuracy across diverse domains including fashion [8], music [35], and micro-videos [7]. Early approaches, such as VBPR [18]), used matrix factorization to combine multimodal information with ID embeddings. Recent approaches adopt deep generative models [45, 52], graph neural networks (e.g., FREEDOM [55], LATTICE [53]), and self-supervised learning [43, 56] to better exploit modality interactions. Meta-learning [36] and attention mechanisms [14] further enhance adaptation and interpretability. Despite these advancements, challenges persist, particularly in learning effectively from limited or potentially biased and noisy user-item interaction data [8, 31, 50] and in achieving efficient fusion of diverse multimodal knowledge [27, 29, 34, 54]. Traditional multimodal approaches often rely on late-fusion techniques, such as concatenation, summation, or element-wise multiplication, to combine features from separate unimodal encoders [53, 55, 56]. However, these methods lack explicit mechanisms for controlling information flow and offer limited transparency into how downstream models utilize the resulting

composite representations. This raises questions about whether observed gains reflect genuine semantic understanding or merely increased model capacity. Moreover, the effectiveness of such fusion strategies in preserving cross-modal semantics remains unclear, highlighting a critical gap this work aims to address.

2.2 Large Vision-Language Models

Building on the capabilities of LLMs [6, 11, 37], Large Vision-Language Models (LVLMs) [1, 9, 13, 17, 20, 24, 28, 47, 49] extend this paradigm by integrating computer vision capabilities, enabling joint reasoning over textual and visual inputs. This multimodal integration makes LVLMs well-suited for tasks such as image captioning [46] and visual question answering (VQA) [3]. These tasks fundamentally involve verbalizing visual content, either by generating descriptive text or answering queries based on image understanding, and require sophisticated alignment and interpretation of visual and textual signals to achieve a deep semantic understanding of multimodal data. This ability to process and understand multimodal information makes LVLMs particularly relevant for domains such as MMRSs. A critical capability that underpins the versatility of both LLMs and LVLMs is In-Context Learning (ICL) [6, 11, 37]. ICL allows models to adapt to new tasks by leveraging a few illustrative examples provided directly within the input prompt, avoiding the need for computationally expensive retraining or fine-tuning. Specifically, Few-Shot ICL [6, 37] enables effective generalization with only a limited number of labeled instances, which is highly advantageous in data-scarce scenarios. Leveraging these advancements, our work explores the application of LVLMs embeddings within MMRSs. While prior work has explored multimodal integration in recommender systems using various strategies [5, 34], to the best of our knowledge, this is the first systematic use of LVLMs embeddings for this purpose. Unlike traditional approaches that fuse features from separate unimodal encoders, LVLMs features are inherently multimodal (*multimodal-by-design* [4, 5]), as they are learned from models jointly trained across modalities, capturing cross-modal interactions natively. Our approach builds upon and contributes to Multimodal Recommender Systems research.

3 Preliminaries

This work investigates the task of personalized recommendation in both classical and multimodal settings. The goal is to predict user preferences for unseen items based on prior interactions and, when available, multimodal content. In the classical setting, the task is formalized as the learning of a function:

$$f : \mathcal{U} \times \mathcal{I} \to \mathbb{R}, \tag{1}$$

which predicts a relevance score $\hat{y}_{ui} \in \mathbb{R}$ expressing the likelihood that user $u \in \mathcal{U}$ will prefer item $i \in \mathcal{I}$. Users and items are encoded as latent vectors $\mathbf{z}_u, \mathbf{z}_i \in \mathbb{R}^d$, where d denotes the dimensionality of the latent space. The relevance is computed via a scoring function $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

$$\hat{y}_{ui} = \phi(\mathbf{z}_u, \mathbf{z}_i), \tag{2}$$

where ϕ can be a dot product, cosine similarity, or a parameterized function. User embeddings may be directly learned or constructed by aggregating the embeddings of previously interacted items:

$$\mathbf{z}_u = g(\{\mathbf{z}_{i_j} \mid i_j \in \mathcal{I}_u\}), \tag{3}$$

with $\mathcal{I}_u \subseteq \mathcal{I}$ denoting the set of items interacted with user u, and $g(\cdot)$ being an aggregation function.

This framework is extended to a multimodal setting in which each item $i \in \mathcal{I}$ is associated with a set of modality-specific features $\{\mathbf{x}_i^{(m)} \in \mathbb{R}^{d_m} \mid m \in \mathcal{M}\}$, extracted from pre-trained models. Each modality $m \in \mathcal{M}$ is processed by a corresponding encoder $h^{(m)} : \mathbb{R}^{d_m} \to \mathbb{R}^d$, resulting in

$$\mathbf{z}_i^{(m)} = h^{(m)}(\mathbf{x}_i^{(m)}). \tag{4}$$

The item representation is enriched by the available modalities, yielding a content-aware recommendation function:

$$f(u, \{\mathbf{x}_i^{(m)}\}_{m \in \mathcal{M}}) = \phi(\mathbf{z}_u, \mathbf{z}_i) = \hat{y}_{ui} \tag{5}$$

where $\mathbf{z}_i$ integrates the multimodal features.

To investigate the role of multimodal representations, embeddings from three categories of feature extractors are considered: classical unimodal encoders, general-purpose multimodal encoders, and LVLMs. For classical models, features are extracted either from the penultimate layer or, in the case of transformers-based encoders, from the `[CLS]` (short for "Classification") token, which is a special classification token prepended to the input sequence whose final hidden state is typically used as a summary representation of the entire input [10]. In recommendation context, visual representations are derived from product images, while textual representations are obtained from item descriptions. For general-purpose multimodal encoders, embeddings are taken from the shared projection layer that aligns information across modalities into a unified latent space.

Differently from multimodal encoders, LVLMs are primarily designed for generative tasks such as VQA, rather than for embedding extraction. To adapt these models for use in recommendation, embedding extraction is reformulated as a VQA task: each item image is paired with a structured query designed to find salient attributes. This formulation enables the application of ICL [6, 37], a paradigm in which models generalize to new tasks by conditioning on a small set of illustrative examples embedded within the input prompt, thereby avoiding the need for task-specific fine-tuning. Using one-shot prompting, where a single example response is provided, LVLMs are guided to generate semantically informative textual descriptions (see Section 4.4). These outputs function as content representations suitable for recommendation, while also providing access to the internal states of the model that reflect its multimodal semantic understanding.

Due to the autoregressive nature of LVLMs, embeddings are extracted from the final hidden state associated with the last output token, specifically, the `[EOS]` (End-of-Sequence) token. This special token denotes the termination of a sequence and is commonly used in autoregressive models to signal the point at which text generation should stop. The hidden state corresponding to the `[EOS]` token captures the model's aggregated understanding of the entire input context and its expected response, providing a semantically aligned representation. Unlike mean pooling [25], which averages token embeddings across a sequence without regard to their order or contextual role, using the `[EOS]` representation [26, 48] preserves the model's causal structure and is suitable for downstream recommendation tasks. The resulting embeddings are inherently multimodal, eliminating the need for late-fusion of unimodal features. Additionally, to assess the multimodal comprehension of LVLMs, their

generated textual representations can be integrated into classical recommenders as auxiliary side information, enabling a quantitative evaluation of their informativeness. This dual use of LVLMs, as latent embeddings and interpretable text, supports a comprehensive assessment of their semantic relevance in recommendation.

4 Experimental Setting

This section outlines the experimental setup for validating the research questions. It begins with a description of the dataset used in the evaluation, followed by the multimodal feature extractors for item representations. Next, we introduce the trained recommendation models, and finally, we explain how to process outputs from LVLMs into content features for recommendation tasks.

4.1 Datasets

For this study, we selected three distinct product categories, Baby, Pets, and Clothing, from the comprehensive Amazon Reviews 2023 dataset [21]. This dataset provides extensive user-item interactions and rich multimodal metadata, including both visual and textual descriptions. These particular categories were chosen primarily because they exhibit inherent differences in the distribution and interaction patterns between users and items within the larger dataset, providing distinct characteristics for our analysis. Subsequently, specific preprocessing steps were applied to each of these selected category subsets. To ensure a minimum level of user and item activity within each set, k-core filtering was performed: the Baby and Pets categories were processed as 5-core, requiring every user and item to have at least 5 associated reviews. The Clothing category was processed as a 10-core, needing at least 10 reviews per user and item. Furthermore, to align with our focus on multimodal information, we exclusively included items that possessed both an associated textual description and at least one accompanying image. In cases where multiple images were available for a selected item, the first image identified by the `large` key within the metadata was consistently utilized. Descriptive statistics summarizing these curated category datasets are presented in Table 1. Finally, each processed category was partitioned into standard training, validation, and test sets following an 80%/10%/10% split, respectively, to facilitate model development and evaluation.

4.2 Multimodal Feature Extractors

To evaluate the extent to which MMRSs effectively leverage the semantic content of multimodal inputs, we conduct experiments using a diverse set of pre-trained feature extractors spanning visual and textual modalities. In addition to these learned representations, we also include synthetic noisy baselines generated from random noise and multivariate Gaussian distributions to assess model sensitivity to semantically uninformative features. We selected established encoders to perform unimodal feature extraction. For visual features, we employ **ResNet50 (RNet50)** [16], from which we extract activations from the final pooling layer as image embeddings. We also utilize the **Vision Transformer (ViT)** [12], a more recent transformer-based architecture for image understanding that models spatial relationships using self-attention. For ViT, item representations are derived from the output embedding of the `[CLS]` token in the final transformer layer, specifically using the *vit-base-patch16-224* checkpoints. For textual content, we use **Sentence-Bert (Sbert)** [40], a BERT-based sentence encoder fine-tuned for semantic textual similarity. Sbert maps item descriptions to dense vector representations via mean pooling over token embeddings, and we considered the *all-mpnet-base-v2* checkpoints.

Table 1: Statistics of the datasets.

Datasets	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{R}\rvert$	Sparsity (%)
Baby	108,897	17,836	330,771	99.983%
Pets	461,828	61,060	1,772,584	99.993%
Clothing	398,796	75,616	2,109,975	99.993%

Furthermore, to represent classical multimodal models trained with explicit vision-language alignment, we include **CLIP** [38]. This model is trained via contrastive learning to embed images and text into a shared latent space, and we extract the final projection from both the image and text encoders onto this shared space, utilizing the *clip-vit-large-patch14* checkpoints.

To incorporate more recent, inherently multimodal-by-design representations, we employed two LVLMs. Our choice was guided by their performance on general vision-language benchmarks [1, 49, 51], their proficient instruction-following capabilities, which are crucial for our VQA-based embedding extraction methodology, and their open availability at the time of our research. These models, **Phi-3.5-vision-instruct (Phi-3.5-VI)** [1] and **Qwen2-VL-7B-Instruct (Qwen2-VL)** [49], offer a practical balance of advanced capabilities and computational cost for our experiments. Specifically, Phi-3.5-VI is a lightweight multimodal autoregressive model designed to process and generate coherent image-grounded textual responses. For each item, we input the image along with a prompt instructing the model to describe its visual content. We then extract the final token embedding (i.e., the `[EOS]` representation) from the last hidden layer, which captures the semantics of both the input and the generated description by leveraging the autoregressive structure inherent to LVLMs. We consider the *Phi-3.5-vision-instruct* checkpoints. Similarly, Qwen2-VL is a vision–language model capable of complex cross-modal reasoning. As with Phi-3.5-VI, we provide both the image and a natural language prompt, and extract the `[EOS]` token embedding from the final hidden layer, relying on the *Qwen2-VL-7B-Instruct* checkpoints. These selected models span a range of architectural paradigms, offering a comprehensive suite for analyzing how the nature and quality of multimodal representations influence downstream recommendation performance.

4.3 Recommender Models

Here are the models selected as baselines. We compare our approach against a broad set of baseline methods to provide a comprehensive performance landscape, spanning traditional collaborative filtering techniques, content-aware models, and recent multimodal recommendation frameworks. Within traditional collaborative filtering, we include foundational models such as **Item-kNN** [41], a memory-based approach that calculates item similarity based on user interaction history; **BPR-MF** [23], a widely used matrix factorization model optimized for personalized ranking; and **LightGCN** [19], a simplified graph-based model focusing on essential neighborhood

Prompt for Baby Products

Imagine you're creating metadata for an image database of baby products. Your task is to select five keywords that best represent the image content. Fill in the blanks: **[Category]**, **[Age Group]**, **[Purpose]**, **[Material]**, **[Usage Context]**.

For example, your answer will be: [Category] {Feeding}, [Age Group] {Infant}, [Purpose] {Hygiene}, [Material] {Silicone}, [Usage Context] {Home}.

Prompt for Pets Items

Imagine you're creating metadata for an image database of pet-related items. Your task is to select five keywords that best represent the image content. Fill in the blanks: **[Category]**, **[Pet Type]**, **[Purpose]**, **[Material]**, **[Usage Context]**.

For example, your answer will be: [Category] {Toys}, [Pet Type] {Dog}, [Purpose] {Entertainment}, [Material] {Rubber}, [Usage Context] {Outdoor}.

Prompt for Clothing Items

Imagine you're creating metadata for an image database of clothing items. Your task is to select five keywords that best represent the image content. Fill in the blanks: **[Type]**, **[Color]**, **[Wear Location]**, **[Material]**, **[Style]**.

For example, your answer will be: [Type] {Dress}, [Color] {Red}, [Wear Location] {Torso}, [Material] {Cotton}, [Style] {Casual}.

Figure 1: Domain-specific prompts designed to guide LVLMs in generating structured keyword-based descriptions for items from the selected datasets.

aggregation. To evaluate the impact of incorporating basic content features, we include **Attribute Item-kNN** [15, 42], an extension of Item-kNN that utilizes item content, such as textual information, to compute similarity, which can be particularly beneficial for cold-start items. Representing various approaches to multimodal integration, we benchmark against several state-of-the-art multimodal recommender systems. **VBPR** [18] is an early content-aware extension of BPR-MF that integrates pre-extracted visual features. **LATTICE** [53] is a graph-based framework that learns enhanced item embeddings by mining latent structures and combining similarity graphs across different modalities using a graph convolutional network; these embeddings can then be used as input to standard latent factor models. **BM3** [56] is a self-supervised model that learns robust multimodal representations through contrastive learning, avoiding reliance on computationally expensive data augmentations. Finally, **FREEDOM** [55] is a multimodal model that improves recommendation by refining the user-item interaction graph using a frozen multimodal item-item similarity graph derived from LATTICE principles. This diverse selection enables a rigorous analysis of multimodal representation impact, isolating performance gains from semantic content versus model capacity, and assessing the efficacy of different multimodal integration strategies across various recommendation paradigms to address our research questions.

4.4 LVLMs Prompting and Answer Processing

We employed the LVLMs Qwen2-VL and Phi-3.5-VI for their proven performance in multimodal tasks like VQA. Each item image was described using concise, domain-specific prompts designed to produce structured, essential outputs without excessive irrelevant detail. For each of the three datasets, a tailored prompt format was defined based on visual inspection of item distributions. Each prompt included five representative keywords, a number selected for its balance between descriptiveness and discriminative power in qualitative evaluations. As shown in Figure 1, a single in-context example guided the generation process, enabling the models to adapt without fine-tuning. The keyword-based outputs adhered to the intended structure and reflected the LVLMs' nuanced understanding of the specified attributes (Figure 2). For instance, on the Clothing dataset with Qwen2-VL, the process yielded 3,239 unique keywords.

To assess the utility of LVLMs in recommendation, we adopt a twofold evaluation strategy. First, we evaluate the impact of the latent representations extracted from LVLMs within MMRSs , specifically, the embeddings associated with the `[EOS]` token that encapsulates the generated response (see Section 4.2). Second, to assess the semantic informativeness and multimodal understanding embedded in the generated descriptions, and the corresponding `[EOS]` representations, we incorporate them as additional content signals

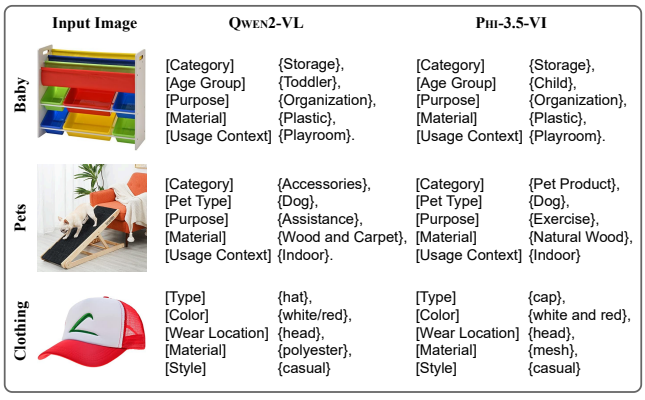
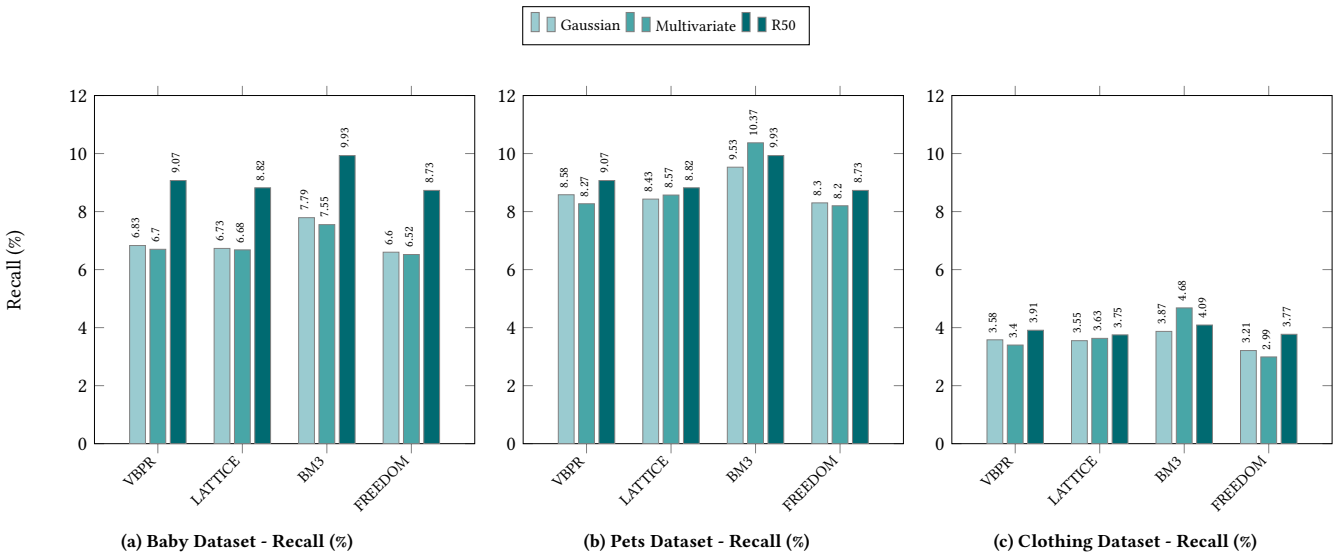


Figure 2: Examples of structured descriptions from LVLMs Qwen2-VL and Phi-3.5-VI for items in the selected datasets.



(a) Baby Dataset - Recall (%) (b) Pets Dataset - Recall (%) (c) Clothing Dataset - Recall (%)

Figure 3: Recall@20 (%) performance for four MMRSs on (a) Baby, (b) Pets, and (c) Clothing datasets. Models are evaluated using item representations from: (i) Gaussian noise, (ii) multivariate structured noise, and (iii) ResNet50 visual embeddings.

into a classical hybrid model (e.g., Attribute Item-kNN). For comprehensive benchmarking, this approach is compared against both standard collaborative baselines and MMRSs that do not utilize such descriptive content. To enable this integration, raw textual descriptions are transformed into structured keyword-based features through minimal preprocessing. This involves correcting formatting inconsistencies, applying lemmatization, and clustering keywords based on semantic similarity. The top-50 most frequent keywords per category are retained, while less frequent terms are mapped to a generic "Other" token. The final representation consists of 255 one-hot encoded features, which mitigates sparsity and improves the robustness of the Attribute Item-kNN model.

4.5 Implementation Details

Feature extraction was performed using the `Ducho` framework [5]. We extracted visual features with the RNet50, ViT, Sbert, and CLIP backbones, and integrated LVLMs features Qwen2-VL and Phi-3.5-VI via HuggingFace[2] models to obtain the `[EOS]` embeddings of the generated descriptions.

Recommendation training and evaluation were conducted using the `Elliot` framework [2]. We tested both similarity-based and learning-based recommenders. For Item-kNN and Attribute Item-kNN, we evaluated multiple similarity functions: `cosine`, `jaccard`, `dot`, `asym`, and `tversky`. The number of neighbors was varied within the range $\{5, 100\}$ and different feature weighting schemes were evaluated, including `TF_IDF` and `BM25`. In the specific case of Attribute Item-kNN, discrete keywords derived from the LVLMs text were required, as described in Section 4.4. In particular, lemmatization was performed using the NLTK toolkit [33], keywords were grouped applying agglomerative clustering using cosine distance, via the Sentence Transformers library[3]. For learning-based recommenders, we conducted a hyperparameter search across learning rates $\{0.0001, 0.0005, 0.001, 0.005, 0.01\}$, regularization values $\{10^{-2}, 10^{-1}\}$, and latent dimensions $\{64, 128, 256\}$, resulting in 30 explorations for each configuration.

Model performance was evaluated using standard top-K recommendation metrics: Recall@20, nDCG@20, and Hit Ratio@20. The best hyperparameters for each model were selected based on Recall@20 from the validation set to ensure a fair comparison. All experiments were conducted on a compute node with an NVIDIA A100 GPU, involving over *7,000* distinct training runs.

[2]https://huggingface.co
[3]https://sbert.net

5 Results

This section presents the results of the experimental evaluation, structured around the following research questions (RQs):

RQ1: *Do multimodal recommender systems truly benefit from the semantic content of input features, or are observed performance gains mainly relates to increased model complexity?*
RQ2: *Is there a principled and embedding-agnostic way to provide multimodal content to recommender systems?*
RQ3: *Do LVLMs embeddings faithfully capture semantic item content relevant to recommendation?*

5.1 Impact of Item Features in Recommendation (*RQ1*)

An important question is whether MMRSs genuinely leverage the semantic information embedded in item representations to enrich the user-item interaction matrix, or the observed performance improvements are primarily attributable to the increased model capacity, such as a larger number of parameters, required to process additional input modalities. To address this, we evaluate four representative MMRS architectures: VBPR, LATTICE, BM3, and FREEDOM (see Section 4.2), under controlled input settings. Specifically, we

examine model behavior when multimodal inputs are represented by: (i) pure Gaussian noise matching the dimensionality of standard item embeddings; (ii) multivariate noise designed to match the statistical properties (e.g., mean and variance) of real embeddings; and (iii) visual features extracted from RNet50, a widely adopted feature extractor in multimodal recommendation.

This experimental setup allows us to isolate the effect of model capacity, as all MMRSs receive input embeddings of identical dimensionality. As shown in Figure 3, which reports Recall@20 scores across the three datasets, VBPR, LATTICE, and FREEDOM consistently achieve higher performance when provided with semantically meaningful features (e.g., RNet50 representations). This indicates that these models benefit from the informative content of the input features, rather than from increased capacity alone. Interestingly, BM3 performs similarly on the Pets and Clothing datasets regardless of input type, with a slight improvement on Baby when using RNet50 features. While it does benefit from semantic features, the performance gap between noise-based and semantic inputs remains relatively small compared to other models, particularly on Pets and Clothing. This indicates that BM3's effectiveness may stem more from its architecture than in multimodal content. Its ability to utilize minimal structural cues in complex inputs might reduce its reliance on rich semantic features after a certain complexity threshold is reached.

With regard to **RQ1:** "**Do multimodal recommender systems truly benefit from the semantic content of input features, or are observed performance gains mainly relates to increased model capacity?**", our evaluation shows that MMRSs do indeed benefit from semantically meaningful item representations. This positive effect is apparent across most models; however, it is less pronounced for the BM3 model. This suggests that the performance improvements seen with BM3 may be more influenced by its architectural complexity rather than the quality of the input features.

5.2 LVLMs Representations for Multimodal Recommendation (*RQ2*)

Previous sections have demonstrated that multimodal representations can notably improve the performance of RSs. A widely adopted approach involves combining modality-specific features, typically extracted using dedicated encoders such as RNet50 for visual inputs and Sbert for textual data. However, this late-fusion approach has limitations, including potential semantic misalignment and limited control over the distribution of modality-specific information within the fused latent representation. To evaluate these methods, we analyze various multimodal architectures in the recommendation pipeline, including: (i) RNet50 and ViT as traditional vision backbones; (ii) a baseline that concatenates embeddings from RNet50 and Sbert, in line with previous works [34, 53, 55, 56]; and (iii) CLIP, a multimodal backbone. Since it is not trained on the Amazon Reviews 2023 dataset, its visual and textual representations may not align semantically within the shared latent space. Therefore, consistent with (ii), we concatenate CLIP's visual and textual embeddings to preserve the informational content of both modalities. Additionally, we examine Qwen2-VL and Phi-3.5-VI for their ability to produce multimodal representations without fusion. These models use structured queries to generate concise and content-rich descriptions of item images [1, 49, 51], providing unified embeddings as direct inputs for recommendation models.

Our empirical evaluation, presented in Table 2, shows that across all tested recommender models (VBPR, LATTICE, BM3, FREEDOM), LVLMs embeddings consistently rank among the top two performers. A critical aspect of our analysis is the statistical significance of performance differences between the leading and runner-up extractors for each model. For instance, with the VBPR model, Qwen2-VL not only achieves the highest scores across all datasets and metrics, but these scores are also statistically significantly superior ($p < 0.05$, indicated by * in the table) to those of the second-best extractor, Phi-3.5-VI. This highlights that even within the LVLMs family, specific model choices can yield significant performance variations. Similarly, for the BM3 model, the best-performing LVLM extractor (Qwen2-VL on Pets; Phi-3.5-VI on Baby and Clothing) notably outperforms the respective second-best option in each case.

Although LVLMs demonstrate strong performance, the traditional RNet50 – Sbert combination remains a robust baseline in certain contexts. Notably, for the FREEDOM model, RNet50 – Sbert consistently delivers the best results across all datasets, surpassing LVLMs extractors which hold the second position. In the LATTICE model, RNet50 – Sbert performs better on Baby than the second-best LVLM, Qwen2-VL. However, LVLMs hold the significantly leading positions on the other datasets for LATTICE. These cases show that the performance gap between top and succeeding extractors is often statistically significant, highlighting the need for careful extractor selection based on the recommendation model and dataset characteristics.

Despite these model-specific nuances where traditional methods can occasionally excel, a broader perspective is offered by the Borda count analysis (Table 3). This analysis provides an aggregated view of extractor performance. In essence, for each specific evaluation scenario (a given recommender model, dataset, and metric), the extractors are ranked. Points are awarded based on these ranks (e.g., the top-performing extractor gets the most points, the second gets fewer, and so on). These points are then summed across all scenarios for each extractor. The resulting total score offers a consolidated measure of which extractor performs the most consistently well across the diverse range of tested conditions. This aggregate view reveals that, on average, Qwen2-VL and Phi-3.5-VI secure the highest rankings across the majority of extractor-dataset-model-metric combinations. This evidence strongly supports the efficacy of LVLMs embeddings, attributing their success to the creation of semantically grounded, multimodal-by-design representations.

In response to **RQ2:** "**Is there a principled and embedding-agnostic way to provide multimodal content to recommender systems?**", the answer is yes, using LVLMs. Experimental results indicate that multimodal-by-design embeddings associated with LVLMs generated descriptions consistently outperform classical fusion approaches, such as the combination of RNet50 and Sbert, as well as alternative multimodal-by-design solutions like CLIP. These embeddings also outperform unimodal visual backbones such as RNet50 and ViT, demonstrating the semantic richness of LVLMs representations. By avoiding late-fusion of independently derived modality-specific embeddings, LVLMs representations enable a more coherent and controlled integration of multimodal signals, reducing the risk of cross-modal misalignment.

Table 2: Top-20 performance (Recall, nDCG, HR as %) of MMRS models using different feature extractors on Baby, Pets, and Clothing datasets. Boldface/underline denote best/second-best extractor performance within each MMRSs model. ⋆ indicates that the improvement of the best-performing extractor over the second-best is statistically significant ($p < 0.05$) within the specific set of evaluated extractors for that MMRSs model.

Models	Extractors	Baby			Pets			Clothing		
		Recall ↑	nDCG ↑	HR ↑	Recall ↑	nDCG ↑	HR ↑	Recall ↑	nDCG ↑	HR ↑
VBPR	Qwen2-VL	**7.690⋆**	**3.460⋆**	**8.140⋆**	**9.340⋆**	**4.320⋆**	**10.180⋆**	**4.060⋆**	**1.820⋆**	**4.810⋆**
	Phi-3.5-VI	7.640	3.440	8.100	9.260	4.280	10.100	3.990	1.800	4.720
	CLIP	7.490	3.340	7.940	9.170	4.220	10.010	3.960	1.780	4.690
	RNet50 – Sbert	7.550	3.370	8.020	9.230	4.250	10.070	3.940	1.780	4.670
	RNet50	7.400	3.300	7.860	9.070	4.180	9.900	3.910	1.760	4.630
	ViT	7.390	3.270	7.840	8.990	4.140	9.810	3.940	1.780	4.660
LATTICE	Qwen2-VL	7.410	3.210	7.860	9.030	4.080	9.830	**4.050⋆**	**1.800⋆**	**4.790⋆**
	Phi-3.5-VI	7.330	3.210	7.790	**9.070⋆**	**4.110⋆**	**9.880⋆**	3.850	1.710	4.560
	CLIP	7.150	3.130	7.610	8.740	3.930	9.530	3.771	1.680	4.480
	RNet50 – Sbert	**7.450**	**3.260⋆**	**7.890**	9.020	4.090	9.830	3.910	1.750	4.630
	RNet50	7.240	3.120	7.680	8.820	3.970	9.610	3.750	1.670	4.440
	ViT	7.200	3.130	7.630	8.820	3.990	9.640	3.780	1.680	4.480
BM3	Qwen2-VL	7.630	3.340	8.080	**10.360**	**4.950**	**11.320**	4.680	2.190	5.570
	Phi-3.5-VI	**7.680⋆**	**3.370**	**8.130⋆**	10.290	4.920	11.250	**4.700⋆**	**2.210⋆**	**5.610⋆**
	CLIP	7.670	3.350	**8.130⋆**	9.510	4.340	10.370	3.990	1.790	4.750
	RNet50 – Sbert	7.660	3.290	8.100	9.510	4.330	10.370	3.950	1.760	4.690
	RNet50	7.640	3.290	8.080	9.930	4.600	10.850	4.090	1.860	4.870
	ViT	7.600	3.320	8.050	9.700	4.470	10.580	4.150	1.900	4.930
FREEDOM	Qwen2-VL	7.780	3.400	8.240	9.120	4.070	9.900	4.130	1.810	4.880
	Phi-3.5-VI	7.530	3.220	7.980	8.950	3.980	9.740	4.020	1.750	4.760
	CLIP	6.230	2.520	6.620	7.740	3.380	8.430	2.910	1.210	3.450
	RNet50 – Sbert	**7.810**	**3.420**	**8.270**	**9.470⋆**	**4.320⋆**	**10.300⋆**	**4.220⋆**	**1.870⋆**	**4.990⋆**
	RNet50	7.470	3.200	7.920	8.730	3.890	9.510	3.770	1.630	4.470
	ViT	7.420	3.190	7.860	8.840	3.940	9.620	3.900	1.700	4.620

Table 3: Aggregated Borda count scores for feature extractors. Scores are derived from Recall@20 and general performance ranks across all MMRSs models and datasets, with higher scores indicating superior overall extractor performance.

Rank	Extractor	Baby	Pets	Clothing	Overall
1	Qwen2-VL	14.0	18.0	18.0	50.0
2	Phi-3.5-VI	15.0	16.0	15.0	46.0
3	RNet50 – Sbert	16.0	11.5	10.5	38.0
4	RNet50	7.0	6.5	3.0	16.5
5	ViT	2.0	5.5	8.5	16.0
6	CLIP	6.0	2.5	5.0	13.5

5.3 Semantic Alignment of LVLMs Representations (*RQ3*)

As previously discussed, multimodal-by-design representations generated by LVLMs frequently improve the performance of MMRSs compared to other multimodal configurations. In particular, the benefits extend beyond improvements in embedding quality alone. As detailed in Sections 3 and 4.4, these embeddings are derived from the final hidden state corresponding to the `[EOS]` token in the LVLMs output. By exploiting the autoregressive nature of LVLMs, this token captures the full semantic content of the generated description, effectively representing its multimodal understanding. Consequently, each LVLM embedding is intrinsically linked to a coherent textual description produced by the model itself.

To evaluate the semantic richness and utility of the generated descriptions, we incorporate them as additional content features into the classical hybrid model Attribute Item-kNN. As detailed in Section 4.4, LVLMs generated descriptions are pre-processed into keyword-based features and converted into one-hot vectors for seamless integration. We then assess the impact of these features on recommendation performance relative to both traditional collaborative filtering baselines (Random, MostPop, Item-kNN, BPR-MF, LightGCN) and multimodal recommendation systems (VBPR, LATTICE, BM3, FREEDOM) that directly utilize LVLMs embeddings.

The comparative results presented in Table 4 demonstrate that incorporating LVLMs textual descriptions into Attribute Item-kNN yields notable performance improvements compared to all baseline methods. While these gains are moderate on the Baby dataset, they are particularly pronounced on the Pets and Clothing datasets, with the exception of BM3 on Pets which retains a performance advantage when utilizing LVLMs embeddings. In general, Attribute Item-kNN, especially when leveraging Phi-3.5-VI attributes, not only surpasses all classical baselines but also matches or exceeds the performance of several advanced MMRSs. For example, on the Clothing dataset, Attribute Item-kNN consistently outperforms models such as VBPR and LATTICE. Overall, Attribute Item-kNN demonstrates

Table 4: Top-20 performance (Recall, nDCG, HR as %) for all configurations of RSs, extractors, and datasets. Boldface/underlined denote best/second-best values for each recommendation model. ⋆ indicates $p < 0.05$ significance.

Models	Extractors	Baby			Pets			Clothing		
		Recall ↑	nDCG ↑	HR ↑	Recall ↑	nDCG ↑	HR ↑	Recall ↑	nDCG ↑	HR ↑
Random	—	1.046	0.362	0.115	0.036	0.014	0.039	0.028	0.010	0.033
MostPop	—	5.968	2.358	4.173	4.173	1.780	4.512	1.971	0.081	2.253
Item-kNN	—	6.785	3.159	7.233	9.229	4.523	10.061	3.997	1.973	4.786
BPR-MF	—	7.031	2.903	7.440	8.787	4.024	9.610	3.572	1.595	4.255
LightGCN	—	7.468	3.168	7.918	9.156	4.122	10.003	3.801	1.638	4.500
Attribute Item-kNN	Qwen2-VL	6.965	3.256	7.401	9.357	4.598	10.195	4.073	2.012	4.860
	Phi-3.5-VI	6.980	3.260	7.410	9.370	4.600	10.200	4.070	2.010	4.860
VBPR	Qwen2-VL	7.690	**3.460⋆**	8.140	9.340	4.320	10.180	4.060	1.820	4.810
	Phi-3.5-VI	7.640	3.440	8.100	9.260	4.280	10.100	3.990	1.800	4.720
LATTICE	Qwen2-VL	7.410	3.210	7.860	9.030	4.080	9.830	4.050	1.800	4.790
	Phi-3.5-VI	7.330	3.210	7.790	9.070	4.110	9.880	3.850	1.710	4.560
BM3	Qwen2-VL	7.630	3.340	8.080	**10.360**	**4.950**	**11.320**	4.680	2.190	5.570
	Phi-3.5-VI	7.680	3.370	8.130	10.290	4.920	11.250	**4.700⋆**	**2.210⋆**	**5.610⋆**
FREEDOM	Qwen2-VL	**7.780⋆**	3.400	**8.240⋆**	9.120	4.070	9.900	4.130	1.810	4.880
	Phi-3.5-VI	7.530	3.220	7.980	8.950	3.980	9.740	4.020	1.750	4.760

strong competitiveness, with consistently high nDCG scores that highlight the notable ranking capabilities of kNN-based models. This effect is particularly evident when the model is enhanced with LVLMs content, emphasizing the considerable informativeness and practical utility of these multimodal textual descriptions.

In light of previous observations, we positively address **RQ3:** "**Do LVLMs embeddings faithfully capture semantic item content relevant to recommendation?**". By extracting keywords from LVLMs textual descriptions and incorporating them as features into a straightforward content-based recommender, Attribute Item-kNN, we demonstrate consistent performance improvements. These gains, observed across multiple datasets, frequently allow this content-aware approach to outperform classical collaborative filtering methods and, in several cases, advanced MMRSs that directly leverage LVLMs embeddings. This provides strong quantitative evidence of the semantic informativeness and practical utility of LVLMs textual content. The dual use of LVLMs outputs, as latent embeddings and textual features, enhances even simple models, elevating their performance to competitive, and often statistically superior, levels. These findings underscore the value of LVLMs in generating rich multimodal representations for recommendation.

6 Conclusion

This research demonstrates that the effectiveness of MMRSs is primarily influenced by the semantic content of multimodal inputs rather than solely by the increased model complexity. Our study shows that LVLMs provide a robust and systematic approach to create items embeddings *multimodal-by-design*. These embeddings effectively capture cross-modal semantics without requiring explicit fusion, consistently improving recommendation performance across various MMRS architectures and often outperforming traditional late-fusion techniques. A key contribution of this work is the validation of the semantic relevance of the representations generated by LVLMs. The ability to convert these embeddings into structured textual descriptions enhances classical recommender models when used as supplementary content. This provides direct evidence of the models' deep understanding of multimodal data and their practical usefulness. The dual capability of offering both powerful latent representations and interpretable textual outputs underscores the value of LVLMs. Our findings support the adoption of high-quality, semantically aligned multimodal representations as a fundamental element in advancing recommendation systems. LVLMs present a promising foundation for developing more powerful, interpretable, and effective MMRSs. Future research may focus on tailoring LVLMs specifically for recommendation systems, exploring two primary aspects: first, utilizing LVLMs directly to generate recommendations, and second, enhancing the quality of explanations provided alongside those recommendations.

Acknowledgments

This work has been carried out while *Matteo Attimonelli* was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with *Politecnico Di Bari*. We thank the CINECA award under the ISCRA initiative for providing high-performance computing resources and support. This work was partially supported by the following projects: LUTECH DIGITALE 4.0. "VAI2C"- Virtual Artificial Intelligence Contact Center. "IDENTITA" CUP D93C22001020008. "REACH-XY": RESEARCH ACTIONS FOR REDUCING THE IMPACT ON AGRICULTURAL AND NATURAL ECOSYSTEMS OF THE HARMFUL PLANT PATHOGEN XYLELLA FASTIDIOSA, CUP B93C22001920001. "Patto Territoriale sistema universitario pugliese" CUP F61B23000370006- cod. id. PATTI_TERRITORIALI_WP1. Natuzzi S.p.A. del Contratto di Sviluppo Industriale ai sensi dell'art. 9 del Decreto del Ministro dello Sviluppo Economico del 09.12.2014. P+ARTS - Partnership for Artistic Research in Technology and Sustainability, funded by the European Union Next- GenerationEU (NRRP – M4C1, Investment 3.4, INTAFAM00037, CUP: G43C24000640006)

GenAI Usage Disclosure

No content generated by AI technologies has been used in this assessment.

References

[1] Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *CoRR* abs/2404.14219 (2024).

[2] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2405–2414. doi:10.1145/3404835.3463245

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*. IEEE Computer Society, 2425–2433.

[4] Matteo Attimonelli, Danilo Danese, Angela Di Fazio, Daniele Malitesta, Claudio Pomo, and Tommaso Di Noia. 2024. Ducho meets Elliot: Large-scale Benchmarks for Multimodal Recommendation. *arXiv preprint arXiv:2409.15857* (2024).

[5] Matteo Attimonelli, Danilo Danese, Daniele Malitesta, Claudio Pomo, Giuseppe Gassi, and Tommaso Di Noia. 2024. Ducho 2.0: Towards a More Up-to-Date Unified Framework for the Extraction of Multimodal Features in Recommendation. In *WWW (Companion Volume)*. ACM, 1075–1078.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

[7] Desheng Cai, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2022. Heterogeneous Hierarchical Feature Aggregation Network for Personalized Micro-Video Recommendation. *IEEE Trans. Multim.* 24 (2022), 805–818.

[8] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation. In *SIGIR*. ACM, 765–774.

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *NeurIPS*.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.

[11] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, and Zhifang Sui. 2024. A Survey on In-context Learning. In *EMNLP*. Association for Computational Linguistics, 1107–1128.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*. OpenReview.net.

[13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 Herd of Models. *CoRR* abs/2407.21783 (2024).

[14] Hong Fang, Jindong Liang, and Leiyuxin Sha. 2025. Enhanced multimodal recommendation systems through reviews integration. *Knowl. Inf. Syst.* 67, 4 (2025), 3459–3486.

[15] SongJie Gong, HongWu Ye, and XiaoYan Shi. 2008. A Collaborative Recommender Combining Item Rating Similarity and Item Attribute Similarity. In *2008 International Seminar on Business and Information Management*, Vol. 2. 58–60. doi:10.1109/ISBIM.2008.190

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society, 770–778.

[17] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient Multimodal Learning from Data-centric Perspective. *CoRR* abs/2402.11530 (2024).

[18] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *AAAI*. AAAI Press, 144–150.

[19] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *SIGIR*. ACM, 639–648.

[20] Musashi Hinck, Matthew L. Olson, David Cobbley, Shao-Yen Tseng, and Vasudev Lal. 2024. LLaVA-Gemma: Accelerating Multimodal Foundation Models with a Compact Language Model. *CoRR* abs/2404.01331 (2024).

[21] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian J. McAuley. 2024. Bridging Language and Items for Retrieval and Recommendation. *CoRR* abs/2403.03952 (2024).

[22] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *ICDM*. IEEE Computer Society, 263–272.

[23] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.

[24] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *CoRR* abs/2405.02246 (2024).

[25] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. In *ICLR*. OpenReview.net.

[26] Chaofan Li, Minghao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Defu Lian, Yingxia Shao, and Zheng Liu. 2025. Making Text Embedders Few-Shot Learners. In *ICLR*. OpenReview.net.

[27] Fan Liu, Huilin Chen, Zhiyong Cheng, Anan Liu, Liqiang Nie, and Mohan S. Kankanhalli. 2023. Disentangled Multimodal Representation Learning for Recommendation. *IEEE Trans. Multim.* 25 (2023), 7149–7159.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *NeurIPS*.

[29] Han Liu, Yinwei Wei, Fan Liu, Wenjie Wang, Liqiang Nie, and Tat-Seng Chua. 2024. Dynamic Multimodal Fusion via Meta-Learning Towards Micro-Video Recommendation. *ACM Trans. Inf. Syst.* 42, 2 (2024), 47:1–47:26.

[30] Kang Liu, Feng Xue, Dan Guo, Le Wu, Shujie Li, and Richang Hong. 2023. MEGCF: Multimodal Entity Graph Collaborative Filtering for Personalized Recommendation. *ACM Trans. Inf. Syst.* 41, 2 (2023), 30:1–30:27.

[31] Xiaohao Liu, Zhulin Tao, Jiahong Shao, Lifang Yang, and Xianglin Huang. 2022. EliMRec: Eliminating Single-modal Bias in Multimedia Recommendation. In *ACM Multimedia*. ACM, 687–695.

[32] Yong Liu, Susen Yang, Chenyi Lei, Guoxin Wang, Haihong Tang, Juyong Zhang, Aixin Sun, and Chunyan Miao. 2021. Pre-training Graph Transformer with Multimodal Side Information for Recommendation. In *ACM Multimedia*. ACM, 2853–2861.

[33] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. *CoRR* cs.CL/0205028 (2002). https://arxiv.org/abs/cs/0205028

[34] Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, Felice Antonio Merra, Tommaso Di Noia, and Eugenio Di Sciascio. 2025. Formalizing Multimedia Recommendation through Multimodal Deep Learning. *Trans. Recomm. Syst.* 3, 3 (2025), 37:1–37:33.

[35] Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra. 2017. A Deep Multimodal Approach for Cold-start Music Recommendation. In *DLRS@RecSys*. ACM, 32–37.
[36] Xingyu Pan, Yushuo Chen, Changxin Tian, Zihan Lin, Jinpeng Wang, He Hu, and Wayne Xin Zhao. 2022. Multimodal Meta-Learning for Cold-Start Sequential Recommendation. In *CIKM*. ACM, 3421–3430.
[37] Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. 2023. Bidirectional Language Models Are Also Few-shot Learners. In *ICLR*. OpenReview.net.
[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763.
[39] Al Mamunur Rashid, István Albert, Dan Cosley, Shyong K. Lam, Sean M. McNee, Joseph A. Konstan, and John Riedl. 2002. Getting to know you: learning new user preferences in recommender systems. In *IUI*. ACM, 127–134.
[40] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP/IJCNLP (1)*. Association for Computational Linguistics, 3980–3990.
[41] Badrul Munir Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW*. ACM, 285–295.
[42] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. 2002. Methods and metrics for cold-start recommendations. In *SIGIR*. ACM, 253–260.
[43] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2023. Self-Supervised Learning for Multimedia Recommendation. *IEEE Trans. Multim.* 25 (2023), 5107–5116.
[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023).
[45] Nhu-Thuat Tran and Hady W. Lauw. 2022. Aligning Dual Disentangled User Representations from Ratings and Textual Content. In *KDD*. ACM, 1798–1806.
[46] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*. IEEE Computer Society, 3156–3164.
[47] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. *Trans. Mach. Learn. Res.* 2022 (2022).
[48] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving Text Embeddings with Large Language Models. In *ACL (1)*. Association for Computational Linguistics, 11897–11916.
[49] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *CoRR* abs/2409.12191 (2024).
[50] Wei Yang, Zhengru Fang, Tianle Zhang, Shiguang Wu, and Chi Lu. 2023. Modal-aware Bias Constrained Contrastive Learning for Multimodal Recommendation. In *ACM Multimedia*. ACM, 6369–6378.
[51] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. *CoRR* abs/2409.02813 (2024).
[52] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative Knowledge Base Embedding for Recommender Systems. In *KDD*. ACM, 353–362.
[53] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *ACM Multimedia*. ACM, 3872–3880.
[54] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. 2023. Latent Structure Mining With Contrastive Modality Fusion for Multimedia Recommendation. *IEEE Trans. Knowl. Data Eng.* 35, 9 (2023), 9154–9167.
[55] Xin Zhou and Zhiqi Shen. 2023. A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation. In *ACM Multimedia*. ACM, 935–943.
[56] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap Latent Representations for Multimodal Recommendation. In *WWW*. ACM, 845–854.