

# Revisiting CORE for Session-based Recommendation: Dual-Attention and Hard-Negative Extensions

Liran Smadja, Itay Mizikov  
*Ben-Gurion University, Beer-Sheva, Israel*  
`{smadjali, itaymiziki}@post.bgu.ac.il`

**Project Repository:** <https://github.com/Itaymizik/RecSys---CORE>

## Abstract

Session-based recommendation addresses the problem of predicting the next item a user will interact with from the short sequence of actions observed in an ongoing session, often without reliable long-term user profiles. This setting is practically important because many recommendations must be delivered to anonymous or first-time users, and because user intent can shift across sessions. Recent surveys position session-based recommender systems as a distinct paradigm that emphasizes short-term yet dynamic preference modeling and timely adaptation to evolving session context.

This work is motivated by a specific modeling tension highlighted by CORE: when a session embedding is produced by a non-linear encoder, it may not lie in the same representation space as item embeddings, and this mismatch can lead to inconsistent scoring behavior during decoding. CORE proposes to unify the representation space for encoding and decoding via a representation-consistent encoder based on linear combination, together with a robust distance measuring strategy designed to reduce overfitting in a consistent representation space.

Building on these foundations, we revisit CORE from the perspective of representation consistency and propose two extensions that preserve its geometric principle. First, we introduce a Dual-Attention encoder with adaptive fusion of global session context and recent intent while keeping the session representation as a linear combination of item embeddings. Second, we propose a hard-negative-aware hybrid objective that combines cross-entropy with a pairwise ranking term focused on hard negatives. The remainder of this paper presents the background, formalizes these extensions, and outlines an empirical evaluation under the CORE protocol.

Across the five CORE benchmarks, our extensions yield modest but consistent rank-quality gains: CORE-DA improves MRR@20 over CORE-trm on four out of five datasets, and CORE-DAHN achieves the best MRR@20 on Diginetica and RetailRocket under a fixed seed. In addition, we evaluate on an extra dataset outside the CORE suite (eShop2008). On eShop2008, CORE-TRM remains the strongest baseline, while CORE-DA, CORE-HN, and CORE-DAHN are competitive but do not surpass it, suggesting that the benefits of dual attention and hard-negative mining can be dataset-dependent.

## 1 Introduction

**Research Topic:** Session-based Recommendation Systems

Recommender systems are widely used to mitigate information overload by helping users discover relevant content. In many real-world deployments, however, personalization cannot rely on a stable user profile: the platform may not know who the user is, the user may be new, or long-term histories may be inaccessible due to privacy, business constraints, or system design. Session-based recommendation focuses on this regime by tailoring suggestions to short-term needs inferred from an ongoing session typically a short sequence of interactions. This scenario is both

practically relevant and technically challenging due to shifting intent and limited within-session evidence [1].

Recent surveys characterize session-based recommender systems as a specialized paradigm that captures short-term yet dynamic user preferences [2]. Modern approaches rely on learned representations and neural architectures including sequential networks and graph neural networks to encode sessions and score candidate items [3].

Within this landscape, CORE addresses a key geometric issue: when session embeddings are produced by non-linear encoders, they may not lie in the same representation space as item embeddings. Since decoding scores items by comparing session and item embeddings, this mismatch can manifest as an “inconsistent prediction” issue [4]. CORE resolves this by unifying the representation space through a representation-consistent encoder (forming session embeddings as linear combinations of item embeddings) and a robust distance measuring approach to prevent overfitting [4].

Building on CORE, we propose two extensions that preserve its geometric principle. First, we introduce a Dual-Attention encoder with adaptive fusion of global session context and recent intent while keeping the session representation as a linear combination of item embeddings. Second, we propose a hard-negative-aware hybrid training objective that combines cross-entropy with a pairwise ranking term focused on confusing items. Together, these improvements aim to enhance CORE’s ranking quality while maintaining its representation consistency.

**Paper Organization.** Section 2 reviews related work on session-based recommendation and CORE. Section 3 introduces the problem setting and the CORE framework (representation-consistent encoding and robust decoding). Section 4 presents our proposed Dual-Attention encoder and the hard-negative hybrid objective. Section 5 describes the evaluation protocol, datasets, metrics, and experimental setup. Section 6 reports results and ablations, and Section 7 discusses the findings and limitations. Section 8 concludes and outlines future directions.

## 2 Related Work

### 2.1 Problem Framing and Evaluation in Session-based Recommendation

Session-based recommendation addresses the task of predicting the next item a user will interact with, given only the short sequence of actions observed within the current session. The setting is practically demanding for at least three reasons that distinguish it from the more classical collaborative-filtering or content-based paradigms. First, user identity is frequently absent: recommendations must be generated for anonymous or first-time visitors for whom no long-term interaction history is available. Second, user intent within a single session is inherently dynamic: a user may begin browsing one product category and pivot to a semantically different one within the same visit, so a model that captures only average session preferences will systematically miss these late-session shifts. Third, sessions are short by nature, and the resulting within-session evidence is sparse; even a few irrelevant clicks can distort a naive aggregation of session content. These challenges have been carefully articulated in the handbook treatment of session-based recommenders, which motivates the problem through practical deployment scenarios and identifies anonymity, short-term yet dynamic preference modeling, and timely adaptation as central design axes [1].

Evaluation in session-based recommendation has converged on a set of standard conventions that differ from long-horizon user modeling. Because the ground truth is a single next item, hit-based metrics such as Recall@ $K$  (equivalently Hit Rate@ $K$ ) and rank-sensitive measures such as MRR@ $K$  are the dominant choices, with  $K=20$  being widely used. Survey work on session-based systems reinforces this framing, arguing that the problem constitutes a *distinct paradigm* relative to static or long-term preference modeling, and calling for clearer problem statements

that explicitly acknowledge the within-session dynamics and evaluation challenges specific to this regime [2]. Positioning new methods relative to these conventions rather than those of general collaborative filtering is therefore essential for ensuring that reported improvements are meaningful and commensurable across published work.

## 2.2 Sequential Encoders for Session Modeling

A natural first family of neural session encoders treats each session as a strictly ordered sequence and applies recurrent or attention-based architectures to propagate dependencies forward in time. Early recurrent approaches demonstrated that gating mechanisms can selectively retain or discard signals from past interactions, which is valuable when sessions contain noise or spurious clicks. Attention-based and Transformer-style encoders later emerged as a more expressive alternative, allowing non-local dependencies between any two positions in the session and enabling explicit queries over all past interactions to compute an aggregated session state. Surveys on the evolution of session modeling highlight that the transition from purely recurrent to attention-based sequential encoders improved the ability to model informative interactions under short and noisy conditions, while also better capturing recency effects when positional encodings or explicit recency biases are introduced [2, 3].

A persistent challenge within this family is the tension between recency and global context. Many attention-based encoders aggregate all interactions with a single learned importance vector, which may not adequately reflect late-session intent shifts where the last few items carry disproportionate signal about where the user is heading next. Conversely, architectures that focus exclusively on the most recent interaction risk discarding useful global cues embedded in the session’s broader history. This recency-versus-context dilemma motivates our Dual-Attention encoder, which computes two separate importance weight vectors one capturing global session context and one explicitly driven by the most recent interaction and fuses them via a learnable gating mechanism, an approach not present in prior sequential encoder designs [2].

Compared to CORE, sequential encoders in this broader family typically do not constrain the session embedding to lie within any particular geometric subspace, and the resulting representation-space mismatch can produce inconsistent scoring behavior at inference time. CORE-trm addresses this by ensuring that the final session vector is produced as a linear combination of item embeddings, but it relies on a single learned importance vector over contextualized token representations [4]. Our work inherits CORE-trm’s Transformer-based contextualization and its linear-combination constraint, while modifying only how importance weights are computed and fused: the session embedding remains by construction a linear combination of item embeddings and therefore stays in the same representation space as the candidates it is compared against, strictly extending rather than replacing the CORE-trm design [4].

## 2.3 Graph-based Session Modeling

A second major family of neural approaches represents sessions as directed graphs in which nodes correspond to items and edges encode transition or co-occurrence patterns. By applying graph neural network (GNN) message-passing over these session graphs, models in this family can capture multi-hop relational signals that go beyond strict sequential adjacency: an item visited early in a session may have a strong transitional relationship with an item visited several steps later, and graph representations can surface this without requiring the recurrence depth of a sequential model. Recent survey work highlights that graph-based formulations have become an important strand of session-based recommendation precisely because they offer richer inductive biases about item-to-item transitions, including the ability to model repeated interactions and non-linear navigation patterns within a session [3, 1].

Despite these modeling advantages, graph-based approaches introduce several engineering and theoretical challenges. Session-level graphs are typically constructed dynamically one per session so the graph topology varies across instances and must be handled by architectures robust to variable-size and variable-density inputs. The interaction between GNN message-passing and the item embedding space also raises a representational concern: after multiple rounds of neighborhood aggregation through non-linear activation functions, the aggregated session vector is no longer a simple linear function of the original item embeddings. This non-linearity can reintroduce the very representation-space inconsistency that CORE was designed to address, and the additional complexity of dynamic graph construction and multi-round propagation increases computational overhead while complicating attribution of performance gains to any single mechanism [4, 3].

Compared to our approach, graph-based methods trade simplicity and geometric transparency for potentially richer relational expressiveness. We deliberately remain within the CORE design philosophy of architectural minimalism and representation-space consistency: we modify only the importance-weighting mechanism within a linear-combination encoder and adjust the training objective, making no changes to session-graph topology or multi-hop propagation. This design choice allows empirical gains to be attributed cleanly to the two proposed mechanisms rather than to increased architectural complexity, which is especially important given that our goal is a controlled, interpretable extension of CORE rather than a full architecture redesign [4].

## 2.4 Training Objectives and Negative Sampling

Most next-item session models are optimized with a cross-entropy classification objective over the full item catalog, which treats the true next item as the positive class and all other items as negatives. This full-softmax formulation is consistent in the sense that the model is trained with the same candidate set it will face at inference, but it can be expensive on large catalogs and, more importantly, it assigns non-negligible loss weight to the many trivially easy negatives items that are semantically or representationally distant from the session and therefore confidently rejected even at the start of training. The gradient signal from easy negatives is small in magnitude and contributes little to learning meaningful discriminative boundaries, a problem that persists even under sampled-softmax approximations unless the sampling distribution is carefully designed [2].

The recognition that not all negatives contribute equally motivates pairwise ranking objectives as a complementary training paradigm. Rather than classifying the positive item over the full catalog in a single multinomial step, pairwise methods directly optimize a margin or smooth surrogate between a positive item and one or more negatives, focusing the training signal on the relative ordering of items. The canonical formulation in this tradition is Bayesian Personalized Ranking (BPR) [5], which maximizes the posterior probability that an observed positive item is ranked above a uniformly sampled unobserved item. BPR has been widely adopted as a ranking-aware training objective in recommender systems and remains a standard baseline for pairwise methods because of its simplicity, probabilistic interpretation, and straightforward gradient structure [5].

Our hybrid objective extends this pairwise intuition in a targeted way. Instead of sampling negatives uniformly, we mine *hard negatives* the top- $K$  highest-scoring incorrect items under the current model and apply a BPR-style smooth pairwise penalty specifically to these confusing candidates. The resulting combined loss is a weighted sum of the standard full-catalog cross-entropy and the pairwise term over mined hard negatives, with a scalar coefficient  $\lambda$  controlling their relative contribution [5]. Crucially, the hard-negative mining step is treated as a non-differentiable selection operator (stop-gradient), so it does not alter the inference-time scoring function; at test time, items are still ranked by cosine similarity between the session embedding and item embeddings in exactly the same way as in CORE, preserving the inference procedure while strengthening the training signal on the most confusing items [4, 5].

## 2.5 Representation Consistency and CORE-style Methods

The motivation for representation consistency arises from a geometric observation about the standard encode-then-decode pipeline used in neural recommendation. In this pipeline, an encoder transforms the observed session into a single session embedding  $\mathbf{h}_s$ , and a decoder scores each candidate item  $v$  by comparing  $\mathbf{h}_s$  with the item’s embedding  $\mathbf{h}_v$  typically via dot product or cosine similarity. When the encoder is non-linear, the session embedding generally lies in a region of the embedding space that is not spanned by any linear combination of item embeddings, meaning that the encoding space and the decoding space are geometrically misaligned. CORE formalizes this as an *inconsistent prediction* problem: identical or closely related sessions may be mapped to disparate regions, producing unstable similarity behavior with respect to item embeddings [4].

CORE addresses this by proposing two complementary mechanisms. The Representation-Consistent Encoder (RCE) constructs the session embedding as a weighted linear combination of item embeddings,  $\mathbf{h}_s = \sum_{i=1}^m \alpha_i \mathbf{h}_{v_i}$ , where the weights  $\{\alpha_i\}$  are produced by a neural sub-network but the final session vector is explicitly constrained to the affine span of the items appearing in the session. CORE instantiates RCE with two variants: a mean-pooling variant with uniform weights and a Transformer-based variant (CORE-trm) with learned attention-based importance weights. The Robust Distance Measuring (RDM) component complements RCE by applying dropout to item embeddings during scoring and introducing a tunable temperature parameter, reducing overfitting within the consistent representation space and connecting the training loss to a tuplet-loss-style objective [4].

Our Dual-Attention extension builds directly on the CORE-trm variant of RCE. We compute two importance weight vectors a global-context attention vector derived from all position-aware token representations, and a recent-intent attention vector conditioned on the last item’s contextualized representation as a query each yielding a separate session embedding that is itself a linear combination of item embeddings. A learnable softmax gate then fuses the two, and the final session vector is a convex combination of two linear combinations, which is itself a linear combination of item embeddings; the representation-consistent form required by CORE is therefore preserved by construction [4]. Similarly, the hard-negative hybrid objective modifies only the training loss: the pairwise BPR term is computed using the same cosine-similarity scoring function that CORE uses for cross-entropy, and the stop-gradient mining step ensures that the inference-time computational graph is identical to CORE-trm. Representation consistency is thus a property of the *model architecture* rather than the *training loss*, and our hybrid objective strengthens supervision on confusing negatives without violating this property [4, 5].

## 2.6 Summary of Differences

The preceding subsections reveal a landscape of trade-offs that positions our approach as a principled and controlled extension of CORE. Typical sequential models whether recurrent or Transformer-based prioritize rich temporal dependency modeling but do not constrain their session embeddings to remain in the item embedding space, introducing the representation-space mismatch that CORE identifies as a root cause of inconsistent scoring; our approach inherits the Transformer’s sequential inductive bias while resolving this mismatch through the same linear-combination constraint as CORE, differing only in how importance weights are computed [4, 2]. Typical graph-based models go further in expressive power by capturing multi-hop relational signals through GNN message-passing, but this expressiveness comes at the cost of dynamic graph construction overhead, non-linear aggregation across multiple propagation rounds, and a loss of the geometric transparency that representation-consistent encoders provide; our method deliberately forgoes this additional complexity, retaining CORE’s minimalist architecture so that any empirical gain can be attributed directly to the proposed mechanisms [3, 1].

CORE itself, as a direct baseline, offers strong simplicity and geometric clarity through

its linear-combination encoder and robust distance measuring strategy, but its single-weight aggregation in CORE-trm does not explicitly disentangle global session signals from recent intent, and its training objective is a standard cross-entropy over the full catalog that treats all non-target items as equally informative negatives regardless of their proximity to the current session representation [4]. CORE-DA addresses the first limitation by computing separate global-context and recent-intent attention vectors and fusing them via a learnable gate, while preserving the linear-combination form of the encoder; CORE-DAHN additionally addresses the second limitation by supplementing cross-entropy with a BPR-style pairwise term over mined hard negatives, strengthening the training signal on the most confusing items without altering inference [4, 5]. A further point of contrast concerns ablation clarity: CORE-DAHN reduces *exactly* to CORE-trm when the recent-intent branch is disabled, the gate is fixed to the global branch, the residual mean-pooling path is removed, and the pairwise coefficient  $\lambda$  is set to zero, making the empirical comparison between our models and CORE a controlled ablation rather than an architecture-replacing redesign [4]. This design philosophy incremental, geometrically principled, and fully interpretable relative to the baseline reflects a broader methodological preference for targeted modifications over black-box performance improvements, and positions CORE-DA and CORE-DAHN as the natural next step within the representation-consistent paradigm that CORE established [4, 2, 3].

## 3 Background

### 3.1 Problem Setting

We adopt the standard session-based recommendation setting. Let  $\mathcal{V}$  denote the item set. A session  $s$  is an ordered sequence of interacted items,

$$s = [v_1, v_2, \dots, v_m], \quad v_i \in \mathcal{V}, \quad (1)$$

where  $m$  is the session length. The goal is to predict the next interacted item  $v_{m+1}$  given the observed prefix  $[v_1, \dots, v_m]$ . In many practical cases, user identity and long-term histories are not available or not reliable, and the model must infer short-term intent from the within-session interactions alone [1].

A common neural formulation introduces an item embedding table  $\mathbf{E} \in R^{|\mathcal{V}| \times d}$ , where each item  $v$  is mapped to an embedding  $\mathbf{h}_v \in R^d$ . Given a session, the model constructs a session representation  $\mathbf{h}_s$  (encoding) and then scores candidate items (decoding) by comparing  $\mathbf{h}_s$  with candidate item embeddings.

### 3.2 Representation Consistency and the CORE Motivation

A core observation in CORE is that the encoding-decoding interface can be geometrically inconsistent. Many session encoders are non-linear (for example, recurrent encoders), and thus the resulting session embedding is not guaranteed to lie in the same linear space spanned by item embeddings. Yet decoding frequently relies on direct similarity computations between  $\mathbf{h}_s$  and item embeddings. CORE argues that this mismatch can lead to an “inconsistent prediction” issue: even when sessions share an objective, non-linear encoding can map them to different regions of space, producing unstable or inconsistent similarity behavior with respect to item embeddings [4].

CORE addresses this by unifying the representation space for encoding and decoding, with two key components: a representation-consistent encoder (RCE) and a robust distance measuring (RDM) strategy for decoding [4]. The emphasis is not on a more complex encoder architecture, but rather on ensuring that the scoring space is coherent by construction.

### 3.3 CORE: Representation-Consistent Encoder

Given a session  $s = [v_1, \dots, v_m]$ , CORE first obtains item embeddings  $\mathbf{h}_{s,i}$  for each interaction via the embedding table. The representation-consistent encoder defines the session embedding as a weighted linear combination of item embeddings:

$$\mathbf{h}_s = \sum_{i=1}^m \alpha_{s,i} \mathbf{h}_{s,i}. \quad (2)$$

The weights  $\alpha_{s,i}$  reflect item importance within the session and are produced by a deep neural network operating on the sequence of item embeddings, but constrained so that  $\mathbf{h}_s$  remains a linear combination in the original embedding space. CORE instantiates this idea with two simple variants [4]. In the mean-pooling variant, the encoder ignores order and importance and simply uses

$$\alpha_{s,i} = \frac{1}{m}, \quad (3)$$

so that the session representation is the average of its item embeddings. In the Transformer-based variant, CORE applies a stack of self-attention blocks over the sequence  $[\mathbf{h}_{s,1}; \dots; \mathbf{h}_{s,m}]$  to obtain contextualized token representations  $\mathbf{F} \in R^{m \times d'}$ , and then computes normalized importance weights via a learned vector  $\mathbf{w} \in R^{d'}$ :

$$\boldsymbol{\alpha} = \text{softmax}(\mathbf{w} \mathbf{F}^\top). \quad (4)$$

In both variants, the resulting session embedding  $\mathbf{h}_s$  is explicitly formed as a linear combination of item embeddings and therefore lies in the same representation space as the items. This consistency has a direct implication for common dot-product decoding: if one scores a target item  $v_t$  by  $\mathbf{h}_s^\top \mathbf{h}_t$  while  $\mathbf{h}_s$  is a weighted sum of item embeddings, then the score decomposes into a weighted sum of item-to-item dot products, which aligns session-to-item scoring with item-to-item similarity in the same space [4].

### 3.4 CORE: Robust Distance Measuring for Decoding

CORE further observes that simply operating in a consistent space can lead to overfitting of embeddings if the decoder is not regularized appropriately. To mitigate this, CORE revisits the widely used dot-product decoder from the perspective of triplet loss optimization and introduces a controllable temperature margin together with dropout on candidate item embeddings [4]. Let  $\mathbf{h}_s$  denote the session embedding and  $\mathbf{h}_v$  the embedding of item  $v$ . CORE applies dropout to item embeddings, yielding  $\mathbf{h}'_v$ , and measures similarity via cosine similarity with temperature  $\tau$ :

$$\ell = -\log \frac{\exp(\cos(\mathbf{h}_s, \mathbf{h}'_{v+})/\tau)}{\sum_{v \in \mathcal{V}} \exp(\cos(\mathbf{h}_s, \mathbf{h}'_v)/\tau)}, \quad (5)$$

where  $v^+$  is the ground-truth next item. This formulation replaces the fixed margin implicit in standard dot-product cross-entropy with a tunable temperature  $\tau$ , and uses dropout directly on item embeddings to improve robustness in the consistent representation space. When embeddings are  $\ell_2$ -normalized, cosine similarity and squared Euclidean distance are monotonically related, so this decoder can also be interpreted as operating on (negative) distances in the consistent space.

## 4 Methodology

### 4.1 Proposed Methodology

CORE highlights a geometric tension: decoding typically compares a session representation to item embeddings, yet common non-linear encoders do not guarantee that the session embedding

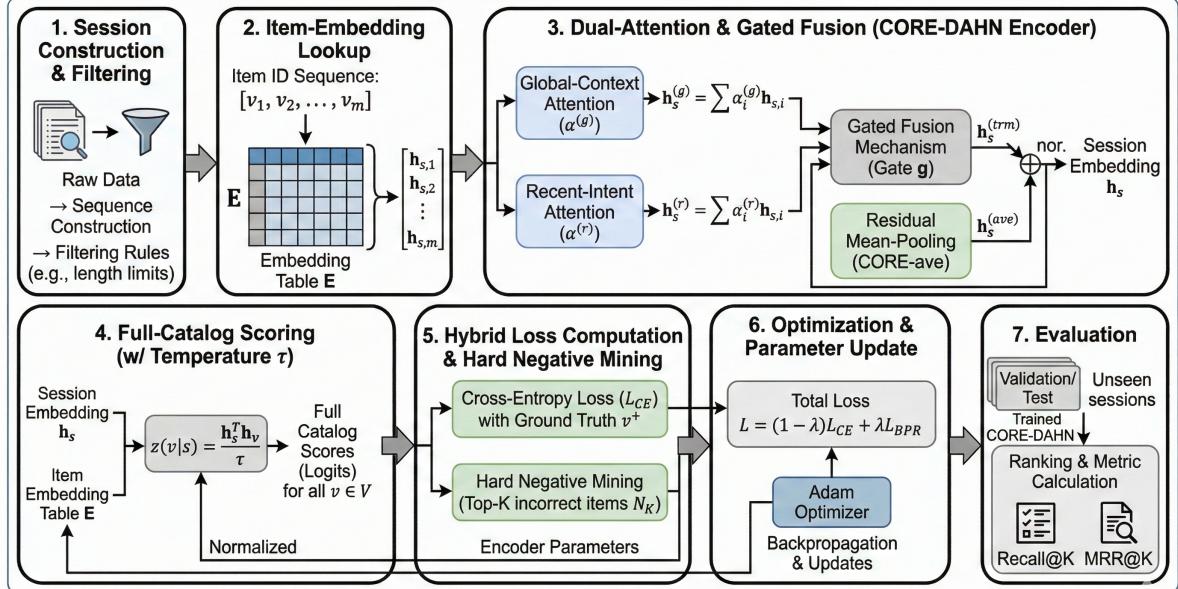
lies in the same representation space as the item embeddings. CORE resolves this by constructing the session representation as a weighted linear combination of item embeddings (Eq. 2) and by using robust distance measuring strategies to reduce overfitting in the consistent space [4]. Our methodology follows the same design principle: we do not change the representational *form* of the session embedding (it remains a linear combination in the item space), but we modify how the importance weights are computed in the Transformer-based variant, and how training exposes the model to difficult negatives. Concretely, we build on CORE-trm [4] and introduce two complementary improvements:

**Contribution 1: Dual-Attention with Adaptive Fusion.** CORE-trm learns item importance weights using self-attention blocks and then applies a softmax-based weighting scheme. While this captures sequential patterns, it can under-emphasize a frequent phenomenon in sessions: the *last interactions* may reflect the most current short-term intent. We therefore compute two sets of importance weights, one capturing a global session context and one explicitly emphasizing the last-item-driven recent intent, and then fuse the resulting session representations using a learnable gate. This preserves representation consistency because the final session embedding is still a linear combination of item embeddings.

**Contribution 2: Hard-Item Negatives with a Hybrid Objective.** CORE and many session-based models are trained with a next-item classification objective that contrasts the positive next item against a large set of negatives. However, many negatives are trivially easy, especially in large item catalogs. We therefore adopt *hard negative mining* by selecting the highest-scoring incorrect items under the current model, and add a pairwise ranking term that explicitly encourages the positive item to score higher than these hard negatives. This leaves inference unchanged (still based on similarity in the consistent space) but strengthens training signals on confusing items.

We denote our architecture-improved model as CORE-DA (Dual Attention), our training-improved model as CORE-HN (Hard Negatives), and the combined model as CORE-DAHN when both are enabled.

**Relation to CORE-trm.** Our architecture is a controlled extension of CORE-trm that preserves the representation-consistent aggregation form. If we disable the recent-intent branch (using only the global attention weights), fix the gating mechanism to always select the global branch, and remove the residual mean-pooling path, the resulting encoder reduces to the Transformer-based representation-consistent encoder used in CORE-trm [4]. Similarly, if we set  $\lambda = 0$  and do not mine hard negatives, the training objective reduces to the standard cross-entropy loss employed in CORE. Thus, CORE-DAHN recovers CORE-trm as a special case.



Overall experimental pipeline used in CORE-DAHN, from data preparation to training and evaluation.

Figure 1: Overall experimental pipeline used in CORE-DAHN, from data preparation to training and evaluation.

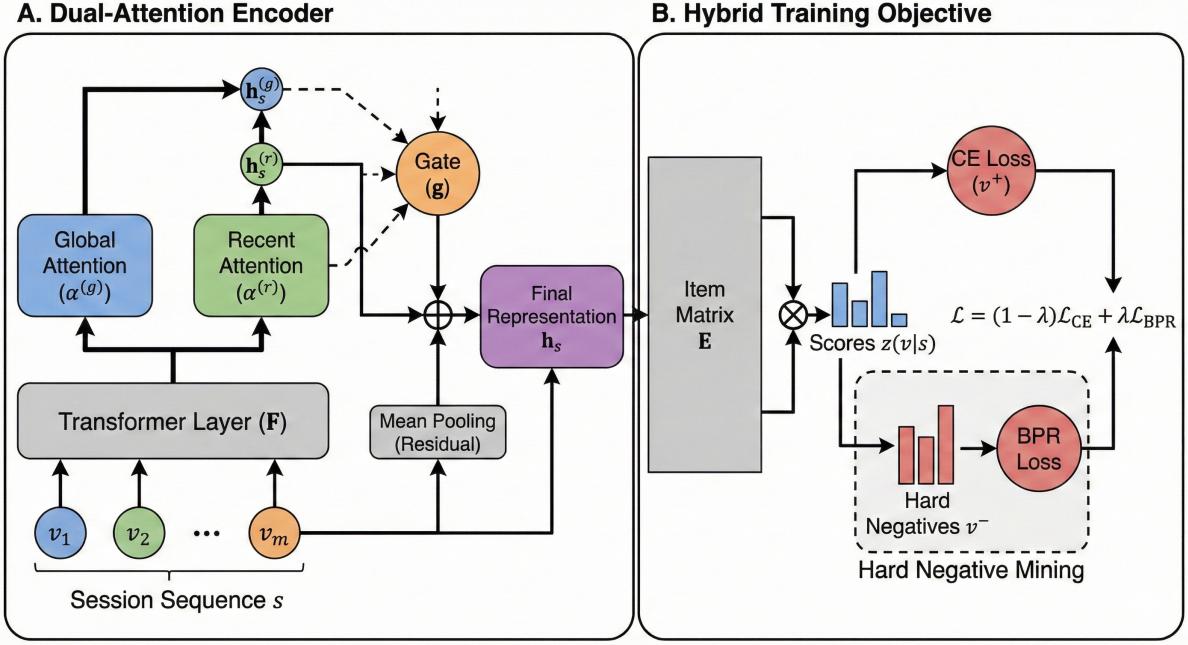


Figure 2: The overall architecture of CORE-DAHN. **(A) Dual-Attention Encoder:** Illustrates the parallel computation of global context ( $\alpha^{(g)}$ ) and recent intent ( $\alpha^{(r)}$ ), fused via an adaptive gate. **(B) Hybrid Training:** Shows the hard-negative mining process where the model is optimized using both Cross-Entropy on the positive item and BPR loss on hard negatives.

#### 4.1.1 Dual-Attention Weighting in a Consistent Space

Let a session be represented by item embeddings  $\mathbf{H}_s = [\mathbf{h}_{s,1}, \dots, \mathbf{h}_{s,m}] \in R^{m \times d}$ , where  $\mathbf{h}_{s,i}$  is the embedding of item  $v_i$ . Following CORE-trm, we obtain contextualized representations

$\mathbf{F} \in R^{m \times d'}$  using  $L$  Transformer self-attention blocks with positional encodings [4]. Instead of producing a single importance vector, we compute two attention score vectors:

**Global-context scores.** Following the CORE-trm aggregation style, we inject positional information explicitly when estimating global importance. Let  $\mathbf{p}_i \in R^{d'}$  denote the positional embedding of position  $i$  (or an equivalent position signal already used in the Transformer input). We compute global attention scores by an MLP over the concatenation of each contextualized token representation and its position signal:

$$e_i^{(g)} = \text{MLP}_g([\mathbf{f}_i; \mathbf{p}_i]), \quad \alpha_i^{(g)} = \text{softmax}(e^{(g)})_i. \quad (6)$$

This branch aims to capture the overall session context, while remaining compatible with CORE's representation-consistent linear aggregation.

**Recent-intent scores.** To explicitly emphasize recency, we use the last valid token representation as a *query* signal. Let  $\ell(s)$  be the last valid index in the padded sequence (i.e., the last interacted item in the session), and let  $\mathbf{f}_{\ell(s)}$  denote its contextualized representation. We compute recent-intent scores by conditioning each token on this last-item query:

$$e_i^{(r)} = \text{MLP}_r([\mathbf{f}_i; \mathbf{f}_{\ell(s)}]), \quad \alpha_i^{(r)} = \text{softmax}(e^{(r)})_i. \quad (7)$$

Intuitively, this branch prioritizes past interactions that best align with the most recent interaction, capturing late-session intent shifts without changing the representation-consistent form of the encoder.

**Session representations and gating.** Using the same (dropout-perturbed) item embeddings as CORE, we compute two session embeddings:

$$\mathbf{h}_s^{(g)} = \sum_{i=1}^m \alpha_i^{(g)} \hat{\mathbf{h}}_{s,i}, \quad \mathbf{h}_s^{(r)} = \sum_{i=1}^m \alpha_i^{(r)} \hat{\mathbf{h}}_{s,i}. \quad (8)$$

We then fuse them by a learnable gate  $\mathbf{g} \in R^2$ :

$$\mathbf{g} = \text{softmax}\left(\text{MLP}_{\text{gate}}([\mathbf{h}_s^{(g)}; \mathbf{h}_s^{(r)}])\right), \quad \mathbf{h}_s^{(\text{trm})} = g_1 \mathbf{h}_s^{(g)} + g_2 \mathbf{h}_s^{(r)}. \quad (9)$$

Finally, to preserve CORE's emphasis on stable consistent representations, we include the CORE-ave residual path and normalize:

$$\mathbf{h}_s^{(\text{ave})} = \sum_{i=1}^m \alpha_i^{(\text{ave})} \hat{\mathbf{h}}_{s,i}, \quad \mathbf{h}_s = \text{norm}\left(\mathbf{h}_s^{(\text{trm})} + \mathbf{h}_s^{(\text{ave})}\right), \quad (10)$$

where  $\alpha^{(\text{ave})}$  is uniform (mean pooling) as in CORE-ave [4]. This construction keeps  $\mathbf{h}_s$  in the item embedding space by design.

#### 4.1.2 Hard Negative Mining with a Hybrid CE+BPR Objective

Given the final session embedding  $\mathbf{h}_s$  and item embedding matrix  $\mathbf{E}$ , we score all items by cosine-style similarity via normalized embeddings (equivalently a dot product after  $\ell_2$  normalization), scaled by a temperature  $\tau$ :

$$z(v | s) = \frac{\mathbf{h}_s^\top \mathbf{h}_v}{\tau}, \quad \mathbf{h}_v \in \mathbf{E}. \quad (11)$$

---

**Algorithm 1** Training Step for CORE-DAHN (Dual Attention + Hard Negatives)

---

**Require:** Batch of sessions  $\{s_b\}$  with positives  $\{v_b^+\}$ , item embeddings  $\mathbf{E}$ , temperature  $\tau$ ,  $K$ ,  $\lambda$

- 1: Compute contextual token representations  $\mathbf{F}_b \leftarrow \text{Transformer}(\mathbf{H}_{s_b})$
- 2: Compute  $\alpha_b^{(g)}, \alpha_b^{(r)}$  and  $\mathbf{h}_{s_b}^{(g)}, \mathbf{h}_{s_b}^{(r)}$
- 3: Fuse with gating to obtain  $\mathbf{h}_{s_b}^{(\text{trm})}$ ; add CORE-ave residual to get  $\mathbf{h}_{s_b}$
- 4: Compute logits  $z(v | s_b) = \mathbf{h}_{s_b}^\top \mathbf{h}_v / \tau$
- 5:  $\mathcal{L}_{\text{CE}} \leftarrow$  cross-entropy over  $z(\cdot | s_b)$  and  $v_b^+$
- 6: Mine hard negatives  $\mathcal{N}_K(s_b) \leftarrow \arg \text{TopK}_{v \in \mathcal{V} \setminus \{v_b^+\}} z(v | s_b)$  ▷ stop-gradient selection
- 7:  $\mathcal{L}_{\text{BPR}} \leftarrow \frac{1}{K} \sum_{v^- \in \mathcal{N}_K(s_b)} \log(1 + \exp(z(v^- | s_b) - z(v_b^+ | s_b)))$
- 8: **return**  $\mathcal{L} = (1 - \lambda)\mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{BPR}}$

---

The standard next-item loss is the cross-entropy over the full catalog:

$$\mathcal{L}_{\text{CE}} = -\log \frac{\exp(z(v^+ | s))}{\sum_{v \in \mathcal{V}} \exp(z(v | s))}, \quad (12)$$

where  $v^+$  is the ground-truth next item.

To focus learning on confusing alternatives, we mine hard negatives by selecting the top- $K$  highest-scoring *items* under the current model, excluding the ground-truth next item:

$$\mathcal{N}_K(s) = \arg \text{TopK}_{v \in \mathcal{V} \setminus \{v^+\}} z(v | s). \quad (13)$$

In practice, the TopK selection is treated as a non-differentiable operator (stop-gradient): we first compute scores under the current model and then use the selected indices as negatives for the pairwise term.

We then add a pairwise ranking term over these negatives [5]:

$$\mathcal{L}_{\text{BPR}} = \frac{1}{K} \sum_{v^- \in \mathcal{N}_K(s)} \log(1 + \exp(z(v^- | s) - z(v^+ | s))), \quad (14)$$

which encourages the positive item to score higher than mined hard negatives (a smooth separation objective rather than a hard margin).

## 4.2 A Minimal Running Example

Consider a session  $s = [v_1, v_2, v_3]$  where  $v_3$  is the most recent interaction. CORE-style encoding constructs  $\mathbf{h}_s$  as a weighted sum of  $\mathbf{h}_{s,1}, \mathbf{h}_{s,2}, \mathbf{h}_{s,3}$ . In CORE-DA, the global attention may assign higher weight to  $v_2$  if it best represents the session’s overall context, while the recent-intent attention can place extra weight on  $v_3$  to reflect a late-stage shift in intent. The gating module then adapts the mixture based on how consistent the two signals are: when global and recent views agree, the gate tends to average them; when the last item indicates a clear pivot, the gate can emphasize the recent view. During training with CORE-HN, if the model currently ranks an incorrect item  $u$  very highly for this session,  $u$  will be selected into  $\mathcal{N}_K(s)$  and will contribute explicitly to the pairwise penalty in Eq. (??), pushing the model to rank  $v^+$  above  $u$  under the same session embedding  $\mathbf{h}_s$ .

## 5 Evaluation

### 5.1 Research Questions and Hypotheses

Our empirical study is structured around three questions. First, we ask whether improving importance-weight estimation via dual attention yields gains over CORE-trm across standard

Table 1: Statistics of the five CORE benchmark datasets (as reported in [4]).

Dataset	#Interactions	#Items	#Sessions	Avg. Length
Diginetica	786,582	42,862	204,532	4.12
Nowplaying	1,085,410	59,593	145,612	9.21
RetailRocket	871,637	51,428	321,032	6.40
Tmall	427,797	37,367	66,909	10.62
Yoochoose	1,434,349	19,690	470,477	4.64

benchmarks. Second, we ask whether hard-negative-aware training improves ranking quality beyond standard cross-entropy training, particularly by reducing confusions among high-scoring incorrect items. Third, we analyze whether the two improvements are complementary, i.e., whether combining them outperforms applying each in isolation.

## 5.2 Preparing Experimental Data

To ensure comparability with CORE, we follow the same evaluation protocol and public benchmarks as reported in the CORE study, which includes five datasets and a standard train/validation/test split ratio of 8:1:1 [4]. In addition, following the course requirement to implement the reproduced method on an extra dataset beyond the original paper’s benchmark suite, we evaluate the same models on eShop2008 [4]. We apply the same top- $K$  evaluation settings (e.g.,  $K = 20$ ) and report Recall@20 and MRR@20 consistently across all datasets.

## 5.3 Dataset Statistics

Table 1 summarizes basic statistics of the five CORE benchmarks (number of interactions, items, sessions, and average session length), following the dataset preprocessing described in CORE [4]. Short average session lengths (e.g., Diginetica and Yoochoose) emphasize recency and intent shifts, whereas longer sessions (e.g., Nowplaying and Tmall) amplify the need for robust importance weighting under longer prefixes.

## 5.4 References to the Open-Source Code

Our experiments build on the open-source CORE implementation released by the authors of CORE [4]. We implement CORE-DA, CORE-HN, and CORE-DAHN as modular extensions to the CORE-trm codepath.

## 5.5 Reproducibility: Reported vs. Reproduced CORE Baselines

To satisfy the reproduction requirement, we compare our reproduced results for CORE-ave and CORE-trm against the numbers reported in the CORE paper under the same benchmarks and protocol [4]. Table 2 shows that our implementation closely matches the reported performance, with small deviations likely due to implementation details and randomness.

Table 2: Reproducibility check: reported CORE results [4] vs. our reproduced results (Recall@20 / MRR@20, %).

Dataset	CORE-ave (rep.)	CORE-ave (ours)	CORE-trm (rep.)	CORE-trm (ours)
Diginetica	50.21 / 18.07	50.28 / 17.99	52.89 / 18.58	52.87 / 18.66
Nowplaying	20.31 / 6.62	20.24 / 6.64	21.81 / 7.35	21.73 / 7.45
RetailRocket	59.18 / 37.52	58.58 / 37.18	61.85 / 38.76	61.80 / 38.59
Tmall	44.67 / 31.85	45.02 / 31.62	44.48 / 31.72	44.99 / 31.63
Yoochoose	58.83 / 25.05	59.02 / 25.07	64.61 / 28.24	64.58 / 28.21

## 5.6 Performance Evaluation Metrics

We focus on top- $K$  ranking metrics standard in session-based recommendation. Because next-item prediction has a single ground-truth item per test session, Recall@K coincides with Hit Rate@K: it measures the fraction of test sessions for which the target item appears in the top- $K$  recommendation list. We also report MRR@K (mean reciprocal rank), which additionally rewards placing the target item higher in the ranked list [3]. Following CORE, our primary metrics are Recall@20 and MRR@20 [4].

Formally, for a test set of sessions  $\mathcal{S}$  with ground-truth next items  $\{v_s^+\}$  and rank  $\text{rank}_s(v_s^+)$  in the predicted list, we compute:

$$\text{Recall@K} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbf{1}[\text{rank}_s(v_s^+) \leq K], \quad (15)$$

$$\text{MRR@K} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbf{1}[\text{rank}_s(v_s^+) \leq K] \cdot \frac{1}{\text{rank}_s(v_s^+)}. \quad (16)$$

## 5.7 Benchmarking Methods and Data

We compare to CORE-ave and CORE-trm [4]. Our aim is not to re-argue the full landscape, but to position the proposed improvements relative to CORE and to the strong baselines already considered in CORE.

## 5.8 Experimental Setup

All configurations are trained under a fixed, reproducible setup with a single random seed (`seed` = 2020). Unless otherwise noted, we use 20 epochs with early stopping (`stopping_step` = 5), batch size 2048, and Adam with learning rate  $10^{-3}$ . For CORE-HN and CORE-DAHN, we enable hard negatives with `hard_neg_k` = 5 and mixing coefficient  $\lambda$  = 0.2. We evaluate once on the held-out test split using Recall@20 and MRR@20.

For the additional eShop2008 dataset, we additionally sweep the item-dropout ratio  $\rho \in \{0.15, 0.20, 0.25, 0.30\}$  and select the setting per architecture based on validation MRR@20, then report the corresponding test results. This follows the same motivation as our earlier dropout sensitivity checks, namely that  $\rho$  can affect generalization in the robust distance measuring component.

## 5.9 Statistical Tests for Model Evaluation

This draft reports single-run results (one seed per configuration). A complete statistical analysis (mean  $\pm$  std across multiple seeds, paired tests, and/or confidence intervals) is left to future work, and is particularly important because the observed differences are often within a few tenths of a percentage point.

## 5.10 Ablation Study

We isolate each contribution by comparing:

- (i) CORE-DA against CORE-TRM to measure the effect of the dual-attention encoder.
- (ii) CORE-HN against CORE-TRM to measure the effect of hard-negative-aware training.
- (iii) CORE-DAHN to test whether the two improvements are complementary.

## 6 Results

Table 3 reports single-run test results across five CORE benchmarks. We report Recall@20 and MRR@20 as percentages. Best MRR@20 per dataset is highlighted.

Table 3: Main results on CORE benchmarks (single run with `seed` = 2020). Metrics are Recall@20 / MRR@20 (%). Best MRR@20 per dataset is bolded.

Dataset	CORE-ave	CORE-trm	CORE-DA	CORE-HN	CORE-DAHN
Diginetica	50.28 / 17.99	52.87 / 18.66	52.90 / 18.66	52.52 / 18.46	52.67 / <b>18.69</b>
Nowplaying	20.24 / 6.64	21.73 / 7.45	21.77 / <b>7.67</b>	21.70 / 7.43	21.46 / 7.64
RetailRocket	58.58 / 37.18	61.80 / 38.59	61.95 / 38.80	61.82 / 38.69	61.83 / <b>38.85</b>
Tmall	45.02 / 31.62	44.99 / 31.63	45.01 / 31.63	44.92 / <b>31.76</b>	44.92 / 31.75
Yoochoose	59.02 / 25.07	64.58 / 28.21	64.60 / 28.29	64.56 / <b>28.31</b>	64.43 / 28.19
eShop2008	—	67.84 / <b>18.89</b>	67.61 / 18.82	67.83 / 18.82	67.81 / 18.78

**Overall baselines.** Across all datasets, CORE-trm substantially improves upon CORE-ave, consistent with the role of attention-based importance weighting in CORE’s representation-consistent space. The gap is particularly pronounced on larger benchmarks such as Yoochoose and RetailRocket, where order-aware weighting improves both Recall@20 and MRR@20.

**Effect of Dual-Attention (CORE-DA).** Dual attention yields small but generally positive rank-quality gains (MRR@20) over CORE-trm on four out of five datasets (Nowplaying, RetailRocket, Tmall, Yoochoose), with the largest observed increase on Nowplaying and RetailRocket (about +0.2 absolute MRR points). On Diginetica, CORE-DA matches CORE-trm on MRR@20 and slightly improves Recall@20.

**Effect of Hard Negatives (CORE-HN).** Hard-negative-aware training exhibits a dataset-dependent trade-off. It improves MRR@20 on Tmall and Yoochoose (and slightly on RetailRocket), but can reduce MRR@20 on Diginetica and Nowplaying. In several cases, hard negatives slightly decrease Recall@20 while sharpening the ranking of the correct item when it remains in the top- $K$  list, consistent with the intuition that focusing on hard confounders can reallocate probability mass away from easy negatives.

**Combined model (CORE-DAHN).** The combined model achieves the best MRR@20 on Diginetica and RetailRocket, indicating that the two mechanisms can be complementary in some regimes. However, complementarity is not universal: on Nowplaying, CORE-DA remains best; on Tmall and Yoochoose, CORE-HN remains best in MRR@20, while CORE-DAHN does not consistently add further gains and can slightly reduce Recall@20. This suggests that the interaction between the dual-attention encoder and hard-negative mining is sensitive to dataset characteristics and hyperparameter settings (e.g.,  $\lambda$  and  $K$ ).

**Additional dataset (eShop2008).** On eShop2008, CORE-trm achieves the best MRR@20 (18.89), while CORE-DA and CORE-HN are close (18.82) and CORE-DAHN is slightly lower (18.78). This indicates that, on this dataset, the CORE-trm weighting mechanism already provides a strong bias for ranking quality, and that the interaction between dual-attention fusion and hard-negative mining may require dataset-specific tuning (e.g.,  $\lambda$ ,  $K$ , or dropout  $\rho$ ) to translate into gains.

## 7 Discussion

Our improvements were designed to strengthen CORE-trm along two orthogonal axes while preserving CORE’s geometric principle.

**Dual attention improves rank quality with minimal disruption.** CORE-DA tends to increase MRR@20 while leaving Recall@20 nearly unchanged relative to CORE-trm. This

pattern is consistent with the intended behavior: by explicitly modeling global session context and recent intent and adapting their fusion via a gate, the model can place the true next item higher in the ranked list without fundamentally changing the candidate set recalled into the top- $K$ .

**Hard negatives sharpen decisions but can hurt recall.** CORE-HN often improves MRR@20 (notably on Tmall and Yoochoose) while slightly reducing Recall@20. This is consistent with the training signal: mining top-scoring incorrect items and penalizing them more strongly can improve separation from confounders, but may also make the distribution more peaked around a narrower set of candidates and thus slightly harm top- $K$  coverage.

**Why the combined model is not always additive.** CORE-DAHN wins on Diginetica and RetailRocket in MRR@20, but does not dominate elsewhere. A plausible explanation is that hard-negative mining is performed using the current model’s scoring function; when the encoder changes (dual attention + gating + residual), the set of mined negatives and the gradients induced by the hybrid objective can interact with the gating dynamics. If the gate overreacts to mined confounders, it may bias the fusion toward representations that are locally good for separating a few hard negatives but suboptimal for global top- $K$  coverage, especially in datasets with highly popular items or frequent intent shifts. This motivates future tuning and analyses of  $\lambda$ ,  $K$ , and gate behavior per dataset.

Finally, the additional eShop2008 experiment shows that our extensions are not universally improving: CORE-trm remains best there, suggesting sensitivity to dataset properties and motivating dataset-adaptive tuning of the hybrid-loss weight and the gating behavior.

## 8 Conclusions and Future Work

We revisited CORE through the lens of its original motivation: ensuring representation consistency between session encoding and item decoding. Building on CORE-trm, we introduced two improvements that preserve this principle. First, we proposed a Dual-Attention mechanism with adaptive gating that fuses global context with explicit recent-intent emphasis while keeping the session embedding as a linear combination of item embeddings. Second, we proposed a hard-negative-aware hybrid objective that strengthens learning against confusing items by combining cross-entropy with a pairwise ranking term.

Empirically, Dual-Attention (CORE-DA) yields consistent, modest improvements in rank quality (MRR@20) across most benchmarks while maintaining similar Recall@20 to CORE-trm. Hard negatives (CORE-HN) provide dataset-dependent benefits, improving MRR@20 on Tmall and Yoochoose but slightly harming some datasets. The combined approach (CORE-DAHN) achieves the best MRR@20 on Diginetica and RetailRocket, but does not universally dominate, suggesting that complementarity depends on dataset characteristics and hyperparameters.

The additional eShop2008 evaluation further highlights that these gains are not universal: on eShop2008, the strongest results remain with CORE-trm, indicating that the benefits of our extensions can be dataset-dependent.

### 8.1 Limitations

This article reports single-run results with a fixed random seed and therefore does not establish statistical significance. Moreover, hard negative mining introduces additional computation due to top- $K$  selection; practical deployments on very large catalogs may require approximate retrieval or cached candidate sets.

**Hard negatives and false negatives.** Hard-negative mining may occasionally select unobserved yet relevant items as negatives (false negatives). In this work we follow the standard implicit-feedback assumption and only exclude the ground-truth next item; exploring additional filtering or debiasing strategies is left for future work.

## 8.2 Future work

Two directions appear especially promising. First, we will run multiple seeds per configuration and report mean $\pm$ std with statistical tests to verify that the observed gains are robust. Second, we will explore dataset-adaptive hard-negative settings (e.g., tuning  $\lambda$  and  $K$ , or scheduling  $\lambda$  over training) and analyze gate behavior to better understand when dual attention and hard-negative mining are complementary. Finally, we plan to explore multi-interest consistent-space variants and graph-informed weighting mechanisms that influence importance estimation while preserving CORE’s representation principle [4, 3].

## References

- [1] D. Jannach, M. Quadrana, and P. Cremonesi, “Session-based recommender systems,” in *Recommender Systems Handbook* (F. Ricci, L. Rokach, and B. Shapira, eds.), pp. 301–334, Springer US, 3 ed., 2022.
- [2] S. Wang, L. Cao, Y. Wang, Q. Z. Sheng, M. A. Orgun, and D. Lian, “A survey on session-based recommender systems,” *ACM Computing Surveys*, vol. 54, no. 7, 2021.
- [3] Z. Li, C. Yang, Y. Chen, X. Wang, H. Chen, G. Xu, L. Yao, and Q. Z. Sheng, “Graph and sequential neural networks in session-based recommendation: A survey,” *arXiv preprint arXiv:2408.14851*, 2024.
- [4] Y. Hou, B. Hu, Z. Zhang, and W. X. Zhao, “CORE: Simple and effective session-based recommendation within consistent representation space,” in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’22)*, pp. 1796–1801, ACM, 2022.
- [5] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, “BPR: Bayesian personalized ranking from implicit feedback,” in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pp. 452–461, AUAI Press, 2009.