

EchoRec: Multi-Item Prediction-Empowered Generative Recommendation via Cycle-Consistent Preference Alignment

Haokai Ma*
haokai.ma1997@gmail.com
National University of Singapore
Singapore, Singapore

Aoqi Hu*
huaoqi800@gmail.com
National University of Singapore
Singapore, Singapore

Yueao Xing
sd513@bupt.edu.cn
Beijing University of Posts and
Telecommunications
Beijing, China

Ruobing Xie
xrbsnowing@163.com
Tencent
Beijing, China

Yonghui Yang
yyh.hfut@gmail.com
National University of Singapore
Singapore, Singapore

Teng Tu
teng.tu@u.nus.edu
National University of Singapore
Singapore, Singapore

Lei Meng
lmeng@sdu.edu.cn
Shandong University
Jinan, China

Tat-Seng Chua
dcscts@nus.edu.sg
National University of Singapore
Singapore, Singapore

Abstract

Generative recommendation autoregressively generates the semantic IDs of the target item, unifying preference modeling and index retrieval within the shared token space. Recent attempts have introduced Multi-Token Prediction (MTP) into this field, yet they primarily inherit its efficiency merit, leaving its potential as dense supervision unexplored. Unlocking this potential hinges on whether future behaviors qualify as informative supervision. Our analysis reveals that future behaviors carry a semantic echo of the current one far above that of random pairs, which nevertheless decays along horizons under intent transitions, making them informative yet order-dependent signals. Motivated by this, we propose **EchoRec**, which empowers MTP with cycle-consistent holistic preference alignment across multi-horizon for generative recommendation. It comprises two synergistic modules. Horizon-aware Preference Generation (HPG) sequentially chains lightweight auxiliary branches upon the base recommender, where each branch conditions on its predecessor to respect preference evolution. Verifiable Holistic-Preference Alignment (VHA) further consolidates them into the holistic preference and echoes it back through cycle-consistent projectors to suppress spurious alignment, with theoretical guarantees that exclude the rank-collapse form of spurious alignment under an invertible transport, enabling the holistic preference to be retained in the decoding representation. All auxiliary components serve as disposable scaffolding discarded at inference, introducing negligible

online serving overhead. Extensive experiments on three datasets demonstrate the superiority of our **EchoRec**, together with its naturally acquired multi-item generation ability. Our code and datasets will be available upon acceptance.

CCS Concepts

• Information systems → Recommender systems.

Keywords

Multi-Token Prediction, Generative Recommendation

ACM Reference Format:

Haokai Ma, Aoqi Hu, Yueao Xing, Ruobing Xie, Yonghui Yang, Teng Tu, Lei Meng, and Tat-Seng Chua. 2027. EchoRec: Multi-Item Prediction-Empowered Generative Recommendation via Cycle-Consistent Preference Alignment. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Generative recommendation (GR) has recently emerged as a promising paradigm that tokenizes each item into a tuple of discrete semantic IDs (SIDs) and generates the target item token by token [20, 24, 26]. Instead of ranking candidates over the entire corpus, GR unifies preference modeling and index retrieval within a single process, where the shared token space anchors semantically related items to overlapping SIDs [7, 8, 13]. It enables recommenders to function well in sparse scenarios by exploiting external semantics, attracting increasing attention from e-commerce and streaming services.

Drilling down into this topic, existing works generally build the GR pipeline around two key components: item tokenization and autoregressive generation [9]. Regarding the former, TIGER [27] utilizes Residual Quantized Variational Autoencoder (RQ-VAE) to generate SIDs for items, DiscRec [17] injects position embeddings

*Equal Contribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, Woodstock, NY

© 2027 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/XXXXXXX.XXXXXXX>

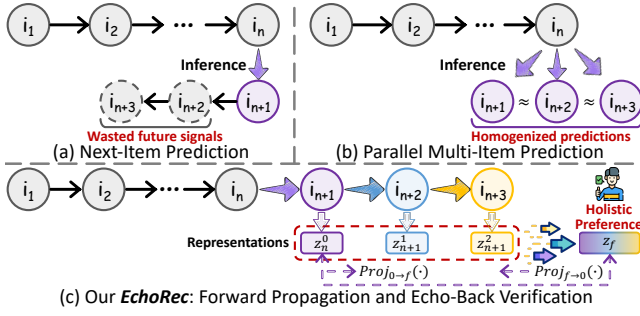


Figure 1: Conceptual comparison of future-behavior modeling paradigms. (a) Next-item prediction underutilizes future-behavior supervision. (b) Parallel multi-item prediction risks homogenizing the horizons. (c) EchoRec propagates supervision across horizons and cycle-consistently aligns the decoding representation with holistic preference.

into token sequences for item-aware alignment and Pctx [35] incorporates user contexts during SID generation for personalization. For the latter, EAGER [30] and LETTER [29] equip the Transformer-based generator with modality-aware contrastive objectives to integrate behavioral and semantic information, while ETEGRec [18] devises an alternating optimization technique to foster sequence-item alignment and preference-semantic alignment simultaneously.

Inspired by the remarkable capabilities of Multi-Token Prediction (MTP) in advanced Large Language Models (LLMs) [6, 16], several pioneering efforts have introduced it into recommendation. For instance, GReF [15] generates multiple ordered future items with MTP for reranking efficiency, RPG [12] applies MTP to expand the token number of items to boost expressiveness, and GPR [32] and EGA-V2 [33] adopt MTP to capture concurrent interests and generate advertisement content in industrial scenarios. Notably, beyond the decoding efficiency that MTP is best known for, it also serves as dense supervision that densifies training signals and encourages pre-planned representations [1, 34]. However, existing efforts primarily transplant the former merit, leaving the latter potential as future-behavior supervision largely unexplored.

However, translating such potential into GR is non-trivial, as existing attempts leave two fundamental challenges unresolved: (1) **How to exploit future behaviors as sequentially dependent supervision across horizons?** Unlike consecutive tokens tightly coupled by grammar in LLMs, consecutive behaviors are driven by evolving user intents, where i_{n+2} depends on the intent transition triggered by i_{n+1} rather than on the historical sequence alone. Existing MTP-empowered recommenders either repurpose MTP as an efficiency-oriented device [12, 32, 33] or decode multiple items simultaneously from the same context for reranking efficiency [15], degenerating multi-horizon prediction into homogenized next-item prediction tasks that disregard preference evolution. (2) **How to guarantee that the incorporated future signals are actually internalized?** Even equipped with future-aware supervision, such signals shape the decoding representation only indirectly through auxiliary objectives. A natural remedy is to align this representation toward the aggregated holistic target via projectors, yet such one-way alignment is inherently *unverifiable*, as the projector can absorb the discrepancy between the two subspaces to satisfy it

while leaving the representation oblivious to the holistic preference, a failure mode that we term **spurious alignment**. To our knowledge, no existing GR study offers a mechanism to rule it out.

To address these challenges, we propose **EchoRec**, a framework that empowers MTP with cycle-consistent holistic preference alignment across multi-horizon for GR. Drawing an analogy to the acoustic echo, **EchoRec** regards preference modeling as an echoing process: user preferences are propagated forward across future horizons as sequentially dependent supervision, and echoed back for verifying their internalization. Specifically, we devise the *Horizon-aware Preference Generation* (HPG) module to extend the next-item objective to multiple future horizons through sequential MTP branches, which conditions on its predecessor’s representation to capture intent transition between adjacent horizons, thereby delivering the structural-level supervision that respects preference evolution. To further guarantee actual internalization, we design a *Verifiable Holistic-Preference Alignment* (VHA) module to realize the echo-back verification, which consolidates all branches into the holistic preference and pulls the decoding representation toward it. Considering that such one-way alignment remains unverifiable, VHA further imposes a bidirectional cycle-consistency constraint on the projector pair, which shapes the round-trip geometry of the transport (*cf.* Lemma 1) and excludes the rank-collapse form of spurious alignment (*cf.* Theorem 2), suppressing the spurious alignment that plagues the one-way objective. Notably, EchoRec remains lightweight on both sides, as the auxiliary branches of HPG reuse the shared output head and decoding graph at negligible training overhead, while the projectors of VHA serve as disposable scaffolding discarded at inference, preserving the decoding efficiency during online serving.

We conduct extensive experiments on three real-world datasets to demonstrate the effectiveness of our **EchoRec**. We further conduct analyses on multi-horizon prediction, component ablation, and efficiency to verify its robustness, universality and the superiority of each design. The main contributions are as follows:

- We propose **EchoRec**, which empowers MTP with cycle-consistent holistic preference alignment to capture the horizon-aware preference. To our knowledge, it is among the pioneering efforts to unlock MTP as sequentially dependent future-behavior supervision in generative recommendation.
- We devise HPG to structurally inject sequentially dependent supervision across horizons and VHA to verifiably internalize the holistic preference via cycle-consistent alignment, with theoretical guarantees that characterize the round-trip geometry and exclude the rank-collapse form of spurious alignment.
- We conduct extensive experiments on three datasets, where **EchoRec** consistently improves diverse SOTA backbones while exhibiting the multi-item generation ability.

2 Related Works

Generative Recommendation. Inspired by the remarkable achievement of generative models, several pioneering recommenders assign each item a set of discrete but shared tokens, referred to as *semantic ID* (SID), where tokenization discretizes item

semantics into SIDs and generation predicts the target item conditioned on them. Specifically, VQ-Rec [10] and SeeDRec [25] respectively utilize transferable “codes” and “sememe” to represent items from the content information. TIGER [27] discretizes item representations into SIDs via RQ-VAE and predicts the target item with the Transformer-based generator. EAGER [30] and LETTER [29] design contrastive alignment objectives to integrate behavioral and semantic information. DiscRec [17] incorporates item-level position embeddings to enable the item-aware alignment and disentangle semantic and collaborative signals, while Pctx [35] devises a context-aware tokenizer for personalized SID generation. To further unify the two stages into a cohesive objective, ETEGRec [18] fosters mutual enhancement between sequence-item and preference-semantic levels via alternating optimization. However, these efforts uniformly follow the next-item objective, leaving future behaviors beyond the immediate target untouched, which is the focus of our work.

Multi-Token Prediction for Recommendation. Multi-Token Prediction (MTP) extends the next-token objective of large language models to predict multiple tokens within a single step, alleviating their inefficiency and limited long-range dependency modeling [5, 19, 28, 34]. Inspired by its superiority, several attempts have integrated MTP into recommendation. GReF [15] generates multiple ordered future items simultaneously for real-time reranking efficiency. RPG [12] allocates longer SIDs to each item and decodes them in parallel to enhance item expressiveness. GPR [32] maps both advertisements and organic content into a shared multi-level SID space and integrates MTP to capture concurrent interests, while EGA-V2 [33] chains MTP branches across modalities to jointly generate advertisement sequences and creative content. Nevertheless, these works predominantly inherit the efficiency merit of MTP, whereas our work unlocks its potential as sequentially dependent future-behavior supervision with verifiable internalization.

3 Preliminary

3.1 Multi-Token Prediction

As the name implies, Multi-token prediction (MTP) extends the conventional next-token prediction objective by predicting multiple future tokens within a single step, thereby providing longer-range dependencies during training. Given the token sequence $x_{1:t}$, the generic MTP objective can be written as $\mathcal{L}_{\text{MTP}} = -\sum_t \log P_\theta(x_{t+1:t+n} | x_{1:t})$, where n denotes the prediction horizon. A representative implementation [6] adopts a shared Transformer trunk $f_s(\cdot)$ with n independent prediction heads $f_h^{(i)}(\cdot)$ to predict these future tokens in parallel from the same contextual representation. Let $z_{1:t} = f_s(x_{1:t})$ represent the hidden state produced by $f_s(\cdot)$, the i -th future token can be predicted as $P_\theta(x_{t+i} | x_{1:t}) = \text{softmax}(f_u(f_h^{(i)}(z_{1:t})))$, where $f_u(\cdot)$ denotes the shared unembedding matrix. Instead of predicting all future tokens independently from the same state, DeepSeek-V3 [16] sequentially chains MTP modules to preserve the causal dependencies among future predictions and sequentially predict additional n tokens. To be specific, the k -th MTP module combines the representation from its predecessor with the embedding of the $(i+k)$ -th token to generate the input $\mathbf{h}_i^k$ of the additional Transformer block:

$$\mathbf{h}_i^k = M_k [\text{RMSNorm}(\mathbf{h}_i^{k-1}); \text{RMSNorm}(\text{Emb}(i_{i+k}))], \quad (1)$$

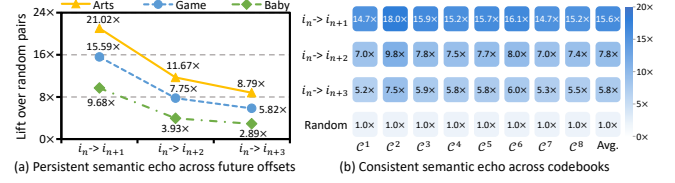


Figure 2: Analysis of the usefulness of future information for next-item prediction. (a) shows the SID overlap lift of (i_n, i_{n+h}) over random pairs across three datasets, which remains well above random even at larger offsets. (b) presents the codebook-wise lift on Game, where the semantic echo stays consistent across all codebooks.

where $\text{RMSNorm}(\cdot)$, $M_k \in \mathbb{R}^{d \times 2d}$, and $\text{Emb}(\cdot)$ denote the Root Mean Square Layer Normalization, the projection matrix and the shared embedding layer, respectively. The Transformer block $\text{TRM}_k(\cdot)$ then produces the output representation $\mathbf{h}_{i+k}^k$, upon which the shared output head predicts the $(i+k+1)$ -th token. In this work, we adopt this sequential design to preserve coherent causal dependencies across multiple prediction depths, providing denser supervision over future tokens while keeping the auxiliary objective compatible with the generation process.

3.2 Motivation Analysis on Future Behaviors

Exploiting future behaviors as auxiliary supervision is appealing yet non-trivial, which hinges on two fundamental questions: (1) *whether future behaviors convey informative signals for next-item prediction, rather than mere noise induced by drifting intents* and (2) *whether multiple future behaviors can be faithfully forecast in parallel from the shared context, rather than collapsing into homogenized variants?* We conduct two motivation analyses to respectively answer them.

3.2.1 Semantic Echo across Future Behaviors. To answer the first question, we measure the SID overlap of each observed pair (i_n, i_{n+h}) with offset $h \in \{1, 2, 3\}$ on three datasets, and illustrate the *lift* in Figure 2, which is defined as the ratio of this overlap to that of randomly sampled item pairs. The *lift* of 1.0x indicates that the future item i_{n+h} is semantically indistinguishable from a random one, whereas a higher *lift* indicates the future behaviors are semantically anchored to the current interaction i_n . As shown in Figure 2 (a), **future behaviors i_{n+h} persistently echo the related semantics of the current interaction i_n across all offsets**: the *lift* reaches up to 21.02x at $h=1$ and remains 2.89x–8.79x even at $h=3$ across three datasets. This indicates that interactions beyond i_{n+1} still echo the semantics of the current one rather than degenerating into noise, thus potentially constituting informative supervision for preference modeling. Besides, the consistent decay along h reveals that this echo is progressively reshaped by intent transitions, suggesting that different horizons carry order-dependent signals and should be modeled as sequentially dependent targets rather than being treated equally. Moreover, Figure 2 (b) shows that **such semantic echo remains consistent across all codebooks**, where all eight codebooks exhibit comparable *lift* on Game across all offsets, indicating that such semantic echo is not driven by certain dominant codebooks but consistently spans all semantic subspaces.

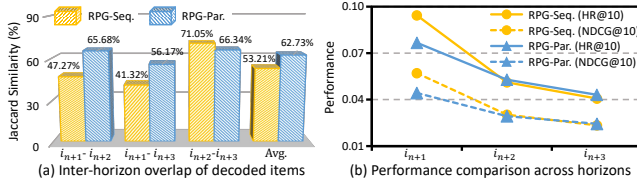


Figure 3: Analysis of the homogenization pitfall of parallel MTP structures on Game. (a) reports the inter-horizon Jaccard similarity of the top-10 decoded items, where *RPG-Par.* exhibits higher overlap than *RPG-Seq.* on average. (b) compares the performance across horizons, where *RPG-Par.* merely matches the rollout without horizon-specific supervision yet underperforms at the immediate horizon i_{n+1} .

3.2.2 Homogenization Pitfall of Parallel MTP Structures. To answer the second question, we instantiate two multi-horizon variants upon the same base recommender [12], where *RPG-Par.* decodes three future items from the shared context following the parallel MTP structure [6], while *RPG-Seq.* autoregressively rolls out the prediction as the input sequence to decode the next horizon. Figure 3 (a) reports the Jaccard similarity between the top-10 predictions decoded at different horizons on Game, showing that **parallel MTP homogenizes multi-horizon predictions**: *RPG-Par.* yields higher average inter-horizon overlap than *RPG-Seq.* (62.73% vs. 53.21%), especially on the pairs involving the immediate horizon i_{n+1} (65.68% vs. 47.27% for $i_{n+1}-i_{n+2}$ pair and 56.17% vs. 41.32% for $i_{n+1}-i_{n+3}$ pair). This indicates that predicting all horizons from the same context blurs the intermediate intent transition, whereas even the naive rollout that conditions each horizon on its predecessor differentiates the predictions across horizons better. Figure 3 (b) further compares the performance of these two variants at each horizon, revealing that **such homogenization fails to translate into performance gains**. Despite the dedicated supervision on each horizon, *RPG-Par.* merely matches the unsupervised rollout of *RPG-Seq.* at i_{n+2} and i_{n+3} , yet underperforms at the immediate horizon i_{n+1} that directly determines the recommendation quality. This clearly indicates that the future signals injected by parallel MTP are not effectively absorbed into horizon-specific predictions, and instead interfere with the immediate objective.

4 Methodology

4.1 Problem Formulation

We define the *behavioral sequence* of user $u \in \mathcal{U}$ as $S_u = \{i_1^u, i_2^u, \dots, i_n^u\}$, where $i_t^u \in \mathcal{V}$ denotes the t -th behavior of S_u and n denotes the sequence length. We adopt optimized product quantization [4, 10] to represent item i_t as a k -token tuple $\phi(i_t) = \{c_{(t,1)}, c_{(t,2)}, \dots, c_{(t,k)}\}$, where $\phi(i_t)$ denotes the SID of item i_t and $c_{(t,j)} \in C^j$ denotes the code selected from the j -th codebook C^j . Given S_u and the SIDs of all involved items, our goal is to predict the k -token tuple $\phi(i_{n+1}^u) = \{c_{(n+1,1)}, c_{(n+1,2)}, \dots, c_{(n+1,k)}\}$ of the target item i_{n+1}^1 that user u will interact with.²

¹For brevity, we omit the superscript u in i_t^u in the following sections.

²The key notations are summarized in Appendix.

4.2 Overall Structure

The structure of **EchoRec** is illustrated in Figure 4, which regards preference modeling as an acoustic echoing process. Specifically, the Horizon-aware Preference Generation module (*cf.* Section 4.4) propagates user preferences forward across future horizons, extending the next-item objective through sequential MTP branches where each branch conditions on its predecessor to capture the intent transition. The Verifiable Holistic-Preference Alignment module (*cf.* Section 4.5) then echoes them back for verification, consolidating the multi-horizon representations into a holistic preference and pulling the decoding representation toward it under a cycle-consistency constraint that suppresses spurious alignment.

4.3 Base Recommender

We adopt RPG [12] as the base recommender to tokenize each item using a long semantic ID and generate its tokens in parallel. Following Section 4.1, each item i_t is encoded as $\phi(i_t) = (c_{(t,1)}, \dots, c_{(t,k)})$, with $c_{(t,j)} \in C^j$. To avoid inflating the context length under such long SIDs, we aggregate all token embeddings of item i_t into a single representation $v_{i_t} = \text{Agg}(e_{(t,1)}, \dots, e_{(t,k)})$ via pooling. Given a user sequence S_u , the corresponding item representations $H = [v_1, v_2, \dots, v_n]$ are encoded by multiple Transformer blocks into the sequence representation s at position n . The shared output head $g(\cdot)$ then splits its output into k codebook-specific vectors, and the token-level distribution of j -th codebook is computed as

$$P_j = \text{softmax}(E^j \cdot [g(s)]_j / \tau), \quad j = 1, \dots, k, \quad (2)$$

where $[g(s)]_j$ denotes the j -th split vector, E^j is the embedding matrix of C^j , and τ is the temperature. Assuming the tokens of target item i_t to be conditionally independent given s , we optimize the token-level MTP loss $\mathcal{L}_{\text{RPG}} = -\sum_{j=1}^k \log P_j(c_{(t,j)})$ to score the target SID against all token combinations across codebooks [12]. For inference, the graph-constrained decoding samples an initial beam of valid SIDs from $\{P_j\}_{j=1}^k$, propagates it over a similarity graph $\mathcal{G}$, scores each candidate by $\sum_{j=1}^k \log P_j(c_j)$, and retains the top- B candidates per step, yielding the top- K recommendations without exhaustive scoring over the entire item pool.

4.4 Horizon-aware Preference Generation

Although existing MTP-based recommenders improve decoding efficiency or item expressiveness, most of them still follow the next-item prediction paradigm, optimizing only toward the next item i_{n+1} given a sequence S_u . However, user behaviors are inherently sequential and evolving, where the later interactions beyond i_{n+1} may reveal dynamic intent transitions and future-oriented dependencies that cannot be characterized by the single target. Meanwhile, the attempt that generates multiple future items [15] decodes them in parallel from the same context, overlooking these inter-item dependencies and collapsing these items into homogenized candidates. To exploit the intent transition between adjacent horizons under the sequentially dependent supervision, we introduce a *Horizon-aware Preference Generation* (HPG) module, which structurally extends the next-item objective to multiple future horizons.

Specifically, given a behavioral sequence S_u of user u , we follow the base recommender in Section 4.3 to convert each item i into its aggregated representation v_i . The original branch (MTP-0) encodes

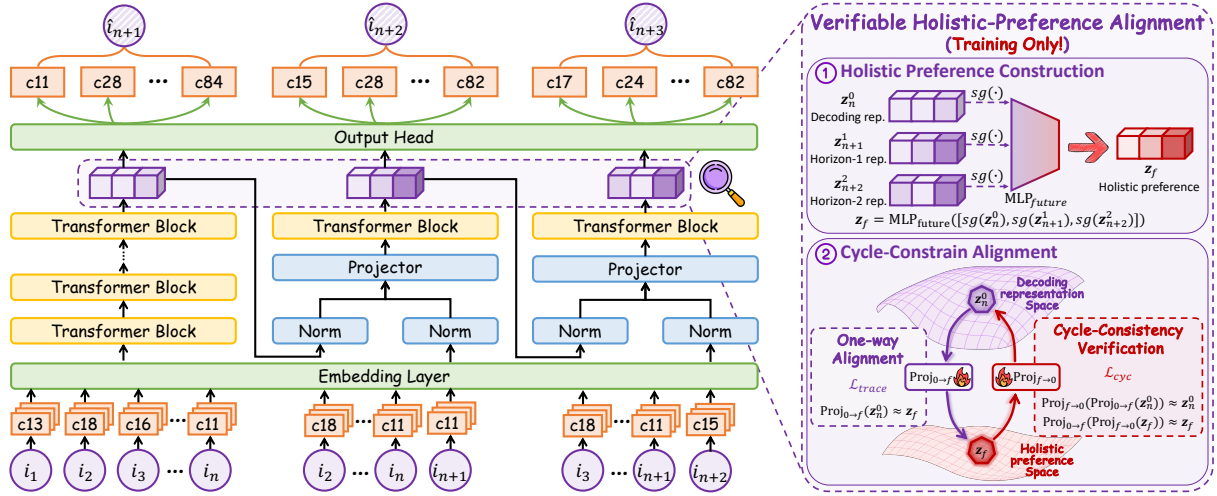


Figure 4: Overall structure of our EchoRec.

the item-level sequence $H^0 = [v_1, v_2, \dots, v_n]$ as follows:

$$Z^0 = \text{MTRM}^0(H^0) = [z_1^0, z_2^0, \dots, z_n^0], \quad (3)$$

where $\text{MTRM}^0(\cdot)$ denotes the multi-layer Transformer backbone within MTP-0. It then feeds the last representation z_n^0 into the shared output head $g(\cdot)$ to produce the token-level distributions $\{P_j^0\}_{j=1}^k$ as in Eq. (2) for graph-constrained decoding, where the superscript b in P_j^b indexes the branch hereafter. MTP-0 thus reproduces the base recommender [12], upon which HPG appends lightweight auxiliary branches as additional supervision pathways while leaving the original architecture and decoding pipeline intact, as detailed below.

Beyond the original branch, HPG appends auxiliary branches to predict the subsequent items beyond i_{n+1} , which not only enriches the supervision for capturing preference transitions but also enables multi-item generation when required. Since both auxiliary branches share the same architecture, we take MTP-1 as an example. Following the sequential MTP design [16], MTP-1 generates the SID of item i_{n+2} by additionally conditioning on the intermediate item i_{n+1} , extending the preference generation from a single-step objective to horizon-aware preference modeling while capturing intent transition between adjacent horizons. It fuses the predicted representations $[z_1^0, \dots, z_n^0]$ from MTP-0 with the embeddings of the shifted sequence $S_u^1 = [i_2, \dots, i_{n+1}]$ as follows:

$$H^1 = \text{Proj}^1([\text{RMSNorm}([z_1^0, \dots, z_n^0]); \text{RMSNorm}([v_2, \dots, v_{n+1}])]), \quad (4)$$

where $\text{Proj}^1(\cdot)$ aligns the concatenated representation into the hidden space of MTP-1. The one-layer Transformer block $\text{TRM}^1(\cdot)$ then encodes the projected representation H^1 into the future-aware representations $Z^1 = \text{TRM}^1(H^1) = [z_2^1, \dots, z_{n+1}^1]$, whose last entry z_{n+1}^1 is fed into the shared output head $g(\cdot)$ to generate $\{P_j^1\}_{j=1}^k$ for decoding i_{n+2} over the similarity graph $\mathcal{G}$. Likewise, MTP-2 conditions on Z^1 and the further shifted sequence $S_u^2 = [i_3, \dots, i_{n+2}]$ to predict i_{n+3} , thereby chaining the branches into sequentially dependent horizons. MTP-1 and MTP-2 predict over the same item pool and SID space as MTP-0, which allows them to reuse the token embeddings $\{E^j\}_{j=1}^k$, the output head $g(\cdot)$, and the decoding graph $\mathcal{G}$ of MTP-0, while only introducing $\text{Proj}^b(\cdot)$ and $\text{TRM}^b(\cdot)$ per branch, keeping the auxiliary branches lightweight.

4.5 Verifiable Holistic-Preference Alignment

Although HPG injects future-behavior supervision at the *structural* level, such supervision is distributed across several branch-wise objectives and thus reaches the original branch only indirectly through the auxiliary pathways. Consequently, the decoding representation z_n^0 that directly generates the recommendation is never explicitly required to internalize the consolidated multi-horizon preference, leaving the future signals under-exploited at the *representation* level. To bridge this gap, we devise a *Verifiable Holistic-Preference Alignment* (VHA) module to consolidate them into a holistic target and explicitly pull z_n^0 toward it in *representation* space. Specifically, we aggregate the predicted representations of the three branches in Section 4.4 into a holistic preference representation z_f as:

$$z_f = \text{MLP}_{\text{future}}([sg(z_n^0), sg(z_{n+1}^1), sg(z_{n+2}^2)]), \quad (5)$$

where $\text{MLP}_{\text{future}}(\cdot)$ is a two-layer projection and $sg(\cdot)$ denotes the stop-gradient operation, which prevents the alignment objective from back-propagating into the other branches to reshape their supervised signals. Here, the future representations z_{n+1}^1 and z_{n+2}^2 convey the evolving intents across horizons, while the present one z_n^0 serves as an anchor against over-committing to distant and noisier future intents. We then align the decoding representation toward this holistic preference via a representation-level objective:

$$\mathcal{L}_{\text{trace}} = 1 - \cos(\text{Proj}_{0 \rightarrow f}(z_n^0), z_f), \quad (6)$$

where $\text{Proj}_{0 \rightarrow f}(\cdot)$ transports z_n^0 into holistic preference space. Note that the stop-gradient in Eq. (5) acts only on the construction of z_f , whereas the z_n^0 in Eq. (6) remains differentiable. Hence, the gradient of $\mathcal{L}_{\text{trace}}$ on the z_f side merely calibrates $\text{MLP}_{\text{future}}(\cdot)$, while that on the z_n^0 side directly drives the decoding representation to encode the holistic preference. Meanwhile, the detached anchor prevents z_f from trivially collapsing onto z_n^0 to shortcut the objective.

However, **such one-way alignment is inherently unverifiable**, as minimizing Eq. (6) is consistent with two opposite outcomes, a faithful one where z_n^0 truly internalizes the holistic preference, and a spurious one where $\text{Proj}_{0 \rightarrow f}(\cdot)$ overfits a shortcut mapping from an uninformed z_n^0 to z_f . This alignment objective alone cannot distinguish between them. Since z_n^0 and z_f reside in

different subspaces, this *spurious alignment* is generic rather than incidental: the forward projector $\text{Proj}_{0 \rightarrow f}(\cdot)$ can simply absorb the discrepancy between the two subspaces within its own mapping, satisfying Eq. (6) while leaving z_n^0 oblivious to the holistic preference (cf. Theorem 2). To suppress such spurious alignment, we require the transport between z_n^0 and z_f to be invertible, which prevents $\text{Proj}_{0 \rightarrow f}(\cdot)$ from discarding any direction of the holistic space, so that a low alignment loss reflects the genuine encoding of the holistic preference. To this end, we introduce a backward projector $\text{Proj}_{f \rightarrow 0}(\cdot)$ and impose a bidirectional cycle-consistency constraint that enforces both round trips to return to their origins:

$$\mathcal{L}_{\text{cyc}} = \left[1 - \cos(\text{Proj}_{f \rightarrow 0}(\text{Proj}_{0 \rightarrow f}(\text{sg}(z_n^0))), \text{sg}(z_n^0)) \right] + \left[1 - \cos(\text{Proj}_{0 \rightarrow f}(\text{Proj}_{f \rightarrow 0}(\text{sg}(z_f))), \text{sg}(z_f)) \right]. \quad (7)$$

Since both round trips are detached, $\mathcal{L}_{\text{cyc}}$ reshapes only the projector pair rather than the representations, jointly driving them toward a mutually invertible transport between the two subspaces [2, 31]. We abbreviate $z \triangleq z_n^0$, $P \triangleq \text{Proj}_{0 \rightarrow f}$ and $Q \triangleq \text{Proj}_{f \rightarrow 0}$, which map between $\mathbb{R}^D$ and $\mathbb{R}^r$ with $r < D$. Let $\Lambda = \text{diag}(\rho_1, \dots, \rho_r)$ collect the canonical correlations between z and z_f , let $\bar{Q}$ be the scale-calibrated representative of Q , and let $\text{cond}(\cdot)$ denote the ratio of the largest to the smallest singular value. We formalize this intuition as follows (cf. Appendix for the full proofs):

Lemma 1 (Cycle-consistency characterizes the round-trip geometry). *Let P^* and Q^* minimize the matching risks $\mathbb{E}\|Pz - z_f\|^2$ and $\mathbb{E}\|Qz_f - z\|^2$. Then P^*Q^* has eigenvalues $\{\rho_i^2\}_{i=1}^r$, where ρ_i denotes the i -th canonical correlation between z and z_f , and the scale-fixed counterpart of Eq. (7) satisfies $\tilde{\mathcal{L}}_{\text{cyc}} = 2\sum_{i=1}^r (1 - \rho_i^2)^2$ and $\|P^*Q^* - I\|_{\text{op}} \leq \sqrt{\tilde{\mathcal{L}}_{\text{cyc}}/2}$ in the canonical reduction of Appendix. A small $\tilde{\mathcal{L}}_{\text{cyc}}$ corresponds to a well-conditioned round trip between the subspaces.*

The overall alignment objective of VHA is formulated as $\mathcal{L}_{\text{VHA}} = \lambda_{\text{trace}} \mathcal{L}_{\text{trace}} + \lambda_{\text{cyc}} \mathcal{L}_{\text{cyc}}$, where λ_{cyc} balances the alignment objective against the invertibility that makes it verifiable. Building on Lemma 1, we further characterize how this invertible transport constrains the minimizers of $\mathcal{L}_{\text{trace}}$:

Theorem 2 (Cycle-consistency drives projected alignment). *At the pair of Lemma 1, one-way alignment may settle on a direction other than the back-projected preference whenever $\Lambda^{-1}z_f$ and Λz_f are not positively collinear. Let $A \triangleq P\bar{Q}$ be invertible, and set $\Pi \triangleq \bar{Q}A^{-1}P$, $z^\circ \triangleq \bar{Q}A^{-1}z_f$, and $z^* \triangleq \bar{Q}z_f$. Then every minimizer $z_{\min}$ of $\mathcal{L}_{\text{trace}}$ obeys $\Pi z_{\min} = \lambda z^\circ$ for some $\lambda > 0$, with $\|z^\circ - z^*\| \leq \text{cond}(\bar{Q})\|A^{-1} - I\|_{\text{op}}\|z^*\|$.*

That is, once the transport is invertible, the projected component of the alignment minimizer is pinned to a single direction, and its deviation from the back-projected preference shrinks as the round trip approaches the identity, which excludes the rank-collapse form of spurious alignment.

4.6 Optimization Objective

Following the base recommender in Section 4.3, each branch $b \in \{0, 1, 2\}$ of HPG utilizes its predicted representation z_{n+b}^b to generate the semantic ID of the target i_{n+b+1} with a token-level MTP loss:

$$\mathcal{L}_b = - \sum_{j=1}^k \log \text{softmax} \left(E^j \cdot [g(z_{n+b}^b)]_j / \tau \right)_{c_{(n+b+1,j)}}, \quad (8)$$

Table 1: Statistics of three real-world datasets.

Dataset	#User	#Item	#Interactions	Sparsity
Game	22,970	30,642	310,554	99.96%
Baby	40,501	49,783	485,607	99.98%
Arts	20,965	76,985	280,931	99.98%

where E^j is the embedding matrix of the j -th codebook, and $c_{(n+b+1,j)}$ denotes the j -th SID of the target item i_{n+b+1} within MTP- b . Hence, the overall objective of HPG is $\mathcal{L}_{\text{HPG}} = \mathcal{L}_0 + \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2$, where λ_1 and λ_2 control the contribution of each future-behavior supervision term. Together with $\mathcal{L}_{\text{VHA}}$ in Section 4.5, the final optimization objective of **EchoRec** is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{HPG}} + \mathcal{L}_{\text{VHA}}, \quad (9)$$

where the former propagates preferences across horizons and the latter verifiably echoes them back to the decoding representation.³

Further Discussion. Due to space limitations, we provide the computational complexity analysis, comparison with MTP-related recommenders, and discussion of EchoRec’s threefold novelty in Appendix.

5 Experiments

5.1 Experimental Setups

5.1.1 Dataset. To evaluate the effectiveness of **EchoRec**, we construct three datasets based on the Amazon Reviews ’23⁴. Following the traditional GR protocol [12, 21–23], we chronologically organize each user’s historical interactions and extend the leave-one-out evaluation to the specific multi-item prediction task by reserving the last three interacted items as test set, using the fourth-to-last item as the validation set, and treating all the remaining interactions as training set. The statistical details are summarized in Table 1, and the complete construction process is elaborated in Appendix.

5.1.2 Baselines. To evaluate the effectiveness of **EchoRec**, we compare it with six representative generative recommenders, including **EAGER** [30], **TIGER** [27], **LETTER** [29], **ETEGRec** [18], **SETRec** [14], and **RPG** [12]. Due to space limitations, detailed descriptions of them are provided in Appendix.

5.1.3 Evaluation Metrics. We choose Hit Rate (HR@K) and Normalized Discounted Cumulative Gain (NDCG@K) for evaluation, where $k \in \{10, 20\}$. Following conventional GR works [3, 11, 17], we adopt the all-ranking paradigm to avoid selection bias.

5.1.4 Implementation Details. All methods are trained and evaluated under identical environments on each dataset, sharing the same data splits, pre-computed SID representations, and evaluation protocol, where EchoRec retains its three-horizon training objective within the training split yet sees no extra supervision beyond the baselines, and is evaluated solely with MTP-0. Each method is independently run three times with different seeds. Due to space limitations, more details are provided in Appendix.

³Due to the space limitation, the detailed training and inference pseudocode and training strategies are provided in Appendix.

⁴<https://amazon-reviews-2023.github.io/>

Table 2: Performance comparison of EchoRec against six baselines across three datasets. * denotes statistically significant improvements of EchoRec over its corresponding backbone ($p < 0.01$ with paired t-tests).

Algorithm	Game				Baby				Arts			
	H@10	N@10	H@20	N@20	H@10	N@10	H@20	N@20	H@10	N@10	H@20	N@20
EAGER [30]	0.0266	0.0132	0.0376	0.0160	0.0172	0.0094	0.0241	0.0111	0.0024	0.0016	0.0030	0.0018
TIGER [27]	0.0831	0.0443	0.1275	0.0554	0.0220	0.0110	0.0352	0.0143	0.0583	0.0364	0.0791	0.0417
LETTER [29]	0.0361	0.0179	0.0544	0.0225	0.0210	0.0108	0.0352	0.0144	0.0177	0.0095	0.0262	0.0116
ETEGRec [18]	0.0631	0.0328	0.0968	0.0413	0.0206	0.0106	0.0330	0.0137	0.0384	0.0214	0.0557	0.0258
RPG [12]	0.0931	0.0564	0.1249	0.0644	0.0362	0.0243	0.0452	0.0266	0.0625	0.0421	0.0802	0.0466
EchoRec (RPG)	0.1035*	0.0628*	0.1382*	0.0715*	0.0426*	0.0295*	0.0528*	0.0321*	0.0716*	0.0498*	0.0911*	0.0547*
<i>Rel. Imp.</i>	+11.17%	+11.35%	+10.65%	+11.02%	+17.68%	+21.40%	+16.81%	+20.68%	+14.56%	+18.29%	+13.59%	+17.38%
SETRec [14]	0.1157	0.0610	0.1722	0.0752	0.0322	0.0164	0.0511	0.0212	0.0746	0.0412	0.1086	0.0498
EchoRec (SETRec)	0.1231*	0.0657*	0.1839*	0.0810*	0.0373*	0.0193*	0.0587*	0.0247*	0.0838*	0.0468*	0.1197*	0.0559*
<i>Rel. Imp.</i>	+6.40%	+7.70%	+6.79%	+7.71%	+15.84%	+17.68%	+14.87%	+16.51%	+12.33%	+13.59%	+10.22%	+12.25%

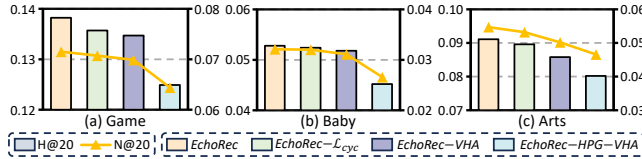


Figure 5: Results of EchoRec and its ablation versions on Game, Baby, and Arts. All components are effective.

5.2 Performance Comparison

To demonstrate the effectiveness of **EchoRec**, we implement it with two diverse backbones and compare them with six GR baselines across three datasets in Table 2, from which we observe that: (1) **EchoRec** consistently and significantly outperforms its base recommender on all datasets and metrics, implying that the sequentially dependent supervision of HPG and the verifiable alignment of VHA effectively convert the informative signals within future behaviors into recommendation gains. (2) The relative improvements on NDCG generally exceed those on HR, indicating that **EchoRec** not only retrieves more correlated target items but ranks them higher, as the holistic preference internalized by VHA refines the decoding representation toward fine-grained preference discrimination. (3) Among the baselines, the methods with stronger tokenization or decoding mechanism (e.g., SETRec and RPG) generally surpass the earlier generative recommenders, yet their relative strengths fluctuate across datasets, reflecting the sensitivity of the single next-item objective to data characteristics. In contrast, **EchoRec** delivers stable gains over its backbone on every dataset, showing the value of exploiting future behaviors beyond this single target.

5.3 Ablation Study

To verify the contribution of each design within our **EchoRec**, we quantify the contribution of each module in Section 5.3.1, and further justify the superiority of the tailored HPG in Section 5.3.2.

5.3.1 Component Ablation. To verify the contribution of each module, we compare **EchoRec** with three ablated variants in Figure 5, where *EchoRec- $\mathcal{L}_{cyc}$* keeps the one-way alignment without the cycle-consistency constraint, *EchoRec-VHA* retains only the multi-horizon supervision, and *EchoRec-HPG-VHA* equals the base recommender. We observe that: (1) *EchoRec-VHA* consistently surpasses *EchoRec-HPG-VHA* across all datasets, indicating that the

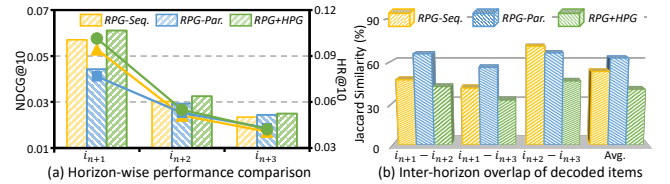


Figure 6: Multi-horizon prediction performance of RPG+HPG against RPG-Seq. and RPG-Par. on Game, in terms of (a) the horizon-wise performance and (b) the inter-horizon overlap of the decoded items at three future horizons.

sequentially dependent supervision of HPG indeed converts future behaviors into informative signals. (2) Appending the one-way alignment brings further gains, indicating that the **consolidated z_f carries holistic information beyond the decoding representation rather than trivially replicating z_n^0** , since aligning toward such a shortcut target would yield no improvement. This empirically rules out the target-side shortcut concern in Eq. (5). (3) **EchoRec** achieves the best performance on all datasets, indicating that the cycle-consistency constraint further suppresses the spurious alignment that the one-way objective alone cannot exclude.

5.3.2 Multi-Horizon Structure Ablation. To justify the sequential chaining design of HPG, we follow the protocol of Section 3.2.2 to replace it with two alternative multi-horizon structures upon the same base recommender, where *RPG-Par.* decodes all horizons from the shared context and *RPG-Seq.* rolls out the prediction without horizon-specific supervision. As shown in Figure 6, the horizon-wise performance measures whether the future horizons are accurately predicted without compromising the immediate one, while the inter-horizon overlap examines whether the predictions across horizons are genuinely differentiated rather than homogenized replicas. We observe that: (1): Sequential chaining within HPG improves the future horizons without compromising the immediate one. *RPG+HPG* achieves the best performance at i_{n+1} and i_{n+2} on both metrics, and remains competitive at i_{n+3} . In contrast, *RPG-Seq.* lags behind across future horizons for lacking any supervision, while *RPG-Par.* suffers from interference at i_{n+1} , a degradation that never emerges on *RPG+HPG* whose future signals are absorbed without sacrificing the immediate objective. (2): Horizon-specific supervision differentiates the predictions across horizons. *RPG+HPG*

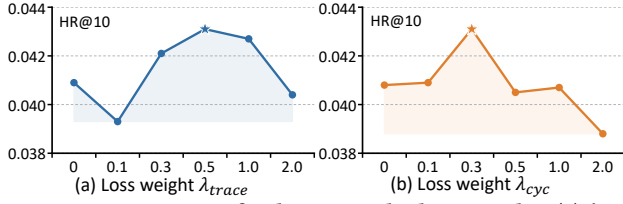


Figure 7: Sensitivity of EchoRec to the loss weights (a) λ_{trace} and (b) λ_{cyc} on Baby, where $\star$ marks the adopted value.

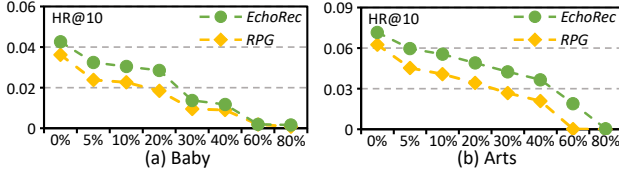


Figure 8: Robustness analysis of EchoRec against RPG on Baby and Arts, where the proportion η of historical interactions is randomly replaced with sampled items.

exhibits the lowest average inter-horizon overlap among all variants, even below the unsupervised rollout of *RPG-Seq.*, indicating that each auxiliary branch captures the intent transition of its own horizon rather than replicating the immediate prediction, which resolves the homogenization pitfall identified in Section 3.2.2. These results justify the sequential chaining of HPG, which additionally endows **EchoRec** with the multi-item generation ability to forecast a coherent yet non-redundant trajectory within a single model.

5.4 Robustness Analysis

5.4.1 Robustness Analysis on Important Parameters. We examine the sensitivity of **EchoRec** to the two loss weights of VHA on Baby, varying λ_{trace} and λ_{cyc} within $\{0, 0.1, 0.3, 0.5, 1.0, 2.0\}$ in Figure 7. Both parameters exhibit a unimodal trend that peaks at moderate values ($\lambda_{trace} = 0.5$ and $\lambda_{cyc} = 0.3$): a smaller weight under-exploits the one-way alignment or the cycle constraint, while an over-large one distracts the optimization from the primary next-item objective. **EchoRec** outperforms its base recommender under all configurations, implying its robustness against the hyper-parameter selection rather than relying on delicate tuning.

5.4.2 Robustness Analysis on Noisy Interactions. To evaluate the robustness of **EchoRec** against noisy behavioral sequences, we randomly replace a proportion $\eta \in \{0\%, 5\%, 10\%, 20\%, 30\%, 40\%, 60\%, 80\%\}$ of items in each historical sequence with items uniformly sampled from the corpus, where $\eta = 0\%$ corresponds to the original input. As reported in Figure 8, **EchoRec** consistently outperforms RPG under mild to moderate noise, and the relative improvement first enlarges with the noise ratio (e.g., from 17.68% at $\eta = 0\%$ to 54.89% at $\eta = 20\%$ on HR@10 of Baby) before narrowing under heavier corruption. We attribute this to the holistic preference internalized by VHA, where the decoding representation is aligned to the consolidated multi-horizon preference rather than individual interactions, sporadically corrupted items are less likely to distort prediction. Only under extreme corruption ($\eta \geq 60\%$) do both methods degrade to comparable and unusable performance, which is expected as the behavioral evidence itself becomes unreliable.

Table 3: Efficiency comparison results of **EchoRec** (RPG) against SETRec and RPG, where “#Tra.” and “#Inf.” denote the average training and inference latency per batch, while “#Mem.” denotes the GPU memory usage during inference.

Dataset	Algorithm	#Tra. (ms)	#Inf. (ms)	#Mem. (MiB)
Game	SETRec	125.61	41.57	1464.07
	RPG	30.11	87.44	181.04
	EchoRec (RPG)	61.87	85.46	181.04
Baby	SETRec	139.81	46.46	2231.32
	RPG	11.25	64.13	171.54
	EchoRec (RPG)	32.81	64.21	171.54
Arts	SETRec	141.68	48.33	2964.74
	RPG	12.50	67.32	215.40
	EchoRec (RPG)	32.97	68.96	215.40

5.4.3 Robustness Analysis on Diverse Backbones. Since **EchoRec** only appends the auxiliary branches and optimization objectives upon the base recommender while leaving the recommender itself unmodified, it can be seamlessly plugged into diverse backbones. To verify such robustness, we further instantiate it upon SETRec [14], whose order-agnostic set identifiers differ fundamentally from RPG. As reported in Table 2, **EchoRec** (SETRec) consistently and significantly outperforms SETRec across all datasets and metrics, exhibiting the same trend as **EchoRec** (RPG). Considering that these two backbones diverge in both tokenization and decoding strategy, these consistent gains indicate that the multi-horizon supervision of HPG and the verifiable alignment of VHA jointly capture the backbone-agnostic value of future behaviors, rendering **EchoRec** a plug-and-play framework for generative recommenders.

5.5 In-depth Analysis

5.5.1 Computational Complexity Analysis. To verify the complexity analysis in Appendix, we compare the training latency, inference latency per batch, and the inference GPU memory usage of **EchoRec** against SETRec and RPG in Table 3. We notice that: (1) **The training overhead of EchoRec remains affordable.** Compared with RPG, **EchoRec** increases the training latency (e.g., from 11.25ms to 32.81ms on Baby) due to the auxiliary branches and alignment objectives, whereas such training-only cost is well amortized as a one-off investment. (2) **EchoRec introduces negligible inference overhead.** The inference latency of **EchoRec** stays nearly identical to RPG on all datasets, empirically confirming that **all auxiliary components serve as disposable scaffolding discarded at inference.** (3) **The memory footprint of EchoRec stays lightweight at serving.** Its inference memory usage remains nearly identical to RPG and substantially below SETRec, as all auxiliary components are discarded and only MTP-0 is activated during serving. Overall, **EchoRec delivers its gains at marginal training cost while fully preserving the online serving efficiency.**

5.5.2 Case Study. To illustrate how **EchoRec** benefits from the future-behavior supervision, we show two representative cases from Baby. Figure 9 (a) presents a next-item recommendation case, where RPG recommends generally relevant items such as booster seats and baby wipes, whereas **EchoRec** accurately ranks “Diapers (Size 5, 96 Count)” at the top, with a same-series “Diapers (Size 5, 25 Count)” immediately after. We attribute this to the holistic

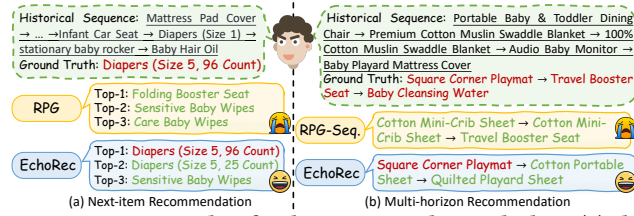


Figure 9: Case study of EchoRec on Baby, including (a) the next-item recommendation and (b) the multi-horizon recommendation, where the ground-truth items are marked in red.

preference internalized by VHA, which captures the temporal progression of user demands rather than merely matching the semantics of recent interactions. Figure 9 (b) illustrates a multi-horizon recommendation case, where **EchoRec** correctly generates “Square Corner Playmat” at the immediate horizon and continues with diverse yet semantically coherent items, whereas the predictions of *RPG-Seq* remain close to each other across the first two horizons. These cases qualitatively imply that the multi-horizon supervision of HPG and the verifiable alignment of VHA enable **EchoRec** to forecast a reasonable trajectory of future behaviors.

6 Conclusion

In this paper, we investigate how to unlock the dense-supervision potential of MTP to enhance generative recommendation, supported by the empirical observation that future behaviors semantically echo the current interaction and thus qualify as informative supervision. Considering that this echo decays under intent transitions and thereby demands sequentially dependent modeling across horizons, we propose **EchoRec**, which regards preference modeling as an acoustic echoing process. Specifically, HPG converts future behaviors into sequentially dependent supervision through chained MTP branches where each horizon conditions on its predecessor, while VHA consolidates them into the holistic preference and echoes it back through a cycle-consistent projector pair, with theoretical guarantees that exclude the rank-collapse form of spurious alignment. Extensive experiments on three datasets demonstrate the effectiveness of our **EchoRec**, together with its naturally acquired multi-item generation ability. In the future, we plan to explore adaptive horizon selection that calibrates the supervision strength against the decaying semantic echo, and to investigate a mesh-style prediction architecture that captures the fine-grained correlation evolution across both tokens and items.

References

- [1] Gregor Bachmann and Vaishnavh Nagarajan. 2024. The pitfalls of next-token prediction. In *Proceedings of the 41st International Conference on Machine Learning*. 2296–2318.
- [2] Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. 2025. Cycle consistency as reward: Learning image-text alignment without human preferences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22934–22946.
- [3] Junchen Fu, Xuri Ge, Alexandros Karatzoglou, Ioannis Arapakis, Suzan Verberne, Joemon M Jose, and Zhaochun Ren. 2026. Differentiable Semantic ID for Generative Recommendation. *arXiv preprint arXiv:2601.19711* (2026).
- [4] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. 2013. Optimized product quantization. *IEEE transactions on pattern analysis and machine intelligence* 36, 4 (2013), 744–755.
- [5] Anastasios Gerontopoulos, Spyros Gidaris, and Nikos Komodakis. 2025. Multi-token prediction needs registers. *arXiv preprint arXiv:2505.10518* (2025).
- [6] Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. 2024. Better & faster large language models via multi-token

- prediction. *arXiv preprint arXiv:2404.19737* (2024).
- [7] Ruidong Han, Bin Yin, Shangyu Chen, He Jiang, Fei Jiang, Xiang Li, Chi Ma, Mincong Huang, Xiaoguang Li, Chunzhen Jing, et al. 2025. Mtgr: Industrial-scale generative recommendation framework in meituan. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 5731–5738.
- [8] Yingzhi He, Yan Sun, Junfei Tan, Yuxin Chen, Xiaoyu Kong, Chunxu Shen, Xiang Wang, An Zhang, and Tat-Seng Chua. 2026. Reasoning over semantic ids enhances generative recommendation. *arXiv preprint arXiv:2603.23183* (2026).
- [9] Min Hou, Le Wu, Yuxin Liao, Yonghui Yang, Zhen Zhang, Yu Wang, Changlong Zheng, Han Wu, and Richang Hong. 2026. A survey on generative recommendation: Data, model, and tasks. *AI Open* (2026).
- [10] Yupeng Hou, Zhankui He, Julian McAuley, and Wayne Xin Zhao. 2023. Learning vector-quantized item representation for transferable sequential recommenders. In *Proceedings of the ACM Web Conference 2023*. 1162–1171.
- [11] Yupeng Hou, Haven Kim, Clark Mingxuan Ju, Eduardo Escoto, Neil Shah, and Julian McAuley. 2026. Expressiveness Limits of Autoregressive Semantic ID Generation in Generative Recommendation. *arXiv preprint arXiv:2605.06331* (2026).
- [12] Yupeng Hou, Jiacheng Li, Ashley Shin, Jinsung Jeon, Abhishek Santhanam, Wei Shao, Kaveh Hassani, Ning Yao, and Julian McAuley. 2025. Generating long semantic ids in parallel for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 956–966.
- [13] Peiyu Hu, Wayne Lu, and Jia Wang. 2026. From ids to semantics: A generative framework for cross-domain recommendation with adaptive semantic tokenization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 14874–14882.
- [14] Xinyu Lin, Haihan Shi, Wenjie Wang, Fuli Feng, Qifan Wang, See-Kiong Ng, and Tat-Seng Chua. 2025. Order-agnostic identifier for large language model-based generative recommendation. In *Proceedings of the 48th international ACM SIGIR conference on research and development in information retrieval*. 1923–1933.
- [15] Zhijie Lin, Zhuofeng Li, Chenglei Dai, Wentian Bao, Shuai Lin, Enyun Yu, Haoxiang Zhang, and Liang Zhao. 2025. GReF: A Unified Generative Framework for Efficient Reranking via Ordered Multi-token Prediction. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 5879–5887.
- [16] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).
- [17] Chang Liu, Yimeng Bai, Xiaoyan Zhao, Yang Zhang, Fuli Feng, and Wenge Rong. 2025. Discrec: Disentangled semantic-collaborative modeling for generative recommendation. *arXiv preprint arXiv:2506.15576* (2025).
- [18] Enze Liu, Bowen Zheng, Cheng Ling, Lantao Hu, Han Li, and Wayne Xin Zhao. 2025. Generative recommender with end-to-end learnable item tokenization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 729–739.
- [19] Xiaohao Liu, Xiaobo Xia, Weixiang Zhao, Manyi Zhang, Xianzhi Yu, Xiu Su, Shuo Yang, See-Kiong Ng, and Tat-Seng Chua. 2026. L-mtp: Leap multi-token prediction beyond adjacent context for large language models. *Advances in Neural Information Processing Systems* 38 (2026), 102569–102600.
- [20] Zhanyu Liu, Shiyao Wang, Xingmei Wang, Rongzhou Zhang, Jiaxin Deng, Honghui Bao, Jinghao Zhang, Wuchao Li, Pengfei Zheng, Xiangyu Wu, et al. 2025. Onerec-think: In-text reasoning for generative recommendation. *arXiv preprint arXiv:2510.11639* (2025).
- [21] Haokai Ma, Yunshan Ma, Ruobing Xie, Lei Meng, Jialie Shen, Xingwu Sun, Zhanhui Kang, and Tat-Seng Chua. 2026. Large Language Model Empowered Recommendation Meets All-Domain Continual Pre-Training. *IEEE Transactions on Knowledge and Data Engineering* (2026), 1–14. doi:10.1109/TKDE.2026.3717059
- [22] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Zhanhui Kang. 2024. Plug-In Diffusion Model for Sequential Recommendation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, February 20–27, 2024, Vancouver, Canada*. 8886–8894. doi:10.1609/AAAI.V38I8.28736
- [23] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Jie Zhou. 2024. Triple Sequence Learning for Cross-domain Recommendation. *ACM Trans. Inf. Syst.* 42, 4 (2024). doi:10.1145/3638351
- [24] Haokai Ma, Ruobing Xie, Lei Meng, Fuli Feng, Xiaoyu Du, Xingwu Sun, Zhanhui Kang, and Xiangxu Meng. 2026. Negative Sampling in Recommendation: A Survey and Future Directions. *ACM Trans. Inf. Syst.* (2026). doi:10.1145/3793855
- [25] Haokai Ma, Ruobing Xie, Lei Meng, Yimeng Yang, Xingwu Sun, and Zhanhui Kang. 2024. SeedRec: Sememe-based Diffusion for Sequential Recommendation. In *IJCAI*. 2270–2278.
- [26] Gustavo Penha, Edoardo D’Amico, Marco De Nadai, Enrico Palumbo, Alexandre Tamborrino, Ali Vardasbi, Max Lefarov, Shawn Lin, Timothy Heath, Francesco Fabbri, et al. 2025. Semantic ids for joint generative search and recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 1296–1301.
- [27] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonathan Samost, et al.

2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems* 36 (2023), 10299–10315.
- [28] Mohammad Samragh, Arnav Kundu, David Harrison, Kumari Nishu, Devang Naik, Minsik Cho, and Mehrdad Farajtabar. 2025. Your llm knows the future: Uncovering its multi-token prediction potential. *arXiv preprint arXiv:2507.11851* (2025).
- [29] Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2024. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2400–2409.
- [30] Ye Wang, Jiahao Xun, Minjie Hong, Jieming Zhu, Tao Jin, Wang Lin, Haoyuan Li, Linjun Li, Yan Xia, Zhou Zhao, et al. 2024. Eager: Two-stream generative recommender with behavior-semantic collaboration. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3245–3254.
- [31] Hongye Xu and Bartosz Krawczyk. 2026. Two-Way Is Better Than One: Bidirectional Alignment with Cycle Consistency for Exemplar-Free Class-Incremental Learning. *arXiv preprint arXiv:2606.05675* (2026).
- [32] Jun Zhang, Yi Li, Yue Liu, Changping Wang, Yuan Wang, Yuling Xiong, Xun Liu, Haiyang Wu, Qian Li, Enming Zhang, et al. 2025. GPR: Towards a Generative Pre-trained One-Model Paradigm for Large-Scale Advertising Recommendation. *arXiv preprint arXiv:2511.10138* (2025).
- [33] Zuowu Zheng, Ze Wang, Fan Yang, Jiangke Fan, Teng Zhang, Yongkang Wang, and Xingxing Wang. 2025. Ega-v2: An end-to-end generative framework for industrial advertising. *arXiv preprint arXiv:2505.17549* (2025).
- [34] Qimin Zhong, Hao Liao, Siwei Wang, Mingyang Zhou, Xiaoqun Wu, Rui Mao, and Wei Chen. 2025. Understanding and Enhancing the Planning Capability of Language Models via Multi-Token Prediction. *arXiv preprint arXiv:2509.23186* (2025).
- [35] Qiyong Zhong, Jiajie Su, Yunshan Ma, Julian McAuley, and Yupeng Hou. 2025. Pctx: Tokenizing Personalized Context for Generative Recommendation. *arXiv preprint arXiv:2510.21276* (2025).