

Beyond Noisy Signals: Dual-Level Denoising for Multi-modal Sequential Recommendation

A Appendix

A.1 Fourier Transform

Discrete Fourier Transform. Discrete Fourier Transform (DFT) is a fundamental tool in digital signal processing [10] that converts a sequence from the time domain to the frequency domain. Given a sequence $\{x_j\}$ with $j \in \{0, 1, \dots, n-1\}$, the DFT transforms it into the frequency domain according to:

$$\tilde{x}_k = \sum_{j=0}^{n-1} x_j \exp\left(-\frac{2\pi i}{n} jk\right), \quad 0 \leq k \leq n-1, \quad (1)$$

where i is the imaginary unit, and $\tilde{x}_k$ denotes the spectrum of the sequence at frequency step $\omega_k = 2\pi k/n$. The DFT is a one-to-one transformation, and the original sequence can be recovered from its spectrum via the Inverse DFT (IDFT):

$$x_j = \frac{1}{n} \sum_{k=0}^{n-1} \tilde{x}_k \exp\left(\frac{2\pi i}{n} jk\right), \quad 0 \leq j \leq n-1. \quad (2)$$

Fast Fourier Transform. The Fast Fourier Transform (FFT) is an efficient algorithm for computing the DFT, which reduces the time complexity from $O(n^2)$ in the naive DFT to $O(n \log n)$ by factorizing the DFT matrix into a product of sparse factors [1]. The inverse DFT can likewise be efficiently computed via the Inverse FFT (IFFT). Since FFT converts input signals into the frequency domain where periodic characteristics are more easily captured, it is widely adopted in sequential modeling to filter noise [13–15].

In this work, we apply FFT and IFFT to the multi-modal item embedding sequence $\mathbf{H} \in \mathbb{R}^{n \times d}$ for sequence-level denoising. Specifically, FFT projects the sequence from the time domain into the frequency domain, where a learnable complex-valued filter modulates the spectrum to suppress noise-induced components; IFFT then reconstructs the purified sequence back to the time domain for downstream preference modeling.

A.2 Details of Baselines

We compare our method with well-known SR baselines with two categories:

ID-based Models:

- **SASRec** [4] employs a unidirectional self-attention mechanism to capture sequential dependencies among items for next-item prediction.
- **BERT4Rec** [7] adopts a bidirectional Transformer with a masked item prediction task to learn richer sequential representations.
- **FEARec** [2] augments self-attention with frequency-domain components via a ramp structure to capture both low- and high-frequency sequential signals.
- **BSARec** [6] leverages the Fourier transform to integrate frequency information into self-attention, mitigating the oversmoothing problem in sequential recommendation.

Multi-modal Models:

Table 1: Efficiency comparison with DMMD4SR. *Time/E* (s) denotes the average training time per epoch, and *MU* (GB) indicates GPU memory usage.

Datasets	Model	Recall@20	NDCG@20	Time/E	MU
Beauty	DMMD4SR	0.1129	0.0466	42s	18.13G
	DDMSR	0.1579	0.0671	14s	9.32G
Sports	DMMD4SR	0.0672	0.0262	62s	18.48G
	DDMSR	0.0975	0.0417	26s	12.36G
Toys	DMMD4SR	0.1159	0.0488	36s	18.13G
	DDMSR	0.1660	0.0695	11s	8.56G
MicroLens	DMMD4SR	0.1216	0.0490	136s	18.44G
	DDMSR	0.1497	0.0637	53s	11.15G

- **UniSRec** [3] learns transferable universal sequence representations from item textual descriptions via a parametric whitening and MoE-enhanced adaptor, enabling cross-domain recommendation.
- **MISSRec** [8] designs an interest-aware Transformer encoder-decoder that captures sequence-level multi-modal user interests and adopts a dynamic fusion module for user-adaptive item representations.
- **MP4SR** [11] proposes a multi-modal pre-training framework that leverages contrastive learning to capture correlations among different modality sequences of users and items for enhanced sequential recommendation.
- **TedRec** [9] transforms textual and ID embeddings into the frequency domain via Fourier transform, achieving effective sequence-level text-ID semantic fusion through element-wise multiplication.
- **HM4SR** [12] proposes a hierarchical time-aware Mixture-of-Experts framework that integrates explicit temporal information with multi-modal features to model dynamic user interests.
- **DMMD4SR** [5] introduces a diffusion model-based multi-level denoising framework that progressively mitigates domain shift noise and interest-agnostic noise in multi-modal representations for sequential recommendation.

For a fair comparison, we train UniSRec, MP4SR and MISSRec from scratch without pre-training on additional datasets.

A.3 Efficiency Analysis

In this section, we comprehensively compare the training efficiency of our proposed DDMSR against the generative diffusion-based baseline, DMMD4SR, from both theoretical complexity and empirical perspectives.

Let B denote the batch size, L the maximum sequence length, d the hidden feature dimension, N the item vocabulary size, and k the number of retained neighbors in the sparse similarity graph.

Theoretical Training Complexity. The denoising module of DMMD4SR heavily relies on multiple generative diffusion networks. During the training phase, for each batch, DMMD4SR not only computes the reconstruction loss but also executes a T -step autoregressive reverse sampling across M independent diffusion networks (e.g., text, image, and sequence SDNs) to obtain denoised representations for subsequent sequence encoding. Since the single-step prediction of each network is dominated by MLPs with a complexity of $O(B \cdot L \cdot d^2)$, the total time complexity per training batch reaches:

$$O_{\text{DMMD4SR}} = O(M \cdot T \cdot B \cdot L \cdot d^2) \quad (3)$$

Furthermore, storing the multi-step computational graphs and gradients for all modality experts incurs an extremely high memory footprint.

In contrast, DDMSR introduces a highly efficient discriminative spatial-frequency denoising paradigm. For the Graph Feature Denoiser, graph aggregation over the entire item vocabulary is performed via sparse matrix multiplication, incurring a per-forward cost of $O(N \cdot k \cdot d)$. For the sequence Frequency Filter, high-frequency noise is suppressed in a single one-shot operation: across F stacked filter layers, the Fast Fourier Transform (FFT) and its inverse contribute $O(F \cdot B \cdot d \cdot L \log L)$, while the sequential gating mechanism contributes $O(F \cdot B \cdot L \cdot d^2)$. The combined per-batch training complexity of DDMSR is therefore:

$$O_{\text{DDMSR}} = O(N \cdot k \cdot d + F \cdot B \cdot L \cdot d \cdot (\log L + d)). \quad (4)$$

It is worth noting that although Eq. (4) contains the term $O(N \cdot k \cdot d)$, this operation amounts to a single, highly optimized sparse aggregation step whose practical wall-clock overhead is negligible compared to the repeated multi-step MLP evaluations in diffusion-based denoising. The training complexity of DDMSR can therefore be effectively regarded as $O(F \cdot B \cdot L \cdot d \cdot (\log L + d))$. Moreover, during online inference, the item-level graph aggregation can be pre-computed entirely offline, eliminating the $O(N \cdot k \cdot d)$ term from the online critical path and reducing its effective inference-time cost to $O(1)$.

Empirical Efficiency. The theoretical advantages are strongly validated by the empirical training overhead presented in Table ?? . Across all evaluated datasets, DDMSR consistently achieves a 2.5× to 3× speedup in training time per epoch (Time/e) compared to DMMD4SR (e.g., reducing the training time from 42s to 14s on the Beauty dataset). Additionally, by completely circumventing the cumbersome multi-step diffusion graphs, DDMSR reduces GPU memory usage (MU) by roughly 30% to 50%. Most importantly, while substantially reducing both computational and memory costs, DDMSR still yields significant performance improvements in Recall@20 and NDCG@20. This firmly demonstrates the superiority, scalability, and practical value of our proposed method in real-world recommender systems.

A.4 More Hyperparameter Sensitivity Analysis

A.5 Hyperparameter Sensitivity Analysis

Due to space limitations in the main text, we present detailed hyperparameter sensitivity analyses in this section. Specifically, we investigate the impact of the number of top- K neighbors in the semantic graph, the number of frequency-domain denoising layers

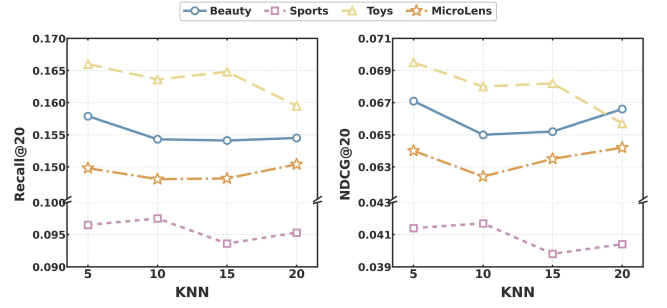


Figure 1: Effects of top- K neighbors.

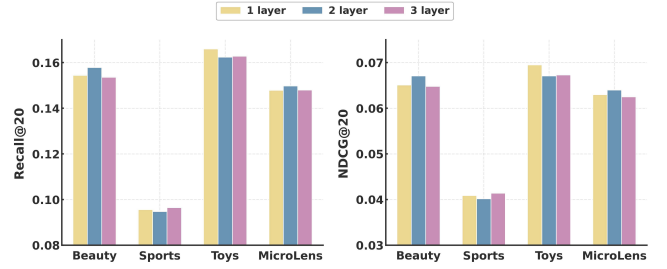


Figure 2: Effects of frequency-domain denoising layers.

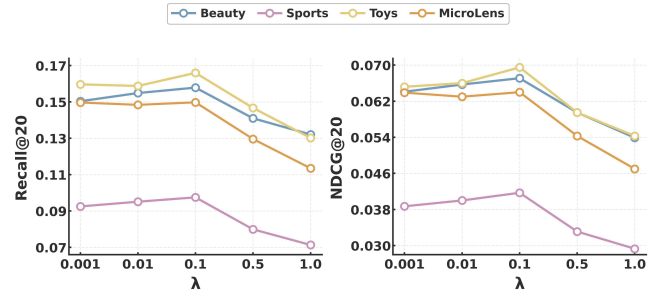


Figure 3: Effects of the contrastive learning weight λ .

L_f , and the contrastive learning weight λ on the overall recommendation performance.

Impact of Top- K Neighbors. Figure 1 shows the impact of the number of retained neighbors ($K \in \{5, 10, 15, 20\}$) during item semantic graph construction. The optimal K varies across datasets: Sports performs best at $K = 10$ and MicroLens at $K = 20$, while Beauty and Toys achieve peak performance at $K = 5$. We attribute this to differences in item vocabulary scale and interaction density. For datasets with fewer items and sparse interactions (e.g., Beauty and Toys), a smaller K better preserves the most relevant neighbors and avoids introducing semantic noise. In contrast, datasets with larger item sets or denser interactions (e.g., Sports and MicroLens) benefit from a larger K to capture richer semantic relationships. Overall, the performance remains relatively stable across different K values, demonstrating the robustness of our graph-based denoising module.

Table 2: Robustness under Noise (Recall@20)

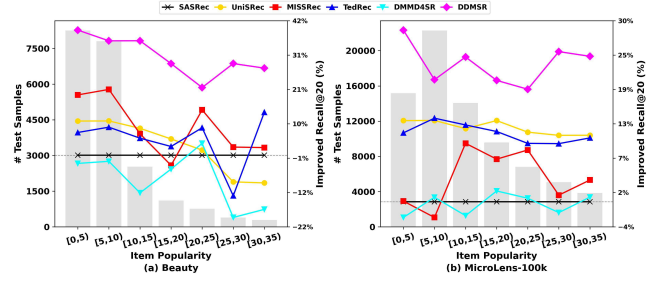
Dataset	Method	Clean R@20	$n_r = 0.1$		$n_r = 0.3$		$n_r = 0.5$	
			R@20	Drop	R@20	Drop	R@20	Drop
Beauty	UniSRec	0.1265	0.1047	-17.2%	0.0961	-24.0%	0.0677	-46.5%
	DMMD4SR	0.1109	0.0958	-13.6%	0.0902	-18.7%	0.0640	-42.3%
	DDMSR	0.1579	0.1405	-11.0%	0.1337	-15.3%	0.1039	-34.2%
MicroLens	UniSRec	0.1374	0.1139	-17.1%	0.1079	-21.4%	0.0756	-45.0%
	DMMD4SR	0.1219	0.1012	-17.0%	0.0962	-21.1%	0.0661	-45.8%
	DDMSR	0.1499	0.1337	-10.8%	0.1302	-13.1%	0.1011	-32.6%

Impact of Frequency-Domain Denoising Layers. Figure 2 illustrates the influence of the number of frequency filter layers ($L_f \in \{1, 2, 3\}$). The optimal L_f is dataset-dependent: Beauty and MicroLens achieve their best performance at $L_f = 2$, Toys at $L_f = 1$, and Sports at $L_f = 3$. Generally, performance improves when increasing L_f from 1 to 2 on most datasets, but further deepening the network to 3 layers yields marginal gains or even slight degradation. This reveals a critical trade-off between noise suppression and information preservation. Datasets with heavier interaction noise (e.g., Sports) require deeper filtering, while relatively cleaner datasets (e.g., Toys) may suffer from over-smoothing or information loss when excessive filtering layers are applied.

Impact of Contrastive Learning Weight λ . The hyperparameter λ controls the magnitude of the multi-modal contrastive alignment objective, balancing the auxiliary cross-modal regularization with the primary sequential recommendation task. We search for the optimal λ within $\{0.001, 0.01, 0.1, 0.5, 1.0\}$. As depicted in Figure 3, performance steadily improves as λ increases from 0.001, reaching a consistent peak at $\lambda = 0.1$. However, increasing λ beyond 0.1 (e.g., to 0.5 or 1.0) leads to a sharp performance drop. This trend aligns with our intuition: a moderate contrastive weight effectively bridges the heterogeneity gap between modalities and enforces semantic consistency. Conversely, an excessively large λ causes the model to over-focus on the alignment task, dominating the gradient updates and inducing negative transfer to the main temporal preference modeling. Hence, we set $\lambda = 0.1$ across all main experiments.

A.6 Noise Robustness

As clarified, our feature denoising targets semantic artifacts, not random numerical fluctuations. Injecting synthetic random noise corrupts the representation manifold, causing graph aggregation to inadvertently amplify unstructured interference. Furthermore, realistically simulating multi-modal semantic noise (e.g., altering backgrounds) is complex and introduces confounders. Therefore, to strictly evaluate noise robustness, we conducted sequence-level perturbations to simulate the common noise in recommendation systems. Specifically, we replaced a proportion ($n_r \in \{0.1, 0.3, 0.5\}$) of actual user interactions with random items to mimic severe accidental clicks. While all models degrade due to sequence corruption, DDMSR exhibits the strongest resilience. As shown in Table 2, under extreme 50% noise on MicroLens, DDMSR’s Recall@20 drops by only 32.6%, significantly outperforming DMMD4SR (45.8% drop) and UniSRec (45.0% drop). This rigorously proves that our frequency filter actively attenuates anomalous signals rather than acting as a standard transformation.

**Figure 4: Performance improvement across different item popularity (Long-tail) groups.**

A.7 Long-Tail Analysis

Pure ID-based models (e.g., SASRec) inherently suffer from representation collapse on data-sparse items. In Fig. 4, we establish the pure ID model SASRec as the baseline to evaluate relative performance improvements. While other multi-modal models exhibit moderate robustness on infrequent items, DDMSR decisively dominates. It achieves massive, unprecedented performance gains on the least frequent items (e.g., the $[0, 5]$ and $[5, 10]$ bins). This directly confirms that our graph-based semantic denoising, coupled with cross-modal alignment, successfully leverages neighborhood aggregation and modality complementarity to enrich high-quality multi-modal representations for long-tail items where interactions are extremely scarce.

References

- [1] David H. Bailey. 1993. Computational Frameworks for the Fast Fourier Transform (Charles Van Loan). *SIAM Rev.* 35, 1 (1993), 142–143. doi:10.1137/1035017
- [2] Xinyu Du, Huanhuan Yuan, Pengpeng Zhao, Jianfeng Qu, Fuzhen Zhuang, Guanfeng Liu, Yanchi Liu, and Victor S. Sheng. 2023. Frequency Enhanced Hybrid Attention Network for Sequential Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23–27, 2023*. 78–88. doi:10.1145/3539618.3591689
- [3] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards Universal Sequence Representation Learning for Recommender Systems. In *KDD ’22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14–18, 2022*. ACM, 585–593. doi:10.1145/3534678.3539381
- [4] Wang-Cheng Kang and Julian J. McAuley. 2018. Self-Attentive Sequential Recommendation. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17–20, 2018*. 197–206. doi:10.1109/ICDM.2018.00035
- [5] Weihai Lu and Li Yin. 2025. DMMD4SR: Diffusion Model-based Multi-level Multimodal Denoising for Sequential Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27–31, 2025*. ACM, 6363–6372. https://doi.org/10.1145/3746027.3755861
- [6] Yehjin Shin, Jeongwhan Choi, Hyowon Wi, and Noseong Park. 2024. An Attentive Inductive Bias for Sequential Recommendation beyond the Self-Attention. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20–27, 2024, Vancouver, Canada*. 8984–8992. doi:10.1609/AAAI.V38I8.28747
- [7] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3–7, 2019*. 1441–1450. doi:10.1145/3357384.3357895
- [8] Jinpeng Wang, Ziyun Zeng, Yunxiao Wang, Yuting Wang, Xingyu Lu, Tianxiang Li, Jun Yuan, Rui Zhang, Hai-Tao Zheng, and Shu-Tao Xia. 2023. MISSRec: Pre-training and Transferring Multi-modal Interest-aware Sequence Representation for Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023– 3 November 2023*.

- ACM, 6548–6557. doi:10.1145/3581783.3611967
- [9] Lanling Xu, Zhen Tian, Bingqian Li, Junjie Zhang, Daoyuan Wang, Hongyu Wang, Jinpeng Wang, Sheng Chen, and Wayne Xin Zhao. 2024. Sequence-level Semantic Representation Fusion for Recommender Systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21–25, 2024*. ACM, 5015–5022. doi:10.1145/3627673.3680037
- [10] C. K. Yuen. 1978. Review of "Theory and Application of Digital Signal Processing" by Lawrence R. Rabiner and Bernard Gold. *IEEE Trans. Syst. Man Cybern.* 8, 2 (1978), 146. doi:10.1109/TSMC.1978.4309918
- [11] Lingzi Zhang, Xin Zhou, Zhiwei Zeng, and Zhiqi Shen. 2025. Multimodal Pre-training for Sequential Recommendation via Contrastive Learning. *Trans. Recomm. Syst.* 3, 1 (2025), 1–23. <https://doi.org/10.1145/3682075>
- [12] Shengzhe Zhang, Liyi Chen, Dazhong Shen, Chao Wang, and Hui Xiong. 2025. Hierarchical Time-Aware Mixture of Experts for Multi-Modal Sequential Recommendation. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025– 2 May 2025*. ACM, 3672–3682. doi:10.1145/3696410.3714676
- [13] Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Filter-enhanced MLP is All You Need for Sequential Recommendation. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*. 2388–2399. doi:10.1145/3485447.3512111
- [14] Tian Zhou, Ziqing Ma, Xue Wang, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, and Rong Jin. 2022. FiLM: Frequency improved Legendre Memory Model for Long-term Time Series Forecasting. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. http://papers.nips.cc/paper_files/paper/2022/hash/524ef58c2bd075775861234266e5e020-Abstract-Conference.html
- [15] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In *International Conference on Machine Learning, ICML 2022, 17–23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research)*. PMLR, 27268–27286. <https://proceedings.mlr.press/v162/zhou22g.html>