Cold-Start Audiobook Recommendation via Cross-Domain Sub-Tower Fusion

Kirandeep Kaur*
University of Washington
Seattle, United States
kaur13@cs.washington.edu

Amit Goyal
Amazon Music
San Francisco, United States
goyalam@amazon.com

Abstract

For music streaming services expanding into audiobooks, cold-start personalization presents a critical challenge: as audiobooks are a newly introduced content type, the vast majority of existing users have no audiobook listening history. This domain-level cold-start scenario differs from traditional item or user cold-start scenarios, since personalization must begin before any behavioral data exists in the target domain. Yet these same users possess rich engagement histories in the platform's established offerings of music and podcasts, creating an opportunity to transfer cross-modal signals for early-stage audiobook recommendations. We present a lightweight framework designed for scalability and minimal retraining, showing that cross-modal transfer can yield strong personalization even in sparse domains. Our framework, studied in the context of a large-scale music streaming service, adopts a two-tower design with two key design choices: (1) the user side is frozen and structured into modality-specific sub-towers, preserving signals without retraining overhead; and (2) an adaptive fusion mechanism integrates these signals, while the item side learns audiobook embeddings. To further enrich content representations, we incorporate BAAI's BGE model for text encoding, which injects semantic knowledge into the towers. This combination yields consistent and substantial relative gains: offline precision exceeds +100% over popularity baselines and +50% over single-domain based collocation methods, with strong complementarity between modalities. Our method scales to millions of users with minimal training cost and generalizes to public datasets, enabling both open research and industrial adoption. Large-scale A/B testing in the US marketplace demonstrates a ∼10% improvement in first audiobook listens compared to popularity baselines. These results demonstrate that frozen multi-modal sub-towers with pretrained text enrichment offer a principled alternative for cross-domain cold-start personalization, providing a generalizable architecture for efficient content expansion across any streaming platform diversifying into new media types.

CCS Concepts

• **Information systems** → **Personalization**; **Recommender systems**.

*Work done during internship at Amazon Music.

WSDM Companion '26, Boise, ID, USA

ACM ISBN 979-8-4007-2358-2/2026/02
https://doi.org/10.1145/3779211.3793160

Keywords

cross-domain recommendations, cold-start, multi-modal fusion, frozen user towers, large-scale retrieval

ACM Reference Format:
Kirandeep Kaur and Amit Goyal. 2026. Cold-Start Audiobook Recommendation via Cross-Domain Sub-Tower Fusion. In *The Nineteenth ACM International Conference on Web Search and Data Mining (WSDM Companion '26), February 22–26, 2026, Boise, ID, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3779211.3793160

1 Introduction

Recommender systems have transformed how users discover digital content, yet personalization remains especially difficult when platforms expand into an entirely new content domain. Audiobooks exemplify this *domain-level cold-start* scenario: as long-form, high-commitment media often hidden behind paywalls or limited previews, they generate sparse engagement signals and discourage exploratory consumption [31]. When such a domain is newly introduced, the vast majority of users arrive with no interaction history, making traditional collaborative or content-based methods ineffective [12, 25]. Meanwhile, leading platforms are rapidly converging across modalities—Spotify integrating podcasts and audiobooks [2], and Netflix experimenting with interactive games [24]—where the majority of the users possess rich interaction trails in established ecosystems but none in the new one. This transition creates both a challenge and an opportunity: how can we transfer behavioral and semantic signals from mature modalities such as music and podcasts to personalize recommendations *before any audiobook interactions exist*?

To study this problem at scale, we use behavioral logs from a large commercial audio platform containing listening histories across music, podcasts, and newly launched audiobooks. These interactions naturally form a heterogeneous cross-domain structure: music captures long-horizon stylistic and affective cues, while podcasts provide high-precision topical and narrative signals. Notably, music activity is nearly universal among audiobook listeners, whereas podcast engagement is sparser but semantically richer. Figure 1 illustrates the dominant pathways in this data, and their complementary fusion, highlighting the value of leveraging multiple modalities for early-stage personalization.

Recent advances in cross-domain recommendations highlight the potential of leveraging behavioral overlap across modalities. For example, Spotify's 2T-HGNN [2] system integrates audiobook and podcast co-listening graphs to mitigate sparsity and improve quality for cold-start users. However, such graph-based methods presuppose an established cross-domain ecosystem—where users already interact with both modalities—making them less suitable for the *domain-level cold-start* setting. Music, due to its ubiquity and

stable long-term patterns, and podcasts, due to their strong semantic signals, offer complementary behavioral structure that can be utilized without relying on graph construction.

We introduce **MAP-2T**, a cross-domain two-tower framework designed for the *domain-level cold-start* setting, where personalization must be bootstrapped from existing modalities. On the user side, each modality—music and podcasts—is processed by a frozen sub-tower that calibrates embedding scale and distribution, preserving pre-learned behavioral patterns without retraining. An adaptive feature-gated fusion mechanism then balances long-term stylistic preferences captured by music with topical and narrative coherence derived from podcasts, yielding richer and more stable user representations. On the item side, audiobook embeddings are learned through a lightweight tower and semantically enriched with a pretrained text encoder to expand coverage when user–item co-listening data are sparse. These design choices are motivated by industrial deployment requirements: low retraining overhead, robustness to missing modalities, and bounded inference latency at platform scale.

We show that incorporating both music and podcast signals within the user tower produces unified embeddings that capture complementary aspects of listening behavior, yielding more robust user representations and consistently improving audiobook recommendations. In summary, this work makes four primary contributions: 1) We propose a cross-domain two-tower framework, **MAP-2T**, that leverages pretrained music and podcast embeddings to address the *domain-level cold-start* problem in audiobook recommendation. 2) We design a feature-gated fusion mechanism that dynamically balances long-term stylistic preferences from music with topical coherence from podcasts, outperforming naive concatenation by **+6.7% Precision@3**. 3) Through large-scale offline evaluation on millions of real-world interactions, we demonstrate substantial gains: across multi-signal user cohorts, **MAP-2T** improves Precision@3 by over **+100%** relative to popularity baselines and remains robust under missing or imputed modalities. 4) Our online A/B test evaluation, conducted in the US marketplace with millions of customers, demonstrates a ∼10% incremental improvement in first audiobook listens relative to the popularity-based baseline.

2 Related Work

Personalization across emerging content domains draws upon a long line of research in recommender systems, from early work on transfer learning for cold-start users to recent advances in cross-modal and large-scale architectures. Yet, most existing studies assume at least minimal user interaction in the target domain, leaving the *domain-level user cold-start* problem largely unexplored. This section reviews prior efforts that inform our approach.

Cross-domain recommendation (CDR) leverages behavioral or content signals from related domains to mitigate sparsity [20]. Early approaches such as Collective Matrix Factorization (CMF) [26] share latent factors across domains with overlapping entities, while later neural models incorporate cross-layer connections or joint metric constraints to enable bidirectional feature transfer [6, 8, 16, 17]. Graph-based extensions propagate signals through domain graphs [14, 32], and adversarial methods aim to learn domain-invariant embeddings by reducing distribution gaps [7, 15]. These methods nevertheless assume shared users, shared items, or established interaction structures. **MAP-2T** instead targets the stricter case where the audiobook domain is newly introduced with no interactions at all, constructing representations solely from pretrained music and podcast encoders.

A closely related line of research focuses on using transfer learning to handle cold-start users who lack historical interactions in a given domain. Traditional collaborative and content-based approaches [12, 25] struggle in this regime due to the absence of explicit feedback. Neural transfer learning methods such as EMCDR [17] and TMCDR [33] learn an explicit mapping function from a source-domain user embedding to a target-domain embedding, using overlapping users as anchors. Subsequent models—PTUPCDR [34] and SSCDR [10]—extend this paradigm through meta-learning or semi-supervised objectives to better generalize across users. Meta-learning approaches like MeLU [13] and MAMO [5] adaptively fine-tune user models in a few-shot setting, but these methods incur significant retraining costs and rely on user-specific adaptation. **MAP-2T** departs from these frameworks in both design and assumption. Instead of learning a transfer function or meta-network, it employs frozen modality-specific user towers pretrained on mature domains (music and podcasts) and fuses their outputs via a lightweight gating mechanism. This design enables zero-shot transfer to the audiobook domain without requiring overlapping users, auxiliary adaptation data, or per-user fine-tuning—making it more scalable and better aligned with industrial scale.

Unlike short-form content like music or video clips, audiobooks and long-form media pose unique challenges due to sparse engagement patterns and higher user commitment [18, 31]. These properties make learning user preferences from co-consumption patterns difficult. Spotify's 2T-HGNN model [2] addresses this by constructing a heterogeneous item graph linking podcasts and audiobooks, jointly optimized with a two-tower retrieval framework. Parallel efforts in podcast recommendation [30] and contextual long-form retrieval [3, 11] emphasizes the importance of intent understanding and narrative structure. **MAP-2T** extends this line of inquiry to an even earlier stage—*the pre-graph regime*—where a new domain (e.g., audiobooks) lacks sufficient co-listening data to construct reliable item graphs or fine-tune retrieval models.

3 MAP-2T: Cross-Domain Two-Tower Architecture for Cold-Start Audiobook Recommendation

This section presents **MAP-2T**, a modular dual-tower retrieval model that transfers user preference signals from mature modalities (music, podcasts) into the audiobook domain.

3.1 Problem Formulation

We formalize cross-domain audiobook recommendation as a *representation retrieval* problem, where user preferences inferred from established modalities (music and podcasts) are transferred to a newly introduced content domain (audiobooks). Let $\mathcal{U}$ denote the set of users and $\mathcal{A}$ the catalog of audiobooks. Our goal is to learn:

$$s(u, a) = f_\theta(u, a) \in \mathbb{R},$$

(a) Music → Audiobook (Stylistic Continuity): User engaging with emotionally resonant music such as *"Fix You" by Coldplay* often transition to audiobooks with similar affective tone, e.g., *"It Ends With Us" by Colleen Hoover*. This reflects a stylistic pathway where mood and emotional intensity guide cross-domain recommendation.

(b) Podcast → Audiobook (Semantic Continuity): Listeners of topic-focused podcasts such as *The Huberman Lab* (neuroscience and wellness) show higher affinity for semantically related audiobooks like *"Why We Sleep" by Matthew Walker, PhD*. This represents semantic continuity based on conceptual overlap.

(c) Combined Music + Podcast → Audiobook (Complementary Fusion)

Figure 1: Cross-domain behavioral pathways connecting music, podcasts, and audiobooks demonstrate complementary fusion, where affective and semantic signals jointly predict self-improvement audiobooks like *"Atomic Habits" by James Clear* or *"The Subtle Art of Not Giving a F*ck" by Mark Manson*. Together, these panels highlight how stylistic, semantic, and multimodal cues collectively motivate cross-domain recommendation in MAP-2T.

that predicts ratings for audiobooks $a \in \mathcal{A}$ for user u such that those most aligned with the user's cross-modal preferences receive higher scores.

User–modality interactions. Each user u is associated with a history of items consumed in two auxiliary modalities:

$$\mathcal{H}_u^{(m)} = \{i_1^{(m)}, i_2^{(m)}, \ldots, i_{n_u^{(m)}}^{(m)}\}, \quad \mathcal{H}_u^{(p)} = \{i_1^{(p)}, i_2^{(p)}, \ldots, i_{n_u^{(p)}}^{(p)}\},$$

from which we extract *top-k* engagement signals (e.g., most listened artists or most played podcasts). Each entity has a pretrained embedding $\mathbf{x}_i^{(m)}! \in !\mathbb{R}^{d_m}$ or $\mathbf{x}_i^{(p)}! \in !\mathbb{R}^{d_p}$. We obtain compact modality representations via mean aggregation:

$$\mathbf{x}_u^{(m)} = \frac{1}{k_m} \sum_{i=1}^{k_m} \mathbf{x}_i^{(m)}, \qquad \mathbf{x}_u^{(p)} = \frac{1}{k_p} \sum_{i=1}^{k_p} \mathbf{x}_i^{(p)}.$$

Together, they form a compact behavioral summary of user identity across modalities.

Audiobook representation. Each audiobook $a \in \mathcal{A}$ is initialized with a pretrained text embedding $\mathbf{x}_a \in \mathbb{R}^d$ derived from its metadata (title, author, and description). These embeddings serve as semantic anchors and are refined during training to align with cross-modal user representations.

Scoring and retrieval objective. The model maps users and audiobooks to a shared latent space and computes preference via dot-product::

$$s(u, a) = \mathbf{u}_u^\top \mathbf{v}_a.$$

Training uses positive audiobook engagements with a temperature-scaled softmax loss [4, 29] to align users with relevant audiobooks and push away non-relevant ones. This objective encourages user representations to align with the audiobooks they engage with and diverge from non-relevant ones.

Domain-level cold-start setting: Unlike standard cross-domain recommendation [8, 17], our setting is a stricter *domain-level cold-start* case: (i) users have no audiobook history, (ii) music and podcast encoders are pretrained independently with no parameter sharing, and (iii) alignment across modalities must be learned solely from audiobook feedback.

Design choices: We bootstrap personalization entirely from existing cross-modal behavior. We (1) freeze pretrained music/podcast towers and train only lightweight projection and gating layers to preserve behaviors and reduce cost, (2) first calibrate modality embeddings via affine alignment and normalization before fusing them to avoid scale mismatch, and (3) explicitly model missing-modality cases through imputation so the same system serves music-only, podcast-only, and multi-signal users.

3.2 Architecture Overview

MAP-2T adopts a dual-tower retrieval design (Figure 2), consisting of a *User Tower* that aggregates multi-modal user representations and an *Item Tower* that refines audiobook representations for retrieval alignment within a shared latent space $\mathbb{R}^d$. All modules are lightweight and operate on pretrained embeddings, enabling scalable deployment in production environments [4, 19].

3.2.1 User Tower: Modality-Aware Fusion. Users express complementary behavioral patterns across music and podcast modalities. The User Tower processes these two signals through dedicated towers before integrating them via a feature-gated fusion mechanism.

Music-side user representation. The music-domain input $x^{(m)} \in \mathbb{R}^{d_m}$ encodes a user's aggregate stylistic profile. To align it with the shared space, a linear projection followed by feature-wise calibration

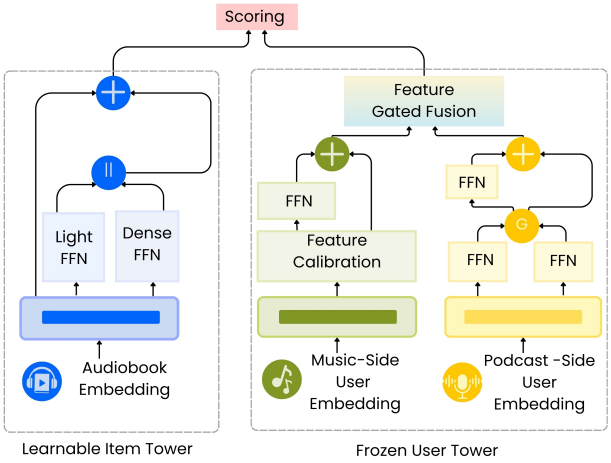

Figure 2: MAP-2T Architecture: The framework consists of two-tower setup: a *User Tower* (left) that integrates modality-specific user representations and an *Item Tower* (right) that refines audiobook embeddings for alignment in a shared latent space $\mathbb{R}^d$.

is applied:

$$z^{(m)} = W_m x^{(m)} + b_m, \qquad \tilde{z}^{(m)} = \gamma^{(m)} \odot z^{(m)} + \beta^{(m)},$$

where $W_m \in \mathbb{R}^{d \times d_m}$ and $\gamma^{(m)}, \beta^{(m)} \in \mathbb{R}^d$ are learnable affine parameters. A residual feed-forward network refines the calibrated features:

$$u^{(m)} = \mathrm{LN}\Big(\tilde{z}^{(m)} + f_{\mathrm{FFN}}(\tilde{z}^{(m)})\Big).$$

This design, conceptually similar to [23], corrects for scale and shift mismatches between pretrained spaces while preserving long-horizon stylistic identity.

Podcast-side user representation. Podcast-derived representations $x^{(p)} \in \mathbb{R}^{d_p}$ capture users' topical focus and cognitive intent—what subjects they choose to listen to and how they engage with them. To retain this semantic richness while enabling flexible adaptation, we process $x^{(p)}$ through two parallel feed-forward transformations that capture complementary views of the signal:

$$h^{(1)} = f_{\mathrm{FFN}_1}(x^{(p)}), \qquad h^{(2)} = f_{\mathrm{FFN}_2}(x^{(p)}).$$

A learnable gating mechanism determines, for each feature dimension, how to combine these two perspectives. The gate is computed as:

$$g = \sigma\big(f_{\mathrm{gate}}(x^{(p)})\big),$$

where f_{gate} is a lightweight feed-forward transformation followed by a sigmoid activation that outputs weights $g \in [0,1]^{d_p}$. Each gate dimension thus reflects how much the model should rely on the abstract path $h^{(1)}$ versus the concrete path $h^{(2)}$. The gated combination is then obtained as:

$$\tilde{u}^{(p)} = g \odot h^{(1)} + (1 - g) \odot h^{(2)}.$$

The blended representation is refined through a residual normalization layer:

$$u^{(p)} = \mathrm{LN}\Big(\tilde{u}^{(p)} + f_{\mathrm{FFN}}(\tilde{u}^{(p)})\Big).$$

Intuition. This mixture-of-paths formulation functions as a soft mixture-of-experts [8, 9], where the gating vector performs fine-grained, per-dimension arbitration between different semantic abstractions. By dynamically weighting these transformations, the model can emphasize topical coherence for certain listeners (e.g., science or business enthusiasts) while maintaining generalization for more stylistically diverse users.

Feature-gated fusion. The modality-specific vectors $u^{(m)}, u^{(p)} \in \mathbb{R}^d$ are combined through an adaptive, per-feature gate that captures complementarity and disagreement between modalities:

$$z = [\, u^{(m)} \,\|\, u^{(p)} \,\|\, |u^{(m)} - u^{(p)}| \,\|\, u^{(m)} \odot u^{(p)} \,],$$

$$\alpha = \sigma\big(W_2 \phi(W_1 z + b_1) + b_2\big), \quad \alpha \in [0,1]^d,$$

$$u = \alpha \odot u^{(m)} + (1 - \alpha) \odot u^{(p)}.$$

The absolute difference term accentuates divergence, while the Hadamard product rewards alignment, allowing the fusion gate α to prioritize stylistic (music) or semantic (podcast) cues per dimension. Unlike static concatenation or domain-level weighting [16, 17], this formulation performs fine-grained, context-sensitive blending.

When podcast histories are unavailable, we assign a representative embedding corresponding to the most popular podcast in the corpus. This choice is guided by empirical observation: users without explicit podcast engagement tend to exhibit behavioral similarity to those consuming widely popular shows. Using this representative embedding ensures full user coverage while maintaining semantic consistency across the podcast sub-tower, allowing the fusion gate to learn modality interactions without introducing distributional shifts.

3.2.2 Item (Audiobook) Tower. Audiobook representations are initialized from pretrained text embeddings $x^{(ab)} \in \mathbb{R}^d$ based on title and author metadata. These are refined through two complementary feed-forward adapters:

$$h_{\mathrm{dense}} = f_{\mathrm{FFN}_{\mathrm{deep}}}(x^{(ab)}), \qquad h_{\mathrm{light}} = f_{\mathrm{FFN}_{\mathrm{shallow}}}(x^{(ab)}),$$

whose outputs are concatenated and mixed:

$$z = f_{\mathrm{mixer}}([\, h_{\mathrm{dense}} \,\|\, h_{\mathrm{light}} \,]), \qquad v = \mathrm{LN}(z + x^{(ab)}).$$

The shallow branch maintains the pretrained semantic geometry essential for zero-shot generalization, while the denser branch provides controlled adaptation for retrieval performance. Residual addition stabilizes training by constraining updates to low-magnitude deviations from the original embedding [1].

3.2.3 Training Objective. We train the model using a standard two-tower retrieval setup with temperature-scaled cosine similarity and a softmax cross-entropy loss [4, 29]. To improve generalization, we add three lightweight regularizers: (i) *label smoothing* [28] to reduce overconfidence, (ii) a KL divergence term [21, 27] that pulls predictions toward a tempered popularity prior, and (iii) an *entropy bonus* [22] that encourages score diversity and broader coverage. The final objective is:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{retrieval}} + \lambda_{\mathrm{ls}} \mathcal{L}_{\mathrm{smooth}} + \lambda_{\mathrm{kl}} \mathcal{L}_{\mathrm{pop\_reg}} + \lambda_{\mathrm{ent}} \mathcal{L}_{\mathrm{div}},$$

where λ_{ls}, λ_{kl}, and λ_{ent} control the relative contribution of each regularizer.

4 Empirical Evaluation

We evaluate **MAP-2T** across multiple user cohorts, baselines, and ablation settings to quantify the effectiveness of cross-modal alignment. Let $\mathcal{U}$ denote the set of all audiobook users, and $\mathcal{U}_m \subset \mathcal{U}, \mathcal{U}_p \subset \mathcal{U}$ represent users with *music* and *podcast* histories, respectively. We further define:

$$\mathcal{U}_{m\cap p} = \mathcal{U}_m \cap \mathcal{U}_p, \quad \mathcal{U}_{m\setminus p} = \mathcal{U}_m \setminus \mathcal{U}_p,$$

Multi-signal users ($\mathcal{U}_{m\cap p}$) engage with both music and podcasts on the platform. Their histories provide complementary behavioral cues. This subset represents the ideal case for cross-domain recommendation, as both modalities contribute to user representation. Single-signal users ($\mathcal{U}_{m\setminus p}$) have rich music activity but no recorded podcast engagement. In practical terms, they reflect the majority cold-start population, where cross-domain information is partially missing. For such users, we impute the missing podcast embedding with a neutral centroid representation (computed from the most popular podcasts) to preserve dimensional consistency and enable inference.

All results are reported in relative terms with respect to the strongest non-parametric baseline.

5 Experimental Setup

Training and data variants. We train **MAP-2T** on $\mathcal{U}$, the full audiobook customer base, where users without podcast histories ($u \in \mathcal{U}_{m\setminus p}$) are assigned an imputed podcast embedding computed as the centroid of the most frequently streamed podcast representation. This approach preserves embedding dimensionality and allows inference for users with partial modality coverage. During evaluation, we distinguish between two data configurations:

- **MAP-2T(imputed_p):** all users use imputed podcast embeddings, simulating the most challenging cold-start setting where true podcast histories are absent;
- **MAP-2T(full):** only users missing podcast signals are imputed, preserving authentic cross-domain structure for others.

Baselines. We benchmark against two production-oriented baselines representing popularity-driven and co-occurrence-based paradigms.

- **Popularity.** A frequency-based ranking model that always recommends the most consumed audiobooks across the platform.
- **ColloPop (Artist-based Collocation with Popularity Weighting).** For each audiobook a, ColloPop identifies music artists co-consumed by users who also listened to a, constructing an implicit artist–audiobook association matrix. The score for recommending audiobook a to user u is defined as:

$$s(u,a) = \sum_{i\in\mathcal{A}_u} \text{coocc}(i,a) \cdot \text{pop}(a)^{\alpha},$$

where $\mathcal{A}_u$ is the set of artists listened to by u, $\text{coocc}(i,a)$ is their co-occurrence count, and $\text{pop}(a)$ is audiobook popularity with smoothing exponent α.This hybrid heuristic highlights artists strongly linked to an audiobook while using global popularity to reduce long-tail sparsity.

Table 1: Relative improvements (%) over the Popularity baseline across models and metrics at different K.

Models	K	Relative Improvement (%)			
		Precision	Recall	F1 Score	Hit Rate
Collopop-Artist	3	+11.5	+17.2	+15.7	+19.9
MAP-2T (imputed_p)	3	+20.2	+13.6	+16.2	+12.8
MAP-2T (full)	3	**+26.3**	**+28.3**	**+28.8**	**+24.5**
Collopop-Artist	5	+7.6	+4.7	+6.2	+9.2
MAP-2T (imputed_p)	5	+18.9	+1.4	+1.6	–0.1
MAP-2T (full)	5	**+24.2**	**+14.3**	**+19.4**	**+10.8**

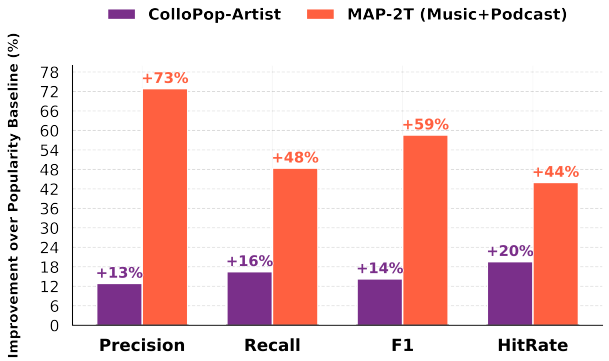

Figure 3: Relative comparison at $K = 3$ for users with both music and podcast histories. Bars represent percentage improvements of MAP-2T over popularity and ColloPop-Artist baselines across Precision, Recall, F1, and Hit Rate.

5.1 Results

Empirical results (Table 1) reveal a consistent pattern: **MAP-2T**(full) delivers the strongest performance across all evaluation metrics and retrieval depths, outperforming both popularity-based and non-parametric co-occurrence baselines. Its gains are most pronounced in precision and recall, indicating that cross-modal alignment reliably surfaces audiobooks aligned with user intent rather than merely popular items. The imputed variant follows the same trend, maintaining sizeable improvements despite missing-modality inputs—evidence that the learned representation space captures transferable behavioral signals rather than overfitting to fully observed cases. Overall, the results show that lightweight fusion provides stable benefits even as candidate sets expand, offering a scalable and modality-robust alternative to harder-to-deploy graph or fully retrained architectures.

5.1.1 Evaluation on Music and Podcast Signals. For users who engage with both music and podcasts, the system has access to its richest behavioral signal: music offers stylistic and mood preferences, while podcasts capture narrative and topical interests. This combined profile allows the recommender to construct a denser, more semantically grounded representation of user intent. As shown in Figure 3, **MAP-2T** leverages this complementary information to achieve consistently large gains over both popularity and ColloPop-Artist across Precision, Recall, F1, and Hit Rate. The pattern indicates that cross-modal fusion is extracting genuinely non-redundant cues

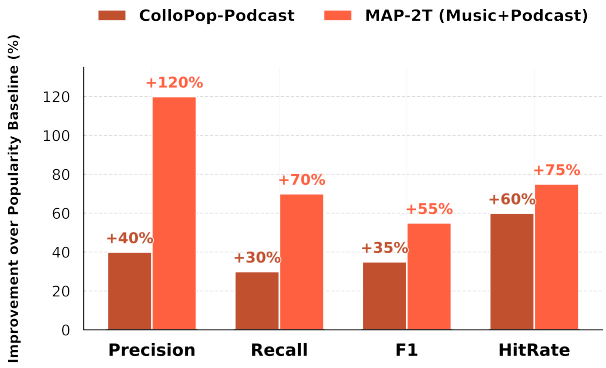

Figure 4: Relative improvements (%) at $K = 3$ for users with both music and podcast histories. MAP-2T(music_podcast) outperforms both the popularity and ColloPop-Podcast baselines across all metrics.

from the two modalities, yielding sharper and more reliable ranking performance than co-occurrence–based approaches.

5.1.2 Expanded Evaluation against Podcast-based ColloPop Model and Impact of Podcast Signals on Training. To precisely quantify the contribution of podcast signals during training, we construct a reduced yet semantically rich dataset containing only users with both *music* and *podcast* histories. To contextualize improvements, we extend the ColloPop framework to a **ColloPop-Podcast** variant that leverages co-listening relationships among podcasts instead of artist co-occurrences. This baseline captures topical similarity in spoken-word media but does not align these signals with musical behavior, providing a strong unimodal reference for comparison.

When trained solely on users with complete multi-modal histories, **MAP-2T**(music_podcast) yields consistent and significant gains across all evaluation metrics. At $K = 3$, it achieves approximately **+120% Precision** and **+70% Recall** improvements over the popularity baseline, while outperforming the podcast-only ColloPop variant by roughly **+55% Precision** and **+30% Recall**. This reveals that podcast signals act as powerful complementary features. When both signals are modeled jointly, the learned embeddings capture *behavioral depth*: a fusion of how users feel and what they seek cognitively. This combination yields semantically coherent user vectors that generalize across diverse audiobook themes and authorial styles.

Training exclusively on dual-signal users also alters the geometry of the learned representation space. From an operational perspective, these gains are noteworthy. **MAP-2T** achieves substantial quality uplift through lightweight multi-modal alignment, requiring no additional supervision or architectural retraining.

5.2 Ablation on Input Modalities

To disentangle the contribution of each modality, we progressively extend from a single-signal baseline to multi-modal variants, quantifying relative improvements in ranking performance across precision, recall, F1, and hit rate metrics.

Starting from Music Signals. Using music-only embeddings establishes a strong baseline given their ubiquity and long-term behavioral stability. Extending this model to incorporate audiobook interactions yields minimal change (Precision: +0.2%), suggesting that architectural augmentation without new behavioral evidence offers limited benefit. However, integrating podcast histories produces notable gains: Precision increases by +13.1%, F1 by +4.7%, and Hit Rate by +2.5%. These improvements indicate that podcast signals inject semantically aligned contextual cues that refine ranking calibration and broaden personalization coverage. In essence, while music embeddings effectively capture enduring taste patterns, their discriminative power plateaus without cross-modal enrichment.

Starting from Podcast Signals. When beginning with podcast-only representations, which are thematically closer to audiobooks, we observe stronger initial alignment and higher precision relative to music-only inputs. Incorporating audiobook data further enhances performance, yielding improvements of +34.6% in Precision, +26.1% in F1, and +16.8% in Hit Rate. Adding music signals amplifies these effects substantially, culminating in +52.6% Precision, +38.5% F1, and +27.7% Hit Rate compared to the podcast-only variant. These results reveal two complementary dynamics: podcasts provide semantic proximity and fine-grained topical alignment with audiobooks, whereas music contributes broad behavioral coverage that enhances recall and robustness. The combination of both allows the model to construct richer, more balanced user representations that generalize effectively across content modalities and intent dimensions.

Together, these findings confirm that modality complementarity is critical for robust cross-domain recommendation: podcasts enhance semantic alignment, while music broadens behavioral scope, and their integration yields superior overall generalization.

5.2.1 Ablation on Fusion Strategies: Concatenation vs. Gated Fusion. Beyond the choice of input modalities, an equally important consideration is *how* to combine them. We compare two fusion strategies: (1) a simple concatenation of modality embeddings, and (2) a learnable gated fusion mechanism that adaptively reweights each modality's contribution on a feature-wise basis.

While both approaches yield comparable overall coverage, gated fusion consistently improves ranking sharpness at the top of the list. Precision@3 rise corresponds to a relative improvement of +6.7%, while HitRate@3 remains nearly unchanged . This pattern suggests that gated fusion enhances the prioritization of relevant items rather than expanding the retrieval set. By learning to amplify predictive dimensions (e.g., podcast features aligned with audiobook semantics) and suppress less informative ones, the gating mechanism enables fine-grained modulation across modalities.

Overall, these results highlight that the benefit of gating lies not in increasing recall, but in improving ranking calibration and confidence—key for practical recommendation scenarios where precision at the top-K directly drives engagement and trust.

6 Conclusion

This work investigated how cross-domain behavioral signals from music and podcasts can be effectively leveraged to enhance audiobook recommendations in cold-start settings. Empirically, we observed that the two modalities contribute complementary strengths: music embeddings capture stable, long-term user preferences that broaden recommendation coverage, while podcast embeddings provide semantically aligned cues that sharpen ranking precision. Together, they enable richer and more discriminative user representations that

transfer effectively across content boundaries. From an industry perspective, these findings highlight the practical utility of embedding alignment and lightweight fusion as scalable mechanisms for improving personalization under data sparsity. A key limitation of this study lies in the absence of a large, publicly available benchmark for cross-modal and audio-centric recommendations. Developing such shared, multimodal datasets would represent an important contribution for the community, enabling reproducible research and standardized evaluation of cross-domain models.

References

[1] Wenjie Bi and et al. 2022. Monolith: Real-Time Recommendation System With Collisionless Embedding Table. In *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[2] Oskar Celma and et al. 2024. Personalized Audiobook Recommendations at Spotify Through Graph Neural Networks. In *Proceedings of the 18th ACM Conference on Recommender Systems*.

[3] Shubham Chinchalikar, Supriya Patil, and Smita Shinde. 2018. Emotion-Based Audiobook Recommendation Using NLP. In *Proceedings of the 2nd International Conference on Intelligent Systems*.

[4] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 191–198.

[5] Cheng Dong, Yaqiong Xie, Yue Wu, and Hui Xiong. 2020. MAMO: Memory-Augmented Meta-Optimization for Cold-start Recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1083–1091.

[6] Ali M. Elkahky, Yang Song, and Xiaodong He. 2015. A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems. In *Proceedings of the 24th International Conference on World Wide Web*. 278–288.

[7] Min Gao, Xiangnan He, Yongfeng Li, and Xiangnan Zhang. 2019. Adversarial training for cross-domain recommendation. In *Proceedings of The Web Conference*.

[8] Guangneng Hu, Yu Zhang, and Qiang Yang. 2018. Conet: Collaborative cross networks for cross-domain recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 667–676.

[9] Nan Jiang, Yifan Cao, Weijie Chen, Peng Chen, and Peng Zhang. 2022. ADIN: Adaptive Domain Interest Network for Multi-Domain Recommendation. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*.

[10] Gyuyoung Kang, Youngjoong Kim, Nosup Park, and Kyoung-Sook Han. 2019. Semi-supervised learning for cross-domain recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

[11] Jun Komesu and et al. 2021. Goal-Aware Long-Form Content Recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*.

[12] Xin Lam, Tho Vu, Minh Le, and Son Duong. 2008. Addressing cold-start in recommender systems: A semi-supervised co-training algorithm. In *Proceedings of the 2008 International Conference on Machine Learning and Applications*. 63–70.

[13] Hoyeop Lee, Jinsung Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. 2019. Melu: Meta-learning user preference for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1073–1082.

[14] Dong Li, Peng Yang, Peng Zhang, Guannan Wang, and Philip S. Yu. 2021. Dual-Target Cross-Domain Recommendation with Heterogeneous Information Networks. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*. 1232–1241.

[15] Dong Li, Peng Yang, Peng Zhang, Guannan Wang, and Philip S. Yu. 2022. GACL: Graph Adversarial Contrastive Learning for Cross-Domain Recommendation. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*. 1359–1368.

[16] Chen Ma, Zheng Ma, Ying Liu, and Jiliang Tang. 2019. Dual metric learning for cross-domain recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1051–1060.

[17] Tong Man, Hua Shen, Senzhang Liu, Xiaolong Jin, and Xueqi Cheng. 2017. Cross-Domain Recommendation: A Neural Transfer Learning Approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 2464–2470.

[18] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Being Accurate Is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*.

[19] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Shi, and et al. 2019. Deep Learning Recommendation Model for Personalization and Recommendation Systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 322–326.

[20] Weike Pan and Qiang Yang. 2010. Transfer learning for cross-domain recommendation. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*.

[21] Donghyun Park and Alexander Tuzhilin. 2019. Accurate and diverse recommendations via item-specific weighting. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 81–89.

[22] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *Proceedings of the International Conference on Learning Representations (ICLR), Workshop Track*.

[23] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (New Orleans, Louisiana, USA) *(AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, Article 483, 10 pages.

[24] Netflix Inc. / Press Release. 2024. Netflix Begins Expansion into Interactive Games. Press / public announcement; fill URL or publisher as needed.

[25] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 253–260.

[26] Ajit P Singh and Geoffrey J Gordon. 2008. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 650–658.

[27] Harald Steck. 2011. Item popularity and recommendation accuracy. In *Proceedings of the 5th ACM Conference on Recommender Systems*. 125–132.

[28] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826.

[29] Xinyang Yi, Liangjie Hong, Eugene Zhong, Yongfeng Shi, Dian Liu, Hongbo Dou, Chih-Jen Song, Yi Wang, and Ed H Chi. 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2130–2138.

[30] Rui Yin, Linhong Chen, Xinyi Fu, and Weinan Zhang. 2022. Modeling Podcast Listening Preferences for Personalized Recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems*.

[31] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep Learning based Recommender System: A Survey and New Perspectives. *Comput. Surveys* 52, 1 (2019), 1–38.

[32] Qi Zhao, Yu Wei, Liang Cao, and Liqiang Nie. 2020. Cross-domain recommendation with graph neural networks. In *Proceedings of The Web Conference*. 2598–2604.

[33] Jing Zhu, Xiangnan He, Dawei Yin, and Peng Cui. 2021. Transfer Meets Hybrid: A Meta-Learning Approach for Cross-Domain Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[34] Jing Zhu, Xuemeng Liu, Xiangnan He, Wayne Xin Zhao, and Peng Cui. 2022. Personalized Transfer for Cross-Domain Recommendation. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*.