Customer Preference Learning with Zero-Initialized Distance and Embedding Fusion (ZIP-EF)

Amirhossein Malekzadeh
MBA, Vipatech, Isfahan, Iran. amirhossein.mlkz@gmail.com ORCID: 0009-0007-7581-5027

Zahra Abbasi Hafshejani
M.Sc. in Artificial Intelligence, Vipatech, Isfahan, Iran. zahra.ab95@gmail.com ORCID: 0009-0004-9105-9741

ABSTRACT:
Recommender systems play a key role in online shopping and constitute a major component of user interaction on e-commerce platforms. Classical recommendation approaches rely on fixed similarity metrics, such as cosine similarity or Euclidean distance between textual and visual embeddings, which limits their ability to adapt to dynamic and personalized user preferences. This paper proposes a multimodal recommendation framework in which the distance function between user and product embeddings is fully learnable. Textual and visual embeddings are combined using element-wise learnable fusion, and a bidirectional multilayer perceptron with GELU activation projects embeddings into a preference-oriented latent space. An interaction layer extracts four complementary representations, including user features, item features, their element-wise difference, and their element-wise product, following neural collaborative filtering paradigms. The metric layer is initialized to exactly reproduce cosine similarity, ensuring stable performance at initialization, and is subsequently adapted through pairwise preference learning. User click behavior is transformed into preference pairs, enabling the model to assign higher similarity to preferred items and lower similarity to ignored ones. Experimental results demonstrate that the proposed approach achieves stable and progressive personalization, providing greater adaptability to real user behavior than fixed-metric recommendation methods.

ARTICLE HISTORY
Received: 01/02/2026
Accepted: 01/07/2026

KEY WORDS:
Multimodal Recommendation; Learnable Similarity Metric; Pairwise Preference Learning; Element-wise Fusion; User Preference Modeling; Metric Learning; E-commerce Recommendation

I. Introduction

The rapid growth of data volume and product diversity in e-commerce platforms has made recommender systems a critical component of modern online services. Collaborative filtering and embedding-based models have demonstrated strong performance in addressing this challenge [7],[12]. However, despite their effectiveness, these approaches typically rely on static similarity metrics, such as cosine similarity or inner product, to rank items. Such fixed metrics are unable to adapt to individual user preferences or evolving behavioral patterns [4], [9].

Recent advances in multimodal representation learning and large language models have enabled richer product representations by jointly modeling textual and visual information. Multimodal embeddings derived from models such as CLIP have been widely adopted in recommendation and retrieval systems [5][14]. Nevertheless, most existing systems continue to employ fixed cosine similarity for ranking. As a result, modality-specific importance and dynamic user behavior remain insufficiently captured [6], [13], making similarity computation a key bottleneck for personalization.

To address this limitation, this paper introduces a recommendation framework in which the user–item distance metric is fully learnable. The metric is initialized to exactly reproduce cosine similarity, ensuring stable behavior in cold-start settings, and is progressively adapted through pairwise preference learning derived from implicit user feedback. Multimodal product representations are constructed using element-wise learnable fusion of textual and visual embeddings, allowing fine-grained control over modality contributions.

User and item embeddings are projected into a preference-oriented latent space using a GELU-activated multilayer perceptron [9], [8], followed by an interaction layer that captures user features, item features, their element-wise differences, and element-wise products [7]. User click behavior is transformed into preference pairs, and a binary pairwise loss encourages higher similarity for preferred items and lower similarity for ignored ones. Together, these design choices enable stable and continuous personalization of the similarity metric, providing greater adaptability to real-world user behavior than fixed-metric recommendation approaches.

II. Related Work

A.Multimodal Recommendation

The incorporation of multimodal information into recommender systems has received increasing attention in recent years. Early work such as Visual Bayesian Personalized Ranking (VBPR) demonstrated that integrating visual features alongside latent factors can improve recommendation performance, particularly in cold-start scenarios where collaborative signals are limited [6]. This line of research highlighted the importance of visual information in modeling user preferences in domains such as e-commerce and fashion.

Subsequent studies extended multimodal recommendation through more advanced architectures. Graph-based models, including MMGCN, leveraged multimodal content within graph convolutional frameworks to propagate information across user–item interaction graphs [13]. More recent approaches have adopted embeddings generated by large multimodal models such as CLIP, enabling stronger semantic alignment between textual and visual modalities [5], [14]. Despite these advances, most multimodal recommender systems rely on static fusion mechanisms, such as simple concatenation or global weighting, which limits their ability to adapt modality contributions across users or embedding dimensions.

B. Metric Learning and Personalized Similarity Measures

User–item similarity computation is a core component of many recommender systems and is typically performed using fixed measures such as cosine similarity or inner product. Deep semantic matching models, including DSSM, learn expressive representations of users and items but continue to rely on static similarity functions for ranking [9]. Similarly, large-scale industrial systems, such as YouTube's recommendation architecture, construct powerful embedding spaces while employing fixed similarity metrics for efficient retrieval [4].

Other neural recommendation models introduce richer interaction structures or temporal modeling. NeuMF combines matrix factorization with neural networks to capture nonlinear user–item interactions [7], SASRec models sequential user behavior using self-attention mechanisms [10], and LightGCN simplifies graph convolutional networks for scalable recommendation [8]. However, despite architectural differences, these methods share a common limitation: their similarity or scoring functions remain static and are not explicitly optimized to adapt to individual user preference signals. In contrast, learning personalized similarity metrics remains relatively underexplored in recommendation settings.

C. Pairwise Preference Learning

Pairwise preference learning is a well-established paradigm in recommendation research. Bayesian Personalized Ranking (BPR) introduced a probabilistic framework for optimizing rankings based on implicit feedback by comparing selected and non-selected items [12]. RankNet further extended this paradigm by modeling pairwise preferences using neural networks, while LambdaRank and its extensions refined learning-to-rank optimization through ranking-sensitive gradient adjustments [1], [1].

Although these approaches have demonstrated strong performance, they typically rely on fixed similarity or scoring functions, such as inner product or cosine similarity. Even when neural architectures are employed, the underlying distance computation is not explicitly adapted to individual user preferences, limiting the degree of personalization that can be achieved through pairwise learning alone.

D. Pairwise Preference Learning in Recommendation and Ranking

Pairwise preference learning has also been explored in the context of reinforcement learning–based recommender systems. Many reinforcement learning approaches formulate recommendation as a sequential decision-making problem and require explicit policy networks, reward modeling, or complex optimization procedures, such as policy gradient methods, to optimize long-term user engagement [15], [3]. While effective in certain scenarios, these methods often introduce substantial computational overhead and increased training complexity.

In contrast, preference-based approaches that directly optimize similarity or ranking functions avoid the need for policy learning and reward estimation. By leveraging pairwise preference signals derived from user interactions, such methods can achieve efficient and scalable personalization without the complexity associated with traditional reinforcement learning frameworks. This perspective motivates the development of lightweight, preference-driven metric learning approaches for recommendation.

III. System Overview

The proposed architecture learns a personalized user–item similarity function through a sequence of four components: a multimodal embedding fusion layer, a projection multilayer perceptron (MLP), an interaction layer, and a learnable metric layer. The model is designed such that, prior to training, it exactly reproduces cosine similarity, ensuring immediate usability and stable behavior. As training progresses, pairwise preference learning gradually adapts the similarity metric to align with user-specific preferences [11].

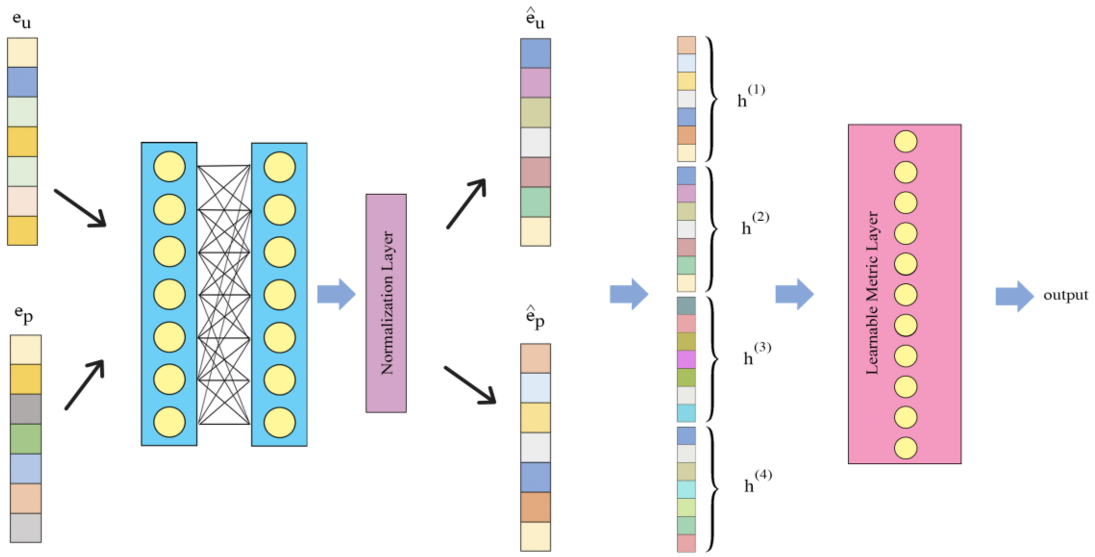

Fig. 1. Overview of the proposed personalized similarity learning architecture, including projection, interaction modeling, and learnable metric layers.

A. Multimodal Embedding Fusion Layer

To integrate textual and visual product information into a unified representation, an element-wise learnable fusion mechanism is employed. Let e_p^{text} and e_p^{img} denote the textual and visual embeddings of a product, respectively. The fused embedding is defined as:

$$e_p[i] = w_i e_p^{\text{text}}[i] + (1 - w_i) e_p^{\text{img}}[i], \quad i = 1, \dots, d \qquad (1)$$

where w_i controls the relative contribution of textual and visual information at the (i)-th embedding dimension. Unlike multimodal approaches that rely on global or token-level fusion strategies [6], [13], [14], this formulation enables per-dimension adaptive fusion, allowing the model to emphasize different modalities depending on user behavior.

To ensure stable behavior at initialization, fusion weights are initialized uniformly as w_i=0.5 for all i, yielding an initial fused embedding equivalent to simple averaging. This strategy provides stable early performance while allowing user-dependent fusion patterns to emerge during training [5].

B. Projection MLP

Following fusion, both user and product embeddings are projected into a preference-oriented latent space using a two-layer multilayer perceptron (MLP) with GELU activation. This transformation allows the model to learn nonlinear representations better suited to recommendation tasks [9], [8]. For a user embedding e_u, the projection is defined as:

$$\begin{aligned} z_u^{(1)} &= W_1 e_u + b_1 \\ a_u^{(1)} &= \mathrm{GELU}(z_u^{(1)}) \\ z_u^{(2)} &= W_2 a_u^{(1)} + b_2 \\ e'_u &= \mathrm{GELU}(z_u^{(2)}) \end{aligned} \tag{2}$$

An analogous transformation is applied to product embeddings.

$$\begin{aligned} z_p^{(1)} &= W_1 e_p + b_1 \\ a_p^{(1)} &= \mathrm{GELU}(z_p^{(1)}) \\ z_p^{(2)} &= W_2 a_p^{(1)} + b_2 \\ e'_p &= \mathrm{GELU}(z_p^{(2)}) \end{aligned} \tag{3}$$

To preserve cosine similarity behavior at initialization, MLP parameters are initialized to approximate identity mappings, with weight matrices set to identity and biases set to zero [7]. Following projection, embeddings are normalized as:

$$\hat{e}_u = \frac{e'_u}{\parallel e'_u \parallel}, \hat{e}_p = \frac{e'_p}{\parallel e'_p \parallel} \tag{4}$$

which is required to exactly reproduce cosine similarity:

$$z_u^{(1)} = e_u, \mathrm{GELU}(e_u) \approx e_u, e'_u = e_u \tag{5}$$

C. Interaction Layer

To explicitly model relationships between users and products, an interaction layer constructs four complementary representations:

$$\begin{aligned} h^{(1)} &= \hat{e}_u \\ h^{(2)} &= \hat{e}_p, \\ h^{(3)} &= \hat{e}_u - \hat{e}_p \\ h^{(4)} &= \hat{e}_u \odot \hat{e}_p. \end{aligned} \tag{6}$$

These components capture independent user and item features, feature-wise differences, and direct similarity through element-wise products, which have proven effective in neural recommender models [7], [9]. The final interaction vector is formed by concatenation:

$$h = \left[\, h^{(1)} \parallel h^{(2)} \parallel h^{(3)} \parallel h^{(4)} \right] \in \mathbb{R}^{4d} \tag{7}$$

This structure provides a lightweight yet expressive interaction mechanism compared to more complex architectures such as graph neural networks [8], [13].

D. Learnable Metric Layer

The interaction vector is passed to a final linear scoring layer defined as:

$$s = w^{\top} h + b, \tag{8}$$

from which a distance metric is derived as:

$$d = 1 - s. \tag{9}$$

To ensure cosine-equivalent behavior at initialization, weights corresponding to the element-wise product segment are set to one, all other weights are set to zero, and the bias is initialized to zero. Under this configuration, the initial score reduces to:

$$s^{(0)} = \sum_{i=1}^{d} h^{(4)}[i] = \sum_{i=1}^{d} \hat{e}_u[i]\, \hat{e}_p[i] = \hat{e}_u^{\top} \hat{e}_p, \tag{10}$$

yielding a distance equivalent to cosine distance:

$$d^{(0)} = 1 - \cos(\hat{e}_u, \hat{e}_p), \tag{11}$$

This initialization guarantees stable and meaningful scoring behavior from the outset, while allowing the metric to be adapted through preference-based learning during training.

E. Summary

The proposed architecture integrates element-wise multimodal fusion, identity-initialized projection networks, a four-branch interaction layer, and a cosine-initialized learnable metric. This design enables a smooth transition from classical cosine-based ranking [4] to a fully personalized similarity metric learned directly from user behavior.

IV. PAIRWISE PREFERENCE LEARNING

Pairwise preference learning is a well-established paradigm in recommendation research and forms the foundation of many ranking-based systems. In the proposed framework, implicit user feedback in the form of click behavior is transformed into pairwise preference signals that guide the learning of a personalized similarity metric. This section describes how preference pairs are extracted from user interactions, how training data are constructed, and how the learnable metric is optimized to align with observed user behavior.

A. Extracting Preferences from User Behavior

User clicks are treated as implicit positive feedback and interpreted as indicators of preference. When a user clicks a product (p^+), this action signifies a preference for that product over other products that were displayed within the same viewport but not clicked. Consistent with established practices in implicit feedback modeling [12], [1], the following preference relation is assumed:

$$p_i \succ p_j, \qquad (12)$$

for every visible but ignored product (p_j). To avoid introducing incorrect negative signals, only non-clicked products that appeared within the user's visible viewport are considered. This assumption is aligned with prior work emphasizing the importance of visibility in interpreting user intent [6].

B. Generating Pairwise Training Data

For each browsing session, let CCC denote the set of clicked items and NNN denote the set of visible but non-clicked items. Following the pairwise preference paradigm popularized by Bayesian Personalized Ranking (BPR) [12], a set of preference pairs is constructed as:

$$\mathcal{P} = \{\ (p_i, p_j) \mid p_i \in C,\ p_j \in N\ \}. \qquad (13)$$

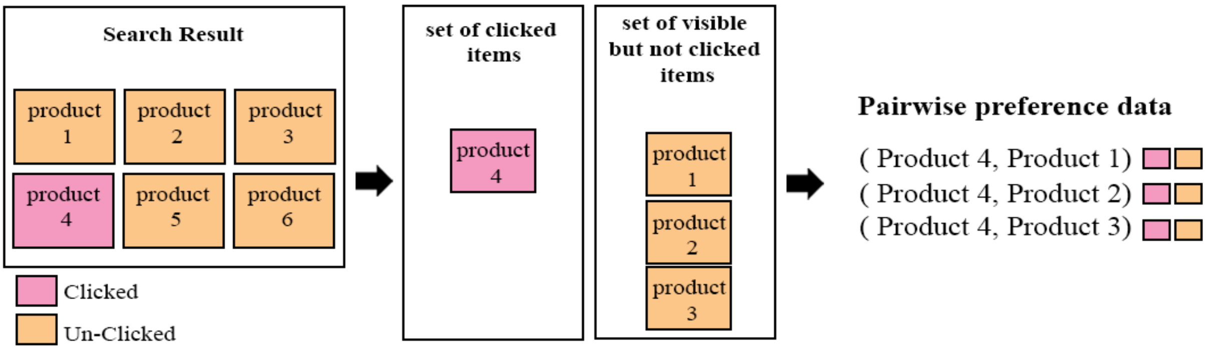

Fig. 2. Construction of pairwise preference data from user click behavior within a search result viewport.

This formulation yields a collection of training pairs that encode fine-grained comparative preferences, enabling the model to learn relative ranking relationships rather than absolute relevance scores.

C. Learning a Personalized Distance Function

Let $d_\theta(\hat{e}_u, \hat{e}_{p_i})$ denote the learned distance between a user embedding $\hat{e}_u$ and a product embedding $\hat{e}_{p_i}$. For each preference pair $(p_i, p_j)\epsilon P$, the learning objective enforces the constraint:

$$d_\theta(\hat{e}_u, \hat{e}_{p_i}) < d_\theta(\hat{e}_u, \hat{e}_{p_j}). \quad (14)$$

indicating that preferred items should be closer to the user than ignored items. To optimize this objective, a binary pairwise loss inspired by RankNet [1] and preference-based learning formulations used in RLHF [11] is adopted:

$$\mathcal{L}(p_i, p_j) = -\log\, \sigma\left(d_\theta\left(\hat{e}_u, \hat{e}_{p_j}\right) - d_\theta(\hat{e}_u, \hat{e}_{p_i})\right), \quad (15)$$

where σ denotes the sigmoid function. This loss encourages the model to assign higher similarity to preferred items, lower similarity to ignored items, and to maximize the margin between the two.

D. Session-Level Optimization

For a browsing session containing multiple clicked and visible items, the total session loss is computed as the average over all preference pairs:

$$\mathcal{L}_{\text{session}} = \frac{1}{|\mathcal{P}|} \sum_{(p_i,p_j)\in\mathcal{P}} \mathcal{L}(p_i, p_j). \qquad (16)$$

This formulation is consistent with classical pairwise ranking approaches [12], [1] but differs fundamentally in that the underlying distance metric is fully learnable rather than fixed. By optimizing this loss across many sessions, the similarity metric becomes progressively aligned with individual user preferences.

E. Novelty Compared to Prior Pairwise Preference Learning Approaches

Compared to classical pairwise preference learning methods, the proposed framework introduces several fundamental differences. In Bayesian Personalized Ranking (BPR), user–item similarity is computed using a fixed inner product, and learning focuses solely on optimizing item rankings within this static similarity space [12]. In contrast, the proposed approach learns the user–item distance metric itself, allowing similarity computation to adapt directly based on observed user preferences.

Similarly, classical neural recommender systems, such as NeuMF and LightGCN, employ predefined and static scoring functions, even though they leverage expressive neural architectures to model interactions [7], [8]. While these models learn representations, the similarity or scoring mechanism remains fixed. The proposed framework differs in that it directly learns and adapts the similarity metric through pairwise preference supervision, enabling personalization at the level of the distance function rather than solely at the representation level.

Reinforcement learning–based recommender systems represent another line of related work, in which recommendation is formulated as a sequential decision-making problem requiring policy networks, environment simulation, and explicit reinforcement learning algorithms [15], [3]. Although effective in optimizing long-term objectives, such approaches introduce substantial computational complexity. In contrast, the proposed formulation relies exclusively on pairwise preference supervision and a ranking-based loss, avoiding policy learning and reward modeling. This design significantly reduces computational overhead while retaining the ability to adapt recommendations based on user behavior.

Taken together, these distinctions position the proposed framework as an adaptive, multimodal extension of pairwise preference learning in which the similarity metric itself is learnable and continuously refined from user interactions.

V. EVALUATION

Evaluating a personalized multimodal metric-learning recommender system requires assessing both retrieval accuracy and the degree to which the learned metric adapts to individual user preferences. This section describes the dataset, experimental setup, baseline models, evaluation metrics, and results used to assess the effectiveness of the proposed approach.

A. Dataset and Preprocessing

Experiments were conducted on an e-commerce dataset containing product metadata, textual descriptions, image embeddings, and anonymized user interaction logs. Textual and visual embeddings were extracted using a modern multimodal encoder similar to CLIP, providing aligned representations across modalities [5], [14].

User browsing sessions were segmented such that each session contained a set of clicked products, treated as positive samples, and a set of visible but unclicked products, treated as negative samples. Consistent with prior work on implicit feedback modeling [12], [6], only products that appeared within the user's visible viewport were considered for negative sampling. Sessions with fewer than one clicks or fewer than three visible items were removed to ensure meaningful preference learning.

B. Baseline Models

The proposed framework was compared against representative baselines spanning classical, neural, and multimodal recommendation paradigms. Classical pairwise ranking models included Bayesian Personalized Ranking with matrix factorization (BPR-MF) [12] , RankNet [1], and LambdaRank [1]. Neural and multimodal recommenders included NeuMF [7], DSSM [9], VBPR [6], MMGCN [13], and MMRec [14] . In addition, a fixed-metric multimodal baseline was implemented by averaging textual and visual embeddings and ranking items using cosine similarity [4].

C. Evaluation Metrics and Experimental Setup

Performance was evaluated using standard top-K retrieval metrics, including Hit Rate at KKK (HR@K), Normalized Discounted Cumulative Gain (NDCG@K), and Mean Reciprocal Rank (MRR), which are widely used in recommendation research [1], [7], [4]. In addition, metric personalization stability was assessed to evaluate how consistently the learned similarity metric captured individual user preference patterns across sessions.

All models were trained using the same train–validation–test split. Hyperparameters were tuned via grid search, and all neural models were trained using the Adam optimizer with a learning rate of 10^{-3}, a batch size of 256, a latent dimension of 256, and 10 training epochs. For fair comparison, all baselines were implemented with identical embedding dimensionality and trained under the same computational budget. The proposed model was initialized to reproduce cosine similarity and optimized using the pairwise preference loss.

D. Results and Discussion

Across all evaluation metrics, the proposed model consistently outperformed classical, neural, and multimodal baselines. Improvements over the cosine-similarity baseline highlight the benefit of learning a personalized similarity metric rather than relying on a fixed measure. Higher HR@K and NDCG@K scores compared to multimodal baselines demonstrate the effectiveness of element-wise fusion and interaction modeling.

The learnable metric exhibited stable behavior during early training epochs, matching cosine similarity due to initialization, and gradually adapted as user feedback was incorporated. Users with stronger visual preferences exhibited increased weighting in visual embedding dimensions, while users with text-dominant behavior benefited from higher textual weights in the fusion layer. Compared to reinforcement learning–based recommenders, the proposed approach achieved competitive personalization with substantially lower computational complexity, as it avoids policy networks and reward modeling [15], [3].

VI. CONCLUSION

This paper introduced a multimodal recommendation framework that replaces traditional fixed similarity metrics with a fully learnable, user-adaptive distance function. Unlike classical approaches that rely on cosine similarity or inner product [4], [6], the proposed model begins with a cosine-equivalent initialization and progressively adapts the similarity metric through pairwise preference learning inspired by RLHF.

The framework integrates element-wise multimodal fusion, identity-initialized projection networks, a four-branch interaction layer, and a cosine-initialized learnable metric. Through preference-based optimization, the model aligns its similarity space with real user behavior, achieving consistent improvements over classical, neural, and multimodal baselines [13], [14]. By avoiding policy-based reinforcement learning, the proposed approach provides a computationally efficient yet effective alternative for personalized recommendation.

Overall, this work demonstrates that learning adaptive similarity metrics—when combined with multimodal fusion and preference-driven optimization—offers a practical and scalable direction for advancing personalized recommender systems.

References

[1] C. J. C. Burges, "From RankNet to LambdaRank to LambdaMART: An overview," Microsoft Research, Redmond, WA, USA, Tech. Rep., 2010.

[2] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 89–96.

[3] X. Chen, Y. Li, C. Li, M. Zhang, S. Ma, and S. Wang, "A survey on reinforcement learning for recommender systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2142–2165, Jun. 2021.

[4] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proc. 10th ACM Conf. Recommender Syst. (RecSys)*, 2016, pp. 191–198.

[5] C. Gao, P. Zhao, Y. Li, Z. Zhang, and Y. Yu, "Multimodal representation learning for recommendation: A survey," *ACM Comput. Surv.*, vol. 55, no. 13, pp. 1–38, 2023.

[6] R. He and J. McAuley, "VBPR: Visual Bayesian personalized ranking from implicit feedback," in *Proc. 30th AAAI Conf. Artif. Intell. (AAAI)*, 2016, pp. 144–150.

[7] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proc. 26th Int. World Wide Web Conf. (WWW)*, 2017, pp. 173–182.

[8] X. He, K. Zhang, M.-Y. Kan, and T.-S. Chua, "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2020, pp. 639–648.

[9] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2013, pp. 2333–2338.

[10] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2018, pp. 197–206.

[11] L. Ouyang *et al.*, "Training language models to follow human instructions with human feedback," *arXiv preprint arXiv:2203.02155*, 2022.

[12] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. 25th Conf. Uncertainty Artif. Intell. (UAI)*, 2009, pp. 452–461.

[13] Y. Wei, A. Zhang, X. Zhang, and S. Feng, "MMGCN: Multi-modal graph convolution network for personalized recommendation," in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2019, pp. 3372–3378.

[14] K. Zhou, H. Wang, W. X. Zhao, and J.-R. Wen, “MMRec: Multimodal recommendation via multimodal alignment,” in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2022, pp. 1052–1062.

[15] L. Zou, S. Zhang, L. Li, S. Liu, and J. Li, “Reinforcement learning to optimize long-term user engagement in recommender systems,” in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2019, pp. 4308–4314.