GateDiffInt: Gate-Mediated Controllable Diffusion and Multi-Intent LLM Distillation for User Behavior Modeling

Jialong Duan*
Fudan University
Shanghai, China
duanjialong@xiaohongshu.com

Zichen Zhang*✉
Xiaohongshu Inc.
Shanghai, China
zhangzichen1@xiaohongshu.com

Zirui Tu
Xiaohongshu Inc.
Shanghai, China
tuzirui@xiaohongshu.com

Zheng Zhang
Xiaohongshu Inc.
Shanghai, China
zhangzheng104@xiaohongshu.com

Zepeng Li
Xiaohongshu Inc.
Shanghai, China
lizepeng@xiaohongshu.com

Qingyao Cui
Xiaohongshu Inc.
Beijing, China
cuiqingyao@xiaohongshu.com

Qinwen Wang
Xiaohongshu Inc.
Beijing, China
wangqinwen@xiaohongshu.com

Yudan Liu✉
Xiaohongshu Inc.
Beijing, China
liuxiaobo@xiaohongshu.com

Luo Yang
Xiaohongshu Inc.
Beijing, China
yangluo1@xiaohongshu.com

Yao Hu
Xiaohongshu Inc.
Beijing, China
xiahou@xiaohongshu.com

Abstract

Existing recommendation ranking models, whether traditional feature-interaction methods or sequential approaches with behavior-sequence modeling, typically encode intent signals implicitly in model parameters or hidden states, making it difficult to explicitly disentangle structured intents that vary in strength and temporal scale. More fundamentally, noise and intent in behavior sequences are not independent but mutually reinforcing: on the one hand, behavioral noise persistently dilutes and distorts genuine intents; on the other, the absence of a structured intent prior leaves sequence denoising without a well-defined target. We term this mutual reinforcement Noise–Intent Coupling (NIC). To address NIC, we propose GateDiffInt, an intent interaction framework specifically designed for the ranking stage in industrial recommender systems that uses the final conversion task as a shared signal to jointly align sequence denoising and intent extraction. GateDiffInt introduces a controllable forward diffusion process with dual gating to enhance and denoise user behavior sequences. Building upon the denoised representations, the framework employs a large language model as a teacher to explicitly distill four categories of structured intents—long-term, short-term, latent, and conversion intents—from the behavior sequences, which are then aligned through a lightweight student model. The resulting enhanced sequence representations and structured intent representations are deeply fused via an attention mechanism to produce intent-aware sequence representations, which are subsequently used for final conversion rate prediction. Extensive experiments on both public benchmark datasets and large-scale industrial datasets demonstrate that GateDiffInt consistently and substantially outperforms strong baseline models. Moreover, in large-scale online A/B testing within real-world industrial recommendation scenarios, the ranking model powered by GateDiffInt achieves substantial improvements in GMV and has been successfully deployed to serve the primary traffic for hundreds of millions of daily active users, validating both the effectiveness and practical deployability of the proposed framework in production systems.

1 Introduction

Conversion rate (CVR) prediction operates on the behavior sequences that users generate over a feed, and such sequences inherently carry two entangled ingredients: *noise* from random browsing, accidental clicks, and conformity-driven interactions, and *intents* that evolve in parallel across multiple temporal scales—stable long-term preferences, transient short-term needs, latent comparison, and imminent conversion [32, 46]. Most prior studies treat denoising and intent extraction as separable subproblems: one line extracts intents directly from noisy sequences and leaves the noise to be implicitly absorbed by ever-stronger encoders; another denoises sequences under generic reconstruction targets and defers intent to downstream tasks. Both overlook a more fundamental property: in CVR sequences, noise and intent are not independent but *mutually reinforcing*—noise persistently dilutes and distorts genuine intents, while the absence of a structured intent prior deprives denoising of any criterion for separating informative weak signals from dispensable noise. We term this phenomenon *Noise–Intent Coupling* (NIC): intuitively, intent extraction without clean input has no material, and denoising without an intent target has no aim.

For analysis and naming, we give NIC a lightweight, *informal* characterization that assumes no specific noise-generating model. Let $\mathcal{I}^\star$ denote a user's underlying intent structure, and *conceptually* decompose the observed sequence as $\mathcal{S} = \mathcal{S}^\star \oplus \mathcal{N}$ (a true

*Zichen Zhang and Jialong Duan contributed equally to this work.

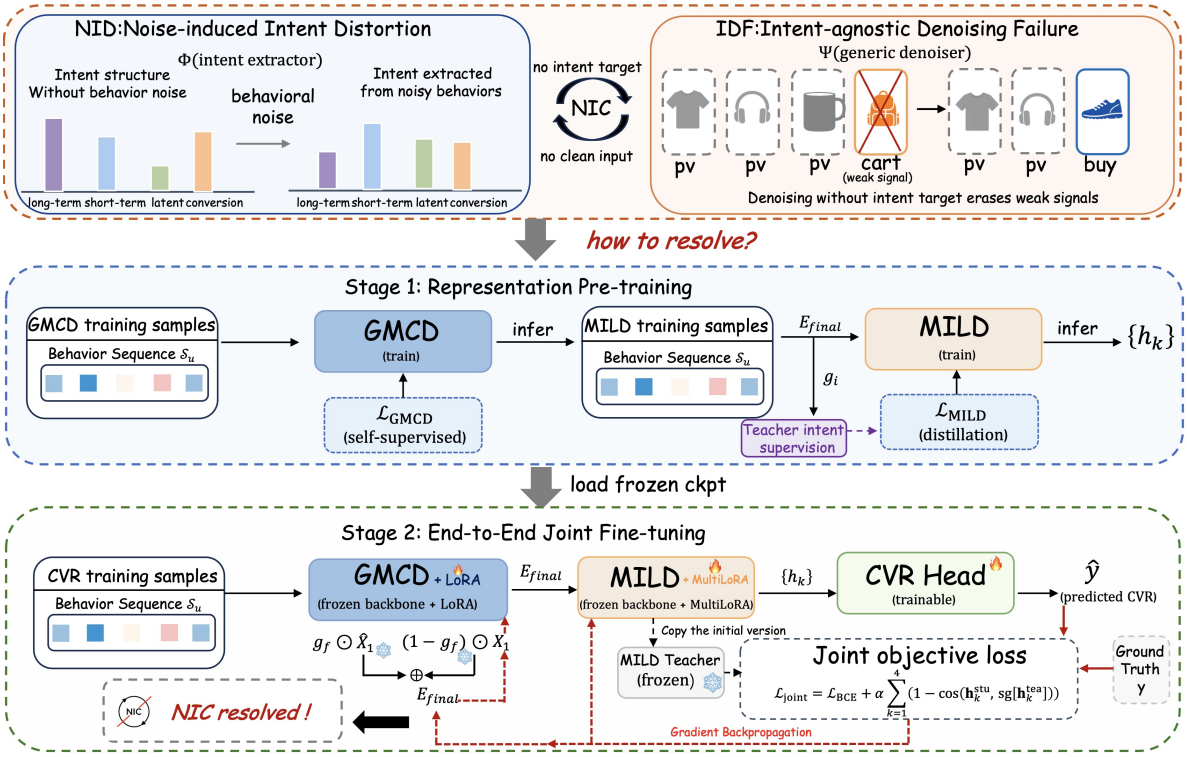

Figure 1: Overview of GateDiffInt. Top: the two sub-phenomena of NIC (NID and IDF). Stage 1 pretrains GMCD and MILD separately; Stage 2 loads frozen checkpoints with LoRA and a trainable CVR head, optimizing $\mathcal{L}_{\text{joint}}$ to align denoising and intent extraction toward the conversion task, resolving NIC.

signal $\mathcal{S}^\star$ and a noise component $\mathcal{N}$); let Φ and Ψ denote the intent extractor and the denoiser, respectively. This characterization serves only to *locate* the two sub-phenomena of NIC, not to impose a strict modeling assumption. NIC manifests as a *mutual conditional degradation* of Φ and Ψ, which decomposes into two separately measurable sub-phenomena. ***(i) Noise-induced Intent Distortion (NID).*** As $\mathcal{N}$ grows, both attention-based [19, 41] and LLM-based [24, 27] extractors let $\Phi(\mathcal{S})$ drift systematically toward high-frequency yet shallow signals, pulling the output away from $\mathcal{I}^\star$—short-term and latent intents dominate while stable long-term preferences are diluted [3, 21, 59]. ***(ii) Intent-agnostic Denoising Failure (IDF).*** Mainstream diffusion-based recommenders [25, 49] optimize generic reconstruction by design and are not conditioned on $\mathcal{I}^\star$, and thus lack a criterion for what to keep: a weak but informative signal such as an early price comparison may be smoothed away together with incidental browsing, or conversely systematic noise may be retained. Even recent diffusion–multi-interest hybrids [22, 36], though touching both sides, connect Φ and Ψ through a single forward pass and provide no explicit channel for the two to suppress each other. NID and IDF are therefore not independent problems to be solved in sequence—any one-sided effort leaves a backlash on the other.

Revisiting existing work through the lens of NID and IDF, each line of research closes only one end of the loop. Sequential and multi-interest models [2, 4, 10, 23, 42, 53, 58, 59] strengthen the representation of the raw sequence through attention or multi-vector modeling, confronting NID head-on; diffusion-based recommenders [25, 28–30, 49, 50, 52, 54] remove noise under generic reconstruction targets, confronting IDF head-on; recent LLM-based recommenders [1, 11, 24, 27] inject stronger semantic priors into intent extraction, yet still consume unpurified raw sequences; and even diffusion–multi-interest hybrids [22, 36] merely chain the two sides in a single forward pass, lacking a bidirectional mediator. Overall, no existing framework offers a task-grounded shared signal by which denoising and intent extraction iteratively shape each other toward the final objective.

Guided by this diagnosis, we propose **GateDiffInt**, a joint framework that uses the supervisory signal of the final conversion task as a shared guide to align the denoiser Ψ and the intent extractor Φ under a single objective. The framework comprises three cooperating modules (Figure 2). ***(i) Gate-Mediated Controllable Diffusion (GMCD)*** applies behavior-importance-differentiated masks and forward noise across positions, and performs denoising via mask-aware forward diffusion and deterministic DDIM reverse sampling [16, 38]: a *fusion gate* performs position-wise fusion of the denoised and original representations to yield the sequence representation E_{final}, while an interpretable position-importance signal is exposed as a reference for downstream intent extraction. ***(ii) Multi-Intent LLM Distillation (MILD)*** distills a frozen LLM

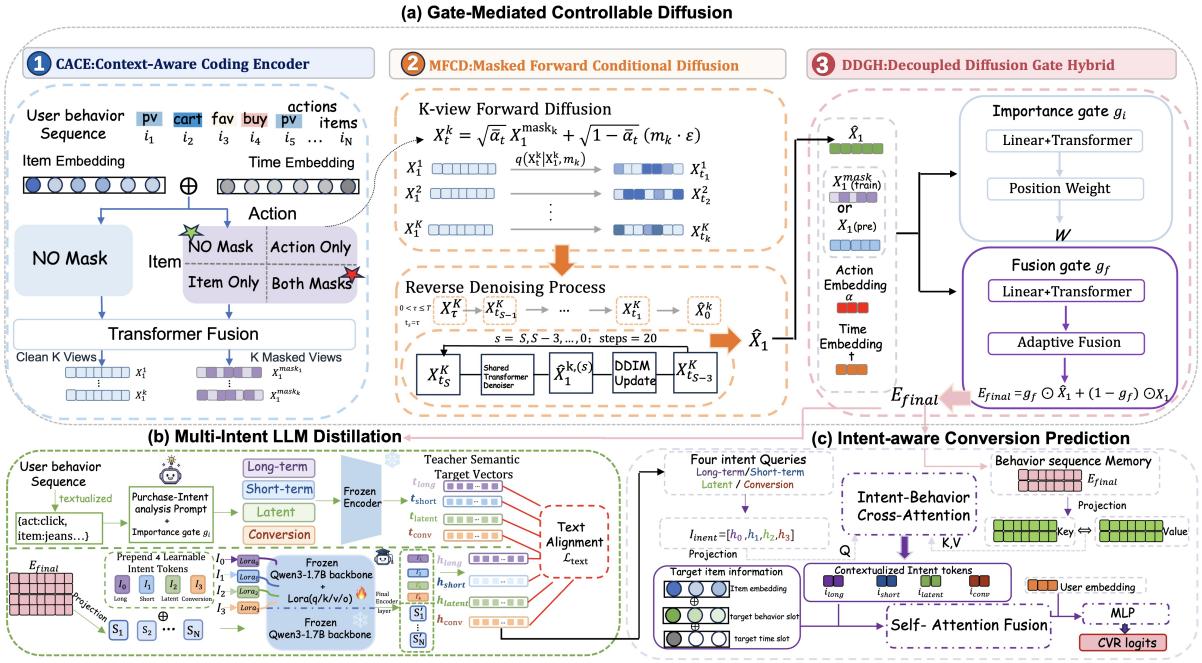

Figure 2: Overall architecture of GateDiffInt: (a) Gate-Mediated Controllable Diffusion Module, (b) Multi-Intent LLM Distillation, and (c) Intent-aware Conversion Prediction.

teacher into a lightweight student that produces four structured intent vectors (long-term, short-term, latent, and conversion)—whose categories follow Goal Systems Theory [20, 21, 40]—and employs per-intent LoRA routing [17] to keep the four heads from collapsing into a single intent, preserving their distinguishability, which we qualitatively illustrate via a t-SNE visualization (Figure 4). ***(iii) Intent-aware Conversion Prediction*** fuses the denoised sequence with the four intents through cross-attention to produce the final conversion prediction.

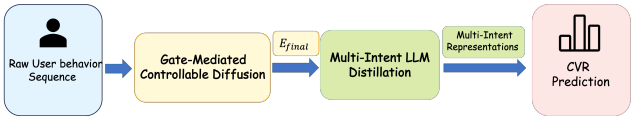

Figure 3: Overall pipeline of GateDiffInt.

Resolving NIC hinges on a two-stage optimization (Figure 1). *In the first stage*, GMCD and MILD are pretrained separately: GMCD, driven by behavior-importance-differentiated masks and forward noise, learns to preserve reliable signals and suppress noise (mitigating IDF), while MILD distills the four structured intents on top of the enhanced representations (mitigating NID). *In the second stage*, a CVR head is attached for joint fine-tuning: the denoising backbone is frozen and only lightweight LoRA is injected into the sequence encoder and the intent extractor, so that the task-specific supervisory signal refines both at once—the encoder, while preserving sequence reconstruction and enhancement, further down-weights unimportant positions, and intent extraction is drawn closer to the conversion objective—thereby aligning the enhanced sequence representation and the structured intents toward a shared objective and closing the mutual-shaping loop between Φ and Ψ. By using a frozen LLM as the intent teacher and obtaining a lightweight student through distillation, MILD injects the world-knowledge priors of LLMs while meeting the online-latency constraints of industrial ranking, which also distinguishes it from paradigms that employ LLMs directly as recommenders [1, 11, 27]. The main contributions of this work are as follows:

- **We identify and empirically characterize *Noise–Intent Coupling* (NIC) in CVR sequence modeling.** To our knowledge, we are the first to explicitly identify the mutual reinforcement in which noise degrades intent and unstructured intent degrades denoising, decomposing it into two separately measurable subphenomena, NID and IDF, and empirically confirming their existence and magnitude through a diagnostic study (Section 4.1.2).
- **We propose GateDiffInt, which resolves NIC via a task-grounded shared signal.** Through a two-stage design, the supervisory signal of the final conversion task serves as a shared guide connecting denoising and intent extraction, refining the sequence encoding and the intent extractor simultaneously via lightweight LoRA so that the two align toward a shared objective.
- **We validate GateDiffInt at both benchmark and industrial scale.** On two public benchmarks and a production dataset from an industrial recommender serving hundreds of millions of daily active users, GateDiffInt consistently outperforms representative sequential and multi-interest ranking baselines under a unified CVR ranking task; component-wise ablations further isolate the individual gains of controllable diffusion and multi-intent LLM

distillation, a 14-day online A/B test corroborates the offline improvements, and the framework has been deployed at scale.

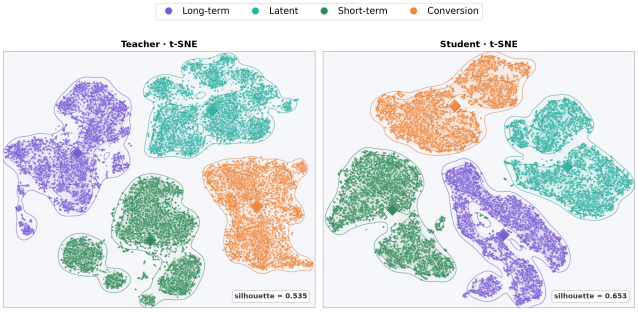

Figure 4: t-SNE of the four distilled intent vectors (long-term, short-term, latent, conversion) on Taobao; the well-separated clusters qualitatively suggest intent disentanglement.

2 Related Work

We review related work along four lines and position GateDiffInt through the lens of NID and IDF.

Sequential and multi-interest models. Beyond feature-interaction models [5, 13, 18, 26, 33, 39, 47, 48], DIN, DIEN, DSIN, BST, and HSTU [4, 10, 55, 58, 59] model sequential dynamics via attention, interest evolution, and Transformers, while MIND, ComiRec, and DMIN [2, 6, 9, 23, 35, 42, 53] further capture interest diversity with multiple latent vectors. However, all of them operate directly on the raw noisy sequence, implicitly assuming that noise can be absorbed by the encoder [3], and their interests emerge at the co-occurrence level without alignment to a semantic structure—hence they are prone to NID under noise. In contrast, GateDiffInt first explicitly purifies the sequence via controllable diffusion and then distills semantically labeled structured intents from an LLM.

Diffusion-based recommenders. DiffRec and DiffuRec [25, 49] pioneer the use of diffusion for sequence denoising, followed by guided, curriculum-based, and conditional variants [28–30, 50, 52, 54]. Yet their forward noising and reconstruction targets are largely position-agnostic and lack fine-grained control aligned with behavioral importance and intent structure; absent an intent prior, they cannot decide which weak signals to keep—i.e., they suffer from IDF. GMCD instead applies behavior-reliability-differentiated masks and forward noise, and fuses the denoised and original representations position-wise via gating, thereby preserving latent-intent weak signals while suppressing noise.

LLMs for recommendation and intent modeling. LLMs have recently been used to unify recommendation paradigms or to enhance sequential models [1, 11, 24, 27], but mainstream practice targets generative recommendation or post-hoc explanation [51, 56], incurring high online latency, still consuming raw sequences, and rarely producing structured intents directly consumable by ranking models. Knowledge distillation [7, 8, 12, 15] can transfer such capabilities but has not been organically combined with sequence denoising. MILD positions the LLM as an offline structured-intent teacher and distills it into a lightweight student via per-intent LoRA, injecting world-knowledge priors while meeting online-latency constraints—fundamentally different from employing LLMs directly as recommenders.

Diffusion–multi-interest hybrids. DiffuMIN and InDiRec [22, 36] combine diffusion with multi-interest modeling and thus touch both the denoising and intent sides, but connect them through a one-way denoise-then-extract pass, lacking a bidirectional mediator for the two to correct each other and thus failing to close the NIC loop. GateDiffInt instead uses the supervisory signal of the final conversion task as a shared guide and, through two-stage optimization, aligns denoising and intent extraction toward a common objective, mechanistically closing the mutual-shaping loop between Φ and Ψ.

3 Methodology

In this section, we briefly review the background and then detail the proposed GateDiffInt framework and its key components.

Let $\mathcal{U}$ and $\mathcal{V}$ denote the sets of users and items, respectively, and let $\mathcal{A}$ be the set of interaction types. For each user $u \in \mathcal{U}$, we take the most recent L interactions in chronological order to form the behavioral sequence

$$\mathcal{S}_u = \{(i_\ell,\, a_\ell,\, \Delta t_\ell)\}_{\ell=1}^{L}, \tag{1}$$

where $i_\ell \in \mathcal{V}$ is the item, $a_\ell \in \mathcal{A}$ the action type, and $\Delta t_\ell \in \mathbb{R}_{\geq 0}$ the time interval to the current timestamp; sequences shorter than L are left-padded. Given a candidate item $v^{\text{tgt}} \in \mathcal{V}$, conversion-rate (CVR) prediction estimates the purchase probability

$$\hat{y} = f_\Theta(\mathcal{S}_u,\, v^{\text{tgt}},\, u) \in [0, 1], \tag{2}$$

where f_Θ is the model parameterized by Θ, optimized by the binary cross-entropy between $\hat{y}$ and the ground-truth label $y \in \{0, 1\}$. Since the behavioral sequence entangles noise with multi-scale intents, GateDiffInt comprises three cooperating modules: **GMCD** performs controllable denoising and passes a fused representation E_{final} together with an interpretable importance signal g_i to the downstream MILD; **MILD** distills four structured intent vectors $\{\mathbf{h}_k\}_{k=1}^{4}$ (long-term, short-term, latent, and conversion) from E_{final}; and an **Intent-aware Conversion Prediction** module fuses them to produce the prediction. The overall data flow is $\mathcal{S}_u \to (E_{\text{final}},\, g_i) \to \{\mathbf{h}_k\} \to \hat{y}$.

3.1 GMCD

User behavioral sequences can be viewed as noisy observations of true preferences: shallow actions are noise-heavy, whereas deep actions, though sparse, carry strong signals. GMCD performs controllable denoising in the latent space and outputs a position-wise aligned fused representation $E_{\text{final}} \in \mathbb{R}^{L \times d}$ as the primary sequence representation for the downstream MILD, together with an interpretable importance signal g_i that the MILD teacher can consult. GMCD consists of three cascaded sub-modules:

- **Context-Aware Coding Encoder (CACE)**: fuses item and temporal information and applies action-aware masking to produce a clean view X_1 and a masked view X_1^{mask};
- **Masked Forward Conditional Diffusion (MFCD)**: performs mask-aware forward diffusion and partial-start deterministic reverse sampling to obtain the denoised reconstruction $\hat{X}_1$;
- **Decoupled Diffusion Gate Hybrid (DDGH)**: fuses $\hat{X}_1$ and X_1 via a fusion gate g_f into E_{final}, and emits an interpretable

Figure 5: Multi-intent behavior timelines of two e-commerce users (Feb–Jun 2026). Right panels show items linked to long-term, short-term, latent, and conversion intents.

importance signal via an importance gate g_i for the downstream MILD teacher.

Training builds self-supervision from masking and noise injection, while inference uses deterministic sampling for efficiency.

3.1.1 CACE. For each position ℓ, the item and temporal embeddings are linearly projected, concatenated with a learnable fusion token, and fed into a position-wise shared Transformer that attends only within each position (never across positions), decoupling feature fusion from sequential modeling. Its output is the fused representation $\mathbf{z}_\ell$, and the clean view is

$$X_1 = [\mathbf{z}_1, \ldots, \mathbf{z}_L] \in \mathbb{R}^{L\times d}. \tag{3}$$

For every non-padding position, CACE independently samples two Bernoulli masks: an item mask for item-level reconstruction and an action mask set by behavioral reliability—higher for shallow actions (e.g., pv) and lower for strong actions (e.g., buy):

$$M_\ell^{\text{item}} \sim \text{Bernoulli}(p^{\text{item}}), M_\ell^{\text{act}} \sim \text{Bernoulli}\big(\pi[a_\ell]\big). \tag{4}$$

Masked items are replaced by a learnable <MASK> token and re-encoded (temporal information retained), yielding the masked view X_1^{mask}. Padding positions are never masked, and the two masks jointly determine the per-position noise strength in the subsequent forward diffusion. We sample K=2 masked views per sequence during training; at inference no masking is applied and X_1 is used as the reverse-sampling start.

3.1.2 MFCD. MFCD runs diffusion in the CACE latent space, sharing one denoising network f_θ across training and inference. Under the DDPM [16] parameterization with a cosine schedule and X_1 as the reconstruction target, each masked view is noised as

$$X_t^k = \sqrt{\bar\alpha_t}\, X_1^{\text{mask}_k} + \sqrt{1-\bar\alpha_t}\cdot\big(m_k(\ell)\,\varepsilon\big), \varepsilon \sim \mathcal{N}(0, I), \tag{5}$$

where the noise multiplier $m_k(\ell)$ is set by the mask state at ℓ (baseline for no mask, strongest for dual mask), so more-masked positions receive stronger noise. The timestep- and position-conditioned f_θ predicts x_0, giving the reconstruction loss

$$\mathcal{L}_{\text{denoise}} = \frac{1}{K}\sum_{k=1}^{K}\big\|\hat X_1^k - \text{sg}[X_1]\big\|_2^2. \tag{6}$$

An InfoNCE regularizer over the denoised outputs of two masked views enhances sequence-level discriminability:

$$\mathcal{L}_{\text{cl}} = -\frac{1}{K(K-1)B}\sum_{b=1}^{B}\sum_{\substack{p,q=1\\ p\neq q}}^{K}\log\frac{\exp\big(\text{sim}(\mathbf{z}_b^{(p)}, \mathbf{z}_b^{(q)})/\tau\big)}{\sum_{b'=1}^{B}\exp\big(\text{sim}(\mathbf{z}_b^{(p)}, \mathbf{z}_{b'}^{(q)})/\tau\big)}, \tag{7}$$

where $\mathbf{z}^{(k)} = \phi(\text{Pool}(\hat X_1^k))$ and sim is cosine similarity. At inference, starting from an intermediate step $s = \lfloor \rho T \rfloor$,

$$X_s = \sqrt{\bar\alpha_s}\, X_1 + \sqrt{1-\bar\alpha_s}\,\varepsilon, \tag{8}$$

a deterministic DDIM [38] update ($\eta = 0$) is applied along $t_1 > t_2 > \cdots > t_{K'} = 0$ (with $t_1 = s$):

$$\hat\varepsilon_{t_i} = \frac{X_{t_i} - \sqrt{\bar\alpha_{t_i}}\,\hat X_1^{(t_i)}}{\sqrt{1-\bar\alpha_{t_i}}}, X_{t_{i+1}} = \sqrt{\bar\alpha_{t_{i+1}}}\,\hat X_1^{(t_i)} + \sqrt{1-\bar\alpha_{t_{i+1}}}\,\hat\varepsilon_{t_i}. \tag{9}$$

The resulting $\hat X_1$ is passed together with X_1 to DDGH.

3.1.3 DDGH. DDGH fuses $\hat X_1$ and X_1 via dual gates that share the same architecture but have independent parameters: $[\hat X_1; X_1^{\text{mask}}; \mathbf{e}^a; \mathbf{e}^t]$ is concatenated and passed through a lightweight Transformer to produce gate values,

$$g = \sigma\Big(\text{MLP}\big(\text{Trans}_{\text{gate}}([\hat X_1;\, X_1^{\text{mask}};\, \mathbf{e}^a;\, \mathbf{e}^t])\big)\Big) \in [0,1]^L. \tag{10}$$

The fusion gate g_f performs position-wise weighted fusion,

$$E_{\text{final}} = g_f \odot \hat X_1 + (1-g_f)\odot X_1, \tag{11}$$

while the importance gate g_i does not participate in the fusion; instead it produces an interpretable position-importance signal provided to the downstream MILD teacher as a reference. Both gates are supervised by self-generated targets with BCE. Let the per-position denoising error be $u_\ell = \|\hat X_{1,\ell} - X_{1,\ell}\|_2^2/d$ with sequence-wise normalized form $\tilde u$; g_f is supervised by $1-\tilde u$, and g_i further incorporates a behavioral prior b_ℓ (λ balances the prior and the reconstruction error):

$$t_\ell^{(i)} = \text{clip}\big(\lambda b_\ell + (1-\lambda)\tilde u_\ell,\ [t_{\min}, t_{\max}]\big), \tag{12}$$

$$\mathcal{L}_{\text{fusion}} = \text{BCE}\big(g_f,\ \text{sg}[1-\tilde u]\big), \mathcal{L}_{\text{imp}} = \text{BCE}\big(g_i,\ \text{sg}[t^{(i)}]\big). \tag{13}$$

The overall GMCD objective sums the denoising, contrastive, and dual-gate losses, where λ_{cl}, λ_f, and λ_i are balancing coefficients:

$$\mathcal{L}_{\text{GMCD}} = \mathcal{L}_{\text{denoise}} + \lambda_{\text{cl}}\mathcal{L}_{\text{cl}} + \lambda_f\mathcal{L}_{\text{fusion}} + \lambda_i\mathcal{L}_{\text{imp}}. \tag{14}$$

GMCD thus completes encoding, diffusion-based denoising, and gated fusion, passing E_{final} and the importance signal g_i to MILD.

3.2 MILD

Users pursue multiple goals of different temporal scales and priorities. Following Goal Systems Theory [21] and subsequent consumer-behavior studies [20, 40], we distinguish four categories of intent (Figure 5):

- **Long-term intent**: rooted in identity and lifestyle, stable across contexts;
- **Short-term intent**: triggered by events or state changes within a clear time window;
- **Latent intent**: reflected by repeated browsing and comparison without conversion;
- **Conversion intent**: the most probable immediate buy decision given prior signals and current context.

On the teacher side, MILD employs a frozen Gemini 3.5 Flash (via API) [44] to generate textual descriptions of the four intents from a user's behavioral sequence as semantic supervision; the importance signal g_i from GMCD is supplied to the teacher as a position-importance hint alongside the sequence, so that the teacher attends more to reliable behaviors when generating the descriptions. On the student side, a frozen Qwen3-1.7B backbone [45] extracts the four intent vectors via per-intent LoRA [17] routing, and Qwen3-Embedding-8B [57] encodes the teacher texts into the semantic space for alignment.

Concretely, K=4 learnable intent tokens $\{I_0, \ldots, I_{K-1}\}$ are prepended, and E_{final} is projected into the LLM latent space and concatenated:

$$\mathbf{H}^{(0)} = [I_0, \ldots, I_{K-1},\ \text{Proj}(E_{\text{final}})], \tag{15}$$

where Proj maps from dimension d to d_{LLM}. A 2D attention mask lets each intent token attend only to itself and all behavior positions (intent tokens are mutually invisible), while behavior positions follow a causal mask and cannot attend to any intent token. During adaptation, LoRA is injected only into the $\{q, k, v, o\}$ projections; the k-th intent token activates only the k-th independent LoRA path, and behavior tokens receive no LoRA updates. The hidden states at the first K positions are taken as the intent vectors:

$$[\mathbf{h}_{\text{long}}, \mathbf{h}_{\text{short}}, \mathbf{h}_{\text{latent}}, \mathbf{h}_{\text{conv}}] = \text{Backbone}(\mathbf{H}^{(0)}, M)[:, :K, :], \tag{16}$$

where M is the attention mask above.

Teacher descriptions are encoded by Qwen3-Embedding-8B into targets $\mathbf{t}^{\text{tea}}_{b,k}$; student intents are mapped by a projection head $\text{Head}(\cdot)$ into the teacher embedding space and aligned via a masked cosine distance:

$$\mathcal{L}_{\text{MILD}} = \frac{1}{\sum_{b,k} m_{b,k}} \sum_{b,k} m_{b,k} \Big(1 - \cos\left(\text{Head}(\mathbf{h}_{b,k}),\ \mathbf{t}^{\text{tea}}_{b,k}\right)\Big), \tag{17}$$

where b indexes samples in a batch, k indexes intent categories, and $m_{b,k} \in \{0, 1\}$ indicates whether valid supervision is available; unsupervised positions are excluded from the loss.

MILD thus extracts the four structured intent vectors via per-intent LoRA routing and semantic distillation, and passes them together with E_{final} to the downstream intent-aware conversion prediction module.

3.3 Intent-aware Conversion Prediction

Given E_{final} and the four intent vectors $\{\mathbf{h}_k\}_{k=1}^4$, this module fuses them with candidate item information to produce the conversion probability $\hat{y}$.

Cross-attention is performed with intents as queries and the behavioral sequence as keys and values. On the sequence side, E_{final} is concatenated with behavior embeddings and normalized; on the intent side, a linear transform with nonlinear activation yields the query, followed by multi-head cross-attention to obtain contextualized intents:

$$\mathbf{e}^{\text{new}}_\ell = \text{LN}\big([W_e E_{\text{final},\ell};\ \mathbf{e}^{\text{beh}}_\ell]\big), \tag{18}$$

$$\mathbf{i} = \text{LN}\big(\phi(W_i \mathbf{h})\big), \tag{19}$$

$$\mathbf{i}^{\text{out}} = \text{LN}\big(\mathbf{i} + \text{CrossAttn}(Q = \mathbf{i},\ K = V = \mathbf{e}^{\text{new}})\big). \tag{20}$$

The contextualized intents $\{\mathbf{i}^{\text{out}}_k\}_{k=1}^4$ are concatenated with candidate item features and passed through self-attention to obtain the fused representation $\mathbf{T}^{\text{out}}$, which is then fed together with the user embedding into an MLP to produce the final prediction under the binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$:

$$\hat{y} = \sigma\big(\text{MLP}\big([\text{flatten}(\mathbf{T}^{\text{out}});\ \mathbf{e}^{\text{user}}]\big)\big). \tag{21}$$

GateDiffInt is trained in two stages (Figure 1): GMCD and MILD are first pretrained separately; checkpoints are then loaded and a CVR head is attached for end-to-end joint fine-tuning. During fine-tuning, the GMCD backbone is frozen and only a low-rank LoRA is injected into the encoder fusion Transformer; MILD retains its per-intent MultiLoRA; the CVR head is fully trainable. To preserve the distilled intent geometry, a teacher-anchoring regularizer is introduced:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{BCE}} + \alpha \sum_{k=1}^{4} \big(1 - \cos(\mathbf{h}^{\text{stu}}_k,\ \text{sg}[\mathbf{h}^{\text{tea}}_k])\big), \tag{22}$$

where $\mathbf{h}^{\text{tea}}_k$ comes from the frozen MILD copy after the first stage, and α is a balancing coefficient. Since DDIM reverse sampling does not propagate gradients and the fusion gate is frozen at this stage, CVR gradients reach the encoder LoRA only through the $(1 - g_f) \odot X_1$ branch, enabling lightweight refinement.

4 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of GateDiffInt[1].

4.1 Experimental Setup

4.1.1 Datasets. We evaluate on two public benchmarks and one industrial dataset. **Taobao**[2] records behavior logs (click, add-to-cart, favorite, and buy) of about one million users from Nov. 25 to Dec. 3, 2017, yielding **1.67M** training and **0.53M** test samples over **5** features; positives are real buy behaviors, and negatives are sampled at a 1:1 ratio, preferentially as hard negatives from interacted-but-unpurchased items. **Amazon-Electronics**[3] [14] consists of review records, from which item embeddings are obtained directly from review texts via a large language model, giving **1.56M**/**0.30M**

[1] Code is available at: https://github.com/chen667/GateDiffInt
[2] https://tianchi.aliyun.com/dataset/dataDetail?dataId=649
[3] https://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Electronics_5.json.gz

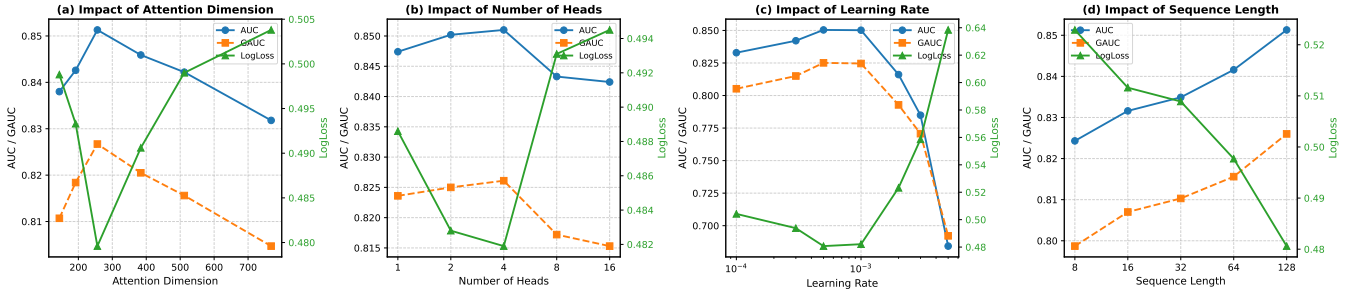

Figure 6: Hyper-parameter sensitivity of our model on the Taobao dataset: (a) attention dimension, (b) number of heads, (c) learning rate, and (d) sequence length. Left axis: AUC and GAUC; right axis: LogLoss.

train/test samples over **9** features; ratings ≥ 4 are treated as positive, and negatives are drawn by popularity$^{0.75}$ sampling [34] from non-interacted items. **Industrial** is sampled from the click logs of a large-scale recommender serving hundreds of millions of users, using one full day of logs for training/validation and a subset of the next day for testing, yielding **20M**/**2M** train/test samples over **56** features. To prevent user-level leakage, all datasets reserve disjoint user sets exclusively for GMCD pre-training and MILD distillation, with the remaining users forming the CVR evaluation set under the leave-one-out protocol [19, 41].

4.1.2 NIC Diagnostic Analysis. To show that Noise–Intent Coupling (NIC) is a *measurable* phenomenon rather than a conceptual abstraction, we isolate its two sub-phenomena—NID and IDF—on the public Taobao benchmark (3 seeds; mean±std). **NID (Noise-induced Intent Distortion).** We inject replacement noise at ratio $\gamma \in \{0, 0.1, \ldots, 0.5\}$ into each test sequence, replacing a γ fraction of non-padding, non-purchase positions with popularity-sampled random items while protecting the target and strong behaviors. Using the clean (γ=0) output as reference, we measure the intent drift

$$D(\gamma) = \frac{1}{K} \sum_{k=1}^{K} \Big(1 - \cos\big(\mathbf{h}_k(\mathcal{S}_0),\, \mathbf{h}_k(\mathcal{S}_\gamma)\big)\Big), \tag{23}$$

over the K=4 intent vectors. The denoising-free variant (*w/o GMCD*) drifts sharply and monotonically as noise grows, whereas GateDiffInt stays markedly stable—its drift is consistently ~6.5× lower across all γ—confirming both the existence of NID and its mitigation by our task-grounded denoising. **IDF (Intent-agnostic Denoising Failure).** On a weak-signal subset (cart/fav interactions occurring more than the median interval before the final purchase; $N_w \approx 1.8\times10^5$), we compare our controllable diffusion against a *vanilla* diffusion that is identical in encoder and training data but strips all task grounding (plain DDPM, no gating/conditioning). We report the weak-signal reconstruction fidelity $\cos(\hat{X}_{\text{denoised}}, X_{\text{clean}})$. Vanilla diffusion attains only 0.44—its intent-agnostic denoising severely distorts the encoded weak signals—whereas GateDiffInt faithfully preserves them at 0.98, a 2.2× improvement, confirming IDF and its mitigation. As weak-signal positions are only a small fraction of each sequence, this large fidelity gap is diluted into the moderate AUC gap seen in Table 3. These diagnostics empirically ground NIC: noise distorts intent when denoising is absent (NID), and generic denoising erases informative weak signals when it is not intent-grounded (IDF)—both addressed by GateDiffInt.

4.1.3 Baselines. We compare GateDiffInt against representative sequential baselines under a unified CVR ranking task:

- **DIN, DIEN, DSIN** [10, 58, 59]: target attention and interest evolution over historical behaviors;
- **BST, HSTU** [4, 55]: Transformer self-attention for long-range dependencies;
- **DMIN** [53]: multi-vector modeling of interest diversity.

We do not include diffusion-based/LLM-based recommenders as standalone baselines, as they target top-k or generative recommendation and are misaligned with target-conditioned pointwise CVR ranking; instead, the benefits of controllable diffusion and multi-intent LLM distillation are isolated via component-wise ablations. All baselines follow the implementation practice of FuxiCTR [61], use the same input features and sequence length as ours, and are optimized with BCE; since Taobao has few features, we additionally train an item encoder on it.

4.1.4 Evaluation Metrics. We adopt AUC (Area Under the ROC Curve), GAUC (Group AUC), and LogLoss as the offline evaluation metrics, and GMV (Gross Merchandise Volume) as the business metric for online A/B testing. AUC measures overall ranking ability, LogLoss measures the calibration of predicted probabilities, while GAUC further evaluates ranking quality within each user group and has been shown to be more consistent with online performance [37, 60]. Its calculation is given in Equation (24):

$$\text{GAUC} = \frac{\sum_{u=1}^{U} \#samples(u) \times \text{AUC}_u}{\sum_{u=1}^{U} \#samples(u)}. \tag{24}$$

where U denotes the number of users, $\#samples(u)$ is the number of samples for the u-th user, and AUC_u is the AUC computed on the samples of user u.

4.1.5 Implementation Details. All methods use the same backbone architectures as in the original papers. We optimize all models with Adam using a learning rate of 5×10^{-4}. Each model is trained for up to 15 epochs with early stopping (patience = 4). The batch size is 2048 for baselines, 96 for GateDiffInt on public datasets, and 64 for GateDiffInt on the production dataset. All models are implemented in PyTorch. The maximum sequence length is 128 for Taobao (average length ≈ 101), 32 for Amazon-Electronics (average

Table 1: Performance comparison on the public datasets. The best results are highlighted in bold, and the second-best results are underlined. Improvements over the strongest baseline are statistically significant with p-value < 0.01 under pairwise t-tests.

Model	Taobao			Amazon-Electronics		
	AUC	GAUC	LogLoss	AUC	GAUC	LogLoss
DIN	0.8068±0.0039	0.7919±0.0035	0.5353±0.0039	0.7141±0.0105	0.6996±0.0106	0.6228±0.0091
DIEN	0.8388±0.0011	0.8152±0.0010	0.4974±0.0031	0.7523±0.0037	0.7399±0.0037	0.6002±0.0058
DSIN	0.8254±0.0073	0.8033±0.0061	0.5134±0.0078	0.7612±0.0021	0.7489±0.0025	0.5915±0.0049
BST	0.8378±0.0021	0.8126±0.0012	0.5005±0.0015	0.7619±0.0050	0.7491±0.0048	0.5889±0.0077
DMIN	0.8397±0.0013	0.8174±0.0039	0.4963±0.0040	0.7636±0.0033	0.7527±0.0026	0.5878±0.0068
HSTU	0.8304±0.0032	0.8075±0.0031	0.5064±0.0043	0.7829±0.0016	0.7681±0.0019	0.5693±0.0047
Ours	**0.8515±0.0014**	**0.8275±0.0013**	**0.4836±0.0018**	**0.8016±0.0076**	**0.7792±0.0033**	**0.5414±0.0046**

length ≈ 9), and 1024 for the production dataset. For the diffusion process, timesteps are uniformly sampled from 1 to 1000 in the forward process, and the reverse process starts at timestep 700 with DDIM sampling [16, 38].

4.2 Performance on Public Datasets

4.2.1 Overall Performance. Table 1 reports the overall performance on the two public datasets. Sequential models consistently outperform DIN on both Taobao and Amazon-Electronics, with stronger sequential modeling yielding better results. On Taobao, DIEN and DMIN are the strongest baselines, while HSTU performs best on Amazon-Electronics. Notably, HSTU underperforms on Taobao due to limited feature dimensions and short sequence length, but benefits from richer review-based embeddings on Amazon-Electronics. GateDiffInt achieves the best results across all metrics on both datasets, with statistically significant improvements over the strongest baselines, demonstrating the effectiveness of unifying controllable diffusion-based sequence enhancement with multi-intent LLM distillation.

4.2.2 Impact of Hyperparameters. For simplicity, we study the sensitivity of GateDiffInt to key hyperparameters on the Taobao dataset, including attention dimension, number of attention heads, learning rate, and sequence length (Figure 6).

The learning rate is the most sensitive: the model performs best within 5×10^{-4} to 1×10^{-3}, while lower rates cause underfitting and higher rates ($\geq 2 \times 10^{-3}$) lead to instability. The attention dimension shows an inverted-U curve peaking at 256. The model is relatively robust to the number of heads (best at 4), and performance improves steadily with longer sequences. Overall, the default setting (dimension=256, heads=4, learning rate=5×10^{-4}) lies in a favorable range.

4.3 Performance on Industrial Dataset

Table 2 reports the results on the production dataset. Baseline performance improves with stronger sequential modeling, and HSTU achieves the best results among all baselines. Unlike its weaker performance on Taobao, HSTU shows a clear advantage here, benefiting from the longer sequences and richer features in the production data. GateDiffInt significantly outperforms all baselines, achieving relative improvements of 3.00% in AUC, 1.82% in GAUC, and 2.81% in LogLoss over HSTU. These consistent gains in both ranking quality and calibration further validate the effectiveness of our approach in real-world industrial scenarios.

Table 2: Performance comparison on the production dataset. The best results are highlighted in bold. Improvements over the strongest baseline are statistically significant with p-value < 0.01 under pairwise t-tests.

Model	AUC	GAUC	LogLoss
DIN	0.7682 ± 0.0026	0.7053 ± 0.0029	0.6078 ± 0.0034
DIEN	0.7712 ± 0.0033	0.7111 ± 0.0024	0.5915 ± 0.0027
DSIN	0.7843 ± 0.0021	0.7246 ± 0.0017	0.5883 ± 0.0011
BST	0.7978 ± 0.0014	0.7352 ± 0.0018	0.5626 ± 0.0016
DMIN	0.8026 ± 0.0022	0.7398 ± 0.0014	0.5437 ± 0.0013
HSTU	0.8239 ± 0.0019	0.7476 ± 0.0024	0.5118 ± 0.0015
Ours	**0.8486 ± 0.0020**	**0.7612 ± 0.0012**	**0.4974 ± 0.0014**

4.4 Online Deployment

GateDiffInt has been deployed in the ranking stage of the home feed of a major industrial platform serving hundreds of millions of daily active users. The system follows a two-stage paradigm: GMCD and MILD are pre-trained and updated weekly, while they are jointly fine-tuned with the CVR model via LoRA on a daily basis—the weekly cadence suffices because sequence enhancement and intent extraction capture relatively stable patterns, whereas the daily fine-tuning adapts to the latest data distribution. At serving time, the latest user behaviors are retrieved in real time and passed sequentially through GMCD, the MILD student, and the CVR head to produce the final scores. In a 14-day online A/B test against the production baseline, GateDiffInt achieves a statistically significant **GMV** improvement of **+1.13%** in the home-feed scenario, and has since been rolled out to other businesses within the same scenario, bringing a cumulative **GMV** gain of **+5.13%** to the feed scenario. These results confirm the practical value and deployability of GateDiffInt at industrial scale.

4.5 Ablation Study

To assess each component, we ablate GateDiffInt on Taobao (Table 3). Removing both GMCD and MILD causes the largest drop, and removing either alone also degrades performance—MILD more so. The joint drop is markedly larger than the sum of the two individual drops, indicating that the two modules are complementary rather

than redundant. Within GMCD, replacing controllable diffusion with a vanilla variant, or disabling the fusion gate or importance gate, all hurt performance, with the fusion gate contributing most. Within MILD, collapsing the four per-intent LoRAs into one, or the four intent types into a single type, likewise degrades results. These confirm the necessity of gate-mediated control and multi-intent disentanglement.

Table 3: Ablation study results in Taobao Dataset.

Variant	AUC	GAUC	LogLoss
GateDiffInt	**0.8515**	**0.8275**	**0.4836**
GateDiffInt w/ Vanilla Diffusion	0.8326	0.8158	0.5112
GateDiffInt w/o GMCD & MILD	0.7154	0.6295	0.7128
GateDiffInt w/o GMCD	0.8313	0.8139	0.5241
GateDiffInt w/o MILD	0.8278	0.8037	0.5092
GateDiffInt w/o Fusion Gate	0.8358	0.8170	0.5073
GateDiffInt w/o Importance Gate	0.8424	0.8215	0.4983
GateDiffInt w/o Multi-LoRA	0.8467	0.8253	0.4981
GateDiffInt w/o Multi-Intent	0.8374	0.8133	0.5022

5 Conclusion

In this paper, we identify and empirically characterize *Noise–Intent Coupling* (NIC) in CVR sequence modeling—the mutual reinforcement between noise-induced intent distortion (NID) and intent-agnostic denoising failure (IDF)—and propose GateDiffInt, which uses the supervisory signal of the final conversion task as a shared guide to align Gate-Mediated Controllable Diffusion (GMCD) with Multi-Intent LLM Distillation (MILD). Offline experiments on public and industrial datasets and a 14-day online A/B test show that GateDiffInt consistently improves ranking quality and calibration while delivering real business value. Future work will extend NIC to multi-task CVR/CTR modeling [31, 43] and further develop diagnostic characterizations of NID and IDF.

References

[1] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song (Eds.). ACM, 1007–1014. doi:10.1145/3604915.3608857

[2] Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Controllable Multi-Interest Framework for Recommendation. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 2942–2951. doi:10.1145/3394486.3403344

[3] Hong Chen, Yudong Chen, Xin Wang, Ruobing Xie, Rui Wang, Feng Xia, and Wenwu Zhu. 2021. Curriculum Disentangled Recommendation with Noisy Multi-feedback. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 26924–26936. https://proceedings.neurips.cc/paper/2021/hash/e242660df1b69b74dcc7fde711f924ff-Abstract.html

[4] Qiwei Chen, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. 2019. Behavior Sequence Transformer for E-commerce Recommendation in Alibaba. *CoRR* abs/1905.06874 (2019). arXiv:1905.06874 http://arxiv.org/abs/1905.06874

[5] Weiyu Cheng, Yanyan Shen, and Linpeng Huang. 2020. Adaptive Factorization Network: Learning Adaptive-Order Feature Interactions. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 3609–3616. doi:10.1609/AAAI.V34I04.5768

[6] Minjin Choi, Hye-young Kim, Hyunsouk Cho, and Jongwuk Lee. 2024. Multi-intent-aware Session-based Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 2532–2536. doi:10.1145/3626772.3657928

[7] Yu Cui, Feng Liu, Pengbo Wang, Bohao Wang, Heng Tang, Yi Wan, Jun Wang, and Jiawei Chen. 2024. Distillation Matters: Empowering Sequential Recommenders to Match the Performance of Large Language Models. In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024*, Tommaso Di Noia, Pasquale Lops, Thorsten Joachims, Katrien Verbert, Pablo Castells, Zhenhua Dong, and Ben London (Eds.). ACM, 507–517. doi:10.1145/3640457.3688118

[8] Yingpeng Du, Zhu Sun, Ziyan Wang, Haoyan Chua, Jie Zhang, and Yew-Soon Ong. 2025. Active large language model-based knowledge distillation for session-based recommendation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25)*. AAAI Press, Article 1290, 9 pages. doi:10.1609/aaai.v39i11.33263

[9] Yingpeng Du, Ziyan Wang, Zhu Sun, Yining Ma, Hongzhi Liu, and Jie Zhang. 2024. Disentangled Multi-interest Representation Learning for Sequential Recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, Ricardo Baeza-Yates and Francesco Bonchi (Eds.). ACM, 677–688. doi:10.1145/3637528.3671800

[10] Yufei Feng, Fuyu Lv, Weichen Shen, Menghan Wang, Fei Sun, Yu Zhu, and Keping Yang. 2019. Deep Session Interest Network for Click-Through Rate Prediction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 2301–2307. doi:10.24963/IJCAI.2019/319

[11] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). In *RecSys '22: Sixteenth ACM Conference on Recommender Systems, Seattle, WA, USA, September 18 - 23, 2022*, Jennifer Golbeck, F. Maxwell Harper, Vanessa Murdock, Michael D. Ekstrand, Bracha Shapira, Justin Basilico, Keld T. Lundgaard, and Even Oldridge (Eds.). ACM, 299–315. doi:10.1145/3523227.3546767

[12] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge Distillation: A Survey. *Int. J. Comput. Vis.* 129, 6 (2021), 1789–1819. doi:10.1007/S11263-021-01453-Z

[13] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 1725–1731. doi:10.24963/ijcai.2017/239

[14] Ruining He and Julian J. McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao (Eds.). ACM, 507–517. doi:10.1145/2872427.2883037

[15] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR* abs/1503.02531 (2015). arXiv:1503.02531 http://arxiv.org/abs/1503.02531

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=nZeVKeeFYf9

[18] Tongwen Huang, Zhiqi Zhang, and Junlin Zhang. 2019. FiBiNET: combining feature importance and bilinear feature interaction for click-through rate prediction. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). ACM, 169–177. doi:10.1145/3298689.3347043

[19] Wang-Cheng Kang and Julian J. McAuley. 2018. Self-Attentive Sequential Recommendation. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*. IEEE Computer Society, 197–206. doi:10.1109/ICDM.2018.00035

[20] Catalina E. Kopetz, Arie W. Kruglanski, Zachary G. Arens, Jordan Etkin, and Heather M. Johnson. 2012. The dynamics of consumer behavior: A goal systemic perspective. *Journal of Consumer Psychology* 22, 2 (2012), 208–223. arXiv:https://myscp.onlinelibrary.wiley.com/doi/pdf/10.1016/j.jcps.2011.03.001 doi:10.1016/j.jcps.2011.03.001

[21] Arie W. Kruglanski, James Y. Shah, Ayelet Fishbach, Ron Friedman, Woo Young Chun, and David Sleeth-Keppler. 2002. A theory of goal systems. Advances in Experimental Social Psychology, Vol. 34. Academic Press, 331–378. doi:10.1016/S0065-2601(02)80008-9

[22] Weijiang Lai, Beihong Jin, Yapeng Zhang, Yiyuan Zheng, Rui Zhao, Jian Dong, Jun Lei, and Xingxing Wang. 2025. Modeling Long-term User Behaviors with Diffusion-driven Multi-interest Network for CTR Prediction. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems, RecSys 2025, Prague, Czech Republic, September 22-26, 2025*, Mária Bieliková, Pavel Kordík, Markus Schedl, Marco de Gemmis, Sole Pera, Rodrigo Alves, Olivier Jeunen, and Vito Ostuni (Eds.). ACM, 289–298. doi:10.1145/3705328.3748045

[23] Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-Interest Network with Dynamic Routing for Recommendation at Tmall. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 2615–2623. doi:10.1145/3357384.3357814

[24] Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian J. McAuley. 2023. Text Is All You Need: Learning Language Representations for Sequential Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye (Eds.). ACM, 1258–1267. doi:10.1145/3580305.3599519

[25] Zihao Li, Aixin Sun, and Chenliang Li. 2024. DiffuRec: A Diffusion Model for Sequential Recommendation. *ACM Trans. Inf. Syst.* 42, 3 (2024), 66:1–66:28. doi:10.1145/3631116

[26] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. 2018. xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Yike Guo and Faisal Farooq (Eds.). ACM, 1754–1763. doi:10.1145/3219819.3220023

[27] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. 2024. LLaRA: Large Language-Recommendation Assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 1785–1795. doi:10.1145/3626772.3657690

[28] Xiao Lin, Xiaokai Chen, Chenyang Wang, Hantao Shu, Linfeng Song, Biao Li, and Peng Jiang. 2024. Discrete Conditional Diffusion for Reranking in Recommendation. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw (Eds.). ACM, 161–169. doi:10.1145/3589335.3648313

[29] Qidong Liu, Fan Yan, Xiangyu Zhao, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and Feng Tian. 2023. Diffusion Augmentation for Sequential Recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos (Eds.). ACM, 1576–1586. doi:10.1145/3583780.3615134

[30] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Zhanhui Kang. 2024. Plug-In Diffusion Model for Sequential Recommendation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 8886–8894. doi:10.1609/AAAI.V38I8.28736

[31] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 1930–1939. doi:10.1145/3219819.3220007

[32] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. 2018. Entire Space Multi-Task Model: An Effective Approach for Estimating Post-Click Conversion Rate. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 1137–1140. doi:10.1145/3209978.3210104

[33] Kelong Mao, Jieming Zhu, Liangcai Su, Guohao Cai, Yuru Li, and Zhenhua Dong. 2023. FinalMLP: An Enhanced Two-Stream MLP Model for CTR Prediction. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 4552–4560. doi:10.1609/AAAI.V37I4.25577

[34] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 3111–3119. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

[35] Aditya Pal, Chantat Eksombatchai, Yitong Zhou, Bo Zhao, Charles Rosenberg, and Jure Leskovec. 2020. PinnerSage: Multi-Modal User Embedding Framework for Recommendations at Pinterest. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 2311–2320. doi:10.1145/3394486.3403280

[36] Yuanpeng Qu and Hajime Nobuhara. 2025. 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, SIGIR 2025, Padua, Italy, July 13-18, 2025*, Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne (Eds.). ACM, 1552–1561. doi:10.1145/3726302.3730010

[37] Xiang-Rong Sheng, Liqin Zhao, Guorui Zhou, Xinyao Ding, Binding Dai, Qiang Luo, Siran Yang, Jingshan Lv, Chi Zhang, Hongbo Deng, and Xiaoqiang Zhu. 2021. One Model to Serve All: Star Topology Adaptive Recommender for Multi-Domain CTR Prediction. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong (Eds.). ACM, 4104–4113. doi:10.1145/3459637.3481941

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=St1giarCHLP

[39] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. 2019. AutoInt: Automatic Feature Interaction Learning via Self-Attentive Neural Networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 1161–1170. doi:10.1145/3357384.3357925

[40] Jacob Suher, Szu-chi Huang, and Leonard Lee. 2019. Planning for Multiple Shopping Goals in the Marketplace. *Journal of Consumer Psychology* 29, 4 (2019), 642–651. arXiv:https://myscp.onlinelibrary.wiley.com/doi/pdf/10.1002/jcpy.1130 doi:10.1002/jcpy.1130

[41] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 1441–1450. doi:10.1145/3357384.3357895

[42] Qiaoyu Tan, Jianwei Zhang, Jiangchao Yao, Ninghao Liu, Jingren Zhou, Hongxia Yang, and Xia Hu. 2021. Sparse-Interest Network for Sequential Recommendation. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, Liane Lewin-Eytan, David Carmel, Elad Yom-Tov, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 598–606. doi:10.1145/3437963.3441811

[43] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 269–278. doi:10.1145/3383313.3412236

[44] Gemini Team. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *CoRR* abs/2507.06261 (2025). arXiv:2507.06261 doi:10.48550/ARXIV.2507.06261

[45] Qwen Team. 2025. Qwen3 Technical Report. *CoRR* abs/2505.09388 (2025). arXiv:2505.09388 doi:10.48550/ARXIV.2505.09388

[46] Hao Wang, Tai-Wei Chang, Tianqiao Liu, Jianmin Huang, Zhichao Chen, Chao Yu, Ruopeng Li, and Wei Chu. 2022. ESCM2: Entire Space Counterfactual Multi-Task Model for Post-Click Conversion Rate Estimation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio

Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 363–372. doi:10.1145/3477495.3531972

[47] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & Cross Network for Ad Click Predictions. In *Proceedings of the ADKDD'17, Halifax, NS, Canada, August 13 - 17, 2017*. ACM, 12:1–12:7. doi:10.1145/3124749.3124754

[48] Ruoxi Wang, Rakesh Shivanna, Derek Zhiyuan Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed H. Chi. 2021. DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 1785–1797. doi:10.1145/3442381.3450078

[49] Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. 2023. Diffusion Recommender Model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 832–841. doi:10.1145/3539618.3591663

[50] Cheng Wu, Liang Su, Chaokun Wang, Shaoyun Shi, Ziqian Zhang, Ziyang Liu, Wang Peng, Wenjin Wu, and Peng Jiang. 2025. Learning Multiple User Distributions for Recommendation via Guided Conditional Diffusion. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, Toby Walsh, Julie Shah, and Zico Kolter (Eds.). AAAI Press, 12845–12853. doi:10.1609/AAAI.V39I12.33401

[51] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2024. A survey on large language models for recommendation. *World Wide Web (WWW)* 27, 5 (2024), 60. doi:10.1007/S11280-024-01291-2

[52] Zihao Wu, Xin Wang, Hong Chen, Kaidong Li, Yi Han, Lifeng Sun, and Wenwu Zhu. 2023. Diff4Rec: Sequential Recommendation with Curriculum-scheduled Diffusion Augmentation. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdulmotaleb El Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 9329–9335. doi:10.1145/3581783.3612709

[53] Zhibo Xiao, Luwei Yang, Wen Jiang, Yi Wei, Yi Hu, and Hao Wang. 2020. Deep Multi-Interest Network for Click-through Rate Prediction. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 2265–2268. doi:10.1145/3340531.3412092

[54] Zhengyi Yang, Jiancan Wu, Zhicai Wang, Xiang Wang, Yancheng Yuan, and Xiangnan He. 2023. Generate What You Prefer: Reshaping Sequential Recommendation via Guided Diffusion. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/4c5e2bcbf21bdf40d75fddad0bd43dc9-Abstract-Conference.html

[55] Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Jiayuan He, Yinghai Lu, and Yu Shi. 2024. Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024 (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR / OpenReview.net, 58484–58509. https://proceedings.mlr.press/v235/zhai24a.html

[56] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2024. Large Language Models for Recommendation: Progresses and Future Directions. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw (Eds.). ACM, 1268–1271. doi:10.1145/3589335.3641247

[57] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *CoRR* abs/2506.05176 (2025). arXiv:2506.05176 doi:10.48550/ARXIV.2506.05176

[58] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep Interest Evolution Network for Click-Through Rate Prediction. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 5941–5948. doi:10.1609/AAAI.V33I01.33015941

[59] Guorui Zhou, Xiaoqiang Zhu, Chengru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep Interest Network for Click-Through Rate Prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Yike Guo and Faisal Farooq (Eds.). ACM, 1059–1068. doi:10.1145/3219819.3219823

[60] Han Zhu, Junqi Jin, Chang Tan, Fei Pan, Yifan Zeng, Han Li, and Kun Gai. 2017. Optimized Cost per Click in Taobao Display Advertising. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. ACM, 2191–2200. doi:10.1145/3097983.3098134

[61] Jieming Zhu, Jinyang Liu, Shuai Yang, Qi Zhang, and Xiuqiang He. 2020. FuxiCTR: An Open Benchmark for Click-Through Rate Prediction. *CoRR* abs/2009.05794 (2020). arXiv:2009.05794 https://arxiv.org/abs/2009.05794