Collapse Hunter: Tackling the Dimensional Degeneration in Generative Ranking

Haoran Xin[1†] **Junwei Pan**[3*] **Yongqi Zhou**[3] **Tianqu Zhuang**[3] **Yilong Sun**[3] **Zhixiang Feng**[3]
Shoujun Liu[3] **Gong Chen**[3] **Shudong Huang**[3] **Haijie Gu**[3] **Hui Xiong**[1,2*]

[1]Thrust of Artificial Intelligence,
The Hong Kong University of Science and Technology (Guangzhou)
[2]Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology, Hong Kong SAR, China
[3]Tencent Inc., China

hxin883@connect.hkust-gz.edu.cn; jonaspan@tencent.com; xionghui@ust.hk

Abstract

Generative ranking models (GRMs) have emerged as a scalable alternative to traditional recommendation pipelines by unifying items and user actions into a single token sequence and formulating recommendation as next-token prediction. Despite their promise, we identify a geometric failure mode: attention interactions with low-cardinality action tokens can induce dimensional collapse in high-rank item representations, according to the interaction-collapse theory. To pinpoint the origin of collapse, we interleave action and item tokens and view the attention mechanism in GRMs as a mixture of directional interaction channels between token types, and find that collapse concentrates in the action-attending-item channel, where low-rank action gating drags the item value stream toward a low-dimensional subspace. To address this, we theoretically show that nonlinearity mitigates collapse, and propose **Selective VAlue Nonlinearity (SVAN)**, a simple yet effective approach that applies nonlinear activation to the value component in the collapsed channel, protecting item representations while preserving the remaining attention geometry. SVAN consistently improves the ranking quality over strong baselines. Further analyses confirm that SVAN alleviates dimensional collapse, yielding up to a 12.58% effective-rank gain, and enhances representation discriminability. We release our code at xhran2010/Collapse-Hunter.

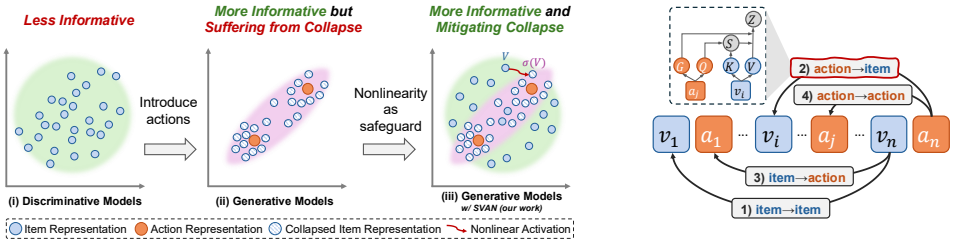
1 Introduction

Recommendation systems [19, 44, 54] have become a cornerstone of modern platforms, powering personalized experiences across e-commerce, social media, and content streaming. A longstanding challenge is to capture complex user behaviors and high-order dependencies among diverse signals. Classical discriminative models [31, 56, 39] address this with carefully engineered multi-field features and specialized interaction modules, but their design typically couples modeling capacity to handcrafted feature pipelines. Recently, generative recommenders [8, 48, 17, 30] have emerged as compelling alternatives: by sequentializing the signals into a unified token stream and training in a generative fashion, they simplify complex cascaded pipelines and exhibit favorable scaling behavior.

A representative paradigm interleaves item tokens with action tokens (*e.g.,* click/cart/purchase) [48], enabling generative modeling of the joint distribution over items and actions. Following this line,

*Corresponding authors.
†Work done when Haoran Xin was an intern at Tencent.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) (b)

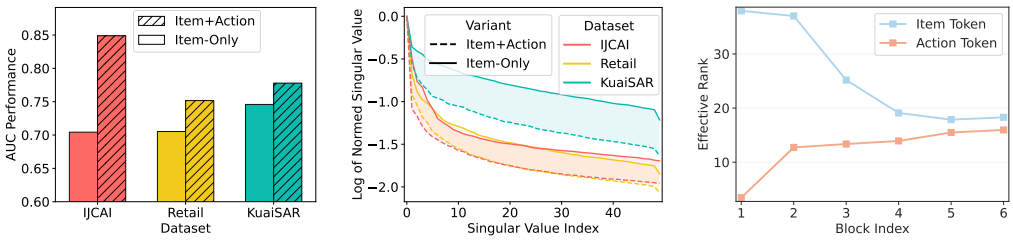
Figure 1: (a) low-cardinality actions can drag item representations toward a low-dimensional subspace, *i.e.,* causing dimensional collapse, while SVAN introduces a nonlinear safeguard to preserve item geometry. (b) Heterogeneous attention decomposition, where "→" is the attending direction, the highlighted *action→item* is the collapse-prone channel and $\boldsymbol{G} = \sigma(\boldsymbol{X}\boldsymbol{W}_G)$ is the gating term.

recent work further refines action modeling. Huang et al. [17], Guo et al. [10] improve efficiency by introducing action-oriented sequences that reduce the sequence length. Badrinath et al. [1] increases flexibility by conditioning generation on controllable actions. However, these efforts largely overlook a critical side effect: actions have *orders-of-magnitude lower cardinality* than items, and their interactions may distort the representation geometry of items.

In this paper, we study generative ranking models (GRMs) that incorporate both items and actions in the sequence and predict user actions on candidate items, and take a principled look at how actions affect representation learning in GRMs. Generally, actions have low cardinality (dozens of types) compared to items (millions to billions). Let d denote the hidden dimension and $|\mathcal{A}|$ the number of action types. While item embeddings can in principle occupy a d-dimensional space, action embeddings can span at most a $|\mathcal{A}|$-dimensional subspace, and $|\mathcal{A}| \ll d$. Building on interaction-collapse theory [12], we identify an unexplored failure mode: *through attention interactions, low-rank action embeddings can constrain item embeddings to a low-dimensional subspace, causing dimensional collapse* (Figure 1(a)(i-ii)). We validate this effect by showing that, while injecting actions improves ranking quality, the singular-value spectrum of item representations becomes more skewed, as shown in Figure 2(a-b).

To further pinpoint where collapse arises, we follow a representative sequence organization in GRMs that interleaves item and action tokens [48], which not only delivers superior ranking performance but also cleanly decouples item–action interactions by token type. Building on this organization, we introduce a novel heterogeneous view of GRM attention in which the aggregation naturally factorizes into directional interaction channels indexed between item and action tokens. Concretely, the attention can be decomposed into four channels: *item→item*, *item→action*, *action→item*, and *action→action*, where "→" denotes the attending direction, *i.e.,* queries/gates on the left attend keys/values on the right, as illustrated in Figure 1(b). In the attention, the value stream primarily governs representational expressiveness, while the remaining components work together to modulate interaction strength [9]. Accordingly, we diagnose collapse by examining the geometry of the value representations within each channel. Our theoretical analysis reveals that only the *action→item* channel exhibits collapse. Intuitively, *action→item* couples high-rank item values with low-rank gating terms derived from low-cardinality action tokens, which confines the training updates of item values to a low-dimensional subspace, repeatedly dragging item representations toward it. In contrast, the remaining channels either (i) couple token types of comparable cardinality on both sides, or (ii) place low-cardinality tokens on the value side, which is already low-rank. Neither case collapses a high-rank value stream.

This tension motivates a remedy that intervenes in the collapse-prone channel to preserve geometry. To this end, we devise **Selective VAlue Nonlinearity (SVAN)**, a simple yet effective approach that introduces interaction-aware nonlinear activations to mitigate collapse in GRMs. Specifically, we theoretically show that, in the identified collapse-prone channel, applying nonlinear activations to the high-rank value stream can expand the effective update subspace, thereby alleviating collapse (illustrated in Figure 1(a)(iii)). Guided by this insight, we propose to use nonlinearity as a selective safeguard: we apply it to the value stream in the collapsed channel, while leaving the remaining attention geometry unchanged. SVAN is parameter-free, straightforward to implement, and generalizable to various GRM architectures. Experiments show that SVAN consistently improves over strong traditional and generative baselines. Ablation results further confirm the necessity of selectivity:



(a) (b) (c)

Figure 2: (a) AUC w/ vs. w/o actions in the sequence: incorporating actions improves performance. (b) Singular value spectrum of item embeddings: introducing actions induces dimensional collapse. (c) Collapse level of action and item tokens across different layers measured by effective rank.

activating the value stream on the collapsed channel is consistently more effective than activating remaining collapse-free channels. Additional analyses illuminate the underlying mechanism by demonstrating that SVAN mitigates collapse compared with linear and universally nonlinear variants, yielding up to an 12.58% effective-rank gain, and improves representation discriminability.

Our main contributions are as follows: **1)** We identify and empirically characterize the action-induced dimensional collapse issue in GRMs; **2)** From a heterogeneity perspective, we decompose attention into interaction channels and pinpoint the collapse-prone channel that requires targeted remedy; **3)** We propose Selective VAlue Nonlinearity (SVAN), which selectively applies nonlinear activations to the value stream in the collapsed channel, achieving a targeted and geometry-preserving intervention.

2 Preliminaries

2.1 Task Formulation

We consider generative ranking models (GRMs) that include both items and actions in the sequence and cast recommendation as a next token prediction problem. Concretely, let $v_i \in \mathcal{V}$ denote the i-th item exposure and $a_i \in \mathcal{A}$ the corresponding user action (e.g., click, skip, like), where $\mathcal{V}$ and $\mathcal{A}$ are the set of items and actions, respectively. We organize each user's interaction history as a token sequence $(x_0, x_1, \ldots, x_{T-1})$ over $\mathcal{V} \cup \mathcal{A}$ that contains both item exposures and the actions taken on them, and parameterize a next-token distribution $p_\theta(x_t \mid x_{<t})$. Generative ranking is formulated as follows: given a history prefix h_i and a candidate item $v \in \mathcal{V}$, the model estimates the distribution over the target action on v, denoted $p_\theta(a \mid h_i, v)$. The predicted probability of the target action then serves as the ranking score for item v.

2.2 GRM Architecture

Similar to Transformers [35], each layer in a GRM first projects the input sequence embedding $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ into query, key, and value tensors:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q, \qquad \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K, \qquad \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V, \tag{1}$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{d \times d}$. It then performs attention aggregation,

$$\boldsymbol{S}\boldsymbol{V} = \phi\big(\boldsymbol{Q}\boldsymbol{K}^\top\big)\,\boldsymbol{V}, \tag{2}$$

where ϕ denotes the activation function. Finally, GRMs commonly incorporate a gating mechanism [48, 28] to regulate the information flow:

$$\boldsymbol{H} = \mathrm{FFN}(\boldsymbol{Z}) = \mathrm{FFN}(\boldsymbol{S}\boldsymbol{V} \odot \sigma(\boldsymbol{X}\boldsymbol{W}_G)), \tag{3}$$

where $\boldsymbol{W}_G \in \mathbb{R}^{d \times d}$, σ is the activation function, and FFN is a feedforward network. For clarity, we omit positional signals and layer normalization.

3 Analysis

In this section, we conduct an in-depth empirical analysis of GRMs and uncover two key phenomena: (i) introducing low-cardinality action tokens induces dimensional collapse in item token representations; (ii) collapse intensifies with network depth due to repeated item–action interactions.

3.1 Action Tokens Induce Dimensional Collapse

GRMs incorporate user actions (*e.g.,* click, cart, purchase) alongside items in the sequence, which introduces a cardinality imbalance: action vocabularies are typically tiny ($|\mathcal{A}| < 10$), whereas item catalogs scale to millions. Consequently, action embeddings are inherently confined to a low-dimensional subspace ($\leq |\mathcal{A}|$ dimensions), far below the hidden dimension d available to item embeddings. According to interaction-collapse theory [12], when high-rank and low-rank representations interact repeatedly through multiplication, the high-rank side tends to be pulled toward the low-rank subspace. In GRMs, this manifests as item representations progressively collapsing toward the action subspace, known as dimensional collapse [20, 14].

To empirically characterize this effect, we compare two GRM input variants: **(i) Items-Only:** sequences contain only item tokens; **(ii) Items+Actions:** sequences additionally contain action tokens. For each setting, we collect the learned item embedding table and examine its singular value spectrum, alongside the corresponding ranking performance. As shown in Figures 2(a) and 2(b), action tokens provide richer behavioral signals and substantially improve AUC (*e.g.,* a 20.5% gain on IJCAI). However, they also markedly skew the spectrum of item representations: across all three datasets, the Item+Action variant exhibits a much lower spectral than the Item-only baseline (*e.g.,* a 3.04% effective-rank drop on KuaiSAR), indicating that item embeddings are confined to a lower-dimensional subspace. This collapse suggests that the model may rely on fewer effective degrees of freedom, potentially limiting expressiveness and degrading downstream performance.

> ***Finding 1*** Although semantically informative, low-cardinality action tokens can induce dimensional collapse in item representations, potentially reducing the expressiveness of GRMs.

3.2 Collapse Accumulates with Depth

The finding above shows that introducing action tokens already induces noticeable collapse in item representations. Since deep GRMs stack multiple layers, each performing another round of item–action interaction, we next examine whether this effect compounds with depth, and how the collapse evolves across layers. Specifically, for each layer $\ell = 1, \ldots, L$, we extract the hidden states of items and actions separately, forming matrices H^{ℓ}_{item} and H^{ℓ}_{act}, and compute their effective rank:

$$r_{\text{eff}}(H) = \exp\Big(-\sum_i p_i \log p_i\Big), \quad p_i = \frac{\sigma_i^2}{\sum_j \sigma_j^2}, \tag{4}$$

where σ_i is the i-th singular value of H, and larger r_{eff} indicates a less collapsed spectrum. As shown in Figure 2(c), the effective rank of item representations decreases with depth (*e.g.,* a 32.0% effective-rank drop from layer 2 to 3), whereas action representations exhibit the opposite trend. Moreover, item representations maintain a higher r_{eff} than action representations at every depth, with a large gap in shallow layers that narrows in deeper layers. This pattern reflects that high-cardinality tokens consistently retain more information than low-cardinality tokens, yet are still dragged toward collapse as interactions accumulate.

> ***Finding 2*** Dimensional collapse of item tokens intensifies with depth as item–action interactions stack, while items consistently maintain a higher effective rank than actions across all layers.

4 Selective Value Nonlinearity

Building on these insights, we propose **Selective VAlue Nonlinearity (SVAN)** to mitigate dimensional collapse of item representations induced by low-cardinality actions in GRMs. In a nutshell, we

employ the nonlinear activation to the value component in the collapse-prone channel to prevent item representations from collapsing.

4.1 Heterogeneous Attention Interaction

GRMs jointly model items and actions in a unified token sequence, and several sequence organizations have been explored in the literature, such as interleaving item and action tokens [48], absorbing actions into items [17], and conditioning generation on action tokens [1]. SVAN adopts the item–action interleaved organization, *i.e.,* $(v_0, a_0, v_1, a_1, \dots)$, which assigns an explicit token type $t_i \in \{I, A\}$ to every position. We choose this organization for two reasons. First, it delivers superior ranking performance over alternative organizations, as empirically validated in Section 5.2. Second, by giving every position a well-defined token type, it lets us cleanly separate the otherwise entangled item–action interactions in GRMs into directional channels indexed by token type, enabling the heterogeneous decomposition below. Notably, this interleaved organization is orthogonal to the underlying GRM backbone, as it operates purely at the input level and leaves the model architecture untouched. We further validate this orthogonality in Appendix D.1, where SVAN generalizes consistently across diverse backbones (e.g., HSTU, Llama, Qwen).

As shown in Figure 1(b), we decompose the attention aggregation in Equation (2) into four blocks according to the types of the query/gate and the key/value:

$$\boldsymbol{Z}^{\bullet\to\diamond} = \boldsymbol{S}^{\bullet\to\diamond}\boldsymbol{V}^{\diamond} \odot \sigma(\boldsymbol{X}^{\bullet}\boldsymbol{W}_G) = \phi\left(\boldsymbol{Q}^{\bullet}(\boldsymbol{K}^{\diamond})^{\top}\right)\boldsymbol{V}^{\diamond} \odot \sigma(\boldsymbol{X}^{\bullet}\boldsymbol{W}_G), \quad \bullet, \diamond \in \{I, A\}, \tag{5}$$

where $\bullet$ and $\diamond$ denote the token types on the query/gate side and the key/value side, respectively; $\boldsymbol{Q}^{\bullet}$ denotes the submatrix of $\boldsymbol{Q}$ indexed by type $\bullet$ (analogously for $\boldsymbol{X}^{\bullet}$, $\boldsymbol{K}^{\diamond}$, $\boldsymbol{V}^{\diamond}$); and $\to$ indicates the attention direction, with queries/gates on the left attending to keys/values on the right. This decomposition yields four heterogeneous attention interactions: $\boldsymbol{Z}^{I\to I}$, $\boldsymbol{Z}^{I\to A}$, $\boldsymbol{Z}^{A\to I}$, $\boldsymbol{Z}^{A\to A}$. For example, $\boldsymbol{Z}^{I\to A}$ corresponds to item token queries attending to action tokens: item queries compute attention weights over action keys, then aggregating and gating the action-side value representations.

4.2 Selective Value Nonlinearity

Based on the heterogeneous decomposition, we can next identify the collapse-prone channel that requires targeted remedy and introduce our proposed SVAN approach.

Identifying Collapse-Prone Interaction Channels. We next identify which heterogeneous interaction channels can geometrically induce dimensional collapse. Recall that each channel follows the multiplicative update in Equation (5), where the attention-aggregated value stream is modulated element-wise by a gating term computed from the query-side tokens. Since the value stream primarily governs representational expressiveness, whereas the other components (query, key, and gating) mainly control interaction strength [9], we next diagnose collapse by analyzing the geometry of the value representations in each channel. Crucially, the multiplicative update imposes a directional constraint from the gate on the value update: when the gate is low-rank (*e.g.,* due to low-cardinality token types), it restricts the directions along which the value stream can be updated, and this restriction is most consequential when the value stream itself is high-rank. Accumulated over training, such restricted updates progressively drag the high-rank value stream toward a low-dimensional subspace. This mechanism leads to

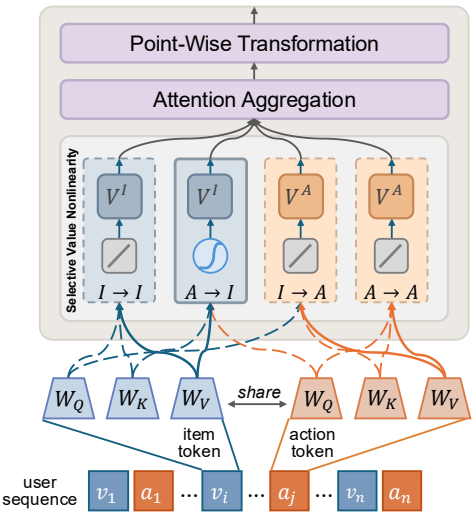


Figure 3: Overview of SVAN. We decompose attention into four heterogeneous interaction channels and identify $(A \to I)$ as the unique collapse-prone channel: a low-rank action gating confines the updates of the high-rank item value stream, pulling it toward a low-dimensional subspace. SVAN applies nonlinear activations to the value stream in this channel to counteract collapse while leaving all other channels unchanged.

interaction-induced collapse as characterized in prior work [12]. We next formalize this phenomenon at the level of heterogeneous interaction channels.

Definition 4.1 (**Collapse-Prone Channel**). Consider a single heterogeneous interaction in Equation (5), with value term $\boldsymbol{V}^{\diamond}$ and value-side gradient update $\Delta\boldsymbol{V}^{\diamond}$. We say this channel is collapse-prone if its high-rank value stream repeatedly receives low-rank updates, *i.e.,*

$$\operatorname{rank}(\boldsymbol{V}^{\diamond}) \gg \operatorname{rank}(\Delta\boldsymbol{V}^{\diamond}), \qquad \operatorname{rank}(\Delta\boldsymbol{V}^{\diamond}) \ll d. \tag{6}$$

Theorem 4.2 (**Low-Rank Gating Constrains Value Updates**). *Consider a heterogeneous interaction channel in Equation* (5) *with value term $\boldsymbol{V}^{\diamond}$ and gating term $\sigma(\boldsymbol{X}^{\bullet}\boldsymbol{W}_G)$. If the gating term is low-rank, then the value-side gradient update $\Delta\boldsymbol{V}^{\diamond}$ is confined to a low-dimensional subspace $\mathcal{S} \subseteq \mathbb{R}^d$ determined by the gating term, where d is the hidden dimension, i.e.,*

$$\operatorname{rank}(\Delta\boldsymbol{V}^{\diamond}) \;\le\; \dim(\mathcal{S}) \;\ll\; d. \tag{7}$$

The proof of Theorem 4.2 is provided in Appendix B.1. A key takeaway is the asymmetric learning dynamics: collapse arises when a high-rank value is repeatedly updated through a low-rank gating.

Corollary 4.3. *Among the four heterogeneous channels in Section 4.1, $(A \to I)$ is the only collapse-prone interaction channel induced by low-rank action tokens.*

To see this, note that item tokens have extremely high cardinality, so $\boldsymbol{V}^I$ typically exhibits high-rank geometry (up to the embedding dimension d). In contrast, action tokens have very low cardinality (only a few to dozens of types), which induces low-rank structure in the gate $\sigma(\boldsymbol{X}^A\boldsymbol{W}_G)$; hence, by Theorem 4.2, the updates of $\boldsymbol{V}^I$ in this channel are low-rank. Among the four channels, only

$$(A \to I): \quad \boldsymbol{Z}^{A\to I} = \boldsymbol{S}^{A\to I}\boldsymbol{V}^I \odot \sigma(\boldsymbol{X}^A\boldsymbol{W}_G)$$

matches Definition 4.1, since its high-rank item value stream repeatedly receives low-rank updates induced by the action gate. The other channels violate at least one premise: (i) $I \to I$ involves high-cardinality tokens on both sides, making the gate diverse rather than low-rank, so the value updates are not confined to a low-dimensional subspace; (ii) $I \to A$ and (iii) $A \to A$ use $\boldsymbol{V}^A$ on the value side, whose geometry is already low-rank, and thus do not satisfy $\operatorname{rank}(\boldsymbol{V}^{\diamond}) \gg \operatorname{rank}(\Delta\boldsymbol{V}^{\diamond})$. Therefore, $\boldsymbol{Z}^{A\to I}$ *constitutes the unique collapse-prone channel*, calling for a targeted remedy.

Mitigating Collapse via Selective Nonlinearity. Having isolated the collapse-prone interaction channel, we next introduce a targeted, geometry-preserving remedy. Motivated by prior observations that nonlinearity can effectively alleviate collapse with minimal modification [46], we propose to apply a nonlinear activation to the value component in the collapsed interaction channel, so as to counteract the contraction induced by multiplicative modulation. We begin by introducing:

Theorem 4.4. *For the collapse-prone interaction channel $\boldsymbol{Z}^{A\to I}$, replacing the value component by a pointwise nonlinear activation, i.e., $\boldsymbol{V}^I \mapsto \sigma(\boldsymbol{V}^I)$, reweights the update by the Jacobian of σ:*

$$\Delta'\boldsymbol{V}^I \;=\; \boldsymbol{J} \odot \Delta(\boldsymbol{V}^I), \qquad \boldsymbol{J} := \sigma'(\boldsymbol{V}^I). \tag{8}$$

Consequently, the nonlinear update lies in the diagonally reweighted span:

$$\mathcal{S}_{\mathrm{nl}} \;:=\; \operatorname{span}\Big(\bigcup_p \boldsymbol{J}_p\,\mathcal{S}\Big), \tag{9}$$

where p indexes tokens (rows). In general $\mathcal{S}_{\mathrm{nl}} \supseteq \mathcal{S}$, and $\dim(\mathcal{S}_{\mathrm{nl}})$ can be substantially larger than $\dim(\mathcal{S})$, thereby mitigating the dimensional collapse of value components.

The proof of Theorem 4.4 is provided in Appendix B.2. Intuitively, this nonlinearity breaks the purely linear contraction induced by repeated low-rank modulation, thereby mitigating the progressive concentration of singular values. To this end, we define an interaction-aware nonlinear operator:

$$\boxed{\overline{\sigma}(X;\bullet,\diamond) = \begin{cases}\sigma(X) & \text{if } \bullet = A,\ \diamond = I,\\ X & \text{otherwise,}\end{cases}} \tag{10}$$

and use it to construct a selectively nonlinear value stream,

$$\overline{\boldsymbol{V}}^{\diamond} = \overline{\sigma}(\boldsymbol{V}^{\diamond};\bullet,\diamond)\,. \tag{11}$$

The resulting heterogeneous interaction becomes

$$\overline{\boldsymbol{Z}}^{\bullet\to\diamond} = \boldsymbol{S}^{\bullet\to\diamond}\overline{\boldsymbol{V}}^{\diamond} \odot \sigma(\boldsymbol{X}^{\bullet}\boldsymbol{W}_G), \tag{12}$$

which applies nonlinear activations only to the $(A \to I)$ channel and leaves the others unchanged.

4.3 Discussion

The Necessity of Selectivity. A natural alternative to SVAN is to apply nonlinear activations to the value stream universally, *i.e.,* replacing $\boldsymbol{V}$ with $\sigma(\boldsymbol{V})$ in all channels. We argue that such a blanket modification is generally unnecessary and can even harm effective dimensionality of representations. Concretely, a coordinate-wise activation may introduce additional contraction in the value space: for instance, SiLU compresses responses in the negative-input region, which can reduce the spread and diversity of representations. This is particularly undesirable for channels that are not collapsed (*e.g.,* $(A \to A)$), where extra contraction provides little benefit but may limit the model's capacity. In contrast, SVAN applies activations selectively—only on the collapsed channel $(A \to I)$, yielding the minimal intervention to counteract collapse while preserving the geometry of the remaining channels.

Generalization to Multi-Layer GRMs. Although the high-/low-cardinality distinction is most explicit at the input layer (original item/action embeddings), the collapse mechanism is inherently depth-amplifying. Each additional layer introduces a new round of heterogeneous mixing. Consequently, the collapse can accumulate across layers. Empirically, our preliminary collapse analysis (Figure 2(c)) shows that the effective rank of item-position hidden states decreases as the depth increases. Importantly, at every depth, item positions consistently maintain a higher effective rank than action positions (*Finding 2*), indicating that the relative rank ordering between item and action positions persists even after multiple rounds of mixing. This makes SVAN directly extensible to deep GRMs: since token types are tied to positions, we can decompose each layer into the same four heterogeneous channels, and the only channel that continues to satisfy the collapse-prone channel configuration remains $(A \to I)$. Formally, for each layer $\ell = 1, \ldots, L$, we apply the same selective rule $\overline{\boldsymbol{V}}^{(\ell)\diamond} = \overline{\sigma}(\boldsymbol{V}^{(\ell)\diamond}; \bullet, \diamond)$, while all other channels remain unchanged. Intuitively, SVAN serves as a repeated safeguard at every depth, preventing the progressive rank decay induced by stacking the same collapsed heterogeneous interaction. Further discussion on complexity is in Appendix C.

5 Experiments

5.1 Setup

Datasets and Evaluation Protocols. We conducted experiments on three widely adopted public datasets: IJCAI [43], Retail [43] and KuaiSAR [34]. We first followed the 5-core setting to ensure the data quality. Following [48], we left out the last 5 behaviors as the target and split the dataset into training, validation, and test sets following the ratio of 70%/10%/20%. Basic statistics are summarized in Table 4. As for evaluation, we use AUC and LogLoss. All reported results are averaged over 3 random seeds. We use the AUC performance on the validation set for early stopping.

Baselines We compare SVAN with recommendation models including: 1) *conventional models*: DIN [56], DCNv2 [39], DIEN [57], and DeepFM [11]; 2) *generative models*: PinRec [1], SASRec [21], GenRank [17], and HSTU [48]. Their descriptions and implementation details are in Appendix A.

5.2 Results

Overall Performance. Table 1 summarizes the overall ranking results. SVAN consistently delivers the strongest performance across all three datasets, achieving the best AUC and the lowest LogLoss among all compared methods. In particular, SVAN improves AUC over the best baseline by 0.81%, 0.28%, and 1.36%, and reduces LogLoss by 4.15%, 0.45%, and 2.53% on the three datasets, respectively. These margins are practically significant: a 0.001 absolute AUC improvement is widely acknowledged in the recommendation literature to translate into measurable profit gains [17, 25], indicating that the gains delivered by SVAN are non-trivial. We also observe that generative models generally outperform traditional feature-interaction baselines, highlighting the benefit of modeling long-range dependencies and leveraging action supervision within a unified sequence. HSTU is a particularly strong performer, which we attribute to two complementary factors: (i) its pointwise aggregated attention architecture, which effectively handles long action-rich sequences; and (ii) its item–action interleaving organization, which preserves both items and actions as first-class tokens and enables fine-grained target-aware interactions, unlike the non-interleaving designs adopted by

Method	IJCAI		Retail		KuaiSAR	
	AUC↑	LogLoss↓	AUC↑	LogLoss↓	AUC↑	LogLoss↓
DIN	0.8373±0.0021	0.5826±0.0013	0.7351±0.0011	0.6580±0.0019	0.7326±0.0085	0.2119±0.0031
DCNv2	0.8242±0.0007	0.5859±0.0014	0.7126±0.0008	0.6766±0.0017	0.7783±0.0019	0.2283±0.0028
DIEN	0.8218±0.0011	0.5613±0.0024	0.7208±0.0038	0.6783±0.0076	0.7258±0.0017	0.2247±0.0035
DeepFM	0.8195±0.0013	0.5878±0.0004	0.7014±0.0009	0.6826±0.0014	0.7623±0.0022	0.2362±0.0018
PinRec	0.8231±0.0018	0.5650±0.0035	0.7428±0.0009	0.6519±0.0098	0.7462±0.0011	0.2087±0.0005
SASRec	0.8167±0.0023	0.5691±0.0064	0.7412±0.0019	0.6541±0.0040	0.7507±0.0064	0.2157±0.0023
GenRank	0.8284±0.0019	0.5737±0.0015	0.7454±0.0008	0.6484±0.0011	0.7343±0.0047	0.2031±0.0024
HSTU	<u>0.8489±0.0006</u>	<u>0.5379±0.0013</u>	<u>0.7517±0.0015</u>	<u>0.6434±0.0021</u>	<u>0.7779±0.0029</u>	<u>0.2012±0.0033</u>
SVAN	**0.8558±0.0003**	**0.5156±0.0008**	**0.7538±0.0014**	**0.6405±0.0019**	**0.7885±0.0027**	**0.1961±0.0030**

Table 1: Overall ranking performance (*mean* ± *std*). Best are in **bold** and runner-up are <u>underlined</u>.

other generative baselines. SVAN further outperforms HSTU consistently, suggesting that selectively placing nonlinearity along collapsed paths improves the ranking performance.

Ablations on Other Attention Components Following [48], we apply nonlinear activations to the query/key projections by default. We further ablate these activations to assess their necessity. As shown in Table 2, removing the nonlinear activations on Q/K consistently degrades performance, reducing AUC by 1.07%, 0.89%, and 0.93% on the three datasets, respectively, which confirms their importance and aligns with the design choice in [48]. We also examine the role of the gating mechanism. Since the gate multiplicatively modulates the value update, one may expect that removing it could weaken item–action multiplicative interactions and thus alleviate collapse. However, eliminating the gating term consistently hurts performance, reducing AUC by 0.82%, 0.61%, and 5.40% on the three datasets, respectively. This indicates that collapse cannot be addressed by simply discarding the gate; effective remedies must preserve its benefits.

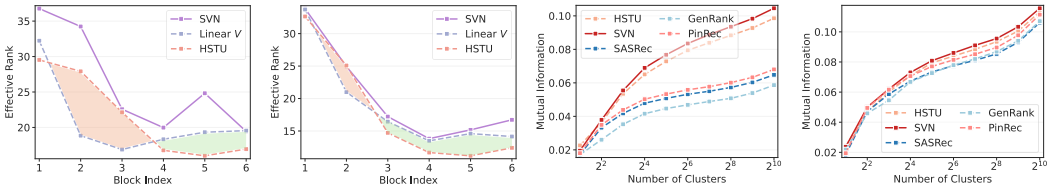
Variants	IJCAI		Retail		KuaiSAR	
	AUC↑	LogLoss↓	AUC↑	LogLoss↓	AUC↑	LogLoss↓
Nonlinear Type (HSTU)						
SiLU	0.8489	0.5379	0.7517	0.6434	0.7779	0.2012
ReLU	0.8482	0.5421	0.7529	0.6449	0.7822	0.2069
GELU	0.8470	0.5407	0.7543	0.6389	0.7848	0.2007
Nonlinear Type (SVAN)						
SiLU	0.8558	0.5156	0.7538	0.6405	0.7885	0.1961
ReLU	0.8496	0.5409	0.7537	0.6400	0.7914	0.2002
GELU	0.8492	0.5329	0.7551	0.6319	0.7913	0.1995
Ablations on Q / K / gating term						
Linear Q / K	0.8466	0.5353	0.7471	0.6482	0.7812	0.2137
w/o gating	0.8488	0.5203	0.7492	0.6470	0.7459	0.2143

Table 2: Ranking performance of HSTU and SVAN with different activations, along with ablations on other attention components (*i.e.,* Q, K, and the gating term).

Nonlinear Strategy on V SVAN adopts a selective nonlinear strategy on the value stream. To assess whether this choice is indeed the most effective, we evaluate several variants that differ in where the nonlinear activation is applied to V, and obtain the following observations. **(i) Making V fully linear degrades performance.** Disabling the activation on V entirely leads to consistently worse results compared to SVAN (Table 3 (a,c)). Specifically, making the value stream fully linear reduces AUC by 1.52%, 0.24%, and 0.53% on the three datasets, respectively. A fully linear value stream is more vulnerable to collapse, particularly in the collapsed channel where low-cardinality action tokens modulate high-rank item values and progressively constrain them to a low-rank subspace. **(ii) V should be activated, but in a targeted way.** Comparing Table 3(c) with (a,b), selectively activating V only on the collapsed action→item channel yields the best results, outperforming both fully linear and fully nonlinear strategies. We have discussed the underlying mechanism in Section 4.3 and will provide additional analysis in Section 5.3. Moreover, we test alternative selective schemes that activate V on other channel subsets while keeping the remaining channels linear; the results in Table 3(d-f) consistently underperform (c), indicating that the best practice is to activate V only on the collapsed channel. Notably, the fully activated variant performs worse than most selective strategies, which again suggests that indiscriminate nonlinearity could hurt performance.

Variant	Channel Nonlinearity of V				IJCAI		Retail		KuaiSAR	
	$I \to I$	$A \to I$	$I \to A$	$A \to A$	AUC↑	LogLoss↓	AUC↑	LogLoss↓	AUC↑	LogLoss↓
Reference Baselines										
(a) SVAN	✗	✓	✗	✗	0.8558	0.5156	0.7538	0.6405	0.7885	0.1961
(b) HSTU	✓	✓	✓	✓	0.8489	0.5379	0.7517	0.6434	0.7779	0.2012
Nonlinear Strategy on V										
(c) Linear V	✗	✗	✗	✗	0.8428	0.5495	0.7520	0.6430	0.7843	0.1966
(d) Act $I \to I$	✓	✗	✗	✗	0.8445	0.5494	0.7516	0.6422	0.7856	0.2054
(e) Act $I \to A$	✗	✗	✓	✗	0.8451	0.5455	0.7518	0.6413	0.7856	0.2032
(f) Act $A \to A$	✗	✗	✗	✓	0.8413	0.5506	0.7522	0.6426	0.7820	0.2001

Table 3: Ranking performance of variants under different value-stream nonlinear activation strategies.



(a) Effective rank - IJCAI (b) Effective rank - Retail (c) Discrim - IJCAI (d) Discrim - Retail

Figure 4: (a-b) Effective rank of item value representations before activation across layers on two datasets; lower values indicate more severe collapse. The shaded region " " indicates layers dominated by interaction-induced collapse, while " " indicates layers where an effective-dimensionality bottleneck dominates. (c-d) Representation discriminability measured by the mutual information between the quantized representations and labels. Results on the KuaiSAR dataset are in Appendix D.

Nonlinear Type To examine whether our gains depend on a particular activation, we replace the default nonlinear activation in both HSTU and SVAN with three commonly used functions: SiLU, ReLU, and GELU. Table 2 reports the results. Across all datasets and activation choices, SVAN consistently outperforms its HSTU counterpart, indicating that the benefits of selective nonlinearity are not tied to a specific nonlinear type and thus generalize well.

5.3 Analysis

Selective Nonlinearity Mitigates Dimensional Collapse We investigate whether SVAN 's selective nonlinearity can mitigate dimensional collapse. We compare SVAN with HSTU and a variant that makes V fully linear. Following Equation (4), we randomly sample a batch from the validation set and compute the effective rank of the item-token value representations before selective activation at each layer. As shown in Figure 4(a-b), SVAN consistantly attains a higher effective rank than both variants across layers, validating its effectiveness. Averaged over layers, SVAN achieves an effective-rank improvement of 12.58% and 4.85% over the best-performing variant on the IJCAI and Retail datasets, respectively. In shallow layers (e.g., layers 1–3), the fully linear V variant quickly loses rank because high-cardinality item values are directly exposed to low-cardinality action signals without the protection of nonlinearity, making collapse more likely. In deeper layers, the fully linear-V variant can exhibit a comparable (or even higher) effective rank than HSTU (e.g., layers 4–6). One plausible explanation is that applying nonlinear activations everywhere may introduce an implicit bottleneck after repeated aggregations, since excessive activations can compress or sparsify representations, leading to a lower measured rank in later layers, as dicussed in Section 4.3. Overall, SVAN strikes a favorable trade-off: it alleviates collapse while preserving effective dimensionality, suggesting that targeted value nonlinearity is preferable to either global activation or full linearization.

Selective Nonlinearity Improves Representation Discriminability We next analyze the discriminability of the learned representations. Concretely, follwing [36, 9], on each validation dataset we extract the hidden representation of the candidate-item token h for every sample and assess their representation space. We then apply K-means to quantize the space, mapping each h to a discrete cluster assignment $Q(h)$, and measure discriminability via the mutual information between $Q(h)$ and the ground-truth action label a: $\text{Discrim}(h, a) = \text{MI}(Q(h), a)$. As reported in Figure 4(c-d),

SVAN consistently achieves higher MI than the baselines across cluster granularities, indicating that selective nonlinearity improves representation discriminability. Notably, SASRec and GenRank perform worse, likely because they entangle action semantics with item identity and thus struggle to capture fine-grained behavior structure, while HSTU improves upon them but still suffers from reduced discriminability due to collapse when all value components are activated. In contrast, SVAN introduces nonlinear activations only on the collapsed channel, which mitigates collapse while preserving representation diversity, leading to more action-separable candidate representations.

6 Related Works

Generative Recommendation. With the remarkable advances in large language models (LLMs) [45, 5, 18], the recommender systems community has increasingly explored the potential of generative modeling to reshape traditional recommendation paradigms. Consequently, generative recommendation has recently emerged as a compelling paradigm, departing from discriminative scoring by casting recommendation as sequential generation. An influential research line [48] explicitly introduces action tokens into the sequence, typically interleaving items and actions to enable target-aware interactions. Follow-up studies refine this behavior-tokenization paradigm for greater flexibility and efficiency [1, 17, 10]. Along a complementary direction, another line [30, 40, 15, 8] represents items as semantic IDs, often learned via quantization such as RQ-VAE [22], and performs next-token prediction over these codes. In this work, we focus on the action-tokenization paradigm, identify a fundamental failure mode, *i.e.,* dimensional collapse induced by low-cardinality action tokens, and mitigate it with a simple yet effective approach.

Dimensional Collapse. Dimensional collapse confines learned representations to a low-dimensional subspace, limiting their expressive capacity. It has been studied in self-supervised and federated learning, graph learning, and Transformer attention [20, 14, 32, 53, 38]. In recommendation, remedies include decorrelation, spectrum balancing, and uniformity-promoting objectives or layers [52, 2, 42, 27, 51, 47]. Interaction-collapse analyses show that low-information fields can constrain other embeddings [12, 26]. Architectural remedies use diverse or decoupled embeddings, nonlinear feature generation, and rank-aware architectures [24, 9, 46, 37, 23, 4, 33, 7]. Related bottlenecks arise in unified sequence–feature modeling, diffusion recommendation, and semantic or LLM-derived representations [58, 3, 41, 50, 6, 55, 49, 16]. Building on these insights, we identify collapse induced by low-cardinality action tokens in generative ranking and selectively apply value nonlinearity to the action-query/item-value channel.

7 Conclusion

In this paper, we identify an action-induced dimensional collapse issue in modern GRMs. We decompose the attention interactions into heterogeneous channels, and propose Selective VAlue Nonlinearity (SVAN) to selectively apply nonlinear activations to the value stream in the collapse-prone channel. Extensive experiments validate that SVAN consistently improves ranking performance over strong baselines while effectively mitigating collapse and improving representation discriminability.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China (Grant No.92370204), in part by the National Key R&D Program of China (Grant No.2023YFF0725001), in part by Guangdong Provincial Key Laboratory of Frontier Basic Science for All-domain Intelligence, in part by the guangdong Basic and Applied Basic Research Foundation (Grant No.2023B1515120057), in part by the Key-Area Special Project of Guangdong Provincial Ordinary Universities (2024ZDZX1007).

References

[1] Anirudhan Badrinath, Prabhat Agarwal, Laksh Bhasin, Jaewon Yang, Jiajing Xu, and Charles Rosenberg. Pinrec: Outcome-conditioned, multi-token generative retrieval for industry-scale recommendation systems. *arXiv preprint arXiv:2504.10507*, 2025.

[2] Huiyuan Chen, Vivian Lai, Hongye Jin, Zhimeng Jiang, Mahashweta Das, and Xia Hu. Towards mitigating dimensional collapse of representations in collaborative filtering. In *Proceedings of the 17th ACM international conference on web search and data mining*, pages 106–115, 2024.

[3] Jialei Chen, Yuanbo Xu, and Yiheng Jiang. Unlocking the Power of Diffusion Models in Sequential Recommendation: A Simple and Effective Approach. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 155–166. ACM, 2025. doi: 10.1145/3711896.3737172. URL https://doi.org/10.1145/3711896.3737172.

[4] Jin Chen, Shangyu Zhang, Bin Hu, Chao Zhou, Junwei Pan, Gengsheng Xue, Wentao Ning, Gengyu Weng, Wang Zheng, Shaohua Liu, Zeen Xu, Chengyuan Mai, Shijie Quan, Tingyu Jiang, Lifeng Wang, Shudong Huang, Chengguo Yin, Haijie Gu, and Jie Jiang. RankUp: Towards High-rank Representations for Large Scale Advertising Recommender Systems. arXiv preprint arXiv:2604.17878, 2026. URL https://arxiv.org/abs/2604.17878.

[5] Shuting Cui, Ying Sun, Yuting Zhang, Qingxin Meng, and Hengshu Zhu. Llm-enhanced career knowledge graph understanding for job mobility prediction. *ACM Transactions on Management Information Systems*, 2026.

[6] Yu Cui, Feng Liu, Zhaoxiang Wang, Changwang Zhang, Jun Wang, Can Wang, and Jiawei Chen. SpecTran: Spectral-Aware Transformer-based Adapter for LLM-Enhanced Sequential Recommendation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 234–245. ACM, 2026. doi: 10.1145/3805712.3809701. URL https://doi.org/10.1145/3805712.3809701.

[7] Yu Cui, Yi Xu, Jiahao Wang, Hao Zhang, Yu Zhang, Xiaoyi Zeng, Can Wang, Jinxin Hu, and Jiawei Chen. SpecFormer: Mitigating Embedding and Attention Collapse via Spectral-Aware Transformer for Recommendation. arXiv preprint arXiv:2607.24025, 2026. URL https://arxiv.org/abs/2607.24025.

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

[9] Ningya Feng, Junwei Pan, Jialong Wu, Baixu Chen, Ximei Wang, Qian Li, Xian Hu, Jie Jiang, and Mingsheng Long. Long-sequence recommendation models need decoupled embeddings. *arXiv preprint arXiv:2410.02604*, 2024.

[10] Hao Guo, Erpeng Xue, Lei Huang, Shichao Wang, Xiaolei Wang, Lei Wang, Jinpeng Wang, and Sheng Chen. Action is all you need: Dual-flow generative ranking network for recommendation. *arXiv preprint arXiv:2505.16752*, 2025.

[11] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, Xiuqiang He, and Zhenhua Dong. Deepfm: An end-to-end wide & deep learning framework for ctr prediction. *arXiv preprint arXiv:1804.04950*, 2018.

[12] Xingzhuo Guo, Junwei Pan, Ximei Wang, Baixu Chen, Jie Jiang, and Mingsheng Long. On the embedding collapse when scaling up recommendation models. *arXiv preprint arXiv:2310.04400*, 2023.

[13] Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

[14] Junlin He, Jinxiao Du, and Wei Ma. Preventing dimensional collapse in self-supervised learning via orthogonality regularization. *Advances in Neural Information Processing Systems*, 37: 95579–95606, 2024.

[15] Yupeng Hou, Jiacheng Li, Ashley Shin, Jinsung Jeon, Abhishek Santhanam, Wei Shao, Kaveh Hassani, Ning Yao, and Julian McAuley. Generating long semantic ids in parallel for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 956–966, 2025.

[16] Bin Huang, Xin Wang, Junwei Pan, Yongqi Zhou, Yifeng Zhou, Zhixiang Feng, Shudong Huang, Haijie Gu, and Wenwu Zhu. Asymmetric Generative Recommendation via Kronecker Residual Bridge and Multi-Faceted Hierarchical Quantization. arXiv preprint arXiv:2605.14512, 2026. URL https://arxiv.org/abs/2605.14512.

[17] Yanhua Huang, Yuqi Chen, Xiong Cao, Rui Yang, Mingliang Qi, Yinghao Zhu, Qingchang Han, Yaowei Liu, Zhaoyu Liu, Xuefeng Yao, et al. Towards large-scale generative ranking. *arXiv preprint arXiv:2505.04180*, 2025.

[18] Yang Ji and Ying Sun. OCP: Outlier-centric probing for dynamic structured pruning of LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6108–6124, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.276. URL https://aclanthology.org/2026.acl-long.276/.

[19] Jie Jiang, Shangyu Zhang, Shijie Quan, Zhongren Wang, Tianqu Zhuang, Junwei Pan, Bo Fu, Yilong Sun, Jieying Lin, Jushuo Chen, Xiaotian Li, Wenjie Li, Zhixiang Feng, Xian Hu, Huiting Deng, Hua Lu, Jinpeng Wang, Boqi Dai, Xiaoyu Chen, Bin Hu, Lili Huang, Yanwen Wu, Yeshou Cai, Qi Zhou, Huang Tang, Chunfeng Yang, Chengguo Yin, Tingyu Jiang, Lifeng Wang, Shudong Huang, Dapeng Liu, Yuekui Yang, Shu-Tao Xia, Lei Xiao, and Haijie Gu. Lfm4ads: Large foundation model for ads recommendation. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management*, CIKM '26, page 8310–8318, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725395. doi: 10.1145/3799682.3840163. URL https://doi.org/10.1145/3799682.3840163.

[20] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

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

[22] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11523–11532, 2022.

[23] Guoming Li, Shangyu Zhang, Junwei Pan, Wentao Ning, Jin Chen, Gengsheng Xue, Chao Zhou, Shudong Huang, Haijie Gu, and Menglin Yang. Expand More, Shrink Less: Shaping Effective-Rank Dynamics for Dense Scaling in Recommendation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 2569–2579. ACM, 2026. doi: 10.1145/3770855.3818049. URL https://doi.org/10.1145/3770855.3818049.

[24] Zhutian Lin, Junwei Pan, Haibin Yu, Xi Xiao, Ximei Wang, Zhixiang Feng, Shifeng Wen, Shudong Huang, Dapeng Liu, and Lei Xiao. Crocodile: Cross Experts Covariance for Disentangled Learning in Multi-Domain Recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 1839–1849. ACM, 2025. doi: 10.1145/3746252.3761332. URL https://doi.org/10.1145/3746252.3761332.

[25] Xiaoliang Ling, Weiwei Deng, Chen Gu, Hucheng Zhou, Cui Li, and Feng Sun. Model ensemble for click prediction in bing search ads. In *Proceedings of the 26th international conference on world wide web companion*, pages 689–698, 2017.

[26] Junwei Pan, Wei Xue, Ximei Wang, Haibin Yu, Xun Liu, Shijie Quan, Xueming Qiu, Dapeng Liu, Lei Xiao, and Jie Jiang. Ads recommendation in a collapsed and entangled world. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5566–5577, 2024.

[27] Shaowen Peng, Kazunari Sugiyama, Xin Liu, and Tsunenori Mine. Balancing Embedding Spectrum for Recommendation. *ACM Transactions on Recommender Systems*, 3(4):1–25, 2025. doi: 10.1145/3718488. URL https://doi.org/10.1145/3718488.

[28] Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *arXiv preprint arXiv:2505.06708*, 2025.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[30] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315, 2023.

[31] Steffen Rendle. Factorization machines. In *2010 IEEE International conference on data mining*, pages 995–1000. IEEE, 2010.

[32] Yujun Shi, Jian Liang, Wenqing Zhang, Chuhui Xue, Vincent YF Tan, and Song Bai. Understanding and mitigating dimensional collapse in federated learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2936–2949, 2023.

[33] Chen Sun, Beilin Xu, Boheng Tan, Jiacheng Wang, Yuefeng Sun, Rite Bo, Ying He, Yaqiang Zang, and Pinghua Gong. OCP: Orthogonal Constrained Projection for Sparse Scaling in Industrial Commodity Recommendation. arXiv preprint arXiv:2603.18697, 2026. URL `https://arxiv.org/abs/2603.18697`.

[34] Zhongxiang Sun, Zihua Si, Xiaoxue Zang, Dewei Leng, Yanan Niu, Yang Song, Xiao Zhang, and Jun Xu. Kuaisar: A unified search and recommendation dataset. 2023. doi: 10.1145/3583780.3615123. URL `https://doi.org/10.1145/3583780.3615123`.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] Elena Voita, Rico Sennrich, and Ivan Titov. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. *arXiv preprint arXiv:1909.01380*, 2019.

[37] Jiancheng Wang, Mingjia Yin, Hao Wang, and Enhong Chen. Understanding DNNs in Feature Interaction Models: A Dimensional Collapse Perspective. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4267–4272. ACM, 2026. doi: 10.1145/3805712.3809983. URL `https://doi.org/10.1145/3805712.3809983`.

[38] Junxuan Wang, Xuyang Ge, Wentao Shu, Zhengfu He, and Xipeng Qiu. Dimensional Collapse in Transformer Attention Outputs: A Challenge for Sparse Dictionary Learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 126304–126331. PMLR, 2026. URL `https://proceedings.mlr.press/v306/wang26bh.html`.

[39] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*, pages 1785–1797, 2021.

[40] Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2400–2409, 2024.

[41] Yuhao Wang, Junwei Pan, Xinhang Li, Maolin Wang, Yuan Wang, Yue Liu, Dapeng Liu, Jie Jiang, and Xiangyu Zhao. Empowering Large Language Model for Sequential Recommendation via Multimodal Embeddings and Semantic IDs. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 3209–3219. ACM, 2025. doi: 10.1145/3746252.3761169. URL `https://doi.org/10.1145/3746252.3761169`.

[42] Wei Wu, Chao Wang, Dazhong Shen, Chuan Qin, Liyi Chen, and Hui Xiong. AFDGCF: Adaptive Feature De-correlation Graph Collaborative Filtering for Recommendations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1242–1252. ACM, 2024. doi: 10.1145/3626772.3657724. URL https://doi.org/10.1145/3626772.3657724.

[43] Lianghao Xia, Yong Xu, Chao Huang, Peng Dai, and Liefeng Bo. Graph meta network for multi-behavior recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 757–766, 2021.

[44] Haoran Xin, Ying Sun, Chao Wang, and Hui Xiong. Llmcdsr: Enhancing cross-domain sequential recommendation with large language models. *ACM Transactions on Information Systems*, 43(5):1–33, 2025.

[45] Haoran Xin, Anhao Zhao, Ying Sun, Jin Li, Xiaoyu Shen, and Hui Xiong. Escaping the kl agreement trap in on-policy distillation. *arXiv preprint arXiv:2606.09471*, 2026.

[46] Mingjia Yin, Junwei Pan, Hao Wang, Ximei Wang, Shangyu Zhang, Jie Jiang, Defu Lian, and Enhong Chen. From feature interaction to feature generation: A generative paradigm of ctr prediction models. *arXiv preprint arXiv:2512.14041*, 2025.

[47] Yonghong Yu, Zhiyu Wang, Yujie Liao, Li Zhang, and Rong Gao. Wasserstein distance-based graph contrastive learning for recommendation. *Expert Systems with Applications*, 297:129427, 2026. doi: 10.1016/j.eswa.2025.129427. URL https://doi.org/10.1016/j.eswa.2025.129427.

[48] Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, et al. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *arXiv preprint arXiv:2402.17152*, 2024.

[49] Taolin Zhang, Junwei Pan, Jinpeng Wang, Yaohua Zha, Tao Dai, Bin Chen, Ruisheng Luo, Xiaoxiang Deng, Yuan Wang, Ming Yue, Jie Jiang, and Shu-Tao Xia. Towards Scalable Semantic Representation for Recommendation. arXiv preprint arXiv:2410.09560, 2024. URL https://arxiv.org/abs/2410.09560.

[50] Weizhi Zhang, Liangwei Yang, Wooseong Yang, Henry Peng Zou, Yuqing Liu, Ke Xu, Sourav Medya, and Philip S. Yu. LLMInit: A Free Lunch from Large Language Models for Selective Initialization of Recommendation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2016–2024. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-industry.141. URL https://aclanthology.org/2025.emnlp-industry.141/.

[51] Yabin Zhang, Jiakai Tang, and Xu Chen. Alleviating Dimensional Collapse Problem in Deep Recommender Models by Designing Uniformity Layers. In *Database Systems for Advanced Applications*, volume 14852 of *Lecture Notes in Computer Science*, pages 148–163. Springer Nature Singapore, 2025. doi: 10.1007/978-981-97-5555-4_10. URL https://doi.org/10.1007/978-981-97-5555-4_10. DASFAA 2024.

[52] Yifei Zhang, Hao Zhu, Zixing Song, Piotr Koniusz, Irwin King, et al. Mitigating the popularity bias of graph collaborative filtering: A dimensional collapse perspective. *Advances in Neural Information Processing Systems*, 36:67533–67550, 2023.

[53] Yifei Zhang, Hao Zhu, Menglin Yang, Jiahong Liu, Rex Ying, Irwin King, and Piotr Koniusz. Understanding and mitigating hyperbolic dimensional collapse in graph contrastive learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1984–1995, 2025.

[54] Yuting Zhang, Ziliang Pei, Chao Wang, Ying Sun, and Fuzhen Zhuang. Enhancing llm-based recommendation with preference hint discovery from knowledge graph. *CoRR*, abs/2601.18096, 2026.

[55] Yuheng Zheng, Yu Cui, Bin Wu, Jian Zhang, Ye Feng, Can Wang, and Jiawei Chen. IMFuse: Instance-Aware Multi-Layer Fusion for LLM-Enhanced Sequential Recommendation. arXiv preprint arXiv:2607.27002, 2026. URL https://arxiv.org/abs/2607.27002.

Dataset	IJCAI	Retail	KuaiSAR
#Users	422,195	147,894	25,675
#Items	598,024	98,778	561,624
#Interactions	35,450,356	7,658,037	954,1871
#Action types	4	4	5
Avg. seq. length	83.97	51.78	371.64

Table 4: Dataset statistics.

[56] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1059–1068, 2018.

[57] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5941–5948, 2019.

[58] Yifeng Zhou, Yuehong Hu, Zhixiang Feng, Junwei Pan, Kaihui Wu, Hanyong Li, Shangyu Zhang, Shudong Huang, Zhangbin Zhu, Chengguo Yin, Haijie Gu, and Jie Jiang. TokenFormer: Unify the Multi-Field and Sequential Recommendation Worlds. arXiv preprint arXiv:2604.13737, 2026. URL `https://arxiv.org/abs/2604.13737`.

Appendix

A Experimental Setups

A.1 Baseline Descriptions

We first compare our model with the following conventional recommendation models:

- **DIN** [56] introduces a local activation unit to adaptively calculate user interest representations by performing a weighted sum of historical behaviors based on their relevance to the target item, rather than using a fixed-length vector.
- **DCNv2** [39] utilizes a cross network with a mixture of low-rank experts to efficiently learn explicit high-order feature interactions, combining them with a deep network to capture implicit interactions for improved expressiveness and cost-efficiency.
- **DIEN** [57] employs an interest extractor layer with an auxiliary loss and an interest evolving layer based on a GRU with an attentional update gate to capture interest trajectories relative to the target item, focusing on the dynamic evolution of user interests.
- **DeepFM** [11] is an end-to-end framework that integrates factorization machines for low-order feature interactions and deep neural networks for high-order interactions, sharing the same input embeddings to eliminate the need for manual feature engineering.

We also compare our model with the following generative recommendation models:

- **PinRec** [1] is a transformer-based generative recommendation model that utilizes outcome-conditioning to optimize for diverse business metrics.
- **SASRec** [21] applies a self-attention mechanism to sequential recommendation, adaptively assigning weights to previous items to model long-term semantics and predict the next item without relying on recurrent or convolutional operations.
- **GenRank** [17] is a generative ranking architecture designed for large-scale settings that utilizes an action-oriented sequence organization to iteratively predict user behaviors.
- **HSTU** [48] is a high-performance encoder for generative recommenders that replaces standard softmax attention with a pointwise aggregated attention mechanism to effectively handle user action sequences.

A.2 Implementation Details

For conventional models, we use a learning rate of $1e{-}3$, batch size 128, embedding dimension 50, and the Adam optimizer. Unless otherwise specified, the DNN hidden units are set to $[64, 32]$. For DIN, we use adopt DICE as the activation for the attention MLP. For DCNv2, we use the parallel architecture; the hidden units of both the parallel and stacked DNN branches are set to $[64, 32]$, the activation is ReLU, and the number of cross layers is 3. For DIEN, the attention network uses hidden units $[80, 40]$, and the auxiliary network uses hidden units $[100, 50]$; we adopt ARGRU as the GRU variant with bilinear attention, and use ReLU as the activation. For DeepFM, we use ReLU as the activation.

For generative baselines, we use 6 Transformer blocks with 1 attention head, learning rate $1e{-}3$, batch size 128, embedding dimension 50, and the Adam optimizer. For PinRec, we instantiate the backbone with a GPT-2 style Transformer [29]. For SASRec, we merge item and action tokens by element-wise addition. For GenRank and HSTU, we use SiLU as the default activation function. To prevent information leakage, we apply an action masking on candidate items for GenRank, SASRec, and PinRec. In addition, for all generative models we adopt the candidate mask introduced in [17] to prevent candidate items from attending to each other.

B Proof

B.1 Proof of Theorem 4.2

We prove the claim for the heterogeneous interaction channel in Eq. (5). Let n be the number of tokens within a sequence. For clarity, we denote:

$$\boldsymbol{Z} := \boldsymbol{Z}^{\bullet\to\diamond} = (\boldsymbol{S}\boldsymbol{V}) \odot \boldsymbol{G} \in \mathbb{R}^{n\times d}, \quad \boldsymbol{S} := \boldsymbol{S}^{\bullet\to\diamond} \in \mathbb{R}^{n\times n}, \quad \boldsymbol{V} := \boldsymbol{V}^{\diamond} \in \mathbb{R}^{n\times d}, \quad \boldsymbol{G} := \sigma(\boldsymbol{X}^{\bullet}\boldsymbol{W}_G) \in \mathbb{R}^{n\times d}. \tag{B.1}$$

We consider one training step and focus on how the interaction between $\boldsymbol{V}$ and $\boldsymbol{G}$ constrains the update direction of $\boldsymbol{V}$. The argument applies to an arbitrary layer $\ell \in \{1, \dots, L\}$; we omit the layer superscript (ℓ) when no ambiguity arises, and all quantities below are understood as layer-specific. Let $\mathcal{L}$ denote the loss of the whole network.

Let $\boldsymbol{E} := \nabla_{\boldsymbol{Z}}\mathcal{L} \in \mathbb{R}^{n\times d}$ be the gradient at $\boldsymbol{Z}$. The chain rule yields

$$\nabla_{\boldsymbol{V}}\mathcal{L} \;=\; \boldsymbol{S}^{\top}(\boldsymbol{E} \odot \boldsymbol{G}). \tag{B.2}$$

Writing the gradient row-wise, for each value-side token index $p \in \{1, \dots, n\}$ with row embedding $\boldsymbol{v}_p \in \mathbb{R}^d$, we have

$$\nabla_{\boldsymbol{v}_p}\mathcal{L} \;=\; \sum_{i=1}^{n} S_{i,p}(\boldsymbol{e}_i \odot \boldsymbol{g}_i), \tag{B.3}$$

where $\boldsymbol{e}_i, \boldsymbol{g}_i \in \mathbb{R}^d$ are the i-th rows of $\boldsymbol{E}$ and $\boldsymbol{G}$, respectively.

Consider a channel with a low-rank gate at layer ℓ, *i.e.,* $\operatorname{rank}(\boldsymbol{G}) = r^{(\ell)} \ll d$. For $(A \to I)$ at the input layer, $r^{(1)} \le |\mathcal{A}|$ since each gate row is determined solely by its action type; at deeper layers, the action hidden states empirically remain low-rank (Figure 2(c)). Writing $r := r^{(\ell)}$, there exist r vectors $\boldsymbol{b}_1, \dots, \boldsymbol{b}_r \in \mathbb{R}^d$ such that

$$\forall i \in \{1, \dots, n\}, \qquad \exists \beta_{i1}, \dots, \beta_{ir} \in \mathbb{R}, \qquad \boldsymbol{g}_i = \sum_{j=1}^{r} \beta_{ij}\, \boldsymbol{b}_j. \tag{B.4}$$

Next, we capture the fact that, in generative ranking, supervision is on action logits with a very small label space. Concretely, the backpropagated vector $\boldsymbol{e}_i$ originates from a readout over a small action vocabulary, and task-related gradients are known to concentrate in low-dimensional subspaces [13]. Since the dimension of this subspace may vary with depth, we allow the signal rank to be layer-dependent. We formalize this as:

Assumption B.1 (Low-dimensional downstream signal)**.** For each layer ℓ, there exists a subspace $\mathcal{U}^{(\ell)} \subseteq \mathbb{R}^d$ with $\dim(\mathcal{U}^{(\ell)}) = k^{(\ell)} \ll d$ and basis vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_{k^{(\ell)}}$ such that $\boldsymbol{e}_i \in \mathcal{U}^{(\ell)}$ for all query-side indices i, i.e., $\boldsymbol{e}_i = \sum_{m=1}^{k^{(\ell)}} \alpha_{im}\boldsymbol{u}_m$.

This assumption holds exactly at the readout layer, for example, when the local loss is a cross-entropy/BCE over $|\mathcal{A}|$ action types with linear classifier weights $\boldsymbol{W}_{\text{out}} \in \mathbb{R}^{d\times|\mathcal{A}|}$, since then $\boldsymbol{e}_i \in \operatorname{span}\{\boldsymbol{W}_{\text{out}}(:,a) : a \in \mathcal{A}\}$ and thus $k \le |\mathcal{A}|$. For intermediate layers, the gradient is further mixed by the subsequent blocks during backpropagation, so $k^{(\ell)}$ may exceed $|\mathcal{A}|$ but still remains far below d, as empirically verified in Appendix B.3. Below we write $k := k^{(\ell)}$.

Combining Assumption B.1 and (B.4), for $i \in \{1,\dots,n\}$,

$$\boldsymbol{e}_i \odot \boldsymbol{g}_i \;=\; \Big(\sum_{m=1}^{k} \alpha_{im}\boldsymbol{u}_m\Big) \odot \Big(\sum_{j=1}^{r} \beta_{ij}\boldsymbol{b}_j\Big) \;=\; \sum_{m=1}^{k}\sum_{j=1}^{r} \alpha_{im}\beta_{ij}\,\big(\boldsymbol{u}_m \odot \boldsymbol{b}_j\big). \tag{B.5}$$

Therefore all interaction gradients $\boldsymbol{e}_i \odot \boldsymbol{g}_i$ lie in the fixed subspace

$$\mathcal{S} \;:=\; \operatorname{span}\{\boldsymbol{u}_m \odot \boldsymbol{b}_j :\; m = 1,\dots,k,\; j = 1,\dots,r\} \;\subseteq\; \mathbb{R}^d, \qquad \dim(\mathcal{S}) \le kr. \tag{B.6}$$

Substituting (B.5) into (B.3) yields that, for every value-token index p, the *signal term* $\sum_i S_{i,p}\,(\boldsymbol{e}_i \odot \boldsymbol{g}_i)$ lies in $\mathcal{S}$. Therefore, when the gate is low-rank ($\operatorname{rank}(\boldsymbol{G}) = r \ll d$) and the downstream signal is low-dimensional ($\dim(\mathcal{U}) = k \ll d$), all value-side updates $\Delta\boldsymbol{V} \propto -\nabla_{\boldsymbol{V}}\mathcal{L}$ have their rows in the interaction subspace $\mathcal{S}$, *i.e.,* $\operatorname{rank}(\Delta\boldsymbol{V}) \le \dim(\mathcal{S}) \le kr$. Since both k and r are governed by low-cardinality action signals, $\dim(\mathcal{S})$ remains small relative to d; Appendix B.3 empirically confirms that $k, r \ll d$ and that the resulting update rank stays well below d at every layer. Since this argument holds at every layer ℓ under Assumption B.1, it establishes Theorem 4.2. Note that the theorem constrains the *update* of $\boldsymbol{V}$ rather than the rank of $\boldsymbol{V}$ itself: a high-rank $\boldsymbol{V}$ that repeatedly receives such low-rank updates is collapse-prone in the sense of Definition 4.1. □

B.2 Proof of Theorem 4.4

We now prove that, in a collapsed channel, replacing $\boldsymbol{V}$ by $\sigma(\boldsymbol{V})$ in the value stream mitigates collapse by breaking the fixed low-dimensional update subspace.

Following Corollary 4.3, consider the collapsed channel $\boldsymbol{Z}^{A\to I}$ and write it generically as

$$\boldsymbol{Z} \;=\; (\boldsymbol{S}\,\sigma(\boldsymbol{V})) \odot \boldsymbol{G}, \tag{B.7}$$

where σ acts elementwise. As above, define $\mathcal{L}$ as the loss function, and let $\boldsymbol{E} := \nabla_{\boldsymbol{Z}}\mathcal{L}$.

Let $\boldsymbol{v}_p \in \mathbb{R}^d$ be the p-th row of $\boldsymbol{V}$. Since σ is applied elementwise, the Jacobian of σ at $\boldsymbol{v}_p$ is the diagonal matrix

$$\boldsymbol{J}_p \;:=\; \operatorname{Diag}\big(\sigma'(\boldsymbol{v}_p)\big) \in \mathbb{R}^{d\times d}. \tag{B.8}$$

A direct chain-rule calculation from (B.7) gives

$$\nabla_{\boldsymbol{v}_p}\mathcal{L} \;=\; \boldsymbol{J}_p\Big(\sum_{i=1}^{n} S_{i,p}\,(\boldsymbol{e}_i \odot \boldsymbol{g}_i)\Big) \in \mathbb{R}^d. \tag{B.9}$$

Compared to the linear-value case (B.3), the signal term is now *left-multiplied by* $\boldsymbol{J}_p$, which depends on $\boldsymbol{v}_p$ and varies across value tokens.

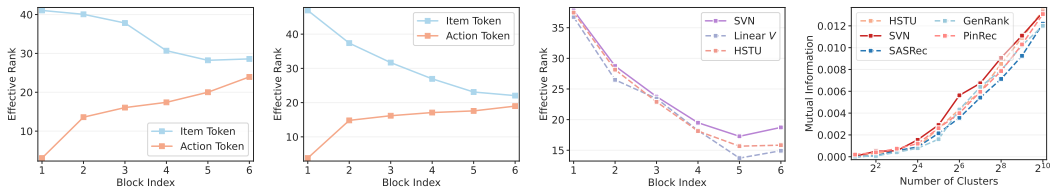
Let $\mathcal{S}$ be the (collapsed) signal subspace defined in (B.6) for the linear-value channel, i.e., $\sum_i S_{i,p}(\boldsymbol{e}_i \odot \boldsymbol{g}_i) \in \mathcal{S}$ for all p. Then (B.9) shows that the nonlinear-value signal lives in the *token-dependent* subspace $\boldsymbol{J}_p\mathcal{S}$. Define the aggregate nonlinear signal span

$$\mathcal{S}_{\text{nl}} \;:=\; \operatorname{span}\Big(\bigcup_{p=1}^{n} \boldsymbol{J}_p\mathcal{S}\Big) \;\subseteq\; \mathbb{R}^d. \tag{B.10}$$

In contrast to Theorem 4.2, where all signal directions lie in a *fixed* low-dimensional $\mathcal{S}$, here the signal subspace can expand with training through varying Jacobians $\{\boldsymbol{J}_p\}$.

We state a sufficient condition under which $\mathcal{S}_{\text{nl}}$ becomes large (up to full $\mathbb{R}^d$), thus removing the hard low-rank bottleneck.

Assumption B.2 (Activation diversity)**.** There exist indices $p_1,\dots,p_d$ and vector $\boldsymbol{s} \in \mathcal{S}$ such that (i) $\boldsymbol{s}$ has all coordinates nonzero; and (ii) the vectors $\{\boldsymbol{J}_{p_t}\boldsymbol{s}\}_{t=1}^{d}$ are linearly independent.



(a) Token rank - IJCAI (b) Token rank - KuaiSAR (c) Collapse - KuaiSAR (d) Discrim - KuaiSAR

Figure 5: Additional experimental results.

Assumption B.2 is mild in practice: for piecewise-linear activations (*e.g.,* ReLU), $\boldsymbol{J}_p$ is a 0–1 diagonal mask, and varying sign/activation patterns across item values typically yields diverse masks; for smooth activations (*e.g.,* SiLU/GELU), $\boldsymbol{J}_p$ remains diagonal with continuously varying entries, which also generically produces diverse scalings across tokens.

Lemma B.3 (Span expansion)**.** *Under Assumption B.2, we have* $\mathcal{S}_{\text{nl}} = \mathbb{R}^d$.

Proof. Since $\boldsymbol{s} \in \mathcal{S}$, for each t, $\boldsymbol{J}_{p_t}\boldsymbol{s} \in \boldsymbol{J}_{p_t}\mathcal{S} \subseteq \mathcal{S}_{\text{nl}}$. By Assumption B.2(ii), the set $\{\boldsymbol{J}_{p_t}\boldsymbol{s}\}_{t=1}^{d}$ is linearly independent, hence spans $\mathbb{R}^d$. Therefore $\mathbb{R}^d \subseteq \mathcal{S}_{\text{nl}}$, implying equality. □

Lemma B.3 shows that with the nonlinear value transform, the interaction gradients can span a much larger subspace than the fixed low-dimensional $\mathcal{S}$ in Theorem 4.2 (indeed, all of $\mathbb{R}^d$ under a mild diversity condition). This breaks the purely linear low-rank confinement induced by repeated low-rank gating, thereby mitigating the progressive concentration of singular values in the collapsed channel.

Applying this argument to $\boldsymbol{Z}^{A\to I}$ (which Corollary 4.3 identifies as collapse-prone, owing to a low-rank action gate acting on a high-rank item value stream) proves Theorem 4.4. □

B.3 Empirical Support for Assumption B.1

To examine whether the low-rank premises hold beyond the readout layer, we measure, for each layer ℓ of the default HSTU backbone on KuaiSAR, the effective rank (Equation (4)) of the backpropagated signal $\boldsymbol{E}^{(\ell)}$, the action gate $\boldsymbol{G}^{(\ell)}$, and the resulting value-side gradient $\nabla_{\boldsymbol{V}^{(\ell)}}\mathcal{L}$ in the $(A \to I)$ channel. As shown in Table 5: (i) the gate stays at an effective rank of 3.8–4.3 across layers, close to $|\mathcal{A}|$, supporting $r^{(\ell)} \ll d$; (ii) the backpropagated signal remains well below d at every layer, and decreases toward deeper layers that are closer to the readout, supporting $k^{(\ell)} \ll d$; (iii) the value-side update stays at 12.3–23.0, *i.e.,* at most 46% of d, confirming that the update is substantially lower-dimensional than d at every layer, as predicted by Theorem 4.2. Note that the update rank is not monotone in depth and exceeds that of $\boldsymbol{E}^{(\ell)}$ in the deepest layers; this is consistent with the theory, since $\mathcal{S}$ is spanned by the products $\boldsymbol{u}_m \odot \boldsymbol{b}_j$ and is bounded by $k^{(\ell)}r^{(\ell)}$ rather than $k^{(\ell)}$.

Layer ℓ	$\text{ERank}(\boldsymbol{E}^{(\ell)})$	$\text{ERank}(\boldsymbol{G}^{(\ell)})$	$\text{ERank}(\nabla_{\boldsymbol{V}^{(\ell)}}\mathcal{L})$
1	26.4	3.8	19.5
2	23.9	3.9	23.0
3	20.2	4.0	13.6
4	13.4	4.2	14.8
5	8.1	4.3	12.3

Table 5: Layer-wise effective ranks of the backpropagated signal, the action gate, and the value-side gradient in the $(A \to I)$ channel ($d = 50$).

C Dicussion on Complexity and Implementation

SVAN introduces no additional learnable parameters and requires only an extra element-wise nonlinear activation. The selective nonlinearity in Equation (11) can be implemented with a simple

Backbone	IJCAI		Retail		KuaiSAR	
	AUC↑	LogLoss↓	AUC↑	LogLoss↓	AUC↑	LogLoss↓
HSTU						
⊢ *w/o SVAN*	0.8489	0.5379	0.7517	0.6434	0.7779	0.2012
⊢ *w/ SVAN*	**0.8558**	**0.5156**	**0.7538**	**0.6405**	**0.7885**	**0.1961**
Llama						
⊢ *w/o SVAN*	0.8207	0.5784	0.7511	0.6443	0.7505	0.2147
⊢ *w/ SVAN*	**0.8482**	**0.5644**	**0.7533**	**0.6416**	**0.7572**	**0.2067**
Qwen						
⊢ *w/o SVAN*	0.8304	0.5592	0.7550	0.6383	0.7469	0.2167
⊢ *w/ SVAN*	**0.8483**	**0.5254**	**0.7554**	**0.6343**	**0.7648**	**0.2075**

Table 6: AUC and LogLoss of SVAN applied to different backbones. The interleaving conversion is performed at the input level only.

type-based mask: using the query/value token types, we identify whether a given interaction corresponds to the collapsed channel $(A \to I)$, and apply the nonlinear activation to the value stream only in that case (otherwise leaving it unchanged). In practice, this amounts to a lightweight modification to existing GRM implementations.

D Experimental Results

D.1 SVAN on Other Backbones

SVAN relies only on a per-position token type and is therefore portable across GRM backbones. In addition to HSTU (our default backbone), we evaluate it on three additional backbones — HSTU, Llama, and Qwen — after a sequence-level conversion to interleaved input (no architectural change). Table 6 reports AUC and LogLoss w/ and *w/o SVAN* across the three datasets; SVAN delivers consistent gains in every case.

D.2 Additional Experimental Results

We additionally report layer-wise effective-rank results for item and action tokens on IJCAI and KuaiSAR (Section 3.2), and dimensional collapse/discriminability results on KuaiSAR (Section 5.3). These results support the same conclusions. In Figure 5(a–b), the effective rank of item tokens decreases in deeper layers, whereas action tokens show the opposite trend; item tokens nonetheless remain higher-rank than action tokens throughout. In Figure 5(c–d), SVAN consistently improves both dimensional collapse and representation discriminability over the base models, further validating its effectiveness.

E Limitations and Broader Impact

Our analysis is centered on attention-based GRMs with an interleaved item–action organization, which is the dominant paradigm in recent generative recommendation [48]; extending the heterogeneous channel view to alternative paradigms such as semantic-ID-based generative retrieval [30] is an interesting direction left for future work. Our experiments span three public benchmarks of different scales and domains, while validation on industrial-scale deployment with billion-level catalogs would further strengthen the empirical picture; we expect SVAN to remain effective at such scales given its parameter-free design and architecture-agnostic nature. We also note that SVAN currently applies a uniform selective nonlinearity across layers, and layer-adaptive variants that modulate intervention strength based on per-layer collapse severity could potentially yield additional gains.

This work improves the representational quality of generative ranking models, which can translate into more accurate recommendation pipelines and better user experience. Potential negative impacts are those generic to recommender systems—such as filter bubbles and amplification of popularity bias—which are inherited from the underlying deployment rather than introduced by SVAN, and

standard mitigation strategies (e.g., diversity-aware ranking and fairness-aware objectives) remain fully compatible with our method.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 explicitly list our three contributions: (i) identifying action-induced dimensional collapse in GRMs and pinpointing the collapse-prone $(A \to I)$ channel via heterogeneous interaction decomposition (Section 4.1, Theorem 4.1); (ii) the SVAN method that selectively applies value-side nonlinearity (Section 4); and (iii) consistent empirical gains in ranking performance, effective rank, and discriminability on three datasets (Section 5). The scope is restricted to attention-based generative ranking with interleaved item/action tokens.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix E

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theorems, lemmas, and assumptions are numbered and cross-referenced. Theorem 4.1 (collapsed channel) and Theorem 4.3 (selective nonlinearity) are stated with their assumptions (e.g., Assumption B.1 on low-dimensional downstream signal, Assumption B.2 on activation diversity). Full proofs are provided in Appendices B.1 and B.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The datasets (IJCAI, Retail, KuaiSAR) are public and described with statistics in Table 4. Section 5 and Appendices A and A.2 specify the architecture (number of Transformer blocks, attention heads, embedding dimensions), optimizer (Adam), learning rate, batch size, baseline hyperparameters, action/candidate masking, and evaluation protocol. Code is also publicly released at `https://github.com/xhran2010/Collapse-Hunter`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A public code repository is provided at `https://github.com/xhran2010/Collapse-Hunter`, including training/evaluation scripts and configuration files for all baselines and SVAN. All three datasets (IJCAI, Retail, KuaiSAR) are publicly available; preprocessing instructions are included in the repository.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 and Appendix A.2 specify the train/validation/test splits, optimizer (Adam), learning rate ($1e{-}3$), batch size (128), embedding dimension (50), number of Transformer blocks/heads, activation choices, and the use of action and candidate masking. Early stopping is performed using the validation AUC.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All main results are reported as mean $\pm$ standard deviation over 3 random seeds (Section 5); the variability captured is across random seeds (initialization and data shuffling). Standard deviations are computed using the standard sample formula.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix A.2 reports the compute setup used for all experiments (single-GPU training on a commodity NVIDIA GPU). Because SVAN introduces no additional learnable parameters and only one extra elementwise activation in a single channel, training cost matches the underlying GRM backbones (Appendix C).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics. The work uses publicly available recommendation datasets, does not collect new human-subject data, and does not raise ethical concerns beyond those generic to recommender systems.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix E

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases neither pretrained generative models with high misuse risk nor scraped datasets; it releases a method (an architectural modification to a value stream) and code for ranking experiments on existing public benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All baselines and datasets are credited with citations to the original papers (e.g., DIN, DCNv2, DIEN, DeepFM, SASRec, HSTU, GenRank, PinRec; IJCAI, Retail, KuaiSAR). The released datasets are used under their respective public terms; we use them only for non-commercial research as permitted.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: the paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects; it uses pre-existing public recommendation datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects, so no IRB approval is required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology of this paper; the proposed SVAN is an architectural modification to attention-based generative ranking models and does not rely on any LLM. Any use of LLMs was limited to writing-aid purposes that do not affect the methodology, scientific rigor, or originality of the work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.