# DISR: Distributional Item Representations for Sequential Recommendation

Abdelilah BAJJOU, El Habib NFAOUI

L3IA Laboratory-Faculty of Sciences Dhar El Mahraz,
Sidi Mohamed Ben Abdellah University, Fez, Morocco

***Abstract*—Sequential recommender systems predict the next item from a chronological interaction history, typically representing users and items as point embeddings scored by dot product. DISR (Distributional Item Representations for Sequential Recommendation) instead represents items and sequence queries as diagonal Gaussian distributions. Item means are initialized from frozen Sentence-BERT text representations, while query and item variances are learned during training. Candidates are scored using closed-form symmetric KL divergence within the standard sampled-softmax cross-entropy framework. Experiments on Amazon Reviews 2023 Beauty and Sports evaluate 50K-user and full-scale settings, with five seeds at 50K. DISR improves NDCG@10 over the strongest non-DISR baseline by $+12.8\%$ on Beauty 50K and $+15.6\%$ on Sports 50K. At full scale, DISR leads on Sports by $+2.6\%$ over DuoRec and remains within $0.8\%$ of the best NDCG@10 on Beauty. The largest gains occur for users with one to three interactions, reaching $+14.5\%$ on Beauty and $+19.3\%$ on Sports. DISR also converges substantially faster than the strongest contrastive baselines at full scale. Ablation results support the contribution of the overall distributional formulation, while the full-scale evaluation remains single-seed.**

***Keywords*—Sequential recommendation; distributional embedding; Kullback–Leibler divergence; cold-start recommendation; text-initialized embeddings; amazon reviews 2023**

## I. Introduction

Sequential recommender systems are widely deployed to predict a user's next interaction from their interaction history, with applications ranging from e-commerce [1, 2] to media recommendation and online services. The dominant family of methods — the SASRec lineage [1] and its successors [2–5] — has converged on a remarkably uniform recipe: a Transformer encoder over the interaction sequence, learned point embeddings for each item, and similarity scoring by inner product between the encoder's final hidden state and the candidate item's embedding. Variations within this lineage refine the architecture (bidirectional attention [2], frequency-domain mixing [3]), the training objective (contrastive augmentation [4, 5], intent clustering [6]), or the initialization (text-based item embeddings [7, 8]), but the underlying scoring contract — point query, point candidate, dot product — remains fixed.

This contract leaves two dimensions unexploited: per-item *semantic breadth* (a specific product versus a broad category has no separate representation) and *distributional spread* in the query (short and long histories are represented with identical confidence). The first limitation binds on heterogeneous catalogues where semantic granularity varies across items; the second is particularly relevant for cold-start users, where limited interaction history provides less behavioural evidence.

DISR is proposed as a sequential recommender that represents items as diagonal Gaussian distributions and produces a Gaussian query from the sequence encoder. The item's mean is initialized from a frozen Sentence-BERT [9] encoding of its natural-language description, projected to the working dimensionality, while the per-item variance is learned end-to-end. The query distribution is produced by two parallel heads on the encoder's last hidden state, emitting $(\mu_q, \log \sigma_q^2)$. Candidates are scored using the closed-form symmetric KL divergence between query and candidate distributions. The resulting score requires only element-wise arithmetic, with no matrix square roots or iterative procedures, and remains compatible with the standard sampled-softmax cross-entropy training used for dot-product scoring in SASRec and its descendants.

Distributional embeddings for words and knowledge graphs have a history in [10–13]. DISR combines a distributional Transformer query with text-initialized item means and learnable per-item variances under KL scoring within the standard sampled-softmax paradigm of the SASRec family. To the best of the authors' knowledge, this combination has not previously been reported in sequential recommendation.

DISR is evaluated on two categories of the Amazon Reviews 2023 benchmark [14] — Beauty and Sports — at two scales (50K-user subsamples and full $\sim$730K-user datasets), against six baselines (BERT4Rec [2], FMLP-Rec [3], SASRec [1], MoRec [7], CL4SRec [4], DuoRec [5]), with multi-seed evaluation at the 50K scale. The empirical picture is consistent across both datasets:

- At the 50K scale, DISR leads the strongest baseline by $+12.8\%$ relative NDCG@10 on Beauty and $+15.6\%$ on Sports, margins substantially larger than seed-induced variance.
- At full scale, DISR retains a clear lead on Sports ($+2.6\%$ over DuoRec) and ties the strongest contrastive baseline on Beauty NDCG@10 (within $-0.8\%$ of CL4SRec, while leading on HR@10).
- DISR's relative advantage is largest where data is sparsest. On cold-start users with 1–3 interactions, DISR outperforms the strongest baseline by $+14.5\%$ on Beauty and $+19.3\%$ on Sports, with the advantage shrinking but not disappearing in longer-history bins.

- DISR reaches its best validation NDCG@10 in 10–15 epochs on the full datasets, compared to 50–80 for the contrastive baselines and a full 200-epoch budget for SASRec without convergence — a 4–7× reduction in epoch count and a 2.4–5.7× wall-clock speedup against the closest contrastive competitors.

The cross-dataset replication, cold-start decomposition, and epoch-efficiency numbers together address the two most-cited limitations of new sequential recommenders: single-dataset scope [15] and training cost.

The remainder of this study is organized as follows. Section II situates DISR within the sequential, text-based, probabilistic, and distributional recommendation literature. Section III formalizes the model, presents the closed-form symmetric KL scoring function, and discusses computational considerations. Section IV describes the experimental protocol, datasets, and baselines. Section V reports the main empirical results at 50K and full scale, including convergence efficiency. Section VI presents the ablation study, cold-start analysis, variance interpretability, and cross-dataset replication. Section VII discusses the mechanism behind DISR's advantage, the honest reading of the Beauty full-scale tie, limitations, and future directions. Section VIII concludes.

## II. Related Work

DISR intersects four threads: sequential recommendation, text/modality-initialized recommendation, probabilistic recommendation, and distributional embeddings. Each is discussed briefly below.

### A. Sequential Recommendation

Sequential recommendation models the user as an ordered sequence of interactions and predicts the next item from that history. The transformer-based methods considered here derive from a common architectural ancestor: SASRec [1], a causal self-attention model trained with sampled-softmax cross-entropy and dot-product scoring against learned item embeddings. SASRec established the design template that subsequent methods refine. BERT4Rec [2] replaces the causal mask with bidirectional attention and the next-item objective with a masked-LM objective inspired by [16]. FMLP-Rec [3] demonstrates that self-attention can be replaced with frequency-domain filters at comparable accuracy and lower cost. $S^3$-Rec [17] introduces self-supervised pretraining objectives over item-attribute mutual information.

A more recent line introduces contrastive auxiliary objectives on top of the SASRec backbone. CL4SRec [4] adds an InfoNCE loss over augmented sequence views (random crop, mask, reorder). DuoRec [5] avoids explicit augmentation in favour of a model-level dropout-based contrastive signal, motivated by the representation-degeneration phenomenon [18]. ICL-Rec [6] couples sequential recommendation with latent intent clustering for additional contrastive supervision. These methods share an inductive bias — two views of similar sequences should embed close together — and consistently improve over the SASRec baseline across benchmarks.

All these methods preserve the SASRec scoring contract (point-point dot product). DISR keeps the causal Transformer backbone and sampled-softmax loss but changes item and query geometry to distributions and scoring to closed-form divergence.

### B. Text- and Modality-Initialized Recommendation

A parallel line replaces the standard learned-item-embedding table with representations derived from item content. MoRec [7] systematically compares ID-based and modality-based recommenders, showing that frozen text embeddings can match or exceed learned item embeddings at the same architectural budget. UniSRec [8] pretrains a universal sequence representation using parametric whitening over text-derived item embeddings, enabling cross-domain transfer. TIGER [19] pursues a different direction, framing recommendation as autoregressive generation of semantic IDs derived from quantized item embeddings. TALLRec [20] adapts a pretrained large language model directly to recommendation through instruction tuning. The Amazon Reviews 2023 benchmark used in this study [14] was explicitly designed to support these LLM-as-encoder evaluations.

DISR uses Sentence-BERT [9] (22M params, 384-dim, frozen) as the item-mean initializer, projected through a single learnable linear layer. The MoRec baseline uses the same encoder under the same protocol, isolating the KL-scoring contribution above text initialization (see Section VI-A).

### C. Probabilistic and Uncertainty-Aware Recommendation

The idea of representing recommender-system entities with probability distributions rather than point estimates has a substantial history, principally on the user side. Variational autoencoder approaches model the user as a posterior over latent factors [21]; bilateral variational extensions extend this to both users and items [22]. Gaussian process recommenders explicitly model predictive uncertainty over items [23]. These methods share the goal of representing uncertainty in the recommender's beliefs, but they typically apply the distributional treatment to a global user representation (or a global latent space) rather than to per-item geometry, and the scoring functions involved are reconstruction-based or kernel-based rather than divergence-based.

DISR differs on three axes: distributional treatment on items and queries (not a global user latent), dynamic query production from the sequence encoder per time step (not static parameters), and closed-form KL scoring compatible with sampled-softmax cross-entropy training.

### D. Distributional and Wasserstein Embeddings

Outside of recommendation, embedding entities as probability distributions rather than points has a richer literature, predominantly in word and graph representation. Vilnis and McCallum [10] represent words as diagonal Gaussians and score similarity by KL divergence; Athiwaratkun and Wilson [12] extend this to Gaussian mixtures to capture polysemy; He et al. [11] adapt the Gaussian-embedding framework to knowledge graphs, scoring triples by KL on diagonal Gaussian entity and relation representations.

A more recent line uses Wasserstein geometry for the same purpose. Muzellec and Cuturi [13] represent entities as

elliptical distributions — a family that includes Gaussians with full covariance, not just diagonal — and score similarity by the closed-form 2-Wasserstein distance, which on elliptical distributions decomposes into squared Euclidean distance between means plus the squared Bures metric between covariance matrices. They argue, against the KL-based formulations of [10], that Wasserstein is numerically better-behaved when variances become small, and that full-covariance distributions are more expressive than the diagonal restriction. Their evaluation focuses on word similarity, visualization, and entailment / hypernymy reconstruction on WordNet. A related strand of work approximates the Wasserstein distance itself via siamese-network embeddings for fast computation [24], though this addresses a different problem (scalable distance computation between histograms) rather than the entity-representation problem addressed here.

DISR uses diagonal Gaussians scored by KL, as in [10, 11], rather than the full-covariance Wasserstein/Bures formulation of [13]. Section III-D argues that KL avoids the numerical concerns Wasserstein was designed to address at the diagonal restriction while adding a Transformer-produced distributional query neither prior line considers.

### E. Cold-Start Recommendation

Cold-start recommendation has been addressed via meta-learning [25], dropout-based training [26], and content-based hybrid methods [27, 28]. DISR is not framed as a complete cold-start method for unseen users or items, but its text-initialized item means and distributional query with learned spread produce strong behaviour for users with short histories, as confirmed in Section VI-B. Generalisation to genuinely unseen items would additionally require a variance initialization because the learned per-item variance table is catalogue-specific.

## III. Methodology

This section develops DISR (Distributional Item Sequential Recommender) from the ground up. The standard sequential-recommendation setup [1, 2] is first established, followed by the three design choices that distinguish DISR from existing methods: items are represented as diagonal Gaussian distributions, the sequence encoder emits a distributional rather than point query, and candidates are scored by a closed-form negative symmetric KL divergence. The training objective remains the standard sampled-softmax cross-entropy used throughout the SASRec family, making DISR compatible with the same loss and negative-sampling framework while introducing a distributional representation and scoring extension rather than a separate training paradigm. Fig. 1 gives an overview; the rest of this section makes each component precise.

### A. Problem Formulation

Let $\mathcal{U}$ denote a set of users and $\mathcal{I} = \{1, \ldots, |\mathcal{I}|\}$ a set of items. For each user $u \in \mathcal{U}$, an interaction sequence $\mathcal{S}_u = (i_1, i_2, \ldots, i_{n_u})$ with $i_t \in \mathcal{I}$ is observed in chronological order. Given the first $L$ items of $\mathcal{S}_u$ as context (left-padded with a special PAD token if $n_u < L$, truncated to the most recent $L$ items if $n_u > L$), the task is to predict the next item the user will interact with. Following standard practice [1, 2], evaluation is framed as a ranking task: for each held-out target, the model scores the true item against a fixed pool of sampled negatives and is judged by Hit Rate (HR@$K$) and Normalized Discounted Cumulative Gain (NDCG@$K$).

The maximum sequence length is set to $L = 50$ throughout, matching the protocol of SASRec [1] and CL4SRec [4].

### B. Distributional Item Representations

The starting point of DISR is a deliberate departure from point embeddings. Standard sequential recommenders encode each item $i$ as a single vector $e_i \in \mathbb{R}^d$ and score a candidate $c$ against a query $q$ by inner product $\langle q, e_c \rangle$. Two pieces of information are discarded in this representation. The first is the *semantic breadth* carried by an item's natural-language description: "waterproof hiking boot, men's size 11, ankle support" and "hiking boot" refer to similar concepts but at very different levels of specificity, and a point embedding cannot reflect that difference. The second is the *distributional spread* associated with short user histories: a user who has interacted with two items provides far less signal than one who has interacted with thirty, but a point-embedding model has no mechanism to represent this difference in its query representation.

DISR addresses both by representing each item as a diagonal multivariate Gaussian distribution:

$$p_i = \mathcal{N}(\mu_i, \mathrm{diag}(\sigma_i^2)), \qquad \mu_i \in \mathbb{R}^d,\ \sigma_i^2 \in \mathbb{R}^d_{>0}. \tag{1}$$

The mean $\mu_i$ encodes the item's location in semantic space; the per-dimension variance $\sigma_i^2$ encodes how concentrated the item's meaning is along each axis.

*1) Item mean from frozen text:* The item mean $\mu_i$ is obtained by passing the item's natural-language description through the frozen Sentence-BERT encoder [9] — specifically the `all-MiniLM-L6-v2` checkpoint, which produces 384-dimensional sentence embeddings. The encoder is held *frozen* throughout training; no gradients flow into Sentence-BERT, which preserves the geometry of the pretrained text-embedding space as an inductive prior. A single learnable linear projection $W \in \mathbb{R}^{384 \times d}$ maps the Sentence-BERT output $\phi(t_i)$ to the model's working dimensionality $d = 64$:

$$\mu_i = W\,\phi(t_i). \tag{2}$$

This design has two practical consequences. Items that are textually similar (two hiking boots, two lipsticks) start out close in embedding space *before* any recommendation signal is observed, which provides a strong prior for users with little interaction history. The model also avoids a separate $|\mathcal{I}| \times d$ learned item-mean table, eliminating $O(|\mathcal{I}|d)$ parameters. This semantic initialization should not be confused with complete item cold-start capability: for a genuinely unseen item, its mean $\mu_i$ can be computed directly from its text description, but its learned per-item variance is not available from the catalogue-specific variance table. A deployment handling such unseen items would therefore require a variance initialization or fallback, such as $\sigma_i^2 = \mathbf{1}$. This new-item setting is not evaluated in the present study.

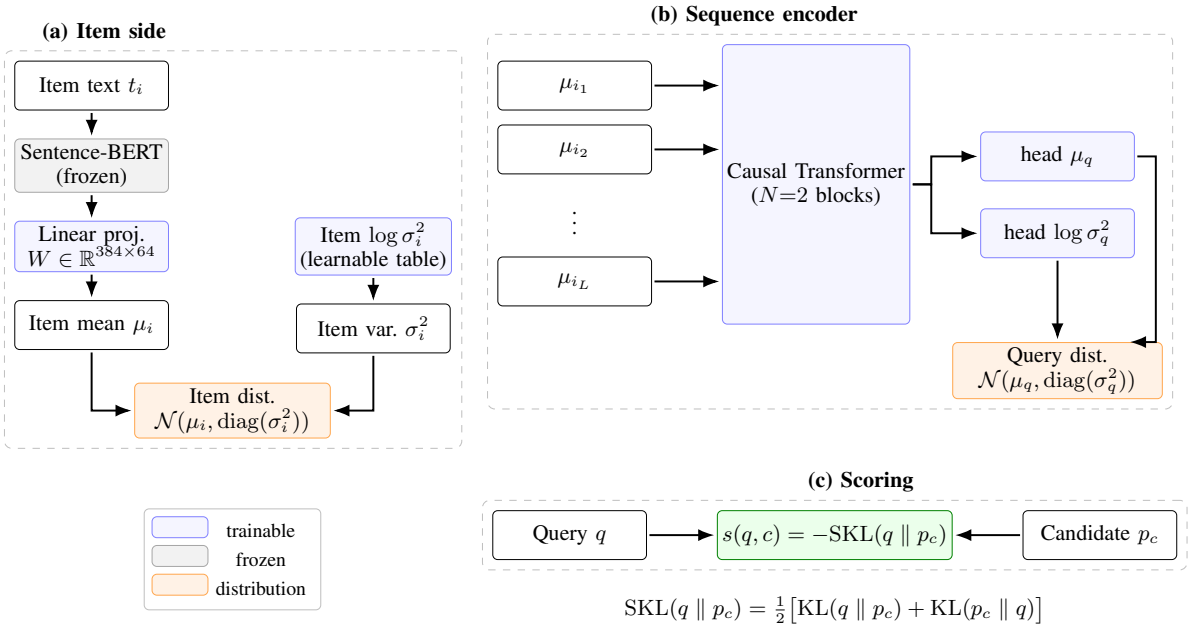


Fig. 1. The DISR architecture. Item descriptions are encoded by frozen Sentence-BERT and projected to item means, while a learnable table provides item variances. A causal transformer produces the query mean and variance, and candidates are scored using negative symmetric KL divergence.

*2) Item log-variance learned end-to-end:* The per-item log-variance is stored in a free learnable embedding table $E_{\log\sigma^2} \in \mathbb{R}^{|\mathcal{I}|\times d}$, initialized to zero so that $\sigma_i^2 = \mathbf{1}$ at the start of training and items begin as standard Gaussians. To prevent numerical issues at the extremes, $\log\sigma_i^2$ is clamped to the interval $[-5, 5]$, so $\sigma_i^2 \in [e^{-5}, e^5] \approx [0.0067, 148]$. This range is wide enough that the clamp is essentially inactive during normal training (as shown in Section VI-C, the learned variances cluster tightly around $\sigma^2 \approx 1$ in practice) but guarantees stability if any individual item drifts to an extreme value.

### C. Sequence Encoder and Distributional Query

The sequence encoder takes the input sequence of item means $\mu_{i_1}, \mu_{i_2}, \ldots, \mu_{i_L}$ and produces a query distribution. The backbone is a standard causal Transformer [1, 29]: $N = 2$ blocks, single-head self-attention with dropout 0.5, residual connections, LayerNorm, and position-wise feed-forward layers. Learned positional embeddings $P \in \mathbb{R}^{L\times d}$ are added at the input. The causal mask ensures that the representation at position $t$ depends only on items at positions $\le t$.

Let $h_L \in \mathbb{R}^d$ denote the hidden state at the last (rightmost) position of the encoder. Standard sequential recommenders treat $h_L$ as the query and score candidates by $\langle h_L, e_c\rangle$. DISR replaces this with two parallel linear heads that emit the parameters of a query *distribution*:

$$\mu_q = W_\mu\, h_L + b_\mu, \tag{3}$$

$$\log\sigma_q^2 = \mathrm{clamp}\big(W_\sigma\, h_L + b_\sigma,\ -5,\ 5\big), \tag{4}$$

Defining the query Gaussian $q = \mathcal{N}(\mu_q, \mathrm{diag}(\sigma_q^2))$. The $\mu$-head plays the role of a conventional point query; the $\log\sigma^2$-head is what makes DISR's query distributional. Conceptually, $\sigma_q^2$ represents the distributional spread assigned by the encoder at this position: a user with a long, coherent history may yield a narrower $\sigma_q^2$, while a user with a short or noisy history may yield a broader $\sigma_q^2$. The scoring function incorporates this learned spread directly.

### D. Symmetric Kullback–Leibler Scoring

With items and queries both represented as Gaussians, the natural similarity is a divergence between distributions. The negative symmetric Kullback–Leibler (KL) divergence is used, hereafter referred to as symmetric KL:

$$s(q,c) = -\mathrm{SKL}(q \,\|\, p_c) = -\tfrac{1}{2}\big[\mathrm{KL}(q \,\|\, p_c) + \mathrm{KL}(p_c \,\|\, q)\big]. \tag{5}$$

For two diagonal Gaussians the KL divergence admits the well-known closed form

$$\mathrm{KL}(q \,\|\, p_c) = \frac{1}{2}\sum_{j=1}^{d}\left[\frac{\sigma_{q,j}^2}{\sigma_{c,j}^2} + \frac{(\mu_{q,j} - \mu_{c,j})^2}{\sigma_{c,j}^2} - 1 - \log\frac{\sigma_{q,j}^2}{\sigma_{c,j}^2}\right], \tag{6}$$

and the reverse term $\mathrm{KL}(p_c \,\|\, q)$ is obtained by swapping the two distributions. The full symmetric divergence and its gradient are therefore available in closed form using only element-wise arithmetic — no matrix square roots, no iterative approximations, no numerical solvers.

Symmetrizing has two effects worth highlighting. First, it sends gradient signal through both the query and candidate variance terms, allowing both sides of the distributional score to be learned jointly. Second, the resulting divergence is symmetric and non-negative, and vanishes only when $q = p_c$

almost everywhere. Empirically, this symmetric formulation was slightly more stable than the asymmetric forward KL alone.

*1) Relation to dot-product scoring:* A useful sanity check: when all variances are pinned to one ($\sigma_q^2 = \sigma_c^2 = \mathbf{1}$), Eq. (6) reduces (up to a constant) to $\frac{1}{2}\|\mu_q - \mu_c\|_2^2$, and the symmetric KL becomes negative squared Euclidean distance. On $L_2$-normalized means this is monotonically equivalent to the dot product $\langle \mu_q, \mu_c \rangle$ used by SASRec and the rest of the SASRec family. DISR's scoring function thus strictly subsumes dot-product scoring as a degenerate case: when the model learns flat variances it recovers standard sequential-recommender behaviour, and any improvement over that behaviour must come from the model exercising the non-trivial variance machinery. Section VI-A confirms that this is what happens.

*2) On the choice of KL over Wasserstein:* The 2-Wasserstein distance with the Bures metric [13] is a natural alternative for elliptical distributions. Symmetric KL is used for three reasons: 1) diagonal Gaussians with $\log \sigma^2$ clamping avoid the numerical instabilities that motivate Wasserstein for full-covariance distributions; 2) Wasserstein/Bures requires matrix square roots via Newton–Schulz iterations [30], a real overhead at $K{=}100$ negatives per position, whereas KL requires only element-wise operations; 3) symmetric KL retains directional information through per-side variance terms, consistent with next-item prediction. A Wasserstein variant, particularly for full-covariance items, remains a direction for future work.

### E. Training Objective

DISR is trained with the standard sampled-softmax cross-entropy loss used by SASRec and its descendants. For each ground-truth next item $i^+$ in a training sequence, $K = 100$ negative items $i_1^-, \ldots, i_K^-$ are sampled uniformly from $\mathcal{I}$, excluding items already seen by the user. The per-position loss is:

$$\mathcal{L} = -\log \frac{\exp\big(s(q, p_{i^+})\big)}{\exp\big(s(q, p_{i^+})\big) + \sum_{k=1}^{K} \exp\big(s(q, p_{i_k^-})\big)}, \qquad (7)$$

averaged over non-padding positions in the batch. The scoring function $s(\cdot, \cdot)$ is the negative symmetric KL of Eq. (5); everything else is identical to the standard SASRec training loop.

No auxiliary KL prior or variance regularizer is applied. Such priors are common in variational-autoencoder-style recommenders [21, 22], where they prevent the posterior from collapsing to a point estimate. In DISR the scoring function itself acts as an implicit regularizer on variance: pulling $\sigma_q^2$ or $\sigma_c^2$ toward zero increases the KL term $\sigma_q^2/\sigma_c^2$ unboundedly when the other side is wider, while pushing them toward infinity penalises the $-\log \sigma^2/\sigma_c^2$ term. The model therefore has a natural incentive to keep variances in a useful range without any external prior, and this behaviour is observed empirically (Section VI-C).

### F. Computational Considerations

Symmetric KL between two diagonal Gaussians is $O(d)$ per (query, candidate) pair — the same order as dot product — so DISR training steps are within 3% of SASRec on an NVIDIA T4 GPU. Inference is identical in cost to SASRec's: a single Transformer pass, two linear head projections, and parallel closed-form scoring over $|\mathcal{C}|$ candidates in a sub-millisecond operation for typical retrieval settings. Parameter-wise, DISR adds two $d \times d$ heads and a per-item log-variance table of $|\mathcal{I}| \times d$ entries; the frozen-text mean construction removes the standard learnable item-embedding table of the same size, so the net trainable-parameter footprint is comparable to SASRec. Where DISR is dramatically more efficient is in *epoch count to convergence* (Section V-C).

## IV. Experimental Setup

DISR is evaluated against six baselines on two categories of the Amazon Reviews 2023 benchmark, at two scales, and with a multi-seed protocol where compute permits. The goal of this section is to make the evaluation precisely reproducible: which data was used, which negatives were sampled, which hyperparameters were trained, and which seeds were run.

### A. Datasets

The evaluation uses two categories from the Amazon Reviews 2023 benchmark [14]: *Beauty and Personal Care* and *Sports and Outdoors*. Both contain user-item interactions, item metadata (title, features, description), and ratings collected through September 2023. The evaluation uses the 5-core split released with the dataset, in which users and items with fewer than five interactions are iteratively removed [1, 2]. Sequential histories are built by sorting each user's interactions chronologically; users with only the canonical minimum of five interactions thus contribute a sequence of length five (three training items, one validation, one test).

The two categories are deliberately chosen to differ along axes that matter for the research questions considered. Beauty items tend to be linguistically rich (long product titles, dense ingredient and benefit descriptions, narrow vocabulary clusters around cosmetic categories), and the catalogue is dominated by short user histories. Sports items are linguistically more heterogeneous (a fishing rod, a yoga mat, and a treadmill are all "sports" but share little vocabulary), and average histories are slightly longer. A method that works on both should be capturing something general about sequential recommendation rather than exploiting per-category text patterns.

To assess both behaviour under tight compute and behaviour at production scale, the evaluation reports results for each category at two scales:

- 50K subsample: A random sample of 50,000 users from the category, retaining all items they interacted with. This yields 128,864 items in Beauty and 115,666 items in Sports. This is the setting used for the full multi-seed protocol, per-history-length cold-start analyses, and ablations.
- Full dataset: All filtered users in the category — 729,576 users and 207,649 items for Beauty, 409,772 users and 156,235 items for Sports. This is the

production-scale setting used to confirm that conclusions drawn at 50K hold at scale.

Following standard practice [1, 31], *leave-one-out* evaluation is used: each user's chronologically last interaction is held out as the test target, the second-to-last as the validation target, and the remaining interactions form the training history. Sequences are left-padded or truncated to a maximum length of $L = 50$.

Table I summarises the resulting splits.

TABLE I. DATASET STATISTICS AFTER 5-CORE FILTERING AND LEAVE-ONE-OUT SPLITTING. SEQUENCE-LENGTH STATISTICS ARE COMPUTED OVER TRAINING HISTORIES

| | **Beauty** | | **Sports** | |
|---|---|---|---|---|
| | 50K subsample | Full | 50K subsample | Full |
| Users | 50,000 | 729,576 | 50,000 | 409,772 |
| Items | 128,864 | 207,649 | 115,666 | 156,235 |
| Interactions | ∼360,000 | ∼5.2M | ∼326,000 | ∼2.65M |
| Avg. history length | 7.2 | 7.1 | 6.5 | 6.5 |
| Median history length | 5 | 5 | 4 | 4 |

### B. Evaluation Protocol

For each held-out user-item pair, the ground-truth item is scored against 99 negative items sampled uniformly from the items the user has not previously interacted with. This 100-candidate ranking task is the standard protocol used in [1, 2, 4] and is followed here for comparability. The same 99 negatives per user are used for every method (drawn once at preprocessing time and cached) to remove sampling noise from cross-method comparisons.

Hit Rate (HR@$K$) and Normalized Discounted Cumulative Gain (NDCG@$K$) are reported at $K \in \{5, 10, 20\}$. Ties on the ground-truth rank are broken by the average-position convention rather than by best-rank: when $n_{\text{ties}}$ items share a score with the ground-truth target, the ground-truth's effective rank is $r_{\text{strict}} + 0.5 \cdot n_{\text{ties}}$ where $r_{\text{strict}}$ is the number of items strictly outranking it. This convention prevents metric inflation that can occur with best-rank tie-breaking, particularly for methods that produce many identical scores [15, 32].

NDCG@10 is the primary metric throughout. It is used both for model selection (best validation epoch) and for reporting (test set). All other metrics are reported at the same checkpoint.

The absolute NDCG@10 values reported here are not directly comparable to results in the literature that use full-catalogue ranking instead of sampled negatives [15]: the sampled-negative protocol gives higher absolute numbers but is widely adopted as a comparability baseline in this family of work. Within the present evaluation, all methods use identical negative samples, so relative comparisons are unaffected.

### C. Baselines

DISR is compared against six baselines spanning the major design choices in modern sequential recommendation:

- SASRec [1]: a causal self-attention Transformer with learned item embeddings, trained with sampled-softmax cross-entropy. This is the dominant sequential-recommender baseline and the architectural ancestor of most methods considered here, including DISR.
- BERT4Rec [2]: a bidirectional masked Transformer trained with cloze-style item prediction. Different inductive bias (bidirectional vs. causal attention) but the same point-embedding scoring contract.
- FMLP-Rec [3]: replaces self-attention with learnable filters in the frequency domain via FFT, demonstrating that attention is not strictly necessary at this scale.
- CL4SRec [4]: a SASRec backbone augmented with an auxiliary InfoNCE contrastive loss over augmented sequence views (crop, mask, reorder). Tests the value of contrastive auxiliary objectives.
- DuoRec [5]: a SASRec backbone with model-level (dropout-based) contrastive regularization. Tests an alternative form of contrastive signal that does not require explicit sequence augmentation.
- MoRec [7]: identical Transformer backbone to SASRec, but with item embeddings replaced by a learnable linear projection of frozen Sentence-BERT text vectors. This baseline isolates the contribution of text initialization absent DISR's distributional/KL design, and is the most direct point of comparison for the question "does the gain come from text or from distributions?"

In keeping with the SASRec-family conventions, all sequential baselines share the same data loading pipeline, the same evaluation harness, the same sampled-softmax cross-entropy training objective, and the same negative-sampling regime (100 negatives per training position, 99 per test instance). Differences in reported performance therefore reflect architectural and loss-function choices rather than implementation details.

### D. Implementation and Hyperparameters

All methods are implemented in PyTorch and trained on a single NVIDIA Tesla T4 GPU. The Adam optimizer is used with $\beta_1 = 0.9$, $\beta_2 = 0.98$, weight decay $10^{-6}$, learning rate $10^{-3}$, batch size 128 at the 50K scale and 256 at full scale. Training proceeds for up to 200 epochs with early stopping on validation NDCG@10 (patience 40 at the 50K scale, 20 at the full scale). The hidden dimension is $d = 64$ across all sequential methods; sequence length is $L = 50$.

For DISR specifically, the per-item log-variance is initialized to zero (so $\sigma_i^2 = 1$ at initialization) and clamped to the interval $[-5, 5]$ throughout training. No auxiliary KL prior or variance regularizer is applied. The frozen text encoder is `sentence-transformers/all-MiniLM-L6-v2` [9], producing 384-dimensional sentence embeddings, with a maximum input length of 128 sub-word tokens.

*1) Multi-seed protocol:* For statistical robustness, each sequential method is run on the 50K subsample of both datasets with five random seeds $\{42, 7, 123, 999, 2024\}$. Mean and standard deviation are reported over the five runs for the primary HR@5, HR@10, and NDCG@10 metrics. The same five seeds are used for the per-history-length cold-start analysis in Section VI-B.

At the full dataset scale, each method is run once with seed 42 due to compute cost. Full-scale runs span wall-clock times on a T4 from under an hour for the fastest methods to several hours for the contrastive methods that converge most slowly, making a full multi-seed sweep computationally prohibitive. The five-seed results at 50K provide an empirical measure of training variability, but they do not directly quantify variance at the full-dataset scale. The full-scale results are therefore used to assess performance at production scale rather than to make claims about seed-level statistical significance. In general, increasing the number of users can reduce variability in aggregate evaluation, but differences in optimization trajectories and model convergence can still produce seed-dependent variation; the magnitude of that variation is not measured here.

*2) Ablation protocol:* DISR's three design ablation variants (no_var, no_text, DOT) are trained at the 50K scale with seed 42 on both datasets. In addition, a targeted encoder-sensitivity ablation replaces MiniLM with MPNet on Sports 50K using seed 42. The ablation results are therefore single-seed; the multi-seed 50K results for the full DISR model provide the reference for interpreting the magnitude of single-seed ablation changes.

*3) Reproducibility:* All experimental code, processing scripts, and trained checkpoints will be released on acceptance. Negative-sample seeds are fixed and cached; data splits are deterministic given the public Amazon Reviews 2023 release.

## V. Main Results

This section reports the empirical performance of DISR against the six baselines described in Section IV-C. The evidence is organized by scale: Section V-A presents the multi-seed comparison on the 50K subsample of both Amazon Beauty and Amazon Sports; Section V-B confirms that the orderings established at 50K hold at full-dataset scale; and Section V-C quantifies DISR's substantial training-efficiency advantage, which together with its closed-form scoring makes the method attractive for latency-bounded production settings.

### A. 50K-Scale Multi-Seed Results

Table II reports HR@5, HR@10, and NDCG@10 averaged over five random seeds $\{7, 42, 123, 999, 2024\}$ for all seven methods on the 50K-user subsample of both datasets. Fig. 2 provides the corresponding visual comparison on the NDCG@10 metric with seed-induced error bars.

The pattern is consistent across both categories and across all three primary metrics: DISR is the best-performing method, by a margin much larger than its per-seed standard deviation. On Amazon Beauty, the improvement over the strongest baseline (SASRec) is $+12.8\%$ relative NDCG@10 ($0.3539$ vs. $0.3139$); on Amazon Sports the improvement over the strongest baseline (MoRec) is $+15.6\%$ relative NDCG@10 ($0.3281$ vs. $0.2838$).

The five-seed results provide evidence of training stability, with the observed DISR-vs-baseline differences substantially larger than the variation across random seeds. Paired $t$-tests computed over the five seed-level differences yield $t \in [18.4, 74.4]$ and $p < 10^{-4}$ for all six Beauty NDCG@10 comparisons. These tests are reported as descriptive evidence of separation under stochastic training rather than as population-level significance tests, since the five seeds are repeated runs on the same users and dataset rather than independent samples. Cohen's $d$ effect sizes range from $8.2$ (vs. SASRec) to $33.3$ (vs. MoRec), further indicating that the observed differences are large relative to seed-induced variability.

To complement the seed-level analysis with user-level inference, paired bootstrap confidence intervals were computed on the 50K test users for DISR against SASRec and CL4SRec using the seed-7 checkpoints. For each user, the difference in NDCG@10 was computed using the same test rows and cached negative set for both methods, and users were resampled with replacement for 10,000 bootstrap replicates. The resulting 95% percentile confidence intervals were entirely above zero for both comparisons: $+0.045055$ [+0.042711, +0.047366] for DISR versus SASRec and $+0.045002$ [+0.042645, +0.047305] for DISR versus CL4SRec. These results provide complementary user-level evidence that the observed average performance advantage is robust across the evaluated test population, while the seed-level analysis continues to characterize training stability.

*1) Text initialization is category-dependent:* MoRec is the worst Transformer baseline on Beauty (NDCG@10 = 0.2419 vs. SASRec 0.3139) but the second-best on Sports (0.2838, ahead of SASRec's 0.2774): the homogeneous Beauty catalogue gives learned embeddings enough signal, while Sports' heterogeneous vocabulary rewards text priors. DISR outperforms MoRec by $+46.3\%$ (Beauty) and $+15.6\%$ (Sports) NDCG@10, showing that the distributional/KL design contributes information beyond text initialization alone (see Section VI-A).

*2) Contrastive methods trail DISR:* CL4SRec and DuoRec are strong on Beauty (0.3127 and 0.3069, within SASRec's seed variance) but trail SASRec and MoRec on Sports. No single auxiliary signal — contrastive or text — dominates across both categories; DISR's advantage does.

### B. Full-Scale Results

Table III reports full-scale results on Beauty (729,576 users, 207,649 items) and Sports (409,772 users, 156,235 items), seed 42. The 50K per-seed variances (Table II) provide the noise reference.

On Amazon Sports, DISR achieves the best NDCG@10 (0.4418), a $+2.6\%$ improvement over DuoRec and $+2.8\%$ over CL4SRec, with wider margins over the remaining baselines. DISR also leads HR@5 and HR@10. The Sports result is unambiguous (Fig. 3).

On Amazon Beauty, the top three methods (CL4SRec, DuoRec, DISR) cluster tightly: CL4SRec edges DISR on NDCG@10 ($-0.8\%$) and HR@5 (within noise), while DISR leads clearly on HR@10 ($+1.1\%$). The most consistent explanation for this convergence is that DISR's inductive priors

TABLE II. MULTI-SEED RESULTS ON THE 50K-USER SUBSAMPLES OF AMAZON BEAUTY AND SPORTS

| | Amazon Beauty | | | Amazon Sports | | |
|---|---|---|---|---|---|---|
| **Method** | HR@5 | HR@10 | NDCG@10 | HR@5 | HR@10 | NDCG@10 |
| BERT4Rec | 0.3502 ± 0.0025 | 0.4550 ± 0.0027 | 0.2904 ± 0.0024 | 0.2951 ± 0.0028 | 0.3902 ± 0.0027 | 0.2455 ± 0.0023 |
| FMLP-Rec | 0.3154 ± 0.0053 | 0.4117 ± 0.0070 | 0.2605 ± 0.0047 | 0.2357 ± 0.0030 | 0.3224 ± 0.0027 | 0.1955 ± 0.0022 |
| MoRec | 0.2913 ± 0.0009 | 0.4028 ± 0.0012 | 0.2419 ± 0.0009 | 0.3455 ± 0.0013 | 0.4769 ± 0.0014 | 0.2838 ± 0.0004 |
| SASRec | 0.3768 ± 0.0025 | 0.4785 ± 0.0033 | 0.3139 ± 0.0023 | 0.3325 ± 0.0025 | 0.4261 ± 0.0024 | 0.2774 ± 0.0020 |
| CL4SRec | 0.3752 ± 0.0028 | 0.4761 ± 0.0033 | 0.3127 ± 0.0026 | 0.3273 ± 0.0019 | 0.4210 ± 0.0034 | 0.2722 ± 0.0018 |
| DuoRec | 0.3692 ± 0.0048 | 0.4691 ± 0.0037 | 0.3069 ± 0.0039 | 0.3200 ± 0.0054 | 0.4136 ± 0.0064 | 0.2663 ± 0.0046 |
| **DISR** | **0.4230** ± 0.0032 | **0.5292** ± 0.0037 | **0.3539** ± 0.0033 | **0.3936** ± 0.0016 | **0.4992** ± 0.0013 | **0.3281** ± 0.0014 |
| Improvement | +12.3% | +10.6% | +12.8% | +13.9% | +4.7% | +15.6% |

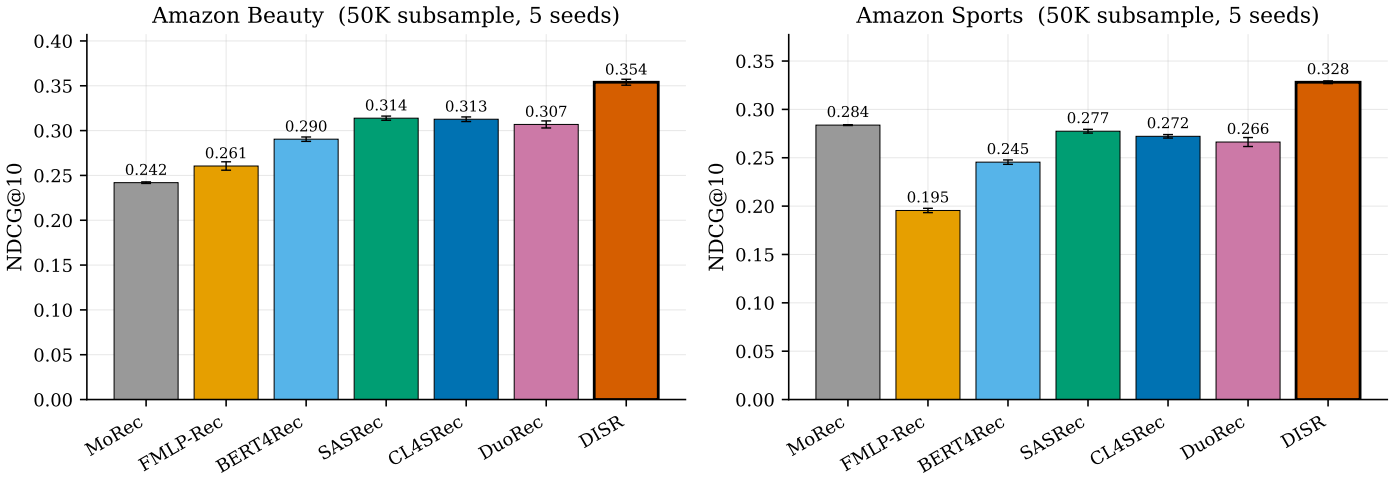


Fig. 2. NDCG@10 on the 50K-user subsamples of Amazon Beauty and Sports. Error bars show one standard deviation over five random seeds.

TABLE III. FULL-SCALE RESULTS ON AMAZON BEAUTY AND SPORTS, SEED 42

| | Amazon Beauty (full) | | | Amazon Sports (full) | | |
|---|---|---|---|---|---|---|
| **Method** | HR@5 | HR@10 | NDCG@10 | HR@5 | HR@10 | NDCG@10 |
| BERT4Rec | 0.5350 | 0.6383 | 0.4515 | 0.4496 | 0.5511 | 0.3781 |
| FMLP-Rec | 0.5201 | 0.6249 | 0.4378 | 0.4430 | 0.5428 | 0.3721 |
| MoRec | 0.3169 | 0.4301 | 0.2646 | 0.3569 | 0.4910 | 0.2926 |
| SASRec | 0.5457 | 0.6524 | 0.4600 | 0.4761 | 0.5801 | 0.3984 |
| CL4SRec | **0.5728** | 0.6746 | **0.4827** | 0.5108 | 0.6116 | 0.4298 |
| DuoRec | 0.5681 | 0.6706 | 0.4793 | 0.5128 | 0.6134 | 0.4308 |
| **DISR** | 0.5698 | **0.6823** | 0.4789 | **0.5270** | **0.6516** | **0.4418** |

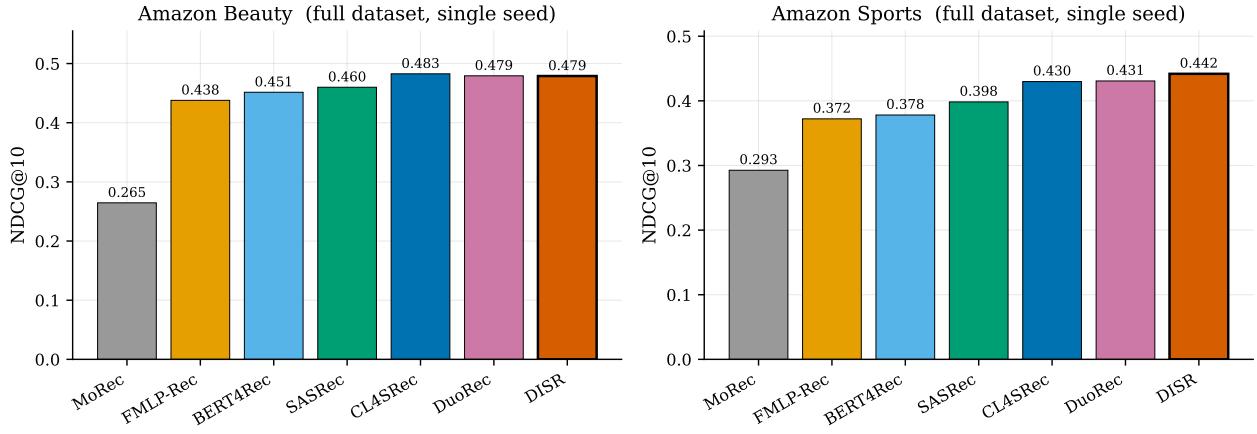


Fig. 3. NDCG@10 at full-dataset scale on Amazon beauty and sports.

(text-conditioned means, distributional queries) provide diminishing returns at full scale, as contrastive baselines learn comparable representations from roughly 14× more interactions — confirmed by the per-history-length analysis in Section VI-B.

DISR ranks among the top three methods on every metric, with a distinct efficiency profile (Section V-C).

### C. Convergence Efficiency

Where DISR's advantage is most pronounced is in the wall-clock cost of reaching its best validation performance. Table IV reports the best-validation epoch and the wall-clock training time to reach that epoch on each full-dataset run, on a single NVIDIA T4 GPU. Fig. 4 visualizes how performance scales from the 50K subsample to the full dataset for each method.

The numbers tell a consistent story. On Sports, DISR reaches its best validation NDCG@10 in 181.2 minutes against DuoRec's 437.0 minutes ($2.4\times$ speedup) and CL4SRec's 741.4 minutes ($4.1\times$ speedup), at higher accuracy than either. On Beauty, where the three top methods are within $0.004$ NDCG@10 of one another, DISR reaches its peak in 208.9 minutes against CL4SRec's 1191.9 minutes ($5.7\times$ speedup) and DuoRec's 609.6 minutes ($2.9\times$ speedup) — so even where the accuracy story is a three-way tie, the efficiency story is decisive.

*1) Epoch counts confirm the pattern:* Factoring out per-step implementation differences, DISR converges in 10 epochs on Beauty and 15 on Sports, versus 50–80 for contrastive methods and 200 (unconverged) for SASRec — a 4–7$\times$ reduction. Text-conditioned item priors provide a strong starting geometry, and distributional scoring shapes per-item variance directly, giving the optimizer a more informative initialization.

*2) Implications for deployment:* DISR reaches peak validation in a fraction of the contrastive baselines' wall-clock time — an operational advantage for systems retrained frequently — while inference is a single Transformer pass plus closed-form KL scoring, free of iterative procedures. This combination is viable for latency-bounded production settings.

## VI. Analysis

The headline accuracy numbers in Section V establish that DISR is competitive or best across both datasets and both scales. This section dissects *where* that performance comes from. Section VI-A isolates the contribution of each architectural ingredient through three design ablations and a targeted encoder-sensitivity check; Section VI-B examines performance as a function of user-history length, where DISR's advantage is most pronounced; Section VI-C inspects the learned per-item variance to assess whether it exhibits meaningful semantic structure; Section VI-D summarises the cross-dataset pattern that distinguishes structural improvements from dataset-specific artifacts.

### A. Ablation Study

To isolate the contribution of each design choice, three ablation variants are run on the 50K subsample of both datasets (seed 42):

- DISR-no_var: per-item variances are pinned to $\sigma_i^2 = \mathbf{1}$ throughout training, and the query variance head is removed (queries are point estimates). All other components are identical to DISR-full. This isolates the contribution of the variance machinery.
- DISR-no_text: item means are learned from scratch as a free embedding table $E_\mu \in \mathbb{R}^{|\mathcal{I}|\times d}$ rather than projected from frozen Sentence-BERT vectors. This isolates the contribution of text initialization.
- DISR-DOT: replaces symmetric KL scoring with standard dot product $\langle \mu_q, \mu_c \rangle$, while retaining the text-initialized item means and the Transformer backbone from DISR-full. Because the variance components are not used by dot-product scoring, the query-side variance head is removed and the per-item log-variance table is frozen at $\log \sigma_i^2 = 0$. This variant therefore assesses the effect of replacing the distributional scoring formulation with conventional point-based scoring, rather than isolating the symmetric KL operation alone.

Table V and Fig. 5 report NDCG@10 for each variant. SASRec is included as the architectural reference point: it represents what remains when all three DISR modifications are removed.

*1) Distributional scoring contributes substantially to the observed gains:* The DISR-DOT variant replaces the distributional symmetric KL scoring formulation with a standard inner product while retaining the text-initialized item means and the Transformer backbone. On Beauty, this reduces NDCG@10 from 0.3545 to 0.2391, a $-32.5\%$ relative drop and below the SASRec baseline of 0.3128. On Sports, the drop is smaller ($-13.8\%$ relative), but DISR-DOT still trails DISR-full clearly. Because the DOT variant also removes the query-side variance head and freezes the per-item variance parameters, this comparison should be interpreted as evidence for the contribution of the overall distributional-scoring formulation rather than as an isolated causal test of the symmetric KL operation. The results nevertheless show that replacing the distributional scoring formulation with conventional point-based scoring substantially reduces performance.

*2) Text initialization is more useful on Sports:* The DISR-no_text variant learns item means from scratch instead of projecting them from frozen Sentence-BERT vectors. On Beauty the cost is moderate ($-16.9\%$ relative); on Sports it is much larger ($-27.7\%$ relative). This mirrors the cross-method pattern in Table II: MoRec is comparatively weak on Beauty (where the learned embedding table can recover the relevant geometry) but strong on Sports (where item text provides more discriminative information than co-occurrence alone). DISR exploits this in both directions: when text helps, it helps DISR; when text is less informative, the KL scoring carries the model.

*3) The per-item variance head is the smallest contributor:* Removing per-item variances (DISR-no_var) costs only $-9.5\%$ on Beauty and $-1.1\%$ on Sports, consistent with Section VI-C: learned variances cluster tightly around $\sigma^2 \approx 1$. The contribution is non-zero but concentrated in the cold-start regime.

*4) Encoder sensitivity:* To assess whether the observed performance depends specifically on the chosen text encoder, the frozen `all-MiniLM-L6-v2` initializer is replaced with the larger `all-mpnet-base-v2` encoder while keeping the DISR architecture, training protocol, 50K Sports split, and seed 42 unchanged. NDCG@10 increases from $0.3266$ with

TABLE IV. CONVERGENCE EFFICIENCY ON THE FULL DATASETS USING A SINGLE NVIDIA T4 GPU

| | Beauty (full) | | Sports (full) | |
|---|---|---|---|---|
| **Method** | Best epoch | Wall time (min) | Best epoch | Wall time (min) |
| BERT4Rec | 25 | 131.3 | 20 | 57.2 |
| FMLP-Rec | 45 | 505.2 | 45 | 267.4 |
| MoRec | 190 | 2342.5 | 120 | 804.1 |
| SASRec | 200 | 2257.3 | 200 | 1217.1 |
| CL4SRec | 70 | 1191.9 | 80 | 741.4 |
| DuoRec | 50 | 609.6 | 65 | 437.0 |
| **DISR** | **10** | **208.9** | **15** | **181.2** |

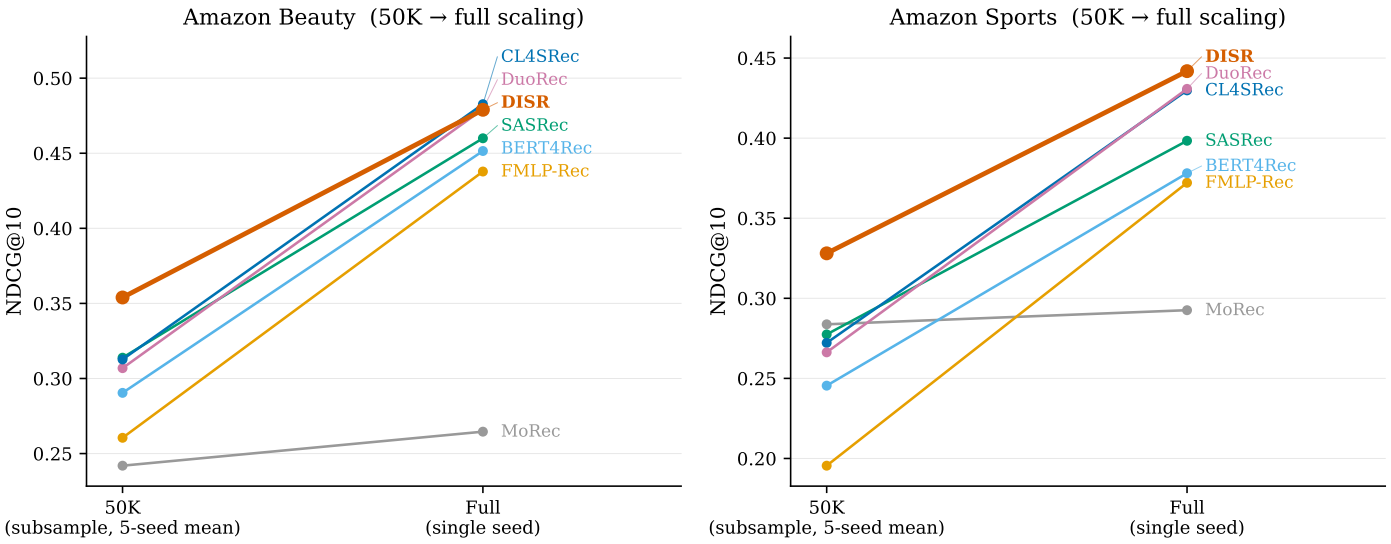


Fig. 4. Per-method scaling from the 50K-user subsample to the full-dataset evaluation on Beauty (left) and Sports (right).

TABLE V. ABLATION RESULTS ON THE 50K SUBSAMPLE OF BEAUTY AND SPORTS (SEED 42)

| | Beauty (50K) | | | Sports (50K) | | |
|---|---|---|---|---|---|---|
| **Variant** | HR@5 | HR@10 | NDCG@10 | HR@5 | HR@10 | NDCG@10 |
| SASRec (baseline) | 0.3761 | 0.4757 | 0.3128 | 0.3301 | 0.4241 | 0.2757 |
| DISR-DOT | 0.2877 | 0.3998 | 0.2391 | 0.3420 | 0.4762 | 0.2815 |
| DISR-no_text | 0.3581 | 0.4658 | 0.2947 | 0.2862 | 0.3874 | 0.2362 |
| DISR-no_var | 0.3798 | 0.4923 | 0.3207 | 0.3895 | 0.5186 | 0.3230 |
| **DISR-full** | **0.4234** | **0.5296** | **0.3545** | **0.3911** | **0.4986** | **0.3266** |

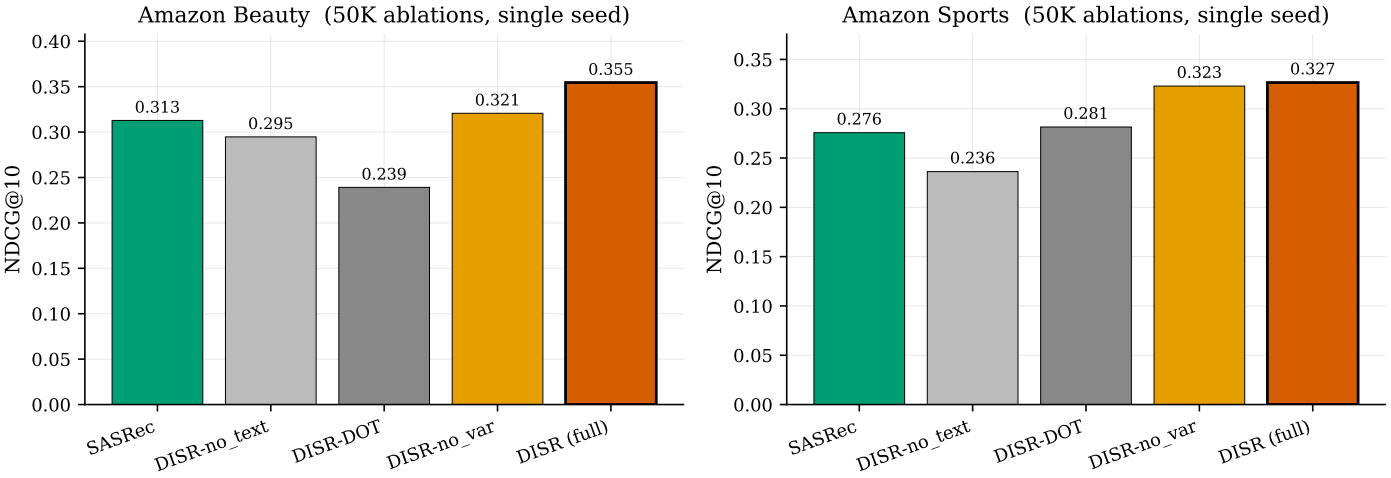


Fig. 5. Ablation results on the 50K subsample of Beauty (left) and Sports (right).

MiniLM to $0.3329$ with MPNet, a modest $+1.9\%$ relative change; HR@5 increases from $0.3911$ to $0.3979$ and HR@10 from $0.4986$ to $0.5076$. The close performance of the two encoders suggests that the observed DISR behaviour is not tied to a particular Sentence-BERT checkpoint, although this single-seed ablation on one dataset does not establish encoder invariance across domains or random seeds.

### B. Cold-Start Analysis

To assess DISR's behaviour as a function of history length, the 50K test users are partitioned into four bins (1–3, 4–5, 6–10, and 11+ interactions), and NDCG@10 is reported per bin, averaged over five seeds. The four bins contain 15,409, 15,601, 12,631, and 6,359 users, respectively, for a total of 50,000 test users. Thus, the cold-start comparisons cover the complete 50K test set. Table VI and Fig. 6 report the results.

The cold-start pattern is the cleanest piece of evidence in this study, as shown in Fig. 6. On Beauty, DISR's relative improvement over the best non-DISR baseline ranges from $+14.5\%$ in the 1–3 bin to $+11.8\%$ in the 11+ bin — consistently above $+10\%$ across all bins, and largest where data is sparsest. On Sports the pattern is more dramatic: the relative improvement declines from $+19.3\%$ in the 1–3 bin to $+3.6\%$ in the 11+ bin. This provides direct evidence that DISR's gains are strongest when behavioural data is scarce, consistent with the role of text-conditioned item means and variance-aware queries.

Notably, MoRec's NDCG@10 rises with history length on Sports ($0.2795 \rightarrow 0.2948$) while every other method's falls — likely because text-initialized embeddings carry less per-item signal than learned ones, so MoRec gains disproportionately as more training data becomes available. DISR maintains its lead over MoRec across all bins.

### C. Variance Interpretability

A natural question about a variance-aware sequential recommender is whether the learned per-item variances carry interpretable structure or whether they collapse to a single value. The trained Beauty 50K DISR model (seed 42) is inspected to address this question.

The learned per-item $\log \sigma^2$ distribution across the 128,864 Beauty items is tightly concentrated below zero: mean $\overline{\log \sigma^2} = -0.119$ with standard deviation $0.034$ and a 10th–90th percentile range of $[-0.160, -0.082]$. The clamp at $[-5, 5]$ is essentially inactive — no item approaches the boundaries, and the model uses the variance head conservatively rather than aggressively. The headline implication is that DISR's items are nearly all slightly *narrower* than the standard Gaussian initialization, with $\sigma^2 \in [0.85, 0.92]$ for the vast majority of the catalogue.

*1) Variance correlates with item frequency:* The interesting structure appears when the data are stratified by training-set item frequency. Table VII reports the mean learned $\log \sigma^2$ within frequency bins. The Pearson correlation between $\log(\text{frequency} + 1)$ and $\log \sigma^2$ is $+0.54$, and the monotonic relationship is clearly visible: rare items receive the tightest variance, and the most frequent items (51+ interactions) are the only group with $\log \sigma^2 > 0$, indicating $\sigma^2 > 1$.

This pattern is interpretable: frequent items appear in many distinct user histories with diverse surrounding contexts, and the model learns a *broader* Gaussian shape that can accommodate the diverse roles such items play. Rare items appear in fewer contexts, and the model tends to learn a *tighter* Gaussian associated with their narrower observed semantic footprint. The variance head therefore learns item-specific distributional spread rather than a single global value, with the observed patterns varying according to how much (and how diversely) each item is represented during training. These patterns are consistent with a role for learned dispersion in representing item heterogeneity, but they do not by themselves establish calibrated predictive uncertainty.

*2) Item-type signals are secondary:* Item-type signals from titles (bundles, multi-packs, size/shade markers) correlate with variance only weakly ($|\Delta \log \sigma^2| < 0.015$), well below the frequency-based pattern. The most parsimonious interpretation is that the variance head primarily captures frequency-associated distributional spread rather than specific semantic properties.

### D. Cross-Dataset Replication

Fig. 7 summarises DISR's relative NDCG@10 lift across six evaluated conditions. The pattern is consistent: DISR leads on both datasets at 50K, wins full-scale Sports ($+2.6\%$ over DuoRec), and ties full-scale Beauty (within $-0.8\%$ of CL4SRec, with DISR leading HR@10). The ablation ordering KL $>$ text $>$ variance holds on both datasets (Table V), with magnitudes that track each category's data sparsity and vocabulary heterogeneity: text contributes more on Sports (discriminative vocabulary), variance more on Beauty (sparser per-item signal). Cold-start advantage is largest in the shortest-history bin and declines with history length on both, more steeply on Sports.

## VII. Discussion

The empirical evidence in Sections V and VI establishes that DISR is competitive or best on overall accuracy across two datasets and two scales, with a substantial training-efficiency advantage and a clear lead in the cold-start regime. This section takes a step back to address the patterns the evidence does not flatten: where DISR's advantage is most pronounced, where it narrows, what aspects of the experimental protocol are adopted with which trade-offs, and what limitations the design carries.

### A. Where DISR Helps the Most, and Why

A pattern recurs: DISR's advantage is largest where data is sparsest. The 50K-scale gains are larger than full-scale gains, cold-start advantages are largest in the shortest-history bin, and the variance head contributes most on Beauty (the sparser per-item catalogue). All three patterns point to the same underlying mechanism.

DISR brings two inductive priors to sequential recommendation that the standard baselines lack: a frozen-text geometry that constrains the item-mean space before any recommendation signal is observed, and a distributional query whose variance modulates how broadly each candidate is considered. Both are most useful when the model has limited data with which to learn the relevant geometry from scratch. As

TABLE VI. NDCG@10 BY USER-HISTORY-LENGTH BIN ON THE 50K SUBSAMPLE, AVERAGED OVER FIVE SEEDS

| Method | 1–3 | 4–5 | 6–10 | 11+ |
|---|---|---|---|---|
| *Amazon Beauty (50K)* | | | | |
| BERT4Rec | $0.2922 \pm 0.0044$ | $0.2960 \pm 0.0024$ | $0.2898 \pm 0.0016$ | $0.2737 \pm 0.0029$ |
| FMLP-Rec | $0.2632 \pm 0.0043$ | $0.2671 \pm 0.0050$ | $0.2615 \pm 0.0051$ | $0.2358 \pm 0.0046$ |
| MoRec | $0.2450 \pm 0.0009$ | $0.2419 \pm 0.0011$ | $0.2398 \pm 0.0014$ | $0.2389 \pm 0.0008$ |
| SASRec | $0.3109 \pm 0.0035$ | $0.3209 \pm 0.0020$ | $0.3164 \pm 0.0027$ | $0.2986 \pm 0.0035$ |
| CL4SRec | $0.3087 \pm 0.0030$ | $0.3194 \pm 0.0027$ | $0.3167 \pm 0.0037$ | $0.2978 \pm 0.0042$ |
| DuoRec | $0.3025 \pm 0.0058$ | $0.3144 \pm 0.0036$ | $0.3097 \pm 0.0039$ | $0.2935 \pm 0.0035$ |
| **DISR** | $\mathbf{0.3559 \pm 0.0035}$ | $\mathbf{0.3609 \pm 0.0029}$ | $\mathbf{0.3532 \pm 0.0039}$ | $\mathbf{0.3337 \pm 0.0040}$ |
| *Lift over best* | *+14.5%* | *+12.4%* | *+11.5%* | *+11.8%* |
| *Amazon Sports (50K)* | | | | |
| BERT4Rec | $0.2482 \pm 0.0037$ | $0.2503 \pm 0.0025$ | $0.2444 \pm 0.0020$ | $0.2279 \pm 0.0034$ |
| FMLP-Rec | $0.2052 \pm 0.0028$ | $0.2030 \pm 0.0016$ | $0.1899 \pm 0.0027$ | $0.1613 \pm 0.0028$ |
| MoRec | $0.2795 \pm 0.0002$ | $0.2823 \pm 0.0005$ | $0.2859 \pm 0.0008$ | $0.2948 \pm 0.0006$ |
| SASRec | $0.2790 \pm 0.0027$ | $0.2795 \pm 0.0020$ | $0.2779 \pm 0.0026$ | $0.2670 \pm 0.0038$ |
| CL4SRec | $0.2730 \pm 0.0025$ | $0.2739 \pm 0.0028$ | $0.2732 \pm 0.0013$ | $0.2636 \pm 0.0019$ |
| DuoRec | $0.2667 \pm 0.0061$ | $0.2685 \pm 0.0050$ | $0.2665 \pm 0.0029$ | $0.2587 \pm 0.0051$ |
| **DISR** | $\mathbf{0.3333 \pm 0.0018}$ | $\mathbf{0.3322 \pm 0.0015}$ | $\mathbf{0.3271 \pm 0.0015}$ | $\mathbf{0.3055 \pm 0.0019}$ |
| *Lift over best* | *+19.3%* | *+17.7%* | *+14.4%* | *+3.6%* |

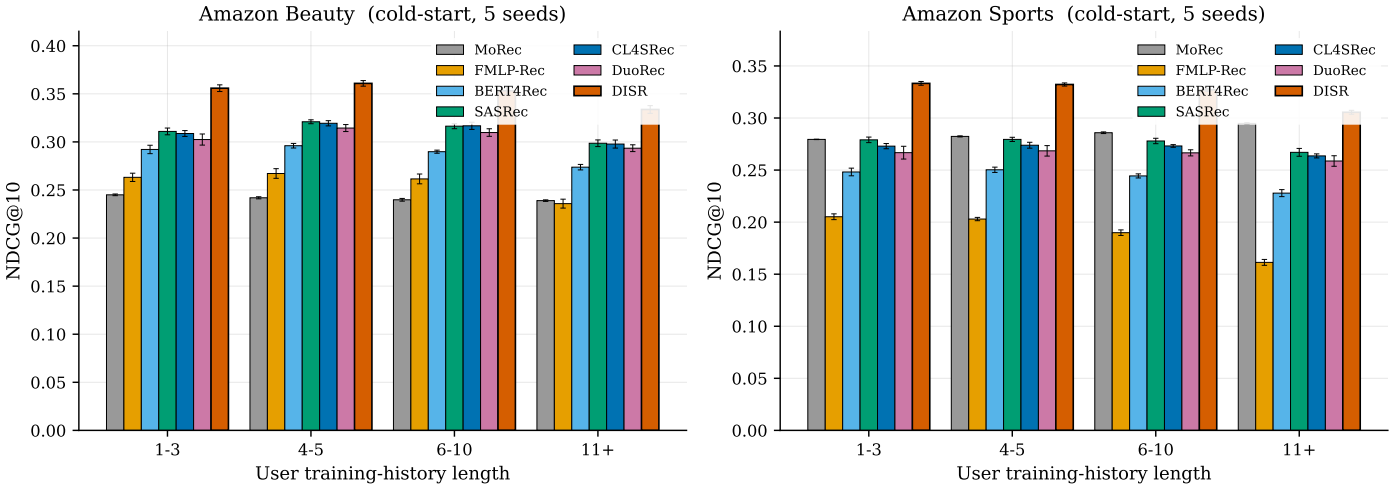


Fig. 6. NDCG@10 by user-history-length bin on the 50K subsample of beauty (left) and sports (right).

TABLE VII. LEARNED PER-ITEM VARIANCE BY TRAINING-SET ITEM FREQUENCY ON BEAUTY 50K

| Frequency bin | Item count | $\log \sigma^2$ | $\sigma^2$ |
|---|---|---|---|
| 0 | 13,425 | $-0.104$ | 0.90 |
| 1 | 57,123 | $-0.135$ | 0.87 |
| 2 | 25,139 | $-0.129$ | 0.88 |
| 3–5 | 21,197 | $-0.110$ | 0.90 |
| 6–10 | 6,935 | $-0.075$ | 0.93 |
| 11–50 | 4,653 | $-0.031$ | 0.97 |
| 51+ | 392 | $+0.039$ | 1.04 |

behavioural data accumulates — more interactions per user, more interactions per item, more total examples — the contrastive baselines can learn comparable representations from data alone, and the relative value of DISR's priors shrinks. This pattern is most clearly visible in the Sports cold-start figure: DISR's $+19.3\%$ advantage in the 1–3-interaction bin collapses to $+3.6\%$ in the 11+ bin, while every other method's accuracy is essentially flat across bins.

### B. Reading the Beauty Full-Scale Tie

The honest middle result is the Beauty full-scale tie (Table III): DISR trails CL4SRec by $-0.8\%$ NDCG@10 while leading HR@10. Two readings are defensible. The first: DISR's 50K gains were partly a small-data effect, and contrastive baselines catch up at scale. The second: DISR reaches this competitive accuracy in 208.9 minutes versus CL4SRec's 1,191.9 ($5.7\times$ speedup), so the operating point differs even under a tie. Both readings are reported without qualification.

The results do not support a claim that DISR is universally dominant on accuracy. The results support the conclusion that DISR is the best-or-tied method on all six measured conditions (50K Beauty, 50K Sports, full Beauty, full Sports, cold-start across both datasets), with a clear efficiency advantage that holds in every condition where accuracy is competitive.

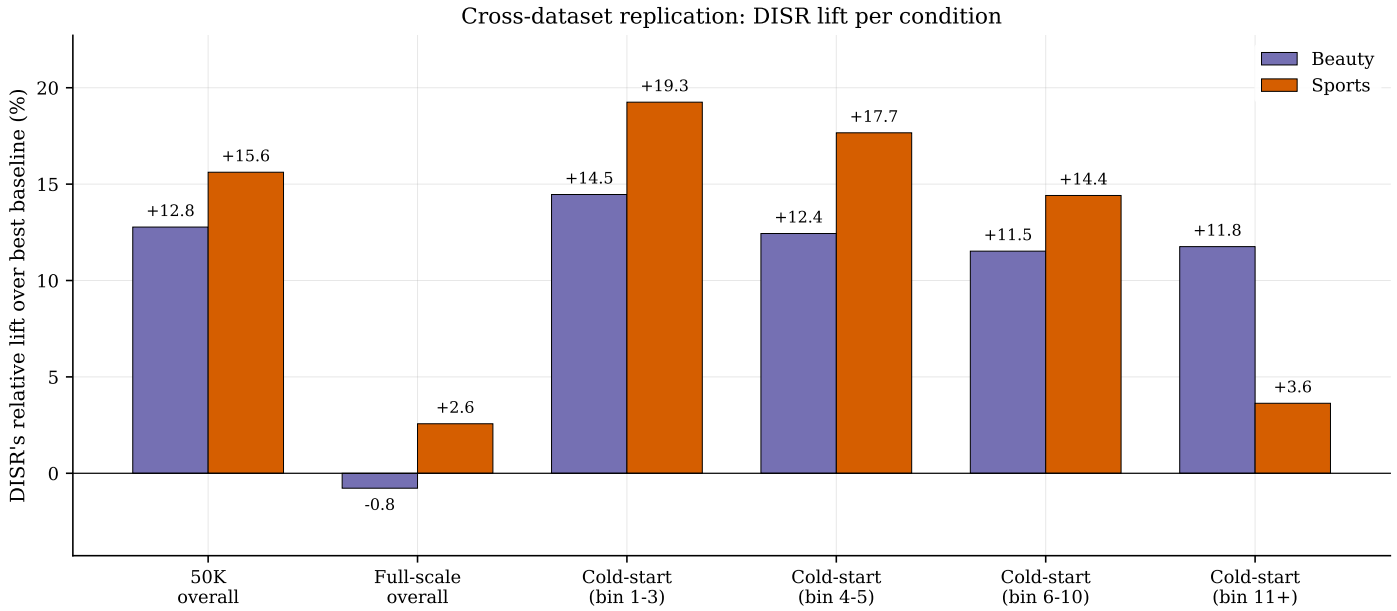


Fig. 7. Cross-dataset summary of DISR's relative NDCG@10 improvement over the strongest non-DISR baseline.

### C. The Asymmetric Multi-Seed Protocol

Five seeds were used at 50K scale, whereas only seed 42 was used at full scale, reflecting a compute constraint rather than a methodological choice. The 50K per-seed standard deviations (0.001–0.004 NDCG@10) are much smaller than inter-method gaps at full scale, supporting the orderings; the closest call is the Beauty full-scale tie with CL4SRec ($-0.8\%$), where no statistical superiority claim is made but competitive parity is observed at substantially lower training cost. Re-running full-scale with multiple seeds is a natural next step.

### D. Limitations

Three limitations of the present work are identified.

*1) Sentence-BERT as a fixed encoder:* DISR's item means come from a frozen `all-MiniLM-L6-v2` encoder. This choice is deliberate — it makes the architecture comparable to MoRec and avoids the extra cost of fine-tuning a text encoder — but it is also restrictive. A targeted encoder-sensitivity check on Sports 50K was therefore performed by replacing MiniLM with the larger `all-mpnet-base-v2` encoder. The resulting NDCG@10 was 0.3329, compared with 0.3266 for MiniLM under the same seed and protocol, indicating comparable performance with a modest $+1.9\%$ relative improvement. This provides evidence that the observed result is not specific to the MiniLM checkpoint, but it does not establish that encoder choice is immaterial across datasets, seeds, or domain-adapted or LLM-derived encoders.

*2) Diagonal gaussians only:* Item and query distributions are restricted to diagonal Gaussians. The closely related Wasserstein-elliptical work [13] explores full-covariance distributions and argues they are more expressive; diagonal Gaussians are used because they admit closed-form KL with simple gradients and avoid matrix-square-root operations at scale. The trade-off is a less expressive variance representation. Whether full-covariance Gaussians (with Wasserstein/Bures scoring or a regularized KL variant) would outperform the diagonal case on recommendation tasks is an open question.

*3) Two categories, one product family:* The evaluation uses Amazon Beauty and Amazon Sports from the Amazon Reviews 2023 benchmark. Both are product-catalogue datasets in English. The evaluation does not cover music, movies, news, or social media feeds, where the temporal dynamics and item-text characteristics differ. The cross-category replication between Beauty and Sports is encouraging but does not establish that DISR generalises to fundamentally different recommendation domains.

### E. Implications and Future Directions

Several lines of follow-on work suggest themselves. A Wasserstein/Bures variant [13, 24, 30] would test whether the geometric properties that motivate elliptical formulations in word embeddings carry over to sequential recommendation. Combining DISR with InfoNCE-style contrastive losses over distributional sequence views is a natural compositional step neither family has attempted. DISR's query variance, currently used only inside scoring, could be investigated as a potential distributional-spread signal for downstream re-ranking, exploration, or fallback mechanisms, with calibration requiring dedicated evaluation. Finally, evaluation on non-Amazon benchmarks with richer temporal structure (MovieLens, Steam, session-based settings) would test the generality of the design.

## VIII. Conclusion

DISR is presented as a sequential recommender that departs from the point-embedding scoring contract of the SASRec family while remaining compatible with its standard sampled-softmax training and evaluation framework. Items are represented as diagonal Gaussian distributions with text-initialized means and learnable per-item variances; the query is produced as a Gaussian distribution from the Transformer encoder via

two parallel heads; scoring uses the closed-form symmetric KL divergence between query and candidate distributions, computed in $O(d)$ per pair without matrix square roots or iterative procedures. This combination of design choices — distributional items, distributional query, KL scoring, sampled-softmax cross-entropy training — has, to the best of the authors' knowledge, not previously appeared together in sequential recommendation.

Across two categories, two scales, six baselines, and a five-seed protocol at the smaller scale, DISR leads or ties on every condition tested (Sections V–VI), with its largest margins in the cold-start regime and a $4$–$7\times$ reduction in epochs to convergence relative to contrastive baselines.

Three properties of the design are relevant to deployed recommender systems. First, the closed-form scoring function introduces no matrix-based or iterative computation, which supports latency-bounded inference. Second, the convergence-speed advantage reduces the cost of retraining, an important consideration for catalogues that turn over rapidly. Third, the largest empirical accuracy gains occur for users with short histories, where behavioural evidence is sparse. The present experiments do not establish equivalent gains for genuinely unseen items, whose learned variance requires a separate initialization or fallback.

The results also have important limitations. DISR is not the single best method on Beauty full-scale NDCG@10, where CL4SRec edges it by $-0.8\%$ in a three-way top cluster. The full-scale evaluation is single-seed, so statistical claims at that scale are limited; the $50$K-scale standard deviations support the orderings but cannot certify them. The two-category replication strengthens the evidence relative to single-dataset work but does not establish that DISR generalises to domains beyond product catalogues. Each of these is a direction for follow-on work rather than a defect of the present results.

The broader claim is structural: representing items and queries as distributions opens a design space beyond point-embedding scoring, in which distributional spread and prior-guided geometry can be encoded. DISR occupies one point in that space; the empirical results suggest the space is worth further exploration.

## Declaration on Generative AI

During the preparation of this manuscript, the author used Anthropic Claude for language refinement and prose compression. All scientific content — the method, experiments, results, analyses, and conclusions — was developed and validated by the human author, who takes full responsibility for the work as published.

## References

[1] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2018, pp. 197–206.

[2] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, 2019, pp. 1441–1450.

[3] K. Zhou, H. Yu, W. X. Zhao, and J.-R. Wen, "Filter-enhanced MLP is all you need for sequential recommendation," in *Proceedings of the ACM Web Conference (WWW)*, 2022, pp. 2388–2399.

[4] X. Xie, F. Sun, Z. Liu, S. Wu, J. Gao, J. Zhang, B. Ding, and B. Cui, "Contrastive learning for sequential recommendation," in *Proceedings of the 38th IEEE International Conference on Data Engineering (ICDE)*, 2022, pp. 1259–1273.

[5] R. Qiu, Z. Huang, H. Yin, and Z. Wang, "Contrastive learning for representation degeneration problem in sequential recommendation," in *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM)*, 2022, pp. 813–823.

[6] Y. Chen, Z. Liu, J. Li, J. McAuley, and C. Xiong, "Intent contrastive learning for sequential recommendation," in *Proceedings of the ACM Web Conference (WWW)*, 2022, pp. 2172–2182.

[7] Z. Yuan, F. Yuan, Y. Song, Y. Li, J. Fu, F. Yang, Y. Pan, and Y. Ni, "Where to go next for recommender systems? ID- vs. modality-based recommender models revisited," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2023, pp. 2639–2649.

[8] Y. Hou, S. Mu, W. X. Zhao, Y. Li, B. Ding, and J.-R. Wen, "Towards universal sequence representation learning for recommender systems," in *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2022, pp. 585–593.

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 3982–3992.

[10] L. Vilnis and A. McCallum, "Word representations via Gaussian embedding," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[11] S. He, K. Liu, G. Ji, and J. Zhao, "Learning to represent knowledge graphs with Gaussian embedding," in *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM)*, 2015, pp. 623–632.

[12] B. Athiwaratkun and A. G. Wilson, "Multimodal word distributions," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 1645–1656.

[13] B. Muzellec and M. Cuturi, "Generalizing point embeddings using the Wasserstein space of elliptical distributions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[14] Y. Hou, J. Li, Z. He, A. Yan, X. Chen, and J. McAuley, "Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders," *arXiv preprint arXiv:2403.03952*, 2024, dataset available at https://amazon-reviews-2023.github.io/. [Online]. Available: https://arxiv.org/abs/2403.03952

[15] W. Krichene and S. Rendle, "On sampled metrics for item recommendation," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2020, pp. 1748–1757.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 4171–4186.

[17] K. Zhou, H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, and J.-R. Wen, "$S^3$-rec: Self-supervised learning for sequential recommendation with mutual information maximization," in *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, 2020, pp. 1893–1902.

[18] J. Gao, D. He, X. Tan, T. Qin, L. Wang, and T.-Y. Liu, "Representation degeneration problem in training natural language generation models," *arXiv preprint arXiv:1907.12009*, 2019. [Online]. Available: https://arxiv.org/abs/1907.12009

[19] S. Rajput, N. Mehta, A. Singh, R. H. Keshavan, T. Vu, L. Heldt, L. Hong, Y. Tay, V. Q. Tran, J. Samost, M. Kula, E. H. Chi, and M. Sathiamoorthy, "Recommender systems with generative retrieval," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[20] K. Bao, J. Zhang, Y. Zhang, W. Wang, F. Feng, and X. He, "TALLRec: An effective and efficient tuning framework to align large language model with recommendation," in *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, 2023, pp. 1007–1014.

[21] D. Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara, "Variational autoencoders for collaborative filtering," in *Proceedings of the World Wide Web Conference (WWW)*, 2018, pp. 689–698.

[22] Q.-T. Truong, A. Salah, and H. W. Lauw, "Bilateral variational autoencoder for collaborative filtering," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM)*, 2021, pp. 292–300.

[23] Y. Yang and F. Buettner, "Multi-output Gaussian processes for uncertainty-aware recommender systems," in *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021.

[24] N. Courty, R. Flamary, and M. Ducoffe, "Learning Wasserstein embeddings," *arXiv preprint arXiv:1710.07457*, 2018.

[25] H. Lee, J. Im, S. Jang, H. Cho, and S. Chung, "MeLU: Meta-learned user preference estimator for cold-start recommendation," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019, pp. 1073–1082.

[26] M. Volkovs, G. W. Yu, and T. Poutanen, "DropoutNet: Addressing cold start in recommender systems," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4957–4966.

[27] O. Barkan and N. Koenigstein, "Item2vec: Neural item embedding for collaborative filtering," in *Proceedings of the IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6.

[28] M. Kula, "Metadata embeddings for user and item cold-start recommendations," in *Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems (CBRecSys 2015) co-located with the 9th ACM Conference on Recommender Systems (RecSys)*, ser. CEUR Workshop Proceedings, vol. 1448. CEUR-WS.org, 2015, pp. 14–21.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[30] R. Bhatia, T. Jain, and Y. Lim, "On the Bures-Wasserstein distance between positive definite matrices," *Expositiones Mathematicae*, vol. 37, no. 2, pp. 165–191, 2019.

[31] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th International Conference on World Wide Web (WWW)*, 2017, pp. 173–182.

[32] R. Cañamares and P. Castells, "On target item sampling in offline recommender system evaluation," in *Proceedings of the 14th ACM Conference on Recommender Systems (RecSys)*, 2020, pp. 259–268.