SKIPGR: ACCELERATING GENERATIVE RECOMMENDATIONS WITH EFFICIENT SKIPPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative Recommendation Models (GRMs) frame recommendation tasks as autoregressive sequence modeling, achieving strong accuracy. However, training on all exposure tokens or sequences can be inefficient, as user logs often contain a large proportion of redundant or uninformative interactions that inflate token budgets without significant learning gains. To address this inefficiency, we propose the concept of semantic token entropy, which maps an item's multimodal content into a compact semantic vocabulary, enabling tractable entropy estimation. We trained a GR-4B model from scratch on an industrial-scale recommendation corpus and analyzed the token entropy. The analysis reveals that high-entropy tokens and sequences lead to larger gradient updates. Leveraging these insights, we developed a skipping policy based on semantic entropy, named SKIPGR, that selectively and adaptively bypasses uninformative tokens and sequences. This approach accelerates convergence and reduces computational costs, resulting in even superior outcomes compared to using the full corpus. Extensive empirical evaluations on both public datasets and large-scale industrial datasets have validated the enhanced performance.

1 INTRODUCTION

Generative Recommendation Models (GRMs) have revolutionized recommendation systems by integrating tasks like item retrieval, ranking, and explanation through autoregressive sequence generation. By capturing user interaction histories as source sequences while viewing target items as tokens, GRMs leverage the scalability associated with large language models (LLMs), driving strong performance across diverse recommendation tasks. Models such as GR (Zhai et al., 2024) and OneRec (Deng et al., 2025) showcase the paradigm's ability to excel in long-sequence prediction, surpassing traditional models like SASRec (Kang & McAuley, 2018) and DIN (Zhou et al., 2018), etc. However, these advancements entail significant computational costs, especially in industrial applications involving vast user sequences. Training GRMs with billions of parameters often demands substantial token budgets, exceeding those of LLMs, and requires extensive GPU time, leading to notable financial and computational implications. This highlights the urgent need for more efficient training methods that maintain high-quality recommendations while alleviating resource demands.

Current GRMs assume uniform informativeness across user behavior sequences, processing entire user histories akin to LLMs (Feng et al., 2025). This approach neglects key differences between user logs and natural language texts. User sequences often contain behavioral noise from repetitive actions or low engagement, and mismatched intentions, such as interactions driven by reward triggers or passive swiping. These elements can make short session segments uninformative or misleading despite appearing valid. Additionally, recommendation data exhibits a long-tail distribution, where rarely viewed items complicate the efficient training of models. These factors imply that much of the training data provides limited value, hindering convergence and increasing computational costs.

We conduct a preliminary analysis on the Alipay short-video corpus. We observe distinct patterns: 10% of users account for 70+% of the *page view* (PV) samples. Further examination reveals that 30–50% of sequences include identifiable uninformative items, with 20% of items being viewed for less than 3. These sequences align with known behavioral patterns such as task-oriented browsing or passive usage. This raises a natural question: *Can we skip such uninformative tokens or sequences without degrading GRM performance?*

To this end, we propose SKIPGR, a dynamic pruning framework that accelerates GRM training by skipping uninformative tokens and sequences via their semantic token entropy. SKIPGR begins with semantic tokenization, converting user items into compact semantic tokens by utilizing multimodal large language model (MM-LLM) embeddings, such as Qwen2.5-VL (Bai et al., 2025), combined with vector quantization. Semantic token entropy is assessed at both the token level, indicating the uncertainty of individual items, and the sequence level, reflecting global diversity within a user's history. This assessment enables the selective skipping of noisy segments or sequences while retaining valuable tokens. Guided by these entropy signals, SKIPGR accelerates convergence and reduces computational costs, even achieving superior results compared to using the full corpus.

Our Contributions. 1) We design a comprehensive approach for calculating token and sequence entropy within the semantic space, enabling more effective identification and skipping of uninformative elements while preserving crucial information for model training. 2) We introduce SKIPGR, a novel framework designed to enhance GRM training efficiency by dynamically skipping uninformative tokens and sequences guided by semantic token entropy, thereby reducing training complexity without sacrificing accuracy. 3) Through extensive empirical evaluations on both public datasets and large-scale industrial datasets from Alipay, we demonstrate that SKIPGR effectively decreases training time and computational costs, while providing lossless or even enhanced performance.

RELATED WORK

Sequential recommendation models. Deep sequential models are a core paradigm in industrial recommendation systems (Chang et al., 2023; Chen et al., 2021; Cao et al., 2022). Early work such as DIN (Zhou et al., 2018) introduced interest-aware embeddings, followed by extensions including DIEN (Zhou et al., 2019), CAN (Bian et al., 2021), and TWIN (Chang et al., 2023; Si et al., 2024). More recently, transformer-based variants like LONGER (Chai et al., 2025) have been optimized for GPU efficiency. While these models improve the expressiveness of sequential recommendation, they are often developed in isolation from upstream and downstream tasks, limiting opportunities for end-to-end optimization.

Generative recommendation models. Generative approaches reformulate recommendations as sequence generation, where the model autoregressively outputs item identifiers given user history (Rajput et al., 2023; Zhai et al., 2024; Deng et al., 2025). GENRE (Cao et al., 2021) demonstrated the feasibility of transformer-based entity generation, and later works such as TIGER (Rajput et al., 2023) and LD-REC (Zheng et al., 2024) applied residual quantization for scalable training. OneRec (Deng et al., 2025) further scaled this formulation, leveraging LLM scaling laws to improve both accuracy and business metrics. Despite these advances, most GRMs assume all user sequences are equally informative, overlooking noise and low-utility behaviors that reduce training efficiency.

Dataset pruning for generative models. While scaling models and datasets has improved performance, pruning studies reveal inefficiencies when all data are used indiscriminately (Qin et al., 2023; Wang et al., 2025a). Recent methods address this by pruning along different dimensions. Rho-1 (Lin et al., 2024) trains selectively on high-utility tokens, achieving results comparable to DeepSeekMath with only 3% of the pre-training data. TokenSkip (Xia et al., 2025) enables LLMs to skip unimportant tokens for efficient CoT compression, while LVPruning (Sun et al., 2025) prunes vision tokens to streamline multimodal LLM training. However, in industrial-scale GRMs with massive token volumes, token and sequence pruning remains largely unexplored; existing work (Petrov et al., 2025) has focused primarily on inference-time item pruning.

2 PRELIMINARIES

Notations. Given a set of N items $\{\Phi_i\}_{i=1}^{N}$, Φ_i and an action set of A actions $\{a_i\}_{i=1}^{A}$, *e.g.*, like, skip, video completion, and share. We use $\mathcal{X}_c$ to denote the token item and use $\mathcal{X}_a$ to denote the collection of a_i. $\mathcal{V} = \mathcal{X}_c \cup \mathcal{X}_a$ is the vocabulary of the recommendation system. For each user u, let $o_t^{(u)} = (\Phi_t^{(u)}, a_t^{(u)})$ is the item-action pair at timestep t. Thus, we use a time-ordered sequence $\boldsymbol{o}^{(u)} = (o_1^{(u)}, \ldots, o_T^{(u)})$ to denote the observed history item-action interaction, with a fixed observed sequence length T. $\mathcal{O} = \{\mathbf{o}_i\}_{i=1}^{M}$ is the colletions of all users' histories sequence. For brevity, when clear from context we drop the user superscript on $\boldsymbol{o}^{(u)}$ and simply write $\boldsymbol{o}$.

Generative Recommendation Models. We focus on transformer-based generative recommendation models (GRMs) (Zhai et al., 2024) parameterized by θ and denoted π_θ. Given a user's interaction context $\boldsymbol{o}_{<t} = (o_0, \ldots, o_{t-1})$, the parameters θ are found by maximizing the expected log-likelihood:

$$\mathbb{E}_{\mathbf{o}\sim\mathcal{O}}\Big[\sum_{t=1}^{T} \log \pi_\theta(o_t \,|\, \mathbf{o}_{<t})\Big]. \tag{1}$$

Token Entropy for Large Language Models. Token entropy (Wang et al., 2025b) measures the uncertainty of the next-token prediction under the large language models' (LLMs) posterior. Given a large language model π_θ. During text generation, the model outputs a probability distribution $\boldsymbol{p}_t \in \mathbb{R}^{|\mathcal{V}|}$ over the vocabulary $\mathcal{V}$, conditioned on the given prompt $\boldsymbol{q}$ and the sequence of previously generated tokens $\boldsymbol{o}_{<t}$. The token-level generation entropy is defined as

$$H_t := -\sum_{j=1}^{|\mathcal{V}|} p_{t,j} \log p_{t,j}, \quad \text{with } (p_{t,1}, \ldots, p_{t,|\mathcal{V}|}) = \boldsymbol{p}_t = \pi_\theta(\cdot|\boldsymbol{q}, \boldsymbol{o}_{<t}) = \texttt{softmax}(\frac{z_t}{\rho}) \tag{2}$$

where $z_t \in \mathbb{R}^{|\mathcal{V}|}$ denotes the pre-softmax logits at time step t, $\rho \in \mathbb{R}$ is the decoding temperature.

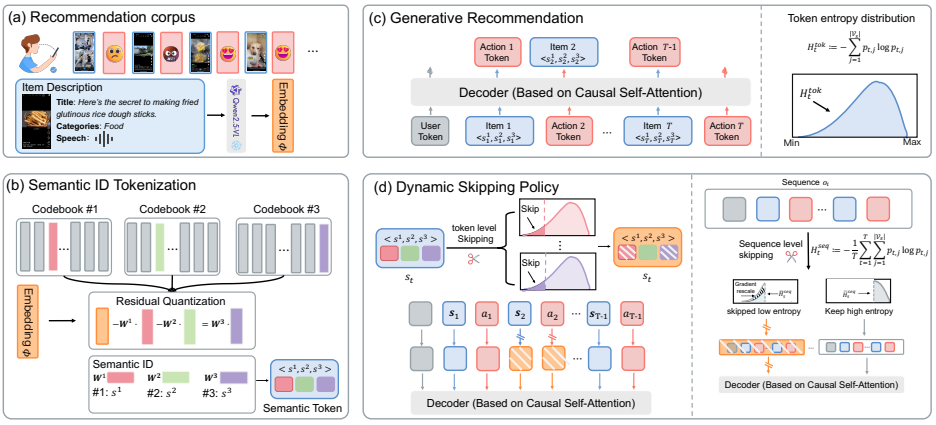

Figure 1: Illustration of the proposed SKIPGR framework. SKIPGR consists of two main skipping policies: token-level and sequence-level. (a) Recommendation Corpus: Multimodal content, e.g., images, descriptions, is embedded using a pretrained model (Qwen2.5-VL) to generate semantic embeddings Φ. (b) Semantic ID Tokenization: The embeddings are tokenized into semantic IDs using residual quantization, producing a sequence of semantic tokens. (c) Generative Recommendation: The GRM generates sequences of item and action tokens autoregressively, using causal self-attention to predict the next token. (d) Dynamic Skipping Policy: The dynamic skipping policy applies skipping at both token and sequence levels. At the token level, subtokens with low entropy are skipped, and at the sequence level, entire low-entropy sequences are skipped. Gradient rescaling is applied to ensure skipped sequences still contribute to training.

3 METHODOLOGY

3.1 PROBLEM DEFINITION

We aim to skip uninformative tokens or sequences during GRM training selectively. Specifically, we discard tokens $o_t \in \mathcal{V}$ or sequences $\boldsymbol{o} \in \mathcal{O}$ that satisfy skipping conditions, allowing the vocabulary $\mathcal{V}$ or sequence set $\mathcal{O}$ to evolve dynamically across epochs. Following dataset pruning practices (Toneva et al., 2019), we define a token-level skipping policy $\mathcal{H}_\tau^{tok}$ and a sequence-level skipping policy $\mathcal{H}_\tau^{seq}$ that assigns an information score to each token $o_{i,t}$ or sequence $\boldsymbol{o}_i$ at epoch τ. The skipping probability is then defined as

$$\mathcal{P}_\tau^{\text{tok}}(o_{i,t}) = \mathbb{1}\big(\mathcal{H}_\tau^{tok}(o_{i,t}) < \varepsilon_x\big), \qquad \mathcal{P}_\tau^{\text{seq}}(\boldsymbol{o}_i) = \mathbb{1}(\mathcal{H}_\tau^{seq}(\boldsymbol{o}_i) < \varepsilon_o), \tag{3}$$

where $\mathcal{H}_\tau^{seq}(\boldsymbol{o}_i) = \frac{1}{T}\sum_{t=1}^{T}\mathcal{H}_\tau(o_{i,t})$ is the mean of the tokens within sequence $\boldsymbol{o}_i$. ε_x and ε_o are adaptive thresholds for tokens and sequences, and $\mathbb{1}(\cdot)$ is the indicator function.

We then construct a skipped training corpus at epoch τ as $\mathcal{S}_\tau = \big\{o_{i,t} \mid \mathcal{P}_\tau^{\text{tok}}(o_{i,t}) = 1 \wedge \mathcal{P}_\tau^{\text{seq}}(\mathbf{o}_i) = 1\big\}$, that is, $\mathcal{S}_\tau \subseteq \mathcal{V}$ retains only the tokens $o_{i,t}$ that are not skipped at token level and whose sequence $\mathbf{o}_i$ is not discarded at sequence level. Thus, minimizing the cross-entropy loss under a skipped corpus leads to the following optimization problem:

$$\min_\theta \mathbb{E}_{\mathbf{o}_i\sim\mathcal{S}_\tau}\frac{1}{|\mathcal{S}_\tau|}\sum_{i=1}^{M}\sum_{t=1}^{T}\mathcal{P}_\tau^{\text{seq}}(\mathbf{o}_i)\,\mathcal{P}_\tau^{\text{tok}}(o_{i,t})\,\big[-\log\pi_\theta(o_{i,t}\mid \boldsymbol{o}_{i,<t})\big], \tag{4}$$

where π_θ is the GRM parameterized by θ, $\boldsymbol{o}_{i,<t}$ is the history before token t in sequence $\mathbf{o}_i$, and $\mathcal{S}_\tau$ denotes the dynamic skipped corpus at training step τ.

3.2 Semantic token entropy in generative recommendation models

While prior work has emphasized the significance of token entropy in enhancing reasoning processes of LLMs, such as in (Wang et al., 2025b), where high-entropy tokens are filtered for post-training using Reinforcement Learning with Verifiable Reward (RLVR; Guo et al. 2025). However, directly applying token entropy in GRMs poses considerable challenges, for the presence of extremely large and redundant corpora, leading to $|\mathcal{V}|$ that can exceed 10B. This motivates us to define entropy at a *semantic token* level (Rajput et al., 2023; Deng et al., 2025), where each item is mapped into a compact but meaningful semantic vocabulary, enabling tractable and informative entropy estimation.

Semantic token entropy. We assume that each item is associated with multimodal content (*e.g.*, titles, descriptions, images, or speech transcripts) $C_i = \{\texttt{text}_i, \texttt{images}_i, \ldots\}$. Let $\mathbb{C}$ denote the multimodal content space. Given multimodal content $C_i \in \mathbb{C}$ for item i, we leverage a pretrained multimodal large language model as encoder $\texttt{MM-LLM}$ as an encoder to transform these multimodal features into a d-dimensional embedding Φ_i

$$\Phi_i = \texttt{MM-LLM}(C_i) \tag{5}$$

where $\texttt{MM-LLM}: \mathbb{C} \mapsto \mathbb{R}^d$ is instantiated by Qwen2.5-VL (Bai et al., 2025).

This embedding Φ_i is then mapped to sequences of sparse IDs $\boldsymbol{s}$ via a hierarchical clustering tokenizer. Since residual quantization variational autoencoder (RQ-VAE; Lee et al. 2022) has demonstrated its effectiveness as a tokenizer for GRMs (Ju et al., 2025). Following this fashion, we adopt Residual Quantization Variational Autoencoder (RQ-VAE) (Lee et al., 2022) with joint codebook optimization (Zhu et al., 2024), which has been shown effective for GRMs (Ju et al., 2025), to discretize the semantic embeddings. Given a quantization depth L, we use the $\texttt{Tokenizer}$ represents $\boldsymbol{\Phi}$ as an ordered sequence of L codes

$$\boldsymbol{s} = \texttt{Tokenizer}(\boldsymbol{\Phi}; \mathcal{Q}, L) = (s^1, s^2, \ldots, s^L) \in [K]^L \tag{6}$$

where $\mathcal{Q}$ is the codebook of size $|\mathcal{Q}| = K$, and s^l represents a *subtoken* at depth l of the semantic hierarchy of $\boldsymbol{\Phi}$. The set of all such subtokens $\{s^1, s^2, \ldots, s^L\}$ together represent the semantic token Φ_i, where lower layers (l) capture coarser, broader semantic concepts, and deeper layers capture more fine-grained details of the token. Starting with 0-th residual $s^0 = \boldsymbol{\Phi}$, the $\texttt{Tokenizer}$ recursively computes s^l, which is the code of the residual q_{l-1} and the text residual q_l as

$$s^l = \arg\min_k \|q_l - q_k^l \mathbf{W}^l\|_2^2; \qquad q_l = q_{l-1} - q_{s^l}^l \mathbf{W}^l \tag{7}$$

where $\mathbf{W}$ is the collection of learnable linear layers with $\mathbf{W}^l \in \mathbb{R}^{d\times d}$, $l = 1, \ldots, L$. The final discretized vector is defined as $\hat{\boldsymbol{\Phi}}^{(l)} = \sum_{i=1}^{l} q_{s^i}^i \mathbf{W}^i$, and $\hat{\boldsymbol{\Phi}} := \hat{\boldsymbol{\Phi}}^{(D)}$ represents the quantized version of $\boldsymbol{\Phi}$. Figure 1 (a) illustrates the process in a high-level overview.

The discrete semantic vocabulary is then given by $\mathcal{V}_s = \{s^1, \ldots, s^{|\mathcal{V}_s|}\}$, where each item token Φ_i is mapped into a semantic token $\tilde{\Phi}_t = \texttt{Tokenizer}(\boldsymbol{\Phi}_t; \mathcal{Q}, L)$. For convenience, let $\tilde{o}_t = (\tilde{\Phi}_t, a_t)$ denote the token and its associated attributes. The GRM π_θ autoregressively predicts distributions over these semantic tokens:

$$\boldsymbol{p}_t = \pi_\theta(\cdot \mid \tilde{o}_{i,<t}). \tag{8}$$

Thus, the semantic generation entropy, capturing the uncertainty of the model's predictions over the semantic tokens, is defined as:

$$H_t^{tok} := -\sum_{j=1}^{|\mathcal{V}_s|} p_{t,j} \log p_{t,j}, \tag{9}$$

where $\mathcal{V}_s$ is the semantic vocabulary produced by the quantizer.

3.3 Dynamic skipping policy in generative recommendation models

Building on the semantic token entropy in Eq. (9), we examine how token entropy influenced the model's performance during standard GRM pre-training. Concretely, we train GR-4B (Zhai et al., 2024) from scratch on user queries from the Alipay short-video platform. Tokens are divided into four groups based on the semantic token entropy.

Figure 2 reveals a consistent trend: tokens with higher entropy are associated with larger gradient magnitudes during training, suggesting that they provide stronger learning signals; in contrast, low-entropy regions align with trivial or redundant behaviors and yield comparatively smaller updates.

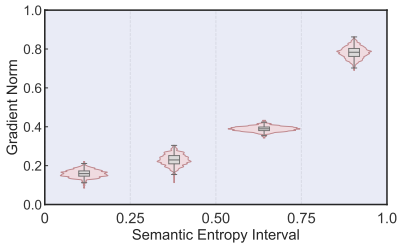

Figure 2: Analysis of the gradient norms on the Alipay short video dataset during training of GR-4B. Tokens are divided into four groups based on the quartiles of their semantic entropy.

This observation motivates a skipping strategy that preferentially retains high-entropy content while eliding less informative regions. Motivated by these observations, we introduce SkipGR, a novel dynamic pruning framework that accelerates GRM training by skipping uninformative tokens and sequences for achieving lossless training acceleration, as sketched in Figure 1 (d). We now present the details of each component.

Token-level skipping policy. In this policy, we skip individual subtokens within each semantic token, rather than skipping the entire item. This selective skipping discards less informative subtokens while retaining the most meaningful semantic information, improving computational efficiency without losing essential details.

At each training step τ, we introduce a *skipping ratio* $r^{\text{tok}} \in (0, 1]$, which controls the expected fraction of subtokens to be skipped. Each semantic token $\tilde{o}_{i,t}$ is represented by a hierarchical sequence of L sub-tokens:

$$\tilde{o}_{i,t} = [s_{i,t}^1, \ldots, s_{i,t}^L], \tag{10}$$

where $s_{i,t}^l \in [K]$ denotes the subtoken at layer l. Conditioned on realized subtokens up to layer $l-1$ and the history $\boldsymbol{o}_{i,<t}$, the GRM emits a layer-wise predictive distribution $\boldsymbol{p}_t^{(l)} = \pi_\theta(\cdot \mid \tilde{o}_{i,t}^{(<l)}, \boldsymbol{o}_{i,<t})$. We define the *layer-wise semantic token entropy*,

$$H_t^{(l)} = -\sum_{j=1}^{K} p_{t,j}^{(l)} \log p_{t,j}^{(l)}, \tag{11}$$

which quantifies model uncertainty at granularity l.

We implement skipping with binary layer-wise skip indicators $\mathcal{P}_\tau^{(l)}(s_{i,t}^l) \in \{0, 1\}$, where 1 means to *skip*. Let $r_\tau^{(l)} \in [0, 1]$ be the per-layer skipping ratios, for each layer l, we compute the $r_\tau^{(l)}$-quantile threshold $\tau_\tau^{(l)}$ over the entropy values $\{H_t^{(l)}\}_{t \in \mathcal{S}_\tau}$, and set:

$$\mathcal{P}_\tau^{(l)}(s_{i,t}^l) = \mathbb{1}(H_t^{(l)} < \bar{H}_\tau^{(l)}), \tag{12}$$

which skips the lowest-entropy $r_\tau^{(l)}$ fraction of subtokens at layer l.

Then, the skipping mechanism applies across all hierarchical subtokens within $\tilde{o}_{i,t}$. Practically, this involves:

$$\mathcal{P}_\tau^{\text{tok}}(s_{i,t}^l) = \begin{cases} 0 & \text{if } s_{i,t}^l \text{ ranks in the top } r^{tok}\% \text{ of } H_t^{tok} \\ 1 & \text{otherwise} \end{cases}, \tag{13}$$

Notably, a token is skipped only if *all* its subtokens are skipped:

$$\mathcal{P}_\tau^{\text{tok}}(o_{i,t}) \;=\; \prod_{l=1}^{L} \mathcal{P}_\tau^{(l)}(s_{i,t}^l), \tag{14}$$

which induces a token-level skipping policy that is conservative at coarse layers yet allows fine-grained retention when uncertainty concentrates at specific depths.

Sequence-level skipping policy. Sequence-level focuses on identifying and skipping continuous whole sequences that contain redundant or noisy information. This improves the efficiency of GRMs by retaining only the most informative sequences, based on their semantic entropy. For each semantic sequence $\tilde{\mathbf{o}}$, we first calculate the sequence-level semantic entropy H^{seq} as its information score,

$$H^{seq} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{|\mathcal{V}_s|} p_{t,j}\log p_{t,j}, \tag{15}$$

Then, the decision to prune a sequence is based on a skipping probability $\mathcal{P}_\tau^{seq}$ at each time step τ, which depends on the entropy of the sequence H^{seq}. We define a soft probability function as follows:

$$\mathcal{P}_\tau^{\text{seq}}(\tilde{\mathbf{o}}) = \begin{cases} r^{seq} & H^{seq}(\tilde{\mathbf{o}}) < \bar{H}_\tau^{seq} \\ 0 & H^{seq}(\tilde{\mathbf{o}}) \geq \bar{H}_\tau^{seq} \end{cases}, \tag{16}$$

where $\bar{H}^{seq}$ is the mean value of all the sequences that adjust dynamically based on the model's learning progress.

For sequences with $H^{\text{seq}}(\tilde{\mathbf{o}}) < \bar{H}_\tau^{\text{seq}}$, we scale their gradients by $1/(1 - r^{\text{seq}})$, which increases their learning signal. This ensures that even lower-entropy, potentially redundant sequences still contribute meaningfully to training, helping the model avoid overlooking important patterns. Sequences with scores greater than or equal to $\bar{H}_\tau^{\text{seq}}$ are retained and not skipped.

4 Experimental Resualts

We first evaluate the performance of SkipGR through extensive experiments on both public and industry datasets. For public datasets, we conduct a series of experiments using MovieLens, Amazon Reviews, and Amazon Product Reviews. We follow sequential recommendation settings in literature, using the last item per user for test, second-to-last for validation, and the rest for training. We report Recall@K and NDCG@K for Amazon Product Reviews, Hit Rate@K and NDCG@K for MovieLens and Amazon Reviews, consistent with recent work. For industry datasets, we perform a series of experiments using the Alipay short video platform, report *gauc* and *wgauc* over the entire corpus. All results are averaged over 5 runs with different seeds. Additional experiments on different skipping probabilities, along with ablation studies, are provided in Appendix **??**.

4.1 Amazon Product Reviews results

Amazon product reviews (He & McAuley, 2016) are a public, real-world benchmark that contains user reviews and item metadata from May 1996 to July 2014. We compare SkipGR for the generative retrieval task with the following sequential recommendation methods: P5 (Geng et al., 2022), Caser (Tang & Wang, 2018), HGN (Wang et al., 2025c), GRU4Rec (Hidasi et al., 2016), BERT4Rec (Sun et al., 2019), FDSA Zhang et al. (2019), SASRec (Kang & McAuley, 2018), GRID (Ju et al., 2025). Among them, the other algorithms all utilize a dual encoder and Maximum Inner Product Search (MIPS) to generate next-item predictions, whereas GRID is a standard generative recommendation model. Therefore, we primarily compare it with GRID.

The results for all baselines are taken from the publicly accessible results made available by Rajput et al. (2023) and Ju et al. (2025). The results are shown in Table 1. We observe that GRID outperforms most of the existing baselines. Our approach maintains over 95% performance after reducing 10% of tokens and 30% of sequences, and even achieves better results on the overall datasets.

Table 1: Performance comparison on sequential recommendation. The last row depicts % improvement with SKIPGR relative to the best baseline.

Methods	Sports and Outdoors				Beauty				Toys and Games			
	Recall @5	NDCG @5	Recall @10	NDCG @10	Recall @5	NDCG @5	Recall @10	NDCG @10	Recall @5	NDCG @5	Recall @10	NDCG @10
P5	0.0061	0.0041	0.0095	0.0052	0.0163	0.0107	0.0254	0.0136	0.0070	0.0050	0.0121	0.0066
Caser	0.0116	0.0072	0.0194	0.0097	0.0205	0.0131	0.0347	0.0176	0.0166	0.0107	0.0270	0.0141
HGN	0.0189	0.0120	0.0313	0.0159	0.0325	0.0206	0.0512	0.0266	0.0321	0.0221	0.0497	0.0277
GRU4Rec	0.0129	0.0086	0.0204	0.0110	0.0164	0.0099	0.0283	0.0137	0.0097	0.0059	0.0176	0.0084
BERT4Rec	0.0115	0.0075	0.0191	0.0099	0.0203	0.0124	0.0347	0.0170	0.0116	0.0071	0.0203	0.0099
FDSA	0.0182	0.0122	0.0288	0.0156	0.0267	0.0163	0.0407	0.0208	0.0228	0.0140	0.0381	0.0189
SASRec	0.0233	0.0154	0.0350	0.0192	0.0387	0.0249	0.0605	0.0318	0.0463	0.0306	0.0675	0.0374
GRID	0.0239	0.0154	0.0363	0.0194	0.0429	0.0285	0.0639	0.0353	0.0373	0.0237	0.0565	0.0300
/ + SKIPGR	0.0245	0.0163	0.0371	0.0205	0.0441	0.0298	0.0639	0.0362	0.0382	0.0240	0.0573	0.0300
Δ	↑2.51%	↑5.84%	↑2.20%	↑5.67%	↑2.79%	↑4.56%	↑0.00%	↑2.55%	↑2.41%	↑1.26%	↑1.42%	↑0.00%

4.2 MOVIELENS AND AMAZON REVIEWS RESULTS

In this section, we evaluate SKIPGR on the MovieLens and Amazon reviews datasets (Liu, 2025). We first follow the procedure outlined in Appendix **??**, including full shuffle and multi-epoch training. For baseline, we use SASRec (Kang & McAuley, 2018) and GR (Zhai et al., 2024), both state-of-the-art GRM implementations. Results are presented in Table 2. We are using identical configurations for our SKIPGR and baseline models following recent works (Zhai et al., 2023; Dallmann et al., 2021), *i.e.* same number of layers, heads, etc. "GR-Large" represents 4x layers and 2x heads of base GR encoders. Results show that 1) SKIPGR with its design efficiency accelerating optimized for recommendations, achieves comparable results to the baseline and even surpasses when using the same configuration, and 2) SKIPGR can further improve the scaling capability of the base model.

Table 2: Evaluations of methods on MovieLens and Amazon Reviews datasets in multi-pass, full-shuffle settings.

Methods	MovieLens-1M				MovieLens-20M				Amazon Reviews			
	HR @10	NDCG @10	HR @200	NDCG @200	HR @10	NDCG @10	HR @200	NDCG @200	HR @10	NDCG @10	HR @200	NDCG @200
SASRec	0.2853	0.1603	0.7528	0.2498	0.2906	0.1621	0.7655	0.2521	0.0292	0.0156	0.1400	0.0350
w / + SKIPGR	0.2794	0.1539	0.7533	0.2490	0.2899	0.1539	0.7673	0.2497	0.0283	0.0132	0.1420	0.0321
Δ	↓2.07%	↓3.99%	↑0.07%	↓0.32%	↓0.24%	↓5.06%	↑0.24%	↓0.95%	↓3.08%	↓15.38%	↑1.43%	↓8.29%
GR	0.3097	0.1720	0.7716	0.2620	0.3252	0.1878	0.7943	0.2774	0.0404	0.0219	0.1710	0.0450
w / + SKIPGR	0.2959	0.1668	0.7742	0.2594	0.3251	0.1874	0.7963	0.2774	0.0447	0.0300	0.1894	0.0498
Δ	↓4.46%	↓3.02%	↑0.34%	↓0.99%	↓0.03%	↓0.21%	↑0.25%	↑0.00%	↑10.64%	↑36.97%	↑10.76%	↑10.67%
GR-Large	0.3294	0.1893	0.7839	0.2771	0.3567	0.2106	0.8076	0.2971	0.0469	0.0257	0.1876	0.0508
w / + SKIPGR	0.3301	0.1902	0.7839	0.2779	0.3542	0.2092	0.8072	0.2838	0.0544	0.0241	0.2139	0.0599
Δ	↑0.21%	↑0.48%	↑0.00%	↑0.29%	↓0.70%	↓0.66%	↓0.05%	↓4.48%	↑15.99%	↓6.23%	↑14.02%	↑17.91%

4.3 ALIPAY SHORT VIDEO DATASET

In this section, we apply SKIPGR to an industrial dataset collected from the recommendation logs of the Alipay short video platform. The dataset has over 12 million users and 2.5 million videos. To ensure a fair comparison, all baseline algorithms adhered strictly to the network architectures as described in their original publications. Table 3 shows the performance comparison under the stay-

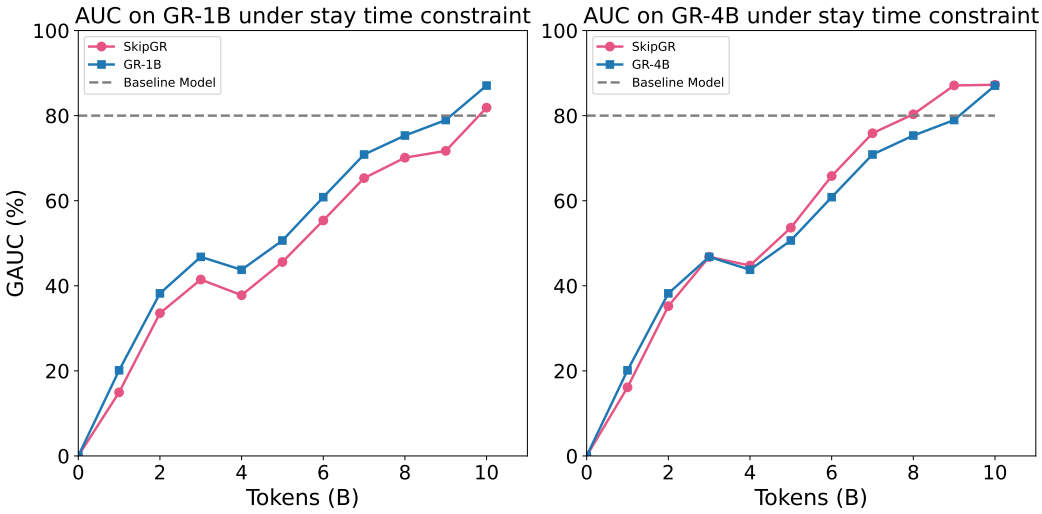

Figure 3: We continual train 1B and 4B GR with 10B Alipay short video tokens. While baselines are trained using a sequential recommendation model. SKIPGR improve AUC on Alipay short video by over 0.05%, achieving the baseline performance 1.1x-1.3x faster.

time of AUC and GAUC. We've observed that SkipGR also demonstrates superior performance on industrial datasets. Figure 3 demonstrates that SKIPGR significant improvements while reducing sequences by 30%∼50% and token computation by approximately 10%. Note that 1% improvement is a significant advancement in the recommendation domain.

Table 3: Evaluations of methods on Alipay datasets under view time

Metric	SASRec	SASRec+SKIPGR	Δ	GR	GR+SKIPGR	Δ	GRID	GRID+SKIPGR	Δ
AUC	0.8012	0.7983	↓0.36%	0.8707	0.8711	↑0.05%	0.8431	0.8509	↑0.93%
GAUC	0.5174	0.5169	↓0.10%	0.6044	0.6038	↓0.10%	0.5782	0.5871	↑ 1.54%

What tokens and sequences are selected? We aim to analyze the tokens and sequences chosen by the GRM in pretraining to explore its working mechanism further. To this end, we visualize the tokens and sequences selection process during the training of GRID. In Appendix, we have highlighted in blue the subtokens and sequences that were retained during actual pretraining. We observe that the majority of subtokens chosen by GRID are closely related to high-level and low-level subtokens, while the tokens that are skipped are mostly mid-level tokens. This may be because during training, high-level subtokens can directly indicate the category of an item, while low-level subtokens can clearly indicate the subtle differences between items in the same category. The discarded sequences are mostly for purposeful incentives for users, whose viewing time is fixed and from which they cannot obtain a fixed set of interests.

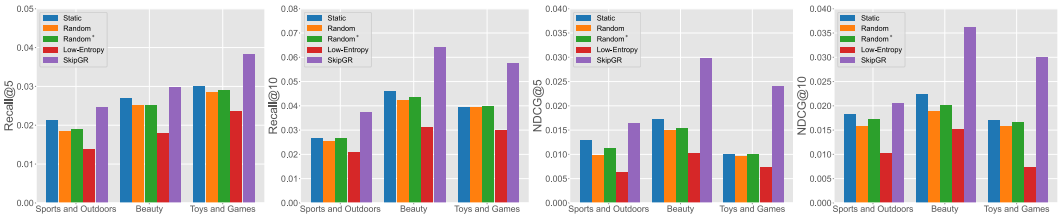

Figure 4: Comparison of skipping conditions. Random* here is using a total dynamic random strategy in each training step. Experiments are conducted on Amazon Product Reviews datasets.

4.4 Ablation Experiments and Analysis

We perform extensive ablation experiments to illustrate the characteristics of SkipGR. If not stated, the experiments are conducted on Amazon Product Reviews by default.

Evaluating the components of SkipGR. We conduct an ablation study by evaluating SkipGR, as shown in Table 4. Specifically, we run GRID, GRID with Qwen2.5-VL embedding, GRID with token skip, GRID with SimVQ, GRID with sequence skip, and GRID with SkipGR on Amazon Product Reviews datasets. The results from these experiments provide insights into the contribution of token and sequence skip to the overall performance of SkipGR.

Table 4: Ablation study on Amazon Product Reviews datasets.

Methods	**Sports and Outdoors**				**Beauty**				**Toys and Games**			
	Recall @5	NDCG @5	Recall @10	NDCG @10	Recall @5	NDCG @5	Recall @10	NDCG @10	Recall @5	NDCG @5	Recall @10	NDCG @10
GRID	0.0239	0.0154	0.0363	0.0194	0.0429	0.0285	0.0639	0.0353	0.0373	0.0237	0.0565	0.0300
/ + Qwen2.5-VL-7B	0.0241	0.0156	0.0365	0.0195	0.0429	0.0285	0.0639	0.0353	0.0373	0.0237	0.0565	0.0300
/ + SimVQ	0.0242	0.0154	0.0361	0.0198	0.0425	0.0286	0.0631	0.0351	0.0370	0.0232	0.0569	0.0293
/ + token skip	0.0241	0.0158	0.0359	0.0196	0.0431	0.0280	0.0635	0.0358	0.0374	0.0239	0.0568	0.0296
/ + sequence skip	0.0235	0.0155	0.0360	0.0194	0.0421	0.0284	0.0630	0.0352	0.0355	0.0232	0.0559	0.0298
/ + SkipGR	0.0245	0.0163	0.0371	0.0205	0.0441	0.0298	0.0639	0.0362	0.0382	0.0240	0.0573	0.0300

From Table 4, we find that each component of SkipGR contributes to improving the overall framework in complementary ways. Incorporating Qwen2.5-VL embeddings yields slight but consistent gains, indicating that high-quality multimodal representations provide a stronger foundation for quantization and entropy estimation. SimVQ achieves results on par with the baseline, showing that the choice of quantization method influences the balance between information fidelity and efficiency. Token-level skipping proves particularly effective, consistently improving Recall and NDCG by removing redundant low-entropy tokens while retaining the most informative ones. In contrast, sequence-level skipping alone may slightly degrade performance, as discarding entire user histories risks losing contextual information. When combined, however, the full SkipGR framework achieves the strongest results across all domains, with notable improvements such as over 4.5% gains on Recall@5 and NDCG@5 in the *Beauty* dataset. These results demonstrate that fine-grained entropy-guided skipping is essential for efficiency, while the integration of both token- and sequence-level strategies ensures robustness and superior performance.

Exploring where to skip. In default setting, tokens or sequences with $H_t^{(l)} < \bar{H}_\tau^{(l)}$ ($H_\tau^{seq} < \bar{H}_\tau^{seq}$) are skiped. Other possible rules are to skip with $H_t^{(l)} > \bar{H}_\tau^{(l)}$ ($H_\tau^{seq} > \bar{H}_\tau^{seq}$), static skip, random skip, or dynamic random skip. We compare the performance of these skipping strategies in Figure 4. Compared to training on the original dataset, the other strategies are far inferior to SkipGR. However, they can provide higher skipping probabilities.

5 Conclusion

In this work, we presented SkipGR, a lightweight yet powerful framework for accelerating generative recommendation models by adaptively skipping uninformative tokens and sequences. By introducing semantic token entropy as a principled measure of informativeness, SkipGR dynamically prunes redundant training signals while retaining high-utility content. This design not only reduces computational overhead but also improves convergence speed, achieving lossless—or even superior—recommendation quality.

Through extensive experiments on public benchmarks and a large-scale industrial dataset, we demonstrated that SkipGR consistently enhances efficiency while maintaining or surpassing baseline performance. These findings suggest that entropy-guided skipping provides a general and scalable strategy for building the next generation of efficient, industrial-scale generative recommendation systems.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Weijie Bian, Kailun Wu, Lejian Ren, Qi Pi, Yujing Zhang, Can Xiao, Xiang-Rong Sheng, Yong-Nan Zhu, Zhangming Chan, Na Mou, Xinchen Luo, Shiming Xiang, Guorui Zhou, Xiaoqiang Zhu, and Hongbo Deng. Can: Feature co-action for click-through rate prediction. *arXiv preprint arXiv:2011.05625*, 2021.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*, 2021.

Yue Cao, XiaoJiang Zhou, Jiaqi Feng, Peihao Huang, Yao Xiao, Dayao Chen, and Sheng Chen. Sampling is all you need on modeling long-term user behaviors for ctr prediction. *arXiv preprint arXiv:2205.10249*, 2022.

Zheng Chai, Qin Ren, Xijun Xiao, Huizhi Yang, Bo Han, Sijun Zhang, Di Chen, Hui Lu, Wenlin Zhao, Lele Yu, Xionghang Xie, Shiru Ren, Xiang Sun, Yaocheng Tan, Peng Xu, Yuchao Zheng, and Di Wu. Longer: Scaling up long sequence modeling in industrial recommenders. *arXiv preprint arXiv:2505.04421*, 2025.

Jianxin Chang, Chenbin Zhang, Zhiyi Fu, Xiaoxue Zang, Lin Guan, Jing Lu, Yiqun Hui, Dewei Leng, Yanan Niu, Yang Song, and Kun Gai. Twin: Two-stage interest network for lifelong user behavior modeling in ctr prediction at kuaishou. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3785–3794, 2023.

Qiwei Chen, Changhua Pei, Shanshan Lv, Chao Li, Junfeng Ge, and Wenwu Ou. End-to-end user behavior retrieval in click-through ratepredictionmodel. *arXiv preprint arXiv:2108.04468*, 2021.

Alexander Dallmann, Daniel Zoller, and Andreas Hotho. A case study on sampling strategies for evaluating neural sequential item recommendation models. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, pp. 505–514, 2021.

Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *arXiv preprint arXiv:2502.18965*, 2025.

Chao Feng, Yanze Zhang, and Chenghao Zhang. Generate the browsing process for short-video recommendation. *arXiv preprint arXiv:2504.08771*, 2025.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, pp. 299–315, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Ruining He and Julian McAuley. 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 '16, pp. 507–517, 2016.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2016.

Clark Mingxuan Ju, Liam Collins, Leonardo Neves, Bhuvesh Kumar, Louis Yufeng Wang, Tong Zhao, and Neil Shah. Generative recommendation with semantic ids: A practitioner's handbook. *arXiv preprint arXiv:2507.22224*, 2025.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. *arXiv preprint arXiv:1808.09781*, 2018.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, yelong shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Not all tokens are what you need for pretraining. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ziyang Liu. Amazon-book and movielens-1m, 2025. URL https://dx.doi.org/10.21227/ztyg-gp55.

Aleksandr V. Petrov, Craig Macdonald, and Nicola Tonellotto. Efficient recommendation with millions of items by dynamic pruning of sub-item embeddings. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2102–2111, 2025.

Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, Zhaopan Xu, Daquan Zhou, Lei Shang, Baigui Sun, Xuansong Xie, and Yang You. Infobatch: Lossless training speed up by unbiased dynamic data pruning. *arXiv preprint arXiv:2303.04947*, 2023.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, Maciej Kula, Ed Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10299–10315, 2023.

Zihua Si, Lin Guan, Zhongxiang Sun, Xiaoxue Zang, Jing Lu, Yiqun Hui, Xingchao Cao, Zeyu Yang, Yichen Zheng, Dewei Leng, Kai Zheng, Chenbin Zhang, Yanan Niu, Yang Song, and Kun Gai. Twin v2: Scaling ultra-long user behavior sequence modeling for enhanced ctr prediction at kuaishou. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 4890–4897. ACM, 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*, CIKM '19, pp. 1441–1450, 2019.

Yizheng Sun, Yanze Xin, Hao Li, Jingyuan Sun, Chenghua Lin, and Riza Batista-Navarro. Lvpruning: An effective yet simple language-guided vision token pruning approach for multi-modal large language models. *arXiv preprint arXiv:2501.13652*, 2025.

Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pp. 565–573, 2018.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Shaobo Wang, Xiangqi Jin, Ziming Wang, Jize Wang, Jiajun Zhang, Kaixin Li, Zichen Wen, Zhong Li, Conghui He, Xuming Hu, and Linfeng Zhang. Data whisperer: Efficient data selection for task-specific llm fine-tuning via few-shot in-context learning. *arXiv preprint arXiv:2505.12212*, 2025a.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025b.

Shuliang Wang, Jiabao Zhu, Yi Wang, Chen Ma, Xin Zhao, Yansen Zhang, Ziqiang Yuan, and Sijie Ruan. Hierarchical gating network for cross-domain sequential recommendation. *ACM Trans. Inf. Syst.*, 43(4), 2025c.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*, 2025.

Jiaqi Zhai, Zhaojie Gong, Yueming Wang, Xiao Sun, Zheng Yan, Fu Li, and Xing Liu. Revisiting neural retrieval on accelerators. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 5520–5531, 2023.

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. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 58484–58509, 2024.

Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S. Sheng, Jiajie Xu, Deqing Wang, Guanfeng Liu, and Xiaofang Zhou. Feature-level deeper self-attention network for sequential recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, pp. 4320–4326, 2019.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. Adapting large language models by integrating collaborative semantics for recommendation. *arXiv preprint arXiv:2311.09049*, 2024.

Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1059–1068, 2018.

Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. Deep interest evolution network for click-through rate prediction. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2019.

Yongxin Zhu, Bocheng Li, Yifei Xin, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. *arXiv preprint arXiv:2411.02038*, 2024.