OT-FLOWREC: OPTIMAL TRANSPORT-GUIDED MULTI-BRIDGE FLOW FOR MULTIMODAL SEQUENTIAL RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

E-commerce user preferences evolve dynamically within sessions, exhibiting concurrent and multi-centric interest distributions. While existing multimodal sequential recommendation methods leverage global attention to compress historical interactions, they inadequately capture the concurrent nature of multi-interest behaviors and fail to model the underlying multi-centric distribution of user intent. To tackle this issue, we propose OT-FlowRec, an optimal transport-guided multi-bridge flow matching framework that explicitly models multi-interest distributions to facilitate multimodal sequential recommendation. Unlike single-path generative models, OT-FlowRec constructs a mixture of parallel, interest-specific transport flows to model the multi-centric distribution characteristics, where each flow preserves a distinct user interest cluster via an optimal transport plan. Then, a geometric transport mechanism is designed to establish minimum-cost paths from the latent multi-interest distribution to target preference prototypes to provide multi-interest awareness. Moreover, a progressive time-sampling warmup strategy dynamically biases optimization weights toward the final states of the generation process, bridging the gap between continuous global distribution fitting and practical top-k ranking metrics. This approach achieves explicit interest disentanglement, ensuring the preservation of a multi-peaked preference representation instead of collapsing it into a unimodal average. Extensive experiments on three public datasets demonstrate that OT-FlowRec outperforms existing state-of-the-art baselines.

1 INTRODUCTION

Multimodal sequential recommendation (MSR) has emerged as a pivotal paradigm for delivering personalized services in e-commerce, driven by its capacity to model the temporal dynamics of user preferences. By synergistically integrating interaction sequences with complementary signals derived from text and images, MSR is advantageous in disambiguating items that share similar behavioral footprints yet differ semantically, meanwhile enriching preference inference from sparse or inherently noisy user interactions. This enables MSR to capture users' nuanced, often concurrent interests—a capability that pure sequential models lack, and has drawn growing research attention. (Pan et al., 2026; Hu et al., 2023; Liu et al., 2024a; Ji et al., 2023; Liu et al., 2024c; Xu et al., 2025; Xv et al., 2024; Wei et al., 2023; Zhang et al., 2024a; Liu et al., 2024b; Fang et al., 2024).

A substantial body of research has been devoted to MSR, which can be categorized into three technical lines: traditional sequential methods, diffusion-based methods, and flow-based methods. Specifically, traditional sequential methods use recurrent networks or self-attention to capture ordered dependencies through discriminative learning (Hidasi et al., 2016; Kang & McAuley, 2018; Sun et al., 2019). Diffusion-based methods learn stochastic denoising processes in continuous spaces, improving robustness to noisy interactions and supporting flexible generation of target-item representations (Xie et al., 2024; Li et al., 2023; Yang et al., 2023; Liu et al., 2025c). Flow-based methods employ vector-field learning and deterministic ODE evolution, reducing the stochasticity and iterative cost of diffusion sampling while directly modeling continuous preference trajectories (Liu et al., 2025b;a; Huang et al., 2025; Li et al., 2025b).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

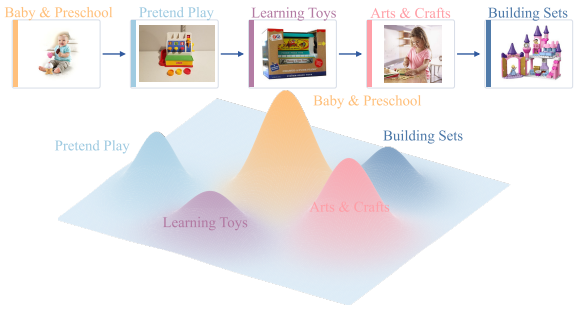

Figure 1: User interaction paths in the Toys dataset. Colors denote latent interest categories; this shows that multiple interests coexist within a session.

Although promising success has been achieved by previous works, two primary limitations remain. (i) **Multi-centric interest distribution collapse.** Existing global attention-based approaches compress multi-peaked interactions into single deterministic embeddings, which fail to describe the intrinsic multi-centric distribution characteristics. Moreover, most existing diffusion-based and flow-based methods construct single-peaked distribution representation, which is insufficient to model the multi-centric distribution of user intent. (Kang & McAuley, 2018; Sun et al., 2019; Yang et al., 2023; Hu et al., 2024; Liu et al., 2025b; Li et al., 2025b) (Figure 1). (ii) **Static transport with misaligned priors.** Prevailing flow-based generative recommenders construct time-dependent conditional paths based on independently sampled prior distributions, without accounting for the underlying optimal transport geometry between users' historical interactions and their target preferences. Consequently, the learned trajectories are directionally unconstrained and lack semantic grounding to the multi-interest prototypes (Lipman et al., 2023; Tong et al., 2024; Liu et al., 2025a).

To address these limitations, we propose OT-FlowRec, an optimal transport-guided multi-bridge flow matching framework. Rather than collapsing a complex interaction sequence into a single representation, a Multi-Prototype Guided Source Distribution (MPSD) module is constructed to represent concurrent user preferences with multi-interest source distribution. Its interest-specific components initialize parallel transport flows, preserving the underlying multi-peaked structure and avoiding unimodal collapse. Considering the advantage of interest prototypes (Wang et al., 2023a) in capturing diverse user semantics, an OT-Guided Multi-Bridge Flow Matching (OT-MBFM) module is further built, where multiple preference prototypes are first constructed. Accordingly, an interest distribution-to-prototype transformation learning process is then designed to map the multiple interest distribution to their matched prototypes through those parallel transport flows. To facilitate the transformation, minimum-cost paths are computed to connect the latent source distribution to target preference prototypes, ensuring explicit awareness of diverse user interests during generation. Furthermore, a mismatch remains between continuous distribution fitting and discrete ranking evaluation during the sampling process. We therefore develop a Progressive Time-sampling Warmup (PTW) that progressively shifts optimization toward the terminal states of the generation trajectory. By focusing training on these later states, PTW narrows the gap between continuous trajectory fitting and top-k ranking. The main contributions are summarized as follows:

- We propose OT-FlowRec, a multi-bridge flow learning framework that explicitly encodes the intrinsic multi-centric distribution of user interests. Unlike global attention mechanisms that compress concurrent interests into single deterministic embeddings, OT-FlowRec constructs a mixture of parallel, interest-specific transport flows, achieving effective multi-interest disentanglement while avoiding the multi-centric interest distribution collapse commonly observed in global attention methods.

- We propose the OT-MBFM module which establishes minimum-cost paths from the latent multi-interest distribution to target preference prototypes, providing directionally constrained trajectories. Specifically, a PTW strategy is designed to improve ranking quality by dynamically shifting optimization weights towards the terminal generative states, which effectively aligns continuous distribution fitting with top-k ranking metrics.

- Extensive experiments on three public datasets demonstrate that our framework outperforms existing state-of-the-art baselines across all evaluation metrics.

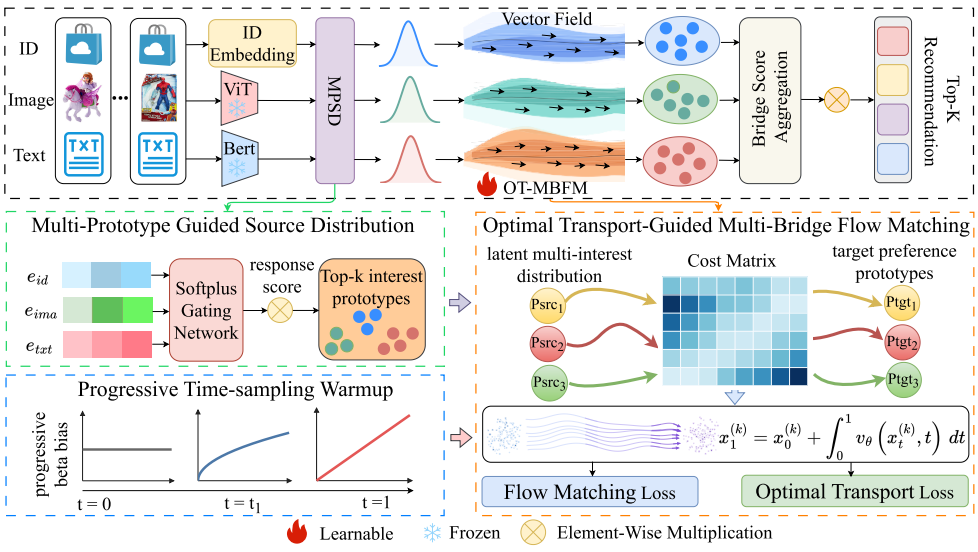

Figure 2: The overall architecture of OT-FlowRec. Top: The end-to-end recommendation pipeline, transforming multimodal item interactions into parallel transport flows. Middle-left: MPSD fuses multimodal features via a softplus gating network and extracts top-k prototypes to construct parallel interest distributions. Bottom-right: OT-MBFM establishes OT couplings via a geometric cost matrix to guide deterministic flow matching along minimum-cost trajectories. Bottom-left: PTW dynamically shifts time-step sampling weights toward terminal generation states ($t \to 1$) via Beta distribution adjustments.

2 Methodology

2.1 Preliminaries

In sequential multimodal recommendation, the user set is defined as $U = \{u_1, u_2, \ldots, u_{|U|}\}$ and the item set is defined as $I = \{i_1, i_2, \ldots, i_{|I|}\}$. For each user $u \in U$, the sequence of historically interacted items is represented by $S_u = [i_1^{(u)}, i_2^{(u)}, \ldots, i_n^{(u)}]$, where n denotes sequence length.

2.2 Overall Framework

Figure 2 shows OT-FlowRec coordinating three modules to transform multimodal interaction sequences into top-k recommendation lists.

Given multimodal interaction sequences, OT-FlowRec fuses item ID embeddings with pre-trained visual and textual features via a softplus gating network. Matching these representations against a learnable prototype space selects top-k responsive prototypes to initialize parallel interest-specific bridges. An optimal transport (OT) matrix couples these bridges with target preference prototypes along minimum-cost trajectories, letting a v_θ-guided ODE solver integrate states over $t \in [0, 1]$ into k terminal representations. During training, a dynamic Beta distribution replaces uniform sampling, shifting weights toward late intervals to align trajectory fitting with discrete ranking objectives. At inference, a bridge score aggregation layer max-pools inner products between terminal states and candidate item embeddings for top-k recommendations.

2.3 Multi-Prototype Guided Source Distribution

We propose the MPSD module to explicitly disentangle concurrent user preferences. Rather than compressing multimodal interactions into a single representation via global attention (Kang & McAuley, 2018; Sun et al., 2019; Lin et al., 2024), MPSD evaluates user responsiveness to multi-interest prototypes. By selecting top-k prototypes to construct independent ODE paths, it forms parallel transport flows that preserve multi-peaked distribution of next-item preferences.

To exploit the multimodal semantics of items, a softplus gating network is designed to dynamically fuse collaborative filtering signals and modal features. For any item i in the sequence and modality $m \in \{id, v, t\}$, its multimodal fused representation $x_i \in \mathbb{R}^d$ is calculated as follows:

$$g_m = \text{Softplus}\left(W_m \left[e_{id} \| e_v \| e_t\right] + b_m\right), \tag{1}$$

$$x_i = g_{id} \odot e_{id} + g_v \odot e_v + g_t \odot e_t, \tag{2}$$

where $W_m \in \mathbb{R}^{d \times 3d}$ and $b_m \in \mathbb{R}^d$ are learnable mapping parameters. This gating mechanism adaptively weights different modalities to accurately capture item characteristics.

To disentangle multiple user interests, a globally learnable interest prototype space $P = [p_1, p_2, \ldots, p_M] \in \mathbb{R}^{M \times d}$ is constructed, where M is the total number of prototypes. The model extracts the most recent fine-grained preference state $h_u \in \mathbb{R}^d$, which combines the last hidden state with the sequence representation and calculates its normalized response score s_j with respect to all candidate prototypes p_j as follows:

$$s_j = \frac{\exp(h_u^\top p_j / \tau)}{\sum_{m=1}^{M} \exp(h_u^\top p_m / \tau)}, \tag{3}$$

where τ is the temperature parameter. This score quantifies the alignment between the user state and latent interests for subsequent prototype selection.

We extract the top-k prototypes with the highest response scores instead of a global weighted average to preserve the multi-peaked nature of diverse user interests. The index set of the selected prototypes K_p is defined as follows:

$$K_p = \text{TopK}\left(s_1, \ldots, s_M\right). \tag{4}$$

Consequently, the representation strategy is transformed into a mixture of parallel, interest-specific transport flows. This design achieves interest disentanglement, thereby maintaining a robust multi-peaked preference representation.

2.4 OT-Guided Multi-Bridge Flow Matching

This section details the OT-MBFM module, which constructs deterministic transport trajectories from the latent multi-interest distribution to target preference prototypes. It minimizes geometric transport cost to derive an optimal coupling that guides continuous representation generation.

Assume that the selected top-k interest prototype matrix $P_{tgt} \in \mathbb{R}^{|K_p| \times d}$ serves as the support set for the target distribution. The geometric transport cost matrix $C \in \mathbb{R}^{n \times |K_p|}$ between the latent multi-interest distribution P_{src} and P_{tgt} is formulated as $C_{j,k} = \|x_j - p_k\|_2^2$. The marginal masses are specified in Appendix A. By introducing Sinkhorn entropy regularization (Cuturi, 2013), the OT plan $\pi^\star \in \mathbb{R}^{n \times |K_p|}$ is obtained as follows:

$$\pi^\star = \arg\min_{\pi \in \Pi(P_{src}, P_{tgt})} \sum_{j,k} \pi_{j,k} C_{j,k} - \epsilon H(\pi), \tag{5}$$

$$H(\pi) = -\sum_{j,k} \pi_{j,k} \log \pi_{j,k}, \tag{6}$$

where $\Pi(\cdot, \cdot)$ is the set of all joint probability measures that satisfy the marginal distribution constraints, and ϵ denotes a hyperparameter controlling smoothness.

Based on $\pi^\star$, the framework softly assigns historical sequence items to corresponding target preference prototypes, thereby forming multiple parallel flow matching bridges. For any time-step $t \in [0, 1]$, a deep neural network $v_\theta(x, t)$ parameterizes the target vector field $v_t(x)$. Guided by the OT plan, the conditional vector field transitions from random pairing to deterministic matching along optimal geometric paths. By establishing minimum-cost transport paths, the OT-MBFM module equips the framework with explicit multi-interest awareness throughout continuous generation.

2.5 PROGRESSIVE TIME-SAMPLING WARMUP

This section details the PTW strategy, designed to align continuous generative objectives with discrete top-k ranking metrics. By dynamically shifting optimization weights toward the terminal states of the generation trajectory, the PTW strategy ensures that the vector field precision is concentrated where practical recommendation ordering is determined.

Specifically, the strategy employs the Beta distribution to replace the standard uniform distribution $U(0,1)$ for sampling the integration time-step t. The shape parameter α is dynamically adjusted as training progresses:

$$t \sim \mathrm{Beta}\left(\alpha, 1\right). \tag{7}$$

This formulation establishes a time-dependent weighting mechanism to prioritize late-stage generation. During the early stages of training, setting $\alpha \approx 1$ reduces the process to uniform sampling, which prompts the model to capture the global vector field. As training advances, α is gradually increased to bias the sampling probability toward the high time-step interval approaching 1. Consequently, this mechanism transitions the optimization focus from global distribution modeling to terminal state refinement, bridging continuous distribution fitting and discrete ranking evaluation.

2.6 OPTIMIZATION

This section details the end-to-end optimization of OT-FlowRec. To guide continuous representation generation and discrete item ranking, the framework minimizes a combined loss function L_{total}:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{FM}} + \lambda_2 \mathcal{L}_{\text{CE}} + \lambda_3 \mathcal{L}_{\text{OT}} + \lambda_4 \mathcal{L}_{\text{Proto}}, \tag{8}$$

where $\mathcal{L}_{\text{FM}}$, $\mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\text{OT}}$, and $\mathcal{L}_{\text{Proto}}$ denote the flow matching loss, the cross-entropy ranking loss, the OT geometric regularization, and the prototype orthogonality constraint, respectively, with $\lambda_\cdot$ representing weighting hyperparameters tuned via Bayesian optimization.

To drive representation generation along the deterministic geometric paths established by the OT-MBFM module, the target vector field $v_t(x)$ at time-step $t \in [0,1]$ is parameterized by a deep neural network $v_\theta(x,t)$. The flow matching loss $\mathcal{L}_{\text{FM}}$ is defined as follows:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{\substack{t\sim\mathrm{Beta}(\alpha,1),\\ (x_0,x_1)\sim\pi^\star}} \left[\left\| v_\theta(x_t,t) - (x_1 - x_0) \right\|_2^2 \right], \tag{9}$$

where $x_t = (1-t)x_0 + t x_1$ represents interpolated state along OT path. This loss guides continuous representation generation along minimum-cost transport paths.

Bridging continuous generation and discrete ranking, $\mathcal{L}_{\text{CE}}$ injects a collaborative filtering signal by aggregating the terminal states $x_1^{(k)}$ over the selected prototypes K_p. Specifically, we define the bridge score for item j as $s_j = \max_{k\in K_p} \langle x_1^{(k)}, e_j \rangle / \tau_{\text{ce}}$ and maximize the likelihood of the ground-truth next item i_{n+1}:

$$\mathcal{L}_{\text{CE}} = -\sum_{u\in\mathcal{U}} \log \frac{\exp\left(s_{i_{n+1}}\right)}{\sum_{j\in\mathcal{I}} \exp\left(s_j\right)}, \tag{10}$$

where τ_{ce} is the temperature and e denotes item embeddings.

To prevent unconstrained generation in semantic spaces, we regularize vector field using the Sinkhorn optimal coupling $\pi^\star$ between source distribution and target prototypes for each bridge:

$$\mathcal{L}_{\text{OT}} = \sum_{j=1}^{n} \sum_{k\in K_p} \pi^\star_{j,k} \left\| x_j - p_k \right\|_2^2 - \epsilon H\left(\pi^\star\right), \tag{11}$$

with entropy regularizer $H(\cdot)$. This regularization guides generation trajectories along minimal-effort transport paths.

Finally, the orthogonality constraint $\mathcal{L}_{\text{Proto}}$ on the row-normalized matrix $\tilde{P}$ maintains semantic capacity against mode collapse:

$$\mathcal{L}_{\text{Proto}} = \| \tilde{P}\tilde{P}^\top - I_M \|_F^2. \tag{12}$$

This constraint disperses prototypes uniformly to capture multi-peaked interest independently.

We optimize the framework using Adam. During inference, the selected branches share the initial state $x_0^{(k)} = h_u$, the Euler ODE solver integrates along the learned vector field:

$$x_1^{(k)} = x_0^{(k)} + \int_0^1 v_\theta(x_t^{(k)}, t \mid p_k)\, dt, \quad \forall k \in K_p, \tag{13}$$

where the resulting terminal states $x_1^{(k)}$ serve as the predicted multi-interest representation. Finally, these representations are matched against candidate item embeddings via max-pooling aggregation to generate the top-k recommendations.

3 Experiments

3.1 Experiment Setting

To evaluate the effectiveness of the proposed OT-FlowRec in sequential multimodal recommendation tasks, we design and conduct comprehensive experiments. Specifically, the experiments are guided by three primary research questions:

RQ1: We investigate whether OT-FlowRec consistently outperforms existing sequential, multimodal, and generative recommendation baselines across real-world datasets characterized by varying sequence lengths and item domains.

RQ2: We explore how the key modules of OT-FlowRec, including MPSD, OT-MBFM, and PTW, contribute to the overall recommendation performance, and we analyze the specific roles they play within the model.

RQ3: We examine how the proposed MPSD, OT-MBFM, and PTW modules improve recommendation performance and generation quality compared with traditional generative recommendation models that rely on single global representations and standard flow matching.

3.1.1 Datasets and Evaluation Metrics

We conduct experiments on three widely used Amazon datasets (McAuley et al., 2015): Beauty, Video Games, and Toys (statistics in Table 1). These datasets cover distinct item domains, representing cosmetics and personal care, gaming consoles and accessories, and children's entertainment products. For a fair comparison, we adopt Hit Rate (H@K) and Normalized Discounted Cumulative Gain (N@K) with $K \in \{10, 20\}$ to assess top-k recall and ranking quality, respectively.

Table 1: Statistics of the evaluation datasets.

Dataset	Users	Items	Inter.	Sparsity	Avg. Len.
Beauty	22364	12102	198502	99.92%	8.8
Video Games	24304	10679	231780	99.91%	9.5
Toys	19413	11925	167597	99.92%	8.6

3.1.2 Baseline Methods

To evaluate OT-FlowRec, we compare it against three categories of baseline methods: (1) **Traditional sequential models**, including RNN-based and Transformer-based architectures (GRU4Rec (Hidasi et al., 2016), SASRec (Kang & McAuley, 2018), BERT4Rec (Sun et al., 2019)) and those integrating side information (MSSR (Lin et al., 2024)); (2) **Diffusion-based models**, spanning foundational frameworks (DreamRec (Yang et al., 2023), DiffuRec (Li et al., 2023)) and advanced variants (DiffMM (Jiang et al., 2024), ADRec (Chen et al., 2025), MoDiCF (Li et al., 2025a), PreferDiff (Liu et al., 2025c)); and (3) **Flow-based models** tailored for recommendation (FMRec (Liu et al., 2025b), FlowCF (Liu et al., 2025a), FindRec (Wang et al., 2025)).

3.1.3 Implementation Details

We implement the proposed method using the widely adopted RecBole (Zhao et al., 2021) framework. For all datasets, we filter out users and items with fewer than 5 interactions and truncate

historical sequences to a maximum length of 50. The models are trained for up to 300 epochs using the Adam optimizer, with a training batch size of 1024 and an evaluation batch size of 4096. We apply an early-stopping strategy if validation performance does not improve for 5 consecutive evaluations. Performance is evaluated using Hit and NDCG at $K \in \{10, 20\}$. Textual and visual features are extracted using BERT (Devlin et al., 2019) (768 dimensions) and CLIP-ViT (Radford et al., 2021) (512 dimensions), respectively, which are then mapped to the hidden dimension d via linear layers. All experiments are conducted on a single NVIDIA RTX A6000 GPU. For further experimental details, see Appendix D.

3.2 RQ1: Overall Performance

Table 2 demonstrates that OT-FlowRec outperforms all three baseline categories across all datasets and top-k metrics, achieving average performance improvements of 25.08% in H@20 and 9.15% in N@20 across the three datasets.

Comparison with traditional sequential models. Traditional sequential baselines compress complex historical interactions into a single deterministic representation via global attention, inducing unimodal collapse that prevents capturing concurrent multi-interest behaviors. Through the MPSD module, OT-FlowRec explicitly disentangles interactions into parallel, interest-specific transport flows to preserve multi-peaked interest distributions. Consequently, compared with MSSR, which is the strongest baseline in this category, OT-FlowRec achieves substantial improvements on Beauty dataset, increasing H@20 by 32.53% and N@20 by 19.86%.

Comparison with diffusion-based models. Continuous diffusion models suffer from unguided and curved diffusion paths. In contrast, OT-FlowRec replaces stochastic denoising with deterministic flow matching along OT paths. By establishing explicit minimum-cost connections between source distributions and target prototypes via the OT-MBFM module, the framework significantly stabilizes generation. Compared with ADRec, the top-performing diffusion baseline, OT-FlowRec demonstrates notable gains on Video Games, improving H@20 by 27.69% and N@20 by 10.63%.

Comparison with flow-based models. Existing flow-based recommendation baselines ignore semantic transport distances by relying on randomly paired Euclidean straight paths and uniform time sampling, which misaligns continuous trajectory fitting with discrete ranking objectives. OT-FlowRec resolves these bottlenecks through OT couplings and the PTW strategy. Biasing optimization weights toward terminal generation states ($t \to 1$) concentrates vector field precision where discrete item ranking occurs. Compared with FlowCF, the strongest flow-based competitor, OT-FlowRec drives NDCG improvements, increasing N@20 by 9.81% over FlowCF on Beauty.

3.3 RQ2: Ablation Study

To assess the individual contributions of core designs in OT-FlowRec, we evaluate three ablation variants across the Beauty, Video Games, and Toys datasets (Table 3).

At the representation level, constructing parallel, interest-specific transport flows is core to disentangling concurrent user behaviors and preserving multi-peaked interest distributions. Removing the multi-prototype mechanism (**w/o MPSD**) compresses complex historical interactions into a single global representation, degrading performance across all evaluation metrics. For instance, on the Beauty dataset, H@10 drops from 0.0891 to 0.0866.

Moreover, establishing minimum-cost geometric paths from the latent multi-interest distribution to target preference prototypes contributes to explicit multi-interest awareness and mitigates trajectory conflicts. Replacing OT guidance with randomly paired Euclidean straight paths (**w/o OT-MBFM**) fails to capture complex behavioral sequences, causing performance declines. Specifically, H@10 decreases from 0.1276 to 0.1239 on the Video Games dataset.

During the optimization phase, dynamically shifting training weights toward terminal generation states is beneficial for continuous distribution fitting with discrete top-k ranking metrics. Substituting the PTW strategy with uniform time-sampling (**w/o PTW**) ignores ranking sensitivity to late-stage integration precision ($t \to 1$). Consequently, this variant impairs ranking accuracy across all datasets, as evidenced by N@20 dropping from 0.0780 to 0.0757 on the Video Games dataset.

Table 2: Overall performance. The best and second-best scores are shown in bold and underlined, respectively.

Baseline	Beauty				Video Games				Toys			
	H@10	H@20	N@10	N@20	H@10	H@20	N@10	N@20	H@10	H@20	N@10	N@20
SASRec	0.0401	0.0660	0.0200	0.0265	0.0847	0.1281	0.0423	0.0532	0.0322	0.0497	0.0164	0.0208
GRU4Rec	0.0562	0.0851	0.0298	0.0370	0.0725	0.1122	0.0376	0.0475	0.0454	0.0676	0.0260	0.0316
BERT4Rec	0.0379	0.0571	0.0200	0.0249	0.0708	0.1091	0.0361	0.0456	0.0274	0.0395	0.0154	0.0184
MSSR	0.0583	0.0956	0.0314	0.0448	0.0709	0.1283	0.0415	0.0515	0.0613	0.0950	0.0318	0.0410
DreamRec	0.0228	0.0417	0.0130	0.0176	0.0415	0.0712	0.0221	0.0295	0.0216	0.0391	0.0111	0.0155
DiffuRec	0.0548	0.0939	0.0263	0.0354	0.0617	0.0795	0.0278	0.0433	0.0725	0.0982	0.0366	0.0430
DiffMM	0.0556	0.0975	0.0225	0.0361	0.0651	0.1064	0.0333	0.0438	0.0416	0.0664	0.0214	0.0276
ADRec	0.0673	0.0945	0.0417	0.0485	0.1059	0.1538	0.0584	0.0705	0.0704	0.0909	0.0421	0.0512
MoDiCF	0.0531	0.0821	0.0263	0.0335	0.0585	0.0948	0.0292	0.0383	0.0496	0.0724	0.0254	0.0312
PreferDiff	0.0594	0.0675	0.0350	0.0471	0.0735	0.0838	0.0552	0.0578	0.0463	0.0489	0.0334	0.0341
FMRec	0.0531	0.1069	0.0358	0.0459	0.0918	0.1370	0.0477	0.0593	0.0838	0.1013	0.0426	0.0501
FlowCF	0.0502	0.1075	0.0349	0.0489	0.0744	0.1050	0.0427	0.0505	0.0650	0.0804	0.0360	0.0450
FindRec	0.0549	0.0770	0.0318	0.0374	0.0911	0.1394	0.0465	0.0586	0.0703	0.0959	0.0440	0.0505
ours	**0.0891**	**0.1267**	**0.0442**	**0.0537**	**0.1276**	**0.1964**	**0.0607**	**0.0780**	**0.0946**	**0.1314**	**0.0455**	**0.0548**
Improv.	32.39%	17.86%	5.99%	9.81%	20.49%	27.69%	3.93%	10.63%	12.88%	29.71%	3.40%	7.03%

Table 3: Ablation study results.

Variant	Beauty				Video Games				Toys			
	H@10	H@20	N@10	N@20	H@10	H@20	N@10	N@20	H@10	H@20	N@10	N@20
w/o MPSD	0.0866	0.1249	0.0431	0.0515	0.1241	0.1948	0.0590	0.0776	0.0934	0.1299	0.0437	0.0529
w/o OT-MBFM	0.0874	0.1253	0.0435	0.0531	0.1239	0.1936	0.0589	0.0764	0.0913	0.1303	0.0446	0.0544
w/o PTW	0.0863	0.1257	0.0428	0.0527	0.1244	0.1919	0.0587	0.0757	0.0942	0.1299	0.0438	0.0529
ours	**0.0891**	**0.1267**	**0.0442**	**0.0537**	**0.1276**	**0.1964**	**0.0607**	**0.0780**	**0.0946**	**0.1314**	**0.0455**	**0.0548**

3.4 RQ3: Qualitative Analysis

Figure 3 shows the vector field directions of both traditional random-pairing flow matching and OT-FlowRec at different integration times. According to Figure 3, traditional flow matching produces conflicting directions and extensive region ambiguity, with directional consistency falling to 0.18 and the ambiguity rate rising to 90.0% at $t = 0.5$. In contrast, OT-FlowRec maintains directional consistency above 0.95 and ambiguity below 1.1% across all three time steps, resulting in more coherent transport trajectories. The comparison of vector field direction demonstrates that OT-guided coupling in OT-FlowRec provides directionally constrained trajectories. The t-SNE and transport-cost analyses are presented in Appendix C(Figures 4 and 5) and training stability is presented in Appendix E.

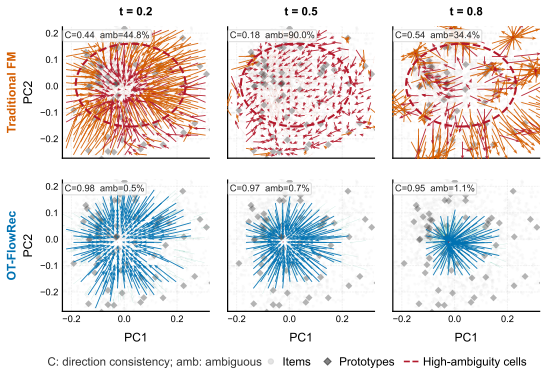

Figure 3: Vector field comparison between traditional flow matching and OT-FlowRec on the Toys.

4 Related Work

4.1 Traditional Sequential Recommendation

Traditional sequential recommendation methods primarily learn behavior transition patterns from user historical interaction sequences (Cen et al., 2020; Li et al., 2024; Zhang et al., 2024b; Chen et al., 2022; Qin et al., 2024; Wang et al., 2019; Li et al., 2019; Tan et al., 2021; Pan et al., 2026;

Park et al., 2025; Duan et al., 2025; Hou et al., 2022). Early approaches utilized Markov chains, RNNs (Hidasi et al., 2016), and CNNs (Tang & Wang, 2018), which were later broadly superseded by Transformer-based architectures like SASRec (Kang & McAuley, 2018) and BERT4Rec (Sun et al., 2019) to better capture sequence dependencies.

To alleviate interaction sparsity, subsequent research incorporated auxiliary multimodal information (e.g., text and images) (Hu et al., 2023; Wang et al., 2024; Liu et al., 2024a; Kim et al., 2025; Ji et al., 2023; Liu et al., 2024c; Xu et al., 2025; Yang et al., 2025; Zhang et al., 2025; Xv et al., 2024; Wei et al., 2023; Zhang et al., 2024a; Liu et al., 2024b; Fang et al., 2024), utilizing techniques like multi-sequence attention (Lin et al., 2024) and modality experts (Guo et al., 2025). However, these methods typically remain within the discriminative modeling paradigm, compressing complex user behaviors into deterministic, single-point representations. This limitation prevents them from fully capturing the multi-peaked distributional nature and generative evolution of dynamic user interests.

4.2 Diffusion-Based Recommendation

Diffusion-based methods shift recommendation from discriminative matching to generative distribution modeling (Xie et al., 2024; Chen et al., 2025; Cui et al., 2025; Jiang et al., 2024; Liu et al., 2023; Li et al., 2025a; Qu & Nobuhara, 2025; Mao et al., 2025; Liu et al., 2025c; Wang et al., 2023b). Early works like DiffuRec (Li et al., 2023) and DreamRec (Yang et al., 2023) reconstruct target item representations via stochastic noise-addition and reverse denoising processes, thereby capturing diverse user interests as continuous distributions rather than fixed vectors.

Subsequent research has enhanced this paradigm by integrating multimodal user-item graphs (Jiang et al., 2024), aligning generation with personalized ranking objectives (Liu et al., 2025c), and employing token-level diffusion to prevent embedding collapse (Chen et al., 2025). Despite these advancements, the inherent reliance on curved probability paths and multi-step stochastic sampling limits their inference efficiency and often leads to error accumulation during the generative process.

4.3 Flow-Based Recommendation

Flow-based models replace the stochastic denoising of diffusion models with deterministic continuous generation, transporting prior distributions to target distributions via time-dependent vector fields (Lipman et al., 2023; Tong et al., 2024). Recent works have adapted flow matching for recommendation by designing discrete flows for implicit feedback (Liu et al., 2025a), utilizing straight trajectories to reduce preference perturbations (Liu et al., 2025b), and employing Stein-guided entropic flows to capture multi-peaked sequential preferences (Wang et al., 2025).

Despite improving inference efficiency and stability over diffusion models, existing flow matching methods typically construct simple, global straight paths. This ignores the explicit geometric transport costs between historical distributions and latent multi-interest prototypes. Furthermore, their continuous distribution fitting objectives often misalign with discrete top-k ranking metrics. To bridge these gaps, our OT-FlowRec proposes OT-MBFM for explicit interest disentanglement, alongside the PTW strategy to directly optimize late-stage generation for practical ranking accuracy.

5 Conclusion

In this paper, we proposed OT-FlowRec to jointly model concurrent user interests in multimodal sequential recommendation. To overcome unimodal collapse, our framework constructed parallel transport flows that disentangled diverse interests and preserved multi-peaked preference distributions. Integrating minimum-cost geometric transport paths with the progressive time-sampling warmup strategy achieved explicit multi-interest awareness and bridged continuous distribution fitting and discrete top-k ranking metrics. Extensive experiments across three datasets demonstrated that OT-FlowRec outperformed state-of-the-art baselines.

AI use statement

AI tools were used to polish the writing of this paper.

REFERENCES

Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2942–2951, 2020. doi: 10.1145/3394486.3403344.

Jialei Chen, Yuanbo Xu, and Yiheng Jiang. Unlocking the power of diffusion models in sequential recommendation: A simple and effective approach. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 155–166, 2025. doi: 10.1145/3711896.3737172.

Yongjun Chen, Zhiwei Liu, Jia Li, Julian McAuley, and Caiming Xiong. Intent contrastive learning for sequential recommendation. In *Proceedings of the ACM Web Conference 2022*, pp. 2172–2182, 2022. doi: 10.1145/3485447.3512090.

Xiaoxi Cui, Weihai Lu, Yu Tong, Yiheng Li, and Zhejun Zhao. Multi-modal multi-behavior sequential recommendation with conditional diffusion-based feature denoising. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1593–1602, 2025. doi: 10.1145/3726302.3730044.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pp. 2292–2300, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Sirui Duan, Mengya Ouyang, Rong Wang, Qian Li, and Yunpeng Xiao. Let long-term interests talk: An disentangled learning model for recommendation based on short-term interests generation. *Information Processing & Management*, 62(2):103997, 2025. doi: 10.1016/j.ipm.2024.103997.

Hong Fang, Leiyuxin Sha, and Jindong Liang. Multimodal recommender system based on multi-channel counterfactual learning networks. *Multimedia Systems*, 30(5):242, 2024. doi: 10.1007/s00530-024-01448-z.

Xu Guo, Tong Zhang, Yufei Xue, Chenxu Wang, Fuyun Wang, and Zitao Cui. M3rec: Selective state space models with mixture-of-modality experts for multi-modal sequential recommendation. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887582.

Balazs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations*, 2016.

Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 585–593, 2022. doi: 10.1145/3534678.3539381.

Guoqing Hu, Zhengyi Yang, Zhibo Cai, An Zhang, and Xiang Wang. Generate and instantiate what you prefer: Text-guided diffusion for sequential recommendation. *CoRR abs/2410.13428*, 2024. doi: 10.48550/arXiv.2410.13428.

Hengchang Hu, Wei Guo, Yong Liu, and Min-Yen Kan. Adaptive multi-modalities fusion in sequential recommendation systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 843–853, 2023. doi: 10.1145/3583780.3614775.

Yinxuan Huang, Ke Liang, Zhuofan Dong, Xiaodong Qu, Tianxiang Wang, Yue Han, Jingao Xu, Bin Zhou, and Ye Wang. Flow matching for denoised social recommendation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25474–25485. PMLR, 2025.

Wei Ji, Xiangyan Liu, An Zhang, Yinwei Wei, Yongxin Ni, and Xiang Wang. Online distillation-enhanced multi-modal transformer for sequential recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 955–965, 2023. doi: 10.1145/3581783.3612091.

Yangqin Jiang, Lianghao Xia, Wei Wei, Da Luo, Kangyi Lin, and Chao Huang. Diffmm: Multi-modal diffusion model for recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7591–7599, 2024. doi: 10.1145/3664647.3681498.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining*, pp. 197–206, 2018. doi: 10.1109/ICDM.2018.00035.

Hye-young Kim, Minjin Choi, Sunkyung Lee, Ilwoong Baek, and Jongwuk Lee. Diff: Dual side-information filtering and fusion for sequential recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1624–1633, 2025. doi: 10.1145/3726302.3729948.

Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Pipei Huang, Huan Zhao, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. Multi-interest network with dynamic routing for recommendation at tmall. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2615–2623, 2019. doi: 10.1145/3357384.3357814.

Jin Li, Shoujin Wang, Qi Zhang, Shui Yu, and Fang Chen. Generating with fairness: A modality-diffused counterfactual framework for incomplete multimodal recommendations. In *Proceedings of the ACM on Web Conference 2025*, pp. 2787–2798, 2025a. doi: 10.1145/3696410.3714606.

Li Li, Mingyue Cheng, Yuyang Ye, Zhiding Liu, and Enhong Chen. Preference trajectory modeling via flow matching for sequential recommendation. *CoRR abs/2508.17618*, 2025b. doi: 10.48550/arXiv.2508.17618.

Zihao Li, Aixin Sun, and Chenliang Li. Diffurec: A diffusion model for sequential recommendation. *ACM Transactions on Information Systems*, 42(3):1–28, 2023.

Zihao Li, Yunfan Xie, Wei Emma Zhang, Pengfei Wang, Lixin Zou, Fei Li, Xiangyang Luo, and Chenliang Li. Disentangle interest trend and diversity for sequential recommendation. *Information Processing & Management*, 61(3):103619, 2024. doi: 10.1016/j.ipm.2023.103619.

Xiaolin Lin, Jinwei Luo, Junwei Pan, Weike Pan, Zhong Ming, Xun Liu, Shudong Huang, and Jie Jiang. Multi-sequence attentive user representation learning for side-information integrated sequential recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 414–423, 2024. doi: 10.1145/3616855.3635815.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Chengkai Liu, Yangtian Zhang, Jianling Wang, Rex Ying, and James Caverlee. Flow matching for collaborative filtering. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1765–1775, 2025a. doi: 10.1145/3711896.3736967.

Feng Liu, Lixin Zou, Xiangyu Zhao, Min Tang, Liming Dong, Dan Luo, Xiangyang Luo, and Chenliang Li. Flow matching based sequential recommender model. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 3108–3116, 2025b. doi: 10.24963/ijcai.2025/346.

Qidong Liu, Fan Yan, Xiangyu Zhao, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and Feng Tian. Diffusion augmentation for sequential recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 1576–1586, 2023. doi: 10.1145/3583780.3615134.

Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 57(2):26:1–26:17, 2024a. doi: 10.1145/3695461.

Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. Multimodal pretraining, adaptation, and generation for recommendation: A survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6566–6576, 2024b. doi: 10.1145/3637528.3671473.

Shuo Liu, An Zhang, Guoqing Hu, Hong Qian, and Tat-Seng Chua. Preference diffusion for recommendation. In *Proceedings of the 13th International Conference on Learning Representations*, 2025c.

Yifan Liu, Kangning Zhang, Xiangyuan Ren, Yanhua Huang, Jiarui Jin, Yingjie Qin, Ruilong Su, Ruiwen Xu, Yong Yu, and Weinan Zhang. Alignrec: Aligning and training in multimodal recommendations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1503–1512, 2024c. doi: 10.1145/3627673.3679626.

Wenyu Mao, Jiancan Wu, Guoqing Hu, Zhengyi Yang, Wei Ji, and Xiang Wang. On efficiency-effectiveness trade-off of diffusion-based recommenders. In *Advances in Neural Information Processing Systems 38*, 2025.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52, 2015. doi: 10.1145/2766462.2767755.

Li-Wei Pan, Wei-Ke Pan, Mei-Yan Wei, Hong-Zhi Yin, and Zhong Ming. A survey on sequential recommendation. *Frontiers of Computer Science*, 20(3):2003606, 2026. doi: 10.1007/s11704-025-41329-w.

Seongmin Park, Mincheol Yoon, Minjin Choi, and Jongwuk Lee. Temporal linear item-item model for sequential recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 354–362, 2025. doi: 10.1145/3701551.3703554.

Xiuyuan Qin, Huanhuan Yuan, Pengpeng Zhao, Guanfeng Liu, Fuzhen Zhuang, and Victor S. Sheng. Intent contrastive learning with cross subsequences for sequential recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 548–556, 2024. doi: 10.1145/3616855.3635773.

Yuanpeng Qu and Hajime Nobuhara. Intent-aware diffusion with contrastive learning for sequential recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1552–1561, 2025. doi: 10.1145/3726302.3730010.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1441–1450, 2019. doi: 10.1145/3357384.3357895.

Qiaoyu Tan, Jianwei Zhang, Jiangchao Yao, Ninghao Liu, Jingren Zhou, Hongxia Yang, and Xia Hu. Sparse-interest network for sequential recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 598–606, 2021. doi: 10.1145/3437963.3441811.

Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 565–573, 2018. doi: 10.1145/3159652.3159656.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Jinpeng Wang, Ziyun Zeng, Yunxiao Wang, Yuting Wang, Xingyu Lu, Tianxiang Li, Jun Yuan, Rui Zhang, Hai-Tao Zheng, and Shu-Tao Xia. Missrec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6548–6557, 2023a.

Maolin Wang, Yutian Xiao, Binhao Wang, Sheng Zhang, Shanshan Ye, Wanyu Wang, Hongzhi Yin, Ruocheng Guo, and Zenglin Xu. Findrec: Stein-guided entropic flow for multi-modal sequential recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3008–3018, 2025. doi: 10.1145/3711896.3736968.

Shoujin Wang, Liang Hu, Yan Wang, Quan Z. Sheng, Mehmet Orgun, and Longbing Cao. Modeling multi-purpose sessions for next-item recommendations via mixture-channel purpose routing networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 3771–3777, 2019. doi: 10.24963/ijcai.2019/523.

Shuhan Wang, Bin Shen, Xu Min, Yong He, Xiaolu Zhang, Liang Zhang, Jun Zhou, and Linjian Mo. Aligned side information fusion method for sequential recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 112–120, 2024. doi: 10.1145/3589335.3648308.

Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. Diffusion recommender model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 832–841, 2023b. doi: 10.1145/3539618.3591663.

Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023*, pp. 790–800, 2023. doi: 10.1145/3543507.3583206.

Wenjia Xie, Rui Zhou, Hao Wang, Tingjia Shen, and Enhong Chen. Bridging user dynamics: Transforming sequential recommendations with schrödinger bridge and diffusion models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2618–2628, 2024. doi: 10.1145/3627673.3679756.

Jinfeng Xu, Zheyu Chen, Wei Wang, Xiping Hu, Sang-Wook Kim, and Edith 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*, pp. 1830–1839, 2025. doi: 10.1145/3726302.3729927.

Guipeng Xv, Xinyu Li, Ruobing Xie, Chen Lin, Chong Liu, Feng Xia, Zhanhui Kang, and Leyu 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*, pp. 3645–3656, 2024. doi: 10.1145/3637528.3671703.

Yimeng Yang, Haokai Ma, Lei Meng, Shuo Xu, Ruobing Xie, and Xiangxu Meng. Curriculum conditioned diffusion for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13035–13043, 2025. doi: 10.1609/aaai.v39i12.33422.

Zhengyi Yang, Jiancan Wu, Zhicai Wang, Xiang Wang, Yancheng Yuan, and Xiangnan He. Generate what you prefer: Reshaping sequential recommendation via guided diffusion. In *Advances in Neural Information Processing Systems 36*, pp. 24247–24261, 2023.

Lingzi Zhang, Xin Zhou, Zhiwei Zeng, and Zhiqi Shen. Multimodal pre-training for sequential recommendation via contrastive learning. *ACM Transactions on Recommender Systems*, 3(1): 1–23, 2024a. doi: 10.1145/3682075.

Shengzhe Zhang, Liyi Chen, Dazhong Shen, Chao Wang, and Hui Xiong. Hierarchical time-aware mixture of experts for multi-modal sequential recommendation. In *Proceedings of the ACM on Web Conference 2025*, pp. 3672–3682, 2025. doi: 10.1145/3696410.3714676.

Xiaokun Zhang, Bo Xu, Zhaochun Ren, Xiaochen Wang, Hongfei Lin, and Fenglong 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*, pp. 1883–1892, 2024b. doi: 10.1145/3626772.3657748.

Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 4653–4664, 2021. doi: 10.1145/3459637.3482016.

A Marginal Distributions for Optimal Transport

For a user with n valid Transformer states, let $\ell = \min(n, L)$, where L is the maximum source-window length. The source and target distributions are

$$P_{src} = \sum_{j=1}^{\ell} a_j \delta_{x_{n-\ell+j}}, \qquad P_{tgt} = \sum_{k=1}^{M} s_k \delta_{p_k}, \qquad a_j = \frac{\exp[-\gamma(\ell - j)]}{\sum_{q=1}^{\ell} \exp[-\gamma(\ell - q)]}, \tag{14}$$

where δ denotes a point mass, γ controls recency weighting, and s_k is the normalized prototype response defined in the main text. Padding positions are excluded. The feasible couplings satisfy

$$\Pi(P_{src}, P_{tgt}) = \{\pi \in \mathbb{R}_{+}^{\ell \times M} : \pi \mathbf{1}_{|K_p|} = \mathbf{a},\ \pi^{\top} \mathbf{1}_{\ell} = \mathbf{s}\}. \tag{15}$$

B Problem Formulation

Given the multimodal context, each item $i \in I$ comprises an ID embedding $e_{id} \in \mathbb{R}^d$, a visual feature $e_v \in \mathbb{R}^d$ and a textual feature $e_t \in \mathbb{R}^d$, where d represents the dimension of the latent feature space. Formally, given the historical interaction sequence S_u, the core task is to learn a continuous generative mapping function $f_\theta(\cdot)$ to predict the item $i_{n+1}^{(u)} \in I$ with the most likely to interact in the next time-step $n + 1$.

C Qualitative Analysis (RQ3)

Figure 4 shows the comparison between t-SNE projections (van der Maaten & Hinton, 2008) of historical item embeddings and the global mean-pooled state with the interest anchors learned by OT-FlowRec. A mean-pooled state represents these clusters with a single central point, whereas the learned anchors align with different clusters and retain the multi-peaked structure of concurrent user interests. This advantage arises from the MPSD model, which preserves interest-specific representations instead of averaging the entire interaction history into a single vector.

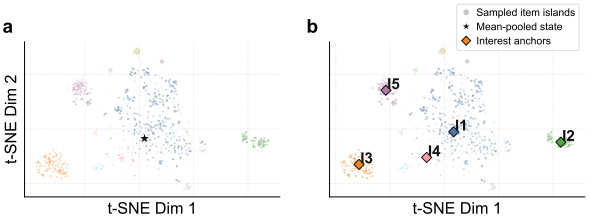

Figure 4: T-SNE visualization of user multi-interest behaviors on the Toys dataset. (a) A single averaged state suppresses minority interests; (b) multi-interest anchors preserve the semantic structure.

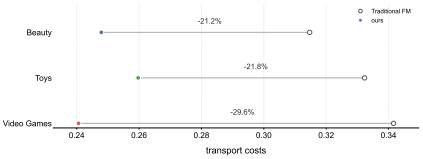

Figure 5: A comparison of transport costs between traditional flow matching and OT-FlowRec.

Beyond preserving the multi-peaked interest structure, Figure 5 compares transport costs of OT-FlowRec with traditional flow matching based on random pairings. OT-FlowRec lowers the average transport cost by 24.2% on three datasets, indicating better efficiency from source distribution to target preference prototypes. The improvement of OT-guided may be attributed to the coupling mechanism defined in Equation (11), which jointly minimizes the trajectory costs of multiple bridges.

D Additional Implementation Details

We further tune OT-FlowRec on each dataset using validation Hit@10. Unless otherwise specified, we fix the number of attention heads to 2, the feed-forward dimension to 512, the time-embedding dimension to 32. The pre-extracted visual and textual features are kept frozen and are projected into the hidden space by two-layer MLPs. The final item representation is obtained by a trainable

softplus-gated fusion of ID, visual, and textual embeddings. The user state is computed as the average of the last Transformer state and the mean-pooled valid sequence states. The entropic OT coupling is solved by Sinkhorn iterations with regularization ϵ_{OT}, and the OT loss uses the expected squared transport cost under this coupling. For multi-prototype bridge inference, the top-k prototypes are used as bridge conditions and their scores are aggregated by max pooling. We optimize softmax cross-entropy over all items.

The hyperparameter search space is summarized in Table 4. Log-uniform parameters are sampled as $\exp(\mathcal{U}[a, b])$. The dataset-specific best configurations selected on the validation set are reported in Table 5. After hyperparameter selection, each final configuration is rerun with five random seeds $\{2022, 2023, 2024, 2025, 2026\}$, and we report the mean performance over these runs.

Table 4: Hyperparameter search space of OT-FlowRec.

Hyperparameter	Search space
Learning rate	$\exp(\mathcal{U}[-9.2, -6.0])$
Number of Transformer layers	$\{1, 2, 3\}$
Hidden dimension d	$\{128, 256\}$
Hidden dropout / attention dropout	$\mathcal{U}[0.1, 0.5]$
Number of prototypes M	$\{32, 64, 96, 128\}$
ODE steps	$\{2, 4, 6, 8\}$
Prototype temperature τ_{proto}	$\exp(\mathcal{U}[-4.2, -1.9])$
Sinkhorn regularization ϵ_{OT}	$\exp(\mathcal{U}[-4.0, -1.6])$
Sinkhorn iterations	$\{5, 10, 20\}$
Warm-up epochs	validation-selected
λ_{Proto}, λ_{OT}	$\exp(\mathcal{U}[-4.6, -0.7])$

Table 5: Best hyperparameters of OT-FlowRec on each dataset.

Dataset	LR	Layers	d	Drop.	Attn Drop.	M	ODE	τ	ϵ_{OT}	Iter.	Warm-up	λ_{Proto}	λ_{OT}
Beauty	5.87×10^{-4}	1	128	0.39755	0.44607	96	6	0.03919	0.12053	10	12	0.11967	0.02177
Toys	6.322×10^{-4}	1	256	0.45730	0.26735	96	6	0.02318	0.117279	10	12	0.056853	0.02742
Video_Games	6.30×10^{-4}	2	128	0.49551	0.340265	96	6	0.084436	0.081655	5	9	0.12009	0.05440

E Additional Experimental Results

Validation curves corroborate these conclusions from the perspective of optimization dynamics. As Figure 6 illustrates, the full model achieves more stable and higher validation N@20 scores across the Beauty, Video Games, and Toys datasets. On the Toys dataset, the full model converges faster during early training and maintains its lead in later stages, indicating that multi-interest modeling and optimal-transport bridge paths more effectively capture the evolution of user interests. Curves for the Video Games dataset exhibit more pronounced fluctuations, reflecting its more complex user interests and item semantics. While the full model maintains optimal performance in later stages, removing the OT-MBFM module causes frequent oscillation and insufficient convergence. This demonstrates that OT-MBFM not only improves final recommendation accuracy but also alleviates directional conflicts during training, yielding a more stable optimization trajectory. Although performance gaps among variants are smaller on the Beauty dataset, the full model still performs best. This suggests that the marginal benefits of individual modules diminish when user interest structures are smooth, whereas the advantages of the optimal-transport-guided multi-bridge flow mechanism become more prominent on datasets with diverse interests and complex semantics.

The training curve for the w/o PTW variant highlights the core contribution of PTW in aligning continuous distribution fitting with practical ranking metrics. By progressively adjusting the beta distribution bias for time-step sampling, PTW dynamically biases optimization weights toward the final states of the generation process, rather than relying on balanced full-path supervision. During early training stages, this variant still exhibits an upward trend comparable to the full model, suggesting that the basic optimal-transport bridge flow alone can learn effective directions for interest evolution. In later stages, however, it frequently plateaus or oscillates, failing to reliably achieve the

optimal validation N@20 of the full model. This performance gap is especially evident on the Video Games and Toys datasets, where the full model maintains a distinct lead. These findings confirm that without dynamically biasing weights toward the final states of the evolution process, the model captures the overall transport path but struggles to refine the local discriminative capability required for practical top-k ranking.

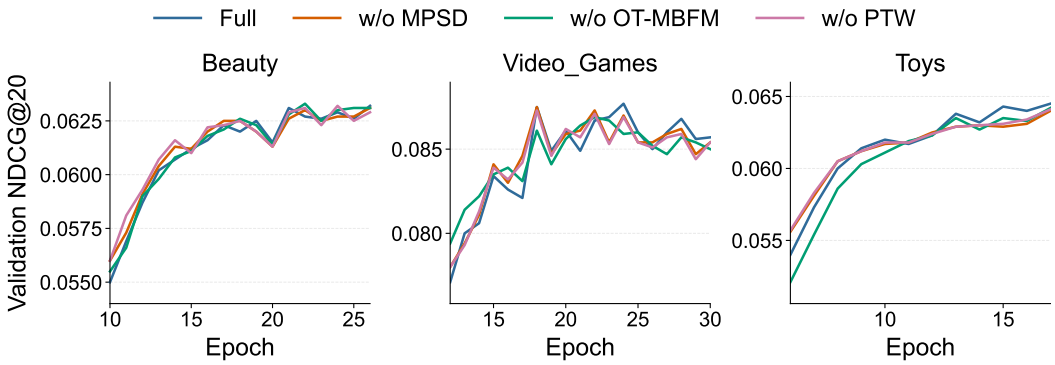

Figure 6: Validation N@20 curves of OT-FlowRec and the ablation variants