AlignDiff: Gradient-Semantic Aligned Diffusion Model for Sequential Recommendation

Anonymous ACL submission

Abstract

Sequential recommendation predicts users' following interactions by modeling dynamically evolving preferences. Recently, diffusion-based sequence recommenders have improved recommendation accuracy through distribution modeling, but they face the dual problems of distribution alignment and semantic guidance distortion. We propose AlignDiff, a Gradient-Semantic Aligned Diffusion Model for Sequential Recommendation addressing these issues via two innovations: (1) By combining a denoising predictor and an energy function network into a siamese denoising network, this network learns the gradient differences between distributions using cross-entropy loss and gradient score-matching loss, explicitly constraining the predicted denoising distribution to fit the ground-truth data distribution; (2) Multi-Level conditional guidance fusing sequence embeddings with attention-derived deep semantic features, efficiently modeling user preferences and correcting the problem of distortion in guidance conditions by mining deep semantic information in user interaction sequences, which guides the model to denoise in the direction of the correct denoising trajectory. Experiments demonstrate that AlignDiff significantly outperforms all baselines on three datasets [1].

1 Introduction

Sequential Recommendation (SR) models user preferences through historical interactions to predict next-item interests. Early methods focus on the temporal correlation of interaction sequences. For example, traditional recurrent neural networks (RNNs) are limited by the vanishing gradient problem in long sequence modeling, leading to gated architectures such as long short-term memory (LSTM) networks (Hidasi et al., 2015; Hidasi and Karatzoglou, 2018). With the advent of self-attention mechanisms (Lin et al., 2020; Tang et al., 2025), dependency modeling has been completely transformed; for example, SASRec (Kang and McAuley, 2018) (Self-Attentive Sequential Recommendation), which employs unidirectional self-attention to learn item transition patterns, and BERT4Rec (Sun et al., 2019), utilizing bidirectional Transformer encoders to jointly model past and future contextual signals for improved next-item prediction.

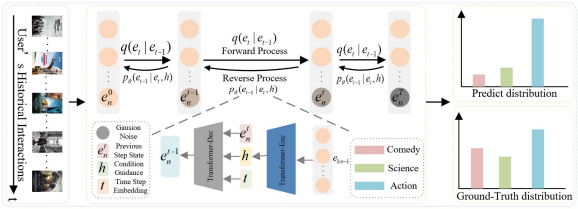

Figure 1: The diffusion model's generated prediction (action movie preference) deviates from the ground-truth distribution (action + comedy), revealing bias between denoised and real user preferences.

While conventional sequential recommendation methods excel with static representations, their limited capacity to model complex preference distributions drives the adoption of generative models—GANs, VAEs, and diffusion models. Unlike discriminative methods, these generative approaches simulate data generation through adversarial training (GANs), latent regularization (VAEs), or iterative denoising (diffusion models), better capturing attribute distributions and preference diversity. Diffusion models particularly surpass traditional architectures: GAN-based methods SeqGAN (Yu et al., 2017), SparseEnNet (Chen et al., 2024), suffer gradient instability, while VAE variants (e.g., ACVAE (Xie et al., 2021), DistVAE (Li et al., 2023), Meta-SGCL (Hao et al., 2024)) face posterior collapse. In contrast, diffusion models like Diff4Rec (Wu et al., 2023) leverage curriculum denoising and fine-grained interaction model-

[1] https://anonymous.4open.science/r/AlignDiff_run

ing to iteratively recover ground-truth data distributions, overcoming instability and collapse issues. This establishes diffusion models as robust frameworks for complex sequential dependencies.

Despite diffusion models' success in sequential recommendation (Wu et al., 2023; Li et al., 2024; Yang et al., 2023), two key limitations remain: 1. Trajectory Deviation from Global Distribution Mismatch: Traditional diffusion models rely on reconstruction loss that optimizes single-step local errors, failing to ensure global alignment between predicted and ground-truth denoising trajectories. Initial prediction misalignment and error accumulation across iterative steps cause the final generated distribution to diverge from ground-truth user preferences (e.g., concentrated action movie predictions vs. actual comedy/action preferences in Figure 1). 2. Semantic Distortion in Conditional Guidance: Shallow interaction sequence modeling compresses multi-dimensional user interests into single dominant semantics. This creates distortion guidance signals (e.g., prioritizing action movies while ignoring comedy/sci-fi preferences in Figure 1), derailing denoising trajectories from true semantic patterns. These biases progressively widen the gap between generated and ground-truth representations, degrading recommendation accuracy.

We propose AlignDiff, a Gradient-Semantic Aligned Diffusion Model for Sequential Recommendation, addressing distribution alignment and semantic guidance distortion in the sequential recommendation to overcome the challenges in current diffusion-based SR models. Specifically, for the problem of conditional semantic distortion, a multi-level conditional guidance module is constructed. The module captures users' preferences in different semantic spaces through multi-head orthogonal attention and further integrates sequence embedding to generate semantically fidelity multi-level conditional guidance. In the model's inference stage, a Classifier-Free Guidance strategy is adopted, which adjusts the guidance ratio parameter and linearly combines conditional and non-conditional prediction results. To address the distribution alignment problem, we propose a siamese denoising network based on a Transformer denoising predictor and an energy function. The network first iteratively denoises the input to generate an initial prediction distribution using the denoising predictor. Then, the energy function network learns the gradient information of the distribution and the gradient distribution information of the labels, constructing a gradient score-matching loss. This loss is combined with cross-entropy loss to jointly constrain the model's denoising path, performing single-step gradient correction on the preliminary prediction items to optimize the gradient difference between the generated distribution and the ground-truth distribution. Our main contributions are as follows:

- To the best of our knowledge, this is the first diffusion sequential recommendation model explicitly aligning denoising gradients to constrain distribution shifts.
- We propose a multi-level conditional guidance module to improve user preference modeling and solve the semantic distortion problem in conditional guidance.
- Extensive experiments showing state-of-the-art performance across three real-world datasets, with HR@K and NDCG@K improvements up to 20.4% and 11.3%.

2 Related Work

2.1 Sequential Recommendation

Early sequential recommendation (SR) models evolved from first-order Markov chains (Rendle et al., 2010) (predicting following items based on immediate predecessors but struggling with long-term dependencies) to deep architectures like GRUs (GRU4Rec) (Hidasi et al., 2015) and CNNs (Caser) (Tang and Wang, 2018) for dynamic pattern modeling. Attention-based models (e.g., BERT4Rec (Sun et al., 2019), TiSASRec (Li et al., 2020)) improved context modeling, while SH-Rec (Ma and Gan, 2024) combined hierarchical user intent and sequence behavior via LSTM/self-attention layers with multi-task learning. However, such models rely on static item embeddings, limiting adaptability to dynamic user preferences or item characteristics.

2.2 Diffusion Model in Recommendation

Diffusion models generate data by iteratively denoising corrupted inputs, offering stable training compared to GANs/VAEs. While effective in continuous spaces (e.g., images), adapting them to discrete data like text or user interactions requires specialized techniques: Diffusion-LM (Li et al., 2022) employs probabilistic discrete diffusion to preserve semantic coherence in text generation. In sequential recommendation, DiffuRec (Li et al.,

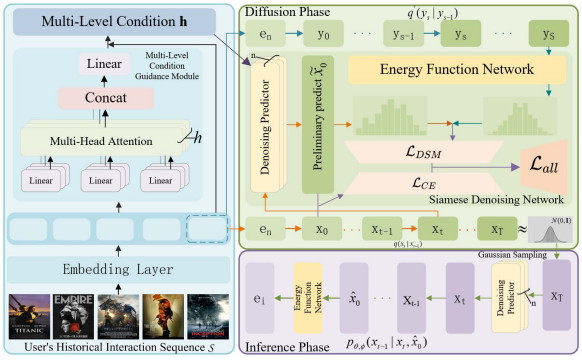

Figure 2: The AlignDiff model integrates a multi-level conditional guidance module (generates Semantic fidelity multi-level conditional signals to steer denoising) and a siamese denoising network (removes noise from input data and outputs final predictions).

2024) injects noise (e.g., masking) into interaction sequences and reconstructs them via reverse diffusion, augmented with contextual signals (time, categories). Extensions like PDRec (Ma et al., 2024) integrate time-aware decay to refine temporal dynamics, while CaDi-Rec (Cui et al., 2024) leverages bidirectional diffusion and contrastive learning to preserve user preferences in augmented views. However, challenges persist: (1) Denoising trajectories often deviate from true user preferences due to biased noise removal; (2) semantic misalignment in guidance signals limits accurate modeling of multidimensional user interests, hindering recommendation precision.

3 Method

In this section, we will introduce the details of the proposed AlignDiff model, including an in-depth overview and the construction of its core components and loss functions.

3.1 Overview

Sequential Recommendation. Let $\mathcal{U}$ represent a set of users and $\mathcal{I}$ denote a collection of discrete items in the dataset. Each user $u \in \mathcal{U}$ has an interaction history truncated to the most recent n items, forming an input sequence $\mathcal{S} = [i_1^u, i_2^u, ..., i_{n-1}^u, i_n^u]$, which is ordered chronologically according to the user's interactions, where $i_j^u \in \mathcal{I}$ denotes the j-th interacted item, n denotes the sequence length, and the total number of items interacted with by user u. Each item $i \in \mathcal{I}$ is associated with a learnable embedding vector $\mathbf{e}_i \in \mathbb{R}^d$. Thus, the sequence is represented as $\mathbf{E}^u = [e_1^u, e_2^u, ..., e_{n-1}^u, e_n^u]$. Sequential recommendation aims to predict the next item i_{n+1}^u by ranking candidate items based on the user's learned preferences from $\mathbf{E}^u$.

The AlignDiff method we propose is shown in Figure 2 and mainly consists of a multi-level conditional guidance module and a Siamese denoising network. The Siamese denoising network is applied in the diffusion and inference phases. The main objective of the diffusion phase is to construct semantically faithful conditional signals based on multi-level conditional guidance, which are then used to guide the denoising process of the Siamese network. In the inference phase, the siamese network trained in the diffusion phase performs denoising starting from data drawn from a standard normal distribution, with the Classifier-Free Guidance method applied throughout.

3.2 Multi-Level Condition Guide Module

We propose a method based on deep interaction sequence multi-level semantic mining to improve the fidelity of condition signals. First, interaction sequences are embedded into continuous vectors to enable gradient optimization and noise injection in the latent space, and a position encoding strategy is adopted to preserve time dependency.

$$\mathbf{X} = \mathbf{E}(\mathcal{S}) + \mathbf{P} \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{|\mathcal{I}| \times d}$ is the item embedding matrix, $\mathbf{P} \in \mathbb{R}^{n \times d}$ is the learnable positional encoding, and n is the sequence length.

Subsequently, the module uses multi-head orthogonal attention to achieve diversified semantic sub-space modeling and avoid capturing redundant features. It also uses a causal masking mechanism to avoid contact with future data. This ensures comprehensive and non-overlapping preference extraction in the semantic sub-space.

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{X}\mathbf{W}_Q, \mathbf{X}\mathbf{W}_K, \mathbf{X}\mathbf{W}_V \tag{2}$$

$$\text{orthoAttention}(\mathbf{X}) = \tag{3}$$

$$\sum_{i=1}^{h} \text{Softmax}\left(\frac{(\mathbf{Q}\mathbf{U}_i)(\mathbf{K}\mathbf{V}_i)^{\mathrm{T}}}{\sqrt{d/h}} + \mathbf{M}_{causal}\right)\mathbf{V}$$

where h is the number of heads, d/h is the head dimension, and $\mathbf{M}_{causal}$ is the lower-triangular mask that prevents future leakage. The orthogonal projection matrix satisfies $\mathbf{U}_i^{\mathrm{T}}\mathbf{U}_j = \delta_{ij}\mathbf{I}$,

$\mathbf{V}_i^{\mathrm{T}}\mathbf{V}_j = \delta_{ij}\mathbf{I}$, $\forall i, j \in [1, h]$. $\mathbf{U}_i$ and $\mathbf{V}_j$ represent a set of orthogonal matrices, δ_{ij} is the Kronecker function that takes the value 1, if i=j and 0 otherwise.

Then, the output of the attention mechanism Y at the last time step is extracted as the user preference representation. A residual connection is adopted between the sequence embedding and the attention output to enhance representation learning ability and enrich feature representation. Finally, the hidden state at the last time step $\mathbf{h}$ is extracted from the optimized output H as a multi-level conditional signal for semantic fidelity.

$$\mathbf{H} = \text{LayerNorm}\big(\mathbf{Y}+\text{ReLU}(\mathbf{Y}\mathbf{W}_1+\mathbf{b}_1)\mathbf{W}_2+\mathbf{b}_2\big) \tag{4}$$

3.3 Siamese Denoising Network

This module consists of a denoising predictor $f_\theta(\cdot)$ and an energy function network $S_\phi(\cdot)$, which work together in the diffusion phase and inference phase to ensure distribution alignment and semantic fidelity.

3.3.1 Diffusion Phase

In the diffusion phase, the model reconstructs target embedding $\mathbf{e}_n$ via reverse denoising guided by sequence $\mathcal{S}$. We initialize $\mathbf{x}_0$ from $\mathbf{e}_n$ through a single diffusion step $q(\mathbf{x}_0|\mathbf{e}_n) = \mathcal{N}(\mathbf{x}_0; \sqrt{\alpha_0}\mathbf{e}_n, (1-\alpha_0)\mathbf{I})$. During the diffusion phase, we adapt an adaptive truncation mechanism (Ho et al., 2020) tailored for sequence recommendation to adjust the amount of noise injected dynamically.

$$\beta_t = \begin{cases} \beta_{\text{min}} + \frac{t}{T}(\beta_{\text{max}} - \beta_{\text{min}}) & t \le \tau \\ \frac{\beta_{\text{max}}}{10} & t > \tau \end{cases} \tag{5}$$

The adaptive noise schedule employs $\beta_{\text{min}} = 0.1$, $\beta_{\text{max}} = 0.9$, and truncation threshold $\tau = 0.8T$ to balance exploration and semantic consistency. Early stages ($t \le 0.8T$): High noise ($\beta_t \approx 0.9$) enables broad item space exploration. Later stages ($t > 0.8T$): Low noise ($\beta_t \approx 0.09$) preserves semantic coherence. During training, t is uniformly sampled from $1, ..., \mathrm{T}$ for stochastic denoising learning. This linear transition ensures smooth noise scaling while preventing distributional abruptness. Noise accumulation follows:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \quad (\bar{\alpha}_t = \textstyle\prod_{s=1}^{t}(1-\beta_s)) \tag{6}$$

The noisy item embedding $\mathbf{x}_t$ is fused with historical interactions to form adjusted embeddings $\mathbf{Z}^u = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n]$, which are fed into the Transformer-based denoising predictor $f_\theta(\cdot)$. The predictor employs stacked self-attention layers with residual connections, augmented by Timestep embeddings $\mathbf{d}_t$ to track diffusion progress and Element-wise fusion of guidance signal $\mathbf{h}$ and Gaussian sampling λ_i to prevent preference oversmoothing, generates the initial denoised prediction $\tilde{\mathbf{x}}_0$.

$$\begin{aligned} \tilde{\mathbf{x}}_0 &= f_\theta(\mathbf{Z}^u) = \text{Transformer}([\mathbf{z}_0, \mathbf{z}_1, ..., \mathbf{z}_n]) \\ \mathbf{z}_i &= \lambda_i \odot (\mathbf{x}_t + \mathbf{d}_t) + (1 + \mathbf{h}) \odot \mathbf{e}_i \end{aligned} \tag{7}$$

where $\mathbf{e}_i$ denotes the item embedding, $\mathbf{d}_t$ represents the diffusion step embedding at timestep t, and $\mathbf{h}$ serves as the multi-level conditional guidance signal. The stochastic weighting factor $\lambda_i \sim \mathcal{N}(\delta, \delta)$ introduces controlled randomness while preserving semantic coherence through element-wise product $\odot$.

After obtaining the initial prediction item representation $\tilde{\mathbf{x}}_0$, in order to further constrain the data distribution of $\tilde{\mathbf{x}}_0$ fit the ground-truth data distribution. When applying noise to the target item $\mathbf{e}_n$, we apply a small noise perturbation $\beta_s^{'} = \gamma\beta_s$ to the target item, where γ is used to control the noise ratio. Here, to avoid confusion, we use $\mathbf{y}$ to denote the target project embedded representation. Similarly, going from $\mathbf{e}_n$ to $\mathbf{y}_0$ is the same process as going from $\mathbf{e}_n$ to $\mathbf{x}_0$. Therefore, the target item after adding noise is as follows:

$$\begin{aligned} q(\mathbf{y}_{1:s}|\mathbf{y}) &= \prod_{s=1}^{s} q(\mathbf{y}_s|\mathbf{y}_{s-1}) \\ q(\mathbf{y}_s|\mathbf{y}_{s-1}) &= \mathcal{N}(\mathbf{y}_s; \sqrt{1-\beta_s^{'}}\mathbf{y}_s, \beta_s^{'}\mathrm{I}) \end{aligned} \tag{8}$$

where $\bar{\alpha}_s^{'} = \prod_{s=1}^{t}(1-\beta_s^{'})$. Similarly, to be able to obtain samples of y_n at any step size, let $\alpha_s^{'} = 1-\beta_s^{'}$, $\bar{\alpha}_s^{'} = \prod_{i=1}^{s}\alpha_i^{'}$. Then, we get:

$$\begin{aligned} q(\mathbf{y}_s|\mathbf{y}_0) &= \mathcal{N}(\mathbf{y}_s; \sqrt{\bar{\alpha}_s^{'}\mathbf{y}_0}, (1-\bar{\alpha}_s^{'})\mathrm{I}) \\ \mathbf{y}_s &= \sqrt{\bar{\alpha}_s}\mathbf{y}_0 + \sqrt{1-\bar{\alpha}_s}\epsilon_y, \epsilon_y \sim \mathcal{N}(0, \mathrm{I}) \end{aligned} \tag{9}$$

The energy function $S_\phi(\cdot)$ refines the denoising process by learning gradient corrections between the noise-perturbed target $\mathbf{y}_s$ and initial prediction $\tilde{\mathbf{x}}_0$. To preserve gradient integrity from the pre-denoised inputs, $S_\phi(\cdot)$ employs simplified activation functions and constructs an implicit gradient

field via score-matching loss. This aligns the predicted distribution with the ground-truth by minimizing the fllowing:

$$\zeta_{\tilde{\mathbf{x}}_0} = S_\phi(\tilde{\mathbf{x}}_0) = \mathbf{W}_n \cdot \text{Dropout} \\ (\text{GELU}(L_{n-1}(\cdots \text{GELU}(L_1(\tilde{\mathbf{x}}_0)) \cdots))) + \mathbf{b}_n \tag{10}$$

The energy function $S_\phi(\cdot)$ employs an MLP architecture (vs. Transformer) with three key design principles: (1) GELU activation for stable second-order optimization, (2) bottleneck structure ($d \to 2d \to 1$) to prevent overfitting while preserving preference signals, and (3) Dropout (rate=0.1) for generalization. This simplified design aligns with its core objective: learning first-order gradient differences between generated and ground-truth distributions rather than modeling high-order interactions.

3.3.2 Inference Phase

In the inference phase, the model starts from a standard normal distribution and undergoes T iterations of denoising to reconstruct the target project representation $\mathbf{e}_n$. In this phase, the denoising predictor first undergoes T iterations of denoising to generate a preliminary prediction of the project representation $\tilde{\mathbf{x}}_0$. However, this distribution deviates from the ground-truth distribution. The energy function is refined through gradient correction to generate the final prediction representation $\hat{\mathbf{x}}_0$. In addition, ClassFier-Free Guidance is used to achieve a balance between personalization and generalization capabilities.

$$\tilde{\mathbf{x}}_0 = f_\theta(\mathbf{Z}_{\mathbf{x}_t}) = (1+w) f_\theta(\mathbf{Z}_{\mathbf{x}_t}) - w f_\theta(\mathbf{Z}_{\mathbf{x}_t}) \\ \mathbf{x}_{t-1} = \tilde{\mu}_t(\mathbf{x}_t, \hat{\mathbf{x}}_0) + \tilde{\beta}_t \epsilon^{'} \tag{11}$$

During inference, the denoising process generates $\mathbf{x}_{t-1}$ using the formula $\tilde{\mu}_t(\mathbf{x}_t, \hat{\mathbf{x}}_0)$, which combines the refined estimate $\hat{\mathbf{x}}_0$ and the current noisy state $\mathbf{x}_t$, along with noise scaling $\tilde{\beta}_t$ and Gaussian noise $\epsilon^{'}$. This iterative refinement repeats until $\tilde{\mathbf{x}}_0$ is obtained. A hyperparameter w modulates the multilevel guidance signal $\mathbf{h}$: larger w prioritizes personalized denoising (via $\mathbf{h}$) over generalization (represented by ϕ) but risks degrading output quality. The energy function $S_\phi(\cdot)$ then performs a single-step gradient correction on $\tilde{\mathbf{x}}_0$, removing residual noise to yield the ground-truth target representation $\hat{\mathbf{x}}_0 = \tilde{\mathbf{x}}_0 - S_\phi(\tilde{\mathbf{x}}_0)$. Finally, $\hat{\mathbf{x}}_0$ is mapped to discrete recommendations by computing inner-product similarity scores with all candidate items, with the highest-scoring item $\mathbf{e}_i$ selected as the user's recommendation.

3.4 Loss Function

In the diffusion process, we first sample diffusion steps t uniformly from $[1, T]$ (where T is the total steps), then perform reverse denoising by decrementing the step-index from t to 1, feeding step embeddings, sequence item distributions, and conditional guidance into the predictor $f_\theta(\cdot)$ for inference. While standard diffusion models use mean squared error (MSE) loss, we adopt cross-entropy loss instead, as MSE is ill-suited for sequential recommendation tasks where target item embeddings are discrete, and similarity is typically measured via dot-product correlations.

$$\hat{y} = \frac{\exp(\tilde{\mathbf{x}}_0 \cdot \mathbf{e}_{n+1})}{\sum_{i \in \mathcal{I}} \exp(\tilde{\mathbf{x}}_0 \cdot \mathbf{e}_i)} \\ \mathcal{L}_{CE} = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} -log\hat{y}_i \tag{12}$$

Here, $\tilde{\mathbf{x}}_0$ is reconstructed by the Transformer-based predictor $f_\theta(\cdot)$, while $\cdot$ represents the inner product operation used to measure the correlation between vectors.

Traditional diffusion models optimize single-step denoising via reconstruction losses (e.g., cross-entropy) but fail to ensure global consistency across multi-step denoising trajectories. To address this, we introduce fractional score-matching loss ($\mathcal{L}_{DSM}$), which aligns the gradient field of the predicted distribution ($\tilde{\mathbf{x}}_0$) with that of the ground-truth distribution ($\mathbf{y}_s$), correcting deviations caused by error accumulation. This loss is defined as:

$$\mathcal{L}_{\text{DSM}} = \mathbb{E}_{y_s, \sigma_s} \left[\frac{1}{2} \left\| S_\phi(\tilde{\mathbf{x}}_0) - \left(\frac{S_\phi(\mathbf{y}_s) - S_\phi(\tilde{\mathbf{x}}_0)}{\sigma_s^2} + S_\phi(\tilde{\mathbf{x}}_0) \cdot \sigma_s^2 \right) \right\|^2 \right] \tag{13}$$

where the energy function $S_\phi(\cdot)$ learns an implicit gradient field to steer $\tilde{\mathbf{x}}_0$ toward the true data manifold ($\mathbf{y}_s$), reducing prediction noise and enhancing generalization. The total loss combines cross-entropy ($\mathcal{L}_{CE}$) and score-matching loss:

$$\mathcal{L}_{all} = \mathcal{L}_{CE} + \eta \mathcal{L}_{DSM} \tag{14}$$

Here, η balances the two objectives, controlling how strictly the generated distribution aligns with ground-truth data.

4 Experiment

In this section, we answer the following research questions to evaluate our propose method systematically:

RQ1. How does AlignDiff compare to state-of-the-art models on standard metrics?

RQ2. How do individual components of AlignDiff affect its performance?

RQ3. How do key hyperparameters influence AlignDiff's performance?

RQ4. Does AlignDiff generate data closer to the ground-truth distribution than baselines?

RQ5. Can the energy function effectively learn gradient information?

RQ6. How does our gradient optimization method compare to prior approaches?

4.1 Experimental Protocol

Datasets. **Amazon Beauty & Toys** [2]: Two Amazon product category datasets spanning 18 years of user behavior (ratings, reviews, purchase sequences), with product metadata (ASIN codes, categories, features) and user profiles capturing behavioral imbalance. **MovieLens-1M** [3]: A cleaned dataset of 1 million explicit movie ratings (1-5 stars) from 6,040 users on 3,952 movies, including timestamped interactions for modeling user interest evolution.

Dataset	#Sequence	#items	#Actions	Avg_len	Sparsity
Beauty	22363	12101	198502	8.53	99.93%
Toys	19412	11924	167597	8.63	99.93%
Movielens-1M	6040	3416	999611	165.50	95.16%

Table 1: Statistical information after preprocessing of the three datasets

Baselines. We evaluate our AlignDiff method by comparing three categories of representative sequential recommendation methods, that is, discriminative, generative and diffusion-based sequential recommendation models. **Discriminative Sequential Recommendation Models:** GRU4Rec (Hidasi and Karatzoglou, 2018), Caser (Tang and Wang, 2018), SASRec (Kang and McAuley, 2018), BERT4Rec (Sun et al., 2019), ComiRec (Cen et al., 2020), STOSA (Fan et al., 2022). **Generative sequential recommendation Models:** SVAE (Sachdeva et al., 2019), ACVAE (Xie et al., 2021), **Diffusion-based sequential recommendation Models:** CadiRec (Cui et al., 2024), DiffuRec (Li et al., 2024). Our experiments on GeForce RTX 3090 GPUs employ 4-layer Transformers (128D embeddings, batch size 512) trained with Adam (η=0.001), using energy-guided diffusion ($\Omega \in$0.1,1, $\nu \in$[0.05,0.5]) and dropout regularization (0.1-0.3).

4.2 Overall Comparision(RQ1)

As shown in table 2, AlignDiff outperforms all baselines across datasets, achieving significant gains (e.g., +20.4% HR@20 and +11.3% NDCG@20 vs. DiffuRec on Amazon-Toys) by aligning denoising paths via gradient differences (energy function + score-matching loss) and multi-level conditional guidance. Traditional sequential models (GRU4Rec/Caser) underperform due to limited long-term dependency modeling, while self-attention-based SASRec and BERT4Rec excel in sequence representation. Among generative models, SVAE struggles with preference modeling, ACVAE improves via adversarial contrastive learning for global distribution alignment, and DiffuRec (prior SOTA) leverages dynamic uncertainty injection. AlignDiff surpasses these by integrating gradient alignment (matching denoised outputs to ground-truth distributions) and multi-level semantic guidance, demonstrating diffusion models' superiority in iterative distribution learning through robust denoising trajectory optimization.

4.3 Ablation Study(RQ2)

We evaluate AlignDiff on three datasets by comparing two variants, with results in Table 3. Ablation Analysis (Table 3) confirms the critical roles of AlignDiff's components: Removing multi-level conditional guidance (ML)—which provides semantic-rich embeddings to align latent user preferences—causes significant performance degradation, as basic interaction sequences fail to capture nuanced preferences. Disabling the energy function (EN)—responsible for gradient alignment between denoised predictions and ground-truth distributions via score-matching loss—disrupts distribution fidelity, leading to suboptimal denoising paths. Both components are indispensable: ML ensures semantically consistent guidance signals, while EN enforces gradient-driven alignment, jointly enabling robust uncertainty handling and generalization. This synergy validates the framework's design superiority in balancing preference modeling and

[2] https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2

[3] http://files.grouplens.org/datasets/movielens/ml-1m.zip

Dataset	Metric	GRU4Rec	Caser	SASRec	BERT4Rec	ComiRec	SVAE	ACVAE	STOSA	CaDiRec	DiffuRec	**AlignDiff**	↑%
Beauty	HR@5	1.0112	1.6118	3.2688	2.1326	2.0495	0.9943	2.4672	3.5457	5.1608	5.5758	**5.8675**	5.2%
	HR@10	1.9370	2.8166	6.2648	3.7160	4.4545	1.9745	3.8832	6.2048	7.2829	7.9068	**8.126**	2.8%
	HR@20	3.8531	4.4048	8.9791	5.7922	7.6968	3.1552	6.1224	9.5939	10.5140	11.1098	**11.3729**	2.4%
	NDCG@5	0.6084	0.9758	2.3989	1.3207	1.0503	0.6702	1.6858	2.5554	3.4473	4.0047	**4.2023**	4.9%
	NDCG@10	0.9029	1.3602	3.2305	1.8219	1.8306	0.9863	2.1389	3.2085	4.2764	4.7494	**4.9261**	3.7%
	NDCG@20	1.3804	1.7595	3.6563	2.3541	2.6451	1.2867	2.7020	3.7609	4.7148	5.5566	**5.6730**	2.1%
Toys	HR@5	1.1009	0.9622	4.5333	1.9260	2.3026	0.9109	2.1897	4.2236	4.9158	5.5650	**6.3244**	13.6%
	HR@10	1.8553	1.8317	6.5496	2.9312	4.2901	1.3683	3.0749	6.9393	7.1549	7.4587	**8.7229**	16.9%
	HR@20	3.1827	2.9500	9.2263	4.5889	6.9357	1.9239	4.4061	9.5096	9.6011	9.8417	**11.8493**	20.4%
	NDCG@5	0.6983	0.5707	3.0105	1.1630	1.1571	0.5580	1.5604	3.1017	3.8417	4.1667	**4.4175**	6.0%
	NDCG@10	0.9396	0.8510	3.7533	1.4870	1.7953	0.7063	1.8452	3.8806	4.4957	4.7724	**5.1904**	8.8%
	NDCG@20	1.2724	1.1293	4.3323	1.9038	2.4631	0.8446	2.1814	4.3789	4.8143	5.3684	**5.9726**	11.3%
ML-1M	HR@5	5.1139	7.1401	9.3812	13.6393	6.1073	1.4869	12.7167	7.0495	15.1846	17.9659	**18.2658**	1.7%
	HR@10	10.1664	13.3792	16.8941	20.5675	12.0406	2.7189	19.9313	14.3941	22.5497	26.2647	**27.3419**	4.1%
	HR@20	18.6995	22.5507	28.318	29.9479	21.0094	5.0326	28.9722	24.9871	31.4655	36.7870	**37.7156**	2.5%
	NDCG@5	3.0529	4.1550	5.3165	8.8922	3.5214	0.9587	8.2287	3.7174	10.5842	12.1150	**12.5781**	3.8%
	NDCG@10	4.6754	6.1400	7.7277	11.1251	5.4076	1.2302	10.5417	6.0771	11.7544	14.7909	**15.4907**	4.7%
	NDCG@20	6.8228	8.4304	10.5946	13.4763	7.6502	1.8251	12.8210	8.7241	15.9640	17.4386	**18.1183**	3.9%

Table 2: Results (%) across three datasets: best in bold, second-best underlined; last column shows AlignDiff's improvement over the top baseline. All experiments were performed five times and the average value was taken.

Dataset	**Ablation**	**HR@5**	**HR@10**	**HR@20**	**NDCG@5**	**NDCG@10**	**NDCG@20**
Beauty	w/o ML	5.8369	7.8229	10.3678	4.1279	4.7627	5.3951
	w/o EN	5.7707	7.9518	10.7033	4.1684	4.8699	5.5596
	AlignDiff	**5.8675**	**8.1260**	**11.0729**	**4.2023**	**4.9261**	**5.6730**
Toys	w/o ML	6.1585	8.6638	11.6357	4.3922	5.0992	5.8229
	w/o EN	5.6382	7.4221	9.8330	4.2255	4.8007	5.4072
	AlignDiff	**6.3244**	**8.7229**	**11.8493**	**4.4175**	**5.1904**	**5.9726**
Ml-1M	w/o ML	16.1037	24.8246	36.2756	10.5078	13.5789	16.2743
	w/o EN	15.8171	24.5478	35.6394	10.3501	13.1469	15.9432
	AlignDiff	**16.3178**	**25.4160**	**36.7156**	**11.0781**	**13.8907**	**16.6183**

Table 3: Ablation experiments on three datasets

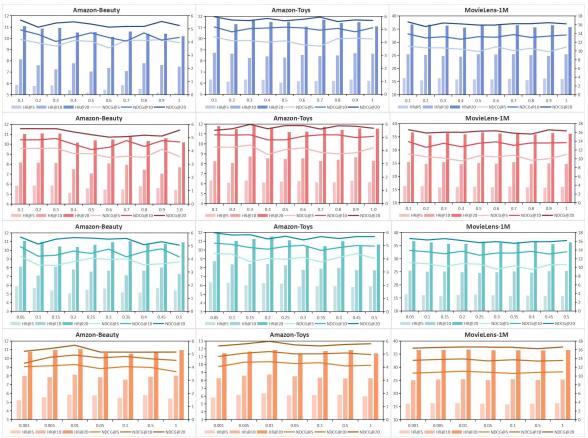

Figure 3: Parameter sensitivity of different Ω, w, ν, η.

Dataset	**Metric**	**1-Layer MLP**	**3-Layer MLP**	**6-Layer MLP**	**Transformer-nased**
Beauty	HR@5	5.6580	**5.8675**	5.7245	5.7769
	HR@10	7.9652	**8.1260**	8.0219	7.8453
	HR@20	10.5314	**11.0729**	10.7256	10.6381
	NDCG@5	4.0945	**4.2023**	4.1429	4.1156
	NDCG@10	4.7549	**4.9261**	4.8126	4.7741
	NDCG@20	5.4826	**5.6730**	5.5016	5.4822

Table 4: The impact of different architectural energy functions on model performance on the Amazon-Beauty dataset.

distributional accuracy.

4.4 Parameter Sensitive Analysis(RQ3)

Hyperparameter analysis (Figures 3) demonstrates AlignDiff's adaptability and stability: Guidance weight w adjusts conditional guidance strength, with sparse Amazon datasets (Beauty/Toys) favoring w = 0.3 to compensate for limited interactions, while dense MovieLens-1M performs best at w = 0.1 to avoid over-constraining rich behavioral patterns. Noise scale ν is unified (ν = 0.05) across datasets, balancing robustness to perturbations without distorting underlying distributions, proving effective regardless of data sparsity. Loss balance $\eta = 0.01$ optimally weights cross-entropy (accuracy) and score-matching (distribution alignment) losses, avoiding underalignment or over-constraint extremes. Energy scale $\Omega = 0.01$ stabilizes gradient-driven denoising across all datasets, preventing overshooting from excessive adjustments.

4.5 Data Distribution Comparison Analysis(RQ4)

Visual (t-SNE/PCA) and quantitative (MMD) analyses confirm that AlignDiff's generated distribution aligns closer to ground-truth data than DiffuRec. As shown in figure 4, PCA shows tighter clustering and preserved structural relationships, attributed to multi-level conditional guidance and gradient alignment via score-matching loss, which refine denoising paths beyond cross-entropy optimization. Lower MMD values (Table 5) and broader low-dimensional overlap with the ground-truth distribution quantitatively validate reduced discrepancy, demonstrating the effectiveness of gradient-driven constraints in enhancing distributional fidelity.

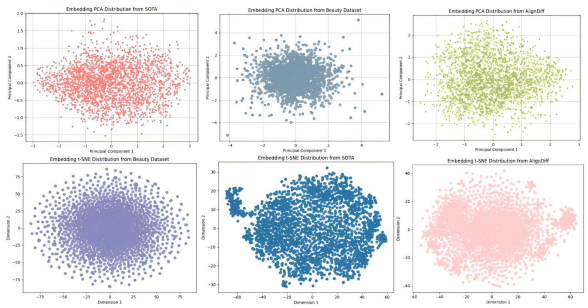

Figure 4: Differences between the original data distribution, the distribution generated by the SOTA model and the data distribution generated by AlignDiff on the Amazon-Beauty dataset using the PCA and T-SNE method.

Dataset	Metric	SOTA(mean±std)	AlignDiff(mean±std)	Improve %	p value
Beauty	MMD($\times 10^{-3}$)	0.37±0.05	0.33±0.03	**10.8%**	0.001
Toys	MMD($\times 10^{-3}$)	0.42±0.04	0.35±0.05	**16.7%**	0.001
ML-1M	MMD($\times 10^{-3}$)	0.34±0.05	0.31±0.02	**8.8%**	0.001

Table 5: MMD significance test on three datasets (p-value < 0.001)

4.6 The Impact of the Energy Function Architecture(RQ5)

Experiments (Table 4) show that a 3-layer MLP-based energy function optimizes performance, balancing gradient alignment, and noise robustness. Overly complex architectures (e.g., 6-layer MLP/Transformer) capture high-frequency noise, causing gradient oscillation, while simplistic 1-layer MLPs fail to decouple noise signals. The energy function's role—modeling gradient differences between generated and ground-truth distributions—requires balancing expressiveness and stability. Single-step correction (Table 6) outperforms multi-step methods, as iterative adjustments risk overcorrection (forcing predictions toward training data's local density peaks), highlighting the need for restrained gradient constraints to preserve generalization while aligning distributions.

4.7 Gradient Optimization Method Comparison(RQ6)

Compared to gradient-enhanced variants (GAD-GP with gradient penalty, GAD-Adv with adversarial training), as shown in table 7. AlignDiff's joint optimization strategy—combining cross-entropy loss (task accuracy) and energy-based score matching (distribution alignment)—achieves superior performance. Unlike fixed-gradient methods (e.g., WGAN-GP), AlignDiff dynamically adjusts gradient alignment strength through learnable energy parameters (Ω), enabling an adaptive balance between recommendation fidelity and distribution shifts. Multi-level conditional guidance further refines gradient alignment across item embeddings and sequence semantics, outperforming single-scale approaches. Results show significant gains on sparse datasets (e.g., +20.4% HR@20 on Amazon-Toys vs. +2.5% on MovieLens-1M), confirming that explicit gradient difference learning via the energy function effectively mitigates recommendation bias caused by data sparsity, while traditional methods struggle with distributional discrepancies in interaction-scarce scenarios.

Steps	Beauty-HR@10	Beauty-NDCG@10	Toys-HR@10	Toys-NDCG@10
1	**8.1260**	**4.9261**	**8.7229**	**5.1904**
2	7.9524	4.6058	8.5219	5.0257
4	7.4428	4.3281	8.2907	4.7528
8	7.1625	4.0817	7.9056	4.4325

Table 6: The proposed model's effect on the model performance at different correction steps.

Method	HR@5	HR@10	NDCG@5	NDCG@10
GAD-GP	5.6416	7.8243	3.9424	4.6308
GAD-Adv	5.2649	7.1526	3.6451	4.2107
AlignDiff	**5.8675**	**8.1260**	**4.2023**	**4.9261**

Table 7: Gradient Method Comparison on Beauty Dataset

5 Conclusion

This study addresses two key limitations of diffusion models in sequential recommendation: 1) semantic distortion guidance representation (failing to capture user behavior patterns) and 2) gradient misalignment between generated and ground-truth data distributions. We propose AlignDiff, which resolves these issues through gradient field alignment and an adaptive guidance mechanism integrating Transformer encoders with multi-level signals (e.g., interaction sequences). Experiments on three benchmark datasets validate the model's effectiveness, with ablation studies confirming component synergies. The work establishes a novel "gradient alignment + multi-level guidance" paradigm for diffusion-based recommendation.

6 Limitations

AlignDiff currently has two limitations: 1) It only uses ID-based interactions and ignores multimodal features to achieve content-aware recommendations. 2) It is difficult to model dynamic changes in user interests.

References

Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2942–2951.

Junyang Chen, Guoxuan Zou, Pan Zhou, Wu Yirui, Zhenghan Chen, Houcheng Su, Huan Wang, and Zhiguo Gong. 2024. Sparse enhanced network: An adversarial generation method for robust augmentation in sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8283–8291.

Ziqiang Cui, Haolun Wu, Bowei He, Ji Cheng, and Chen Ma. 2024. Context matters: Enhancing sequential recommendation with context-aware diffusion-based contrastive learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 404–414.

Ziwei Fan, Zhiwei Liu, Yu Wang, Alice Wang, Zahra Nazari, Lei Zheng, Hao Peng, and Philip S Yu. 2022. Sequential recommendation via stochastic self-attention. In *Proceedings of the ACM web conference 2022*, pages 2036–2047.

Yongjing Hao, Pengpeng Zhao, Junhua Fang, Jianfeng Qu, Guanfeng Liu, Fuzhen Zhuang, Victor S Sheng, and Xiaofang Zhou. 2024. Meta-optimized joint generative and contrastive learning for sequential recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 705–718.

B. Hidasi and A. Karatzoglou. 2018. Recurrent neural networks with top-k gains for session-based recommendations. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 843–852.

B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk. 2015. Session-based recommendations with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

J. Ho, A. Jain, and P. Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 6840–6851.

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

Jiacheng Li, Yujie Wang, and Julian McAuley. 2020. Time interval aware self-attention for sequential recommendation. In *Proceedings of the 13th international conference on web search and data mining*, pages 322–330.

Li Li, Jianbing Xiahou, Fan Lin, and Songzhi Su. 2023. Distvae: distributed variational autoencoder for sequential recommendation. *Knowledge-Based Systems*, 264:110313.

X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto. 2022. Diffusion-lm improves controllable text generation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 4328–4343.

Z. Li, A. Sun, and C. Li. 2024. Diffurec: A diffusion model for sequential recommendation. *ACM Transactions on Information Systems*, 42(3):66:1–66:28.

J. Lin, W. Pan, and Z. Ming. 2020. Fissa: Fusing item similarity models with self-attention networks for sequential recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 130–139.

Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Zhanhui Kang. 2024. Plug-in diffusion model for sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8886–8894.

Yingxue Ma and Mingxin Gan. 2024. Sequential-hierarchical attention network: Exploring the hierarchical intention feature in poi recommendation. page 67.

S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 811–820.

Noveen Sachdeva, Giuseppe Manco, Ettore Ritacco, and Vikram Pudi. 2019. Sequential variational autoencoders for collaborative filtering. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 600–608.

F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1441–1450.

J. Tang and K. Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 565–573.

Min Tang, Shujie Cui, Zhe Jin, Shiuan-ni Liang, Chenliang Li, and Lixin Zou. 2025. Sequential recommendation by reprogramming pretrained transformer. *Information Processing & Management*, 62(1):103938.

Zihao Wu, Xin Wang, Hong Chen, Kaidong Li, Yi Han, Lifeng Sun, and Wenwu Zhu. 2023. Diff4rec: Sequential recommendation with curriculum-scheduled diffusion augmentation. In *Proceedings of the 31st*

ACM international conference on multimedia, pages 9329–9335.

Z. Xie, C. Liu, Y. Zhang, H. Lu, D. Wang, and Y. Ding. 2021. Adversarial and contrastive variational autoencoder for sequential recommendation. In *Proceedings of the Web Conference 2021*, pages 449–459.

Z. Yang, J. Wu, Z. Wang, X. Wang, Y. Yuan, and X. He. 2023. Generate what you prefer: Reshaping sequential recommendation via guided diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 24247–24261.

L. Yu, W. Zhang, J. Wang, and Y. Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2852–2858.