Learning Semantic Co-occurrence and Personalized Focused Representations for Multimodal Sequential Recommendation

Anonymous Author(s)

Abstract

Multimodal sequential recommendation (MSR) leverages textual and visual content of items to capture the evolution of user interests over time. Prior efforts have largely concentrated on multimodal fusion strategies or downstream sequential modeling, while the question of how to obtain item representations genuinely aligned with the recommendation objective has received far less attention. Off-the-shelf encoders pretrained on generic corpora produce representations that emphasize broad semantic similarity rather than the specific cues that drive user choice in a recommendation scenario. In this work, we propose **ScopeRec**, a **S**emantic **co**-occurrence and **pe**rsonalized focus framework for multimodal sequential **Rec**ommendation, which reshapes multimodal representations from two complementary views. From the item view, items co-interacted within a behavior sequence carry rich sequence-level semantic associations, which we term *semantic co-occurrence*. ScopeRec mines such multi-level co-occurrence patterns directly from historical sequences and injects them back into the item representations, yielding embeddings that better reflect users' historical semantic preferences. From the user view, user attention typically concentrates on a few keywords in text or salient regions in images, which we term *semantic focus*. To exploit this property, ScopeRec employs a hierarchical mixture-of-experts (MoE) architecture in which different experts independently capture fine-grained interest from distinct semantic perspectives. A specialized fusion module then adaptively combines representations across modalities and semantic levels in a personalized manner. Experiments on three public MSR benchmarks show that ScopeRec consistently surpasses state-of-the-art baselines, with relative gains of 13.8% to 28.7% on NDCG and Hit Ratio.

CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

Multimodal Sequential Recommendation, Representation Learning, Semantic Co-occurrence, Semantic Focus

ACM Reference Format:

Anonymous Author(s). 2018. Learning Semantic Co-occurrence and Personalized Focused Representations for Multimodal Sequential Recommendation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Sequential recommendation (SR) has emerged as a core paradigm in modern recommender systems, where the goal is to predict the next item a user will interact with by modeling the chronological order of past behaviors [3, 28, 32]. Conventional SR methods are built upon item ID embeddings and have achieved promising results in capturing dynamic user preferences. However, ID-based representations are agnostic to the actual content of items and therefore fail to exploit the rich textual descriptions and visual cues that often serve as the most direct signals of user interest. This observation has motivated the rise of multimodal sequential recommendation (MSR) [16, 17], where textual and visual features are incorporated to enrich item representations and improve recommendation quality [4, 9, 11, 37].

A large body of MSR research has investigated how to better utilize multimodal information [10, 15, 21]. MM-Rec [29] adopts a pretrained VL-BERT as a unified multimodal encoder in a single-tower architecture. UniSRec [9] extracts textual features with pretrained language models and applies an MoE module to enable cross-domain transfer. MISSRec [27] discovers latent interest clusters and fuses modalities with learnable coefficients. More recently, HM4SR [37] explicitly models temporal dynamics and incorporates auxiliary supervisory signals during multimodal fusion. While these efforts have advanced the state of the art, they predominantly focus on the *fusion* or *adaptation* of multimodal features, leaving an important question largely unanswered: *how should item representations be learned so that they are intrinsically aligned with the recommendation task?*

Most existing approaches treat the output of a pretrained encoder as a fixed feature and feed it into the sequential model. Such representations, however, are optimized for generic semantic similarity rather than for distinguishing items that are likely to be co-interacted by a user. As a result, two issues arise. First, the representations capture broad semantic categories but fail to reflect the user-specific structures that emerge from interaction sequences, which limits their utility for interest evolution and long-tail recommendation. Second, the representations are aggregate in nature and treat all semantic components equally, so the cues that actually drive user choice are diluted by irrelevant or redundant content. Effective MSR representations should therefore jointly capture (i) the item-level relational structures revealed by sequential co-consumption, and (ii) the user-level focus patterns that concentrate on a small set of decisive semantic components.

Item view: semantic co-occurrence. Within a user behavior sequence, interacted items tend to be semantically related to their temporal neighbors [38, 39]. As illustrated in Figure 1, the word *pipe* recurs across consecutive items in the Scientific sequence, and

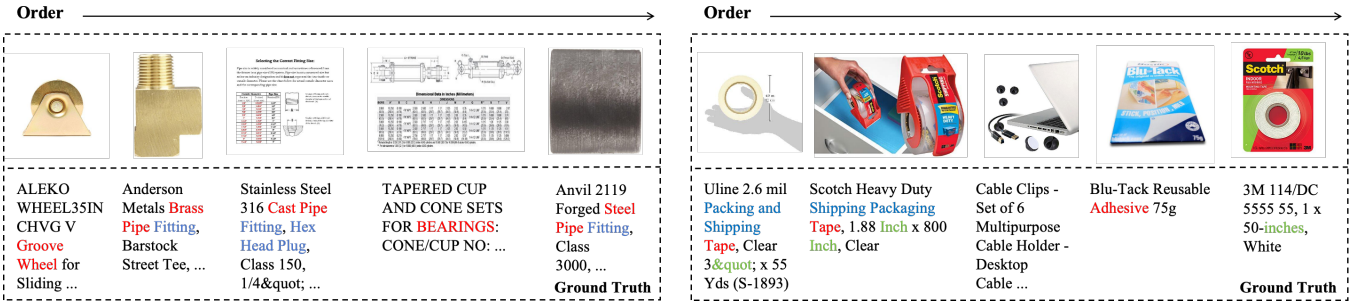


Figure 1: Two samples of user behavior sequences. The left shows a user sequence from the Scientific dataset; the right corresponds to the Office dataset. Words highlighted in red indicate semantic co-occurrence at the sequence level, while items share visual patterns that reflect localized user focus.

items with similar functional descriptions or visual appearance are clustered along the timeline. We refer to such sequence-level associations as **semantic co-occurrence**. Compared with ID-based co-occurrence used in classical collaborative filtering, semantic co-occurrence is intrinsically more transferable, since it abstracts away from specific item identifiers and captures interest patterns at the semantic level. Despite its potential, semantic co-occurrence has not been systematically modeled in existing MSR pipelines.

User view: semantic focus. A user's preference toward an item is rarely uniformly distributed over its content. Rather, it concentrates on a few keywords in text or localized regions in images, which we collectively term **semantic focus**. As shown in Figure 1, terms such as *tape*, *adhesive*, and *inch* dominate the textual modality, while the visual modality is anchored by recurring tape-like objects. Most existing MSR methods [9, 34] encode multimodal content in an aggregated form and do not differentiate between informative and noisy components. The resulting coarse-grained representation introduces noise, blurs modality-level and semantic-level preferences, and fails to characterize how a user's interest unfolds across different stages of behavior.

To address these issues, we propose **ScopeRec**, a unified MSR framework that learns multimodal representations aligned with the recommendation scenario along both views. From the item view, ScopeRec organizes multimodal content chronologically and applies an attention-based relation extraction module to capture sequence-level semantic co-occurrence, which is then injected back into the original representations. Instead of relying on the final encoder layer alone, ScopeRec collects hidden states from multiple intermediate layers, since different layers encode complementary semantic levels [19, 23]. From the user view, ScopeRec introduces a hierarchical MoE module in which multiple experts independently attend to different semantic regions of interest, suppressing irrelevant components. A second, modality- and level-specific MoE further fuses these representations through a gating mechanism that adapts to each user's preference distribution. Together, these designs produce expressive, hierarchical, and user-centered representations for MSR.

We summarize the main contributions as follows.

- We highlight the under-explored problem of learning multimodal representations that are intrinsically aligned with the recommendation task, and propose **ScopeRec**, a unified MSR framework built upon the complementary views of semantic co-occurrence and semantic focus.
- We are, to the best of our knowledge, the first to systematically study sequence-level semantic co-occurrence in MSR. ScopeRec captures such co-occurrence by reorganizing multimodal content along the interaction order and applying an attention-based relation extraction module under a causal mask.
- To characterize fine-grained interest, we propose a hierarchical MoE mechanism that adaptively focuses on the semantic components most relevant to user preferences at multiple semantic levels, effectively suppressing noise and producing more discriminative representations.
- Extensive experiments on three public benchmarks show that ScopeRec consistently outperforms state-of-the-art baselines on NDCG and Hit Ratio. Detailed analyses provide additional insight into how multimodal information can be more effectively exploited in recommendation.

2 Related Work

2.1 Sequential Recommendation

SR [3, 28] aims at modeling users' historical interaction sequences in order to predict their next item of interest. Early studies focused on developing effective sequence encoders. GRU4Rec [8, 25] pioneered the use of gated recurrent units, and NextItNet [35] introduced a convolutional architecture to capture sequential dependencies. With the rise of attention, SASRec [12] and BERT4Rec [24] pushed the field forward by leveraging unidirectional and bidirectional Transformer structures, respectively. Subsequent work incorporated auxiliary context to further enrich behavioral modeling. HyperRec [26] partitions user-item interactions into temporal hypergraphs and combines hypergraph convolution with self-attention to model multi-order correlations and short-term intents simultaneously. AutoMLP [13] integrates IDs with category information via an adaptive search mechanism to balance short-term and long-term preferences. NOVA [15] fuses side information non-invasively to avoid disrupting the ID embedding space. A common limitation across these approaches is that they remain ID-centric and overlook the rich multimodal content of items, which typically

provides more direct and interpretable signals of user interest, especially under cold-start or long-tail scenarios where ID statistics are unreliable.

2.2 Multimodal Sequential Recommendation

The rapid progress of pretrained foundation models [1, 2, 7, 18, 22] has enabled MSR to incorporate semantic features from text and images [30, 33]. CSAN [11] aggregates multimodal features and projects them into joint embeddings for improved item representation. MM-Rec [29] employs pretrained VL-BERT as a unified multimodal encoder in a single-tower SR architecture. UniSRec [9] adopts MoE for cross-domain semantic transfer between textual and ID embeddings. MISSRec [27] clusters user interests and fuses modalities through learnable coefficients. IISAN [4] introduces a decoupled parameter-efficient fine-tuning architecture for modality encoders, and IISAN-VS [5] extends it to asymmetric settings. TeDRec [34] performs sequence-level fusion of ID and text embeddings in the frequency domain via the Fast Fourier Transform, while PMMRec [14] relies on contrastive denoising to improve transferability. HM4SR [37] models temporal dynamics with auxiliary supervisory tasks. CAMMSR [31] utilizes additional classification tasks to assist the high-quality dynamic assignment of modal weights. From a methodological perspective, these works can be grouped into three streams: (i) replacing or augmenting ID embeddings with multimodal features, (ii) designing efficient adaptation strategies for large modality encoders, and (iii) introducing auxiliary objectives or fusion operators on top of frozen modality features.

Despite the progress along these three streams, prior work has paid limited attention to how multimodal representations themselves should be shaped to fit the recommendation task. Two issues remain. First, the sequence-level semantic associations among co-interacted items are largely ignored, even though they encode stable and transferable interest structures. Second, the fine-grained focus that drives user choice within each item is treated implicitly, leaving room for noise and dilution. ScopeRec is designed to address both gaps in a unified manner: it explicitly couples representation learning with the chronological structure of user behaviors, and it exposes the components of an item that actually drive user interest through a hierarchical mixture-of-experts. We view this as a step toward *recommendation-aware* representation learning that complements existing fusion- and adaptation-centric approaches.

2.3 Mixture-of-Experts for Recommendation

Mixture-of-experts has become a versatile building block in recommender systems. UniSRec [9] applies MoE to bridge textual features across domains, while MISSRec [27] uses cluster-style experts to capture distinct interest groups. HM4SR [37] adopts time-aware MoE to model temporal dynamics. ScopeRec departs from these designs in two ways. First, the modality-level MoE operates on representations *enhanced by sequence-level co-occurrence*, so that each expert can disentangle a particular kind of user-relevant focus instead of a generic semantic cluster. Second, the hierarchical fusion MoE is *input-independent*: each expert receives a distinct, modality- and level-specific input rather than the same vector, which prevents the experts from collapsing onto a shared signal.

3 Methodology

We propose ScopeRec, an MSR framework that explicitly mines semantic co-occurrence among items in a user's behavior sequence and enhances modality-aware semantic focus representations. ScopeRec models user interests purely from the textual and visual content of historical interactions, without resorting to user identity features. This design follows the natural evolution of user preferences and yields strong empirical performance, as detailed in Section 4. As shown in Figure 2, ScopeRec consists of four modules. The *semantic co-occurrence mining* module employs a unidirectional attention mechanism to capture cross-item associations and injects them into the initial multimodal representations. The *semantic focus awareness* module adopts a hierarchical MoE architecture to perform fine-grained semantic focusing at multiple semantic levels. The *hierarchical representation fusion* module adaptively assigns importance to representations across modalities and semantic levels through a gating mechanism. Finally, the *next-item prediction* module applies a standard SR backbone to the fused item representations.

3.1 Problem Formulation

We follow the standard setup of MSR. Let $\mathcal{U}$ denote the user set and $\mathcal{V}$ the item set. Each item $v \in \mathcal{V}$ is associated with a textual description txt_v and an image img_v. For a user $u \in \mathcal{U}$, the chronological interaction history is given by $S_u = (v_1, v_2, \dots, v_n)$, with $v_t \in \mathcal{V}$ being the item interacted with at the t-th step. The task is to predict the next item v_{n+1} that the user is most likely to interact with, given the historical sequence and the multimodal content of items. Unlike ID-based SR, no user or item identifier is used by ScopeRec; the prediction relies entirely on the multimodal content of items and the temporal order in which they appear.

3.2 Semantic Co-occurrence Mining

The textual and visual content of items conveys semantics that are more directly indicative of user preferences than ID embeddings. To capture such semantics at multiple granularities, we encode each modality with a pretrained encoder and retain both intermediate hidden states and the final hidden state. Prior studies [19, 23] show that different layers of a pretrained model encode information at different levels of abstraction, ranging from shallow patterns to deep semantics. Therefore, layer-wise hidden states provide complementary cues that benefit user interest modeling.

Formally, recall that $\mathcal{U}$ denotes the user set and $\mathcal{V}$ the item set. For each item $v_i \in \mathcal{V}$, we feed its textual and visual content into a text encoder Enc_T and an image encoder Enc_I:

$$\{h_T^{i(1)}, h_T^{i(2)}, \dots, h_T^{i(l)}\} = Enc_T(\text{txt}_i), \tag{1}$$

$$\{h_I^{i(1)}, h_I^{i(2)}, \dots, h_I^{i(l)}\} = Enc_I(\text{img}_i), \tag{2}$$

where $h_T^{i(l)} \in \mathbb{R}^{d^{txt}}$ and $h_I^{i(l)} \in \mathbb{R}^{d^{img}}$ are the hidden states of the l-th layer for item v_i. The choice of l depends on the depth of the

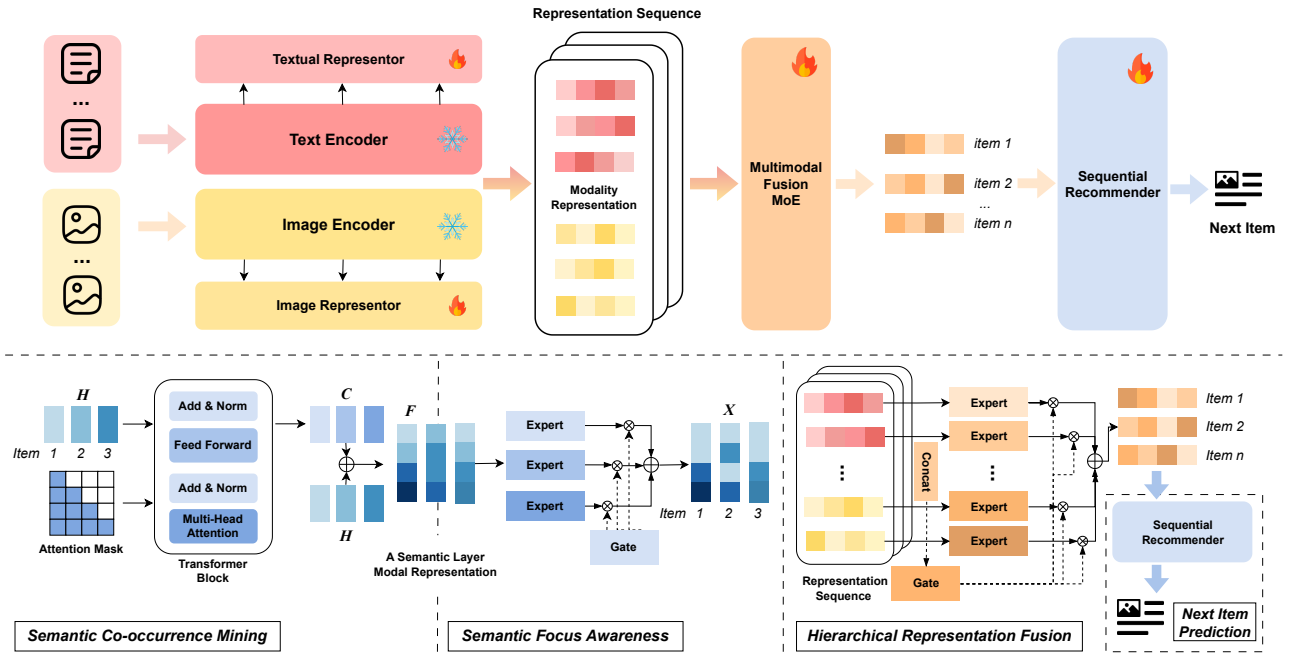


Figure 2: The overall framework of ScopeRec. The pipeline contains four modules: semantic co-occurrence mining, semantic focus awareness, hierarchical representation fusion, and next-item prediction.

backbone encoder and our layer selection strategy. We aggregate three representative layers: a shallow layer encoding basic semantic patterns, a deep layer encoding abstract semantics, and the final layer summarizing the overall item characteristics.

Capturing semantic co-occurrence. Items consumed by the same user in a short time window tend to be semantically related, which is well documented in prior empirical studies of recommendation logs. We refer to such sequence-level semantic associations as *semantic co-occurrence*. Existing MSR methods rarely model this property explicitly, leaving the persistent component of user interest underutilized. We propose to leverage the temporal order of user behavior to expose such associations.

The intuition is best understood by contrast with ID-based co-occurrence. Classical collaborative filtering counts how often two item IDs appear together in user histories and turns those counts into a similarity signal. Such counts are item-specific and become unreliable as soon as the catalog grows or items are introduced. Semantic co-occurrence, instead, operates on the multimodal representation of items, so a co-occurrence pattern observed between two specific items generalizes to any other pair of items with similar text or images. This abstraction is the reason that, as we will show, ScopeRec transfers more gracefully to data-sparse settings than ID-centric baselines.

For a user $u \in \mathcal{U}$, let the interaction sequence be denoted as $S_u = \{v_1, v_2, \dots, v_n\}$. Based on S_u, we organize layer-wise modality representations into temporally ordered sequences as follows:

$$H_{txt}^{(l)} = \{h_T^{1(l)}, h_T^{2(l)}, \dots, h_T^{n(l)}\}, \quad H_{img}^{(l)} = \{h_I^{1(l)}, h_I^{2(l)}, \dots, h_I^{n(l)}\}. \tag{3}$$

We then apply a multi-head self-attention module to mine cross-item dependencies along the sequence. Since future behaviors are unknown at prediction time, we use a causal (lower-triangular) attention mask $Mask$ to prevent information leakage. As illustrated in Figure 2, the white region of the mask blocks attention to future positions, while the blue region allows access to past items only.

We denote the semantic co-occurrence representations as $C^{(l)} = \{c^{1(l)}, \dots, c^{n(l)}\}$. They are produced by a Transformer block that operates on the modality sequence:

$$\begin{aligned} Z^{(l)} &= \mathrm{LN}(H^{(l)} + \mathrm{Attn}(H^{(l)}, Mask)), \\ C^{(l)} &= \mathrm{LN}(Z^{(l)} + \mathrm{FFN}(Z^{(l)})), \end{aligned} \tag{4}$$

where Attn denotes the masked multi-head self-attention layer, LN denotes layer normalization, and FFN denotes a position-wise feed-forward network. Here $H^{(l)}$ is a generic notation that covers both $H_{txt}^{(l)}$ and $H_{img}^{(l)}$.

We then inject the co-occurrence representations back into the original modality representations with a tunable mixing coefficient λ:

$$F^{(l)} = C^{(l)} \cdot \lambda + H^{(l)}. \tag{5}$$

The mixture preserves the intrinsic semantics of items while enhancing their associations with the historical context, producing richer and more sequence-aware item embeddings.

3.3 Semantic Focus Awareness

Item representations enhanced by semantic co-occurrence still encode a wide range of content, much of which is irrelevant to user choice. To extract the components that align with user interest, we

introduce a semantic focus awareness module built on MoE. Each expert in this module acts as an *interest expert* that attends to a distinct semantic region, while a gating network dynamically weighs the experts according to the input representation. This design selectively preserves informative components and suppresses noise.

A further consideration is that hidden states at different layers carry different semantic granularities. Relying solely on the final layer, as most existing methods do, can introduce semantic bias and limit the model's expressiveness. We therefore deploy dedicated MoE modules for each modality and each retained semantic layer.

Concretely, the multi-level representations $F^{(l)}$ enhanced by semantic co-occurrence are routed through the corresponding MoE:

$$x^{(l)} = \sum_{n=1}^{N} g_n(F^{(l)}) \cdot E_n(F^{(l)}), \tag{6}$$

where $x^{(l)}$ is the focus-aware representation, $E_n(\cdot)$ is the n-th interest expert, and $g_n(\cdot)$ is the gating weight assigned to that expert. The gating network produces a softmax-normalized distribution over experts:

$$\mathbf{g}(F^{(l)}) = \text{Softmax}(W^{mg}F^{(l)} + b^{mg}), \quad \mathbf{g} = [g_1, g_2, \dots, g_N]^\top, \tag{7}$$

with learnable parameters $W^{mg} \in \mathbb{R}^{N\times d}$ and $b^{mg} \in \mathbb{R}^N$. Each expert is implemented as a feed-forward layer with a nonlinearity $\phi(\cdot)$ (e.g., ReLU or GELU):

$$E_n(F^{(l)}) = \phi(W_n^{mexp}F^{(l)} + b_n^{mexp}), \tag{8}$$

where $W_n^{mexp} \in \mathbb{R}^{d'\times d}$ and $b_n^{mexp} \in \mathbb{R}^{d'}$ are expert parameters; d is the backbone hidden size and d' is the embedding size of the downstream recommender.

3.4 Hierarchical Representation Fusion

After semantic focus awareness, ScopeRec produces a set of hierarchical representations, one per modality and per semantic level. To translate this set into a single user-centered embedding, we propose an *input-independent* MoE fusion module. Unlike a conventional MoE in which all experts share the same input, our fusion experts receive distinct, modality- and level-specific inputs. This isolates the learning of each semantic source and reduces cross-source interference.

A shared gating network determines how much each source contributes. The gate operates on a concatenation x_{concat} of all sources to leverage their joint context, and its output is a per-expert weight vector. This allows the model to flexibly emphasize modalities and semantic levels that are most informative for the current user.

The fusion is formulated as:

$$m = \sum_{l=1}^{L} g'_l(x_{\text{concat}}) \cdot E_l(x^{(l)}), \tag{9}$$

$$\mathbf{g}'(x_{\text{concat}}) = \text{Softmax}(W^{fg}x_{\text{concat}} + b^{fg}), \tag{10}$$

$$E_l(x^{(l)}) = \phi(W_l^{fexp}x^{(l)} + b_l^{fexp}), \tag{11}$$

where $\mathbf{g}' = [g'_1, g'_2, \dots, g'_L]^\top$, $W_l^{fexp} \in \mathbb{R}^{d'\times d'}$ and $b_l^{fexp} \in \mathbb{R}^{d'}$ are expert parameters, L is the total number of semantic sources from both modalities, and $W^{fg} \in \mathbb{R}^{L\times D}$ and $b^{fg} \in \mathbb{R}^L$ parameterize the gate. Here, $D = Ld'$. The vector m serves as the final multimodal item representation.

3.5 Next-Item Prediction

We adopt conventional SR models as the prediction module, replacing the original ID embedding sequence with the multimodal behavior representations enhanced by semantic collaborative signals. Specifically, given the multimodal embedding sequence of user u's interaction history $\mathbf{E}_u = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n]$, the SR model encodes the sequence to obtain the user behavior representation at each position t, denoted as $\widehat{\mathbf{h}}_{u,t}$. The preference score for a candidate item v at position t is computed via inner product:

$$s_{u,t,v} = \widehat{\mathbf{h}}_{u,t}^\top \mathbf{e}_v \tag{12}$$

During training, we adopt in-batch negative sampling, where items appearing in the current mini-batch serve as negative candidates. To mitigate popularity bias, we apply a logit-adjusted correction by subtracting the log-popularity of each candidate item from its raw score. Additionally, items that have already appeared in the user's historical interaction sequence are masked out from the negative candidate set to avoid false negatives. The adjusted preference score is defined as:

$$C_{u,t,v} = s_{u,t,v} - \log p_v \tag{13}$$

where p_v denotes the empirical popularity (sampling probability) of item v in the training corpus.

For each prediction position t, the model treats the target item $v^*_{u,t}$ (i.e., the $(t+1)$-th item in the interaction sequence) as the positive sample, and the remaining in-batch items excluding user u's historical items as negatives. The training objective is formulated as a multi-class cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{u\in\mathcal{U}}\sum_{t=1}^{n} \mathbb{1}[m_{u,t}=1] \log P(v^*_{u,t} \mid u,t) \tag{14}$$

$$P(v^*_{u,t} \mid u,t) = \frac{\exp\left(C_{u,t,v^*_{u,t}}\right)}{\exp\left(C_{u,t,v^*_{u,t}}\right) + \sum_{j\in\mathcal{B}\setminus\mathcal{I}_u} \exp\left(C_{u,t,j}\right)} \tag{15}$$

where $P(v^*_{u,t} \mid u,t)$ is the predicted probability of the ground-truth item $v^*_{u,t}$ over the in-batch candidate set, $\mathcal{B}$ denotes the item set present in the current training batch, $\mathcal{I}_u$ denotes the set of all items interacted with by user u (used to exclude false negatives from the negative candidate pool), and $m_{u,t} \in \{0,1\}$ is the padding mask that excludes padded positions from the loss computation.

3.6 Testing Time Inference

During the testing phase, the inference procedure is decomposed into two stages: item embedding precomputation and user preference scoring.

Item Embedding Precomputation. To avoid redundant computation at query time, we precompute a static embedding for each item in the catalog offline. Specifically, for each item $v_i \in \mathcal{V}$, its textual and visual content are first processed by the multimodal

encoder Enc_T and Enc_I, respectively, extracting layer-wise hidden states. These hidden states are then passed through the Semantic Co-occurrence module, which applies masked multi-head self-attention over all interaction sequences in which v_i participates, enriching the item representation with sequence-level semantic association signals. The resulting contextualized representations are subsequently fused by the hierarchical representation fusion module to yield a unified embedding $\mathbf{e}_{v_i} \in \mathbb{R}^{d'}$. Since a given item v_i may appear in multiple users' interaction sequences, its semantic co-occurrence representation varies across different sequence contexts, as the masked self-attention module captures distinct neighborhood associations depending on the surrounding items. Rather than discarding this contextual diversity, we aggregate all context-specific embeddings via averaging, which effectively summarizes the item's semantic associations across the full spectrum of observed co-occurrence patterns. This yields a more comprehensive and robust item embedding that reflects not only the item's intrinsic multimodal content but also its generalizable relational semantics across the user population:

$$\mathbf{e}_{v_i} = \frac{1}{|\mathcal{S}_{v_i}|} \sum_{k \in \mathcal{S}_{v_i}} \mathbf{e}_{v_i}^{(k)}, \tag{16}$$

where $\mathcal{S}_{v_i}$ denotes the set of sequence contexts in which item v_i appears, and $\mathbf{e}_{v_i}^{(k)}$ is the embedding produced from the k-th such context. All precomputed embeddings are stored in an item embedding table $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d'}$ for inference-time retrieval.

User Preference Scoring. At query time, given a target user u with interaction history $\mathcal{I}_u = \{v_1, v_2, \dots, v_n\}$, we retrieve the precomputed embeddings of the historical items and feed them as an embedding sequence into the sequential recommendation model (e.g., SASRec). The user behavior representation is extracted from the final sequence position:

$$\hat{\mathbf{h}}_u = \text{UserEncoder}(\mathbf{e}_{v_1}, \mathbf{e}_{v_2}, \dots, \mathbf{e}_{v_n})[-1], \tag{17}$$

where $[\cdot]_{-1}$ denotes taking the output at the last position of the encoder. The preference score for each candidate item $v_j \in \mathcal{V} \setminus \mathcal{I}_u$ is then computed as the inner product between the user representation and the precomputed item embedding:

$$\hat{y}_{u,v_j} = \hat{\mathbf{h}}_u^\top \mathbf{e}_{v_j}. \tag{18}$$

Items belonging to the user's historical interaction set $\mathcal{I}_u$ are excluded from ranking by setting their scores to $-\infty$. Note that the popularity-based logit adjustment applied during training is not used at inference time, as it serves solely as a debiasing mechanism to correct for sampling bias in the in-batch negative sampling objective. The top-K items with the highest preference scores are returned as the final recommendation list:

$$\hat{\mathcal{R}}_u = \underset{v_j \in \mathcal{V} \setminus \mathcal{I}_u}{\arg\,\text{top-}K}\, \hat{y}_{u,v_j}, \tag{19}$$

where $\hat{\mathcal{R}}_u$ denotes the final top-K recommendation list for user u, $\mathcal{V}$ is the full item catalog, $\mathcal{I}_u$ is the set of items previously interacted with by user u, $\hat{y}_{u,v_j}$ is the predicted preference score of user u for candidate item v_j, and K is the predefined size of the recommendation list.

Algorithm 1 Training procedure of ScopeRec.

Require: User set $\mathcal{U}$, item set $\mathcal{V}$ with text txt_v and image img_v; user interaction histories $\{S_u\}_{u \in \mathcal{U}}$; frozen encoders Enc_T, Enc_I; selected layer indices $\mathcal{L}$; co-occurrence coefficient λ; number of experts N; learning rate η; number of epochs T.

Ensure: Trained parameters Θ of the co-occurrence module, modality MoEs, fusion MoE, and SR backbone.

// Offline multimodal feature extraction

1: **for** each item $v_i \in \mathcal{V}$ **do**
2: $\quad \{h_T^{i(l)}\}_{l \in \mathcal{L}} \leftarrow Enc_T(\text{txt}_{v_i})$
3: $\quad \{h_I^{i(l)}\}_{l \in \mathcal{L}} \leftarrow Enc_I(\text{img}_{v_i})$
4: **end for**

// Online training

5: Initialize Θ randomly
6: **for** epoch $= 1, 2, \dots, T$ **do**
7: $\quad$ **for** each mini-batch $\mathcal{B} \subseteq \mathcal{U}$ of user sequences **do**
8: $\quad\quad$ **for** each user $u \in \mathcal{B}$ with $S_u = (v_1, \dots, v_n)$ **do**
9: $\quad\quad\quad$ **for** each layer $l \in \mathcal{L}$ and modality $\star \in \{T, I\}$ **do**
10: $\quad\quad\quad\quad$ Build $H_\star^{(l)} \leftarrow \{h_\star^{1(l)}, \dots, h_\star^{n(l)}\}$ ▷ Eq. (3)
11: $\quad\quad\quad\quad$ $C_\star^{(l)} \leftarrow \text{Transformer}(H_\star^{(l)}, Mask)$ ▷ Eq. (4)
12: $\quad\quad\quad\quad$ $F_\star^{(l)} \leftarrow \lambda \cdot C_\star^{(l)} + H_\star^{(l)}$ ▷ Eq. (5)
13: $\quad\quad\quad\quad$ $x_\star^{(l)} \leftarrow \sum_{n=1}^{N} g_n(F_\star^{(l)}) \cdot E_n(F_\star^{(l)})$ ▷ Eq. (6)
14: $\quad\quad\quad$ **end for**
15: $\quad\quad\quad$ Fuse hierarchical sources: $m \leftarrow \sum_l g'_l(x_{\text{concat}}) \cdot E_l(x^{(l)})$ ▷ Eq. (9)
16: $\quad\quad\quad$ Form embedding sequence $\mathbf{E}_u = [\mathbf{e}_1, \dots, \mathbf{e}_n]$ from $\{m\}$
17: $\quad\quad\quad$ $\hat{\mathbf{h}}_{u,t} \leftarrow \text{SRBackbone}(\mathbf{E}_u)$ for each position t
18: $\quad\quad$ **end for**
19: $\quad\quad$ Compute popularity-adjusted scores $C_{u,t,v}$ ▷ Eq. (13)
20: $\quad\quad$ Compute cross-entropy loss $\mathcal{L}_{\text{CE}}$ ▷ Eq. (14)–(15)
21: $\quad\quad$ $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}_{\text{CE}}$
22: $\quad$ **end for**
23: **end for**
24: **return** Θ

3.7 Overall Training Procedure

We summarize the complete training procedure of ScopeRec in Algorithm 1. The pipeline can be conceptually divided into two phases. The *offline encoding* phase (lines 1–4) extracts layer-wise multimodal hidden states from the frozen text and image encoders for every item in the catalog. Because the encoders are frozen, this phase is executed only once and the resulting features are cached for all subsequent epochs. The *online learning* phase (lines 5–24) trains the proposed semantic co-occurrence mining, semantic focus awareness, and hierarchical fusion modules together with the downstream sequential recommender on top of the cached features. For each mini-batch of user sequences, ScopeRec first arranges the cached layer-wise representations into temporally ordered tensors, then applies a causally masked Transformer block to mine sequence-level co-occurrence (lines 10–12), routes the enhanced representations through modality- and level-specific MoE

Table 1: Statistics of the three benchmark datasets.

Dataset	Users	Items	Interactions
Scientific	12,076	20,314	81,711
Instrument	10,000	19,246	75,022
Office	10,000	22,785	71,276

experts to perform semantic focus awareness (line 13), fuses the resulting hierarchical representations via the input-independent fusion MoE (line 15), and finally feeds the fused multimodal embedding sequence to the SR backbone for next-item prediction (lines 16–17). The training objective is the popularity-debiased cross-entropy loss defined in Eq. (14)–(15). Since gradients do not flow through the frozen encoders, the per-iteration cost is dominated by the lightweight upper modules, which is the key reason ScopeRec attains both strong accuracy and competitive efficiency, as we will quantify in Section 4.5.

4 Experiments

In this section, we report empirical results that answer the following research questions:

- **RQ1**: Does ScopeRec outperform state-of-the-art SR and MSR baselines on standard benchmarks?
- **RQ2**: How does each proposed component contribute to the final performance?
- **RQ3**: How sensitive is ScopeRec to the co-occurrence injection coefficient λ?
- **RQ4**: How efficient is ScopeRec in terms of training time and memory consumption?
- **RQ5**: What do the learned representations look like, and do they reflect the intended notions of semantic co-occurrence and semantic focus?

4.1 Experimental Setup

4.1.1 Datasets. We follow prior work [4, 6, 36] and use three categories from the Amazon review corpus [20]: *Industrial and Scientific*, *Musical Instruments*, and *Office Products*. These categories provide both textual descriptions and product images required for the MSR task. We deduplicate interactions and sort items chronologically per user. Dataset statistics are reported in Table 1.

4.1.2 Baselines. We compare against representative ID-based SR models (GRU4Rec [8, 25] and SASRec [12]) and a comprehensive set of state-of-the-art MSR models, including NOVA [15], UniSRec [9], MISSRec [27], IISAN [4], TeDRec [34], IISAN-VS [5], PMMRec [14], and HM4SR [37].

4.1.3 Evaluation Metrics. We adopt two standard top-K ranking metrics: Hit Ratio (HR@K) and Normalized Discounted Cumulative Gain (NDCG@K), with $K \in \{5, 10\}$.

4.1.4 Implementation Details. We adopt the text and vision branches of CLIP [22] (specifically, the `clip-vit-large-patch14` checkpoint) as the modality encoders. The hidden size shared by all MoE modules and the SR backbone is selected from $\{256, 512, 768, 1024, 2048\}$, with 1024 yielding the best validation performance. The number of experts per modality MoE is searched within $\{2, 3, 4\}$ and fixed at 3. The learning rate is tuned in $[10^{-5}, 10^{-3}]$ for the modality MoEs, the fusion MoE, and the SR backbone, with 10^{-5} shared across all components giving the best results. Following standard MSR protocol, we set the batch size to 72, the maximum sequence length to 10, the number of Transformer blocks in the backbone to 2, and the number of attention heads to 4. We use Adam as the optimizer. All experiments are run on a single NVIDIA A800 GPU.

4.2 Overall Performance (RQ1)

Table 2 reports the results on all three benchmarks. ScopeRec consistently delivers the best score on every metric, improving over the strongest baseline by 13.8% to 28.7%. We observe several patterns worth highlighting.

ScopeRec offers stable gains across datasets. On Instrument, the improvements over the second-best baseline are 15.0% (NDCG@5), 13.8% (NDCG@10), 17.1% (HR@5), and 18.7% (HR@10). This confirms that the proposed model captures user-item dependencies well under complex multimodal settings. Compared with parameter-efficient methods that also avoid ID features (e.g., IISAN and IISAN-VS), ScopeRec yields gains exceeding 27.5%, indicating that semantic co-occurrence mining and semantic focus enhancement are complementary to the recent trend of parameter-efficient adaptation.

Larger gains arise when the modality space is rich. The most significant relative gains are observed on the Office dataset, with up to 28.7% improvement on HR@10 over PMMRec. The Office category contains a larger and more diverse item set, which exposes more co-occurrence patterns and richer cross-item semantics. ScopeRec is able to leverage this richness more effectively than alternatives, suggesting that its representational design scales gracefully with data diversity.

Gains hold under data sparsity. On the Scientific dataset, which has the largest interaction-to-item ratio of the three, ScopeRec still improves over the second-best baseline by 14.4% to 21.1%. The consistent gain in this sparse regime indicates that the model's hierarchical representations and explicit co-occurrence modeling alleviate the cold-start tendency of pure ID-based or coarse multimodal models.

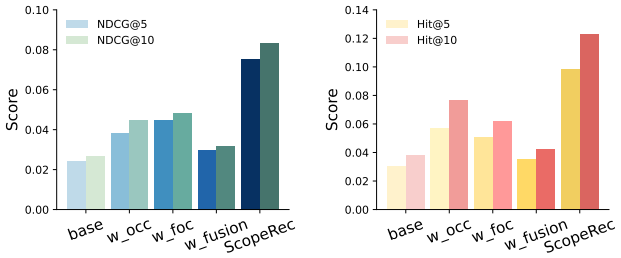
Multimodal models dominate ID-based models. Across the table, multimodal methods clearly outperform GRU4Rec and SASRec, reaffirming the value of multimodal content for capturing the evolution of user preferences. ScopeRec further closes the gap by directly addressing how multimodal representations are learned, rather than only how they are fused.

4.3 Ablation Study (RQ2)

To examine the contribution of each module, we construct four variants of ScopeRec. The *base* variant removes all proposed components, replaces the modality MoE by a plain DNN layer, and replaces the fusion MoE by a simple concatenation. The *w_occ* variant adds the semantic co-occurrence mining module on top of the base. The *w_foc* variant adds only the semantic focus awareness

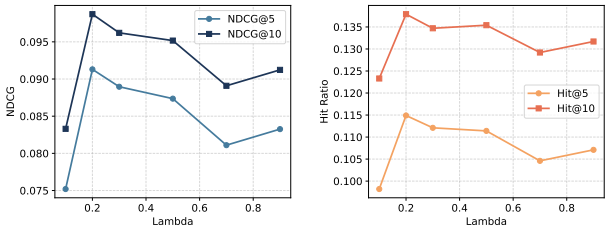
Table 2: Overall performance comparison on the Instrument, Office, and Scientific datasets. Bold marks the best result and underline marks the second best. "Improv." is the relative gain of ScopeRec over the second-best baseline.

Dataset	Metric	GRU4Rec	SASRec	NOVA	UniSRec	MISSRec	IISAN	TeDRec	IISAN-VS	PMMRec	HM4SR	Ours	Improv.
Instrument	NDCG@5	0.0398	0.0590	0.0612	0.0571	0.0645	0.0647	0.0586	0.0659	0.0741	0.0735	**0.0852**	15.0%
	NDCG@10	0.0435	0.0627	0.0654	0.0629	0.0697	0.0701	0.0658	0.0717	0.0795	0.0803	**0.0914**	13.8%
	HR@5	0.0475	0.0656	0.0698	0.0667	0.0762	0.0745	0.0680	0.0813	0.0886	0.0894	**0.1047**	17.1%
	HR@10	0.0591	0.0771	0.0847	0.0782	0.0915	0.0906	0.0813	0.0952	0.1032	0.1044	**0.1240**	18.7%
Office	NDCG@5	0.0082	0.0432	0.0413	0.0382	0.0455	0.0467	0.0403	0.0476	0.0589	0.0581	**0.0752**	27.7%
	NDCG@10	0.0092	0.0455	0.0429	0.0406	0.0483	0.0492	0.0432	0.0511	0.0677	0.0639	**0.0833**	23.0%
	HR@5	0.0115	0.0502	0.0498	0.0491	0.0548	0.0581	0.0482	0.0645	0.0822	0.0786	**0.0982**	19.5%
	HR@10	0.0144	0.0576	0.0565	0.0559	0.0613	0.0680	0.0542	0.0722	0.0958	0.0872	**0.1233**	28.7%
Scientific	NDCG@5	0.0135	0.0273	0.0302	0.0291	0.0337	0.0350	0.0328	0.0378	0.0455	0.0465	**0.0563**	21.1%
	NDCG@10	0.0162	0.0316	0.0369	0.0348	0.0386	0.0414	0.0386	0.0441	0.0527	0.0509	**0.0621**	17.8%
	HR@5	0.0185	0.0354	0.0499	0.0489	0.0518	0.0537	0.0505	0.0564	0.0631	0.0648	**0.0750**	15.7%
	HR@10	0.0268	0.0487	0.0646	0.0627	0.0673	0.0683	0.0657	0.0729	0.0812	0.0793	**0.0929**	14.4%



(a) NDCG **(b) Hit Ratio**

Figure 3: Ablation study on the Office dataset.



(a) NDCG **(b) Hit Ratio**

Figure 4: Effect of the co-occurrence injection coefficient λ on the Office dataset.

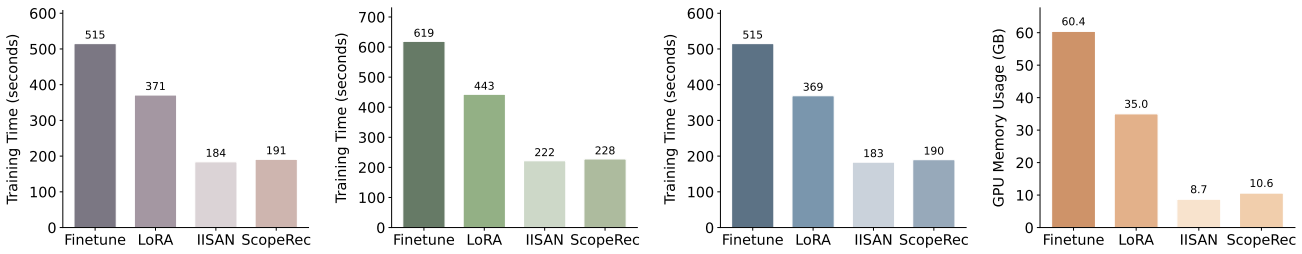
MoE. The *w_fusion* variant adds only the hierarchical representation fusion MoE.

Figure 3 reports the results on Office. All three modules contribute positively, with semantic co-occurrence mining and semantic focus awareness yielding the largest gains. This pattern confirms our central claim: explicit modeling of sequence-level associations and fine-grained focus is essential for accurate MSR. The hierarchical fusion module further provides consistent improvements, supporting the design of input-independent experts that handle modality- and level-specific signals separately. The full ScopeRec, which combines all three modules, is clearly the strongest, indicating that the three components are complementary rather than redundant.

A closer look at the relative gains is revealing. Adding semantic co-occurrence alone (*w_occ*) already lifts NDCG@10 by roughly 35% over the base, which shows that the lack of sequence-level association is indeed a primary bottleneck in standard pipelines. Semantic focus awareness (*w_foc*) brings a similar magnitude of improvement, suggesting that the noise introduced by uniformly aggregating multimodal content is comparably harmful. The two effects are largely orthogonal, since stacking them in the full model produces a further additional gain rather than a saturated curve.

4.4 Parameter Sensitivity (RQ3)

The hyperparameter λ controls how much semantic co-occurrence is injected into the original modality representations. We vary λ from 0.1 to 0.9 while keeping other settings fixed, and report the curves of NDCG and Hit Ratio in Figures 4a and 4b. ScopeRec stays above the second-best baseline of Table 2 across the entire range, which indicates that the co-occurrence module is not sensitive to small changes in λ. The best results are typically obtained for $\lambda \in [0.2, 0.5]$. When λ is small, the co-occurrence signal is too weak to influence the representation, and the model effectively falls back to a strong but non-sequential multimodal baseline. When λ exceeds 0.5, the original semantics of the item are gradually overwhelmed by the co-occurrence signal, which introduces redundancy and slightly degrades performance. The flat plateau in the middle range confirms that ScopeRec is robust within a wide window and easy to tune in practice.



(a) Office Training Time **(b) Scientific Training Time** **(c) Instrument Training Time** **(d) GPU Memory**

Figure 5: Average training time per epoch (seconds) on the three datasets and the corresponding GPU memory consumption (GB) during training.

4.5 Efficiency Analysis (RQ4)

We compare ScopeRec with three popular training paradigms on the modality encoders: full fine-tuning (*Finetune*), *LoRA*, and *IISAN*. All experiments share the same backbone, downstream SR model, and batch size for fairness.

Training time is shown in Figures 5a, 5b, and 5c. ScopeRec is markedly faster than both Finetune and LoRA on every dataset. Against the efficiency-oriented IISAN, ScopeRec only incurs a marginal overhead, which is negligible. The reason is that ScopeRec keeps the backbone modality encoders frozen and trains only the lightweight semantic mining and MoE modules on top of cached features. As a result, no gradients need to flow back through the heavy modality encoders.

GPU memory consumption is reported in Figure 5d. ScopeRec uses roughly one-sixth of the memory required by Finetune and around 70% less than LoRA. This reduction comes from the same decoupled design: by avoiding storage of the encoder activations needed for backpropagation, the computation graph remains small. Overall, ScopeRec achieves a strong balance between expressiveness and efficiency, which makes it readily deployable in practice.

The efficiency profile is particularly important for industrial settings, where the catalog updates frequently and the SR component has to be retrained on a daily or hourly basis. ScopeRec supports such workflows by separating the (slow) encoding of modality content from the (fast) training of the recommender, so that the modality features only need to be computed once per item and can be cached afterwards. The remaining trainable parameters are small enough that a single GPU is sufficient even for catalogs with tens of thousands of items.

4.6 Case Study and Visualization (RQ5)

To probe what the model actually learns, we visualize the textual and visual representations before and after semantic focus awareness, using t-SNE for clustering structure and KDE for density. Results are shown in Figure 6.

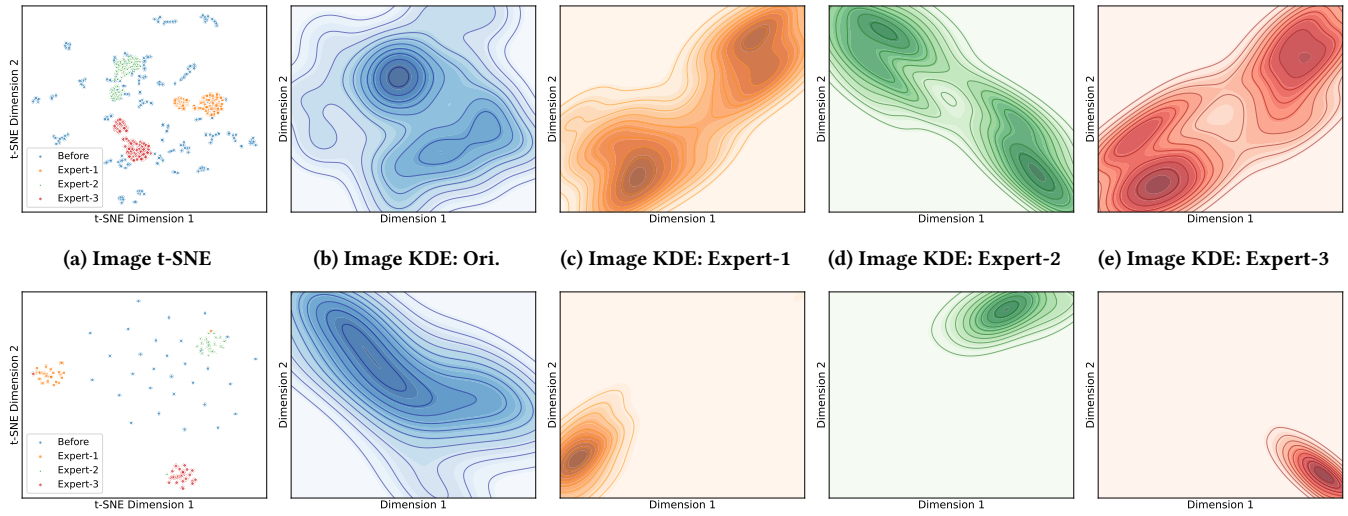
Comparing Figures 6a and 6f with the pre-focusing distributions, we see that the post-focus representations are more concentrated, indicating that the experts indeed extract a compact set of semantic regions. The KDE plots in Figures 6b and 6g (pre-focusing) are diffuse, suggesting that user-relevant cues are mixed with substantial noise. In contrast, the per-expert KDE plots (Figures 6c to 6e and 6h to 6j) display tighter, well-separated modes. Different experts attend to different regions, confirming that the MoE captures complementary aspects of user focus instead of redundant signals. The visualization also explains, at least qualitatively, why semantic focus awareness alone already yields a large lift in the ablation: filtering the dense and noisy raw distribution into a small number of well-separated modes is a strong inductive bias for any subsequent sequential model.

Figure 7 examines a sequence from the Office dataset together with the top-5 predictions of ScopeRec. The historical items strongly center on *tape*: the textual modality features recurring tokens such as *tape*, *inch*, and *adhesive*, while the visual modality contains multiple tape products. These cues exemplify sequence-level semantic co-occurrence. Despite the presence of unrelated items in the history, ScopeRec recommends items that align closely with the user's tape-purchasing intent in both modalities, supporting the design rationale that co-occurrence mining helps the model isolate persistent interests from incidental noise. We note that the visual modality provides cues that complement the textual modality: even when product titles are abbreviated or noisy, recurring visual patterns (here, the rolled shape of tape products) provide an additional anchor that keeps the recommendations on track.

4.7 Discussion

Why does co-occurrence mining help? Conventional MSR pipelines treat each item independently before passing the sequence through the SR backbone. The backbone is then asked to discover both *what* each item is and *how* items relate, with only ID-like position embeddings as guidance. In contrast, the co-occurrence module of ScopeRec runs an attention pass over the modality representations of historical items *before* the SR backbone. Such a design decouples item-level semantics from sequence-level association, and makes the eventual learning of the SR backbone easier. The gains on the Office dataset, which has the richest item space, are consistent with this view: a more diverse catalog brings a larger benefit from a dedicated co-occurrence pass.

Why hierarchical experts? A common practice in MSR is to take only the final-layer output of a frozen encoder. Our results suggest that this is suboptimal: different layers capture complementary aspects, and the optimal layer differs across users and items. The hierarchical MoE addresses this in two ways. The modality-level



(a) Image t-SNE (b) Image KDE: Ori. (c) Image KDE: Expert-1 (d) Image KDE: Expert-2 (e) Image KDE: Expert-3

(f) Text t-SNE (g) Text KDE: Ori. (h) Text KDE: Expert-1 (i) Text KDE: Expert-2 (j) Text KDE: Expert-3

Figure 6: Representation distributions before and after semantic focusing. t-SNE visualizes local clustering structure, while KDE depicts the corresponding local density. Each expert concentrates on a distinct semantic region, suggesting that the experts learn disentangled focus patterns.

Figure 7: Top-5 next-item predictions of ScopeRec on the Office dataset. Both textual and visual cues of historical items revolve around *tape*, and the recommendations capture this intent accurately.

experts disentangle distinct semantic regions within a single layer, while the fusion experts assign personalized weights to multiple layers. Both are critical, as confirmed by the ablation study.

Limitations and future work. ScopeRec currently fixes the layer selection schedule (one shallow, one deep, one final) and does not jointly fine-tune the modality encoders. Extending the framework with adaptive layer selection and light-weight encoder adaptation is a natural next step. A second direction is to couple semantic co-occurrence with cross-domain transfer, where the abstract, ID-agnostic nature of co-occurrence may prove particularly useful for cold-start domains. A third direction is to plug ScopeRec into industrial pipelines that involve multi-task learning, where the co-occurrence signal may be reused as an auxiliary self-supervision objective for upstream retrieval models. Finally, the connection between layer-wise hidden states and recommendation-aware semantics, while empirically validated here, deserves a more thorough theoretical understanding, especially in the context of larger multimodal foundation models such as multimodal large language models.

5 Conclusion

We presented ScopeRec, a multimodal sequential recommendation framework that shifts the focus from multimodal fusion to multimodal representation learning. ScopeRec is built upon two complementary ideas: *semantic co-occurrence*, which captures sequence-level associations among items in user behaviors, and *semantic focus*, which identifies the fine-grained components within each item that actually drive user choice. These ideas are realized through an attention-based co-occurrence mining module under a causal mask, a hierarchical mixture-of-experts that learns disentangled interest experts at multiple semantic levels, and an input-independent fusion module that personalizes the combination across modalities and levels. Extensive experiments on three public MSR benchmarks show that ScopeRec consistently surpasses state-of-the-art baselines by 13.8% to 28.7% across NDCG and Hit Ratio. Efficiency analysis further shows that, thanks to the decoupled training design, ScopeRec is competitive with the most efficient baselines in both training time and memory. We hope that the perspective of recommendation-aware representation learning, instantiated here by sequence-level co-occurrence and user-level focus, will inspire further research on tailoring multimodal pretraining signals for downstream recommendation tasks. As the MSR landscape continues to evolve toward larger foundation models and broader application domains, we believe that representation-centric designs of this kind will become an increasingly important complement to the prevailing fusion-centric and adaptation-centric paradigms.

GenAI Usage Disclosure

We utilized generative AI tools, specifically ChatGPT and Claude, solely for language refinement and grammar enhancement. No AI assistance was employed for technical content, data analysis, or experimental results. All research design, experimentation, analyses, and technical writing were entirely performed by the authors.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[3] Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. 2020. Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. *ACM Transactions on Information Systems (TOIS)* 39, 1 (2020), 1–42.

[4] Junchen Fu, Xuri Ge, Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, Jie Wang, and Joemon M Jose. 2024. IISAN: Efficiently adapting multimodal representation for sequential recommendation with decoupled PEFT. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 687–697.

[5] Junchen Fu, Xuri Ge, Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, Kaiwen Zheng, Yongxin Ni, and Joemon M Jose. 2024. Efficient and effective adaptation of multimodal foundation models in sequential recommendation. *arXiv preprint arXiv:2411.02992* (2024).

[6] Junchen Fu, Fajie Yuan, Yu Song, Zheng Yuan, Mingyue Cheng, Shenghui Cheng, Jiaqi Zhang, Jie Wang, and Yunzhu Pan. 2024. Exploring adapter-based transfer learning for recommender systems: Empirical studies and practical insights. In *Proceedings of the 17th ACM international conference on web search and data mining*. 208–217.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. *ICLR* (2016).

[9] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 585–593.

[10] Hengchang Hu, Wei Guo, Yong Liu, and Min-Yen Kan. 2023. Adaptive multi-modalities fusion in sequential recommendation systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 843–853.

[11] Xiaowen Huang, Shengsheng Qian, Quan Fang, Jitao Sang, and Changsheng Xu. 2018. Csan: Contextual self-attention network for user sequential recommendation. In *Proceedings of the 26th ACM international conference on Multimedia*. 447–455.

[12] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.

[13] Muyang Li, Zijian Zhang, Xiangyu Zhao, Wanyu Wang, Minghao Zhao, Runze Wu, and Ruocheng Guo. 2023. Automlp: Automated mlp for sequential recommendations. In *Proceedings of the ACM web conference 2023*. 1190–1198.

[14] Youhua Li, Hanwen Du, Yongxin Ni, Pengpeng Zhao, Qi Guo, Fajie Yuan, and Xiaofang Zhou. 2024. Multi-modality is all you need for transferable recommender systems. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 5008–5021.

[15] Chang Liu, Xiaoguang Li, Guohao Cai, Zhenhua Dong, Hong Zhu, and Lifeng Shang. 2021. Noninvasive self-attention for side information fusion in sequential recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 4249–4256.

[16] Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2024. Multimodal recommender systems: A survey. *Comput. Surveys* 57, 2 (2024), 1–17.

[17] Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. 2024. Multimodal pretraining, adaptation, and generation for recommendation: A survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 6566–6576.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[19] Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024. Fantastic Semantics and Where to Find Them: Investigating Which Layers of Generative LLMs Reflect Lexical Semantics. arXiv:2403.01509 [cs.CL] https://arxiv.org/abs/2403.01509

[20] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 188–197.

[21] Xingyu Pan, Yushuo Chen, Changxin Tian, Zihan Lin, Jinpeng Wang, He Hu, and Wayne Xin Zhao. 2022. Multimodal meta-learning for cold-start sequential recommendation. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 3421–3430.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[23] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013* (2025).

[24] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[25] Yong Kiam Tan, Xinxing Xu, and Yong Liu. 2016. Improved recurrent neural networks for session-based recommendations. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 17–22.

[26] Jianling Wang, Kaize Ding, Liangjie Hong, Huan Liu, and James Caverlee. 2020. Next-item recommendation with sequential hypergraphs. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 1101–1110.

[27] Jinpeng Wang, Ziyun Zeng, Yunxiao Wang, Yuting Wang, Xingyu Lu, Tianxiang Li, Jun Yuan, Rui Zhang, Hai-Tao Zheng, and Shu-Tao Xia. 2023. Missrec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6548–6557.

[28] Shoujin Wang, Liang Hu, Yan Wang, Longbing Cao, Quan Z Sheng, and Mehmet Orgun. 2019. Sequential recommender systems: challenges, progress and prospects. *arXiv preprint arXiv:2001.04830* (2019).

[29] Chuhan Wu, Fangzhao Wu, Tao Qi, Chao Zhang, Yongfeng Huang, and Tong Xu. 2022. Mm-rec: Visiolinguistic model empowered multimodal news recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 2560–2564.

[30] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, and Edith CH Ngai. 2025. The Best is Yet to Come: Graph Convolution in the Testing Phase for Multimodal Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6325–6334.

[31] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, Yijie Li, Jianheng Tang, Yunhuai Liu, and Edith CH Ngai. 2026. CAMMSR: Category-Guided Attentive Mixture of Experts for Multimodal Sequential Recommendation. *arXiv preprint arXiv:2603.04320* (2026).

[32] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, Wei Wang, Xiping Hu, and Edith Ngai. 2025. NLGCL: Naturally Existing Neighbor Layers Graph Contrastive Learning for Recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 319–329.

[33] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Wei Wang, Xiping Hu, Steven Hoi, and Edith Ngai. 2026. A survey on multimodal recommender systems: Recent advances and future directions. *IEEE Transactions on Multimedia* (2026).

[34] Lanling Xu, Zhen Tian, Bingqian Li, Junjie Zhang, Daoyuan Wang, Hongyu Wang, Jinpeng Wang, Sheng Chen, and Wayne Xin Zhao. 2024. Sequence-level semantic representation fusion for recommender systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 5015–5022.

[35] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M Jose, and Xiangnan He. 2019. A simple convolutional generative network for next item recommendation. In *WSDM*. 582–590.

[36] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. 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*. 2639–2649.

[37] Shengzhe Zhang, Liyi Chen, Dazhong Shen, Chao Wang, and Hui Xiong. 2025. Hierarchical Time-Aware Mixture of Experts for Multi-Modal Sequential Recommendation. In *Proceedings of the ACM on Web Conference 2025*. 3672–3682.

[38] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *arXiv preprint arXiv:2302.04473* (2023).

[39] Jieming Zhu, Xin Zhou, Chuhan Wu, Rui Zhang, and Zhenhua Dong. 2024. Multimodal pretraining and generation for recommendation: A tutorial. In *Companion Proceedings of the ACM Web Conference 2024*. 1272–1275.