FROM NEXT-ITEM TO MULTI-USER GENERATION: REFORMULATING COLD-START RECOMMENDATION VIA PARALLEL GENERATIVE DISCRETE DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most existing recommender systems are built from a user-centric perspective, aiming to recommend the most suitable content to each user. However, this paradigm often fails in cold-start scenarios, where the paucity of interaction feedback compromises both reliable representation learning and sufficient initial exposure within massive candidate pools. This paper advocates a paradigm shift from "Next-Item" to "Multi-User Generation" after revisiting the underlying logic of cold-start recommendation, and reformulates the task as **Item-centric Multi-User Generation**, aiming to proactively identify seed users to break the passive exposure deadlock. However, this transition is challenging, as the multi-modal target user distributions and the non-sequential nature of interactions pose a severe inductive bias mismatch for traditional user-centric methods. To address this challenge, we propose **P**arallel Generative **D**iscrete **D**iffusion (**PD^2**), which employs a permutation-invariant Masked Discrete Diffusion mechanism to construct a direct mapping from item content to the latent user space, while eschewing autoregressive constraints. We also develop Interval-Aware Spatio-Temporal Encoding, incorporating temporal biasing with positional encoding to capture the temporal evolution of group interests and ensure multi-user extrapolation. With this design, PD^2 empowers effective *Multi-Group Look-ahead Prediction* to explore diverse interest groups and *Scalable Parallel Retrieval* to ensure high-efficiency decoding. Extensive experiments on two public real-world datasets and large-scale online A/B tests demonstrate that PD^2 outperforms existing methods, especially in cold-start scenarios.

1 INTRODUCTION

In recent years, personalized recommendation systems have served as an indispensable facility for mitigating information overload. With the advent of deep learning, sophisticated user-centric paradigms, such as Collaborative Filtering (CF) He et al. (2017), Graph Neural Networks He et al. (2020); Ying et al. (2018), and Sequential Modeling Kang & McAuley (2018); Sun et al. (2019), have achieved remarkable success by mining latent patterns from rich historical behaviors and implicit feedback. For online platforms, continuous injection of new items is critical to maintain system vibrancy and satisfy users' continually evolving interests. However, this necessity conflicts with the limitations of these user-centric paradigms Luo et al. (2025); Chen et al. (2024). Since new items lack sufficient historical interaction data, the models struggle to learn meaningful representations or capture latent user-item associations, inevitably triggering the *Matthew Effect*: exposure is unfairly skewed toward popular items via cumulative advantage, creating a feedback cycle that continuously marginalizes high-quality fresh content. This degrades the platform ecosystem by stifling new item survival and ultimately compromises user retention and long-term revenue.

Current approaches mostly utilize auxiliary modalities to bridge the gap left by missing interaction data Chen et al. (2024); Kim et al. (2024). While improving representation quality, these strategies are still bound by the user-centric User-to-Item (U2I) paradigm, where cold-start items function as passive targets submerged in massive candidate pools and consequently suffer from severe exposure suppression. This mechanism limitation leaves the root cause of the *Matthew Effect* unaddressed, thereby failing to guarantee the essential initial exposure for high-quality new items.

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.



(a) User-to-Item (U2I) (b) Item-to-User (I2U)

Figure 1: Paradigm comparison: U2I vs. I2U. (a) User-centric sequential modeling driven by the strong temporal causality of individual behaviors. (b) Item-centric content dissemination, characterized by the concurrent, multi-modal distributions of diverse user groups.

In this context, we argue that the underlying logic of cold-start recommendation should be revisited for motivating a paradigm shift from "Next-Item" to "Multi-User Generation"[1]. This is to transfer the focus from passive U2I retrieval to active Item-to-User (I2U) dissemination, i.e., generating diverse potential users given a cold-start item. On the one hand, leveraging users as supervision targets encourages the model to learn a direct projection from item to user space, thereby accelerating the learning of reliable cold-start item representations. On the other hand, it better matches the dissemination demand of cold-start items. By proactively identifying seed users, I2U fundamentally mitigates the *Matthew Effect* induced by data sparsity and promotes the breakout of new items.

However, this transition is definitely non-trivial. As shown in Fig. 1, distinct from user-side sequences driven by temporal causality, item-side interactions, even if logged sequentially, essentially result from the simultaneous dissemination to diverse user groups. Existing mirror-based strategies Chen et al. (2025b) fail to handle these unique characteristics of I2U cold-start recommendation: (1) **Multi-modality distribution.** Cold-start scenarios necessitate models of strong generalization to precisely capture the multi-modal distribution of diverse potential users from limited interaction histories. (2) **Non-sequential nature.** Unlike U2I sequences driven by strong temporal causality, the target user set in I2U is essentially unordered and concurrent. This necessitates bypassing the autoregressive bottleneck to eliminate spurious sequential dependencies and accurately capture global item-user correlations.

In response to the above challenges, and inspired by the recent success of generative models Ho et al. (2020); Nie et al. (2025), we propose a novel Generative Multi User framework, called **P**arallel Generative **D**iscrete **D**iffusion (**PD^2**). Unlike conventional discriminative matching or autoregressive sequence generation, PD^2 utilizes a discrete diffusion process to effectively model the intricate correlations between item and user groups. Specifically, PD^2 tackles the above issues as follows:

First, to address *distribution multi-modality*, we exploit the conditional generation capability of diffusion models. By performing an denoising process, PD^2 establishes a direct item-to-user mapping within a high-fidelity latent feature space. This generative paradigm empowers the model with strong generalization, enabling the retrieval of diverse user groups across varied interests.

Second, to accommodate the *non-sequential nature*, we propose a permutation-invariant parallel diffusion framework. Tailored to the unordered, set-like structure of I2U interactions, we employ a BERT-style stochastic masking strategy to reconstruct randomly masked tokens at arbitrary positions. This design breaks the serial bottleneck of autoregressive generation, establishing global bidirectional dependencies via parallel reconstruction. Furthermore, we design an Interval-Aware Spatio-Temporal Encoding mechanism by synergizing interval-aware temporal bias with Rotary Positional Embeddings Su et al. (2024). This combination captures the dynamic evolution of group interests while endowing the model with robust length extrapolation capabilities.

This parallel and group-aware framework confers two distinct advantages: (1) **Multi-group Lookahead Prediction.** Moving beyond single-target limitations, PD^2 directly characterizes the joint distribution of user sets within future windows. This captures group co-occurrence patterns and mitigates the inherent stochasticity of future interactions, closely aligning with the concurrent dissemination mechanism of modern platforms. (2) **Scalable Parallel Retrieval.** Leveraging the non-autoregressive mechanism, our approach minimizes inference latency while enabling diverse parallel decoding. This ensures a broad exploration of potential interest clusters for cold-start items.

[1]For a better understanding of the differences between this work and the existing ones, we present a multi-view conceptual comparison in Tab. 1, with additional comparisons with existing works are provided in Appendix A.2

Table 1: Conceptual comparison between user-centric and item-centric formulations. We distinguish broad retrieval directions (U2I and I2U) from specialized prediction formulations. Our work focuses on item-centric parallel multi-user generation, where multiple potential seed users are modeled and retrieved jointly within a future interaction window.

Perspective	Concept	Scope	Mapping Direction	Core Mechanisms	Primary Goal
User-Centric	**U2I** (User-to-Item)	General Paradigm (Broad)	User → Item	Rule-based filtering; Discriminative matching; Generative modeling	Personalized selection (Item passive retrieval)
	Next-Item (Modeling)	Generative paradigm (Specialized)	User history → Target item	Autoregressive Sequence generation	Sequential prediction (Probability estimation)
Item-Centric	**I2U** (Item-to-User)	General paradigm (Broad)	Item → User	Rule-based filtering; Discriminative matching, Look-alike modeling	Content distribution (Item active push)
	Multi-User (Modeling)	Generative paradigm (Specialized)	Item history → Multiple users	**Previously unexplored** **Ours PD^2 is the first**	**Diverse Users generation** (Active exploration)

In summary, the contributions of this paper are fourfold:

- We pioneer the paradigm shift from Next-Item to Multi-User Generation. Unlike the traditional pursuit of single deterministic targets, we reformulate cold-start recommendation as an item-centric multi-user-group generation task, thereby fundamentally alleviating the passive exposure suppression inherent in U2I models.
- We propose the first framework tailored for Multi-User generation. By introducing permutation-invariant Masked Discrete Diffusion, it bypasses the autoregressive bottleneck to enable scalable parallel retrieval, ensuring inference efficiency and retrieval diversity.
- We design the Interval-Aware Spatio-Temporal Encoding, which incorporates temporal bias with Rotary Positional Embeddings. This empowers Multi-group Look-ahead Prediction, which not only captures the dynamic evolution of group interests, but also aligns with the concurrent nature of I2U interactions, effectively modeling the joint distribution of potential user sets.
- Extensive experiments show that PD^2 significantly outperforms existing SOTA models. We further analyze the underlying reasons for performance gain and the impact of key factors like parallel decoding length to provide a clear understanding of the mechanisms supporting PD^2.

2 METHODOLOGY

2.1 PRELIMINARY

Let $\mathcal{U}$ and $\mathcal{V}$ respectively denote the sets of users and items within the dataset $\mathcal{D}$. Both users and items are associated with multi-field features (e.g., IDs, categories). We assume learnable encoding functions that map the sparse features of user $u \in \mathcal{U}$ and item $v \in \mathcal{V}$ into dense vector representations, denoted as $\mathbf{e}_u \in \mathbb{R}^d$ and $\mathbf{e}_v \in \mathbb{R}^d$, respectively, where d is the embedding dimension.

For each target item v, we define its positive interaction history as a chronologically ordered sequence $\mathcal{H}_v = (u_1, u_2, \ldots, u_L)$, where $u_k \in \mathcal{U}$ represents the user positively interacting with item v at position k, and L denotes the sequence length. To explicitly condition the generation process on the item's intrinsic characteristics, we construct the input token sequence $\mathcal{S}_v$ by prepending the item embedding $\mathbf{e}_v$ to the user embedding sequence: $\mathcal{S}_v = [\mathbf{e}_v, \mathbf{e}_{u_1}, \mathbf{e}_{u_2}, \ldots, \mathbf{e}_{u_L}]$, where $\mathcal{S}_v \in \mathbb{R}^{(L+1)\times d}$. For clarity of exposition, we hereinafter refer to the item embedding $\mathbf{e}_v$ as the **item token** and the user embeddings $\{\mathbf{e}_{u_k}\}_{k=1}^{L}$ as **user tokens**.

Distinct from traditional sequential recommendation that predicts a single deterministic next item, the Multi-User Modeling scenario is characterized by the concurrent dissemination of items to diverse user groups. Predicting a single next user u_{L+1} is thus insufficient to capture the multi-modal interest distribution. Consequently, we extend the prediction target from a single user to a future interaction window consisting of N users, denoted as $\mathcal{Y}_v = \{u_{L+1}, u_{L+2}, \ldots, u_{L+N}\}$. The objective is to model the joint probability of this future user set given the context:

$$\theta^* = \arg\max_{\theta} \sum_{v \in \mathcal{V}} \log P_\theta(\mathcal{Y}_v \mid \mathcal{S}_v), \tag{1}$$

where $\mathcal{Y}_v$ represents the diverse user groups likely to interact with the item v in the near future.

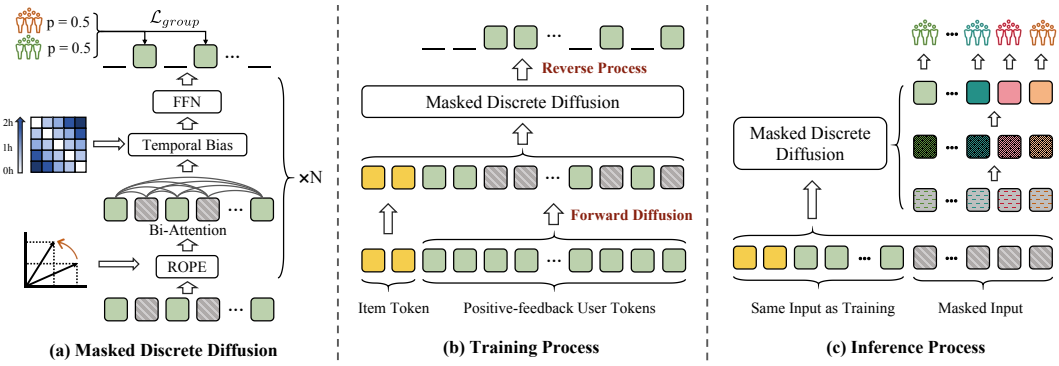

Figure 2: The architecture of our PD^2. (a) The Masked Discrete Diffusion module incorporates interval-aware spatio-temporal encoding to capture multi-modal user distribution. (b) The Training Process employs a parallel masked training strategy, where the model learns to reconstruct randomly masked user tokens within a sequence. (c) The Inference Process utilizes scalable parallel retrieval to simultaneously generate diverse candidate users in a single non-autoregressive forward pass.

2.2 Masked Discrete Diffusion

The core of PD^2 is a masked discrete diffusion model Austin et al. (2021) tailored for set-like user interactions. Unlike continuous diffusion models that add Gaussian noise, PD^2 operates in a discrete state space using a token masking corruption strategy. As shown in Fig. 2, this framework facilitates permutation-invariant modeling via bidirectional context learning. We detail the framework across three key aspects: discrete diffusion, parallel masked training, and scalable parallel retrieval.

2.2.1 Discrete Diffusion

We formulate the generation of the user sequence $\mathcal{S}_v$ as a reverse denoising process. Let $\mathcal{S}_0^v = \mathcal{S}_v$ denote the clean data. We define a *forward diffusion process* that randomly corrupts $\mathcal{S}_0^v$ into a fully masked sequence $\mathcal{S}_1^v$, and a *reverse process* that learns to reconstruct $\mathcal{S}_0^v$ from $\mathcal{S}_1^v$.

Forward Diffusion Process. To clarify the mechanism of our discrete diffusion framework, we first draw an analogy to standard continuous diffusion models Ho et al. (2020). In standard diffusion, the forward process gradually injects Gaussian noise into data $\mathbf{x}_0$ over timestep $t = \{1 \dots T\}$, eventually reaching pure noise $\mathbf{x}_T \sim \mathcal{N}(0, I)$. The generative capability stems from training the model to reverse this process, reconstructing the original data from varying noise levels (different timesteps).

In our masked discrete diffusion, we adapt this paradigm to discrete token sequences. Instead of Gaussian noise, we define the "noise" as a special absorbing state [MASK]. The diffusion process is modeled over a continuous diffusion time $t \in [0, 1]$, where:

$$\mathcal{S}_t^v = \begin{cases} \text{Original sequence } \mathcal{S}_0^v, & t = 0 \\ \text{Partially masked sequence}, & t \in (0, 1) \\ \text{Entirely [MASK]}, & t = 1 \end{cases} \tag{2}$$

As the diffusion time t increases, the likelihood of tokens being absorbed into the mask state increases, resulting in higher noise levels and information loss. Analogous to training standard diffusion models to restore data from varying noise levels, our objective is to enable PD^2 to reconstruct the clean sequence $\mathcal{S}_0^v$ from diverse intermediate corrupted states. Formally, for each training instance, we sample a random diffusion time from a uniform distribution as $t \sim \mathcal{U}(0, 1)$, which dictates the global corruption intensity (analogous to noise level in continuous diffusion). Given the clean state sequence $\mathcal{S}_0^v = (\mathbf{e}_v, e_{u_1}, \dots, e_{u_L})$, the forward process independently corrupts each user token $\{e_{u_k}\}_{k=1}^L$ based on t. Specifically, for each position k, we introduce an independent Bernoulli random variable parameterized by t. The transition probability $q_t(e_{u_k}^{(t)} \mid e_{u_k}^{(0)})$ is defined as:

$$q_t(e_{u_k}^{(t)} \mid e_{u_k}^{(0)}) = \begin{cases} 1 - t & \text{if } e_{u_k}^{(t)} = e_{u_k}^{(0)} \quad \text{(Preserved)}, \\ t & \text{if } e_{u_k}^{(t)} = \text{[MASK]} \quad \text{(Absorbed)}. \end{cases} \tag{3}$$

This implies that any user token has a probability equal to t of being absorbed into the mask state. In implementation, this is efficiently realized by sampling a random probability $p_{u_k} \sim \mathcal{U}(0,1)$ for each user token and masking those where $p_{u_k} < t$. Consequently, the marginal distribution of the entire corrupted sequence $\mathcal{S}_t^v$ at time t can be explicitly factorized as:

$$q_t(\mathcal{S}_t^v \mid \mathcal{S}_0^v) = \underbrace{\mathbb{1}[\mathcal{S}_{t,0}^v = \mathbf{e}_v]}_{\text{Fixed Item Token}} \cdot \prod_{k=1}^{L} q_t(e_{u_k}^{(t)} \mid e_{u_k}^{(0)}). \tag{4}$$

where $\mathcal{S}_{t,0}^v$ denotes the token at index 0 of the sequence $\mathcal{S}_t^v$, $\mathbb{1}[\cdot]$ is the indicator function, which equals 1 if the condition holds (i.e., the item token remains unmasked) and 0 otherwise.

By sampling different t for each sample and training the model to reconstruct $\mathcal{S}_0^v$ from these varying intermediate states $\mathcal{S}_t^v$, PD^2 learns to recover the true data distribution from noise, effectively mirroring the generative capability of standard diffusion models in a discrete sequence.

Reverse Process (Denoising). The reverse process aims to invert the corruption by estimating the posterior distribution $p_\theta(\mathcal{S}_0^v \mid \mathcal{S}_r^v)$. For efficient parallel prediction, we parameterize the reverse prediction over masked positions using a factorized output head:

$$p_\theta(\mathcal{S}_0^v \mid \mathcal{S}_t^v) = \prod_{k \in \mathcal{M}_t} p_\theta(e_{u_k} \mid \mathcal{S}_t^v), \tag{5}$$

where $\mathcal{M}_t = \{\, k \mid \mathcal{S}_{t,k}^v = \text{[MASK]} \,\}_{k=1}^{L}$ denotes the set of masked indices at timestep t. This factorization is a modeling approximation for parallel retrieval rather than a consequence of the independent forward corruption. For each masked position $k \in \mathcal{M}_t$, the model outputs a categorical distribution over the user vocabulary $\mathcal{U}$, conditioned on the entire partially corrupted sequence $\mathcal{S}_t^v$.

2.2.2 Parallel Masked Training

As illustrated in Fig. 2(b), we adopt a BERT-style training strategy that enables efficient parallel optimization over the entire diffusion time $t \in [0, 1]$.

Structure. We instantiate the denoiser $f_\theta(\cdot)$ using a Transformer equipped with bidirectional attention. Distinct from causal GPT-style models constrained by unidirectional masks, our architecture allows each token to attend to the global context. This bidirectional visibility maximizes the utilization of limited informative signals from sparse cold-start histories while intrinsically aligning with the unordered and concurrent nature of I2U interactions.

Given a partially corrupted sequence $\mathcal{S}_t^v$ at time t, the model computes contextualized hidden states $\mathbf{H} = f_\theta(\mathcal{S}_t^v) \in \mathbb{R}^{(L+1)\times d}$. For any masked position $k \in \mathcal{M}_t$, the corresponding hidden state $\mathbf{h}_k \in \mathbf{H}$ serves as the query representation to reconstruct the original user token e_{u_k}.

Training Objective. To efficiently optimize the Evidence Lower Bound (ELBO), we focus on minimizing the reconstruction error of masked user tokens. Given the prohibitive computational cost of standard Cross-Entropy loss over the massive user vocabulary size $|\mathcal{U}|$, we employ Noise Contrastive Estimation (NCE) Gutmann & Hyvärinen (2010) for scalable learning. The objective function is formulated as minimizing the expected NCE loss over the joint distribution of diffusion time t and data samples $\mathcal{S}_0^v$, which is defined as:

$$\mathcal{L} = \mathbb{E}_{t\sim\mathcal{U}(0,1),\mathcal{S}_0^v\sim\mathcal{D}} \left[\sum_{k \in \mathcal{M}_t} \mathcal{L}_{\text{NCE}}(e_{u_k}, \mathbf{h}_k) \right]. \tag{6}$$

where e_{u_k} denotes the ground-truth user at masked position k. The per-token NCE loss is defined as:

$$\mathcal{L}_{\text{NCE}}(u, \mathbf{h}) = -\log \sigma(s_\theta(u, \mathbf{h})) - \sum_{j=1}^{M} \log \sigma(-s_\theta(u_j^-, \mathbf{h})), \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function, $\{u_j^-\}_{j=1}^M$ are negative samples drawn from a noise distribution(e.g., popularity-based sampling), and $s_\theta(u, \mathbf{h}) = \mathbf{e}_u^\top \mathbf{h}$ is the scoring function by dot product. By maximizing the mutual information between the context and the missing users, this objective enables PD^2 to learn high-fidelity user representations without the bottleneck of full softmax normalization. Moreover, we provide the additional theoretical background and analysis in Appendix B.

2.2.3 SCALABLE PARALLEL RETRIEVAL DURING INFERENCE

During inference, leveraging the non-autoregressive nature, PD^2 employs a Scalable Parallel Retrieval strategy to efficiently retrieve a set of potential next users for a target item, as shown in Fig. 2(c).

Scalable Parallel Retrieval. Given an item $\mathbf{e}_v$ and its interaction history of length L, we construct the inference input by appending N mask tokens, where N is a hyperparameter for the retrieval horizon. The input sequence $\mathcal{S}^v_{\text{infer}}$ is defined as:

$$\mathcal{S}^v_{\text{infer}} = [\mathbf{e}_v, \mathbf{e}_{u_1}, \ldots, \mathbf{e}_{u_L}, [\text{MASK}]_1, \ldots, [\text{MASK}]_N]. \tag{8}$$

The model predicts the probability distributions for these N future positions simultaneously in a single forward pass:

$$\mathbf{H}_{\text{infer}} = f_\theta(\mathcal{S}^v_{\text{infer}})_{L+1:L+N} = [\mathbf{h}_{L+1}, \ldots, \mathbf{h}_{L+N}]. \tag{9}$$

Given a target retrieval size K, we treat the N parallel predicted tokens as independent retrieval channels, allocating a local retrieval quota $k' = \lceil K/N \rceil$ to each channel where $\lceil \cdot \rceil$ is the ceiling function. Specifically, for the j-th predicted token $\mathbf{h}_{L+j}$, we retrieve a candidate subset $\mathcal{C}_j$ with the top-k' users of the highest scores:

$$\mathcal{C}_j = \text{Top-k}\left(\{s_\theta(e_u, \mathbf{h}_{L+j}) \mid u \in \mathcal{U}, j \in \{1, ..., N\}\}, k = k'\right). \tag{10}$$

The final recommendation list $\hat{\mathcal{Y}}$ aggregates candidates from all channels and is defined as:

$$\hat{\mathcal{Y}} = \bigcup_{j=1}^{N} \mathcal{C}_j. \tag{11}$$

Advantage and Challenges. By processing all N prediction channels in parallel, inference latency is decoupled from the generation length, circumventing the sequential bottleneck of autoregressive models. This length-agnostic low-latency characteristic is pivotal for real-time industrial serving.

However, naively applying parallel decoding introduces two critical challenges: (1) Mode Collapse: Without guidance, independent predicted tokens tend to converge to the same high-probability user, resulting in redundancy and failing to capture the multi-modal interest distribution. (2) Extrapolation Failure: The disparity between sparse histories L and the extensive retrieval horizon N may expose the model to unseen positional indices, compromising inference accuracy. These challenges necessitate the Interval-Aware Spatio-Temporal Encoding mechanism, which we propose to ensure diversity and accuracy while supporting robust length extrapolation.

2.3 INTERVAL-AWARE SPATIO-TEMPORAL ENCODING

To resolve the challenges above faced by the parallel retrieval phase, we propose the Interval-Aware Spatio-Temporal Encoding mechanism, which comprises three synergistic components: an Interval-Aware Temporal Bias to capture interaction density, a Spatial Position Embedding to support length extrapolation, and a Multi-group Look-ahead to enforce distributional diversity via supervision.

2.3.1 INTERVAL-AWARE TEMPORAL BIAS

Item-centric interactions often exhibit irregular interval densities due to concurrent exposure, which the rigid positional indices fail to capture. Therefore, we explicitly model the temporal dynamics.

Let τ_i and τ_j denote the physical timestamps of user tokens at positions $i, j \in \{1, \ldots, L+N\}$, spanning both the historical (L) and inference (N) horizons. We quantize the continuous time distance into buckets of interval $\Delta\tau$ (e.g., 30 minutes), defined as:

$$\delta_{ij} = \left\lfloor \frac{|\tau_i - \tau_j|}{\Delta\tau} \right\rfloor, \tag{12}$$

where δ_{ij} denotes the discretized relative temporal distance, and $\lfloor \cdot \rfloor$ denotes the floor function. We then introduce a parameter-free temporal bias matrix $\mathbf{B} \in \mathbb{R}^{(L+N)\times(L+N)}$ to modulate the attention mechanism as:

$$\mathbf{B}_{ij} = -\lambda \cdot \delta_{ij}, \tag{13}$$

where λ is a hyperparameter controlling the decay slope. This formulation enforces high affinity within concurrent intervals, effectively capturing temporal evolution of user interests while enabling dynamic interval setting during inference.

2.3.2 SPATIAL POSITION EMBEDDING

While temporal bias captures density, weak structural cues are necessary to preserve the relative order of interactions. Moreover, given that the inference horizon N may exceed the training length L, the model requires robust extrapolation capabilities. To address this, we incorporate Rotary Positional Embeddings (RoPE) Su et al. (2024).

Formally, within the self-attention module of the Transformer denoiser $f_\theta(\cdot)$, let $\mathbf{q}_i, \mathbf{k}_j \in \mathbb{R}^d$ denote the query and key vectors for tokens at positions i and j. RoPE injects relative position information by rotating these vectors via a matrix $\mathcal{R}$: $\tilde{\mathbf{q}}_i = \mathcal{R}_i \mathbf{q}_i$, $\tilde{\mathbf{k}}_j = \mathcal{R}_j \mathbf{k}_j$. The final attention score $\mathcal{A}_{ij}$ then synthesizes both the structural rotation and our interval-aware temporal bias, defined as:

$$\mathcal{A}_{ij} = \frac{(\tilde{\mathbf{q}}_i)^\top \tilde{\mathbf{k}}_j}{\sqrt{d}} + \mathbf{B}_{ij}. \tag{14}$$

This hybrid design facilitates extrapolation to extended inference horizons, leveraging the translation-invariant property of RoPE to generalize naturally beyond observed absolute positions.

2.3.3 MULTI-GROUP LOOK-AHEAD PREDICTION

While encoding captures spatio-temporal dynamics, standard point-wise supervision restricts the model to current states, failing to reflect multi-modal user distributions. We propose a Multi-group Look-ahead Prediction strategy relaxing the supervision signal to a future interaction window.

Let $\mathcal{M}_t$ denote the set of masked indices at diffusion time t. We define the target set $\mathcal{W}_k$ as the subset of masked users appearing in a look-ahead window of size w relative to sequence position k:

$$\mathcal{W}_k = \{e_{u_j} \mid k \le j < \min(k + w, L) \wedge j \in \mathcal{M}_t\}. \tag{15}$$

We posit that any user within this temporal vicinity constitutes a valid realization of the item's target user. Consequently, instead of minimizing the NCE loss against a single fixed token in Eq. (6), we redefine the objective to minimize the expected NCE loss over the uniform distribution of $\mathcal{W}_k$, formulated as:

$$\mathcal{L}_{\text{Group}} = \mathbb{E}_{t, \mathcal{S}_0^v} \left[\sum_{k \in \mathcal{M}_t} \underbrace{\mathbb{E}_{u^+ \sim \mathcal{U}(\mathcal{W}_k)} \left[\mathcal{L}_{\text{NCE}}(e_{u^+}, \mathbf{h}_k) \right]}_{\text{Window-Averaged Loss}} \right]. \tag{16}$$

Here, $\mathcal{U}(\mathcal{W}_k)$ denotes uniform sampling from the future window set, and e_{u^+} is the embedding of the sampled target user. This strategy empowers long-range temporal extrapolation, enabling model to predict targets across extended horizons.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETTINGS

Datasets. Two publicly available datasets are utilized to evaluate the proposed method. The KuaiRand-Pure and KuaiRand-1K Gao et al. (2022) are derived from the recommendation logs of the video-sharing mobile app, Kuaishou. **Metrics.** Performance is measured by Recall@K (R@K) and NDCG@K (N@K) for $K \in \{5, 10\}$, reported in percentage (%). Please refer to Appendix. C for comprehensive descriptions of baselines, dataset, and hyperparameter settings.

3.2 OVERALL PERFORMANCE

We compare PD^2 with nine existing SOTA baselines, and the results are reported in Tab. 2. For a fair comparison, all methods are trained under the Next-User Modeling setting (I2U/Item-centric) and evaluated on predicting users who will interact with a given item. Overall, PD^2 consistently

Table 2: Performance comparison among different methods on KuaiRand-Pure and KuaiRand-1K.

Dataset	Method	Cold				Warm				All			
		R@5	R@10	N@5	N@10	R@5	R@10	N@5	N@10	R@5	R@10	N@5	N@10
KuaiRand-Pure	SASRec	1.23	1.54	0.86	0.96	1.38	1.98	0.89	1.08	1.37	1.96	0.89	1.07
	Bert4Rec	0.86	1.90	0.72	1.10	1.30	2.23	0.87	1.19	1.28	2.21	0.86	1.19
	DreamRec	0.69	1.33	0.53	0.75	0.73	1.35	0.57	0.79	0.73	1.35	0.57	0.79
	DiffuRec	1.12	1.49	0.77	0.89	1.25	2.11	0.76	1.05	1.24	2.08	0.76	1.04
	TIGER	0.53	1.60	0.30	0.61	1.77	2.38	1.23	1.60	1.71	2.34	1.18	1.55
	HSTU	1.45	1.92	1.18	1.49	1.83	2.85	1.22	1.69	1.81	2.80	1.22	1.68
	PAM	1.66	2.71	1.23	1.56	1.97	3.06	1.24	1.68	1.95	3.04	1.24	1.67
	NU	1.61	2.66	1.20	1.52	2.00	3.09	1.26	1.70	1.98	3.07	1.25	1.69
	ReaRec	1.06	1.60	0.97	1.14	1.55	2.45	1.19	1.48	1.52	2.41	1.17	1.46
	PD² (Ours)	**2.14**	**3.46**	**1.30**	**1.85**	**2.16**	**3.61**	**1.31**	**1.85**	**2.16**	**3.60**	**1.31**	**1.85**
KuaiRand-1K	SASRec	4.79	7.65	3.08	4.00	5.19	8.84	3.23	4.41	4.93	8.06	3.13	4.14
	Bert4Rec	5.11	8.23	3.15	4.14	5.24	8.98	3.26	4.46	5.15	8.48	3.19	4.25
	DreamRec	4.43	7.09	2.87	3.73	4.72	8.18	2.92	4.03	4.53	7.46	2.89	3.83
	DiffuRec	5.22	8.18	3.35	4.30	4.98	8.42	3.14	4.25	5.14	8.27	3.28	4.28
	TIGER	5.28	8.55	3.21	4.27	6.25	10.94	3.94	5.44	5.61	9.37	3.46	4.67
	HSTU	6.20	10.40	3.85	5.19	7.98	13.69	4.99	6.82	6.81	11.53	4.24	5.75
	PAM	6.22	10.17	3.83	5.19	7.32	12.95	4.66	6.72	6.60	11.12	4.11	5.71
	NU	6.46	10.52	4.03	5.35	7.09	12.29	4.40	6.06	6.68	11.13	4.16	5.59
	ReaRec	5.41	8.57	3.51	4.53	6.67	11.54	4.13	5.69	5.84	9.59	3.72	4.93
	PD² (Ours)	**9.01**	**14.28**	**5.94**	**7.63**	**10.12**	**16.96**	**6.33**	**8.52**	**9.39**	**15.20**	**6.07**	**7.93**

Here, the best and second best results are marked in **bold** and underlined, respectively. Each experiment is repeated 5 times and the average is reported.

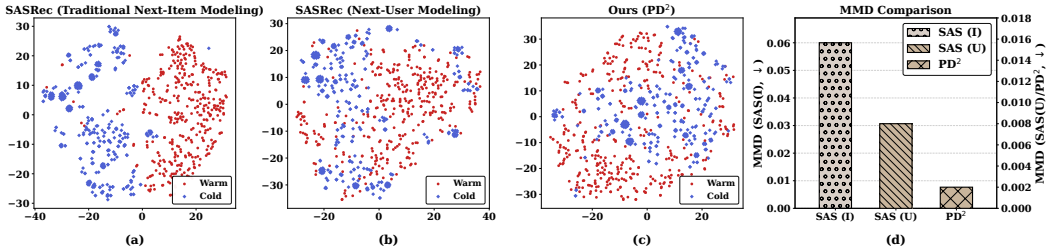

Figure 3: Differences in representation distributions of cold and warm items. (a)-(c) Visualization of embeddings of cold and warm items; (d) Distributional divergence between cold and warm items, measured by Maximum Mean Discrepancy (MMD). Here, SASRec (I) and SASRec (U) denote SASRec trained under Next-Item Modeling and Next-User Modeling, respectively.

outperforms all competing methods on both datasets across all evaluation metrics, demonstrating the effectiveness of the proposed framework for user retrieval. Specifically, compared with the second-best model, PD² improves Recall@10 by **17.3%** and **31.8%** on KuaiRand-Pure and KuaiRand-1K, respectively. The consistent gains across Recall and NDCG indicate that PD² not only retrieves more relevant users but also ranks them more accurately.

More importantly, PD² achieves substantially larger improvements in cold-start scenarios. For cold-start items, it improves Recall@10 by **27.7%** and **35.7%** on the two datasets over the second-best baseline. This highlights its advantage in identifying potential users when item-side interaction history is limited. Meanwhile, the consistent improvements on warm items show that the benefits of PD² are not restricted to sparse settings, but generalize across different item popularity levels. Due to space limit, please refer to Appendix. D for online A/B test.

3.3 Analysis on the Underlying Reasons of Performance Gain

To investigate why item-centric supervision benefits cold-start recommendation, Fig. 3 (a)–(c) visualize the representation distributions of cold-start and warm items under different training formulations using t-SNE Maaten & Hinton (2008). As shown in Fig. 3 (a), SASRec trained under the conventional U2I (Next-Item) formulation exhibits a clear representation gap between cold and warm items. After training the same SASRec backbone under the mirrored I2U formulation, this

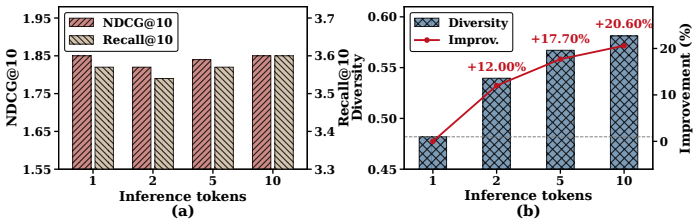

Figure 4: The comparison of the performance among different numbers of inference tokens.

Table 3: Ablation study of key modules on KuaiRand-Pure.

Variants	R@10	N@10
PD^2 (Full)	**3.60**	**1.85**
Components		
(a) w/o MGLP	3.13	1.59
(b) w/o IAST	3.38	1.73
(c) w/o Bi-Attention	3.42	1.76
Interval		
(d) $\Delta\tau = 15$ min	3.51	1.81
(e) $\Delta\tau = 1$ h	3.47	1.78

gap is substantially reduced, as shown in Fig. 3 (b). This observation suggests that using users as supervision targets can provide a more informative learning signal for cold-start item representations.

Fig. 3 (c) further shows that PD^2 extends I2U supervision to parallel multi-user generation, further reduces the discrepancy between cold and warm item representations. Quantitatively, Fig. 3 (d) reports the Maximum Mean Discrepancy (MMD) Borgwardt et al. (2006) between cold and warm item representations, where a larger MMD indicates greater distributional divergence. These results provide representation-level evidence that item-centric user supervision helps reduce the cold–warm representation gap, while the additional design of PD^2 further improves this alignment.

3.4 Ablation Study

Tab. 3 summarizes the ablation results on the KuaiRand-Pure dataset. Removing any component degrades performance, validating their respective contributions. Among them, the two core strategies are most influential: removing Multi-group Look-ahead Prediction (MGLP) reduces Recall@10 and NDCG@10 by 13.1% and 14.1%, respectively, while removing Interval-Aware Spatio-Temporal Encoding (IAST) causes drops of 6.1% and 6.5%. Replacing bidirectional attention with causal attention also decreases Recall@10 and NDCG@10 by 5.0% and 4.9%, confirming that user interactions are better viewed as group-distribution structures with strong interdependencies rather than strictly time-driven causal chains. Moreover, $\Delta\tau = 30$ minutes achieves the best performance, while shorter or longer intervals lead to slight degradation, suggesting that this granularity better captures temporal interaction patterns. More ablation study of hyperparameters can refer to Appendix. D.1.

3.5 Effect of Parallel Decoding

In Fig. 4, we show how parallelly decoding multiple tokens during inference impacts on both performance and diversity on the KuaiRand-Pure dataset. When the objective is to retrieve the top-K users, decoding N tokens in parallel means that each token is responsible for selecting the top-$\frac{K}{N}$ users. We observe that as the number of decoded tokens increases from 1 to 10, the effects on Recall and NDCG are negligible. However, parallel decoding significantly improves the diversity of the retrieved top-k users. Let $\{u_1, u_2, \ldots, u_K\}$ denote the retrieved top-K users, the diversity is measured as follows: $\text{Diversity} = 1 - \frac{2}{K(K-1)} \sum_{1 \le i < j \le K} \cos(\mathbf{e}_{u_i}, \mathbf{e}_{u_j})$ where $\cos(\mathbf{e}_{u_i}, \mathbf{e}_{u_j})$ denotes the cosine similarity between the embeddings of users i and j in the retrieved top-K list. A higher diversity indicates that the retrieved users are more widely distributed across different user groups. This is particularly beneficial for cold-start items, whose target audience is still unknown. By distributing the item to a more diverse set of users each time, the model can more quickly discover the appropriate audience group for the item. we present more experimental results In appendix. D.

4 Conclusion

This work pioneers a paradigm shift in cold-start recommendation from passive "Next-Item" retrieval to proactive "Multi-User generation" via our proposed framework, **PD^2**. By synergizing a permutation-invariant masked diffusion mechanism with interval-aware spatio-temporal encoding, PD^2 circumvents traditional autoregressive bottlenecks to effectively capture non-sequential, dynamic multi-group interests. Extensive offline evaluations and large-scale online A/B tests demonstrate its significant superiority over state-of-the-art baselines.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Gaode Chen, Ruina Sun, Yuezihan Jiang, Jiangxia Cao, Qi Zhang, Jingjian Lin, Han Li, Kun Gai, and Xinghua Zhang. A multi-modal modeling framework for cold-start short-video recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pp. 391–400, 2024.

Gaode Chen, Ruina Sun, Yinjie Jiang, Tianxiang Li, Yinlong Dai, Qifan Shi, Xiaojie Qin, Jialong Fu, Piaoyi Chen, Ronggeng Huang, et al. A cold-start recommendation system at kuaishou designed from the short-video perspective. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 124–132, 2025a.

Hao Chen, Zefan Wang, Feiran Huang, Xiao Huang, Yue Xu, Yishi Lin, Peng He, and Zhoujun Li. Generative adversarial framework for cold-start item recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2565–2571, 2022.

Xiaoshuang Chen, Yibo Wang, Yao Wang, Husheng Liu, and Kaiqiao et al. Zhan. Creator-side recommender system: Challenges, designs, and applications. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 162–170, 2025b.

Yijie Ding, Zitian Guo, Jiacheng Li, Letian Peng, Shuai Shao, Wei Shao, Xiaoqiang Luo, Luke Simon, Jingbo Shang, Julian McAuley, and Yupeng Hou. How well does generative recommendation generalize?, 2026. URL https://arxiv.org/abs/2603.19809.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and machines*, 30(4):681–694, 2020.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. Kuairand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 3953–3957, 2022.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Ruidong Han, Bin Yin, Shangyu Chen, He Jiang, Fei Jiang, Xiang Li, Chi Ma, Mincong Huang, Xiaoguang Li, Chunzhen Jing, et al. Mtgr: Industrial-scale generative recommendation framework in meituan. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5731–5738, 2025.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pp. 173–182, 2017.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 639–648, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Feiran Huang, Zefan Wang, Xiao Huang, Yufeng Qian, Zhetao Li, and Hao Chen. Aligning distillation for cold-start item recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 1147–1157, 2023.

Yanhua Huang, Yuqi Chen, Xiong Cao, Rui Yang, Mingliang Qi, Yinghao Zhu, Qingchang Han, Yaowei Liu, Zhaoyu Liu, Xuefeng Yao, et al. Towards large-scale generative ranking. *arXiv preprint arXiv:2505.04180*, 2025.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pp. 197–206. IEEE, 2018.

Jinri Kim, Eungi Kim, Kwangeun Yeo, Yujin Jeon, Chanwoo Kim, Sewon Lee, and Joonseok Lee. Content-based graph reconstruction for cold-start item recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1263–1273, 2024.

Anton Klenitskiy and Alexey Vasilev. Turning dross into gold loss: is bert4rec really better than sasrec? In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 1120–1125, 2023.

Yu-Ting Lan, Yang Huo, Yi Shen, Xiao Yang, and Zuotao Liu. Next-user retrieval: Enhancing cold-start recommendations via generative next-user modeling. *arXiv preprint arXiv:2506.15267*, 2025.

Wenhao Li, Jie Zhou, Chuan Luo, and Chao et al. Tang. Scene-wise adaptive network for dynamic cold-start scenes optimization in ctr prediction. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 2024.

Zihao Li, Aixin Sun, and Chenliang Li. Diffurec: A diffusion model for sequential recommendation. *ACM Transactions on Information Systems*, 2023.

Yunze Luo, Yuezihan Jiang, Yinjie Jiang, Gaode Chen, Jingchi Wang, Kaigui Bian, Peiyi Li, and Qi Zhang. Online item cold-start recommendation with popularity-aware meta-learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 927–937, 2025.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, and Jun et al. Hu. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=sMyXP8Tanm`.

Aleksandr Vladimirovich Petrov and Craig Macdonald. gsasrec: Reducing overconfidence in sequential recommendation trained with negative sampling. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 116–128, 2023.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315, 2023.

Yuxin Ren, Qiya Yang, Yichun Wu, Wei Xu, Yalong Wang, and Zhiqiang Zhang. Non-autoregressive generative models for reranking recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5625–5634, 2024.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024. ISBN 9798331314385.

Chenglei Shen, Teng Shi, Weijie Yu, Xiao Zhang, and Jun Xu. Genrecedit: Adapting model editing for generative recommendation with cold-start items, 2026. URL https://arxiv.org/abs/2603.14259.

Teng Shi, Chenglei Shen, Weijie Yu, Shen Nie, and Chongxuan et al. Li. Llada-rec: Discrete diffusion for parallel semantic id generation in generative recommendation. *arXiv preprint arXiv:2511.06254*, 2025.

Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, et al. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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.

Jiakai Tang, Sunhao Dai, Teng Shi, Jun Xu, Xu Chen, Wen Chen, Jian Wu, and Yuning Jiang. Think before recommend: Unleashing the latent reasoning power for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

Aaron et al. Van den Oord. Deep content-based music recommendation. *Advances in neural information processing systems*, 2013.

Liwei Wu, Shuqing Li, Cho-Jui Hsieh, and James Sharpnack. Sse-pt: Sequential recommendation via personalized transformer. In *Proceedings of the 14th ACM conference on recommender systems*, pp. 328–337, 2020a.

Tao Wu, Ellie Ka-In Chio, Heng-Tze Cheng, Yu Du, Steffen Rendle, Dima Kuzmin, Ritesh Agarwal, Li Zhang, John Anderson, Sarvjeet Singh, et al. Zero-shot heterogeneous transfer learning from recommender systems to cold-start search retrieval. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 2821–2828, 2020b.

Le Yan, Zhen Qin, Honglei Zhuang, Xuanhui Wang, Michael Bendersky, and Marc Najork. Revisiting two-tower models for unbiased learning to rank. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 2410–2414, 2022.

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

Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 974–983, 2018.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, et al. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *arXiv preprint arXiv:2402.17152*, 2024.

The Appendix is organized as follows:

- Section A: Related Work.
- Section B: Proofs of Theoretical Analysis.
- Section C: Implementation Details.
- Section D: More Experiments Results.
- Section E: Impact of the Recommendation Paradigm.

A Related Work

A.1 Cold-Start Recommendation

The item cold-start problem arises from the scarcity of interaction feedback for newly released items Chen et al. (2025a); Huang et al. (2023); Li et al. (2024). Prevalent solutions address this issue mainly by enriching item representations with auxiliary modalities (e.g., text, images) Chen et al. (2024); Van den Oord (2013); Kim et al. (2024), or by adopting meta-learning to transfer priors from warm to cold domains Luo et al. (2025); Chen et al. (2022); Wu et al. (2020b). Despite these advances, such methods either depend heavily on the quality of side information or remain vulnerable to distribution shifts between warm and cold items. Fundamentally, these approaches are confined to a user-centric U2I representation learning paradigm based on sparse signals. This reliance often yields sub-optimal item representations, overlooking the potential to leverage the far richer and more stable user-side behavioral patterns to bridge the cold-start gap, which is explored in this work.

A.2 Item-to-User (I2U) Recommendation

I2U recommendation is an emerging direction that aims to identify potential users for specific items Chen et al. (2025a;b). Limited existing works typically treat I2U as a mere structural mirror of User-to-Item (U2I) recommendation via heuristic adaptations or two-tower filtering Yan et al. (2022). By simply reversing the input-output roles to encode items for retrieving users, these approaches ignore the unique distributional characteristics inherent to the item-centric view. Recently, as a transposed mirror of SASRec Kang & McAuley (2018), NU Lan et al. (2025) introduces next-user retrieval, but it still essentially follows a discriminative matching paradigm and aim to predict next user rather than the generative Multi-User paradigm advocated in this work. Besides, its rigid sequential dependency modeling and deterministic single-target prediction overlook both the multi-modal user distribution and the non-sequential nature of item-centric user groups. In contrast, we instantiate Multi-User modeling with a diffusion-based generative framework, which constructs a continuous item-conditioned latent user space and enables progressive user generation. Moreover, by discarding the sequential assumption, our framework supports bidirectional context modeling and parallel inference, thereby facilitating the generation of diverse user groups.

A.3 Generative Recommendation

Generative approaches recast recommendation as sequence generation, typically adopting autoregressive (AR) frameworks Wu et al. (2020a); Han et al. (2025); Huang et al. (2025). Despite their merits, AR models are inherently constrained by unidirectional causal masking, which restricts context modeling and hampers inference efficiency. Consequently, research has pivoted towards non-autoregressive paradigms Ren et al. (2024), including continuous Li et al. (2023); Yang et al. (2023) or discrete Nie et al. (2025); Song et al. (2025) diffusion models. LLaDA-Rec Shi et al. (2025) is a recent work that leverages masked discrete diffusion to generate semantic ID (SID) under the conventional U2I scenario. Crucially, our work differs from these user-centric approaches by pioneering multi-user modeling via a generative framework. By leveraging masked diffusion from an item-centric perspective, we bypass AR constraints to enable bidirectional context modeling and parallel multi-token inference, a distinct advantage for cold-start diversity.

B PROOFS OF THEORETICAL RESULTS

We provide theoretical analysis for the PD^2 framework. We first show that the masked diffusion objective $\mathcal{L}_{\mathrm{MD}}$ is a valid negative evidence lower bound (NELBO) that recovers $P^\star$ at optimality. Then we establish that absorbing-state corruption avoids the quantization gap of Gaussian noise.

B.1 CONVERGENCE OF MASKED DISCRETE DIFFUSION

Assumption 1 (Conditional Exchangeability)**.** *For every permutation $\pi \in \mathfrak{S}_L$,*

$$P^\star(u_{\pi(1)}, \ldots, u_{\pi(L)} \mid v) = P^\star(u_1, \ldots, u_L \mid v). \tag{17}$$

Following Sahoo et al. (2024); Ou et al. (2025), the true reverse posterior from t to $s < t$ at masked position k is:

$$q(e_{u_k}^{(s)} \mid e_{u_k}^{(t)} = [\mathrm{M}],\, e_{u_k}^{(0)}) = \tfrac{s}{t}\,\mathbf{1}[e_{u_k}^{(s)} = [\mathrm{M}]] + \tfrac{t-s}{t}\,\mathbf{1}[e_{u_k}^{(s)} = e_{u_k}^{(0)}]. \tag{18}$$

Replacing $e_{u_k}^{(0)}$ with $p_\theta(e_{u_k}^{(0)} \mid \mathcal{S}_t^v)$ (zero masking probability), the per-step KL reduces to:

Lemma 1 (Rao-Blackwellized Loss)**.**

$$D_{\mathrm{KL}}\big[q(e_{u_k}^{(s)} \mid e_{u_k}^{(t)}, e_{u_k}^{(0)}) \,\|\, p_\theta(e_{u_k}^{(s)} \mid \mathcal{S}_t^v)\big] = -\tfrac{t-s}{t} \log p_\theta(e_{u_k}^{(0)} \mid \mathcal{S}_t^v). \tag{19}$$

Proof. We consider two cases based on whether position k is masked at time t.

Case 1: $e_{u_k}^{(t)} \neq [\mathrm{MASK}]$. By the carry-over property of the absorbing process, an unmasked token remains unchanged during the reverse step: both the true posterior $q(e_{u_k}^{(s)} \mid e_{u_k}^{(t)}, e_{u_k}^{(0)})$ and the model $p_\theta(e_{u_k}^{(s)} \mid \mathcal{S}_t^v)$ place all mass on $e_{u_k}^{(t)}$, so $D_{\mathrm{KL}} = 0$.

Case 2: $e_{u_k}^{(t)} = [\mathrm{MASK}]$. The true reverse posterior (Eq. (1) of the main paper) is a two-point distribution on $\{e_{u_k}^{(0)},\ [\mathrm{MASK}]\}$:

$$q(e_{u_k}^{(s)} \mid e_{u_k}^{(t)} = [\mathrm{M}],\, e_{u_k}^{(0)}) \;=\; \frac{t-s}{t}\,\mathbf{1}[e_{u_k}^{(s)} = e_{u_k}^{(0)}] \;+\; \frac{s}{t}\,\mathbf{1}[e_{u_k}^{(s)} = [\mathrm{M}]]. \tag{20}$$

The model posterior, obtained by replacing $e_{u_k}^{(0)}$ with $p_\theta(e_{u_k}^{(0)} \mid \mathcal{S}_t^v)$ under the zero masking probability constraint ($p_\theta([\mathrm{MASK}] \mid \mathcal{S}_t^v) = 0$), assigns:

$$p_\theta(e_{u_k}^{(s)} \mid \mathcal{S}_t^v) \;=\; \frac{t-s}{t}\,p_\theta(e_{u_k}^{(0)} \mid \mathcal{S}_t^v) \;+\; \frac{s}{t}\,\mathbf{1}[e_{u_k}^{(s)} = [\mathrm{M}]]. \tag{21}$$

The KL divergence between equation 20 and equation 21 has exactly two terms:

$$\begin{aligned} D_{\mathrm{KL}} &= \underbrace{\frac{t-s}{t} \log \frac{(t-s)/t}{(t-s)/t \cdot p_\theta(e_{u_k}^{(0)} \mid \mathcal{S}_t^v)}}_{\text{non-mask atom}} \;+\; \underbrace{\frac{s}{t} \log \frac{s/t}{s/t}}_{\text{mask atom} = 0} \\ &= \frac{t-s}{t}\Big(-\log p_\theta\big(e_{u_k}^{(0)} \mid \mathcal{S}_t^v\big)\Big) \;=\; -\frac{t-s}{t} \log p_\theta\big(e_{u_k}^{(0)} \mid \mathcal{S}_t^v\big). \end{aligned} \tag{22}$$

The mask atom contributes zero because the true and model posteriors both assign probability s/t to [MASK]. □

Theorem 1 (Continuous-Time NELBO)**.** *Define the continuous-time masked prediction objective*

$$\mathcal{L}_{\mathrm{MD}}(\theta) = -\mathbb{E}_{t\sim\mathcal{U}(0,1)}\,\mathbb{E}_{\mathcal{S}_0^v,\mathcal{S}_t^v}\left[\frac{1}{t}\sum_{k\in\mathcal{M}_t} \log p_\theta(e_{u_k} \mid \mathcal{S}_t^v)\right]. \tag{23}$$

Then we have: (i) *$\mathcal{L}_{\mathrm{MD}}$ is an upper bound on the negative log-likelihood of themodel distribution Austin et al. (2021); Sahoo et al. (2024).* (ii) *Under sufficient model capacity and infinite data, the global minimizer $P^\star$: $\theta^\star \in \arg\min_\theta \mathcal{L}_{\mathrm{MD}}(\theta) \Rightarrow p_{\theta^\star}(\mathcal{S}_0^v) = P^\star(\mathcal{S}_0^v)$.*

Proof. We prove the two statements separately.

Proof of (i): Upper bound on NLL.

We discretize $[0, 1]$ into T uniform steps with $t(i) = i/T$ and $s(i) = (i-1)/T$. The standard discrete-time NELBO for diffusion models decomposes as Austin et al. (2021):

$$\mathcal{L}_T = \underbrace{\mathcal{L}_{\text{recons}}}_{\text{reconstruction}} + \underbrace{\mathcal{L}_{\text{diffusion}}}_{\text{denoising steps}} + \underbrace{\mathcal{L}_{\text{prior}}}_{\text{prior matching}}. \tag{24}$$

For the absorbing-state process, $\mathcal{L}_{\text{prior}} = 0$ (at $t = 1$ every token is masked with certainty, so the prior is matched exactly).

Applying Lemma 1 to each step, the diffusion loss is:

$$\mathcal{L}_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_q \left[\sum_{k \in \mathcal{M}_{t(i)}} \frac{t(i) - s(i)}{t(i)} \big(- \log p_\theta(e_{u_k} \mid \mathcal{S}^v_{t(i)}) \big) \right]. \tag{25}$$

With uniform discretization, $t(i) - s(i) = 1/T$, so each weight becomes $\frac{1}{T \cdot t(i)}$. Recognizing this as a Riemann sum with mesh $\Delta t = 1/T$:

$$\mathcal{L}_{\text{diffusion}} = \sum_{i=1}^{T} \frac{1}{T} \, \mathbb{E}_q \left[\frac{1}{t(i)} \sum_{k \in \mathcal{M}_{t(i)}} \big(- \log p_\theta(e_{u_k} \mid \mathcal{S}^v_{t(i)}) \big) \right]. \tag{26}$$

Taking $T \to \infty$, the Riemann sum converges to the integral:

$$\mathcal{L}_{\text{diffusion}} \xrightarrow{T \to \infty} \int_0^1 \mathbb{E}_q \left[\frac{1}{t} \sum_{k \in \mathcal{M}_t} \big(- \log p_\theta(e_{u_k} \mid \mathcal{S}^v_t) \big) \right] \mathrm{d}t \; = \; \mathcal{L}_{\text{MD}}(\theta). \tag{27}$$

Convergence at $t \to 0^+$. The integrand contains $1/t$, which diverges as $t \to 0^+$. However, the number of masked positions $|\mathcal{M}_t|$ is $\text{Binomial}(L, t)$, so $\mathbb{E}[|\mathcal{M}_t|] = tL$. Therefore, $\mathbb{E}[|\mathcal{M}_t|/t] = L$, and the integrand remains bounded as $t \to 0^+$. The reconstruction loss $\mathcal{L}_{\text{recons}}$ vanishes in the continuous-time limit and is absorbed into the integral at the boundary $t \to 0^+$.

Since the discrete-time NELBO $\mathcal{L}_T$ is an upper bound on $-\mathbb{E}[\log p_\theta(\mathcal{S}^v_0)]$ for every finite T (by the standard variational argument Austin et al. (2021); Sahoo et al. (2024)), and $\mathcal{L}_T \to \mathcal{L}_{\text{MD}}(\theta)$ as $T \to \infty$ Sahoo et al. (2024), the continuous-time limit is the tightest such bound:

$$-\mathbb{E}_{\mathcal{S}^v_0 \sim P^\star} \big[\log p_\theta(\mathcal{S}^v_0) \big] \; \le \; \mathcal{L}_{\text{MD}}(\theta). \tag{28}$$

Proof of (ii): Consistency.

The objective $\mathcal{L}_{\text{MD}}(\theta)$ is an integral of non-negative KL terms (Lemma 1). At the global optimum $\theta^\star$, every KL term must vanish:

$$D_{\text{KL}} \big[q(e^{(s)}_{u_k} \mid e^{(t)}_{u_k}, e^{(0)}_{u_k}) \parallel p_{\theta^\star}(e^{(s)}_{u_k} \mid \mathcal{S}^v_t) \big] = 0, \quad \forall \, k \in \mathcal{M}_t, \; \forall \, 0 \le s < t \le 1. \tag{29}$$

Since KL divergence equals zero if and only if the two distributions agree almost surely, this requires the pointwise match of single-token posteriors:

$$p_{\theta^\star}(e^{(0)}_{u_k} \mid \mathcal{S}^v_t) = P^\star(e^{(0)}_{u_k} \mid \mathcal{S}^v_t), \quad \forall \, k \in \mathcal{M}_t, \; \forall \, t \in (0, 1]. \tag{30}$$

For any fixed $t \in (0, 1]$, the law of total probability over the discrete state space $\big(\mathcal{U} \cup \{[\text{MASK}]\}\big)^{L+1}$ yields the marginalization identity

$$p_{\theta^\star}(\mathcal{S}^v_0) = \sum_{\mathcal{S}^v_t} p_{\theta^\star}(\mathcal{S}^v_0 \mid \mathcal{S}^v_t) \, q_t(\mathcal{S}^v_t), \tag{31}$$

where we have used the fact that, at the optimum, the model's marginal at time t coincides with the forward marginal $q_t(\mathcal{S}^v_t)$ (a consequence of every per-step reverse transition matching its true counterpart, propagated by induction along the reverse Markov chain).

The remaining task is to verify that the model's joint conditional $p_{\theta^\star}(\mathcal{S}_0^v \mid \mathcal{S}_t^v)$ equals the true joint $P^\star(\mathcal{S}_0^v \mid \mathcal{S}_t^v)$. The factorized form of the reverse process (Eq. (5) of the main paper) gives

$$p_{\theta^\star}(\mathcal{S}_0^v \mid \mathcal{S}_t^v) = \prod_{k \in \mathcal{M}_t} p_{\theta^\star}(e_{u_k}^{(0)} \mid \mathcal{S}_t^v) \cdot \prod_{k \notin \mathcal{M}_t} \mathbf{1}\big[e_{u_k}^{(0)} = e_{u_k}^{(t)}\big]. \tag{32}$$

While equation 32 need not equal $P^\star(\mathcal{S}_0^v \mid \mathcal{S}_t^v)$ pointwise for finite-step reverse processes, the equality holds in the continuous-time limit. Specifically, for an infinitesimal reverse step $\Delta t = t - s$ with $\Delta t \to 0^+$, the per-position transition kernel $q(e_{u_k}^{(s)} \mid e_{u_k}^{(t)} = [\mathrm{M}], e_{u_k}^{(0)})$ unmasks position k with probability $\Delta t/t$; the probability that two or more positions are simultaneously unmasked is $O(\Delta t^2)$ and vanishes in the limit. Hence at most one position is unmasked per infinitesimal step, and equation 32 collapses to a single factor that, by equation 30, matches the corresponding true single-position conditional. Iterating this matched infinitesimal transition along the reverse chain from $t = 1$ to $t = 0$, and invoking the conditional exchangeability assumption (Assumption 1) to remove dependence on the unmasking order, we obtain

$$p_{\theta^\star}(\mathcal{S}_0^v \mid \mathcal{S}_t^v) = P^\star(\mathcal{S}_0^v \mid \mathcal{S}_t^v). \tag{33}$$

Substituting back into equation 31 and applying the total-probability identity to $P^\star$ gives

$$p_{\theta^\star}(\mathcal{S}_0^v) = \sum_{\mathcal{S}_t^v} P^\star(\mathcal{S}_0^v \mid \mathcal{S}_t^v)\, q_t(\mathcal{S}_t^v) = P^\star(\mathcal{S}_0^v), \tag{34}$$

which establishes the claim. □

B.2 Absorbing State vs. Gaussian Noise

Proposition 1 (State-Space Closure of Absorbing Diffusion)**.** *Under the absorbing forward process defined in Eq. (3), the support of q_t satisfies*

$$\mathrm{supp}(q_t) \subseteq \big(\mathcal{U} \cup \{[\mathrm{MASK}]\}\big)^{L+1} \quad \textit{for every } t \in [0, 1]. \tag{35}$$

Consequently, for any masked position $k \in \mathcal{M}_t$, the reverse posterior $p_\theta(e_{u_k}^{(0)} \mid \mathcal{S}_t^v)$ reduces to a categorical distribution over the finite user vocabulary $\mathcal{U}$, i.e., the denoising task is a well-defined classification problem on a finite state space Austin et al. (2021).

Proof. By definition, the absorbing forward kernel

$$q_t(e_{u_k}^{(t)} \mid e_{u_k}^{(0)}) = (1 - t)\, \mathbf{1}[e_{u_k}^{(t)} = e_{u_k}^{(0)}] + t\, \mathbf{1}[e_{u_k}^{(t)} = [\mathrm{MASK}]] \tag{36}$$

maps each token either to itself (with probability $1 - t$) or to $[\mathrm{MASK}]$ (with probability t). For any $t \in [0, 1]$, every coordinate of $\mathcal{S}_t^v$ therefore takes values in $\mathcal{U} \cup \{[\mathrm{MASK}]\}$, which yields the support inclusion. Since the codomain is a finite set, the reverse model $p_\theta(e_{u_k}^{(0)} \mid \mathcal{S}_t^v)$ is a categorical distribution with $|\mathcal{U}|$ atoms. □

Remark 1 (Design Contrast with Continuous (Gaussian) Diffusion). Continuous diffusion models defined directly on the embedding space $\mathbb{R}^d$ (e.g., DreamRec Yang et al. (2023), DiffuRec Li et al. (2023)) employ a Gaussian forward process

$$q_t(\mathbf{x}_t \mid u) = \mathcal{N}\Big(\mathbf{x}_t;\ \sqrt{1 - \beta_t}\, \mathbf{e}_u,\ \beta_t I_d\Big), \tag{37}$$

whose marginal is a Gaussian mixture

$$q_t(\mathbf{x}_t) = \sum_{u \in \mathcal{U}} P^\star(u)\, \mathcal{N}\Big(\mathbf{x}_t;\ \sqrt{1 - \beta_t}\, \mathbf{e}_u,\ \beta_t I_d\Big), \tag{38}$$

supported on all of $\mathbb{R}^d$ for any $t > 0$. Recovering a categorical user therefore requires an auxiliary mapping $\mathbb{R}^d \to \Delta^{|\mathcal{U}|-1}$, whose fidelity depends on (i) the geometric separation of $\{\mathbf{e}_u\}_{u \in \mathcal{U}}$ in $\mathbb{R}^d$, (ii) the noise schedule β_t, and (iii) the calibration of the decoder head. The absorbing formulation in Proposition 1 bypasses this mapping entirely: the reverse problem is, by construction, a finite classification over $\mathcal{U}$, requiring no additional decoding interface beyond the user embedding table. Continuous diffusion requires an extra decoder mapping $\mathbb{R}^d \to \Delta^{|\mathcal{U}|-1}$. When $|\mathcal{U}| \gg d$ or embeddings are learned, the posterior $P(u \mid \mathbf{x}_t)$ is not concentrated on a single user across large regions of $\mathbb{R}^d$, forcing the decoder to resolve genuine ambiguity. The absorbing formulation skips this step: the reverse output is indexed directly by the user embedding table.

C Implementation Details

C.1 Source code

Our code is available at `https://anonymous.4open.science/r/PD2` and will be openly released.

C.2 Experimental Settings

C.2.1 Datasets

Two publicly available datasets are utilized to evaluate the proposed method. The KuaiRand-Pure and KuaiRand-1K Gao et al. (2022) are derived from the recommendation logs of the video-sharing mobile app, Kuaishou. For KuaiRand-Pure, an item uniformly sampled from the candidate pool is randomly inserted into the recommendation list with a fixed probability by replacing one existing item. Users who are exposed to fewer than 10 such random insertions are removed. In total, 27,285 users are retained. All 7,583 items are inserted at least once, resulting in 1,186,059 interaction records, with each item receiving an average of 156 interactions. For KuaiRand-1K, 1,000 users are randomly sampled from the recommendation logs, and videos irrelevant to their recommendation histories are removed, leaving approximately four million videos. It contains 1,000 users, 4,369,953 items, and 11,713,045 interaction records. On average, each item has 2.68 interactions.

C.2.2 Compared Methods

We compare PD^2 against nine SOTA methods, which are detailed as follows:

- **SASRec Kang & McAuley (2018); Klenitskiy & Vasilev (2023):** A representative sequential recommendation model that adopts a transformer architecture to encode sequences. In our implementation, we further enhance the original model with a sampled softmax loss Petrov & Macdonald (2023); Klenitskiy & Vasilev (2023).
- **Bert4Rec Sun et al. (2019):** This method employs bidirectional self-attention and utilizes a cloze objective to encode sequential data.
- **DreamRec Yang et al. (2023):** It utilizes the Transformer to learn users' preferences, which then serve as the diffusion guidance for generating the oracle item for each user.
- **DiffuRec Li et al. (2023):** It employs a diffusion model to learn sequence embeddings in a distributional space and then utilizes an approximator to generate the target representations.
- **TIGER Rajput et al. (2023):** This method first learns a residual quantized autoencoder to generate semantically informative indexing identifiers for items, and then trains an encoder–decoder model on these identifier sequences to predict the next item based on users' historical interactions. In our implementation, we generate the semantic IDs from the embeddings produced by SASRec.
- **HSTU Zhai et al. (2024):** This method introduces point-wise aggregated attention and the M-FALCON algorithm, scaling the model dimensions to GPT-3 Floridi & Chiriatti (2020) levels.
- **PAM Luo et al. (2025)**: A popularity-aware meta-learning method for online item cold-start recommendation. It partitions training instances into meta-learning tasks according to item popularity and adaptively balances behavioral and content features across different popularity levels.
- **NU Lan et al. (2025):** It adopts a SASRec-like framework to model unidirectional temporal correlations within user sequences, aiming to identify the single most probable user for the next interaction.
- **ReaRec Tang et al. (2026):** This method employs an inference-time computing framework for the recommender systems, which enhances user representations through implicit multi-step reasoning.

C.2.3 Metrics & Parameter Settings

For the evaluation metrics, we select two metrics that are widely used in recommender systems: Recall@K (denoted as R@K), and Normalized Discounted Cumulative Gain Järvelin & Kekäläinen

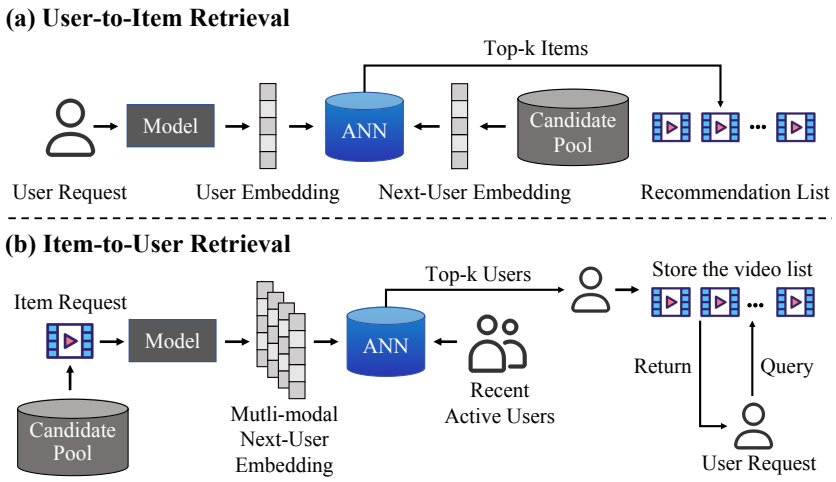

Figure 5: PD^2 online serving pipelines for U2I and I2U paradigms in large-scale industrial recommender systems.

(2002) (NDCG@K), denoted as N@K in this paper. We report the results with K $= \{5, 10\}$, and all metrics are presented in percentage $(\%)$. We define cold-start items as those whose historical interaction lengths fall within the bottom 10% (including items that have no interaction records.).

We adopt the same model dimensions among all methods, and obtain the remaining hyperparameters using the Optuna[2] library to accelerate the tuning process. For methods originally designed for user-to-item (U2I) recommendation, we apply a mirrored item-to-user (I2U) formulation to ensure a fair comparison.

C.3 Online Serving

Fig. 5 illustrates the deployment pipelines of PD^2 in a large-scale industrial recommendation platform. Despite being trained under the Next-User framework (I2U/Item-Centric), PD^2 features a **dual-mode** inference capability, supporting both standard user-to-item retrieval and active item-to-user retrieval.

Standard User-to-Item (U2I) Retrieval. As shown in Fig. 5(a), this pipeline seamlessly integrates with existing retrieval infrastructures. Upon a user request, the system generates a user embedding to query the item Approximate Nearest Neighbor (ANN) index. Crucially, although the retrieval flow is conventional, the item embeddings indexed are derived from our Next-User training. This ensures that cold-start items possess robust representations aligned with the user space, significantly improving retrieval quality without altering the established serving stack.

Active Item-to-User (I2U) Retrieval. This pipeline (Fig. 5(b)) leverages the generative nature of PD^2 for active content distribution. The system performs asynchronous batch retrieval: for each candidate item, PD^2 generates next user embeddings to retrieve top-k target users from an ANN index constructed over currently active users (e.g., active within 30 minutes). The resulting item-user pairs are asynchronously pushed to a Graph Storage cache keyed by user ID. During serving, user requests directly hit this cache to fetch pre-computed recommendations with minimal latency. We provide a comparative analysis of these strategies in Appendix. D.5.

[2] https://optuna.org/

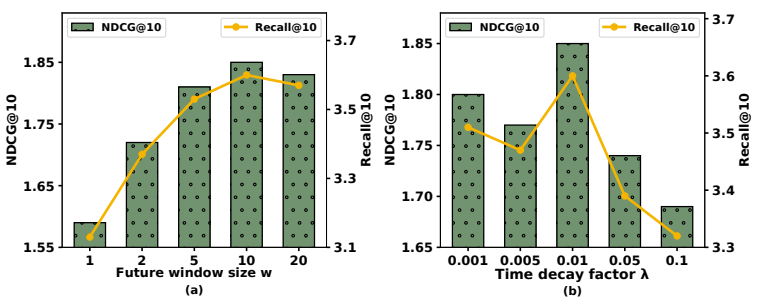

Figure 6: Parameter sensitivity analysis among different future window size w and time decay factor λ.

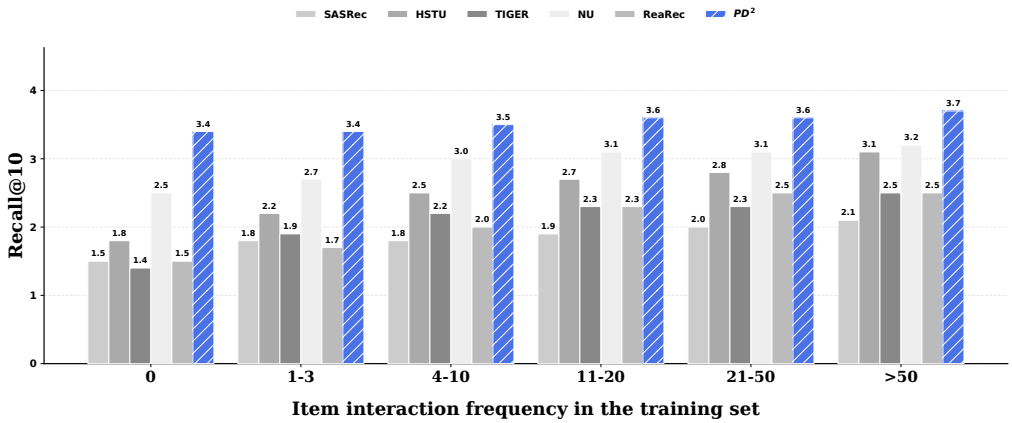

Figure 7: Performance across different item frequency bins.

D More Experiments Results

D.1 Sensitivity Analysis On Hyperparameters

We also analyze the sensitivity of PD^2 to the future window size w and time decay factor λ on KuaiRand-Pure dataset, the results are shown in Fig. 6. We can see that as w increases within a reasonable range, performance continues to improve. However, when w exceeds 20, performance begins to decline. For λ, the optimal value is 0.01, when λ gets larger than this value, the temporal decay becomes excessively strong, causing the current user to mainly attend to only temporally nearby users, which in turn leads to performance degradation.

D.2 Cross-Domain Generalization

To evaluate whether the effectiveness of PD^2 generalizes beyond the short-video domain, we further conduct experiments on two categories from the Amazon-2023 dataset, *Instruments* and *Games*. Following the main experiments, we report Recall@10 (R@10) on cold, warm, and all items, with all results presented in percentage (%). As shown in Tab. 4, PD^2 consistently achieves the best performance across both datasets and all item groups. On *Instruments*, PD^2 achieves a Cold R@10 of 6.17, improving over the strongest baseline NU (5.32) by 15.98%. The relative improvements on Warm and All items are 8.72% and 9.54%, respectively. On *Games*, PD^2 achieves a Cold R@10 of 8.72, corresponding to a 21.11% improvement over NU (7.20), while improving Warm and All R@10 by 10.07% and 10.70%, respectively. These results suggest that the effectiveness of PD^2 can also be observed other scenarios with different item characteristics and interaction patterns.

Table 4: Cross-domain performance comparison on two categories of Amazon-2023. All results are Recall@10 (R@10) reported in percentage (%). Each experiment is repeated five times, and the average result is reported. The best and second-best results are marked in **bold** and underline, respectively.

Dataset	Method	Cold R@10	Warm R@10	All R@10
Instruments	SASRec	3.66	5.32	5.09
	BERT4Rec	3.92	5.45	5.24
	HSTU	4.30	6.02	5.78
	PAM	5.29	6.01	5.91
	NU	<u>5.32</u>	<u>6.19</u>	<u>6.07</u>
	PD^2 (Ours)	**6.17**	**6.73**	**6.65**
Games	SASRec	4.85	7.81	7.56
	BERT4Rec	5.26	8.15	7.90
	HSTU	6.14	8.68	8.46
	PAM	6.68	8.45	8.30
	NU	<u>7.20</u>	<u>9.04</u>	<u>8.88</u>
	PD^2 (Ours)	**8.72**	**9.95**	**9.83**

D.3 Disentangling Training Formulation and Model Architecture

To separate the effect of the training formulation from that of the model architecture, we further evaluate representative models under both user-to-item (U2I) and item-to-user (I2U) training formulations on KuaiRand-1K. All models are evaluated with the same standard U2I retrieval pipeline. Specifically, a user request is first encoded into a user representation, which is then used to retrieve item representations learned under the corresponding training formulation. Different from the main I2U evaluation in Tab. 2, this experiment measures the ability to retrieve relevant items for users under a common downstream serving setting.

Tab. 5 shows that I2U training consistently improves Cold R@10 across all evaluated architectures. SASRec improves from 1.72 to 2.53, BERT4Rec from 1.98 to 2.75, HSTU from 2.15 to 3.07, and NU from 2.47 to 3.38. PD^2 similarly improves from 2.82 under the mirrored U2I formulation to 4.03 under I2U training. At the same time, PD^2 achieves the strongest performance under both training formulations. Under U2I training, PD^2 improves Cold R@10 from the strongest baseline value of 2.47 to 2.82, while under I2U training the corresponding improvement is from 3.38 to 4.03. This suggests that the gain of PD^2 is not explained solely by the change in training direction, since its architectural design also yields additional improvements under a fixed formulation.

The effect of I2U training is most consistent for cold-start items, while the changes on warm and overall items vary across architectures. We therefore interpret these results as evidence that I2U supervision is useful for cold-start representation learning, rather than as evidence that I2U training universally outperforms U2I training in all recommendation settings.

D.4 Further analysis in the cold-start scenario

In Fig. 7, we present a detailed comparison of all methods in cold-start scenarios. Under the item-to-user paradigm, even items with no historical interactions can still be matched to their potential next users by methods such as SASRec and TIGER, yielding non-negligible hit rates. In contrast, under the conventional user-centric paradigm, these methods typically suffer from severe cold-start collapse, with hit rates close to zero Shen et al. (2026); Ding et al. (2026).

Notably, PD^2 demonstrates strong robustness under the item-to-user paradigm. Its performance on zero-interaction items remains consistently strong and is only slightly lower than its performance on popular items. This further validates the compatibility and effectiveness of our method for item-to-user recommendation.

Table 5: Performance under different training formulations on KuaiRand-1K. All models are evaluated using the same standard U2I retrieval pipeline. Results are Recall@10 (R@10) reported in percentage (%). The best and second-best results under each training formulation are marked in **bold** and underline, respectively.

Method	U2I Training			I2U Training		
	Cold R@10	Warm R@10	All R@10	Cold R@10	Warm R@10	All R@10
SASRec	1.72	4.37	4.03	2.53	4.08	3.88
BERT4Rec	1.98	4.19	3.91	2.75	4.13	3.95
HSTU	2.15	<u>5.19</u>	<u>4.80</u>	3.07	<u>5.02</u>	<u>4.77</u>
PAM	2.38	4.15	3.92	3.29	4.66	4.48
NU	<u>2.47</u>	4.52	4.26	<u>3.38</u>	4.57	4.42
PD^2	**2.82**	**5.53**	**5.19**	**4.03**	**5.66**	**5.45**

Table 6: Online A/B testing results on the large-scale production platform.

Comparison	Statistic	LTR (↑)	WTR (↑)	CMTR (↑)	SVTR (↓)
PD^2-U2I vs. Base	Base	4.72%	0.45%	0.64%	31.23%
	PD^2-U2I	5.21%	0.48%	0.69%	26.75%
	Relative change	+10.48%	+7.31%	+7.39%	-14.34%
	95% CI	[9.85, 11.12]%	[5.90, 8.73]%	[6.05, 8.75]%	[-15.02, -13.65]%
	p-value	< 0.01	< 0.01	< 0.01	< 0.01
PD^2-I2U vs. PD^2-U2I	PD^2-U2I	5.21%	0.48%	0.69%	26.75%
	PD^2-I2U	8.75%	0.86%	0.97%	20.84%
	Relative change	+67.91%	+77.64%	+40.82%	-22.10%
	95% CI	[65.80, 70.05]%	[73.90, 81.47]%	[38.10, 43.59]%	[-23.05, -21.12]%
	p-value	< 0.01	< 0.01	< 0.01	< 0.01

D.5 Online A/B Test Results

To evaluate the practical effectiveness of PD^2 in real-world scenarios, we conduct a large-scale online A/B test on a leading short-video platform with over 400 million daily active users. The experiment uses 5% of the online traffic and adopts user-level randomization, where users are randomly assigned to fixed experimental groups and remain in the same group throughout the experiment. We first conduct a 6-day A/A test to verify the stability of the experimental setup, followed by a 7-day A/B test. We consider four online engagement metrics: **L**ike-**T**hrough **R**ate **(LTR)**, Follo**W**-**T**hrough **R**ate **(WTR)**, **C**omment-**T**hrough **R**ate **(CMTR)**, and **S**hort **V**iew **T**hrough **R**ate **(SVTR)**. Higher values are preferred for LTR, WTR, and CMTR, while a lower SVTR indicates fewer short-view behaviors and is therefore preferred.

We compare three serving configurations. PD^2-U2I denotes the conventional user-to-item serving pipeline, where item representations are learned by PD^2 and indexed for standard user-driven retrieval. PD^2-I2U denotes the proactive item-to-user serving pipeline, where candidate items retrieve potential seed users and the resulting item–user pairs are asynchronously cached for subsequent user requests. The two PD^2 serving configurations use the same exploration budget, allowing us to compare different user-allocation strategies without introducing additional exploration traffic.

User-to-item serving. Compared with the production Base system, PD^2-U2I improves LTR, WTR, and CMTR by 10.48%, 7.31%, and 7.39%, respectively, while reducing SVTR by 14.34%. Since this configuration retains the conventional U2I serving direction while using representations learned under our item-centric multi-user training formulation, the results indicate that the learned item representations can also benefit the existing user-driven retrieval pipeline.

Item-to-user serving. We further compare PD^2-I2U with PD^2-U2I to evaluate the practical effect of deploying the model through the proactive item-to-user serving configuration. Under the same

Table 7: Relative online metric changes across different video play-count ranges. All values are relative changes (%) with respect to the corresponding reference system.

(a) PD²-U2I vs. Base

Video play count	LTR (↑)	WTR (↑)	CMTR (↑)	SVTR (↓)
< 200	+15.6	+12.8	+11.9	-18.6
$[200, 2\text{k})$	+12.4	+9.7	+9.2	-16.2
$[2\text{k}, 4\text{k})$	+8.3	+7.3	+7.1	-13.4
$[4\text{k}, 10\text{k})$	+9.4	+7.5	+7.4	-12.5
$\geq 10\text{k}$	+9.8	+7.0	+7.2	-13.8
All	+10.5	+7.3	+7.4	-14.3

(b) PD²-I2U vs. PD²-U2I

Video play count	LTR (↑)	WTR (↑)	CMTR (↑)	SVTR (↓)
< 200	+60.4	+63.7	+34.6	-16.8
$[200, 2\text{k})$	+62.5	+69.3	+38.2	-20.5
$[2\text{k}, 4\text{k})$	+73.2	+81.8	+51.5	-26.9
$[4\text{k}, 10\text{k})$	+65.2	+83.4	+39.7	-22.7
$\geq 10\text{k}$	+67.5	+75.6	+41.9	-21.4
All	+67.9	+77.6	+40.8	-22.1

exploration budget, PD²-I2U further improves LTR, WTR, and CMTR by 67.91%, 77.64%, and 40.82%, respectively, while reducing SVTR by 22.10%.

Unlike conventional U2I serving, the I2U pipeline proactively retrieves potential seed users for candidate items and asynchronously stores the resulting item–user pairs. The selected users do not need to be simultaneously online; the cached candidates are served when the corresponding users subsequently initiate recommendation requests. If a predicted user does not return within the effective serving window, the corresponding item–user pair simply does not result in an exposure. Therefore, this comparison reflects the effect of the complete proactive I2U serving configuration rather than a purely mathematical reversal of the retrieval direction.

The substantially improved engagement metrics suggest that proactive seed-user retrieval can allocate limited cold-start exploration opportunities to more suitable users, providing complementary benefits to the conventional U2I serving pipeline.

Performance across item-popularity levels. To examine whether the online improvements are concentrated in a particular item-popularity range, we further partition videos into five non-overlapping groups according to their play counts. Tab. 7 reports the relative metric changes within each group.

For PD²-U2I versus Base, consistent improvements are observed across all popularity ranges. In particular, videos with fewer than 200 plays obtain the largest LTR improvement of 15.6%, together with a 12.8% improvement in WTR and an 11.9% improvement in CMTR. This indicates that the learned representations are particularly effective for low-exposure items. The additional gains from PD²-I2U are also consistently observed across all five popularity groups. For example, LTR improvements range from 60.4% to 73.2%, while WTR improvements range from 63.7% to 83.4%. The gains are therefore not restricted to a single popularity range, suggesting that the proactive I2U serving mechanism remains effective across items with different levels of historical exposure.

Summary. Overall, the online results provide complementary evidence for two aspects of PD². First, representations learned under the proposed multi-user training formulation can improve the conventional U2I retrieval pipeline. Second, deploying the model through proactive I2U serving yields substantial additional improvements under the same exploration budget. The improvements are statistically significant and remain consistent across different video play-count ranges, support-

Table 8: Token Generation Speed Comparison

Method	2 Tokens	5 Tokens	10 Tokens
Autoregressive Generation	22.164 ms	54.878 ms	114.825 ms
Parallel Generation	11.916 ms	11.931 ms	12.087 ms
Speedup	1.9×	4.6×	9.5×

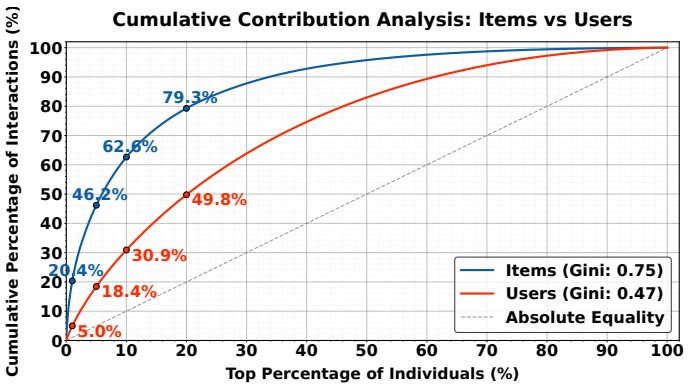

Figure 8: Cumulative distribution of interactions across items and users.

ing the practical applicability of PD^2 in large-scale cold-start recommendation. Following the online A/B test, PD^2 was deployed in the production retrieval pipeline.

D.6 Inference Efficiency

In Tab. 8, we report the inference efficiency of parallel generation. The results show that parallel generation provides a clear advantage over autoregressive generation when generating multiple tokens. Although multi-channel parallel generation is theoretically more computationally complex, the additional computation incurs little practical overhead due to efficient GPU parallelization. Compared with the conventional token-by-token sequential decoding process used by autoregressive models, parallel generation achieves substantially better inference efficiency.

E Impact of the Recommendation Paradigm

E.1 Next-user generation

In Tab. 6, we show that next-user modeling improves online performance, and that its benefits are further amplified when combined with item-to-user inference. To provide a more comprehensive evaluation of this novel paradigm, we also report UE-score, which measures users' overall satisfaction with videos in industrial short-video scenarios and serves as a key intermediate metric for driving long-term user retention, especially LT7. With item-to-user inference, UE-score improves by 33.1% ($p < 0.01$); notably, in industrial short-video scenarios, even a 10% improvement is typically considered substantial.

In addition, Fig. 8 illustrates the popularity bias on the item and user sides. We observe that the bias is much more pronounced on the item side, closely following the Pareto principle: the top 20% of items account for 79.3% of all interaction records. In contrast, this effect is substantially mitigated on the user side, where the top 20% of users account for only 49.8% of the interaction records. This suggests that the item-to-user paradigm does not fully inherit the limitations of user-to-item recommendation (e.g., newly published items are not simply pushed to a small group of highly active head users).

E.2 PARALLEL GENERATION

Fig. 4 show that as the number of inference tokens increases, the diversity of the overall TopK user set generated also increases. In cold-start scenarios, diversity is a metric directly related to core business objectives. Parallel generation can substantially improve recommendation diversity without sacrificing recommendation accuracy, which is crucial in cold-start settings.

In addition, we find that, in recommendation scenarios, the homogeneity of parallel multi-token generation with the discrete diffusion model is much less severe than that observed in LLMs. On average, at least half of the users generated by each channel are unique and do not appear in the outputs of other channels. This may stem from the item-to-user paradigm, where a large number of different types of users in the training data can interact with the same item within a short time window.

To avoid wasting the retrieval quota, we generate the TopK users in parallel for each channel, retain each user's highest ranking across different channels, and finally select the overall TopK users with the highest rankings.

F THE USE OF LARGE LANGUAGE MODELS (LLMS)

We used ChatGPT to refine writing only. The prompt was: "I am preparing a paper for submission to an international conference and would like your help to check for any grammatical issues and refine the wording or sentence structure where necessary to ensure conciseness and precision." Edits were applied paragraph-by-paragraph, and all outputs were verified and revised by the authors; no scientific content, analyses, or references were generated by the tool.