Evaluating Vision-Language Model Embeddings for Multimodal Recommendation

Chi Zhang[1]
[1]ExaWizards Inc.
chi.zhang@exawizards.com

Abstract

Multimodal recommendation systems typically rely on separately trained encoders for visual and textual features, which may lead to representation misalignment. In this paper, we investigate whether vision-language models (VLMs) with joint representations can improve recommendation performance. We evaluate VLM2Vec-derived features across five strategies (visual-only, textual-only, two early-fusion variants, and dual-tower late fusion) on two Amazon datasets using FREEDOM and LATTICE models. Results show that VLM2Vec textual features achieve strong performance comparable to dual-tower fusion, while visual features and early fusion underperform. Our findings could provide practical guidance for feature engineering in multimodal recommendation.

1 Introduction

Recommendation systems play a crucial role in modern information systems, enabling personalized content discovery across various domains from streaming services to e-commerce. While traditional collaborative filtering methods rely solely on user-item interaction patterns [1, 2], they often suffer from cold-start problems and fail to capture rich product semantics. Multimodal recommendation systems address these limitations by leveraging multiple data modalities such as product images, textual descriptions and even audio or video recordings [3, 4].

The dominant paradigm in multimodal recommendation employs independently trained encoders for each modality: convolutional neural networks (CNNs) [5] for images and pretrained language models [6] for text. These features are combined through various fusion strategies—early fusion, late fusion, or hybrid approaches [7]. Recent graph-based methods [8, 9] further enhance this paradigm by propagating multimodal features through user-item interaction graphs.

However, this separate encoding paradigm suffers from fundamental limitations. Independently trained encoders learn representations in different semantic spaces, leading to representation misalignment where visual and textual features may not be properly coordinated [10]. Additionally, existing multimodal recommendation systems usually follow previous models to employ features generated by older pretrained models such as ResNet [5] for vision and BERT [6] for text, potentially missing advances in modern vision-language pretraining.

Recent advances in vision-language models (VLMs) such as CLIP [10] and VLM2Vec [11] offer a promising alternative. These models are pretrained on large-scale image-text pairs to learn unified representations where visual and textual modalities are naturally aligned. Unlike traditional approaches, VLMs enable cross-modal interactions during encoding and can generate features from either modality or fused representations through joint encoding.

In this paper, we conduct a systematic evaluation of VLM2Vec [11]— a state-of-the-art vision-language embedding model—across multiple feature extraction strategies on product recommendation tasks. Our contributions are: (1) systematic evaluation of five VLM2Vec feature strategies on recommendation tasks; (2) comprehensive experiments on two Amazon domains using two state-of-the-art models (FREEDOM and LATTICE); and (3) empirical evidence that VLM2Vec textual features achieve competitive performance with only half the feature dimensions, while visual features consistently underperform.

2 Related Work

Multimodal Recommendation Graph-based collaborative filtering has become a dominant paradigm

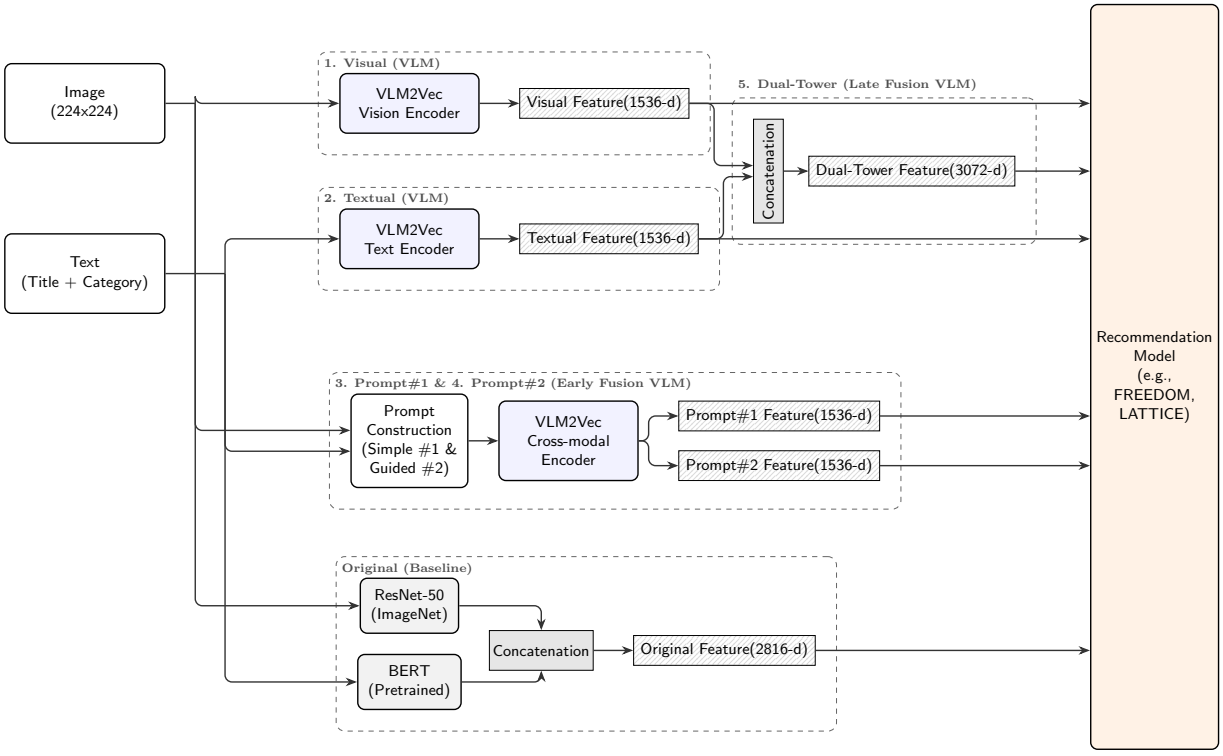

Figure 1 The overview of our proposed feature extraction strategies and recommendation architecture.

for recommendation systems. MMGCN [4] pioneered modality-specific graph convolutions with separate user-item graphs for each modality. LATTICE [9] and Dual-GNN [12] further advanced the field by employing contrastive learning to align multimodal signals and introducing user-user graphs to model preference patterns, respectively. More recently, BM3 [13] simplifies self-supervised frameworks through direct representation perturbation, and FREEDOM [8] advances this paradigm with feature-level denoising and frozen item-item graphs.

Despite these advances, existing methods all rely on separately trained encoders— ResNet [5] for vision and BERT [6] for text— which produce features in disjoint semantic spaces and may not leverage recent advances in vision-language pretraining.

Vision-Language Models Since CLIP [10] revolutionizes vision-language learning by training dual encoders with contrastive objectives on large-scale image-text pairs, VLMs such as BLIP-2 [14] and LLaVA [15] have reached great results on various vision and language tasks. More recently, Qwen2-VL [16] advances dynamic resolution processing and video understanding capabilities. VLM2Vec [11] extends this paradigm by providing flexible feature extraction—visual-only, textual-only, or early-fused representations through shared transformer layers. While VLMs excel at vision-language tasks, their application to recommendation systems remains underexplored.

3 Methodology

Given that existing multimodal recommendation systems predominantly rely on features from older pretrained models such as ResNet and BERT, we investigate whether modern vision-language models can provide superior representations. Additionally, since most approaches process modality features separately, we explore whether VLM-derived fused embeddings— generated through joint encoding before input to recommendation models—can improve performance over traditional late fusion strategies.

3.1 VLM2Vec Feature Extraction

We use VLM2Vec-V2.0 [11] with Qwen2-VL-7B as the base architecture. The model processes image patches through a Vision Transformer and text through a shared transformer encoder, supporting both single-modality and cross-modal encoding. We extract 1536-dimensional embeddings via EOS token pooling. We evaluate five feature

extraction strategies, which can be seen in Figure 1.

Visual Features (Visual) Product images are encoded through VLM2Vec's vision encoder without textual context. Each image is preprocessed to 224×224 resolution and encoded into a 1536-dimensional embedding, isolating the contribution of visual information.

Textual Features (Textual) Product descriptions are encoded through VLM2Vec's text encoder. We concatenate product titles and category labels formatted as: "[TITLE], category: [CATEGORY]".

Fused Features (Prompt#1 and Prompt#2) We design prompt-based early fusion strategies leveraging VLM2Vec's cross-modal encoding:

- **Prompt#1 (Simple)**: Direct concatenation of [image] + [text] + [instruction: "Generate a unified representation for this product."]
- **Prompt#2 (Guided)**: Structured format with explicit task framing: ["You are analyzing a product for retrieval purposes. Here is the product image:"] + [image] + "Product details:" + [text] + [instruction], the same instruction as in Prompt#1.

The usage of two prompts is to test whether explicit task instructions improve VLM's ability to generate representations better for recommendation tasks.

Dual-Tower Features (Dual) Visual and textual features are extracted separately and concatenated into a 3072-dimensional representation, testing whether late fusion outperforms early cross-modal fusion.

Original Features (Original) As baseline, we use ResNet-50 [5] pretrained on ImageNet (2048-dim) and BERT [6] (768-dim), concatenated to form 2816-dimensional representations.

3.2 Experimental Setup

3.2.1 Datasets

We conduct experiments on two Amazon product recommendation datasets [17]: **Baby** and **Sports & Outdoors**. Both contain user-item interactions, product images, and metadata. We apply standard preprocessing: filtering users and items with fewer than 5 interactions, and using temporal split (80% training, 10% validation, 10% test). Table 1 summarizes the statistics.

Table 1 Statistics of the experimental datasets.

Dataset	# Users	# Items	# Interactions	Density (%)
Baby	19,445	7,050	160,792	0.12
Sports	35,598	18,357	296,337	0.05

3.2.2 Recommendation Models

We evaluate using two state-of-the-art multimodal recommendation models:

FREEDOM [8] employs graph-based multimodal fusion with feature-level denoising mechanisms, constructing modality-specific user-item graphs and using learnable attention to weight modalities.

LATTICE [9] leverages contrastive learning to align multimodal features with collaborative signals, constructing item-item similarity graphs and applying contrastive objectives.

3.2.3 Implementation Details

We use the code base [1)] provided by MMRec [18] and follow most settings. Both models use BPR loss [1] for pairwise ranking. We conduct the experiments using the default set of hyperparameters in MMRec, only replacing input features.

We report Recall@20, NDCG@20, and Precision@20 on the test set using the best hyperparameter selected by validation performance. We train on a single NVIDIA A100 GPU with batch size 2048 and early stopping patience of 20 epochs.

4 Results

Table 2 presents the comprehensive comparison across all feature strategies, models, and datasets.

4.1 Analysis

VLM2Vec Features Outperform Traditional Baselines Across both datasets and models, VLM2Vec-derived features demonstrate advantages over traditional separately-trained encoders. The dual-tower strategy achieves the best overall performance, reaching Recall@20 of 0.0987 on Baby (FREEDOM) and 0.1124 on Sports (FREEDOM), surpassing the ResNet+BERT baseline (0.0986 and 0.1089, respectively). Even textual features alone match or exceed the baseline on most metrics, validating that VLM2Vec's joint training learns

1) https://github.com/enoche/MMRec

Table 2 Performance comparison of different features. Best results per dataset-model in **bold**.

Model	Features	R@20	N@20	P@20
Amazon Baby				
FREEDOM	Original	0.0986	0.0420	0.0054
	Visual	0.0762	0.0329	0.0043
	Textual	0.0974	0.0421	0.0054
	Prompt#1	0.0933	0.0405	0.0052
	Prompt#2	0.0938	0.0404	0.0052
	Dual	**0.0987**	**0.0429**	**0.0054**
LATTICE	Original	0.0844	0.0367	0.0047
	Visual	0.0833	0.0363	0.0047
	Textual	0.0875	0.0375	0.0049
	Prompt#1	0.0858	0.0374	0.0048
	Prompt#2	0.0854	0.0369	0.0048
	Dual	**0.0878**	**0.0377**	**0.0049**
Amazon Sports & Outdoors				
FREEDOM	Original	0.1089	0.0485	0.0060
	Visual	0.0840	0.0381	0.0047
	Textual	**0.1125**	0.0496	**0.0062**
	Prompt#1	0.1091	0.0477	0.0060
	Prompt#2	0.1054	0.0465	0.0059
	Dual	0.1124	**0.0501**	**0.0062**
LATTICE	Original	0.0955	0.0427	0.0053
	Visual	0.0943	0.0414	0.0053
	Textual	**0.0997**	**0.0437**	**0.0055**
	Prompt#1	0.0965	0.0433	0.0054
	Prompt#2	0.0969	0.0435	0.0054
	Dual	0.0989	**0.0438**	**0.0055**

more transferable representations than independently trained encoders.

Textual Features Dominate Visual Features A striking pattern emerges: textual features consistently and substantially outperform visual features across all configurations. On Sports with FREEDOM, textual features achieve Recall@20 of 0.1125 compared to visual's 0.0840— a 33.9% relative improvement. Similarly on Baby with FREEDOM, textual (0.0974) significantly outperforms visual (0.0762). This reveals that product recommendation relies heavily on semantic attributes (titles, categories) while visual appearance provides less discriminative signals. Visual features consistently rank as the weakest strategy, suggesting visual similarity does not strongly correlate with user preferences in e-commerce.

Textual Features Competitive with Dual-Tower Remarkably, textual features alone achieve performance nearly equivalent to dual-tower late fusion despite using half the feature dimensions (1536 vs. 3072). On Sports with FREEDOM, textual achieves 0.1125 Recall@20, actually surpassing dual-tower's 0.1124. On Baby with LATTICE, textual (0.0875) approaches dual-tower (0.0878) within 0.3%. This near-parity indicates that adding visual features through late fusion provides minimal marginal gains, making text-only features an attractive practical choice for efficiency.

Late Fusion Outperforms Early Fusion The dual-tower approach consistently outperforms early fusion strategies. On Baby with FREEDOM, dual-tower (0.0987) surpasses both Prompt#1 (0.0933) and Prompt#2 (0.0938). This suggests preserving modality-specific information allows the recommendation model to learn optimal fusion weights during training, while early fusion may compress cross-modal information prematurely. Interestingly, Prompt#1 matches or exceeds Prompt#2 across most configurations, indicating VLM2Vec's pretraining already captures retrieval-relevant representations without elaborate task framing.

Cross-Domain and Model Consistency These patterns hold consistently across both Baby and Sports domains despite different product characteristics. FREEDOM shows larger sensitivity to feature quality (33.9% gap between textual and visual on Sports) compared to LATTICE (5.7%), likely because FREEDOM's adaptive fusion amplifies modality quality differences while LATTICE's contrastive learning provides more robustness.

5 Conclusion

We present one of the first systematic evaluations of VLM-derived features for multimodal product recommendation. Through experiments on two Amazon datasets using FREEDOM and LATTICE models, we find: (1) VLM2Vec features outperform traditional ResNet+BERT baselines; (2) textual features dominate visual features by up to 33.9%; (3) textual features nearly match dual-tower performance with half the dimensions; (4) late fusion outperforms early fusion; and (5) simpler prompting works as well as elaborate task framing.

For practitioners, textual-only VLM2Vec features offer an excellent performance-efficiency tradeoff, while visual features provide limited value for e-commerce recommendation. Future work includes evaluating other VLMs (CLIP, BLIP), testing additional domains (fashion, electronics), and investigating why visual features underperform in recommendation contexts.

References

[1] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 452–461, 2009.

[2] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 639–648, 2020.

[3] Ruining He and Julian McAuley. Vbpr: Visual bayesian personalized ranking from implicit feedback. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 144–150, 2016.

[4] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*, pp. 1437–1445, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, 2019.

[7] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 2, pp. 423–443, 2019.

[8] Yinwei Wei, Xiang Wang, Liqiang Nie, Shaoyu Li, Dingxian Wang, and Tat-Seng Chua. Freedom: Graph-based multimodal feature fusion for recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2468–2478, 2023.

[9] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. Contrastive graph structure learning via information bottleneck for recommendation. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 20407–20420, 2022.

[10] 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. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

[11] Ziyan Jiang, Rui Meng, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[12] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, pp. 1074–1084, 2023.

[13] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, WWW '23, p. 845–854, 2023.

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, pp. 19730–19742, 2023.

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, Vol. 36, pp. 34892–34916, 2023.

[16] 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. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[17] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

[18] Xin Zhou. Mmrec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, pp. 1–2, 2023.