From Raw IDs to Semantic Planning: How Recommender Systems Utilize Information at Scale

Changhong Jin
changhong.jin@ucd.ie
University College Dublin
Dublin, Ireland

Shiqiu Yang
shiqiu.yang@ucdconnect.ie
University College Dublin
Dublin, Ireland

Roger Zhe Li
roger.zhe.li@huawei.com
Huawei Ireland Research Centre
Dublin, Ireland

Yingjie Niu
yingjie.niu1@h-partners.com
Huawei Ireland Research Centre
Dublin, Ireland

Aghiles Salah
aghiles.salah@h-partners.com
Huawei Ireland Research Centre
Dublin, Ireland

Mete Sertkan
mete.sertkan@h-partners.com
Huawei Ireland Research Centre
Dublin, Ireland

Zheng Ju
zheng.ju@ucd.ie
University College Dublin
Dublin, Ireland

Xingsheng Guo
xingsheng.guo2@huawei.com
Huawei Ireland Research Centre
Dublin, Ireland

Huifeng Guo
huifeng.guo@huawei.com
Huawei Technologies
Shenzhen, China

Ruihai Dong
ruihai.dong@ucd.ie
University College Dublin
Dublin, Ireland

Barry Smyth
barry.smyth@ucd.ie
University College Dublin
Dublin, Ireland

Abstract

The evolution of recommender systems can be explored by asking how they utilize information at scale. Throughout most of the historical period under consideration during the past two decades, industrial systems have relied on raw IDs, which are discrete, globally unique, and semantically opaque identifiers that enable exact lookup, logging, and item-specific memorization at scale. Over time, however, recommender systems have sought to utilize richer sources of information, including item content, context, multimodal signals, and cross-domain structure. This development has led to a new stage in which part of such information is no longer used solely as auxiliary features around item identity, but is increasingly encapsulated in semantic IDs that provide a more structured, model-facing form of identity. We argue that this shift goes beyond the rise of generative recommendation over traditional methods. Indeed, it reflects a broader evolution in how recommender systems utilize information under industrial-scale constraints. This paper looks at the past, present, and future to examine three connected questions: why raw IDs dominated the early development of recommender systems, why semantic information is increasingly being encapsulated in IDs today, and what may come next once recommendations move beyond semantic retrieval. In particular, we introduce semantic planning as a possible future direction in which the system first predicts the semantic target of the next exposure, and only then instantiates that target as a specific item or generated creative. We further argue that such a shift may require changes not only in model design but also in evaluation and in the way recommender systems coordinate the objectives of users, platforms, and providers.

CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

Raw IDs, Semantic Information, Semantic IDs, Semantic Planning

1 Introduction

Many real-world recommender systems, especially industrial ones, are multi-stakeholder systems [1]. Their primary role is to help users navigate large choice spaces and information overload [15], but in practice, they should also account for the commercial and strategic objectives of other stakeholders, including platform providers, online product providers, and advertisers [14, 19]. In such settings, recommendation is not only a problem of predicting what a user may like; it is also about optimizing the decisions that need to be made by a business operating at scale.

From this perspective, a useful lens on the past two decades of recommender systems is **how they utilize information at scale**. In this context, the term *information* extends well beyond historical interaction logs. The scope of this field encompasses item content [8, 23, 38, 39], contextual signals [2, 29], multi-view behaviors [18], cross-domain structure [30, 47, 48], and increasingly heterogeneous forms of metadata and feedback [21, 32]. Many influential developments in recommender systems can be viewed as attempts to incorporate richer and richer signals. In particular, the field has repeatedly sought representations that improve generalization while remaining compatible with large vocabularies, growing embedding tables, and production-serving constraints [4, 5, 26, 34].

In this paper, we argue that one useful way to understand recent recommender-system evolution is as a flow from *Raw IDs*, to richer *semantic information* around IDs, and increasingly to *semantic IDs*. Early industrial recommender systems therefore relied heavily on raw IDs as stable categorical inputs and operational keys for lookup, logging, ranking, and memorization at scale [9, 10]. Later systems enriched these pipelines with statistical features [6, 40],

reviews [51], visual signals [8], contextual features [4, 37], and sequential behavior [52], while largely keeping raw IDs as the identity layer. More recently, semantic-ID and tokenization work has explored incorporating part of this richer information into the IDs themselves, in response to large, dynamic catalogs, embedding table growth, and the need for more unified model interfaces toward a "One-Model-Fits-All" effect [27, 28]. What is changing today is therefore not only that recommender systems use richer signals, but that part of their semantic structure is increasingly being moved into the identity layer itself.

This development trajectory can be understood as a sequence of progressively wider unifications. The first stage brought unification within recommendation: semantic IDs allow items from the same domain to share structural identity, enabling better generalization and cold-start handling. The second stage extended this to unification across domains and modalities, as semantic IDs provide a more common substrate for cross-domain transfer and multimodal alignment. The third, ongoing stage concerns the unification of tasks: semantic IDs are increasingly shared between search and recommendation. Yet all three stages remain primarily focused on better user-centered matching over existing items. The critical frontier that none of them fully addresses is explicit multi-stakeholder decision-making, i.e., how to reason over the potentially conflicting objectives of users, platforms, and providers within a single recommendation decision. Under this circumstance, one possible direction is *semantic planning*, which may emerge as the next stage. Once a recommender system can represent not only what an item is, but also what the next exposure is intended to accomplish, recommendation is no longer limited to selecting the nearest available item. The system can instead first determine the semantic target of the next exposure — a representation of what it should accomplish in terms of user need, platform objective, and provider content — and only then instantiate that target as a specific item, or generated creative. Semantic planning is thus the process; the semantic target is its intermediate output. This paper argues that the shift from raw IDs to semantic IDs sets the stage for this transition, and that semantic planning represents its most consequential implication.

2 Past: Raw IDs as Operational Anchors

Raw IDs are discrete, globally unique, and semantically opaque identifiers used to index items in recommender systems. In industrial pipelines, such IDs usually function as stable database keys across retrieval, lookup, logging, storage, and serving [26]. They allow a consistent track & trace on the same item across heterogeneous system components, even when the models, features, and business rules around that item change.

Raw IDs became central because they resolve two practical problems. Firstly, they provide a stable method for connecting catalog entities to the numerous operations required by an industrial recommender system under high concurrency. Crucially, as highlighted in the seminal work on ML technical debt [31], industrial systems often suffer from "implicit coupling", in which multiple infrastructure components rely on these stable anchors for cross-service synchronization. In this sense, raw IDs are not merely a modeling choice, but an architectural necessity to maintain stability in complex and high-concurrency environments. Secondly, they provide models with a direct way to accumulate and aggregate item-specific behavioral evidence. This matters because semantically similar items can still develop different interaction histories and business value. While two items might share identical semantic features, their real-world trajectories, such as price, availability, popularity, or commercial value, may still diverge. Historically, raw IDs provided the most direct mechanism for memorizing such behavioral differences without requiring them to be explained through semantic features [4]. However, this memorization comes with a structural limitation. By treating each item as an atomic, opaque unit, the system is forced to learn each item's nuances from scratch. While this approach provides high fidelity for head items, it also suffers from several limitations, including cold-start [49] and the growing memory and serving overhead caused by increasingly large embedding tables [26].

As recommender systems evolved, richer semantic signals such as text and images were increasingly incorporated into models, but item identity itself remained largely Raw-ID-based. The issue, then, was not a lack of semantic information, but that it remained largely decoupled from the identity layer. This tension between representing items through semantic features and identifying them through arbitrary indices motivates the shift toward semantic IDs. The goal of this shift is to move from unconstrained memorization toward a structured identity that more directly bridges content and behavior [28].

3 Present: From Semantic Information to Semantic IDs

A recommender system can make extensive use of semantic information while keeping item identity organized in largely the same way. Reviews, images, contextual features, and behavioral signals may all enrich the model, while raw IDs still remain the primary identifiers [4, 36]. Semantic IDs mark a further step: **part of that information is encapsulated in the identifiers themselves, turning item identity into a more structured, model-facing object [28]**. Their significance, therefore, lies not in the mere availability of semantic signals, but in reorganizing part of those signals into the ID itself. The key advantage over embedding-based alignment is at the model interface level: semantic IDs allow us to decouple embedding alignment from task-specific models, enabling more flexible integration of heterogeneous item representations. As discrete token sequences, semantic IDs are self-contained and can be directly consumed by any model operating over a shared vocabulary—whether generative or retrieval-oriented—without additional cross-space alignment [12, 35]. This flexibility is particularly consequential in settings where useful structure spans multiple domains, modalities, or task boundaries, precisely the long-standing challenges that have motivated the present stage of recommender system development.

Traditionally, the model attempts to transfer knowledge by aligning or disentangling shared and domain-specific representations in continuous embedding spaces [24, 25, 48]. Recent work suggests that part of the cross-domain structure can be encapsulated in the IDs themselves. This allows items from different domains to be related through a more unified identity space [12], where

domain-agnostic kernels and domain-specific nuances are effectively unified [17] or factorized [13] within the ID structure itself.

A similar shift is taking place in multimodal recommendation. Images, text, and other modalities have long been used as side information, especially in sparse and cold-start settings. At present, however, the goal is increasingly to align and encode part of that multimodal structure into semantic IDs. By doing so, multimodal signals are organized into the identifiers themselves, making them natively accessible to retrieval-, ranking-, and generation-oriented models [16, 44, 45, 53].

Additionally, recent progress has further incorporated collaborative signals into semantic IDs, enabling them to carry information from both users and items [3, 7]. This trend also facilitates the use of semantic IDs in the growing unification of search and recommendation. Recent progress in the community makes this connection more direct by enabling more shared item representations, generation mechanisms, and retrieval interfaces across the two tasks [27, 33, 50]. This does not mean that search and recommendation are already fully unified. It suggests that semantic IDs make it easier to share item representations and retrieval interfaces.

Taken together, these developments share a common character: they unify representation and retrieval around better matching of users to existing items. Each makes it easier for a system to find and present what a user is likely to want. Yet these benefits often come with added complexity in ID stability and lifecycle management and, more importantly, leave the multi-stakeholder structure of industrial recommendation largely untouched. Users, platforms, and providers each have distinct and often conflicting objectives, and none of the above unifications provides a principled mechanism for reasoning over them within a single recommendation decision. This is the boundary of the present stage. The future begins not when better semantic representations of items become available, but when the system must reason explicitly about **what the next exposure should accomplish under competing user, platform, and provider objectives**.

4 Future: Semantic Planning and Implications

The shift from semantic retrieval to semantic planning has implications that extend beyond model design and may open up new R&D avenues. We examine three directions: the introduction of an intermediate planning layer before item selection, the consequent flexibility in how semantic targets are instantiated, and the changes that may be brought to evaluation and industrial deployment.

4.1 From Semantic Retrieval to Semantic Planning

Most current semantic recommender systems remain, at their core, retrieval systems. The generated object is usually an identifier that resolves to a specific item already in the catalog; "generation" is employed principally as a retrieval mechanism [20]. However different their architectures, they are all designed to answer the same question: **which existing item should be shown next?** This framing has a structural consequence: the system's only degree of freedom is the choice among available items, and its only explicit objective is user-item matching. The multi-stakeholder purpose of the next exposure—what it should accomplish for the user, the platform, and the provider—remains implicit, absorbed into item scores rather than reasoned about directly.

Consider a simple hotel recommendation scenario. A user is planning a weekend trip and has recently browsed several hotels. Their behavior may imply an ideal semantic target, such as accommodation in the city center that offers safety assurances, a good breakfast, a gym, and the option of free cancellation. Suppose, however, that no existing hotel in the catalog satisfies all of these conditions. The closest available option may be a centrally located, well-reviewed hotel that includes breakfast but without a gym or free cancellation. A retrieval-oriented system may still recommend this hotel because it is the nearest available match. The user may click the item, read the reviews, and then leave the page after noticing that the cancellation policy is missing. This response is more than a preference among existing hotels. It reveals an existent semantic requirement that the current inventory has failed to satisfy [22].

A planning-oriented recommender would introduce an intermediate decision layer before item selection. Rather than directly asking which item to be shown, it would first ask: **what should the next exposure accomplish?** Semantic IDs may make this intermediate layer more tractable: raw IDs provide little semantic structure that can be transferred across items, while continuous embeddings are harder to use as an explicit and legible vocabulary for planning targets. By contrast, structured semantic IDs can provide a more grounded representation through which such targets can be expressed and through which failures of current inventory to satisfy them can be surfaced more clearly. In the hotel example, the answer may be that the user no longer needs further destination discovery and instead needs reassurance during the booking process. The next exposure may therefore need to emphasize flexibility, reduce perceived risk, or help the user compare low-commitment options. This objective is grounded in the user's current need, but it may also be shaped by platform goals and provider-side content. Semantic planning, therefore, makes the purpose of the next exposure explicit before deciding how that purpose should be instantiated. This distinction separates semantic planning from semantic retrieval as depicted in Figure 1.

Semantic retrieval maps the current user state to an identifier that resolves to a specific item, while semantic planning aligns the current context with a semantic target specifying the role, intent, or desired properties of the subsequent exposure. The system will only instantiate the target as a hotel, an offer, a sponsored message, or another form of content once the target has been determined. This also differs from conversational recommendation [11], which requires explicit dialogue to surface needs, and from intent modeling, which infers user goals but maps them directly to item scores. In semantic planning, the intermediate target is explicitly represented and optimized as a multi-stakeholder decision variable, not merely a latent state before item selection. Within this framework, the semantic structure learned from item IDs may serve as the basis for an intermediate control layer [43]. Such a layer can represent targets such as flexible city-center accommodation, low-risk options for hesitant bookers, or family-friendly alternatives near public transport before mapping them to catalog items. It is important to note that separating the target from the item also enables the system to identify instances in which the current inventory cannot satisfy the inferred need [41].

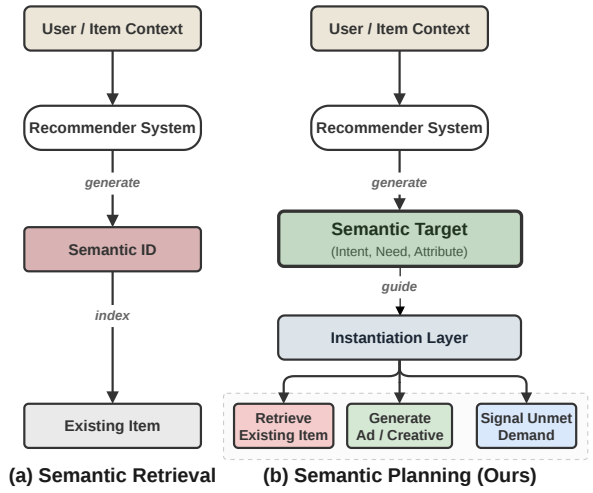


Figure 1: Paradigm shift from semantic retrieval to semantic planning. (a) Semantic Retrieval maps the context directly to a semantic ID to fetch an existing catalog item; (b) Semantic Planning separates the semantic target from the instantiation.

4.2 From Semantic Targets to Instantiation

Once recommender systems can represent what the next exposure should accomplish, recommendations are no longer limited to selecting the closest existing item. The system can instantiate the semantic target through items, messages, offers, or generated creatives, making recommendations a more flexible mechanism for coordinating users, providers, and platforms.

The most direct form of instantiation remains item selection. Once the system predicts a semantic target, it can retrieve the item or offer that best realizes that target. In the hotel example, a target such as low-risk city-center accommodation for hesitant bookers may be instantiated as a hotel with free cancellation, a flexible booking package, or a sponsored listing that clearly highlights refund protection. This differs from ordinary retrieval because the system no longer just does user-item matching. It first makes explicit what the next exposure is intended to achieve, then searches the item, offer, or message that best serves that purpose. Even when the final output is still an item, the intermediate target gives the system a clearer basis for comparing alternative instantiations.

A more flexible form of instantiation becomes possible when the system can generate or adapt the exposure's content. This is especially natural in advertising, where the effectiveness of a recommendation depends not only on which product or service is shown, but also on how it is presented. If the inferred target is to reassure hesitant travelers who want flexible booking options, the system may instantiate that target not only by selecting a hotel with free cancellation, but also by generating a personalized sponsored message that emphasizes low-risk booking, or producing a creative that frames the offer around flexibility and trust. In this setting, generation is no longer limited to producing an identifier for retrieval. It is used to realize or generate a semantic target in the form of a concrete message or creative.

This target-to-instantiation view also makes the multi-stakeholder nature of recommender systems more explicit [1]. For users, semantic targets can lead to exposures better reflecting their current needs or uncertainties, rather than merely presenting the nearest available item. For providers or advertisers, the same target can reveal how an offering should be framed, which user concern should be addressed, or which product attribute should be made more salient. For platforms, target-level information can help allocate inventory, sponsored opportunities, and creative resources more effectively. Ultimately, the system can decide whether a target is best realized through an item, an offer, a message, or a generated creative. When no existing item satisfies the inferred target, this gap itself becomes a signal of unmet demand. Thus, recommender systems evolve from engines for retrieving items into systems for translating semantic objectives into concrete exposures that coordinate information, inventory, and stakeholder value at scale.

4.3 Evaluation and Industrial Deployment

As recommendation shifts towards semantic planning, existing evaluation protocols based on the Cranfield paradigm, which assumes fixed test collections and static relevance judgments, become insufficient [42]. The gold standard must move from surface-level hits toward **planning resonance**: the degree to which a system's trajectory of decisions navigates users toward fulfilled intent [46]. This requires sophisticated "evaluation agents" for user simulation and long-term goal tracking [54]. Crucially, this does not replace long-term proxies like user retention; rather, it provides the missing causal link by explaining *why* an exposure sequence succeeded or failed to align with user and stakeholder intent. Realizing this vision requires resolving significant open challenges:

- **ID stability**: How should semantic IDs be designed to remain consistent across dynamic catalog updates without breaking downstream planning models?
- **Target grounding**: How can the planning layer be constrained to avoid producing targets that are semantically coherent but unsatisfiable by current inventory or platform infrastructure?
- **Objective observability**: How can multi-stakeholder planning quality be supervised and evaluated when provider and platform objectives are only partially observable?

Despite these hurdles, the industrial implications are profound. The platform evolves from a passive traffic distributor into a central agent for interpreting and aggregating demand. By feeding semantic signals of unmet needs back to providers, platforms enable a deeper strategic coupling that goes beyond short-term conversion. This also opens new deployment directions: semantic targets can serve as a shared interface between ranking, advertising, and content systems, allowing heterogeneous platform components to coordinate around a common exposure objective rather than optimizing independently. In this sense, semantic planning opens a path toward recommender systems where every exposure is a purposeful step toward aligning user intent with ecosystem capabilities.

5 Conclusion

One useful way to read the history of recommender systems is through how information is made usable at scale. For two decades,

raw IDs provided the necessary anchors for operational stability and exact memorization. A central trend today is that semantic information is moving from the periphery of item identity closer to the identity layer itself. This transition, however, should not be viewed as a clean break. We anticipate an era of architectural hybridity: Raw IDs will remain essential for auditability, system-of-record requirements, and high-fidelity memorization of head items, while semantic IDs may increasingly serve as a structured vocabulary for unifying information and tasks, and for supporting semantic planning under competing stakeholder objectives. Retrieval-centric designs will remain dominant where stability, interpretability, and strict output control are primary concerns. The core challenge for researchers and practitioners is therefore not to advocate for a single form of identity, but to orchestrate a co-design of semantic structure, planning objectives, and multi-stakeholder objectives.

As recommender systems evolve from passive matching engines to active decision-making systems, the success of the field will be defined by how well we balance the rigid reliability of the past with the generative, intent-driven flexibility of the future. The move toward semantic IDs is not merely a technical optimization of retrieval. It is also an important step toward semantic planning: a paradigm in which systems first determine what an exposure should accomplish, and only then decide how best to realize it.

References

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Augusto Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Model. User Adapt. Interact.* 30, 1 (2020), 127–158.

[2] Gediminas Adomavicius, Bamshad Mobasher, Francesco Ricci, and Alexander Tuzhilin. 2011. Context-Aware Recommender Systems. *AI Mag.* 32, 3 (2011), 67–80.

[3] Honghui Bao, Wenjie Wang, Xinyu Lin, Fengbin Zhu, Teng Sun, Fuli Feng, and Tat-Seng Chua. 2025. Heterogeneous User Modeling for LLM-based Recommendation. In *RecSys*. ACM, 145–154.

[4] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & Deep Learning for Recommender Systems. In *DLRS@RecSys*. ACM, 7–10.

[5] Akira Fukumoto, Aghiles Salah, Sarthak Shrivastava, Alexandru Tatar, Yannick Schwartz, Vincent Michel, and Lee Xiong. 2025. Contrastive Conditional Embeddings for Item-based Recommendation at E-commerce Scale. In *RecSys*. ACM, 931–934.

[6] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. In *IJCAI*. ijcai.org, 1725–1731.

[7] Ruining He, Lukasz Heldt, Lichan Hong, Raghunandan H. Keshavan, Shifan Mao, Nikhil Mehta, Zhengyang Su, Alicia Tsai, Yueqi Wang, Shao-Chuan Wang, Xinyang Yi, Lexi Baugher, Baykal Cakici, Ed H. Chi, Cristos Goodrow, Ningren Han, He Ma, Rómer Rosales, Abby Van Soest, Devansh Tandon, Su-Lin Wu, Weilong Yang, and Yilin Zheng. 2026. PLUM: Adapting Pre-trained Language Models for Industrial-scale Generative Recommendations. In *WWW*. ACM, 8093–8104.

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

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

[10] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*. ACM, 173–182.

[11] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian J. McAuley. 2023. Large Language Models as Zero-Shot Conversational Recommenders. *CoRR* abs/2308.10053 (2023).

[12] Yupeng Hou, Zhankui He, Julian J. McAuley, and Wayne Xin Zhao. 2023. Learning Vector-Quantized Item Representation for Transferable Sequential Recommenders. In *WWW*. ACM, 1162–1171.

[13] Peiyu Hu, Wayne Lu, and Jia Wang. 2026. From IDs to Semantics: A Generative Framework for Cross-Domain Recommendation with Adaptive Semantic Tokenization. In *AAAI*. AAAI Press, 14874–14882.

[14] Dietmar Jannach and Himan Abdollahpouri. 2023. A survey on multi-objective recommender systems. *Frontiers in Big Data* 6 (2023), 1157899. doi:10.3389/fdata.2023.1157899

[15] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender systems: an introduction*. Cambridge University Press.

[16] Jie Jiang, Xinxun Zhang, Enming Zhang, Yuling Xiong, Jun Zhang, Jingwen Wang, Huan Yu, Yuxiang Wang, Hao Wang, Xiao Yan, and Jiawei Jiang. 2026. End-to-End Semantic ID Generation for Generative Advertisement Recommendation. *CoRR* abs/2602.10445 (2026).

[17] Jinqiu Jin, Yang Zhang, Fuli Feng, and Xiangnan He. 2025. Generative Multi-Target Cross-Domain Recommendation. *CoRR* abs/2507.12871 (2025).

[18] Kyungho Kim, Sunwoo Kim, Geon Lee, Jinhong Jung, and Kijung Shin. 2025. Multi-behavior Recommender Systems: A Survey. In *PAKDD (4) (Lecture Notes in Computer Science)*. Springer, 435–452.

[19] Roger Zhe Li. 2023. Metric Optimization and Mainstream Bias Mitigation in Recommender Systems. arXiv:2311.06689 [cs.IR] https://arxiv.org/abs/2311.06689

[20] Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is ChatGPT a Good Recommender? A Preliminary Study. *CoRR* abs/2304.10149 (2023).

[21] Jiawei Liu, Chuan Shi, Cheng Yang, Zhiyuan Lu, and Philip S. Yu. 2022. A survey on heterogeneous information network based recommender systems: Concepts, methods, applications and resources. *AI Open* 3 (2022), 40–57.

[22] Ruochen Liu, Hao Chen, Yuanchen Bei, Qijie Shen, Fangwei Zhong, Senzhang Wang, and Jianxin Wang. 2024. Fine Tuning Out-of-Vocabulary Item Recommendation with User Sequence Imagination. In *NeurIPS*.

[23] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2011. Content-based Recommender Systems: State of the Art and Trends. In *Recommender Systems Handbook*. Springer, 73–105.

[24] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning Disentangled Representations for Recommendation. In *NeurIPS*. 5712–5723.

[25] Tong Man, Huawei Shen, Xiaolong Jin, and Xueqi Cheng. 2017. Cross-Domain Recommendation: An Embedding and Mapping Approach. In *IJCAI*. ijcai.org, 2464–2470.

[26] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep Learning Recommendation Model for Personalization and Recommendation Systems. *CoRR* abs/1906.00091 (2019).

[27] Gustavo Penha, Edoardo D'Amico, Marco De Nadai, Enrico Palumbo, Alexandre Tamborrino, Ali Vardasbi, Max Lefarov, Shawn Lin, Timothy Christopher Heath, Francesco Fabbri, and Hugues Bouchard. 2025. Semantic IDs for Joint Generative Search and Recommendation. In *RecSys*. ACM, 1296–1301.

[28] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Mahesh Sathiamoorthy. 2023. Recommender Systems with Generative Retrieval. In *NeurIPS*.

[29] Aghiles Salah and Hady W. Lauw. 2018. A Bayesian Latent Variable Model of User Preferences with Item Context. In *IJCAI*. ijcai.org, 2667–2674.

[30] Aghiles Salah, Thanh-Binh Tran, and Hady W. Lauw. 2021. Towards Source-Aligned Variational Models for Cross-Domain Recommendation. In *RecSys*. ACM, 176–186.

[31] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. Hidden Technical Debt in Machine Learning Systems. In *NIPS*. 2503–2511.

[32] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and Philip S. Yu. 2017. A Survey of Heterogeneous Information Network Analysis. *IEEE Trans. Knowl. Data Eng.* 29, 1 (2017), 17–37.

[33] Teng Shi, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Enyun Yu. 2025. GenSAR: Unifying Balanced Search and Recommendation with Generative Retrieval. In *RecSys*. ACM, 124–134.

[34] Alex Shtoff, Michael Viderman, Naama Haramaty-Krasne, Oren Somekh, Ariel Raviv, and Tularam Ban. 2024. Low Rank Field-Weighted Factorization Machines for Low Latency Item Recommendation. In *RecSys*. ACM, 238–246.

[35] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Prakash Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer Memory as a Differentiable Search Index. In *NeurIPS*.

[36] Quoc-Tuan Truong, Aghiles Salah, and Hady W. Lauw. 2021. Multi-Modal Recommender Systems: Hands-On Exploration. In *RecSys*. ACM, 834–837.

[37] Moshe Unger, Alexander Tuzhilin, and Amit Livne. 2020. Context-Aware Recommendations Based on Deep Learning Frameworks. *ACM Trans. Manag. Inf. Syst.* 11, 2 (2020), 8:1–8:15.

[38] Chong Wang and David M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *KDD*. ACM, 448–456.

[39] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative Deep Learning for Recommender Systems. In *KDD*. ACM, 1235–1244.

[40] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & Cross Network for Ad Click Predictions. In *ADKDD@KDD*. ACM, 12:1–12:7.

[41] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2023. Generative Recommendation: Towards Next-generation Recommender Paradigm. *CoRR* abs/2304.03516 (2023).

[42] Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. In *EMNLP*. Association for Computational Linguistics, 10052–10065.

[43] Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Yanbin Lu, Xiaojiang Huang, and Yingzhen Yang. 2024. RecMind: Large Language Model Powered Agent For Recommendation. In *NAACL-HLT (Findings) (Findings of ACL)*. Association for Computational Linguistics, 4351–4364.

[44] Yuhao Wang, Junwei Pan, Xinhang Li, Maolin Wang, Yuan Wang, Yue Liu, Dapeng Liu, Jie Jiang, and Xiangyu Zhao. 2025. Empowering Large Language Model for Sequential Recommendation via Multimodal Embeddings and Semantic IDs. In *CIKM*. ACM, 3209–3219.

[45] Yi Xu, Moyu Zhang, Chenxuan Li, Zhihao Liao, Haibo Xing, Hao Deng, Jinxin Hu, Yu Zhang, Xiaoyi Zeng, and Jing Zhang. 2026. MMQ: Multimodal Mixture-of-Quantization Tokenization for Semantic ID Generation and User Behavioral Adaptation. In *WSDM*. ACM, 788–797.

[46] Se-eun Yoon, Zhankui He, Jessica Maria Echterhoff, and Julian J. McAuley. 2024. Evaluating Large Language Models as Generative User Simulators for Conversational Recommendation. In *NAACL-HLT*. Association for Computational Linguistics, 1490–1504.

[47] Feng Yuan, Lina Yao, and Boualem Benatallah. 2019. DARec: Deep Domain Adaptation for Cross-Domain Recommendation via Transferring Rating Patterns. In *IJCAI*. ijcai.org, 4227–4233.

[48] Tianzi Zang, Yanmin Zhu, Haobing Liu, Ruohan Zhang, and Jiadi Yu. 2023. A Survey on Cross-domain Recommendation: Taxonomies, Methods, and Future Directions. *ACM Trans. Inf. Syst.* 41, 2 (2023), 42:1–42:39.

[49] Weizhi Zhang, Yuanchen Bei, Liangwei Yang, Henry Peng Zou, Peilin Zhou, Aiwei Liu, Yinghui Li, Hao Chen, Jianling Wang, Yu Wang, Feiran Huang, Sheng Zhou, Jiajun Bu, Allen Lin, James Caverlee, Fakhri Karray, Irwin King, and Philip S. Yu. 2025. Cold-Start Recommendation towards the Era of Large Language Models (LLMs): A Comprehensive Survey and Roadmap. *CoRR* abs/2501.01945 (2025).

[50] Jujia Zhao, Zihan Wang, Shuaiqun Pan, Suzan Verberne, and Zhaochun Ren. 2026. Unifying Search and Recommendation in LLMs via Gradient Multi-Subspace Tuning. *CoRR* abs/2601.09496 (2026).

[51] Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. Joint Deep Modeling of Users and Items Using Reviews for Recommendation. In *WSDM*. ACM, 425–434.

[52] Guorui Zhou, Xiaoqiang Zhu, Chengru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep Interest Network for Click-Through Rate Prediction. In *KDD*. ACM, 1059–1068.

[53] Jing Zhu, Mingxuan Ju, Yozen Liu, Danai Koutra, Neil Shah, and Tong Zhao. 2025. Beyond Unimodal Boundaries: Generative Recommendation with Multimodal Semantics. *CoRR* abs/2503.23333 (2025).

[54] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2025. A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems. In *WWW*. ACM, 4653–4661.