Generate What You Can Trust: Content Credibility in Generative Recommenders

Zhuo Cai[1] **Guanghao Wu**[1] **Shoujin Wang**[1*]
Peilin Zhou[2] **Victor W. Chu**[1]
[1] Data Science Institute, University of Technology Sydney, Australia
[2] New York University Abu Dhabi, United Arab Emirates
{Zhuo.Cai, Guanghao.Wu}@student.uts.edu.au, Shoujin.Wang@uts.edu.au,
zhoupalin@gmail.com, WingYan.Chu@uts.edu.au

Abstract

Generative recommendation (GR) represents items with semantic IDs (i.e., discrete token sequences) and generates target item tokens as recommendations. Despite its promising results, existing methods predominantly optimize for accuracy while neglecting the credibility of the recommendations they generate. This oversight inevitably exposes users to uncredible content (e.g., fake news) with serious societal consequences, including user distrust, reputation harm to platforms, and broader social instability. To address this critical yet underexplored challenge, we propose CreGR, the first credible GR model that jointly tackles content credibility across the two core stages of GR: *tokenization* and *generation*. In the tokenization stage, we design a new **credibility-aware tokenizer** that explicitly encourages the model to learn discriminative tokens respectively for credible and uncredible items, thereby disentangling credibility signals at the token level. Building on this, in the generation stage, we propose a novel **accuracy-preserving and credibility-oriented generator** grounded in discrete diffusion. Specifically, we introduce an asymmetric masking probability reduction strategy that selectively diminishes the contribution of tokens associated with uncredible content to the generation process, while leaving tokens encoding user preference signals unaffected so as to preserve recommendation accuracy. Experiments on three real-world datasets demonstrate the effectiveness of CreGR. The implementation code and preprocessed datasets of our method are provided in https://github.com/iamZhuoCai/CreGR.

1 Introduction

Generative recommendation (GR) has emerged as a transformative paradigm for recommender systems in today's generative AI era, reformulating the recommendation problem from conventional embedding-based matching to token generation (Rajput et al., 2023; Wang et al., 2024c; Hou et al., 2025a). Compared to traditional methods, GR offers superior generalization that subsumes retrieval and ranking into a single generation process. Despite its remarkable success and rapid adoption by major platforms serving hundreds of millions of users (Deng et al., 2025; Singh et al., 2024; He et al., 2026; Penha et al., 2025), existing GR methods are predominantly optimized for recommendation accuracy while neglecting a critical yet underexplored societal dimension: the **content credibility** of the generated recommendations. As illustrated in Figure 1, existing GR methods may accurately capture user preference yet recommend uncredible content (e.g., fake news) to users. This oversight leads to serious consequences, including eroded user trust, reputational risks to platforms, and even social instability (Cai et al., 2026a; Wang et al., 2024b, 2022b).

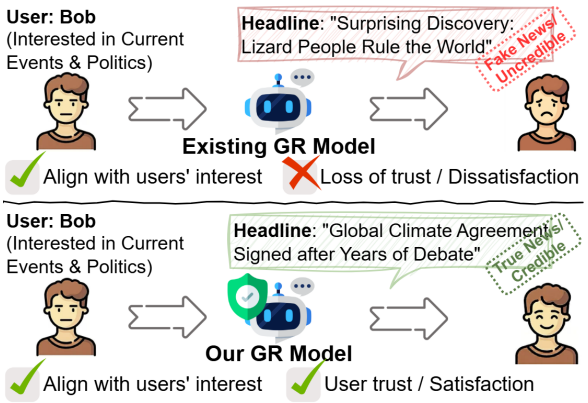

Figure 1: Comparison between existing GR models and our proposed credible GR model. Existing GR models generally neglect the content credibility of recommendations. Our proposed model is capable to improve recommendation credibility while preserving high recommendation accuracy.

There are some recent efforts on developing credible content recommendation methods (Wang et al.,

*Corresponding author: Shoujin.Wang@uts.edu.au

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 48198–48218
October 24-29, 2026 ©2026 Association for Computational Linguistics

2022c, 2024a; Cai et al., 2026a,b; Sallami et al., 2023). For example, some works (Wang et al., 2022c, 2024a) disentangle news content into veracity embeddings and event embeddings, and utilize only event embeddings to compute user-item matching scores, thereby mitigating the influence of uncredible content. Cai et al. (Cai et al., 2026a) employ a diffusion model to improve recommendation credibility. However, these methods all operate within the *embedding matching* paradigm, where recommendations are made by computing similarity scores between user and item embeddings over a large candidate pool. As such, they are inherently incompatible with the emerging GR paradigm, which forgoes embedding matching entirely and instead directly generates the semantic IDs (SIDs) of target items. In this paper, we focus on how to improve the credibility of GR. However, building a credible GR model is non-trivial, primarily due to the following two challenges (CHs):

CH1: How to learn discriminative semantic IDs for credible and uncredible content items in the tokenization stage? Existing methods tokenize items by encoding their metadata (e.g., titles and descriptions) through pre-trained language models and subsequently quantizing the resulting embeddings into discrete tokens (i.e., SIDs) via vector quantization (e.g., RQ-VAE (Lee et al., 2022)). However, these models are pre-trained with objectives such as masked language modeling or next-token prediction, which capture topical and syntactic patterns but are inherently agnostic to content credibility (Chuang et al., 2024; Lin et al., 2022). As a result, credible and uncredible content that share similar topics are mapped to nearby regions in the embedding space. This credibility entanglement causes topically similar but factually divergent items to be assigned similar/nearby tokens, making them difficult to discriminate during the subsequent generation process and potentially leading to the recommendation of uncredible content.

CH2: How to balance recommendation accuracy and credibility in the generation stage? A straightforward solution to mitigate uncredible content is to directly remove uncredible items from training data and retrain the generation model. However, this naive removal strategy inevitably degrades recommendation accuracy, as the valuable information (e.g., topics) in the uncredible items that reflect users' preference is also removed. For instance, a user who frequently engages with uncredible health news articles still reveals a strong interest in wellness topics, which should be leveraged to capture the user's preference and support accurate generation. The empirical evidence is provided in Appendix C. Therefore, there is a fundamental trade-off between recommendation accuracy and credibility in the generation stage.

To address these challenges, in this paper, we propose a novel credible generative recommendation model, `CreGR`. Specifically, to overcome **CH1**, we design a novel credibility-aware tokenizer that introduces credibility regularization to enforce intra-class compactness and inter-class separation between credible and uncredible items. This compels the codebook to disentangle credibility signals at the discrete token level and thereby eliminating entanglement between factually divergent items. To address **CH2**, we propose an accuracy-preserving and credibility-oriented generator grounded in discrete diffusion. Exploiting the fact that discrete diffusion training is conditioned on masked tokens, we design an asymmetric masking probability reduction strategy that selectively reduces the masking probabilities of tokens associated with uncredible content identified in the tokenization stage, while leaving all other tokens encoding valuable user preference signals at their original masking probabilities. We instantiate this strategy through two variants: a *hard* weighting strategy that performs absolute erasure of uncredible token contributions, and a *soft* weighting strategy that applies distance-aware down-weighting based on proximity to cluster centroids. These designs gracefully suppress uncredible generation while preserving the topical preference signals, achieving a balance between recommendation accuracy and credibility.

Overall, our contributions are summarized as:

- We propose a novel credible generative content recommendation model, `CreGR`, which is the first GR model that considers the credibility of the generated recommendations.
- We design a credibility-aware tokenizer and a novel accuracy-preserving and credibility-oriented generator, which work collaboratively to improve the recommendation credibility while preserving high recommendation accuracy.
- Extensive experiments on three real-world datasets demonstrate the effectiveness of our proposed model in generating credible and accurate content recommendations.

2 Problem Formulation

2.1 Credible Content Recommendation

Following prior work (Wang et al., 2022c, 2024a; Cai et al., 2026a), we formulate credible content recommendation as the task of reducing user exposure to uncredible items, such as fake news and misinformation, which degrade user experience and pose adverse societal impacts. Credible content recommendation follows the task of sequential recommendation (Wang et al., 2021, 2019a,b, 2018; Cai et al., 2025). A recommendation model is deemed more credible when its generated recommendation lists contain fewer such items.

Besides, *this task requires developing a recommendation model that inherently ensures the credibility, rather than relying on post-hoc filtering* (i.e., removing identified uncredible items from final recommendation lists). Post-hoc filtering treats the symptom rather than the cause: the underlying model has already learned latent patterns favoring uncredible content during training, and such filtering remains vulnerable to adversarial manipulation where uncredible content is crafted to evade detection (Zhou and Zafarani, 2020).

Definition 1 ***Content credibility.*** *Content credibility is defined as the trustworthiness of the information conveyed by an item. An item is regarded as uncredible if it contains false, misleading, or inaccurate information (e.g., fake news, health misinformation), and credible otherwise.*

2.2 Generative Recommendation

Given a user's interaction history $\mathcal{H} = \{v_1, v_2, \ldots, v_{n-1}\}$, GR formulates the recommendation task as a token generation problem to generate the tokens of the target item v_n that users may be interested in. Each item v_i is represented as a sequence of discrete tokens (i.e., SIDs) $[c_{i,1}, c_{i,2}, \ldots, c_{i,L}]$ using vector quantization techniques (e.g., RQ-VAE), where L denotes the number of quantization levels. Afterwards, the user's interaction sequence can be represented as:

$$\mathcal{S}_{\mathcal{H}} = [c_{1,1}, \cdots, c_{1,L}; \cdots; c_{n-1,1}, \cdots, c_{n-1,L}]. \tag{1}$$

GR aims to learn a probabilistic model $\mathcal{P}_\theta$ that generates the tokens of the target item $v_n = [c_{n,1}, \cdots, c_{n,L}]$ conditioned on $\mathcal{S}_{\mathcal{H}}$: $\mathcal{P}_\theta(v_n|\mathcal{S}_{\mathcal{H}})$.

Existing GR models can be categorized into autoregressive (AR) models that generate the tokens of target items from left-to-right, and non-autoregressive (NAR) models that generate tokens in parallel via discrete diffusion. Our proposed method is based on the second category.

NAR GR. This paradigm generates the tokens of the target item in parallel via discrete diffusion modeling (Shi et al., 2025). It operates through a forward masking process and a reverse denoising process. In the forward process, each token of the target item is independently masked with mask token [M] with probability r, or remains unchanged with probability $1 - r$. In the reverse process, the model $\mathcal{P}_\theta$ predicts all masked tokens simultaneously. Specifically, the reverse process proceeds over T steps. At each step t, the model predicts the original tokens for all currently masked positions. Formally, $\mathcal{P}_\theta(v_n|\mathcal{S}_{\mathcal{H}})$ can be factorized as:

$$\prod_{t=1}^{T}\prod_{l=1}^{L}\begin{cases}\mathcal{P}_\theta(c_{n,l}|v_n^t, \mathcal{S}_{\mathcal{H}}), & \text{if } c_{n,l}^t = [\texttt{M}],\\ 1, & \text{otherwise},\end{cases} \tag{2}$$

where v_n^t is the generated item token sequence at step t, and $c_{n,l}^t$ denotes the generated token at level l and step t. At step t, the model selects the top-$\frac{L}{T}$ predictions to unmask, and the remaining positions are re-masked for subsequent steps.

3 Methodology

As shown in Figure 2, `CreGR` comprises two key components. **(1) Credibility-aware tokenizer** discretizes each item into token sequences while enforcing discriminative representations between credible and uncredible items at the token level. **(2) Accuracy-preserving and credibility-oriented generator** reduces the masking probabilities of tokens encoding uncredible information identified by the tokenizer, while leaving all other tokens unaffected, thereby suppressing the influence of uncredible content in the generation process without compromising recommendation accuracy.

3.1 Credibility-aware Tokenizer

We first employ LLaMA2-7B (Touvron et al., 2023) to encode the title and description of each item v_i into a title embedding $\mathbf{e}_i^{tit}$ and a description embedding $\mathbf{e}_i^{des}$. Other word or sentence embedding techniques (Devlin et al., 2019; Lu et al., 2020, 2019) can also be used, such as BERT (Devlin et al., 2019) to derive items' content embeddings. Although these content embeddings capture rich semantic information, they fail to encode user–item interaction behaviors. To compensate for this limi-

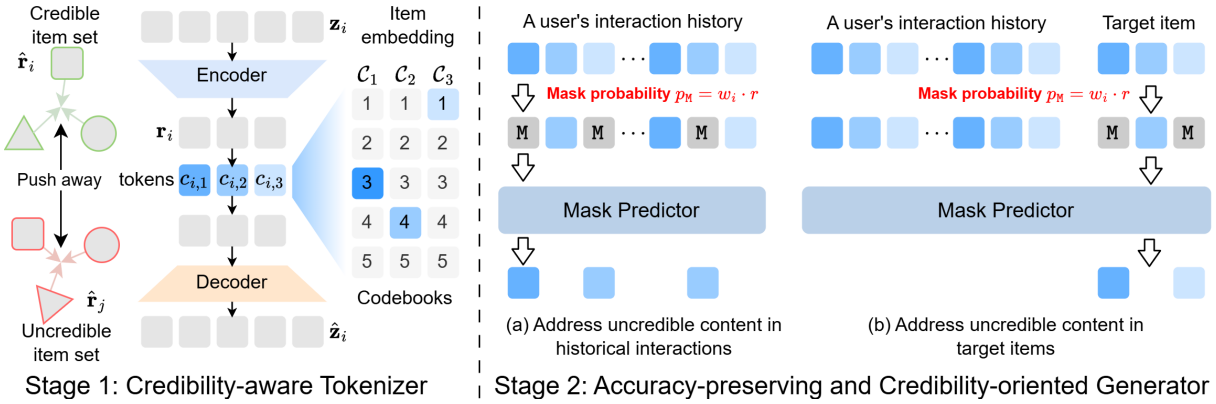

Figure 2: Overall Framework of our proposed model. It consists of two stages: (1) a credibility-aware tokenizer, which tokenizes items into discrete tokens while also disentangling credible and uncredible items in token level; and (2) an accuracy-preserving and credibility-oriented generator, which mitigates the generation of uncredible content items while preserving high recommendaiton accuracy.

tation, we further incorporate collaborative embeddings into the tokenization process. Specifically, we derive item collaborative embeddings $\mathbf{e}_i^{col}$ from a pretrained SASRec model (Kang and McAuley, 2018), a representative sequential recommendation model that captures user behavior patterns and item dependencies. We then apply two separate MLPs to project the content and collaborative embeddings into a unified dimension:

$$\mathbf{z}_i^{con} = [\mathbf{e}_i^{tit} || \mathbf{e}_i^{des}]\mathbf{W}_1, \quad \mathbf{z}_i^{col} = \mathbf{e}_i^{col}\mathbf{W}_2, \tag{3}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times d_2}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_3 \times d_2}$ are learnable weight matrices. $||$ is the embedding concatenation operation. Dimension $d_1 = 2 \times 4096$ and $d_2 = d_3 = 64$ in our implementation. Then, item v_i's embedding is represented as $\mathbf{z}_i = \mathbf{z}_i^{con} || \mathbf{z}_i^{col}$. We then project item embedding into a latent space via an encoder for subsequent vector quantization:

$$\mathbf{r}_i = \mathbf{Encoder}(\mathbf{z}_i), \tag{4}$$

where $\mathbf{r}_i \in \mathbb{R}^d$. The encoder is instantiated with a three-layer MLPs following previous work (Wang et al., 2024c; Shi et al., 2025). Afterwards, we adopt RQ-VAE to hierarchically discretize the latent representation $\mathbf{r}_i$ into a sequence of tokens. Specifically, RQ-VAE maintains L codebooks $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_L\}$, where each codebook $\mathcal{C}_l = \{\mathbf{e}_{l,k}\}_{k=1}^K$ contains K learnable codeword embeddings. The quantization proceeds iteratively: at each level l, the closest codeword to the current residual is selected, and the residual is updated by subtracting the selected codeword:

$$c_{i,l} = \arg\min_{k \in \{1,\ldots,K\}} \|\mathbf{r}_{i,l} - \mathbf{e}_{l,k}\|_2^2, \tag{5}$$

$$\mathbf{r}_{i,l+1} = \mathbf{r}_{i,l} - \mathbf{e}_{c_{i,l}}, \tag{6}$$

where $\mathbf{r}_{i,1} = \mathbf{r}_i$ and $c_{i,l}$ denotes the assigned token index at level l. After this process, the resulting token sequence v_i is $[c_{i,1}, c_{i,2}, \ldots, c_{i,L}]$. Finally, a decoder reconstructs the original embedding from the quantized representation:

$$\hat{\mathbf{z}}_i = \mathbf{Decoder}\left(\sum_{l=1}^L \mathbf{e}_{c_{i,l}}\right). \tag{7}$$

The decoder mirrors the encoder with a symmetrically reversed architecture. The standard RQ-VAE training objective consists of a reconstruction loss and a vector quantization loss:

$$\mathcal{L}_{RQ-VAE} = \mathcal{L}_{Recon} + \mathcal{L}_{RQ}, \tag{8}$$

$$\mathcal{L}_{Recon} = \|\mathbf{z}_i - \hat{\mathbf{z}}_i\|_2^2, \tag{9}$$

$$\begin{aligned}\mathcal{L}_{RQ} = \sum_{l=1}^L [\, &\|\mathrm{sg}[\mathbf{r}_{i,l}] - \mathbf{e}_{c_{i,l}}\|_2^2 \\ &+ \beta \left\|\mathbf{r}_{i,l} - \mathrm{sg}\left[\mathbf{e}_{c_{i,l}}\right]\right\|_2^2],\end{aligned} \tag{10}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator and β is a hyperparameter, which is set to 0.25 following the common practice (Rajput et al., 2023; Wang et al., 2024c; Shi et al., 2025).

However, current vector quantization techniques are agnostic to content credibility, leading to credibility entanglement discussed in **CH 1**. Specifically, the standard *RQ-VAE optimizes purely for reconstruction fidelity, causing the codebook to organize entries along directions that minimize quantization error rather than along the credibility dimension.* As a result, credible and uncredible items whose embeddings differ in factuality but are close in reconstruction distance may still be assigned to the same codebook entry.

To address this, we introduce a credibility regularization that explicitly encourages the tokenizer to separate credible and uncredible items prior to code assignment. To mitigate objective conflict between the quantization loss and credibility regularization—as directly applying regularization on $\mathbf{r}_i$ would cause credible/uncredible items with distinct topical content to collapse toward each other in the embedding space (empirical evidence is provided in Appendix D)—we first project $\mathbf{r}_i$ into a dedicated subspace via a two-layer MLP before computing the regularization: $\hat{\mathbf{r}}_i = \mathbf{MLP}(\mathbf{r}_i)$. Formally:

$$\begin{aligned}\mathcal{L}_{Cred} = &\frac{1}{|\mathcal{I}_{cre}|}\sum_{v_i \in \mathcal{I}_{cre}} \|\hat{\mathbf{r}}_i - \texttt{Cen}(\hat{\mathbf{r}}_i)\|_2^2 \\ &+ \frac{1}{|\mathcal{I}_{unc}|}\sum_{v_j \in \mathcal{I}_{unc}} \|\hat{\mathbf{r}}_j - \texttt{Cen}(\hat{\mathbf{r}}_j)\|_2^2 \\ &- \|\texttt{Cen}(\hat{\mathbf{r}}_i) - \texttt{Cen}(\hat{\mathbf{r}}_j)\|_2^2 ,\end{aligned} \tag{11}$$

where $\mathcal{I}_{cre}$ and $\mathcal{I}_{unc}$ are credible item set and uncredible item set. $\texttt{Cen}(\hat{\mathbf{r}}_i) = \frac{1}{|\mathcal{I}_{cre}|}\sum_{v_i \in \mathcal{I}_{cre}} \hat{\mathbf{r}}_i$ and $\texttt{Cen}(\hat{\mathbf{r}}_j) = \frac{1}{|\mathcal{I}_{unc}|}\sum_{v_j \in \mathcal{I}_{unc}} \hat{\mathbf{r}}_j$ are the centroids of the $\hat{\mathbf{r}}_i$ and $\hat{\mathbf{r}}_j$ for credible and uncredible items. The first two terms encourage items within each credibility group to cluster tightly around their respective centroid (*intra-class compactness*), while the third term pushes the two centroids apart (*inter-class separation*), thus promoting a clear separation between credible and uncredible items in the tokenization space. The rationale for our centroid-based design is provided in Appendix G.

The overall training objective of our designed credibility-aware tokenizer is:

$$\mathcal{L}_{Tokenization} = \mathcal{L}_{RQ-VAE} + \mathcal{L}_{Cred}. \tag{12}$$

3.2 Accuracy-preserving and Credibility-oriented Generator

To address **CH2**, we propose an accuracy-preserving and credibility-oriented generator based on discrete diffusion. Specifically, we design an **asymmetric masking probability reduction strategy** to suppress the contributions of the last tokens of uncredible items during generation, while preserving valuable content-level topical signals.

Our design is motivated by the hierarchical nature of our credibility-aware tokenizer: the earlier levels capture dominant semantic patterns such as topics and writing styles, while the later levels progressively encode finer-grained details through residual refinement (Rajput et al., 2023). Since *factual discrepancies between credible and uncredible items typically manifest as subtle, detail-level differences* (e.g., specific claims, cited sources) rather than coarse topical divergence (Zhou and Zafarani, 2020; Zhou et al., 2019; Liu et al., 2026c; Yang et al., 2023a), such credibility-discriminative signals are more likely to be preserved at the last few token levels after the shared topical information has been absorbed by the preceding levels. Conversely, reducing the masking probabilities of earlier levels would suppress tokens that encode valuable user preference signals, thereby degrading recommendation accuracy. Empirical evidence supporting this design choice is provided in Appendix E.

Specifically, our generator operates through a forward masking process and a reverse denoising process. The forward process progressively masks tokens in the input sequence. For a masking ratio $r \sim U[0,1]$, each token is independently masked with probability r. When $r = 1$, all tokens are replaced with $[\texttt{M}]$; when $r = 0$, the sequence is fully observed. For the last token of uncredible item v_i, we further decrease the masking probability by a weight $w_i \in [0,1)$ to lower its contribution in the generation process. Formally, the masking probability of token $c_{i,l}^r$ at step r is defined as:

$$p_{\texttt{M}}(c_{i,l}^r) = \begin{cases} 0, & \text{if } c_{i,l}^r = [\texttt{M}] \\ w_i \cdot r, & \text{elif } v_i \in \mathcal{I}_{unc} \text{ and } l = L, \\ r, & \text{else,} \end{cases} \tag{13}$$

This asymmetric masking probability reduction strategy suppresses the influence of uncredible signals during generation while leaving all other tokens—which encode valuable user preference signals such as topical interests—unaffected, thereby mitigating the inherent trade-off between recommendation accuracy and credibility.

This asymmetric reduction strategy is instantiated through two variants: **(1) Hard weighting strategy**, which sets $w_i = 0$ for the last tokens of all uncredible items, completely erasing their contributions to the generation process. **(2) Soft weighting strategy**, which assigns w_i based on the credibility degree. Specifically, the credibility degree is estimated by computing the distance difference between the last token and the centroids of the credible and uncredible item clusters. Our motivation is: **the lower the credibility degree** of an item's last token—i.e., the closer it lies to

the uncredible centroid—the more its associated information is contaminated by uncredible signals, and thus **the smaller the weight w_i should be assigned to suppress its contribution**. Formally, w_i in soft version is computed as:

$$\sigma(\mathtt{D}(\mathbf{e}_{c_{i,L}}, \mathtt{Cen}(\mathbf{e}_{c_{j,L}})) - \mathtt{D}(\mathbf{e}_{c_{i,L}}, \mathtt{Cen}(\mathbf{e}_{c_{i',L}}))), \tag{14}$$

where $\mathtt{D}(\cdot,\cdot)$ is the squared Euclidean distance and $\mathbf{e}_{c_{i,L}}$ is item v_i's last token's embedding. $\mathtt{Cen}(\mathbf{e}_{c_{j,L}})$ and $\mathtt{Cen}(\mathbf{e}_{c_{i',L}})$ are centroids of last token embeddings for uncredible and credible items. σ is the Sigmoid activation function.

At the reverse denoising step, a mask predictor $\mathcal{P}_\theta$ is trained to predict all masked tokens ([M]) simultaneously. To mitigate the influence of uncredible items in both users' historical interactions and the target items, we design two training objectives: (1) **historical credibility loss** encourages the model to accurately predict the credible tokens in users' historical interactions, and (2) **target credibility loss** encourages the model to accurately predict the credible tokens in the target items. Formally, the training objectives are defined as:

$$\mathcal{L}_{His} = -\mathbb{E}_r \left[\frac{1}{r} \sum_{c_{i,l} \in \mathcal{S}_{\mathcal{H}}} \log \mathcal{P}_\theta(c_{i,l} | \mathcal{S}^r_{\mathcal{H}}) \right], \tag{15}$$

$$\mathcal{L}_{Tar} = -\mathbb{E}_r \left[\frac{1}{r} \sum_{l=1}^{L} \log \mathcal{P}_\theta(c_{n,l} | v^r_n, \mathcal{S}_{\mathcal{H}}) \right], \tag{16}$$

where **the sum is taken over masked positions**, i.e., $c^r_{i,l} = [\mathtt{M}]$ and $c^r_{n,l} = [\mathtt{M}]$. $\mathcal{P}_\theta$ is a bidirectional Transformer. The overall training objective of our designed generator is:

$$\mathcal{L}_{Generation} = \lambda \mathcal{L}_{His} + \mathcal{L}_{Tar}, \tag{17}$$

where λ controls the strength of historical credibility training objective.

During inference, we begin with a fully masked token sequence $[\mathtt{M}, \cdots, \mathtt{M}]$ and iteratively apply the mask predictor $\mathcal{P}_\theta$ to decode the target item tokens over T diffusion steps. At each step, the $\frac{L}{T}$ token positions with the highest predicted unmasking probabilities are decoded, while all remaining positions are re-masked to [M] and forwarded to the next step. During this process, we employ discrete diffusion-based beam search to obtain the top-N most probable token sequences, which are subsequently mapped back to their corresponding items and returned as the final recommendations, following the practice of (Shi et al., 2025).

Table 1: The statistics of the used three datasets.

Datasets	PolitiFact	GossipCop	MHMisinfo
# Training sequences	103,335	510,149	38,083
# Test sequences	21,490	68,002	8,060
# Items	616	9,529	3,160
# Credible items	306	6,792	2,815
# Uncredible items	310	2,737	345

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. Following prior work (Cai et al., 2026a), we conduct experiments on three real-world datasets: PolitiFact (Shu et al., 2020), (Shu et al., 2020), and MHMisinfo (Nguyen et al., 2025). The PolitiFact and GossipCop datasets are derived from FakeNewsNet repository[1] (Shu et al., 2020). PolitiFact and GossipCop datasets are user-news interaction datasets, where the uncredible items are fake news. MHMisinfo is a user-health information interaction dataset derived from a video-based mental health misinformation dataset[2] (Nguyen et al., 2025), where the uncredible items are health misinformation. To the best of our knowledge, **these are the only publicly available datasets that satisfy both requirements of this task**: user-item interaction data and item-level credibility labels. The statistics of these datasets are reported in Table 1.

Evaluation Metrics. We adopt widely used metrics across three dimensions: recommendation accuracy (NDCG@N, HR@N), content credibility (CR@N, i.e., credible rate, the proportion of credible items in the recommendation list), and their balance (HC@N, which combines HR@N and CR@N) (Cai et al., 2026a). **All metrics are higher-the-better**, and their detailed formulations are provided in Appendix B.2.

4.2 Baselines

We compare our proposed model `CreGR` with the representative and state-of-the-art baselines from three categories: (i) **general recommendation models:** including two traditional models – GRURec (Hidasi et al., 2015), SASRec (Kang and McAuley, 2018), and two diffusion recommenders – DreamRec (Yang et al., 2023b) and PreferDiff (Liu et al., 2025b); (ii) **credible recommendation models:** Rec4Mit (Wang et al., 2022c), HDInt (Wang et al., 2024a), PRISM (Ma et al., 2025), `Disco` (Cai et al., 2026a); and (iii) **generative recom-**

[1]https://github.com/KaiDMML/FakeNewsNet
[2]https://zenodo.org/records/13191247

Table 2: Performance comparison of different methods across datasets. The best performance is in red, and the second-best performance is in blue. The improvement is statistically significant (paired t-test and p-value<0.05).

	Metrics	GRU4Rec	SASRec	DreamRec	PreferDiff	Rec4Mit	HDInt	PRISM	Disco	TIGER	LETTER	LLaDA-Rec	CreGR (hard)	CreGR (soft)
PolitiFact	HR@5	0.2142	0.2158	0.2416	0.2531	0.2039	0.2068	0.2043	0.2678	0.2109	0.2637	0.2720	0.2718	0.2762
	HR@10	0.3390	0.3519	0.3287	0.3554	0.3403	0.3413	0.3317	0.3775	0.3277	0.3838	0.3916	0.3875	0.3920
	NDCG@5	0.1463	0.1386	0.1767	0.1818	0.1348	0.1310	0.1299	0.1983	0.1439	0.1835	0.1942	0.1957	0.1949
	NDCG@10	0.1863	0.1823	0.2047	0.2147	0.1785	0.1741	0.1707	0.2336	0.1815	0.2252	0.2326	0.2329	0.2318
	CR@5	0.9266	0.9059	0.8620	0.8925	0.9118	0.9081	0.8888	0.9823	0.7994	0.9080	0.8887	0.9959	0.9855
	CR@10	0.9122	0.9028	0.8437	0.8981	0.9061	0.9025	0.9150	0.9425	0.8046	0.8968	0.8691	0.9934	0.9826
	HC@5	0.2930	0.2923	0.3096	0.3230	0.2818	0.2842	0.2799	0.3466	0.2761	0.3336	0.3374	0.3517	0.3540
	HC@10	0.3889	0.3955	0.3695	0.3968	0.3887	0.3886	0.3846	0.4192	0.3612	0.4136	0.4120	0.4354	0.4361
GossipCop	HR@5	0.2226	0.3078	0.4619	0.4969	0.2503	0.2154	0.2724	0.5236	0.5517	0.5106	0.5803	0.5605	0.5855
	HR@10	0.3194	0.4706	0.5501	0.6022	0.4054	0.3632	0.3920	0.6143	0.6254	0.6029	0.6380	0.6127	0.6384
	NDCG@5	0.1466	0.1607	0.3415	0.3655	0.1393	0.1141	0.1717	0.3996	0.4531	0.4017	0.4930	0.4762	0.4984
	NDCG@10	0.1778	0.2135	0.3704	0.3999	0.1893	0.1618	0.2103	0.4292	0.4770	0.4317	0.5118	0.4932	0.5157
	CR@5	0.8864	0.8743	0.8464	0.8307	0.9087	0.9032	0.8384	0.9277	0.8536	0.8418	0.8320	0.9601	0.9132
	CR@10	0.8706	0.8526	0.8339	0.8228	0.9065	0.8629	0.8790	0.9039	0.8439	0.8344	0.8236	0.9234	0.8543
	HC@5	0.2964	0.3612	0.4417	0.4525	0.3228	0.2917	0.3032	0.4919	0.4813	0.4614	0.4846	0.5172	0.5131
	HC@10	0.3685	0.4474	0.4744	0.4888	0.4280	0.3944	0.4144	0.5208	0.5039	0.4931	0.5005	0.5266	0.5118
MHMisinfo	HR@5	0.1151	0.1485	0.1819	0.1974	0.1480	0.1364	0.1424	0.2215	0.1617	0.1780	0.1948	0.2197	0.2240
	HR@10	0.1894	0.2592	0.2426	0.2620	0.2537	0.2164	0.2393	0.2822	0.2110	0.2801	0.2422	0.2742	0.2876
	NDCG@5	0.0760	0.0826	0.1313	0.1389	0.0874	0.0835	0.0969	0.1580	0.1131	0.1188	0.1396	0.1397	0.1480
	NDCG@10	0.0998	0.1179	0.1509	0.1598	0.1207	0.1094	0.1289	0.1778	0.1288	0.1513	0.1547	0.1573	0.1683
	CR@5	0.8380	0.8839	0.9002	0.8693	0.8641	0.8333	0.8667	0.9305	0.9118	0.9290	0.9117	0.9579	0.9445
	CR@10	0.8608	0.8915	0.8952	0.8874	0.9056	0.8929	0.9212	0.9264	0.8757	0.9253	0.8914	0.9430	0.9315
	HC@5	0.1806	0.2223	0.2591	0.2715	0.2205	0.2055	0.2144	0.3001	0.2387	0.2574	0.2730	0.3012	0.3039
	HC@10	0.2630	0.3278	0.3147	0.3295	0.3252	0.2915	0.3150	0.3507	0.2848	0.3489	0.3138	0.3467	0.3556

mendation models: TIGER (Rajput et al., 2023), LETTER (Wang et al., 2024c), LLaDA-Rec (Shi et al., 2025). The detailed introduction of these methods is provided in Appendix B.1.

4.3 Implementation Details

For fair comparison, all baseline methods are initialized with hyperparameters suggested in their original papers and subsequently fine-tuned on our datasets to ensure their best performances are reported. We first train the tokenization stage with 10k epochs, and the learning rate is searched in range {1e-5, 5e-5, 1e-4}. We then train the generation stage and set the maximum epoch to 500, with learning rate searched in range {5e-4, 1e-3, 3e-3, 5e-3}. Batch size is searched in range {256, 512, 1024, 2048, 4096}. L, K, T are set to 4, 256 and 4 following prior work (Shi et al., 2025). λ is searched in range {3, 4, 5, 6, 7}. Beam width is fixed to 50. Model parameters are optimized using AdamW (Loshchilov and Hutter, 2017). Each method is run five times, and the average performance is reported. More implementation details are provided in Appendix B.3.

4.4 Performance Comparison

From the results reported in Table 2, we have the following observations:

Credibility Improvement. `CreGR` achieves the best recommendation credibility (CR@5 and CR@10) across all datasets, demonstrating the effectiveness of our credibility-aware tokenization and accuracy-preserving and credibility-oriented generation designs. Notably, the hard weighting strategy (i.e., `CreGR` (hard)) attains higher credibility than its soft counterpart (i.e., `CreGR` (hard)), owing to its more aggressive suppression of uncredible token contributions.

Accuracy Preservation. `CreGR` preserves competitive recommendation accuracy. Although `CreGR` yields slightly lower NDCG than `Disco` on PolitiFact and MHMisinfo datasets, it maintains superior or comparable accuracy across all other cases. This can be attributed to: (1) our accuracy-preserving and credibility-oriented generator selectively reduces the masking probability of only the last token of uncredible items, leaving all other tokens that encode user preference signals unaffected; and (2) the incorporation of collaborative embeddings from a pretrained model further enriches item representations, providing additional accuracy gains.

Credibility-accuracy Balance. `CreGR` achieves the best performance in terms of HC@5 and HC@10, demonstrating its effectiveness in balancing recommendation accuracy and credibility. This achievement can be attributed to our designed accuracy-preserving and credibility-oriented generator which can mitigates uncredible content while preserving recommendation accuracy as much as possible. This balance is of particular practical significance, as both objectives are indispensable in real-world recommendation scenarios.

Table 3: Performance comparison between CreGR (full) and other model variants. The improvement is statistically significant (paired t-test and p-value<0.05).

Datasets	PolitiFact				GossipCop				MHMisinfo			
Methods	HR@5	NDCG@5	CR@5	HC@5	HR@5	NDCG@5	CR@5	HC@5	HR@5	NDCG@5	CR@5	HC@5
CreGR (full)	0.2762	0.1949	0.9855	0.3540	0.5855	0.4984	0.9132	0.5131	0.2240	0.1480	0.9445	0.3039
w/o Title	0.2721	0.1930	0.9758	0.3494	0.5824	0.4952	0.8737	0.4992	0.2035	0.1461	0.9186	0.2820
w/o Description	0.2732	0.1932	0.9380	0.3453	0.5838	0.4967	0.8765	0.5007	0.1993	0.1412	0.9154	0.2777
w/o Collaborative	0.2819	0.1933	0.9446	0.3531	0.5802	0.4953	0.8645	0.4954	0.2192	0.1465	0.9284	0.2978
w/o CT	0.2699	0.1914	0.9768	0.3477	0.5785	0.4971	0.8569	0.4923	0.2031	0.1413	0.9202	0.2818
w/o APCOG	0.2791	0.1979	0.8999	0.3445	0.5850	0.4962	0.8424	0.4898	0.2153	0.1523	0.8302	0.2835

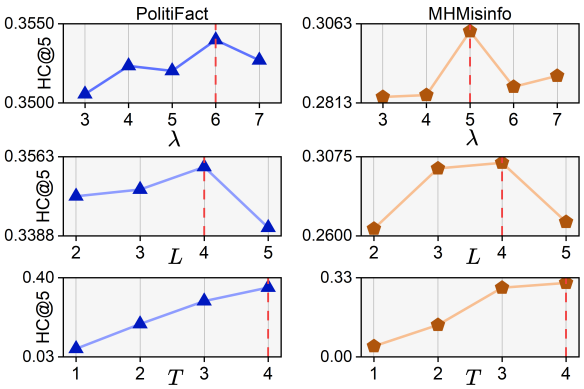

Figure 3: Effect of λ, L, T on HC@5 of CreGR (soft).

4.5 Ablation Study

To evaluate the contribution of each component, we construct five variants and report results in Table 3. All results are based on CreGR (soft).

Variants **w/o Title**, **w/o Desc**, and **w/o Coll** remove title, description, and collaborative embeddings from the tokenization stage, respectively. The consistent performance drops confirm that each information source is indispensable: title and description provide semantic content signals, while collaborative embeddings supply user interaction patterns not captured by content alone.

Variant **w/o CT** replaces credibility-aware tokenizer with standard RQ-VAE tokenizer. The resulting degradation demonstrates that without explicit credibility disentanglement at the token level, the generation stage loses the ability to distinguish credible from uncredible content, jointly harming both recommendation accuracy and credibility.

Variant **w/o APCOG** disables accuracy-preserving and credibility-oriented generator by removing both hard and soft token weighting strategies. While this variant achieves marginal accuracy gains in certain cases, it incurs a substantial drop in recommendation credibility—an unacceptable trade-off given that credibility is a core optimization objective of CreGR.

Table 4: Comparison of collision rate between generative recommendation methods.

Datasets	TIGER	LETTER	LLaDA-Rec	CreGR
PolitiFact	8.0%	12.3%	8.0%	0%
GossipCop	5.9%	5.3%	5.0%	0%
MHMisinfo	10.5%	3.1%	0.5%	0.1%

4.6 In-depth Analysis

4.6.1 Hyperparameters

We vary λ over $\{3, 4, 5, 6, 7\}$, L over $\{2, 3, 4, 5\}$, and T over $\{1, 2, 3, 4\}$ to examine their sensitivity. From Figure 3, we have following observations:

(1) λ. HC@5 first improves then degrades as λ increases, peaking at $\lambda = 6$ and $\lambda = 5$ on PolitiFact and MHMisinfo. This suggests that a moderate λ strikes the best balance between recommendation accuracy and credibility. Overly small or large values will degrades model performance.

(2) L. HC@5 first improves then degrades as L increases, peaking at $L = 4$ for all datasets. A larger L provides finer-grained vector quantization but risks over-segmentation, leading to performance degradation beyond the optimal value. A smaller L is sufficient to capture the hierarchical nature of item's semantic information, thereby harming model performance.

(3) T. HC@5 consistently rises as T increases. Reducing the number of generation steps significantly increases the uncertainty of generation trajectory, and therefore decrease recommendation accuracy and credibility. Larger T ensures finer-grained generation trajectories, leading to optimal recommendation performance.

Due to space constraints, results on GossipCop are omitted; this dataset exhibits a similar trend with optimal performance at λ=5, L=4 and T=4.

4.6.2 Collision rate comparison

The collision rate measures the proportion of distinct items assigned identical token sequences,

Table 5: Performance comparison under different levels of noisy labels.

Method	PolitiFact				GossipCop				MHMisinfo			
	HR@5	NDCG@5	CR@5	HC@5	HR@5	NDCG@5	CR@5	HC@5	HR@5	NDCG@5	CR@5	HC@5
TIGER	0.2109	0.1439	0.7994	0.2761	0.5517	0.4531	0.8536	0.4813	0.1617	0.1131	0.9118	0.2387
LETTER	0.2637	0.1835	0.9080	0.3336	0.5106	0.4017	0.8418	0.4614	0.1780	0.1188	0.9290	0.2574
LLaDA-Rec	0.2720	0.1942	0.8887	0.3374	0.5803	0.4930	0.8320	0.4846	0.1948	0.1396	0.9117	0.2730
CreGR	0.2762	0.1949	0.9855	0.3540	0.5855	0.4984	0.9132	0.5131	0.2240	0.1480	0.9445	0.3039
Label-noise proportion												
5% (CreGR)	0.2733	0.1923	0.9803	0.3509	0.5846	0.4998	0.9100	0.5117	0.2209	0.1501	0.9380	0.3003
10% (CreGR)	0.2720	0.1913	0.9801	0.3498	0.5836	0.4969	0.9095	0.5112	0.2168	0.1471	0.9347	0.2962
15% (CreGR)	0.2717	0.1932	0.9780	0.3493	0.5819	0.4949	0.9068	0.5097	0.2124	0.1453	0.9335	0.2919

where a lower value indicates more discriminative tokenization. As shown in Table 4, CreGR achieves the lowest collision rate across all datasets. We attribute this to two designs: (1) credibility-aware tokenizer encourages discriminative token representations for credible and uncredible items, reducing token sequence collisions between factually divergent items; and (2) collaborative embeddings enrich item representations with user interaction signals, enabling finer-grained distinctions between items with similar content.

4.6.3 Label-noise Performance

In our Limitations section, we point out that exploring semi-supervised or unsupervised approaches is part of our future work, since in real-world scenarios, items' credibility labels may not be always available. These can serve as the solutions when credibility labels are noise or sparse. In the current version, we do not explicitly design a mechanism to resolve the label-noise issue. However, we conducted additional experiments to investigate the performance of CreGR in label-noise scenarios. Specifically, we randomly flipped the labels of a portion (5%, 10%, and 15%) of uncredible items during training (at test time, correct labels are used to accurately evaluate model performance).

As shown in Table 5, CreGR also achieves superior results in the label-noise setting. Specifically, when the noise level is raised to 15%, the recommendation credibility (CR@5) of CreGR only drops slightly, from 0.9855 to 0.9780, still remaining at a very high level (significantly outperforming LETTER's 0.9080). This indicates that although without specific designs, our method still have high potential to achieve great recommendation accuracy and credibility in scenarios where partial labels are noise or unavailable.

Table 6: Tokenization and generation cost comparison.

Stages	Tokenization Total (minutes)		Generation Training/epoch (s)		Generation Inference (s)	
Methods	LLaDA-Rec	CreGR	LLaDA-Rec	CreGR	LLaDA-Rec	CreGR
PolitiFact	55	31	16.8	17.5	117.1	117.6
GossipCop	414	75	189.0	194.8	1006.0	1025.5
MHMisinfo	160	47	5.7	5.9	43.0	43.2

4.7 Efficiency Analysis

We compare the efficiency of CreGR (soft) against LLaDA-Rec, the only other discrete diffusion-based baseline. As shown in Table 6, CreGR achieves higher tokenization efficiency, as LLaDA-Rec employs a multi-head VQ-VAE whose parallel multi-head operations incur substantially higher training overhead. In terms of generation efficiency, CreGR remains comparable to LLaDA-Rec, confirming that our designs introduce negligible computational cost during training and inference.

5 Conclusion

In this paper, we propose a novel credible generative recommendation model, CreGR, to mitigate the recommendations of uncredible content in RSs. Specifically, we design a credibility-aware tokenizer to learn discriminative item tokens via credibility regularization. Afterwards, we propose a novel accuracy-preserving and credibility-oriented generator to reduce the contributions of tokens encoding uncredible information in the generation process. During this process, tokens representing users' preference signals remain unaffected to preserve high recommendation accuracy. Extensive experiments on three real-world datasets demonstrate the effectiveness of CreGR on improving recommendation credibility while preserving high recommendation accuracy.

Limitations

While our proposed model demonstrates promising results in generating credible content recommendations, there are several limitations that warrant further investigation.

Label availability. Our model relies on the availability of labeled data to identify credible and uncredible items during the tokenization and generation stage. In real-world scenarios, such labels may not always be available or may be noisy, which could affect the performance of our model. We plan to explore semi-supervised or unsupervised approaches to mitigate this limitation in future work. For example, Large Language Models (LLMs) (Li et al., 2025; Zhang et al., 2026; Gu et al., 2026) and LLM agents (Huang et al., 2024; Lu et al., 2026a,b,c) can serve as effective tools to handle semi-supervised and unsupervised scenarios, due to their powerful open-world knowledge.

Content modality. Our current model primarily focuses on textual content (e.g., news articles). Although MHMisinfo is a user-video interaction dataset, it doesn't provide the video metadata. It only contains the video titles and descriptions. However, in many recommendation scenarios, items may contain multimodal content (e.g., images, videos) that also contribute to their credibility. Future work could explore extending our model to incorporate multimodal information for a more comprehensive assessment of item credibility via multimodal LLMs (Li et al., 2026; Zhou et al., 2025; Yang et al., 2025, 2026a,c,b) or other techniques.

Dataset Scalability. The datasets used in our experiments (GossipCop: 510,149 sequences; PolitiFact: 103,335 sequences) contain substantially more interaction sequences than commonly used Amazon benchmarks in GR (McAuley et al., 2015; Hou et al., 2024) (Sports: 35,598; Beauty: 22,363; Toys: 19,412), with GossipCop even containing approximately 2-3 times more sequences than larger Amazon datasets (Automotive: 193,651; Phones: 157,212; Tools: 240,799). This indicates that our used datasets provide a robust basis for validating the effectiveness and generalizability of `CreGR`. However, we acknowledge that the item vocabularies of these datasets are relatively limited, as they represent the only publicly available benchmarks that satisfy the dual requirements of user–item interaction data and item-level credibility labels; Amazon datasets, despite their scale, cannot be used as they provide no such credibility annotations. We leave the evaluation on larger-scale datasets as future work, and note that our method can be seamlessly deployed on them once such resources become available.

Ethical Considerations

This paper aims to develop a credibility-aware generative recommendation model to serve societal good by mitigating the dissemination of uncredible content (e.g., fake news and misinformation) through recommender systems (RSs). All datasets used in our experiments are derived from publicly available sources and contain no personally identifiable information.

References

Rui Bing, Guan Yuan, Zhuo Cai, Bohan Li, and Yong Zhou. 2024. Automated modeling of influence diversity with graph convolutional network for social recommendation. In *APWeb-WAIM*, pages 33–49.

Zhuo Cai, Shoujin Wang, Victor W Chu, Usman Naseem, Yang Wang, and Fang Chen. 2025. Unleashing the potential of diffusion models towards diversified sequential recommendations. In *SIGIR*, pages 1476–1486.

Zhuo Cai, Shoujin Wang, Jin Li, Peilin Zhou, Victor W Chu, Fang Chen, Tianqing Zhu, and Charu C Aggarwal. 2026a. Steering diffusion models towards credible content recommendation. In *ICLR*.

Zhuo Cai, Shoujin Wang, Peilin Zhou, Yan Wang, Xiao Liu, Lianyong Qi, Julian McAuley, and Dietmar Jannach. 2026b. Misinformation unlearning for responsible content recommendation. *ACM Transactions on Information Systems*.

Zhuo Cai, Guan Yuan, Shaojie Qiao, Song Qu, Yanmei Zhang, and Rui Bing. 2022. Fg-cf: Friends-aware graph collaborative filtering for poi recommendation. *Neurocomputing*, 488:107–119.

Zhuo Cai, Guan Yuan, Xiaobao Zhuang, Xiao Liu, Rui Bing, and Shoujin Wang. 2023a. All you see is the tip of the iceberg: Distilling latent interactions can help you find treasures. In *ICONIP*, pages 244–257.

Zhuo Cai, Guan Yuan, Xiaobao Zhuang, Senzhang Wang, Shaojie Qiao, and Mu Zhu. 2023b. Adaptive self-propagation graph convolutional network for recommendation. *World Wide Web*, 26(5):3183–3206.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *ICLR*, pages 54158–54183.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Yijie Ding, Zitian Guo, Jiacheng Li, Letian Peng, Shuai Shao, Wei Shao, Xiaoqiang Luo, Luke Simon, Jingbo Shang, Julian McAuley, and 1 others. 2026. How well does generative recommendation generalize? *arXiv preprint arXiv:2603.19809*.

Dengzhao Fang, Jingtong Gao, Yu Li, Xiangyu Zhao, and Yi Chang. 2026. Prism: Purified representation and integrated semantic modeling for generative sequential recommendation. *arXiv preprint arXiv:2601.16556*.

Miriam Fernandez and Alejandro Bellogín. 2020. Recommender systems and misinformation: the problem or the solution?

Miriam Fernández, Alejandro Bellogín, and Iván Cantador. 2021. Analysing the effect of recommendation algorithms on the amplification of misinformation. *arXiv preprint arXiv:2103.14748*.

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*.

Huilin Gu, Shoujin Wang, Jin Li, Zhuo Cai, Lina Yao, Yan Wang, and Longbing Cao. 2026. A survey on trustworthy llm-based recommender systems.

Zitian Guo, Yupeng Hou, Clark Mingxuan Ju, Neil Shah, and Julian McAuley. 2026. Mlps are efficient distilled generative recommenders. *arXiv preprint arXiv:2605.12617*.

Ruining He, Lukasz Heldt, Lichan Hong, Raghunandan Keshavan, Shifan Mao, Nikhil Mehta, Zhengyang Su, Alicia Tsai, Yueqi Wang, Shao-Chuan Wang, and 1 others. 2026. Plum: Adapting pre-trained language models for industrial-scale generative recommendations. In *WWW*, pages 8093–8104.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

Benedikt Hornig, Maria Soledad Pera, and Jan Scholtes. 2024. Misinformation in video recommendations: an exploration of top-n recommendation algorithms. In *ROMCIR@ ECIR*, pages 54–67.

Yupeng Hou, Zhankui He, Julian McAuley, and Wayne Xin Zhao. 2023. Learning vector-quantized item representation for transferable sequential recommenders. In *WWW*, pages 1162–1171.

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*.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Yupeng Hou, Jiacheng Li, Ashley Shin, Jinsung Jeon, Abhishek Santhanam, Wei Shao, Kaveh Hassani, Ning Yao, and Julian McAuley. 2025a. Generating long semantic ids in parallel for recommendation. In *KDD*, pages 956–966.

Yupeng Hou, Jianmo Ni, Zhankui He, Noveen Sachdeva, Wang-Cheng Kang, Ed H Chi, Julian McAuley, and Derek Zhiyuan Cheng. 2025b. Actionpiece: Contextually tokenizing action sequences for generative recommendation. *arXiv preprint arXiv:2502.13581*.

Yupeng Hou, An Zhang, Leheng Sheng, Zhengyi Yang, Xiang Wang, Tat-Seng Chua, and Julian McAuley. 2025c. Generative recommendation models: Progress and directions. In *Companion of WWW*, pages 13–16.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*.

Pengyue Jia, Xiaobei Wang, Yingyi Zhang, Shuchang Liu, Yupeng Hou, Hailan Yang, Xu Gao, Xiaopeng Li, Yejing Wang, Julian McAuley, and 1 others. 2026. From local indices to global identifiers: Generative reranking for recommender systems via global action space. *arXiv preprint arXiv:2604.25291*.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *ICDM*, pages 197–206.

Haven Kim, Yupeng Hou, and Julian McAuley. 2026. Fusid: Modality-fused semantic ids for generative music recommendation. *arXiv preprint arXiv:2601.08764*.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image generation using residual quantization. In *CVPR*, pages 11523–11532.

Jin Li, Shoujin Wang, Qi Zhang, Feng Liu, Tongliang Liu, Longbing Cao, Shui Yu, and Fang Chen. 2026. Revealing multimodal causality with large language models. *NeurIPS*, 38:101384–101430.

Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen. 2025. A survey on diffusion language models. *arXiv preprint arXiv:2508.10875*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL*, pages 3214–3252.

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 *SIGIR*, pages 1923–1933.

Chengyi Liu, Yongqi Zhou, Junwei Pan, Zhixiang Feng, Chengguo Yin, Haijie Gu, Jie Jiang, Yinghao Liu, Yujuan Ding, Qing Li, and 1 others. 2026a. Diffusion language model for recommendation. *arXiv preprint arXiv:2607.21519*.

Enze Liu, Bowen Zheng, Cheng Ling, Lantao Hu, Han Li, and Wayne Xin Zhao. 2025a. Generative recommender with end-to-end learnable item tokenization. In *SIGIR*, pages 729–739.

Shuo Liu, An Zhang, Guoqing Hu, Hong Qian, and Tat-seng Chua. 2025b. Preference diffusion for recommendation. In *ICLR*.

Zhao Liu, Yichen Zhu, Yiqing Yang, Xiao Lv, Guoping Tang, Rui Huang, Qiang Luo, Ruiming Tang, and Guorui Zhou. 2026b. Diffgrm: Diffusion-based generative recommendation model. In *WWW*, pages 5853–5864.

Zhiwei Liu, Paul Thompson, Jiaqi Rong, Baojie Qu, Runteng Guo, Min Peng, Qianqian Xie, and Sophia Ananiadou. 2026c. Misspans: Fine-grained false span identification in cross-domain fake news. *arXiv preprint arXiv:2601.04857*.

Ziwei Liu, Yejing Wang, Qidong Liu, Zijian Zhang, Chong Chen, Wei Huang, and Xiangyu Zhao. 2025c. The best of the two worlds: Harmonizing semantic and hash ids for sequential recommendation. *arXiv preprint arXiv:2512.10388*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mingfei Lu, Mengjia Wu, Feng Liu, Jiawei Xu, Weikai Li, Haoyang Wang, Zhengdong Hu, Ying Ding, Yizhou Sun, Jie Lu, and 1 others. 2026a. Choosing how to remember: Adaptive memory structures for llm agents. *arXiv preprint arXiv:2602.14038*.

Mingfei Lu, Mengjia Wu, Jiawei Xu, Weikai Li, Feng Liu, Ying Ding, Yizhou Sun, Jie Lu, and Yi Zhang. 2026b. From newborn to impact: Bias-aware citation prediction. In *WWW*, pages 7600–7609.

Mingfei Lu, Yi Zhang, Mengjia Wu, and Yue Feng. 2026c. From query to counsel: Structured reasoning with a multi-agent framework and dataset for legal consultation. In *ACL*, pages 17064–17078.

Wenpeng Lu, Fanqing Meng, Shoujin Wang, Guoqiang Zhang, Xu Zhang, Antai Ouyang, and Xiaodong Zhang. 2019. Graph-based chinese word sense disambiguation with multi-knowledge integration. *Computers, Materials and Continua*, 61(1):197–212.

Wenpeng Lu, Yuteng Zhang, Shoujin Wang, Heyan Huang, Qian Liu, and Sheng Luo. 2020. Concept representation by learning explicit and implicit concept couplings. *IEEE Intelligent Systems*, 36(1):6–15.

Zheqi Lv, Shaoxuan He, Tianyu Zhan, Shengyu Zhang, Wenqiao Zhang, Jingyuan Chen, Zhou Zhao, and Fei Wu. 2024. Semantic codebook learning for dynamic recommendation models. In *MM*, pages 9611–9620.

Zihan Ma, Minnan Luo, Yiran Hao, Zhi Zeng, Xiangzheng Kong, and Jiahao Wang. 2025. Bridging interests and truth: Towards mitigating fake news with personalized and truthful recommendations. In *SIGIR*, pages 490–503.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*, pages 43–52.

M Jeffrey Mei, Florian Henkel, Samuel E Sandberg, Oliver Bembom, and Andreas F Ehmann. 2025. Semantic ids for music recommendation. In *RecSys*, pages 1070–1073.

Lingyu Mu, Hao Deng, Haibo Xing, Jinxin Hu, Yu Zhang, Xiaoyi Zeng, and Jing Zhang. 2026. Masked diffusion generative recommendation. *arXiv preprint arXiv:2601.19501*.

Viet Cuong Nguyen, Mini Jain, Abhijat Chauhan, Heather Jamie Soled, Santiago Alvarez Lesmes, Zihang Li, Michael L Birnbaum, Sunny X Tang, Srijan Kumar, and Munmun De Choudhury. 2025. Supporters and skeptics: Llm-based analysis of engagement with mental health (mis) information content on video-sharing platforms. In *ICWSM*, pages 1329–1345.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2026. Large language diffusion models. In *NeurIPS*, pages 50608–50646.

Royal Pathak. 2024. Advancing misinformation awareness in recommender systems for social media information integrity. In *CIKM*, pages 5471–5474.

Royal Pathak and Francesca Spezzano. 2024. Analyzing the interplay between diversity of news recommendations and misinformation spread in social media. In *UMAP Adjunct*, pages 80–85.

Royal Pathak, Francesca Spezzano, and Maria Soledad Pera. 2023. Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks. *ACM Transactions on the Web*, pages 1–26.

Gustavo Penha, Edoardo D'Amico, Marco De Nadai, Enrico Palumbo, Alexandre Tamborrino, Ali Vardasbi, Max Lefarov, Shawn Lin, Timothy Heath, Francesco Fabbri, and 1 others. 2025. Semantic ids for joint generative search and recommendation. In *RecSys*, pages 1296–1301.

Shutong Qiao, Wei Yuan, Tong Chen, Xiangyu Zhao, Quoc Viet Hung Nguyen, and Hongzhi Yin. 2026. When text-as-vision meets semantic ids in generative recommendation: An empirical study. *arXiv preprint arXiv:2601.14697*.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, and 1 others. 2023. Recommender systems with generative retrieval. In *NeurIPS*, pages 10299–10315.

Dorsaf Sallami, Rim Ben Salem, and Esma Aïmeur. 2023. Trust-based recommender system for fake news mitigation. In *UMAP Adjunct*, pages 104–109.

Tal Samuel-Azran and Tsahi Hayat. 2019. Online news recommendations credibility: The tie is mightier than the source. *Comunicar: Media Education Research Journal*, pages 71–80.

Kulin Shah, Bhuvesh Kumar, Neil Shah, and Liam Collins. 2025. Masked diffusion for generative recommendation. *arXiv preprint arXiv:2511.23021*.

Teng Shi, Chenglei Shen, Weijie Yu, Shen Nie, Chongxuan Li, Xiao Zhang, Ming He, Yan Han, and Jun Xu. 2025. Llada-rec: Discrete diffusion for parallel semantic id generation in generative recommendation. *arXiv preprint arXiv:2511.06254*.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Zihua Si, Zhongxiang Sun, Jiale Chen, Guozhang Chen, Xiaoxue Zang, Kai Zheng, Yang Song, Xiao Zhang, Jun Xu, and Kun Gai. 2024. Generative retrieval with semantic tree-structured identifiers and contrastive learning. In *SIGIR*, pages 154–163.

Anima Singh, Trung Vu, Nikhil Mehta, Raghunandan Keshavan, Maheswaran Sathiamoorthy, Yilin Zheng, Lichan Hong, Lukasz Heldt, Li Wei, Devansh Tandon, and 1 others. 2024. Better generalization with semantic ids: A case study in ranking for recommendations. In *RecSys*, pages 1039–1044.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dongsheng Wang, Yuxi Huang, Shen Gao, Yifan Wang, Chengrui Huang, and Shuo Shang. 2025a. Generative next poi recommendation with semantic id. In *KDD*, pages 2904–2914.

Nan Wang, Shoujin Wang, Yan Wang, Quan Z Sheng, and Mehmet Orgun. 2020. Modelling local and global dependencies for next-item recommendations. In *WISE*, pages 285–300.

Nan Wang, Shoujin Wang, Yan Wang, Quan Z Sheng, and Mehmet A Orgun. 2022a. Exploiting intra-and inter-session dependencies for session-based recommendations. *World Wide Web*, 25(1):425–443.

Shijia Wang, Tianpei Ouyang, Qiang Xiao, Dongjing Wang, Yintao Ren, Songpei Xu, Da Guo, and Chuanjiang Luo. 2025b. Progressive semantic residual quantization for multimodal-joint interest modeling in music recommendation. In *CIKM*, pages 6119–6127.

Shoujin Wang, Longbing Cao, Yan Wang, Quan Z Sheng, Mehmet A Orgun, and Defu Lian. 2021. A survey on session-based recommender systems. *ACM Computing Surveys (CSUR)*, 54(7):1–38.

Shoujin Wang, Liang Hu, Longbing Cao, Xiaoshui Huang, Defu Lian, and Wei Liu. 2018. Attention-based transactional context embedding for next-item recommendation. In *AAAI*, volume 32.

Shoujin Wang, Liang Hu, Yan Wang, Longbing Cao, Quan Z. Sheng, and Mehmet Orgun. 2019a. Sequential recommender systems: Challenges, progress and prospects. In *IJCAI*.

Shoujin Wang, Liang Hu, Yan Wang, Quan Z Sheng, Mehmet Orgun, and Longbing Cao. 2019b. Modeling multi-purpose sessions for next-item recommendations via mixture-channel purpose routing networks. In *IJCAI*.

Shoujin Wang, Ninghao Liu, Xiuzhen Zhang, Yan Wang, Francesco Ricci, and Bamshad Mobasher. 2022b. Data science and artificial intelligence for responsible recommendations. In *KDD*, pages 4904–4905.

Shoujin Wang, Wentao Wang, Xiuzhen Zhang, Yan Wang, Huan Liu, and Fang Chen. 2024a. A hierarchical and disentangling interest learning framework for unbiased and true news recommendation. In *KDD*, pages 3200–3211.

Shoujin Wang, Xiaofei Xu, Xiuzhen Zhang, Yan Wang, and Wenzhuo Song. 2022c. Veracity-aware and event-driven personalized news recommendation for fake news mitigation. In *WWW*, pages 3673–3684.

Shoujin Wang, Xiuzhen Zhang, Yan Wang, and Francesco Ricci. 2024b. Trustworthy recommender systems. *ACM Transactions on Intelligent Systems and Technology*, 15(4):1–20.

Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2024c. Learnable item tokenization for generative recommendation. In *CIKM*, pages 2400–2409.

Ye Wang, Jiahao Xun, Minjie Hong, Jieming Zhu, Tao Jin, Wang Lin, Haoyuan Li, Linjun Li, Yan Xia, Zhou Zhao, and 1 others. 2024d. Eager: Two-stream generative recommender with behavior-semantic collaboration. In *KDD*, pages 3245–3254.

Yejing Wang, Shengyu Zhou, Jinyu Lu, Ziwei Liu, Langming Liu, Maolin Wang, Wenlin Zhang, Feng Li, Wenbo Su, Pengjie Wang, and 1 others. 2026. Nezha: A zero-sacrifice and hyperspeed decoding architecture for generative recommendations. In *WWW*, pages 8073–8082.

Yuhao Wang, Junwei Pan, Xinhang Li, Maolin Wang, Yuan Wang, Yue Liu, Dapeng Liu, Jie Jiang, and Xiangyu Zhao. 2025c. Empowering large language model for sequential recommendation via multimodal embeddings and semantic ids. In *CIKM*, pages 3209–3219.

Tianxin Wei, Xuying Ning, Xuxing Chen, Ruizhong Qiu, Yupeng Hou, Yan Xie, Shuang Yang, Zhigang Hua, and Jingrui He. 2025. Cofirec: Coarse-to-fine tokenization for generative recommendation. *arXiv preprint arXiv:2511.22707*.

Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Zhiwei Yang. 2023a. Wsdms: debunk fake news via weakly supervised detection of misinforming sentences with contextualized social wisdom. In *EMNLP*, pages 1525–1538.

Xiaoyu Yang, Jie Lu, and En Yu. 2025. Adapting multimodal large language model to concept drift from pre-training onwards. In *ICLR*.

Xiaoyu Yang, Jie Lu, and En Yu. 2026a. Walking the tightrope: Autonomous disentangling beneficial and detrimental drifts in non-stationary custom-tuning. In *NeurIPS*, volume 38, pages 116167–116193.

Xiaoyu Yang, En Yu, Wei Duan, and Jie Lu. 2026b. Towards robust endogenous reasoning: Unifying drift adaptation in non-stationary tuning. *arXiv preprint arXiv:2604.15705*.

Xiaoyu Yang, En Yu, Wei Duan, and Jie Lu. 2026c. Turning drift into constraint: Robust reasoning alignment in non-stationary multi-stream environments. In *ICML*.

Yuhao Yang, Zhi Ji, Zhaopeng Li, Yi Li, Zhonglin Mo, Yue Ding, Kai Chen, Zijian Zhang, Jie Li, Liu Lin, and 1 others. 2026d. Sparse meets dense: Unified generative recommendations with cascaded sparse-dense representations. In *NeurIPS*, volume 38, pages 93746–93770.

Zhengyi Yang, Jiancan Wu, Zhicai Wang, Xiang Wang, Yancheng Yuan, and Xiangnan He. 2023b. Generate what you prefer: Reshaping sequential recommendation via guided diffusion. In *NeurIPS*.

Wencai Ye, Mingjie Sun, Shaoyun Shi, Peng Wang, Wenjin Wu, and Peng Jiang. 2025. Das: Dual-aligned semantic ids empowered industrial recommender system. In *CIKM*, pages 6217–6224.

Guan Yuan, Peijin Yang, Zhuo Cai, Yuqing Du, Shaojie Qiao, and Qiuyan Yan. 2025. Multibehavior recommendation with nonoverlapping heterogeneous graph collaborative filtering. *IEEE Transactions on Computational Social Systems*, 12(6):4819–4831.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Jiayuan He, and 1 others. 2024. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In *ICML*, pages 58484–58509.

Guixian Zhang, Guan Yuan, Debo Cheng, Lin Liu, Jiuyong Li, and Shichao Zhang. 2025. Mitigating propensity bias of large language models for recommender systems. *ACM Transactions on Information Systems*, 43(6):1–26.

Guixian Zhang, Guan Yuan, Ziqi Xu, Yanmei Zhang, Jing Ren, Zhenyun Deng, and Debo Cheng. 2026. Harnessing llm for noise-robust cognitive diagnosis in web-based intelligent education systems. In *WWW*, pages 4093–4104.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2024. Adapting large language models by integrating collaborative semantics for recommendation. In *ICDE*, pages 1435–1448.

Carolina Zheng, Minhui Huang, Dmitrii Pedchenko, Kaushik Rangadurai, Siyu Wang, Fan Xia, Gaby Nahum, Jie Lei, Yang Yang, Tao Liu, and 1 others. 2025. Enhancing embedding representation stability in recommendation systems with semantic id. In *RecSys*, pages 954–957.

Qiyong Zhong, Jiajie Su, Yunshan Ma, Julian McAuley, and Yupeng Hou. 2025. Pctx: Tokenizing personalized context for generative recommendation. *arXiv preprint arXiv:2510.21276*.

Peilin Zhou, Chao Liu, Jing Ren, Xinfeng Zhou, Yueqi Xie, Meng Cao, Zhongtao Rao, You-Liang Huang, Dading Chong, Junling Liu, and 1 others. 2025. When large vision language models meet multimodal sequential recommendation: An empirical study. In *WWW*, pages 275–292.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40.

Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu. 2019. Fake news detection via nlp is vulnerable to adversarial attacks. *arXiv preprint arXiv:1901.09657*.

A Related Work

A.1 Generative Recommendation

Existing GR methods (Hou et al., 2025c; Rajput et al., 2023; Yang et al., 2026d) generally follow a two-stage design: (i) tokenization stage that represents each item with discrete tokens, and (ii) generation stage that generates a sequence of tokens to represent the recommended items (Rajput et al., 2023; Hou et al., 2025b; Guo et al., 2026; Hou et al., 2026; Ding et al., 2026; Kim et al., 2026; Wei et al., 2025; Zhong et al., 2025).

Item Tokenization. From the perspective of tokenization, existing methods can be broadly categorized into clustering-based and vector quantization-based approaches. Clustering-based methods organize items into hierarchical groups and represent each item with the cluster ID (Si et al., 2024; Wang et al., 2024d; Lv et al., 2024). Vector quantization-based methods learn a shared codebook of discrete tokens and represent each item with the corresponding tokens in the codebook (Wang et al., 2025b; Mei et al., 2025; Jia et al., 2026; Wang et al., 2026; Hou et al., 2023; Wang et al., 2025a; Zheng et al., 2025). For example, TIGER (Rajput et al., 2023) pioneered this direction by utilizing RQ-VAE to quantize item content embeddings into hierarchical semantic IDs. Subsequent work has sought to enrich the tokenization process: Some works (Wang et al., 2024c; Liu et al., 2025c; Wang et al., 2025c; Ye et al., 2025) integrates collaborative signals from pre-trained recommendation models into vector quantization through contrastive alignment, and LC-Rec (Zheng et al., 2024) aligns a large language model with semantic IDs via multi-task learning to improve the model's understanding of discrete tokens. SETRec (Lin et al., 2025) learns order-agnostic identifiers to improve the performance of GR through LLMs.

Nevertheless, these methods follow a two-stage paradigm where the tokenizer and the downstream recommender are optimized separately, leading to objective misalignment between the tokenizer and generator. To bridge this gap, recent efforts have explored end-to-end joint optimization: ETEGRec (Liu et al., 2025a) proposes alternating optimization between the tokenizer and recommender with alignment losses, while DIGER (Fu et al., 2026) introduces differentiable semantic IDs via Gumbel-Softmax to enable direct gradient flow from the recommender back to the tokenizer.

Generation Paradigm. From the perspective of generation, existing GR models can be categorized into autoregressive (AR) models and non-autoregressive (NAR) models. Most existing methods adopt AR generation, which generate tokens sequentially from left to right (Rajput et al., 2023; Wang et al., 2024c; Fang et al., 2026; Qiao et al., 2026). At industrial scale, HSTU (Zhai et al., 2024) introduces hierarchical sequential transducers with significant scalability, and OneRec (Deng et al., 2025) unifies retrieval and ranking within a single generative model with preference alignment. However, the AR paradigm suffers from two inherent limitations: (i) the sequential decoding process is time-consuming, as each token must wait for all preceding tokens to be generated, and (ii) errors made at earlier generation steps propagate irreversibly to subsequent tokens, resulting in error accumulation (Shi et al., 2025).

To address these limitations, NAR models have been proposed (Hou et al., 2025a; Shah et al., 2025; Liu et al., 2026b; Mu et al., 2026). Inspired by discrete diffusion language models (Nie et al., 2026; Li et al., 2025; Liu et al., 2026a), LLaDA-Rec (Shi et al., 2025) introduces a parallel generation paradigm, where all tokens are generated simultaneously conditioned on bidirectional context. This design enables richer inter-token dependency modeling by attending to both preceding and following tokens.

Although GR has shown promising results, existing methods generally focus on improving recommendation accuracy while ignoring the credibility of the recommended content. To fill this gap, in this paper, we propose a novel GR model that explicitly incorporates content credibility into both the tokenization and generation stages.

A.2 Credible Content Recommendation

RSs play a crucial role in helping users quickly locate content they are interested in (Cai et al., 2022, 2023b,a; Bing et al., 2024; Yuan et al., 2025; Zhang et al., 2025; Wang et al., 2020, 2022a). However, RSs have been widely recognized as significant contributors to the spreading of uncredible content, including fake news, rumor, and misinformation, across modern web platforms (Fernandez and Bellogín, 2020; Samuel-Azran and Hayat, 2019). The majority of prior work in this area has centered on two directions: examining whether standard recommendation algorithms amplify uncredible content, and developing mitigation strategies.

Empirical Analysis. The majority of prior work

has centered on examining whether standard recommendation algorithms amplify uncredible content (Fernández et al., 2021; Pathak et al., 2023; Hornig et al., 2024; Pathak and Spezzano, 2024; Pathak, 2024; Fernandez and Bellogín, 2020). For instance, Pathak et al. (Pathak et al., 2023) conduct a systematic evaluation showing that collaborative filtering algorithms disproportionately recommend misinformation to users who have previously engaged with such content, creating self-reinforcing filter bubbles. While these studies provide valuable diagnostic insights, they remain at the analysis level and do not propose concrete algorithmic solutions.

Mitigation Strategies. In contrast, a smaller body of work moves beyond analysis to develop novel mitigation strategies (Wang et al., 2022c, 2024a; Sallami et al., 2023; Ma et al., 2025; Cai et al., 2026b). Wang et al. (Wang et al., 2022c, 2024a) propose veracity-aware news recommendation frameworks that explicitly disentangle items' veracity embeddings from event embeddings and solely utilize event embeddings to reduce the exposure of fake news. Sallami et al. (Sallami et al., 2023) introduce FANAR, a trust-based adaptation of collaborative filtering that excludes untrustworthy neighbors from the recommendation process. More recently, Cai et al. (Cai et al., 2026a) explores to steer diffusion models toward credible content recommendation through two designs: a disentangled diffusion model and the progressively enhancement of credible subspace projection.

Although these mitigation methods have demonstrated effectiveness, they are fundamentally built upon discriminative modeling paradigms that rely on computing embedding similarities between user and item embeddings to make recommendations. As such, they are inherently incompatible with the emerging GR paradigm, which formulates the recommendation task as a sequence generation problem over discrete tokens rather than a matching problem over continuous embeddings. To bridge this gap, in this paper, we propose a novel credibility-aware GR model.

B More Experimental Details

B.1 Baseline Introduction

We compare our proposed `CreGR` with baselines from three categories:

General Recommendation Models

- **GRU4Rec** (Hidasi et al., 2015): A session-based recommendation model that employs Gated Recurrent Units (GRUs) to model sequential user interactions.
- **SASRec** (Kang and McAuley, 2018): A self-attentive sequential recommendation model that leverages the Transformer's self-attention mechanism to capture long-range dependencies in user interaction histories.
- **DreamRec** (Yang et al., 2023b): A diffusion recommender that generates the target item embeddings by applying a continuous diffusion process in the embedding space.
- **PreferDiff** (Liu et al., 2025b): A preference-guided diffusion model for sequential recommendation that incorporates user preference signals into the diffusion process.

Credible Recommendation Models

- **Rec4Mit** (Wang et al., 2022c): A veracity-aware and event-driven personalized news recommendation model designed to mitigate fake news in recommender systems.
- **HDInt** (Wang et al., 2024a): A hierarchical and disentangling interest learning framework for unbiased and true news recommendation.
- **PRISM** (Ma et al., 2025): A diffusion model that bridges user interests and truthfulness for mitigating fake news exposure in recommendations.
- **`Disco`** (Cai et al., 2026a): A disentangled diffusion model that steers the diffusion process towards credible content recommendation.

Generative Recommendation Models

- **TIGER** (Rajput et al., 2023): A generative retrieval model for recommender systems that represents each item with semantic IDs generated by RQ-VAE. It employs Transformer to generate the semantic ID tokens of the target item conditioned on the user's interaction history.
- **LETTER** (Wang et al., 2024c): A generative recommendation model that utilizes items' collaborative signals to guide the tokenization process.
- **LLaDA-Rec** (Shi et al., 2025): A non-autoregressive generative recommendation model based on discrete diffusion. It generates all tokens of the target item in parallel through a discrete diffusion process.

B.2 Evaluation Metric Introduction

We adopt the following evaluation metrics to comprehensively evaluate recommendation accuracy, content credibility, and their balance.

B.2.1 Recommendation Accuracy Metrics

- **HR@N (Hit Rate at N)**: HR@N measures whether the ground-truth item appears in the top-N recommendation list. For a set of test users (i.e., equal to sequences) $\mathcal{U}$, it is defined as:

$$\text{HR@}N = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbf{1}\left(v_u^{gt} \in \mathcal{R}_u^N\right), \quad (18)$$

where v_u^{gt} is the ground-truth next item for user u, $\mathcal{R}_u^N$ is the top-N recommendation list for user u, and $\mathbf{1}(\cdot)$ is the indicator function.

- **NDCG@N (Normalized Discounted Cumulative Gain at N)**: NDCG measures the ranking quality of the recommendation list by assigning higher scores when the ground-truth item appears at higher positions. It is defined as:

$$\text{NDCG@}N = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{\log_2(\text{rank}_u + 1)}, \quad (19)$$

where rank_u denotes the position of the ground-truth item v_u^{gt} in the top-N recommendation list $\mathcal{R}_u^N$. If the ground-truth item does not appear in the list, the NDCG contribution is 0.

B.2.2 Content Credibility Metric

- **CR@N (Credible Rate at N)** (Cai et al., 2026a): CR@N measures the proportion of credible items in the top-N recommendation list. A higher CR@N indicates that the RSs exposes users to fewer uncredible items. It is defined as:

$$\text{CR@}N = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{\left|\mathcal{R}_u^N \cap \mathcal{I}_{cre}\right|}{|\mathcal{R}_u^N|}, \quad (20)$$

where $\mathcal{I}_{cre}$ is the set of credible items and $\mathcal{R}_u^N$ is the top-N recommendation list for user u. CR@N ranges from 0 to 1, where 1 indicates all recommended items are credible.

B.2.3 Combined Metric

- **HC@N** (Cai et al., 2026a): HC@N is a combined metric that evaluates the balance between recommendation accuracy (HR@N) and content credibility (CR@N). It is defined as:

$$\text{HC@}N = \frac{2 \times \text{HR@}N \times (\frac{\text{CR@}N}{2})}{\text{HR@}N + (\frac{\text{CR@}N}{2})}. \quad (21)$$

A high HC@N score requires both high recommendation accuracy and high content credibility, making it a comprehensive indicator of overall recommendation quality. Specifically, CR@N is scaled by a factor of $\frac{1}{2}$ to align its magnitude with HR@N, ensuring that neither metric disproportionately dominates the combined score.

We evaluate all metrics at $N \in \{5, 10\}$, following standard practices in sequential recommendation (Kang and McAuley, 2018).

B.2.4 Collision Rate

Collision rate measures the proportion of items that share identical SIDs (i.e., identical discrete token sequences) with other items:

$$\text{Collision Rate} = 1 - \frac{|\mathcal{SID}_{\text{unique}}|}{|\mathcal{I}|}, \quad (22)$$

where $|\mathcal{SID}_{\text{unique}}|$ is the number of distinct SIDs and $|\mathcal{I}|$ is the total number of items. A lower collision rate indicates better codebook utilization and finer-grained item discrimination.

B.3 More implementation details

Item Content. Items in PolitiFact and GossipCop datasets consist of news articles, for which we use titles and descriptions to derive content representations. In MHMisinfo dataset, although items are videos, only titles and descriptions are available; we therefore rely solely on these fields for content representation. We encode each item's title and description into dense embeddings using LLaMA2-7B (Touvron et al., 2023).

Tokenization Stage. The projection matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ have dimensions $(8092, 64)$ and $(64, 64)$, respectively, each followed by layer normalization. The dimension of collaborative embeddings is 64. The resulting content and collaborative embeddings are concatenated to form the final item embedding. The $\mathbf{Encoder}(\cdot)$ follows the architecture $\{128 \rightarrow 512, \text{ReLU}, 512 \rightarrow 256, \text{ReLU}, 256 \rightarrow 64\}$, and the $\mathbf{Decoder}(\cdot)$ mirrors it symmetrically as $\{64 \rightarrow 256, \text{ReLU}, 256 \rightarrow 512, \text{ReLU}, 512 \rightarrow 128\}$. Prior to computing $\mathcal{L}_{Cred}$ (Equation 11), we pass r_i through a two-layer MLP with architecture $\{64 \rightarrow 64, \text{ReLU}, 64 \rightarrow 64\}$, which mitigates objective conflict between the quantization loss and credibility regularization. Learning rate of the credibility-aware tokenizer is searched in range {5e-5, 1e-4, 5e-4, 1e-3, 5e-3}.

Generation Stage. We adopt a bidirectional Transformer as the mask predictor. The numbers of

Table 7: Recommendation accuracy comparison between CreGR and simply removing uncredible interactions before training.

Method	HR@5	HR@10	NDCG@5	NDCG@10
PolitiFact				
CreGR	0.2762	0.3920	0.1949	0.2318
Simply removing	0.2319	0.3362	0.1661	0.1998
Improv.	-16.04%	-14.23%	-14.78%	-13.81%
GossipCop				
CreGR	0.5855	0.6384	0.4984	0.5157
Simply removing	0.4799	0.5201	0.4138	0.4269
Improv.	-18.04%	-18.53%	-16.97%	-17.22%
MHMisinfo				
CreGR	0.2240	0.2876	0.1480	0.1683
Simply removing	0.1713	0.2282	0.1152	0.1334
Improv.	-23.53%	-20.65%	-22.16%	-20.74%

layers, attention heads, and KV heads are set to $\{2, 8, 8\}$, $\{6, 8, 8\}$, and $\{2, 8, 8\}$ for PolitiFact, GossipCop, and MHMisinfo, respectively. The beam width is fixed at 50 for all datasets. For fair comparison, the embedding dimension d is set to 64 for all methods except continuous diffusion-based baselines (DreamRec (Yang et al., 2023b), PreferDiff (Liu et al., 2025b), PRISM (Ma et al., 2025), and Disco (Cai et al., 2026a)), which require an embedding dimension of 3072 to achieve competitive performance. This high embedding dimension requirement was also explicitly noted in prior work (Liu et al., 2025b; Cai et al., 2026a).

All experiments are conducted on an NVIDIA A40 GPU with 48GB of memory.

C Empirical results of removing interactions between users and uncredible items before training

As mentioned in CH2, simply removing uncredible items from the training data and retraining the generation model inevitably degrades recommendation accuracy. Here we provide the empirical evidence. As shown in Table 7, simply removing uncredible interactions before training significantly degrades recommendation accuracy (up to 23.53%). This is because valuable information (e.g., topics) in the uncredible items that reflect users' preference is also removed.

D Evidence supporting the design of our credibility-aware tokenizer

Prior to computing $\mathcal{L}_{\text{Cred}}$, we project $\mathbf{r}_i$ into a dedicated subspace via a two-layer MLP to obtain $\hat{\mathbf{r}}_i$, rather than applying credibility regularization directly on $\mathbf{r}_i$. This design choice is motivated by the need to mitigate objective conflict between the quantization loss and credibility regularization. Specifically, directly applying regularization on $\mathbf{r}_i$ would cause items with distinct topical content but the same credibility label to collapse toward each other in the embedding space, as this regularization directly pushes the latent representations of items with the same credibility level closer together regardless of their topical differences. This collapse would substantially increase the collision rate—the proportion of distinct items assigned identical token sequences—rendering the token representations insufficiently discriminative for accurate recommendation. To verify this, we conduct an ablation experiment comparing direct application of $\mathcal{L}_{\text{Cred}}$ on $\mathbf{r}_i$ against our MLP-projected $\hat{\mathbf{r}}_i$. As shown in Table 8, bypassing the MLP projection leads to a significant rise in collision rate, empirically confirming the necessity of this design.

Table 8: Collision rate comparison between using $\mathbf{r}_i$ and $\hat{\mathbf{r}}_i$ for $\mathcal{L}_{\text{Cred}}$ on CreGR (soft).

	PolitiFact	GossipCop	MHMisinfo
$\mathbf{r}_i$	48.94%	93.87%	84.01%
$\hat{\mathbf{r}}_i$ (ours)	0%	0%	0.1%

E Evidence supporting the design of our accuracy-preserving and credibility-oriented generator

Our accuracy-preserving and credibility-oriented generator selectively reduces the masking probabilities of the last tokens of uncredible items. This design is motivated by the hierarchical nature of RQ-VAE, whereby fine-grained factual information is more likely to be encoded in the last few tokens after coarser semantic patterns have been captured by the preceding levels. To validate this, we conduct experiments by varying the selected token position from layer 1 to layer 4 (ours). As shown in Table 9, selecting tokens from earlier layers to reduce their masking probabilities leads to consistent degradation in recommendation credibility, confirming that fine-grained factual information is indeed concentrated in the last tokens. This validates the rationale of our design choice to apply masking probability reduction exclusively to the last tokens of uncredible items.

Table 9: The effect of different selected token positions in the generation stage on recommendation credibility (CR@10) of CreGR (soft).

Token position	PolitiFact	MHMisinfo
1	0.9241	0.9101
2	0.9598	0.9026
3	0.9741	0.9121
4 (ours)	0.9826	0.9315

F Case Study

To intuitively illustrate the effectiveness of CreGR, we present a case study from the GossipCop dataset in Figure 4. In case 1, the user has historically interacted with four fake news articles, covering topics such as celebrity relationships (e.g., "Fergie separated from Josh Duhamel", "Chris Pratt files for divorce from Anna Faris") and celebrity family affairs (e.g., "Kylie Jenner admits jealousy over Kendall", "Katherine Jackson bowing out of guardian biz"). Despite all historical interactions being fake news, our model successfully captures the user's underlying topical preferences — celebrity gossip, particularly regarding relationship changes and family dynamics — while steering the recommendations entirely toward credible content. Specifically, the top-ranked recommendation (Item 3631) is a *true news* article about "Chris Pratt and Anna Faris are officially divorced", which precisely matches the user's interest in the same celebrity couple but provides verified factual reporting instead of fabricated claims. Similarly, the second-ranked item (Item 3674) covers a celebrity separation topic that aligns with the user's preference pattern, and the fourth-ranked item (Item 4047) satisfies the user's interest in Kylie Jenner — all being credible content. This case demonstrates that CreGR effectively mitigate uncredible content while still preserving users' preference-related information for improving recommendation accuracy.

G Complexity Analysis

Compared to standard RQ-VAE quantization and discrete diffusion training, the additional computational overhead introduced by CreGR stems from two components: credibility regularization and the soft weighting strategy. For credibility regularization, the complexity of computing the distance between each item's latent representations and the centroids of credible and uncredible clusters is $O(|\mathcal{I}| \cdot d)$, where $|\mathcal{I}|$ is the number of items and d is the codeword embedding dimension. For the soft weighting strategy, the complexity of identifying uncredible last tokens and computing w_i consists of two terms: $O((|\mathcal{S}_{\mathcal{H}}| + |v_n|) \cdot L)$ for scanning token sequences, where $\mathcal{S}_{\mathcal{H}}$ and v_n denote the token sequences of a user's historical interactions and the target item respectively, and L is the codebook depth; and $O(|\mathcal{I}_{\text{unc}}| \cdot d)$ for computing weights over uncredible items, where $|\mathcal{I}_{\text{unc}}|$ is the number of uncredible items. All additional terms scale linearly with the input size and are negligible relative to the dominant cost of standard RQ-VAE and discrete diffusion training, confirming that CreGR maintains high computational efficiency.

Computational efficiency of centroid-based credibility regularization. A key design advantage of our credibility regularization (Eq. 11) is its computational efficiency. By computing distances to the centroids of credible and uncredible item clusters rather than to every individual item, we reduce the complexity of the credibility loss from $O(|\mathcal{I}|^2 \cdot d)$ (pairwise) to $O(|\mathcal{I}| \cdot d)$ (centroid-based).

H Algorithm

We present the training procedure of CreGR in Algorithm 1 (tokenization stage) and Algorithm 2 (generation stage).

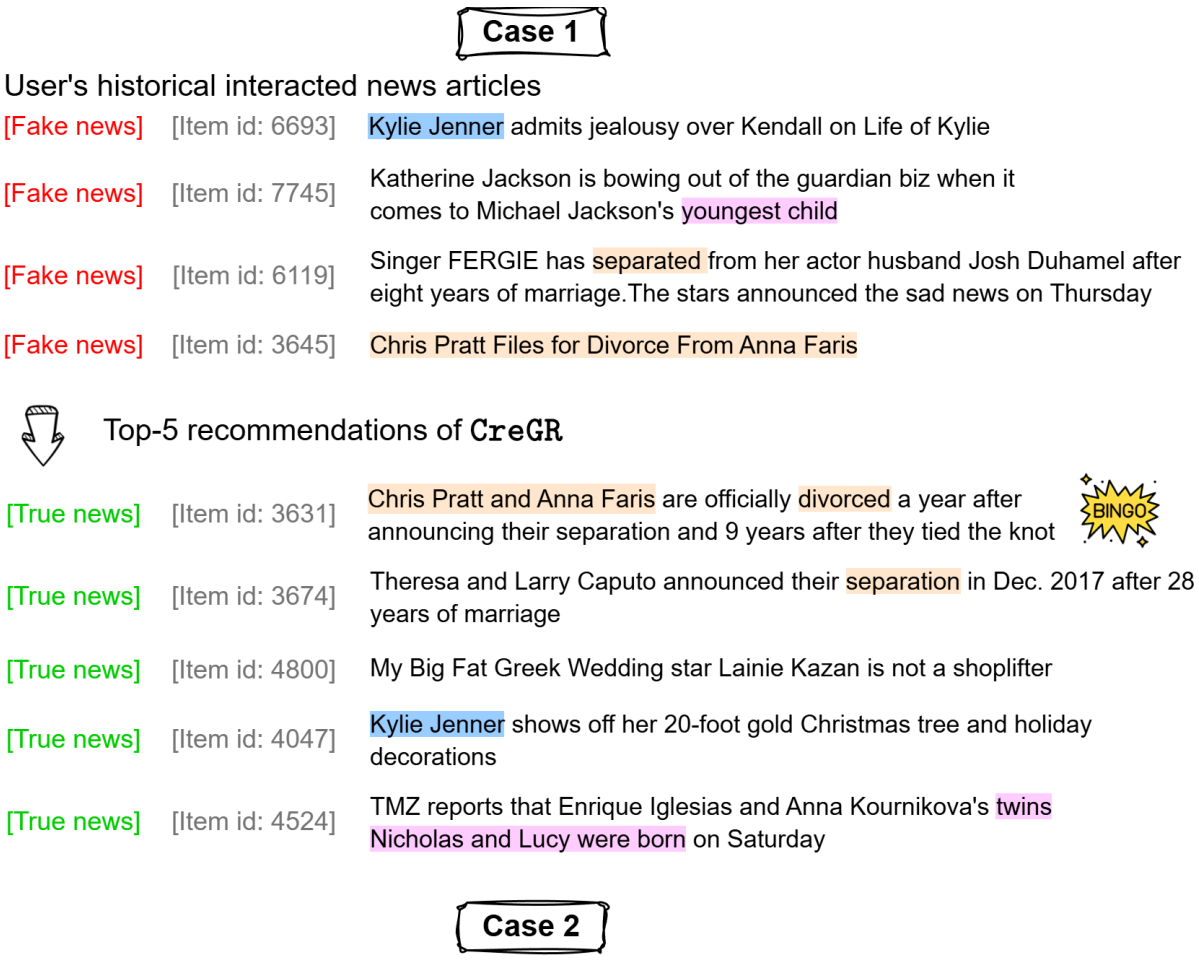

Figure 4: A case study illustrating the effectiveness of CreGR for mitigating uncredible items (fake news) on GossipCop dataset. Items sharing the same background color belong to the same topic cluster. "BINGO" indicates that the news articles is the one the user actually interacted with next.

Algorithm 1 Credibility-aware Tokenizer

Require: Item set $\mathcal{V}$, credible item set $\mathcal{I}_{cre}$, uncredible item set $\mathcal{I}_{unc}$, initialized codebooks $\{\mathcal{C}_1, \ldots, \mathcal{C}_L\}$, the title, description, and collaborative embeddings of each item — $\mathbf{e}_i^{tit}$, $\mathbf{e}_i^{des}$ and $\mathbf{e}_i^{col}$.

Ensure: Semantic ID (a sequence of discrete tokens) of each item: $[c_{i,1}, c_{i,2}, \cdots, c_{i,L}]$. Well-trained codebooks $\{\mathcal{C}_1, \ldots, \mathcal{C}_L\}$.

for each training epoch **do**
  Compute projected embeddings: $\mathbf{z}_i^{con} = [\mathbf{e}_i^{tit} \| \mathbf{e}_i^{des}]\mathbf{W}_1,\ \mathbf{z}_i^{col} = \mathbf{e}_i^{col}\mathbf{W}_2$ (Eq. 3)
  Form item embedding: $\mathbf{z}_i = \mathbf{z}_i^{con} \| \mathbf{z}_i^{col}$
  Encode: $\mathbf{r}_i = \mathbf{Encoder}(\mathbf{z}_i)$ (Eq. 4)
  Initialize $\mathbf{r}_{i,1} = \mathbf{r}_i$
  for $l = 1$ to L **do**
    Assign token: $c_{i,l} = \arg\min_{k \in \{1,\ldots,K\}} \|\mathbf{r}_{i,l} - \mathbf{e}_{l,k}\|_2^2$ (Eq. 5)
    Update residual: $\mathbf{r}_{i,l+1} = \mathbf{r}_{i,l} - \mathbf{e}_{c_{i,l}}$ (Eq. 6)
  end for
  Reconstruct: $\hat{\mathbf{z}}_i = \mathbf{Decoder}\left(\sum_{l=1}^{L} \mathbf{e}_{c_{i,l}}\right)$ (Eq. 7)
  Compute RQ-VAE loss $\mathcal{L}_{RQ-VAE}$ (Eq. 8).
  Compute credibility regularization loss $\mathcal{L}_{Cred}$ (Eq. 11)
  Update parameters by minimizing $\mathcal{L}_{Tokenization} = \mathcal{L}_{RQ\text{-}VAE} + \mathcal{L}_{Cred}$ (Eq. 12)
end for

Algorithm 2 Accuracy-preserving and Credibility-oriented Generator

Require: Token sequences $S_{\mathcal{H}} = [c_{1,1}, \cdots, c_{1,L}; \cdots; c_{n-1,1}, \cdots, c_{n-1,L}]$ of user interaction histories, target item tokens $v_n = [c_{n,1}, \cdots, c_{n,L}]$, uncredible item set $\mathcal{I}_{unc}$, hyperparameter λ.

Ensure: Trained mask predictor $\mathcal{P}_\theta$.

for each training epoch **do**
  Sample masking ratio $r \sim U[0, 1]$
  for each token $c_{i,l}$ in $S_{\mathcal{H}}$ and v_n **do**
    if $c_{i,L}^r = [\mathtt{M}]$ **then**
      Set masking probability $p_{\mathtt{M}}(c_{i,l}^r) = 0$
    else
      if $v_i \in \mathcal{I}_{unc}$ **and** $l = L$ **then**
        if Hard weighting strategy **then**
          $w_i = 0$
        else
          $w_i = \sigma(\mathtt{D}(\mathbf{e}_{c_{i,L}}, \mathtt{Cen}(\mathbf{e}_{c_{j,L}})) - \mathtt{D}(\mathbf{e}_{c_{i,L}}, \mathtt{Cen}(\mathbf{e}_{c_{i',L}})))$ (Eq. 14)
        end if
        Set masking probability $p_{\mathtt{M}}(c_{i,l}^r) = w_i \cdot r$
      else
        Set masking probability $p_{\mathtt{M}}(c_{i,l}^r) = r$
      end if
    end if
    Mask $c_{i,l}$ with probability $p_{\mathtt{M}}(c_{i,l}^r)$ to obtain $c_{i,l}^r$
  end for
  Compute historical credibility loss $\mathcal{L}_{His}$ (Eq. 15)
  Compute target credibility loss $\mathcal{L}_{Tar}$ (Eq. 16)
  Update $\mathcal{P}_\theta$ by minimizing $\mathcal{L}_{Generation} = \lambda\mathcal{L}_{His} + \mathcal{L}_{Tar}$ (Eq. 17)
end for