Information Fusion 128 (2026) 103991

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full Length Article

Multimodal recommender systems: A survey of representation, modeling, and optimization

Lin Pan, Zhiqiang Pan *, Fei Cai, Honghui Chen

National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha, 410000, China

ARTICLE INFO

Keywords:
Deep learning
Information systems
Multimedia recommendation
Multimodal recommender systems

ABSTRACT

With the surge of online information, recommender systems have become indispensable for mitigating information overload and delivering personalized content. Traditional collaborative filtering methods, which depend heavily on user interaction logs, often struggle with sparse feedback, cold-start users, and long-tail items. To address these limitations, recent studies integrate multimodal content such as text, images, and audio, which provide complementary semantic signals. Multimodal recommender systems (MRSs) leverage such heterogeneous data to construct more expressive item representations and generate highly personalized recommendations. However, modality heterogeneity poses fundamental challenges in representation learning and fusion. In light of these issues, this survey provides a comprehensive overview of MRSs, structured around four main aspects. First, it reviews modality-specific representation techniques and fusion mechanisms, comparing encoders for textual, visual, and acoustic data, and contrasting static and dynamic fusion strategies. Second, it categorizes existing methods by modeling paradigms rather than isolated components. The resulting taxonomies cover graph-based and transformer-based models, as well as emerging self-supervised, causality-aware, and large language model (LLM)-enhanced methods, with research outlooks for each paradigm. Third, it analyzes model optimization techniques that shape performance and robustness. Finally, it outlines promising research directions toward efficient, scalable, and reliable MRSs. By synthesizing current progress and highlighting emerging trends, this survey aims to clarify the research landscape and provide forward-looking insights for future methodological innovations. All related materials are available at https://github.com/ppan-lin/MRSs-Survey.

1. Introduction

The accelerated expansion of online information makes recommender systems (RSs) indispensable for mitigating information overload and ensuring personalized content access. As these systems are increasingly deployed across e-commerce [1,2], social media [3], and content platforms [4], a central challenge lies in accurately capturing both user interests and item semantics. Traditional RSs mainly rely on user-item interaction logs, including clicks, ratings, and browsing histories, to infer preferences through collaborative filtering [5,6]. These interaction-driven methods perform well when sufficient behavioral data exist, yet they often degrade in situations of sparse feedback, user cold start, or long-tail item exposure. In such scenarios, the absence of semantic signals prevents the system from fully capturing the underlying user intent and the complex characteristics of items [7,8].

To overcome these limitations, recent studies increasingly leverage multimodal content, including product images, textual descriptions, user reviews, and background music. These modalities provide richer semantic information that more accurately reflects user perception and response. As illustrated in Fig. 1, on short-video platforms, a viewer's decision to continue watching may be influenced by a concise caption (textual), an attractive opening frame (visual), and an engaging soundtrack (acoustic). By combining these textual, visual, and acoustic signals with behavioral traces, multimodal information provides a closer approximation of user cognition. However, integrating multimodal information is not straightforward. Heterogeneity across modalities leads to challenges in feature representation and cross-modal alignment, while incomplete or noisy signals would reduce reliability. Addressing these challenges requires advances in modality-specific encoders, effective fusion architectures, and robust learning paradigms, complemented by optimization strategies that balance accuracy, scalability, and interpretability.

Within this context, multimodal recommender systems (MRSs) have become a rapidly advancing research frontier. A systematic and timely survey is therefore necessary to synthesize recent methodologies and identify evolving trends. To facilitate comparison and highlight underexplored issues, existing surveys can be grouped into three

* Corresponding author.
E-mail addresses: panlin@nudt.edu.cn (L. Pan), panzhiqiang@nudt.edu.cn (Z. Pan), caifei08@nudt.edu.cn (F. Cai), chenhonghui@nudt.edu.cn (H. Chen).

https://doi.org/10.1016/j.inffus.2025.103991
Received 1 October 2025; Received in revised form 16 November 2025; Accepted 19 November 2025
Available online 23 November 2025

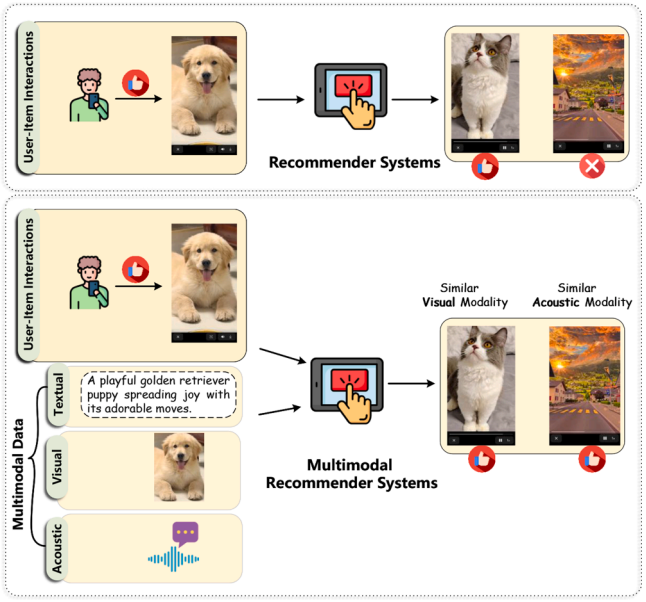


Fig. 1. Illustration of conventional and multimodal recommender systems in short-video recommendation. Conventional systems rely on user-item interactions, while multimodal systems additionally utilize textual, visual, and acoustic data.

categories. (1) Methodology-specific surveys [9–14] provide in-depth analyses of individual paradigms. However, they remain narrow in scope and often overlook multimodal heterogeneity and cross-modal interactions. (2) General recommender system surveys [15–17] cover broader algorithmic families and evaluation practices, yet they typically regard content modalities as auxiliary signals and do not systematically consider modality-specific encoders, fusion mechanisms, or participation patterns. (3) Explicit multimodal surveys [7,18,19] adopt a structural perspective, systematically examining modality types and integration strategies. In contrast to these works, this survey provides three distinctive advantages:

- **Comprehensive and Up-to-date Coverage:** This survey encompasses studies from premier conferences and journals, as well as influential arXiv preprints. As illustrated in Fig. 2, the distribution of references across venues and publication years demonstrates both the comprehensiveness and timeliness of the coverage. Notably, about 60 % of the cited works have been published within the past three years, ensuring that this survey reflects the latest advancements up to 2025.
- **Organization by Modeling Paradigms:** Rather than adopting a structural taxonomy, this survey is organized around modeling paradigms. This paradigm-oriented view clarifies the technological trajectory of MRSs, from graph-based models to large language models (LLMs), and helps readers match paradigms with application scenarios. For instance, self-supervised paradigms are suitable for scenarios with sparse interaction data, while causality-aware paradigms are preferable when the goal is to enhance interpretability and robustness.
- **Tracking Emerging Trends:** This survey follows the evolving research directions in the development of MRSs. For instance, it categorizes multimodal fusion techniques into static and dynamic strategies, emphasizing a growing trend toward treating fusion as a context-aware and adaptive process rather than confining it to early or late stages. Moreover, it highlights the emerging area of LLM-enhanced multimodal recommendation, which has received limited attention in prior surveys despite its increasing importance in recent literature.

To guide readers through the subsequent discussion, this survey is structured as follows:

- Section 1 - Introduction:
  This section reviews the development of RSs and highlights the importance of multimodal information in improving recommendation performance. It also distinguishes this survey from existing studies and outlines its overall structure and key contributions.
- Section 2 - Preliminaries:
  This section introduces the basic problem settings of MRSs, covering interaction-based, sequential-based, and session-based formulations.
- Section 3 - Modality Representation and Fusion:
  This section discusses modality types (i.e., textual, visual, acoustic) and their corresponding encoders, followed by static and dynamic modality fusion strategies.
- Section 4 - Modeling Paradigms:
  This section presents established and emerging modeling paradigms, including graph-based, transformer-based, self-supervised, causality-aware, and LLM-enhanced models. For each paradigm, a focused research outlook is included to discuss paradigm-specific challenges and promising avenues.
- Section 5 - Model Optimization:
  This section introduces recent model optimization techniques, including the loss function, gradient refinement, staged training, and modality balancing.
- Section 6 - Future Directions:
  This section highlights potential research directions for MRSs, emphasizing resource-efficient learning, temporal-causal reasoning, emerging modality integration, continual adaptation, and privacy preservation.
- Section 7 - Conclusion:
  This section provides a summary of the key content of this survey.

Fig. 3 provides a comprehensive structural overview of this survey, with particular focus on Sections 2 to 6.

The primary contributions of this survey are as follows:

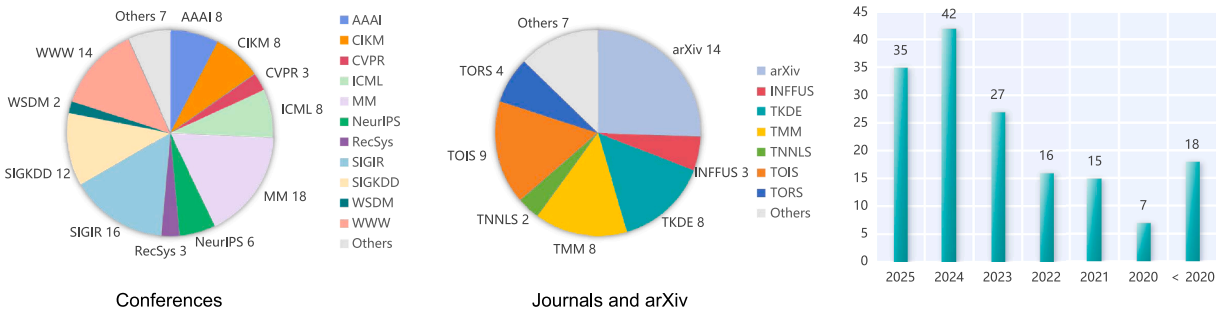
- We present a comprehensive investigation of MRSs, summarizing a wide range of works to trace the field's evolution and provide a foundation for further research.
- We examine modality representation and fusion techniques, detailing how different modalities are encoded and combined, and categorize modeling paradigms ranging from conventional graph-based methods to recent LLM-enhanced approaches.
- We analyze model optimization strategies for MRSs from multiple perspectives, clarifying their mechanisms and providing insights for the design of more effective training frameworks.
- We identify open challenges and emerging opportunities, outlining future research directions that can guide promising developments and foster innovative solutions.

2. Preliminaries

To provide a comprehensive understanding of MRSs, it is essential to first formalize the foundational problem settings. This section introduces three major categories: interaction-based, sequential, and session recommendation. Each represents a distinct user-item interaction scenario and serves as the basis for incorporating multimodal information.

2.1. Interaction-based multimodal recommender systems

Let $\mathcal{U} = \{u\}$ and $\mathcal{I} = \{i\}$ denote the sets of users and items, with cardinalities $|\mathcal{U}|$ and $|\mathcal{I}|$, respectively. The user-item interaction matrix is defined as $\mathcal{R} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $r_{ui} = 1$ indicates that user u has interacted with item i (e.g., via views, clicks), and $r_{ui} = 0$ otherwise. Each item i is associated with a set of multimodal features $\mathcal{M}_i$ (e.g., text,



(a) Proportional distribution of literature sources

(b) Temporal distribution of literature publications

Fig. 2. Reference distribution analysis of surveyed literature.

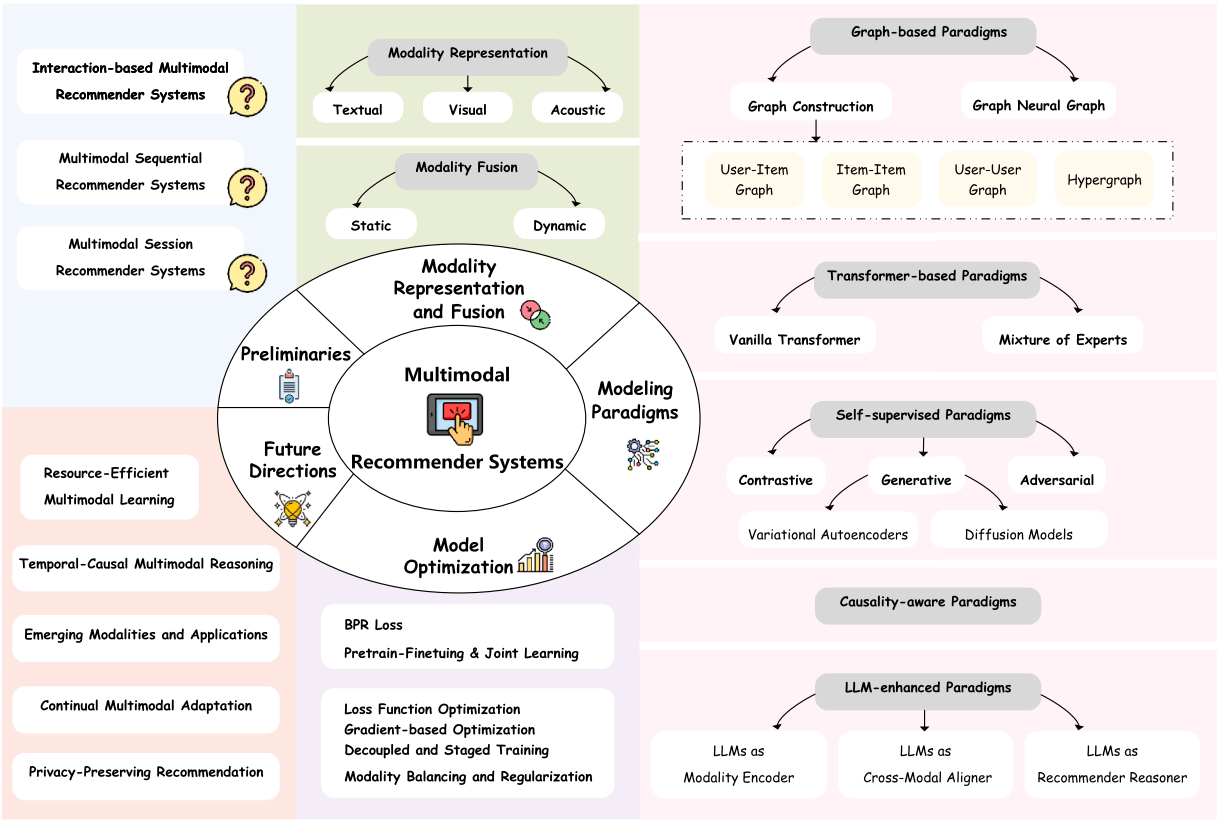


Fig. 3. Structural framework of this survey (Sections 2–6).

image). Interaction-based multimodal recommender systems, denoted as $\mathcal{F}_\theta^{int}$, aim to estimate the probability that user u will interact with a candidate item i by integrating item-level multimodal features $\mathcal{M}$ into the collaborative filtering framework. Unlike traditional collaborative filtering, which relies solely on the interaction matrix $\mathcal{R}$, these models incorporate rich content signals to enhance recommendation quality. The model parameters θ are optimized by minimizing a loss function $\mathcal{L}$ (e.g., Bayesian personalized ranking loss [20]) over the training set $\mathcal{D}_{train}$:

$$\theta^* = \arg\min_\theta \sum_{(u,i)\in\mathcal{D}_{train}} \mathcal{L}(\mathcal{F}_\theta^{int}(u, M_i); r_{ui}). \tag{1}$$

2.2. Multimodal sequential recommender systems

Let $S = \{s_u \mid u \in \mathcal{U}\}$ be the set of user interaction sequences, where $s_u = [i_1, i_2, \ldots, i_n]$ records the chronological interactions of user u. Each interacted item i_j in the sequence is associated with a set of multimodal features $\mathcal{M}_{i_j}$. Multimodal sequential recommender systems, denoted as $\mathcal{F}_\theta^{seq}$, aim to estimate the probability that a user u will interact with a future item i by modeling the temporal dependencies within s_u. In contrast to interaction-based methods, which ignore interaction order, sequential models leverage the temporal structure to capture evolving preferences. The model parameters θ are optimized by minimizing a loss function $\mathcal{L}$ over the training set $\mathcal{D}_{train}$:

$$\theta^* = \arg\min_\theta \sum_{(u,i)\in\mathcal{D}_{train}} \mathcal{L}(\mathcal{F}_\theta^{seq}(s_u, M_i); r_{ui}). \tag{2}$$

2.3. Multimodal session recommender systems

Session-based recommendation is a special case of sequential modeling in anonymous settings. Let $S = \{S\}$ be the set of sessions, where

Table 1
Key abbreviations and their full forms in this survey.

Abbreviation	Full form
RSs	Recommender Systems
MRSs	Multimodal Recommender Systems
LLMs	Large Language Models
MLLMs	Multimodal Large Language Models
ViTs	Vision Transformers
GNNs	Graph Neural Networks
GCNs	Graph Convolutional Networks
MoE	Mixture of Experts
SSL	Self-Supervised Learning
VAEs	Variational Autoencoders
DMs	Diffusion Models
DAG	Directed Acyclic Graph
BPR	Bayesian Personalized Ranking

each session $S = [i_1, i_2, \ldots, i_m]$ records a short sequence of item interactions. Each item i_j is associated with a set of multimodal features $\mathcal{M}_{i_j}$. Multimodal session recommender systems, denoted as $\mathcal{F}_\theta^{ses}$, are designed for anonymous environments where long-term user histories are unavailable. These models rely solely on the short-term contextual behavior observed within the current session to predict the next likely item i. The model parameters θ are optimized by minimizing a loss function $\mathcal{L}$ (e.g., cross-entropy loss [21]) over the training set D_{train}:

$$\theta^* = \arg\min_\theta \sum_{(S,i)\in D_{train}} \mathcal{L}(\mathcal{F}_\theta^{ses}(S, M_i); r_{Si}). \tag{3}$$

To facilitate understanding of the terminology used throughout this survey, Table 1 summarizes key abbreviations and their corresponding full forms.

3. Modality representation and fusion

Modality representation and fusion jointly determine the quality of user and item embeddings in MRSs. Representation learning encodes heterogeneous modalities while preserving their semantic characteristics, and fusion integrates these representations to exploit cross-modal complementarities. Limitations in either stage can distort preference modeling and degrade recommendation accuracy. The following sections review representative encoders for textual, visual, and then discuss major fusion strategies that connect these representations to downstream tasks.

3.1. Modality representation

In MRSs, accurate modality representations are a prerequisite for effective downstream modeling. Each modality exhibits distinct structures and information densities, requiring specialized encoding strategies. Textual data conveys symbolic and sequential semantics, visual data provides perceptual and spatial cues, and acoustic data encodes temporal and spectral dynamics. In practical applications, these modalities do not always co-occur but often appear in different combinations. For instance, e-commerce platforms typically combine textual and visual modalities, while short-video platforms usually integrate textual, visual, and acoustic modalities. A thorough understanding of these modality-specific properties is therefore fundamental to the design of effective encoders. The following subsections review representative encoding methods for textual, visual, and acoustic modalities.

3.1.1. Textual modality

Text is one of the most prevalent modalities in recommendation scenarios, including product descriptions, titles, or user reviews. Due to its structured semantics and inherent interpretability, it serves as a fundamental source for modeling user preferences and understanding content. Early studies adopt static embeddings such as Word2Vec [22] and Sentence2Vec [23] to encode textual semantics, but these methods lack contextual awareness. Recurrent neural networks (RNNs) [24] are later introduced to model sequential structures. However, they face limitations in representing long or complex texts. The emergence of pretrained Transformer-based models, such as BERT [25] and Sentence-Transformer [26], brings richer contextual encoding capabilities, making them suitable for modeling textual features in personalized recommendations. More recently, large language models (LLMs), including T5 [27], ChatGPT [28], and LLaMA [29], have been integrated into MRSs through prompt tuning or lightweight adaptation. These models provide advanced semantic understanding and flexible reasoning capabilities.

3.1.2. Visual modality

Visual modality refers to item images such as product photos, cover images, or thumbnails. These images convey intuitive semantics and are particularly important in domains such as fashion and e-commerce. Traditional approaches rely on convolutional neural networks (CNNs), including VGG [30], GoogLeNet [31], and ResNet [32], to derive semantic representations from images. These networks focus on local patterns and hierarchical abstractions, and their outputs are often combined with other modalities to enhance item representations. Recent studies increasingly adopt vision transformers (ViTs) [33]. By replacing convolution with self-attention, ViTs are able to capture long-range visual dependencies and provide stronger capacity for modeling high-level semantics. Beyond single-modality encoders, vision-language models such as CLIP [34] and BLIP [35] extend the representation space of visual data. Trained on large-scale image-text pairs, these models generate image embeddings that are semantically aligned with textual information, providing stronger generalization and cross-modal transferability.

3.1.3. Acoustic modality

Acoustic modality covers content such as music, speech, and environmental sounds, offering temporally continuous and perceptually rich signals that complement static modalities. Various pretrained models have been introduced to enhance acoustic representation learning. AST [36] leverages transformer architectures to model global spectro-temporal patterns, facilitating long-range acoustic dependency modeling. BEATs [37] adopts masked spectrogram modeling with acoustic tokenizers to enable self-supervised learning of contextual acoustic representations. In addition, diffusion-based models such as Make-An-Audio [38], AudioLDM [39], and AudioSR [40] improve acoustic representations through high-fidelity conditional generation in latent space. These models provide multiple strategies to improve acoustic understanding in recommendation scenarios.

3.2. Modality fusion

Modality fusion effectively combines complementary information from different modalities to construct enriched user and item representations. Prior surveys generally distinguish between **early fusion** and **late fusion**. Early fusion combines low-level modality features at an initial stage, enabling direct feature interactions but often losing modality-specific details and treating each item as a uniform whole. Late fusion, in contrast, merges modality-specific representations at a higher stage, preserving modality individuality but neglecting fine-grained cross-modal dependencies. To overcome the inherent limitations of such fixed-stage designs, recent research increasingly explores **hybrid fusion**, which treats fusion as a context-aware and adaptive process rather than being strictly confined to early or late stages. To better reflect this trend, this survey reclassifies modality fusion methods into **static fusion** and **dynamic fusion**, depending on whether the fusion mechanism remains fixed or adapts to content, user preferences, or task context. Table 2 summarizes the categorization of fusion strategies in MRSs.

3.2.1. Static fusion

Static fusion represents multimodal integration strategies that combine modality features following predetermined rules, without adapt-

Table 2
Multimodal Fusion Categorization in MRSs.

Fusion type	Early fusion	Late fusion	Hybrid fusion
Static fusion	VIP5 [47], LLaRA [48], Molar [49], AB-Rec [8], MLLM-MSR [50], DIMO [61], LPIC [62], BM3 [63], SGFD [66]	GRCN [41], CI2MG [42], MCDRec [43], Modality Debiasing [44], HIRES [45], TMLP [46], DMRL [51], SLMRec [52], MCLN [53], GUME [54], AlignRec [55], MGCE [56], POWERec [57], MIG-GT [58], EVEN [59], FREEDOM [64], LightGT [65], SMORE [67]	ODMT [60]
Dynamic fusion	LATTICE [68], FastMMRec [75]	DGVAE [69], LGMRec [70], DiffMM [71], CKD [72], PromptMM [73], RedNnD [74], FMMRec [76], MoDiCF [77], DCAR-DM [78], DiffCL [79], FGCM [80], MARN [81], DualGNN [82], SOIL [83], UGT [84], MENTOR [85], DOGE [86], PEARL [87], IISAN [88], NoteLLM-2 [89], MICRO [90], MGCL [91], MMIL [92], DA-MRSs [93], SCL [94], ModalSync [95], M3SRec [96], MP4SR [97], HM4SR [98], GMMF [99], MISSRec [100], MGCN [101], MMSBR [104], MILK [105]	MMSR [102], EgoGCN [103], TMFUN [106], COHESION [107]

ing to specific users, items, or contexts. Typical techniques include embedding-level concatenation [41–46], prompt-based concatenation [8,47–50], element-wise summation [51–59], averaging [60–62], projection [63–66], and max-pooling [67]. These methods are simple and computationally efficient, as they avoid learning context-dependent fusion parameters. Their limitation lies in the assumption that modality contributions remain constant, which may lead to suboptimal representations when modalities are noisy, missing, or unevenly informative. Static fusion is thus suited for scenarios with stable and balanced modality quality.

3.2.2. Dynamic fusion

Dynamic fusion emphasizes adaptive mechanisms that assign varying importance to each modality according to the input context. Common techniques include adaptive weighting [68–80], attention-based weighting [81–87], gating networks [88,89], contrastive learning [90–95], and mixture-of-experts models [96–98]. In addition to these conventional techniques, recent studies introduce more innovative fusion strategies. For instance, GMMF [99], MISSRec [100] and MGCN [101] weight different modalities based on individual user preferences. MMSR [102] models user history as a graph of modality nodes and dynamically fuses intra- and inter-modality signals via dual-attention and an update gate. EgoGCN [103] performs edge-wise fusion by modulating each node's features based on messages from other modalities of its neighboring nodes. This allows important inter-modal information to propagate while preserving modality-specific intra-graph processing. MMSBR [104] employs a hierarchical pivot transformer to dynamically integrate visual and textual features layer-by-layer, using trainable pivot tokens to capture complementary semantics from both modalities. MILK [105] samples modality weights from a Dirichlet distribution and cyclically shifts them across environments, ensuring diverse and adaptive modality contributions. TMFUN [106] implements an attention-guided fusion mechanism that adaptively weights modalities in user-item interactions and refines item representations through a multi-step fusion process. COHESION [107] employs a dual-stage fusion strategy, where early fusion dynamically refines modalities using behavior data, and late fusion adaptively aggregates representations through learned attention weights. By dynamically adjusting the fusion process, these methods can better capture nuanced cross-modal dependencies and mitigate issues arising from noisy or missing modalities.

4. Modeling paradigms

Modeling paradigms in MRSs provide the backbone for integrating heterogeneous signals into unified user and item representations. Each paradigm imposes specific inductive biases for organizing, propagating, and optimizing multimodal data. **Graph-based** paradigms emphasize explicit relational structures among users, items, and modality-specific features, enabling message passing across high-order connections. **Transformer-based** paradigms focus on sequential dependencies and modality interactions through attention mechanisms. **Self-supervised** paradigms leverage auxiliary pretext tasks to capture latent structures within multimodal data, while **causality-aware** paradigms aim to disentangle true causal effects from spurious correlations. More recently, **LLM-enhanced** paradigms introduce LLMs as modality encoders, semantic aligners, or reasoning engines, thereby equipping MRSs with powerful cross-modal understanding and generative capabilities. These paradigms provide complementary perspectives that shape the representational capacity, robustness, and interpretability of MRSs.

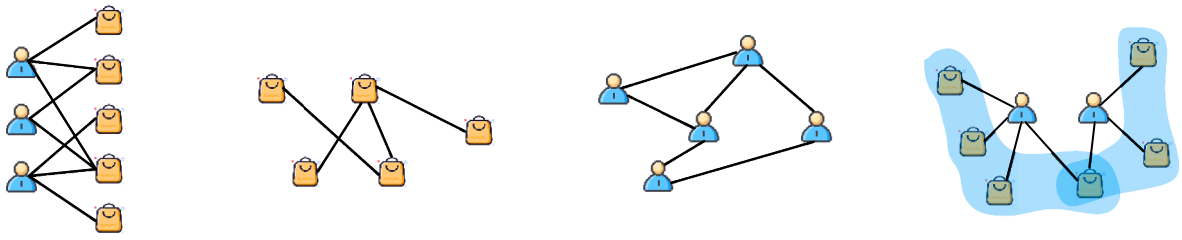
4.1. Graph-based paradigms

Graph-based modeling in MRSs represents users, items, and modality-specific signals as nodes connected by behavioral and semantic relations. This structure allows the model to propagate information across both interaction edges (e.g., user-item clicks) and modality edges (e.g., textual coherence or visual similarity). By iteratively aggregating neighborhood features, the graph captures not only direct user-item interactions but also high-order dependencies that reveal latent affinities, such as shared visual styles or overlapping semantic attributes. The resulting embeddings integrate behavioral patterns with multimodal content in a unified space, providing structured representations that enhance recommendation accuracy and improve cross-modal alignment in downstream tasks.

4.1.1. Graph construction

Graph construction in MRSs organizes users and items to explicitly capture relationships. Depending on node and relation types, these graphs can be homogeneous, connecting entities of the same type, or heterogeneous, encoding multiple entity and relation types. Beyond pairwise connections, higher-order structures such as hypergraphs can model multi-entity interactions and capture group-level behaviors. The choice of graph structure significantly influences how information propagates and how dependencies are learned. Fig. 4 presents the example illustrating these graph types.

Heterogeneous user-item graphs. User-item graphs are bipartite structures where users and items form two disjoint node sets, with edges representing observed interactions such as clicks, ratings, or purchases. These graphs capture collaborative filtering signals from behavioral data [101]. Their representational capacity can be enhanced in multiple ways, such as integrating second-order interest signals [83] or multimodal content [58,77,85,107]. Another enhancement strategy is denoising, which removes spurious or noisy connections. For instance, GRCN [41] employs a soft pruning strategy to adaptively down-weight unreliable user-item edges. FREEDOM [64] applies degree-sensitive edge pruning to refine the bipartite structure, while MCDRec [43] injects diffusion-aware knowledge into the interaction graph for more accurate denoising. PEARL [87] conducts modality-aware hard pruning by removing user-item edges with low modality-specific attention.



(a) User-Item Graph (b) Item-Item Graph (c) User-User Graph (d) Hypergraph

Fig. 4. Illustration of graph structures in MRSs.

Homogeneous item-item graphs. Item-item graphs represent items as nodes, with edges encoding similarity or relatedness. Edge weights are typically derived from co-occurrence statistics [59,61,86,93] or from modality-aware semantic consistency [59,62,67,68,80,85–87,93]. A widely used construction strategy is K-nearest-neighbors (KNN), which retains only the most similar neighbors for each item. Two main design paradigms are adopted. The first constructs modality-specific item-item graphs and treats them as multiple views for downstream modeling [85,91,93,108], enabling the exploitation of complementary modality signals. The second integrates multiple modality signals into a single unified graph [59,67,68,75], facilitating direct modeling of holistic item relationships and reducing the complexity of multi-graph processing. To improve graph quality, recent studies apply denoising or pruning strategies. DA-MRSs [93] removes edges with low similarity or cross-modal inconsistency, while retaining item pairs that frequently co-occur in user interactions. TMLP [46] employs a topological pruning scheme to eliminate spurious connections and refine structural connectivity.

Homogeneous user-user graphs. User-user graphs are homogeneous structures where each node represents a user, and edges encode behavioral similarity derived from interaction patterns. They are often constructed through user co-occurrence [82,87,107], linking two users when they have engaged with the same items. Edge weights can be assigned based on the count or frequency of shared interactions, or on the similarity between the items involved. These graphs capture implicit connections among users who may not have direct ties yet exhibit comparable consumption behaviors.

Hypergraphs. A hypergraph allows a single hyperedge to connect multiple nodes, enabling the modeling of higher-order relationships. In the context of MRSs, hyperedges can capture diverse group-level interactions, such as all items interacted with by a single user in DOGE [86], all items within the same category in HIRES [45], modality-similar items and their engaged users in LGMRec [70], modality-specific user or item clusters with shared preferences in HyperCTR [109], and items grouped by multimodal clustering (e.g., visually similar images or semantically related descriptions) in CI2MG [42]. This flexibility makes hypergraphs well-suited for modeling heterogeneous and interdependent structures in multimodal contexts.

In practice, many MRSs integrate two or more of the above graph types to capture complementary relational patterns, especially when modeling multimodal signals. The resulting multi-graph structures are often processed using graph neural networks, as discussed in the following section.

4.1.2. Graph neural networks

Graph neural networks (GNNs) are extensively adopted in MRSs to capture high-order connectivity from multimodal correlations and user-item interactions. Among them, graph convolutional networks (GCNs) [110] are most prevalent, as they provide a systematic framework for propagating and aggregating information across sparse and intricate interaction graphs. Formally, consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of nodes (e.g., users or items) and $\mathcal{E}$ denotes the set of edges representing observed interactions or inferred relations. Let $A \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ be the adjacency matrix of $\mathcal{G}$, and $D \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ be its diagonal degree matrix, where $D_{ii} = \sum_j A_{ij}$. The normalized adjacency matrix $D^{-1/2}AD^{-1/2}$ ensures scale-invariant message passing and mitigates the effect of nodes with disproportionately high degrees. The layer-wise propagation rule of a standard GCN is given by:

$$H^{l+1} = \sigma(D^{-1/2}AD^{-1/2}H^lW^l), \tag{4}$$

where $H^l \in \mathbb{R}^{|\mathcal{V}|\times d}$ is the node embedding matrix at layer l (H^0 is the initial feature matrix, which may integrate multimodal features), W^l is the trainable weight matrix, and $\sigma(\cdot)$ denotes a non-linear activation function (e.g., ReLU). Through iterative propagation, GCNs aggregate information from progressively larger neighborhoods, thereby capturing both direct and indirect multimodal relationships crucial for recommendation tasks [58,68,91,101,107].

Beyond standard GCNs, several studies enhance their design in MRSs by targeting personalization, modality heterogeneity, and training efficiency. For instance, RedNnD [74] mitigates over-smoothing and preserves personalization by averaging neighbors' multi-layer representations and aligning them with the ego node's initial features during aggregation. MIG-GT [58] addresses modality heterogeneity by introducing modality-independent receptive fields, employing separate GCNs with distinct receptive fields for each modality to better capture their specific characteristics. FastMMRec [75] identifies that GCN-based training can generate harmful training pairs and exacerbate modality isolation. It resolves this by using GCNs solely during inference, improving efficiency and cross-modality integration.

While GCN-based approaches improve graph modeling flexibility and expressiveness, their deployment in large-scale MRSs is limited by the high computational cost of deep layers, feature transformations, and non-linear activations. To improve scalability, many recent studies [43,46,53,55,56,63,67,70,84,86,93] adopt LightGCN [111], which removes feature transformations and activation functions, retaining only the neighborhood aggregation mechanism. The layer-wise propagation in LightGCN is defined as:

$$e_u^{l+1} = \sum_{i\in\mathcal{N}(u)} \frac{1}{\sqrt{|\mathcal{N}(u)||\mathcal{N}(i)|}} e_i^l, \tag{5}$$

where $\mathcal{N}(u)$ is the neighbor set of node u, $|\mathcal{N}(u)|$ is its degree, and $e_u^l \in \mathbb{R}^d$ is the embedding at layer l. The symmetric normalization term balances nodes of varying degrees, reducing computation while preserving semantic propagation over large graphs.

Research outlook. Graph-based paradigm is shifting toward dynamic and adaptive graphs that continuously update with user behavior, contextual shifts, and modality availability. Such adaptability requires models not only to capture current relationships but also to anticipate emerging interaction patterns. Promising directions include temporal GNNs for fine-grained temporal dependencies, self-evolving edge semantics to

reflect changing cross-modal relevance, and scalable learning strategies (e.g., incremental updates, sampling, distributed training) for handling large evolving graphs.

4.2. Transformer-based paradigms

Transformer architectures [112] are a core framework in MRSs for modeling complex dependencies in user sequences. By employing multi-head attention, Transformer effectively integrates heterogeneous modal features and captures long-range interactions across sequences [86,113,114]. Given an input sequence $X = [x_1, x_2, \ldots, x_n]$, each Transformer layer consists of the following key operations. First, the multi-head attention mechanism computes several attention distributions in parallel, allowing the model to capture dependencies across multiple representation subspaces:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O, \tag{6}$$

where each attention head is defined as:

$$head_i = softmax(\frac{QW_i^Q(KW_i^K)^\top}{\sqrt{d}})VW_i^V, \tag{7}$$

where h is the number of heads, and W_i^Q, W_i^K, W_i^V, W^O represent trainable projection matrices. The attention output is further refined through residual connections, layer normalization, and feed-forward layers to stabilize training and enhance representation capacity. Together, these components form the standard Transformer encoder block, which is widely used in recommendation models to learn sequential and multimodal dependencies.

4.2.1. Vanilla transformer

Early Transformer-based models, such as SASRec [115] and BERT4Rec [116], focus on long-range dependencies and bidirectional contextual modeling. Building on these foundations, later studies integrate multimodal information into Transformer frameworks. M3SRec [96] and MP4SR [97] embed visual and textual features directly into the attention mechanism, producing joint representations that blend multiple modalities. PMGT [117] utilizes a pre-trained graph Transformer to construct a unified multimodal representation of project relationships and their associated side information. ODMT [60] combines ID embeddings with multimodal features within a unified Transformer. It projects these combined embeddings into a recommendation-oriented latent space, balancing collaborative filtering signals with modality-aware information. HM4SR [98] separates Transformer layers by modality, generating distinct embedding sequences that strengthen semantic modeling of user behaviors at a modality-specific level. MTSTRec [118] aligns modalities in time through shared tokens, enabling mid-level fusion and fine-grained, stepwise interactions along the sequence. MISSRec [100] employs a Transformer-based context encoder to process modality-specific token sequences, from which informative cues are adaptively selected. Its interest-aware decoder captures item-modality-interest relationships, filtering out redundant interactions while emphasizing the most relevant user preferences. MMSBR [104] integrates a pivot-based mechanism into the Transformer as a hierarchical mixer, extracting and combining signals from heterogeneous modalities. UGT [84] adopts a multi-way Transformer where shared self-attention learns joint contextual features, and modality-specific experts capture complementary patterns from raw multimodal inputs. Some methods also focus on simplifying the Transformer structure. For instance, LightGT [65] employs a light self-attention block by removing multi-head, feed-forward networks, and residual connections to efficiently distill informative multimodal features. MIG-GT [58] reduces full attention cost through a sampling-based strategy, where each node attends only to a subset of sampled vertices.

Research outlook. In multimodal sequential recommendation, vanilla Transformers use multi-head self-attention with a single relative-position encoding, assuming uniform temporal patterns. This limits their ability to handle modality-specific delays, sparsity, and asynchronous signals, causing amplification of outdated or irrelevant information. Allowing each attention head to learn its own temporal kernel or differentiable time shift enables specialization across short-, mid-, and long-term contexts, mitigating cross-modal misalignment and reducing outdated information propagation.

4.2.2. Mixture of experts

Mixture of experts (MoE) is a modular paradigm that allocates model capacity across a collection of specialist sub-networks (i.e., experts), together with a learnable gating mechanism that conditionally routes or weights expert outputs [119]. Within Transformer architectures, MoE modules are typically integrated into or replace feed-forward sublayers, as illustrated in Fig. 5. This integration retains the attention mechanism for contextual representation while introducing conditional and sparsely activated computation. For each input, only a limited number of experts are activated, which balances expressive capacity with manageable inference cost. Experts can specialize by modality, context, or temporal factors. This design further supports hierarchical, cross-modal, and stochastic routing strategies, which improve semantic fusion and robustness in multimodal sequential recommendation. For instance, M3SRec [96] constructs modality-specific MoE modules, where each expert is a feed-forward network specialized for capturing user intents in different modalities. Cross-modal MoE experts then integrate these outputs to achieve deep semantic fusion for sequential recommendation. MP4SR [97] applies MoE in both text and image encoders, with each expert implemented as a linear transformation followed by dropout and layer normalization. A softmax-based gating network adaptively combines their outputs to enhance multimodal sequence representation. VMoSE [120] employs a MoE framework where each expert corresponds to a modality-specific latent variable. Stochastic sampling integrates these experts to adaptively fuse modalities while filtering noisy signals. HM4SR [98] adopts a two-level MoE architecture. The interactive MoE processes all item modalities to extract user-interest features, while the temporal MoE leverages experts conditioned on explicit temporal embeddings to capture dynamic user interests.

Research outlook. MoE mechanisms are attractive for conditional capacity scaling, yet they also bring practical challenges. A central issue is routing instability, as gating networks often concentrate traffic on a few experts while leaving others under-utilized, which causes gradient starvation and unstable training. To address this issue, researchers can enhance gating with confidence estimates and embed latency or compute budget constraints into the routing objective.

4.3. Self-supervised paradigms

Self-supervised learning (SSL) in MRSs enables models to learn from unlabeled inputs by exploiting inherent patterns within users, items, and their multimodal content. By integrating textual, visual, and other features with interaction data, SSL improves representation quality even under sparse feedback. Approaches to SSL in MRSs can be grouped into three main paradigms: contrastive-based, generative-based, and adversarial-based methods, as illustrated in Fig. 6. Among these, contrastive-based methods dominate current research, while adversarial-based methods remain less explored. The following sections introduce each paradigm along with representative strategies.

4.3.1. Contrastive-based

Contrastive learning, initially proposed for vision [121] and language [122] representation learning, trains embeddings by pulling positive pairs closer while pushing negative pairs apart. In MRSs, it facilitates the learning of user-item embeddings that are both modality-consistent and semantically discriminative. A typical framework adopts

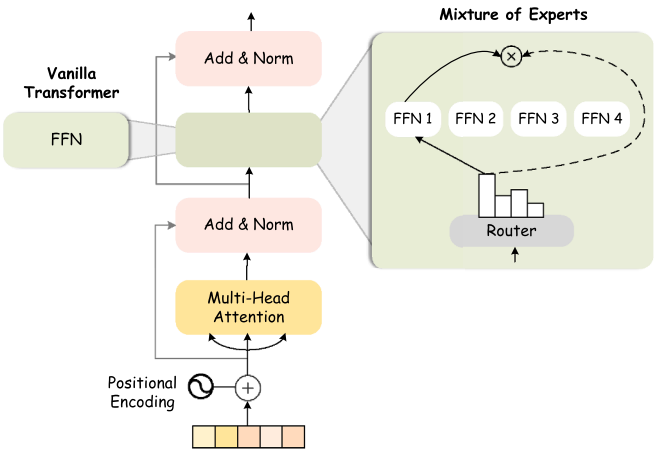


Fig. 5. Architecture of vanilla Transformer block and its mixture-of-experts extension in the feed-forward network (FFN).

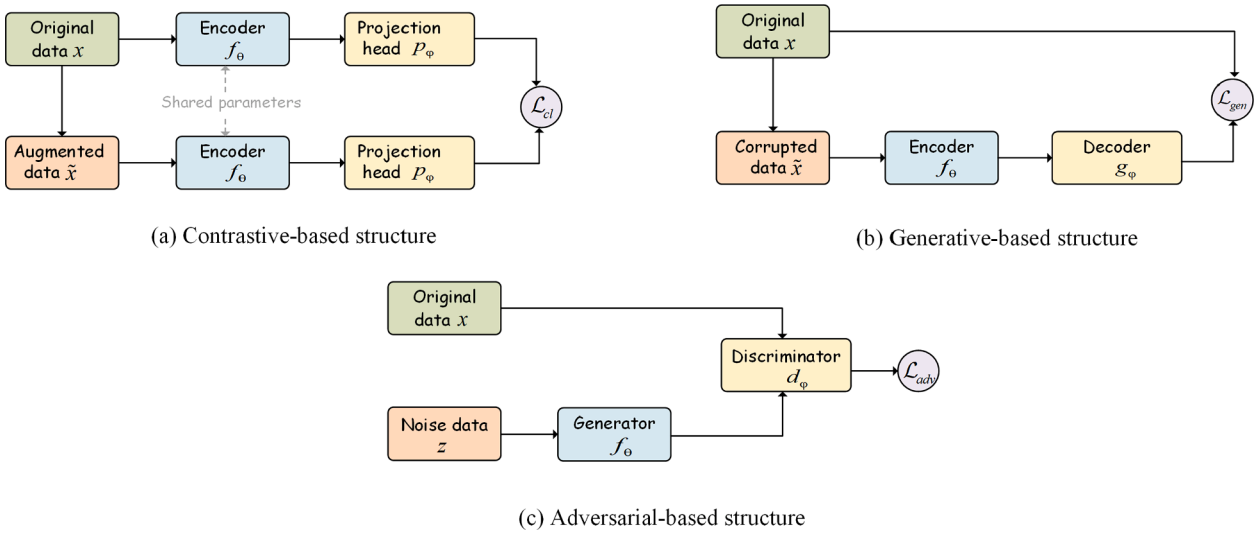


Fig. 6. Illustration of three self-supervised learning paradigms.

an encoder-projection head architecture as shown in Fig. 6(a). Given original multimodal input x and an augmented view $\tilde{x}$, the encoder f_θ extracts modality-specific or cross-modal features, which are then mapped by a projection head p_ϕ into a shared contrastive space. The overall training objective is formulated as:

$$\theta^*, \phi^* = \arg\min_{\theta,\phi} \mathcal{L}_{cl}(p_\phi(f_\theta(x)), p_\phi(f_\theta(\tilde{x}))), \tag{8}$$

where $\mathcal{L}_{\text{cl}}$ denotes a contrastive loss function, typically instantiated as the InfoNCE loss [123].

In MRSs, the quality of learned representations is highly sensitive to the choice of data augmentations. These augmentations create alternative views of users and items that serve as input for contrastive training. Depending on their scope, augmentation strategies fall into two categories, modal-agnostic and modal-specific augmentations. Modal-agnostic augmentations perturb features or interactions without considering modality. Examples include feature dropout, which randomly removes elements within a modality [52]; feature masking, which stochastically masks an entire modality to form complementary subsets [52,124]; edge dropout, which perturbs the interaction graph by removing modality edges [124]; and noise injection, which mitigates overfitting and encourages broader semantic generalization [54,59,80,85]. Several models further refine these strategies with tailored augmentation modules. BM3 [63] employs dropout-based augmentation to bootstrap user and item embeddings and reconstructs the user-item graph. EVEN [98] introduces a sequence-level augmentation that replaces items with temporal placeholders to create time-aware alternatives for contrastive learning. DiffCL [79] leverages both forward and inverse processes of diffusion models to generate semantically consistent contrastive views, effectively mitigating the impact of noisy information during SSL. Modal-specific augmentations are designed to exploit the unique structure of multimodal inputs. For instance, SLMRec [52] constructs fine-grained and coarse-grained feature spaces corresponding to different modalities, and performs alignment between modality pairs to improve fusion consistency. MMGCL [124] introduces modal perturbation, generating hard negatives by perturbing one modality in otherwise aligned pairs. This approach encourages the model to capture cross-modal inconsistencies and enhance modality correlation learning. MMSBR [104] proposes pseudo-modality contrastive learning, where one modality generates pseudo-information in another via data augmentation. The actual-pseudo modality pair is treated as a positive pair, effectively mitigating the impact of noise in visual or textual content. These augmentation strategies collectively enhance view diversity and help contrastive-based models generalize beyond fixed modality configurations.

In addition to constructing auxiliary views, contrastive learning is widely applied in modality alignment and fusion to facilitate the learning of modality-aware user preferences. Specifically, numerous methods adopt cross-modal contrastive objectives to align representations

across different modalities [71,84,92,95,97], or between behavioral signals and modal content [62,67,87,91,106], thereby enforcing consistency among heterogeneous embeddings. Furthermore, LGMRec [70] extends this paradigm by constructing a global hypergraph contrastive framework that captures high-order semantic relations among modalities. To enable finer-grained alignment, DA-MRSs [93] introduces a graded contrastive mechanism that positions multimodally similar items closer than unimodally similar ones, and unimodally similar items closer than dissimilar ones in the representation space. SCL [94] leverages semantic-aware contrastive learning to pre-train multimodal representations by aligning user query-purchase pairs, capturing more fine-grained semantic similarities. Contrastive learning is also leveraged to enhance the alignment between unimodal and fused representations. MICRO [90] and LPIC [62] maximize the agreement between representations from individual modalities and their fused multimodal counterparts. MGCN [101] and SOIL [83] further align user behavioral signals with multimodal fused features to preserve semantic consistency. AlignRec [55] bridges collaborative and content signals by contrasting ID-based user/item embeddings with their multimodal representations. NoteLLM-2 [89] introduces dual-level optimization by separately contrasting visual and multimodal embeddings with their paired notes, preventing the model from over-relying on textual similarity. CI2MG [42] and M3Rec [96] contrast intra-modal and inter-modal embeddings to enhance the consistency and discrimination across different levels of fusion.

Research outlook. Current contrastive learning research in MRSs focuses on constructing stronger positive-negative pairs and enforcing cross-modal alignment. While these strategies can improve performance, they rely on heuristics and may yield uneven gains across modalities and tasks. Effective augmented views in multimodal settings should preserve semantic consistency and collaborative signals, achievable via learnable augmentors and explicit metrics. Reconciling semantic inconsistencies is also critical, as strict alignment can propagate noise. More adaptive objectives, including modality-specific temperatures, uncertainty-aware weighting, and loss functions that down-weight unreliable contrasts, offer a way to maintain the advantages of contrastive learning while reducing semantic conflict.

4.3.2. Generative-based

Generative-based methods in MRSs model the complex joint distribution of multimodal data and extract latent representations by reconstructing or generating data across modalities. In contrast to contrastive-based approaches, which focus on instance-level comparisons, generative methods aim to capture the intrinsic data manifold and inter-modal relationships through probabilistic modeling or iterative denoising. The general structure of generative-based methods is illustrated in Fig. 6(b). Formally, these methods learn the encoder and decoder (or generator) parameters θ, ϕ by reconstructing or denoising multimodal inputs. The learning objective minimizes a loss that quantifies the discrepancy between the original input x and the output generated from its corrupted or augmented variant $\tilde{x}$:

$$\theta^*, \phi^* = \arg\min_{\theta,\phi} \mathcal{L}_{gen}(x, g_\phi(f_\theta(\tilde{x}))), \tag{9}$$

where f_θ and g_ϕ denote the encoder and decoder networks, respectively. $\mathcal{L}_{gen}$ represents a suitable generative loss function, such as reconstruction error (e.g., mean squared error, cross-entropy), variational lower bound objectives, or denoising score matching loss.

Variational autoencoders. Variational autoencoders (VAEs) constitute a foundational generative modeling approach in MRSs. They learn a probabilistic latent space by optimizing the evidence lower bound (ELBO), which simultaneously maximizes reconstruction accuracy and regularizes the latent distribution through the Kullback-Leibler divergence. This process ensures a well-structured latent space and reduces overfitting to noisy inputs. Several extensions leverage this approach to better handle multimodal data. MVGAE [125] applies VAEs to model multiple modalities within the user-item bipartite graph, fusing modality-specific latent variables via a product-of-experts strategy. CMVAE [126] extends the VAE framework by learning a shared latent space through cross-modal reconstruction, where the latent representation of one modality reconstructs the original features of another, and vice versa. VMoSE [120] further adapts the standard VAE for multimodal recommendation by introducing multiple modality-specific encoders that capture uncertainty through learned variances. It replaces conventional latent fusion with a stochastic sampling mechanism that adaptively selects and aggregates modalities based on their precision. DGVAE [69], a graph-based VAE, reconstructs both user-item interactions and user-word preferences by learning disentangled latent variables aligned with semantic prototypes. This alignment enhances interpretability by maximizing mutual information across modalities.

Diffusion models. Diffusion models (DMs) represent another emerging generative modeling paradigm for recommendation, inspired by their success in fields such as image synthesis [127] and text-to-image generation [128]. In recommendation scenarios, these models aim to reconstruct recommendation-relevant variables from noise through a learned denoising process, conditioned on multimodal content and user behavior. DMs are first explored in unimodal recommendation. For instance, CODIGEM [129] generates robust latent representations by modeling non-linear collaborative patterns, while DiffRec [130] retains globally similar yet personalized signals through a denoising diffusion process. These methods highlight the generative potential of diffusion and pave the way for multimodal extensions. MCDRec [43] integrates a modality-conditioned diffusion process to generate enhanced item representations and denoise user-item graphs, effectively bridging multimodal content and collaborative signals. DiffMM [71] utilizes a modality-aware DM to iteratively denoise corrupted user-item interaction graphs, enabling the generation of multimodal collaborative signals for recommendation. MoDiCF [77] designs a modality-aware DM to reconstruct missing modality features from complex distributions and enhance fairness via counterfactual debiasing in incomplete multimodal recommendation. DCAR-DM [78] employs a DM to construct balanced user-item interaction graphs across modalities, enabling consistent and user-preference-aware multimodal alignment. DiffCL [79] employs a DM to generate high-quality contrastive views for graph-based multimodal learning, effectively mitigating noise and enhancing semantic consistency beyond traditional augmentation methods. CCDRec [131] leverages the reverse process of DMs to align multimodal features with collaborative signals and enable adaptive, curriculum-aware negative sampling for recommendation. PMG [132] integrates multimodal preference signals with target features into DMs, allowing controllable generation that aligns with both user-specific multimodal preferences and target semantics.

Research outlook. A critical limitation of generative paradigms in MRSs is the inefficiency of both training and inference. VAEs are prone to posterior collapse, where latent variables degenerate into uninformative states and thus fail to capture multimodal semantics. DMs face an even more pronounced challenge, as the iterative denoising process typically requires hundreds of steps, making them impractical for real-time recommendation. Overcoming these issues requires advances in model design and optimization. For VAEs, alternative formulations such as flow-based or energy-based latent variable models can preserve representational capacity while ensuring tractable optimization. For DMs, accelerated sampling techniques, distillation of the denoising process, and modality-conditioned lightweight architectures can substantially reduce inference latency.

4.3.3. Adversarial-based

Adversarial methods in MRSs are based on generative adversarial networks (GANs) [133], establishing a competitive training mechanism between a generator and a discriminator to enhance representation

learning. As illustrated in Fig. 6(c), the generator f_θ takes noise data z as input and produces synthetic features. The discriminator d_ϕ receives both real multimodal data x and generated features $f_\theta(z)$, aiming to distinguish authentic from synthetic inputs. This adversarial process forces the generator to align its outputs with the real data distribution, capturing modality-invariant and semantically informative representations. Formally, the training objective is formulated as the following min-max optimization problem:

$$\theta^*, \phi^* = \arg\min_{\theta} \max_{\phi} \mathcal{L}_{adv}(d_\phi(x, f_\theta(z))), \tag{10}$$

where the discriminator maximizes its discriminative capability, and the generator minimizes the same objective by producing increasingly indistinguishable representations. Through this optimization, the learned features become robust, generalizable, and better aligned with multimodal data distributions.

The adversarial training mechanism has been employed in several recent MRSs. For instance, AMR [134] improves robustness by perturbing high-level visual features rather than raw images. It formulates a min-max objective between the recommendation model and a feature-space adversary, enhancing generalization under sparse implicit feedback. MARN [81] adopts a double-discriminator framework. The first discriminator identifies shared features across modalities, while the second adversarially aligns these features, facilitating cross-modal knowledge transfer and robust modality-invariant representations. GMMF [99] utilizes a cross-modal GAN to generate visual and textual features from ID embeddings, enabling explicit comparison with real modalities. This facilitates the removal of redundant information and the extraction of complementary cues for enhanced multimodal fusion.

Research outlook. Adversarial-based structures remain relatively underexplored in MRSs. Existing studies primarily emphasize distributional alignment across modalities; yet, the coupling between adversarial objectives and the recommendation goal is often weak. A natural step forward is to design task-aware adversarial training schemes that combine recommendation loss with adversarial loss, ensuring that feature alignment contributes directly to user-item matching. Another direction is to develop discriminators that not only differentiate real from synthetic features but also integrate recommendation relevance signals, thereby enhancing the link between representation robustness and recommendation accuracy.

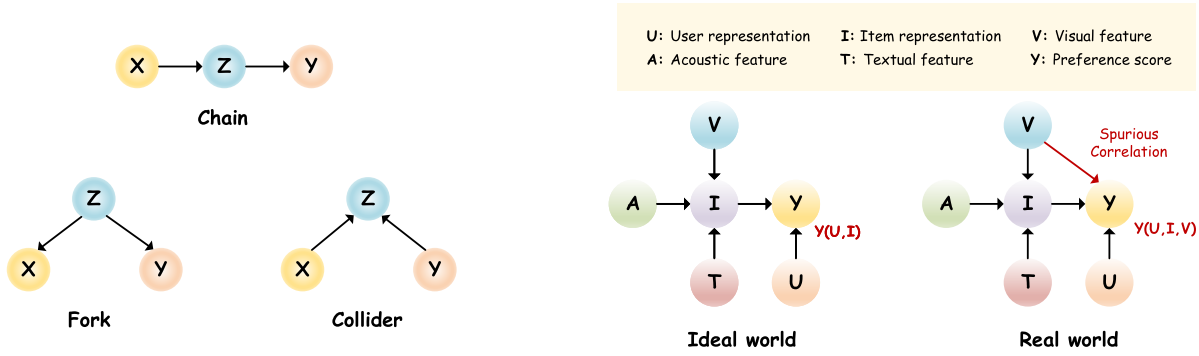
4.4. Causality-aware paradigms

Understanding user behavior is challenging due to the presence of spurious correlations in historical interaction data. In fact, correlation does not imply causation [12]. In an e-commerce example, many users may purchase both a coffee machine and running shoes, not because one causes the other, but because both appear in the same holiday promotion. Such co-occurrences reflect shared external factors rather than direct causal links. To address this, two main frameworks are commonly employed: the structural causal model (SCM) and the Rubin causal model (RCM). SCM encodes causal relations through causal graphs, translates them into structural equations, and applies the do-operator to eliminate confounding effects. RCM focuses on the relationship between a treatment and an outcome, while considering other variables as confounders. It infers causality by comparing predicted counterfactuals with observed outcomes. Despite these differences, both frameworks intend to estimate counterfactual outcomes, representing results that could have occurred under unobserved treatments. In MRSs, most causality-aware approaches adopt SCM to construct causal graphs that capture the underlying generative process of user behavior.

A causal graph is typically represented as a directed acyclic graph (DAG), where nodes correspond to variables and edges indicate direct causal influences. Three fundamental structures commonly appear in DAGs, as shown in Fig. 7(a): the chain, where one variable affects a second, which then influences a third; the fork, where a single variable affects two downstream variables; and the collider, where two variables converge to affect a common outcome. These structures exemplify different patterns of causal dependency and confounding. In MRSs, a basic causal graph often models relationships among user preferences, multimodal item features, and observed interactions, providing a foundation for designing causality-aware models. Fig. 7(b) presents an example of causal structures in ideal and observed recommendation scenarios. In the ideal setting, visual (V), textual (T), and acoustic (A) features affect the item representation (I). The preference score (Y) is then determined by I together with the user representation (U). In this case, multimodal features influence preferences only through I. In the observed world, a direct link from V to Y indicates a spurious effect. Visual features may influence preferences independently of I, due to presentation biases or other unobserved factors. This indicates that models may conflate true causal effects with incidental correlations. While textual and acoustic features remain mediated by I, the direct effect of V must be explicitly adjusted to capture true multimodal influences.

To address the above challenges, causal inference has emerged as a promising paradigm in MRSs, providing a more interpretable and causally grounded understanding of user behavior. By disentangling true causal relationships from spurious correlations, causal methods improve robustness against common data issues. One prominent issue is popularity bias [135–138], which arises when historical user-item interactions exhibit a long-tail distribution, leading models to over-recommend popular items regardless of individual user interests. In addition, exposure bias [139–141] stems from the limitation that users can interact only with items to which they are exposed. As a result, the collected feedback reflects prior exposure mechanisms rather than unbiased user preferences. Another common issue is missing data [142,143], where the absence of interaction does not necessarily indicate disinterest, but may result from unexposed or unrecorded user intentions. Furthermore, recommendation logs often contain noise signals [144]. Interactions in these logs can be influenced by randomness or external factors, and thus may not reflect genuine user preferences. To support such reasoning, the causal graph is typically constructed to model the effects between variables, capturing how changes in one component causally influence another. Based on this structure, techniques like intervention [145] and counterfactual inference [146,147] are employed to obtain unbiased estimations and facilitate causal reasoning.

There exist numerous models built upon causal theory in MRSs. For instance, CausalRec [148] identifies visual bias in visually-aware recommendation as a spurious effect, where visual features act as mediators in user-item interactions. It estimates the total indirect effect to eliminate the direct causal influence of visual attention. This formulation provides a foundation for addressing modality bias in multimodal settings, where one modality may spuriously dominate due to imbalanced or misaligned distributions. Building upon this insight, subsequent works explicitly mitigate such modality bias. EliMRec [149] and Modality Debiasing [44] introduce a counterfactual debiasing framework that explicitly models the causal pathway of modality bias, and subtracts its effects to enhance item-side fairness. MoDiCF [77] treats visibility-induced modality bias as a causal effect and encodes it in an explicit causal graph. It adopts a modality-aware diffusion process to impute missing modalities, and performs counterfactual interventions to subtract visibility's causal impact on learned user and item representations. Furthermore, MCLN [53] employs causal difference to compare preference distributions on multimodal content between interacted and uninteracted items. This method helps identify and reduce preference-irrelevant signals within the interacted items. In contrast to these debiasing-oriented approaches, MGCE [56] further disentangles the causal factors of user interactions by modeling multimodal interest and multimodal conformity separately through a causal graph. This design enables a more fine-grained representation of multimodal user preferences. CKD [72] demonstrates that causal inference can also guide optimization by estimating each modality's causal contribution and adaptively reweighting distillation losses. FMMRec [76] does not explicitly



(a) Three causal structures in directed acyclic graphs

(b) Ideal vs. real-world causal graphs in MRSs

Fig. 7. Illustration of causal structures in MRSs, including three representative causal patterns (left: chain, fork, and collider) and the contrast between ideal and real-world causal graphs (right), where spurious correlations may arise in practice.

estimate causal effects; instead, it adopts a causality-inspired approach by disentangling modality representations and adjusting user representations through relation-aware modeling. The goal is to mitigate the causal impact of sensitive attributes and approximate counterfactual fairness.

Research outlook. Current causality-aware MRSs primarily address bias correction and the disentanglement of modality effects. Reliable causal discovery and counterfactual estimation, however, remain challenging due to embedding distortions and time-varying causal relations. Future work can leverage embedding-aware structure learning, which integrates raw features with latent embeddings via Bayesian posterior edge probabilities, to improve causal structure inference. Further disentanglement of causal factors may be achieved by combining multi-environment invariance with encoder regularization. In parallel, dynamic causality modeling that couples multimodal encoders with temporal point processes or dynamic causal graphs offers a way to track preference evolution and enhance interpretability.

4.5. LLM-enhanced paradigms

Large language models (LLMs) demonstrate advanced capabilities in semantic comprehension, contextual reasoning, and sequential modeling, making them highly suitable for integration into recommendation tasks. Initial LLM-enhanced recommendation methods primarily utilize textual signals, such as item descriptions or user reviews, to improve recommendation accuracy and interpretability [150–152]. Techniques like in-context learning [153], prompt tuning [154], and carefully designed instruction prompts [155] are generally adopted to enable the models to exploit linguistic knowledge while aligning with domain-specific objectives. A systematic comparison of these three adaptation techniques in MRSs is provided in Table 3. These efforts highlight the potential of LLMs to capture complex user-item interactions through natural language representations. Following these developments, subsequent research extends the application of LLMs into multimodal contexts, giving rise to multimodal large language models (MLLMs). By integrating heterogeneous data modalities, MLLMs learn unified representations that improve alignment across modalities and enable a more comprehensive understanding of item content and user preferences. As illustrated in Fig. 8, the functional roles of LLMs within MRSs can be categorized as follows.

LLMs as modality encoder. When deployed as pretrained encoders, LLMs transform raw modality-specific inputs into semantic-rich embeddings. These representations can be directly exploited by downstream recommendation modules for alignment, fusion, or prediction. Compared to conventional encoders trained from scratch or under narrow supervision, LLM-based encoders benefit from large-scale pre-training, providing superior generalization and broader semantic coverage. The details of specific encoders have already been introduced in Section 3.

LLMs as cross-modal aligner. As aligners, LLMs bridge heterogeneous modalities by projecting them into a shared semantic space, often through instructional prompting or parameter-efficient tuning. For instance, VIP5 [47] employs multimodal personalized prompts to unify textual, visual, and user-specific signals within a tokenized space, and is further enhanced through adapter-based fine-tuning. Molar [49] uses an MLLM to encode multimodal item information and appends a special token to guide feature aggregation. AB-Rec [8] employs an MLLM to unify textual, visual, and structural signals into item embeddings, and enhances their integration with ID features through distribution alignment and gradient-balanced optimization. DOGE [86] transforms image patches under textual guidance into high-density semantic representations, thereby capturing fine-grained cross-modal correlations and reducing textual bias. NoteLLM-2 [89] treats image and text as separate modality-compressed tokens and maps visual encoder outputs into the LLM token space through a connector. The LLM then produces modality-specific compressed embeddings that are aligned by in-batch contrastive objectives. LLaRA [48] introduces user sequential behaviors as a new modality and projects their ID-based embeddings into the LLM token space through a learnable adapter. By aligning these behavioral tokens with textual tokens via hybrid prompting and curriculum tuning, the model enables effective cross-modal integration for list-wise recommendation.

LLMs as recommender reasoner. LLMs can also function as reasoning engines that operate on multimodal representations to infer user preferences and produce recommendation outputs. For instance, PMG [132] employs LLMs to infer user preferences from historical interactions and express them as explicit keywords and implicit embeddings, which condition multimodal generation for personalized recommendations. MLLM-MSR [50] uses MLLMs to first summarize multimodal item content, then infer user preferences from these summaries, and fine-tune the model for user-item interaction prediction through prompt-driven sequence modeling. PRIME [156] leverages a two-stage framework in which a lightweight retriever generates candidate items, and an MLLM-based ranker performs fine-grained re-ranking through verbalizer-based inference. LaViC [157] also employs a two-stage framework, first condensing images into a small set of visual tokens via self-distillation. It then jointly processes these tokens with dialogue context in a generative framework, enabling the LLM to reason directly over multimodal inputs and generate contextually relevant recommendations in visually-driven domains. Compared with black-box forward scoring methods that

Table 3
Comparison of LLM adaptation techniques in MRSs.

Techniques	Input format	Model status	Models
In-Context Learning	Original input + Exemplars + Task description	LLM parameters frozen	DOGE [86], NoteLLM-2 [89], PMG [132]
Prompt Tuning	Original input + Learnable prompt tokens	Backbone frozen, prompt tunable	Molar [49], PMG [132]
Instruction Tuning	Original input + Natural instruction	LLM tunable	AB-Rec [8], VIP5 [47], LLaRA [48], Molar [49], MLLM-MSR [50], PRIME [156], LaViC [157], ReasonRec [158]

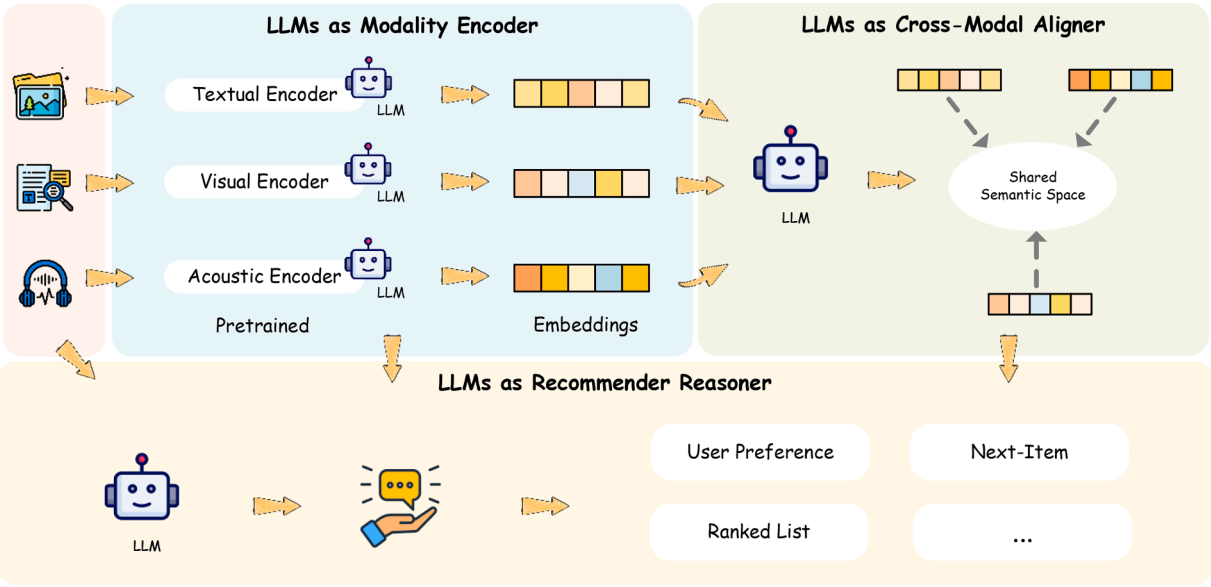


Fig. 8. Illustration of three functional roles of LLMs in MRSs: modality encoders that transform multimodal inputs into embeddings, cross-modal aligners that unify these embeddings, and recommender reasoners that generate user-oriented predictions.

lack interpretability, ReasonRec [158] utilizes an LLM-based reasoning framework where explicit chain-of-thoughts are guided by evidence horizons. It further integrates an uncertainty-aware planner that adaptively delegates tasks to lightweight recommender tools or external signals, achieving both robustness and efficiency in MRSs.

Research outlook. Despite progress, current LLM-enhanced paradigms remain language-centric, often flattening heterogeneous modality signals into a textual space and weakening their inherent complementarity. This language-dominant reliance also heightens vulnerability to hallucinations, where fabricated or semantically inconsistent content may propagate into multimodal reasoning. A more promising direction is LLM-guided co-representations, in which LLMs capture semantics while modality-specific encoders retain structural details. To reduce hallucination risks, external knowledge grounding and cross-modal consistency can be integrated, constraining generated semantics with factual signals from complementary modalities.

5. Model optimization

In modern MRSs, Bayesian personalized ranking (BPR) loss [20] serves as the principal optimization objective for modeling user-item interactions. It is defined over triplets (u, i, j), where user u interacts with item i but not with item j. The set of all such triplets is given by:

$$\mathcal{T} = \{(u, i, j) \mid r_{ui} = 1, r_{uj} = 0\}. \tag{11}$$

The BPR objective aims to maximize the margin between predicted preference scores of positive and negative items:

$$\mathcal{L} = \sum_{(u,i,j)\in\mathcal{T}} -\ln \sigma(y_{ui} - y_{uj}) + \lambda||\theta||_2, \tag{12}$$

where $\sigma(.)$ is the sigmoid function, y_{ui} and y_{uj} are the predicted preference scores of user u for items i and j, λ controls the strength of L_2 regularization, and θ denotes the model parameters. This formulation encourages the model to rank observed items above unobserved ones while reducing overfitting.

Building upon the BPR objective, two dominant optimization paradigms have emerged in MRSs research: **pretraining-finetuning** and **joint learning**. The pretraining-finetuning paradigm [48,50,96,97,99,100,157] first acquires general-purpose multimodal representations with large-scale encoders, and subsequently fine-tunes these representations for recommendation under the BPR objective. In contrast, joint learning combines the BPR loss and auxiliary objectives in a unified training process, directly optimizing multimodal representations for the recommendation task. Representative auxiliary objectives include contrastive loss for representation alignment [42,52,54,59,67,70,71,77,83–85,90–93,98,101,104,106], distillation loss for knowledge transfer across modalities or models [60,66,72,73], cross-entropy loss for classification-oriented objectives [60,61,66,72,104], and mean squared error for reconstruction or regression tasks [42,43]. Recent work therefore concentrates on developing innovative methods under these paradigms to advance model optimization.

5.1. Loss function optimization

Several studies extend the BPR framework to more effectively leverage multimodal information. For instance, MCLN [53] and TMFUN [106] formulate multimodal BPR loss functions that integrate cross-modal information while retaining modality-specific characteristics. MGCE [56] introduces multimodal interest BPR and multimodal conformity BPR to capture heterogeneous user preferences. COHESION [107]

further develops an adaptive BPR loss that assigns greater weight to weaker modalities, thereby improving the robustness of representation learning across the multimodal space.

5.2. Gradient-based optimization

Recent methods refine optimization via gradient manipulation. Mirror Gradient [159] proposes a lightweight gradient modification strategy that alternates between standard gradient descent and a mirror step with scaled gradient inversion. This design regularizes the optimization trajectory toward flatter minima and enhances robustness against inherent noise and frequent information shifts in multimodal inputs. ABRec [8] balances ID and multimodal representations through a gradient modulation strategy, which quantifies each representation's contribution to the training objective and adaptively scales gradient updates. This mechanism suppresses overly dominant ID gradients and ensures sufficient optimization of multimodal components.

5.3. Decoupled and staged training

To improve efficiency and stability, several methods adopt decoupled or staged optimization strategies. For instance, IISAN [88] freezes pretrained backbones and trains lightweight external networks. By caching backbone outputs and tuning modular adapters independently, it lowers computational cost without compromising accuracy. AlignRec [55] employs a two-stage paradigm in which inter-content alignment is first pre-trained, and content-category together with user-item alignments are optimized in the subsequent stage. Such a design addresses the problem of inconsistent learning speeds across modalities. ModalSync [95] proposes a staged co-training framework that alternates between updating the recommender and the feature extractors. Through this alternating process, the model achieves stable and synergistic parameter learning, enabling more effective multimodal integration.

5.4. Modality balancing and regularization

Addressing modality imbalance is a critical challenge in MRSs. POWERec [57] alleviates this issue by introducing weak-modality regularization with hard negatives. It identifies the weakest modality through the smallest rating margin between positive and negative items, and then strengthens it by enlarging the score gap when comparing with a hard negative differing only in that modality. CKD [72] leverages counterfactual inference to estimate the causal effect of each modality on the joint training objective, thereby quantifying its relative contribution. These causal estimates are further exploited to adaptively re-weight the distillation losses, prioritizing the optimization of weaker modalities and mitigating imbalance. MILK [105] addresses modality absence by constructing diverse training environments through cyclic reweighting of modality contributions. It applies invariant learning to minimize prediction variance across these environments and thus ensures stable user preference estimation.

6. Future directions

MRSs are progressing rapidly, yet several critical challenges remain unsolved. Future research needs to tackle technical bottlenecks such as efficiency and temporal reasoning, while also addressing emerging modalities, continual adaptation, and privacy protection. The following directions outline promising avenues with both theoretical and practical implications.

6.1. Resource-efficient multimodal learning

Processing high-dimensional multimodal content incurs substantial computational and storage overhead, which restricts large-scale deployment. Future research should prioritize efficient representation learning strategies that mitigate redundancy while preserving informative features. Potential directions include modality distillation to compress heavy encoders into lightweight alternatives, token pruning to handle long sequences, and cross-modal knowledge transfer to avoid training all modalities jointly. Efficiency-aware training and inference pipelines are essential for scaling MRSs in practical platforms.

6.2. Temporal-causal multimodal reasoning

User preferences are inherently dynamic, with different modalities exerting causal influence at different time points. Current methods primarily capture correlations, while neglecting temporal dependencies shaped by causal structures. Future research may focus on causal sequence models that explicitly represent how exposure to one modality (e.g., images) induces subsequent behaviors mediated by another (e.g., text reviews). Combining temporal point processes, neural ordinary differential equations [160], or dynamic causal graphs with multimodal encoders can provide more faithful modeling of preference evolution and improve both robustness and interpretability.

6.3. Emerging modalities and applications

With new modalities such as 3D objects, physiological signals, and VR interaction traces, MRSs must go beyond conventional text, image, and audio inputs. A key challenge is designing flexible architectures that integrate unseen modalities without full retraining. Solutions include modular encoders, intermediate representations, and meta-learning for adaptation with limited data. Such frameworks are especially promising for personalized healthcare and immersive commerce.

6.4. Continual multimodal adaptation

Most multimodal models are trained once and deployed statically, making them vulnerable to distribution shifts and newly emerging content. Continual learning frameworks provide a promising solution by enabling MRSs to update over time while mitigating catastrophic forgetting. Achieving this requires a combination of lifelong memory modules to preserve modality-specific knowledge, adaptive fusion layers to integrate new modalities, and forgetting mechanisms to attenuate outdated information. Such designs allow MRSs to evolve continuously with user behaviors and content streams in a sustainable and robust manner.

6.5. Privacy-preserving recommendation

Integrating multimodal data raises privacy concerns, particularly for sensitive modalities such as voice, physiological signals, and location. Future systems must safeguard this data while preserving predictive accuracy. Techniques including federated learning, differential privacy, and encrypted representation learning can mitigate risks by preventing exposure of raw multimodal information. Developing privacy-aware training pipelines that balance utility with confidentiality is essential for the large-scale deployment of trustworthy MRSs.

7. Conclusion

Multimodal content enriches RSs with diverse semantic signals, improving the approximation of user preferences. This survey provides a systematic review of methods in MRSs from the perspectives of representation, modeling, and optimization. By categorizing recent advances, this survey clarifies how multimodal signals and training strategies can be aligned to improve recommendation quality. Observed trends include the shift from static fusion to adaptive and context-aware fusion, as well as the evolution from graph-based models to self-supervised and LLM-enhanced paradigms. The proposed paradigm-oriented taxonomy facilitates principled method selection and aims to inspire approaches that balance accuracy, efficiency, and interpretability.

CRediT authorship contribution statement

Lin Pan: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization; **Zhiqiang Pan:** Writing – review & editing, Validation, Supervision, Formal analysis; **Fei Cai:** Writing – review & editing, Supervision; **Honghui Chen:** Writing – review & editing, Supervision.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Á.T. Czapp, M. Jani, B. Domián, B. Hidasi, Dynamic product image generation and recommendation at scale for personalized e-commerce, in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 768–770.

[2] Y. Meng, C. Guo, Y. Cao, T. Liu, B. Zheng, A generative re-ranking model for list-level multi-objective optimization at taobao, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2025, pp. 4213–4218.

[3] K. Vombatkere, S. Mousavi, S. Zannettou, F. Roesner, K.P. Gummadi, Tiktok and the art of personalization: investigating exploration and exploitation on social media feeds, in: Proceedings of the ACM Web Conference 2024, 2024, pp. 3789–3797.

[4] G. Chen, R. Sun, Y. Jiang, T. Li, Y. Dai, Q. Shi, X. Qin, J. Fu, P. Chen, R. Huang, et al., A cold-start recommendation system at kuaishou designed from the short-video perspective, in: Proceedings of the ACM on Web Conference 2025, 2025, pp. 124–132.

[5] X. Yang, Y. Wang, J. Chen, W. Fan, X. Zhao, E. Zhu, X. Liu, D. Lian, Dual test-time training for out-of-distribution recommender system, IEEE Trans. Knowl. Data Eng. 37 (6) (2025) 3312–3326.

[6] H.V. Tran, T. Chen, N. Quoc Viet Hung, Z. Huang, L. Cui, H. Yin, A thorough performance benchmarking on lightweight embedding-based recommender systems, ACM Trans. Inf. Syst. 43 (3) (2025) 1–32.

[7] Q. Liu, J. Zhu, Y. Yang, Q. Dai, Z. Du, X.-M. Wu, Z. Zhao, R. Zhang, Z. Dong, Multimodal pretraining, adaptation, and generation for recommendation: a survey, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6566–6576.

[8] B. Wu, S. Tang, F. Li, B. Han, C. Meng, J. Xiao, J. Gao, Aligning and balancing ID and multimodal representations for recommendation, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2025, pp. 5029–5038.

[9] M.M. Afsar, T. Crump, B. Far, Reinforcement learning based recommender systems: a survey, ACM Comput. Surv. 55 (7) (2022) 1–38.

[10] Y. Wang, W. Ma, M. Zhang, Y. Liu, S. Ma, A survey on the fairness of recommender systems, ACM Trans. Inf. Syst. 41 (3) (2023) 1–43.

[11] C. Gao, Y. Zheng, N. Li, Y. Li, Y. Qin, J. Piao, Y. Quan, J. Chang, D. Jin, X. He, et al., A survey of graph neural networks for recommender systems: challenges, methods, and directions, ACM Trans. Recomm. Syst. 1 (1) (2023) 1–51.

[12] C. Gao, Y. Zheng, W. Wang, F. Feng, X. He, Y. Li, Causal inference in recommender systems: a survey and future directions, ACM Trans. Inf. Syst. 42 (4) (2024) 1–32.

[13] Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasirzadeh, S. Milano, A review of modern recommender systems using generative models (gen-recsys), in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6448–6458.

[14] X. Ren, W. Wei, L. Xia, C. Huang, A comprehensive survey on self-supervised learning for recommendation, ACM Comput. Surv. 58 (1) (2025) 1–38.

[15] L. Wu, X. He, X. Wang, K. Zhang, M. Wang, A survey on accuracy-oriented neural recommendation: from collaborative filtering to information-rich recommendation, IEEE Trans. Knowl. Data Eng. 35 (5) (2022) 4425–4445.

[16] T. Zang, Y. Zhu, H. Liu, R. Zhang, J. Yu, A survey on cross-domain recommendation: taxonomies, methods, and future directions, ACM Trans. Inf. Syst. 41 (2) (2022) 1–39.

[17] Y. Ge, S. Liu, Z. Fu, J. Tan, Z. Li, S. Xu, Y. Li, Y. Xian, Y. Zhang, A survey on trustworthy recommender systems, ACM Trans. Recomm. Syst. 3 (2) (2024) 1–68.

[18] Q. Liu, J. Hu, Y. Xiao, X. Zhao, J. Gao, W. Wang, Q. Li, J. Tang, Multimodal recommender systems: a survey, ACM Comput. Surv. 57 (2) (2024) 1–17.

[19] D. Malitesta, G. Cornacchia, C. Pomo, F.A. Merra, T. Di Noia, E. Di Sciascio, Formalizing multimedia recommendation through multimodal deep learning, ACM Trans. Recomm. Syst. 3 (3) (2025) 1–33.

[20] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback, arXiv preprint arXiv:1205.2618 (2012).

[21] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk, Session-based recommendations with recurrent neural networks, arXiv preprint arXiv:1511.06939 (2015).

[22] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[23] S. Arora, Y. Liang, T. Ma, A simple but tough-to-beat baseline for sentence embeddings, in: International Conference on Learning Representations, 2017, pp. 1–16.

[24] T. Donkers, B. Loepp, J. Ziegler, Sequential user-based recurrent neural network recommendations, in: Proceedings of the 17th ACM Conference on Recommender Systems, 2017, pp. 152–160.

[25] J. Devlin, M.-W. Chang, K. Lee, K. 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, 2019, pp. 4171–4186.

[26] N. Reimers, I. Gurevych, Sentence-bert: sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[27] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (140) (2020) 1–67.

[28] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Adv. Neural Inf. Process. Syst. 35 (2022) 27730–27744.

[29] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[30] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16×16 words: transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[34] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, 2021, pp. 8748–8763.

[35] J. Li, D. Li, C. Xiong, S. Hoi, Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation, in: Proceedings of the 39th International Conference on Machine Learning, 2022, pp. 12888–12900.

[36] Y. Gong, Y.-A. Chung, J. Glass, Ast: audio spectrogram transformer, arXiv preprint arXiv:2104.01778 (2021).

[37] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, F. Wei, BEATs: audio pre-training with acoustic tokenizers, in: Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 5178–5193.

[38] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, Z. Zhao, Make-an-audio: text-to-audio generation with prompt-enhanced diffusion models, in: Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 13916–13932.

[39] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, M.D. Plumbley, AudioLDM: text-to-audio generation with latent diffusion models, in: Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 21450–21474.

[40] H. Liu, K. Chen, Q. Tian, W. Wang, M.D. Plumbley, AudioSR: versatile audio super-resolution at scale, in: ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 1076–1080.

[41] Y. Wei, X. Wang, L. Nie, X. He, T.-S. Chua, Graph-refined convolutional network for multimedia recommendation with implicit feedback, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3541–3549.

[42] Z. Lin, Y. Tan, Y. Zhan, W. Liu, F. Wang, C. Chen, S. Wang, C. Yang, Contrastive intra-and inter-modality generation for enhancing incomplete multimedia recommendation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 6234–6242.

[43] H. Ma, Y. Yang, L. Meng, R. Xie, X. Meng, Multimodal conditioned diffusion model for recommendation, in: Proceedings of the ACM Web Conference 2024, 2024, pp. 1733–1740.

[44] Y. Shang, C. Gao, J. Chen, D. Jin, Y. Li, Improving item-side fairness of multimodal recommendation via modality debiasing, in: Proceedings of the ACM Web Conference 2024, 2024, pp. 4697–4705.

[45] Z. Lin, Y. Tan, J. Chen, H. Zhang, C. Chen, S. Wang, C. Yang, Unified heterogeneous hypergraph construction for incomplete multimedia recommendation, ACM Trans. Inf. Syst. 43 (5) (2024) 1–31.

[46] J. Huang, J. Qin, Y. Yu, W. Zhang, Beyond graph convolution: multimodal recommendation with topology-aware mlps, in: Proceedings of the AAAI Conference on Artificial Intelligence, 39, 2025, pp. 11808–11816.

[47] S. Geng, J. Tan, S. Liu, Z. Fu, Y. Zhang, VIP5: towards multimodal foundation models for recommendation, in: 2023 Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 9606–9620.

[48] J. Liao, S. Li, Z. Yang, J. Wu, Y. Yuan, X. Wang, X. He, Llara: large language-recommendation assistant, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 1785–1795.

[49] Y. Luo, Q. Qin, H. Zhang, M. Cheng, R. Yan, K. Wang, J. Ouyang, Molar: Multimodal LLMs with Collaborative Filtering Alignment for Enhanced Sequential Recommendation, arXiv preprint arXiv:2412.18176 (2024).

[50] Y. Ye, Z. Zheng, Y. Shen, T. Wang, H. Zhang, P. Zhu, R. Yu, K. Zhang, H. Xiong, Harnessing multimodal large language models for multimodal sequential recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2025, pp. 13069–13077.

[51] F. Liu, H. Chen, Z. Cheng, A. Liu, L. Nie, M. Kankanhalli, Disentangled multimodal representation learning for recommendation, IEEE Trans. Multimed. 25 (2022) 7149–7159.

[52] Z. Tao, X. Liu, Y. Xia, X. Wang, L. Yang, X. Huang, T.-S. Chua, Self-supervised learning for multimedia recommendation, IEEE Trans. Multimed. 25 (2022) 5107–5116.

[53] S. Li, D. Guo, K. Liu, R. Hong, F. Xue, Multimodal counterfactual learning network for multimedia-based recommendation, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 1539–1548.

[54] G. Lin, M. Zhen, D. Wang, Q. Long, Y. Zhou, M. Xiao, GUME: graphs and user modalities enhancement for long-tail multimodal recommendation, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 2024, pp. 1400–1409.

[55] Y. Liu, K. Zhang, X. Ren, Y. Huang, J. Jin, Y. Qin, R. Su, R. Xu, Y. Yu, W. Zhang, AlignRec: aligning and training in multimodal recommendations, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 2024, pp. 1503–1512.

[56] S. Li, F. Xue, K. Liu, D. Guo, R. Hong, Multimodal graph causal embedding for multimedia-based recommendation, IEEE Trans. Knowl. Data Eng. 36 (12) (2024) 8842–8858.

[57] X. Dong, X. Song, M. Tian, L. Hu, Prompt-based and weak-modality enhanced multimodal recommendation, Inf. Fusion 101 (2024) 101989.

[58] J. Hu, B. Hooi, B. He, Y. Wei, Modality-independent graph neural networks with global transformers for multimodal recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 39, 2025, pp. 11790–11798.

[59] Y. Qi, Q. Zhang, X. Lin, X. Su, J. Zhu, J. Wang, J. Li, Seeing beyond noise: joint graph structure evaluation and denoising for multimodal recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2025, pp. 12461–12469.

[60] W. Ji, X. Liu, A. Zhang, Y. Wei, Y. Ni, X. Wang, Online distillation-enhanced multimodal transformer for sequential recommendation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 955–965.

[61] X. Zhang, B. Xu, Z. Ren, X. Wang, H. Lin, F. Ma, Disentangling id and modality effects for session-based recommendation, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 1883–1892.

[62] X. Liu, Q. Song, L. Xiao, C. Wang, X. Gao, LPIC: learnable prompts and ID-guided contrastive learning for multimodal recommendation, ACM Trans. Multimed. Comput., Commun. Appl. 21 (9) (2025) 1–16.

[63] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, F. Jiang, Bootstrap latent representations for multi-modal recommendation, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 845–854.

[64] X. Zhou, Z. Shen, A tale of two graphs: freezing and denoising graph structures for multimodal recommendation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 935–943.

[65] Y. Wei, W. Liu, F. Liu, X. Wang, L. Nie, T.-S. Chua, Lightgt: a light graph transformer for multimedia recommendation, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 1508–1517.

[66] F. Liu, H. Chen, Z. Cheng, L. Nie, M. Kankanhalli, Semantic-guided feature distillation for multimodal recommendation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 6567–6575.

[67] R.K. Ong, A.W.H. Khong, Spectrum-based modality representation fusion graph convolutional network for multimodal recommendation, in: Proceedings of the 18th ACM International Conference on Web Search and Data Mining, 2025, pp. 773–781.

[68] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, L. Wang, Mining latent structures for multimedia recommendation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 3872–3880.

[69] X. Zhou, C. Miao, Disentangled graph variational auto-Encoder for multimodal recommendation with interpretability, IEEE Trans. Multimed. 26 (2024) 7543–7554.

[70] Z. Guo, J. Li, G. Li, C. Wang, S. Shi, B. Ruan, LGMRec: local and global graph learning for multimodal recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 8454–8462.

[71] Y. Jiang, L. Xia, W. Wei, D. Luo, K. Lin, C. Huang, Diffmm: multi-modal diffusion model for recommendation, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 7591–7599.

[72] J. Zhang, G. Liu, Q. Liu, S. Wu, L. Wang, Modality-balanced learning for multimedia recommendation, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 7551–7560.

[73] W. Wei, J. Tang, L. Xia, Y. Jiang, C. Huang, Promptmm: multi-modal knowledge distillation for recommendation with prompt-tuning, in: Proceedings of the ACM Web Conference 2024, 2024, pp. 3217–3228.

[74] Z. Chen, J. Xu, H. Hu, Don't lose yourself: boosting multimodal recommendation via reducing node-neighbor discrepancy in graph convolutional network, in: ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5.

[75] J. Xu, Z. Chen, S. Yang, J. Li, E.C.H. Ngai, The Best is Yet to Come: Graph Convolution in the Testing Phase for Multimodal Recommendation, arXiv preprint arXiv:2507.18489 (2025).

[76] W. Chen, L. Chen, Y. Ni, Y. Zhao, Causality-inspired fair representation learning for multimodal recommendation, ACM Trans. Inf. Syst. 43 (6) (2025) 1–29.

[77] J. Li, S. Wang, Q. Zhang, S. Yu, F. Chen, Generating with fairness: a modality-diffused counterfactual framework for incomplete multimodal recommendations, in: Proceedings of the ACM on Web Conference 2025, 2025, pp. 2787–2798.

[78] Y. Xiu, X. Tong, Dual-layer cross-modal alignment recommendation based on the diffusion model, Inf. Fusion 125 (2025) 103472.

[79] Q. Song, J. Hu, L. Xiao, B. Sun, X. Gao, S. Li, DiffCL: a diffusion-based contrastive learning framework with semantic alignment for multimodal recommendations, IEEE Trans. Neural Netw. Learn. Syst. 36 (10) (2025) 18587–18597.

[80] F. Mo, L. Xiao, Q. Song, X. Gao, W. Song, S. Wang, FGCM: modality-behavior fusion model integrated with graph contrastive learning for multimodal recommendation, IEEE Multimed. 32 (3) (2025) 27–37.

[81] X. Li, C. Wang, J. Tan, X. Zeng, D. Ou, D. Ou, B. Zheng, Adversarial multimodal representation learning for click-through rate prediction, in: Proceedings of the ACM Web Conference 2020, 2020, pp. 827–836.

[82] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, L. Nie, Dualgnn: dual graph neural network for multimedia recommendation, IEEE Trans. Multimed. 25 (2021) 1074–1084.

[83] H. Su, J. Li, F. Li, K. Lu, L. Zhu, SOIL: contrastive second-order interest learning for multimodal recommendation, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 5838–5846.

[84] Z. Yi, I. Ounis, A unified graph transformer for overcoming isolations in multimodal recommendation, in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 518–527.

[85] J. Xu, Z. Chen, S. Yang, J. Li, H. Wang, E.C.H. Ngai, Mentor: multi-level self-supervised learning for multimodal recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2025, pp. 12908–12917.

[86] F. Meng, Z. Meng, R. Jin, R. Lin, B. Wu, DOGE: LLMs-enhanced hyper-knowledge graph recommender for multimodal recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2025, pp. 12399–12407.

[87] Y. Dang, W. Chen, Z. Pan, Y. Duan, F. Cai, H. Chen, PEARL: a dual-layer graph learning for multimodal recommendation, Inf. Fusion 122 (2025) 103168.

[88] J. Fu, X. Ge, X. Xin, A. Karatzoglou, I. Arapakis, J. Wang, J.M. Jose, IISAN: efficiently adapting multimodal representation for sequential recommendation with decoupled PEFT, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 687–697.

[89] C. Zhang, H. Zhang, S. Wu, D. Wu, T. Xu, X. Zhao, Y. Gao, Y. Hu, E. Chen, Notellm-2: multimodal large representation models for recommendation, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2025, pp. 2815–2826.

[90] J. Zhang, Y. Zhu, Q. Liu, M. Zhang, S. Wu, L. Wang, Latent structure mining with contrastive modality fusion for multimedia recommendation, IEEE Trans. Knowl. Data Eng. 35 (9) (2022) 9154–9167.

[91] K. Liu, F. Xue, D. Guo, P. Sun, S. Qian, R. Hong, Multimodal graph contrastive learning for multimedia-based recommendation, IEEE Trans. Multimed. 25 (2023) 9343–9355.

[92] W. Yang, Q. Yang, Multimodal-aware multi-intention learning for recommendation, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 5663–5672.

[93] G. Xv, X. Li, R. Xie, C. Lin, C. Liu, F. Xia, Z. Kang, L. Lin, Improving multi-modal recommender systems by denoising and aligning multi-modal content and user feedback, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 3645–3656.

[94] X.-R. Sheng, F. Yang, L. Gong, B. Wang, Z. Chan, Y. Zhang, Y. Cheng, Y.-N. Zhu, T. Ge, H. Zhu, et al., Enhancing taobao display advertising with multimodal representations: challenges, approaches and insights, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 2024, pp. 4858–4865.

[95] S. Liu, C. Li, M. Zhao, L. Zhang, J. Bu, ModalSync: synchronizing user behavior with multimodal features for multimodal pre-training recommendation, in: Proceedings of the ACM on Web Conference 2025, 2025, pp. 2278–2287.

[96] S. Bian, X. Pan, W.X. Zhao, J. Wang, C. Wang, J.-R. Wen, Multi-modal mixture of experts represetation learning for sequential recommendation, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 110–119.

[97] L. Zhang, X. Zhou, Z. Zeng, Z. Shen, Multimodal pre-training for sequential recommendation via contrastive learning, ACM Trans. Recomm. Syst. 3 (1) (2024) 1–23.

[98] S. Zhang, L. Chen, D. Shen, C. Wang, H. Xiong, Hierarchical time-aware mixture of experts for multi-modal sequential recommendation, in: Proceedings of the ACM on Web Conference 2025, 2025, pp. 3672–3682.

[99] F. Xiao, L. Deng, J. Chen, H. Ji, X. Yang, Z. Ding, B. Long, From abstract to details: a generative multimodal fusion framework for recommendation, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 258–267.

[100] J. Wang, Z. Zeng, Y. Wang, Y. Wang, X. Lu, T. Li, J. Yuan, R. Zhang, H.-T. Zheng, S.-T. Xia, Missrec: pre-training and transferring multi-modal interest-aware sequence representation for recommendation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 6548–6557.

[101] P. Yu, Z. Tan, G. Lu, B.-K. Bao, Multi-view graph convolutional network for multimedia recommendation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 6576–6585.

[102] H. Hu, W. Guo, Y. Liu, M.-Y. Kan, Adaptive multi-modalities fusion in sequential recommendation systems, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 843–853.

[103] F. Chen, J. Wang, Y. Wei, H.-T. Zheng, J. Shao, Breaking isolation: multimodal graph fusion for multimedia recommendation by edge-wise modulation, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 385–394.

[104] X. Zhang, B. Xu, F. Ma, C. Li, L. Yang, H. Lin, Beyond co-occurrence: multimodal session-based recommendation, IEEE Trans. Knowl. Data Eng. 36 (4) (2024) 1450–1462.
[105] H. Bai, L. Wu, M. Hou, M. Cai, Z. He, Y. Zhou, R. Hong, M. Wang, Multimodality invariant learning for multimedia-based new item recommendation, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 677–686.
[106] Y. Zhou, J. Guo, H. Sun, B. Song, F.R. Yu, Attention-guided multi-step fusion: a hierarchical fusion network for multimodal recommendation, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 1816–1820.
[107] J. Xu, Z. Chen, W. Wang, X. Hu, S.-W. Kim, E.C.H. Ngai, COHESION: composite graph convolutional network with dual-stage fusion for multimodal recommendation, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2025, pp. 1830–1839.
[108] J. Xu, Z. Chen, J. Li, S. Yang, H. Wang, Y. Li, M. Li, P. Wu, E.C.H. Ngai, Mdvt: enhancing multimodal recommendation with model-agnostic multimodal-driven virtual triplets, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2025, pp. 3378–3389.
[109] L. He, H. Chen, D. Wang, S. Jameel, P. Yu, G. Xu, Click-through rate prediction with multi-modal hypergraphs, in: Proceedings of the 30th ACM International Conference on Information and Knowledge Management, 2021, pp. 690–699.
[110] T.N. Kipf, Semi-Supervised Classification with Graph Convolutional Networks, arXiv preprint arXiv:1609.02907 (2016).
[111] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. 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, 2020, pp. 639–648.
[112] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 1–11.
[113] Q. Song, R. Dian, B. Sun, J. Xie, S. Li, Multi-scale conformer fusion network for multi-participant behavior analysis, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 9472–9476.
[114] Q. Song, B. Sun, S. Li, Multimodal sparse transformer network for audio-visual speech recognition, IEEE Trans. Neural Netw. Learn. Syst. 34 (12) (2023) 10028–10038.
[115] W.-C. Kang, J. McAuley, Self-attentive sequential recommendation, in: 2018 IEEE International Conference on Data Mining (ICDM), 2018, pp. 197–206.
[116] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, P. Jiang, BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 1441–1450.
[117] Y. Liu, S. Yang, C. Lei, G. Wang, H. Tang, J. Zhang, A. Sun, C. Miao, Pre-training graph transformer with multimodal side information for recommendation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2853–2861.
[118] M.-Y. Hong, Y.-J. Hsu, M.-C. Chiang, C. Lin, MTSTRec: multimodal time-aligned shared token recommender, in: Proceedings of the 42rd International Conference on Machine Learning, 2025, pp. 1–22.
[119] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, J. Huang, A survey on mixture of experts in large language models, IEEE Trans. Knowl. Data Eng. 37 (7) (2025) 3896–3915.
[120] J. Yi, Z. Chen, Variational mixture of stochastic experts auto-encoder for multi-modal recommendation, IEEE Trans. Multimed. 26 (2024) 8941–8954.
[121] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning, 2020, pp. 1597–1607.
[122] N. Rethmeier, I. Augenstein, A primer on contrastive pretraining in language processing: methods, lessons learned, and perspectives, ACM Comput. Surv. 55 (10) (2023) 1–17.
[123] O.A. van den, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).
[124] Z. Yi, X. Wang, I. Ounis, C. Macdonald, Multi-modal graph contrastive learning for micro-video recommendation, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1807–1811.
[125] J. Yi, Z. Chen, Multi-modal variational graph auto-encoder for recommendation systems, IEEE Trans. Multimed. 24 (2021) 1067–1079.
[126] J. Yi, Y. Zhu, J. Xie, Z. Chen, Cross-modal variational auto-encoder for content-based micro-video background music recommendation, IEEE Trans. Multimed. 25 (2021) 515–528.
[127] Y. Lu, M. Zhang, A.J. Ma, X. Xie, J. Lai, Coarse-to-fine latent diffusion for pose-guided person image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6420–6429.
[128] M. Yi, A. Li, Y. Xin, Z. Li, Towards understanding the working mechanism of text-to-image diffusion model, Adv. Neural Inf. Process. Syst. 37 (2024) 55342–55369.
[129] J. Walker, T. Zhong, F. Zhang, Q. Gao, F. Zhou, Recommendation via collaborative diffusion generative model, in: International Conference on Knowledge Science, Engineering and Management, 2022, pp. 593–605.
[130] W. Wang, Y. Xu, F. Feng, X. Lin, X. He, T.-S. Chua, Diffusion recommender model, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 832–841.
[131] Y. Yang, H. Ma, L. Meng, S. Xu, R. Xie, X. Meng, Curriculum conditioned diffusion for multimodal recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 39, 2025, pp. 13035–13043.
[132] X. Shen, R. Zhang, X. Zhao, J. Zhu, X. Xiao, Pmg: personalized multimodal generation with large language models, in: Proceedings of the ACM Web Conference 2024, 2024, pp. 3833–3843.
[133] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014) 1–9.
[134] J. Tang, X. Du, X. He, F. Yuan, Q. Tian, T.-S. Chua, Adversarial training towards robust multimedia recommender system, IEEE Trans. Knowl. Data Eng. 32 (5) (2019) 855–867.
[135] T. Wei, F. Feng, J. Chen, Z. Wu, J. Yi, X. He, Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021, pp. 1791–1800.
[136] Z. Zhao, J. Chen, S. Zhou, X. He, X. Cao, F. Zhang, W. Wu, Popularity bias is not always evil: disentangling benign and harmful bias for recommendation, IEEE Trans. Knowl. Data Eng. 35 (10) (2022) 9920–9931.
[137] Z. Liu, Y. Fang, M. Wu, Mitigating popularity bias for users and items with fairness-centric adaptive recommendation, ACM Trans. Inf. Syst. 41 (3) (2023) 1–27.
[138] M. Cai, L. Chen, Y. Wang, H. Bai, P. Sun, L. Wu, M. Zhang, M. Wang, Popularity-aware alignment and contrast for mitigating popularity bias, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 187–198.
[139] X. Wang, R. Zhang, Y. Sun, J. Qi, Combating selection biases in recommender systems with a few unbiased ratings, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 427–435.
[140] Z. Wang, S. Shen, Z. Wang, B. Chen, X. Chen, J.-R. Wen, Unbiased sequential recommendation with latent confounders, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2195–2204.
[141] M. Mansoury, B. Mobasher, H. van Hoof, Mitigating exposure bias in online learning to rank recommendation: a novel reward model for cascading bandits, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 2024, pp. 1638–1648.
[142] S. Mu, Y. Li, W.X. Zhao, J. Wang, B. Ding, J.-R. Wen, Alleviating spurious correlations in knowledge-aware recommendations through counterfactual generator, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1401–1411.
[143] C. Gao, S. Wang, S. Li, J. Chen, X. He, W. Lei, B. Li, Y. Zhang, P. Jiang, CIRS: bursting filter bubbles by counterfactual interactive recommender system, ACM Trans. Inf. Syst. 42 (1) (2023) 1–27.
[144] X. Zhang, H. Jia, H. Su, W. Wang, J. Xu, J.-R. Wen, Counterfactual reward modification for streaming recommendation with delayed feedback, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 41–50.
[145] Y. Zhang, F. Feng, X. He, T. Wei, C. Song, G. Ling, Y. Zhang, Causal intervention for leveraging popularity bias in recommendation, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 11–20.
[146] S. Tang, Q. Li, D. Wang, C. Gao, W. Xiao, D. Zhao, Y. Jiang, Q. Ma, A. Zhang, Counterfactual video recommendation for duration debiasing, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 4894–4903.
[147] X. Wang, Q. Li, D. Yu, Q. Li, G. Xu, Counterfactual explanation for fairness in recommendation, ACM Trans. Inf. Syst. 42 (4) (2024) 1–30.
[148] R. Qiu, S. Wang, Z. Chen, H. Yin, Z. Huang, Causalrec: causal inference for visual debiasing in visually-aware recommendation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 3844–3852.
[149] X. Liu, Z. Tao, J. Shao, L. Yang, X. Huang, Elimrec: eliminating single-modal bias in multimedia recommendation, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 687–695.
[150] S. Kim, H. Kang, S. Choi, D. Kim, M. Yang, C. Park, Large language models meet collaborative filtering: an efficient all-round llm-based recommender system, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 1395–1406.
[151] J. Hu, W. Xia, X. Zhang, C. Fu, W. Wu, Z. Huan, A. Li, Z. Tang, J. Zhou, Enhancing sequential recommendation via llm-based semantic embedding learning, in: Proceedings of the ACM Web Conference 2024, 2024, pp. 103–111.
[152] J. Liu, X. Yan, D. Li, G. Zhang, H. Gu, P. Zhang, T. Lu, L. Shang, N. Gu, Improving LLM-powered recommendations with personalized information, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2025, pp. 2560–2565.
[153] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.
[154] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, arXiv preprint arXiv:2104.08691 (2021).
[155] S. Mishra, D. Khashabi, C. Baral, Y. Choi, H. Hajishirzi, Reframing instructional prompts to gptk's language, arXiv preprint arXiv:2109.07830 (2021).
[156] Z. Yue, H. Zeng, Y. Wang, J. McAuley, D. Wang, Preference-optimized retrieval and ranking for efficient multimodal recommendation, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2025, pp. 3692–3703.
[157] H. Jeon, S. Koide, Y. Wang, Z. He, J. McAuley, Adapting large vision-language models to visually-aware conversational recommendation, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2025, pp. 1037–1048.

[158] Y. Zhang, X. Liu, X. Zeng, M. Liang, J. Yang, R. Jin, W.-Y. Chen, Y. Han, H. Ma, B. Long, et al., ReasonRec: a reasoning-augmented multimodal agent for unified recommendation, in: Proceedings of the 42rd International Conference on Machine Learning, 2025, pp. 1–21.

[159] S. Zhong, Z. Huang, D. Li, W. Wen, J. Qin, L. Lin, Mirror gradient: towards robust multimodal recommender systems via exploring flat local minima, in: Proceedings of the ACM Web Conference 2024, 2024, pp. 3700–3711.

[160] R.T.Q. Chen, Y. Rubanova, J. Bettencourt, D.K. Duvenaud, Neural ordinary differential equations, Adv. Neural Inf. Process. Syst. 31 (2018) 1–13.