CSA-RIT: LEARNING NOISE-RESILIENT MULTIMODAL RECOMMENDATION VIA COARSE SEMANTIC ALIGNMENT AND ROBUST INTEREST TREE

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph-based multimodal recommenders typically build an item–item graph per modality and aggregate neighbor representations by graph convolution. Two things decide the outcome of this aggregation. The first is what the modality features being averaged express, and the second is which items enter a neighborhood. Both are settled before the first propagation step, and repeated smoothing spreads any error into every representation. On the feature side, an item's visual and textual features are typically extracted independently and often inconsistent. We call this content noise. On the graph side, existing methods often introduce noisy links when building item–item graphs. We call this structural noise. We propose CSA-RIT, a noise-resilient framework that treats content and structural noise as two sources of a single pre-propagation error and repairs both. Coarse Semantic Alignment (CSA) places a coarse discrete bottleneck ahead of graph propagation and aligns the same item across modalities only in a shared coarse code, so that the signals to be averaged share a common semantic basis. Robust Interest Tree (RIT) augments each frozen modality graph with a short, order-aware co-occurrence tree, supplying a stable prior that resists this structural noise. On three standard MMRec benchmarks spanning catalog sizes from 7k to 63k items, CSA-RIT is best on every metric, improving over the strongest baseline by 8.9% on average and by 13.7% on the largest split. Ablations confirm complementary gains, and RIT keeps its recall under injected structural noise. Our implementation is available at `https://github.com/paper-ai212/CSA-RIT`.

1 INTRODUCTION

Recommender systems are crucial tools for helping users discover relevant content from massive online information Xu et al. (2025b). In recent years, the prevalence of multimedia platforms has led to a surge of items rich in visual and textual content. Modeling such heterogeneous information provides a more comprehensive understanding of user interests, which gives rise to multimodal recommendation. By integrating content from multiple perspectives, multimodal recommender systems offer greater potential to accurately capture user preferences Wang et al. (2025); Tang et al. (2025); Zhang et al. (2025).

Early approaches, such as VBPR He & McAuley (2016) and VPOI Wang et al. (2017), integrated modality features with item IDs to enrich item representations from multiple perspectives. In recent years, graph neural networks have been extensively studied to capture high-order structure in multimodal recommendation Wei et al. (2020); Zhang et al. (2021).

Within this line of work, FREEDOM Zhou & Shen (2023) argues that learning per-modality item–item graphs is inefficient and instead proposes freezing them, achieving superior performance. Motivated by this success, constructing and leveraging modality-specific item–item graphs has emerged as a prevalent paradigm in multimodal recommendation. These methods typically build item–item graphs from pre-extracted multimodal features and then aggregate neighbor representations by graph convolution. This operator averages item representations over a neighborhood, so its outcome is jointly determined by two inputs, namely the modality features that decide what is being averaged and the adjacency that decides which items enter the neighborhood. In current practice, both in-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

puts are already corrupted before the first propagation step, and repeated smoothing spreads these errors into every representation. The first is content noise, which arises because an item's visual and textual features are extracted independently and are often inconsistent, so even over a correct neighborhood the operator averages signals that are not yet mutually aligned. The second is structural noise, which arises because each item–item graph links an item to its top-k most similar items and therefore returns a full k-sized neighborhood whether or not k genuinely related items exist. Dissimilar items are then forced into that neighborhood, so small similarity-ranking errors at this cutoff become frozen edges, and the averaging operator runs over the wrong support. Appendix Figure 4 illustrates this on a sampled catalog, where the two independently extracted views disagree on which items are similar and the local embedding neighborhood mixes items that are close in one modality but weakly related in user interest. These are not two isolated defects but two sources of a single pre-propagation error that neighborhood smoothing amplifies jointly.

Guided by this view, we propose CSA-RIT, a noise-resilient framework that repairs the two inputs of this aggregation (Figure 1). For the content input, we introduce Coarse Semantic Alignment (CSA). Ahead of graph propagation, a shared encoder–decoder pair and a single-layer finite scalar quantizer (FSQ) Mentzer et al. (2024) form a coarse discrete bottleneck, and no information beyond the shared coarse code q_0 propagates to the graphs. An instance-level consistency regularizer (ICR) of controllable strength pulls the same item together across modalities in this coarse code, so that the signals later averaged by the graphs share a common semantic basis. For the structural input, we develop a Robust Interest Tree (RIT). RIT grows a short co-occurrence tree with an order-aware beam, leaving first-order co-occurrence neighbors unpruned and beam-selecting only higher-order nodes. The tree augments each frozen modality graph, which is retained rather than replaced, through a ReLU-gated, parameter-free enhancement; the separate co-occurrence normalized adjacency is never overwritten. The two components therefore act on one operator rather than being independent add-ons. CSA reduces the content error of the features that are averaged, and RIT reduces the structural error of the neighborhood over which they are averaged. Experiments on three standard MMRec benchmarks, spanning catalog sizes from 7k to 63k items, show that CSA-RIT is best on every metric, that ranking co-moves with q_0 during training, that ablations confirm complementary gains, and that RIT keeps its recall under injected structural noise.

In summary, this work makes the following contributions.

- We model the noise in frozen-graph multimodal recommenders as content noise from unaligned cross-modal features and structural noise from erroneous item–item links, both of which are present before propagation and are amplified by graph convolution.
- We propose CSA, which aligns an item's modalities in a coarse quantized code ahead of propagation. The coarse code is observable during training, since its mean cosine tracks Recall@20 with a Pearson correlation of 0.56 on Baby and 0.75 on Sports, and it keeps rising after validation performance peaks, so further alignment no longer improves ranking.
- We propose RIT, an order-aware co-occurrence tree fused into frozen modality graphs. Its fusion is parameter-free and introduces no smoothing knobs. Under injected structural noise it retains its Recall@20, whereas the backbone and CSA degrade at comparable rates, so the two repairs are not interchangeable.
- We evaluate CSA-RIT with ablations on three benchmarks. Both modules improve on the backbone on all three datasets, and the full model is best on every metric.

2 Related Works

2.1 Multimodal Recommendation Methods

Multimodal recommendation augments collaborative filtering with visual and textual content, alleviating data sparsity and enriching item representations Liu et al. (2024a). Early methods such as VBPR He & McAuley (2016) and VPOI Wang et al. (2017) fuse pre-extracted modality features with CF embeddings, but they neither refine those features nor model item–item structure.

Recent work instead builds an item–item graph per modality and aggregates neighbor representations over it by graph convolution. FREEDOM Zhou & Shen (2023) shows that learning such modality graphs is inefficient and freezes them, retaining strong performance with frozen pre-extracted

features; this frozen-graph recipe is now the dominant setting for multimodal recommendation. Recent methods refine it in different directions: IIMRec Xu et al. (2026) reuses one item–item graph for representation, interaction-graph expansion, and BPR; DHMRec Zhan et al. (2026) disentangles modality-common and modality-specific factors; MoToRec Liu et al. (2026) tokenizes multimodal features with a residual quantizer; SiMGR Lian et al. (2026) injects signed feedback into the multimodal graph; MSCA Li & Wang (2026) and AlignRec Liu et al. (2024b) align representations across views; PNGRec He et al. (2026) splits user–item signals into spectral graphs; and DIGEST Sai et al. (2026) reweights noisy interaction edges. All of them keep the frozen MMRec encoders.

2.2 Applications of Graph Neural Networks in Recommendation

Neighborhood aggregation over a user–item graph captures high-order collaborative signals, which makes graph neural networks a standard tool in recommendation Gao et al. (2022); LightGCN He et al. (2020) shows that a simplified propagation layer suffices for this purpose. Transferred to multimodal recommendation, LATTICE Zhang et al. (2021) propagates over an independently built item graph per modality, MGCN Yu et al. (2023) adds behavior-aware purification and fusion modules, DRAGON Zhou et al. (2023a) fuses modality GNNs with attention, and FREEDOM Zhou & Shen (2023) freezes the modality graphs. Under this frozen-graph recipe, two limitations remain. First, top-k sparsification is sensitive to similarity ranking noise, and repeated neighborhood aggregation can amplify early edge-selection errors. Second, independently encoded modalities of the same item need not share a coarse semantic code before propagation, so graph convolution can mix unaligned views. IIMRec, DIGEST, and PNGRec attack graph noise from different angles, but none of them places a dose-controlled coarse discrete bottleneck in front of graph convolution.

3 Our Approach

We consider a multimodal recommendation scenario comprising a user set $\mathcal{U}$ of size I and an item set $\mathcal{V}$ of size J. The available data consist of the observed user–item interaction matrix $\mathbf{A} \in \{0,1\}^{I\times J}$, along with multimodal feature representations for each item, specifically visual features $\mathbf{E}^v \in \mathbb{R}^{J\times d_v}$ and textual features $\mathbf{E}^t \in \mathbb{R}^{J\times d_t}$. The core challenge lies in effectively integrating these two heterogeneous sources of information to improve recommendation accuracy. Formally, the objective is to learn a scoring function $f(u_i, v_j \mid \mathbf{A}, \mathbf{E}^v, \mathbf{E}^t)$ that estimates the likelihood of a future interaction between user u_i and item v_j.

As illustrated in Fig. 1, CSA-RIT consists of four key steps. (1) Image and text features are extracted via pre-trained networks and projected into a unified space, yielding the visual view and the whitened textual views. (2) Coarse Semantic Alignment maps each modality through a shared encoder, a single-layer FSQ bottleneck, and a shared decoder, so that the codes entering subsequent modules are determined exclusively by the coarse code q_0; ICR is applied on q_0. (3) A Robust Interest Tree is grown on the item co-occurrence graph with an order-aware beam and fused into each frozen modality knn graph by a parameter-free enhancement. (4) GNNs are employed on the decoded coarse codes and ID embeddings to derive refined representations, which are then fused for final prediction.

3.1 Feature Extraction and Transformation

WhitenRec Zhang et al. (2024) pointed out that pre-trained text embeddings reside in an anisotropic semantic space, which can reduce representation discriminability in recommendation. Following Meng et al. (2025), we directly use its preprocessed multimodal features, where the raw text feature $\mathbf{E}^t_{\text{raw}}$ is transformed by PCA Kessy et al. (2018), ICA Jutten & Herault (1991), and ZCA Bell & Sejnowski (1997), yielding three modality views denoted as $\mathbf{E}^{\text{feat}}$ with feat $\in \{v, p, z\}$:

$$\begin{aligned} \mathbf{E}^p &= \text{ICA}\left(\text{PCA}(\mathbf{E}^t_{\text{raw}})\right), \\ \mathbf{E}^z &= \text{ZCA}(\mathbf{E}^t_{\text{raw}}), \\ \mathbf{E}^v &= \mathbf{E}^v_{\text{raw}}. \end{aligned} \tag{1}$$

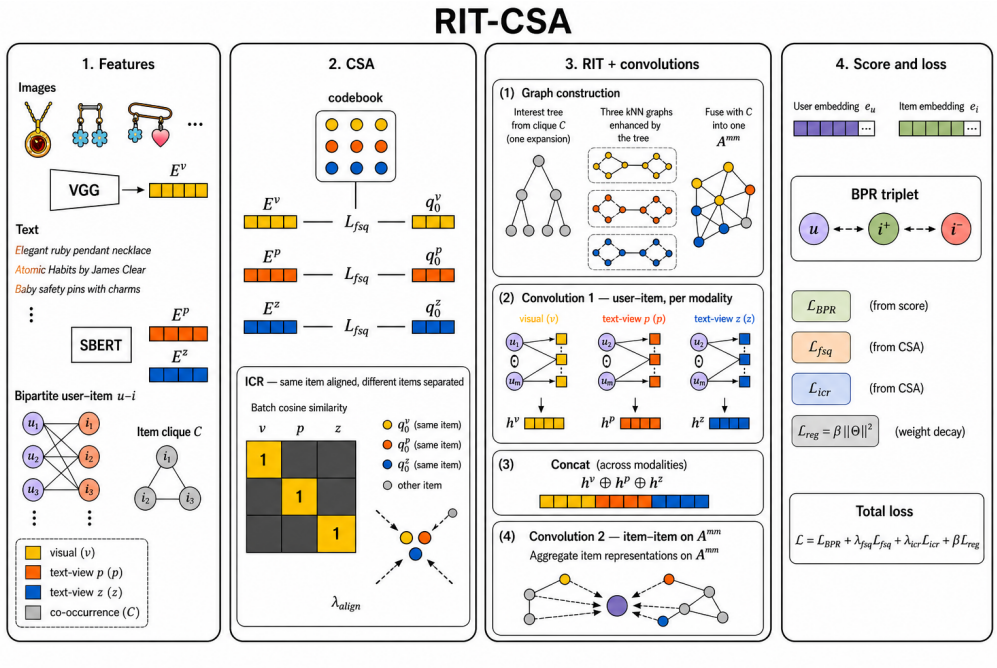


Figure 1: Overview of CSA-RIT.

3.2 Coarse Semantic Alignment

Cross-modal encoders of the same item are not guaranteed to share a coarse semantic code before graph convolution. We therefore place a coarse discrete bottleneck in front of GCN and align modalities only in that code. Let $\mathbf{Z}^m$ denote the latent feature of modality $m \in \mathcal{M} = \{v, p, z\}$ after the modality-specific projection. A shared encoder E and decoder D sandwich a single-layer finite scalar quantizer (FSQ) Mentzer et al. (2024):

$$\mathbf{c}^m = E(\mathbf{Z}^m), \qquad q_0^m = \mathrm{FSQ}(\mathbf{c}^m), \qquad \tilde{\mathbf{e}}^m = D(q_0^m). \tag{2}$$

Only $\tilde{\mathbf{e}}^m$ is fed to the subsequent graph convolution.

FSQ maps $\mathbf{c}^m$ into $(-1, 1)$ and quantizes each coordinate onto a uniform codebook of K_c levels, using the single-layer scale $g = 1$. Then

$$\mathbf{U}^m = 2 \cdot \sigma(1.6 \cdot g \cdot \mathbf{c}^m) - 1, \tag{3}$$

$$\mathcal{C} = \left\{ -1 + \frac{2t}{K_c - 1} \;\middle|\; t = 0, 1, \ldots, K_c - 1 \right\}, \tag{4}$$

and each coordinate of $\mathbf{U}^m$ is replaced by its nearest level in $\mathcal{C}$. The inverse map recovers q_0^m in the encoder space:

$$q_0^m = \frac{1}{1.6 \cdot g} \log\left(\frac{1 + \mathrm{clip}(\mathbf{U}_{\mathrm{st}}^m, -1 + \epsilon, 1 - \epsilon)}{1 - \mathrm{clip}(\mathbf{U}_{\mathrm{st}}^m, -1 + \epsilon, 1 - \epsilon)} \right). \tag{5}$$

The reconstruction term $\mathcal{L}_{\mathrm{fsq}}$ is the MSE between the encoder input and the quantizer reconstruction.

Instance-level consistency on q_0. Let x and y be ℓ_2-normalized q_0 codes of two modalities for the B items in a mini-batch, and $\mathbf{S} = xy^\top$. ICR encourages the same item to align while separating different items:

$$\mathcal{L}_{\mathrm{pair}}(x, y) = \frac{1}{B} \sum_{i=1}^{B} (1 - \mathbf{S}_{ii})^2 + \frac{1}{B(B-1)} \sum_{i \neq j} \mathbf{S}_{ij}^2. \tag{6}$$

We average the three modality pairs (v, p), (v, z), and (p, z):

$$\mathcal{L}_{\mathrm{icr}} = \frac{1}{3} \Big(\mathcal{L}_{\mathrm{pair}}(q_0^v, q_0^p) + \mathcal{L}_{\mathrm{pair}}(q_0^v, q_0^z) + \mathcal{L}_{\mathrm{pair}}(q_0^p, q_0^z) \Big). \tag{7}$$

The scalar λ_{align} is the alignment dose: it should be large enough for q_0 to become a shared coarse code, and not so large that the views collapse. We monitor the mean cosine similarity among $\{q_0^v, q_0^p, q_0^z\}$ (denoted q_0-cos in the experiments) as an observable of this process.

The decoded codes are the features that graph convolution later consumes; the adjacency it uses is produced by RIT.

3.3 Robust Interest Tree Enhanced Homogeneous Graph

After CSA, graph convolution still needs a robust item–item adjacency. Rigid top-k modality graphs freeze ranking errors into edges that graph convolution then amplifies. RIT injects a complementary interest prior that is order-aware in search and parameter-free in enhancement. All item–item matrices below are $J \times J$ and need not be symmetric, as each row is constructed independently; we write $[x]_+ := \max(x, 0)$.

Symmetric normalization. For a (possibly directed) weight matrix $W \in \mathbb{R}^{J\times J}$, let $d_i(W) = \sum_{j=1}^{J} W_{ij}$ be the out-weight of item i. The operator used by later convolution is the symmetrically normalized adjacency

$$\text{Norm}(W)_{ij} = d_i(W)^{-1/2}\, W_{ij}\, d_j(W)^{-1/2}, \tag{8}$$

with a small $\varepsilon > 0$ added to empty rows.

Frozen modality knn graph. For each modality $m \in \mathcal{M}$,

$$S_{ij}^m = \frac{\big(\mathbf{e}_i^m\big)^\top \big(\mathbf{e}_j^m\big)}{\|\mathbf{e}_i^m\|\,\|\mathbf{e}_j^m\|}, \tag{9}$$

where $\mathbf{e}_i^m$ is the i-th row of $\mathbf{E}^m$. The top-k neighborhood (default k=10) is

$$\mathcal{N}_k^m(i) = \arg\text{top-}k_{j\in\mathcal{V}}\, S_{ij}^m. \tag{10}$$

The frozen knn graph is the unweighted indicator of this neighborhood, then normalized:

$$W_{ij}^{\text{knn},m} = \mathbf{1}\big[j \in \mathcal{N}_k^m(i)\big], \qquad \hat{\mathbf{A}}^{\text{knn},m} = \text{Norm}\big(W^{\text{knn},m}\big). \tag{11}$$

Co-occurrence graph. Let C_{ij} be the number of times items i and j co-occur in the same interest set, and let

$$\mathcal{N}_{k_c}^c(i) = \underset{j\in\mathcal{V}}{\arg\text{top-k}_\text{c}}\; C_{ij} \tag{12}$$

(default k_c=10). Tree search uses the weighted restriction

$$\dot{S}_{ij}^c = C_{ij}\, \mathbf{1}\big[j \in \mathcal{N}_{k_c}^c(i)\big], \tag{13}$$

while a separate collaborative prior is the unweighted normalized adjacency on the same support,

$$W_{ij}^c = \mathbf{1}\big[j \in \mathcal{N}_{k_c}^c(i)\big], \qquad \hat{\mathbf{A}}^c = \text{Norm}(W^c). \tag{14}$$

Order-aware beam search. On $\dot{\mathbf{S}}^c$, each item r is expanded as a BFS tree of order n (default n=2). Order 1 keeps the full co-occurrence neighborhood; later orders keep only a beam of width b (default b=5), ranked by the product of edge weights along the walk. Appendix Algorithm 1 states the procedure; we write $T_r = N_r = \bigcup_{o=1}^{n} N_r^{(o)}$ for the resulting node set.

Parameter-free enhancement. Given T_i and $\mathcal{N}_k^m(i)$, form the dense weight matrix

$$W_{ij}^{\text{tree},m} = \frac{1}{2}\, S_{ij}^m\, \mathbf{1}\big[j \in \mathcal{N}_k^m(i)\big] + \frac{1}{2}\, \big[S_{ij}^m\big]_+\, \mathbf{1}\big[j \in N_i\big]. \tag{15}$$

The rectifier applies only to the tree term, so a negative similarity cannot enter as a tree edge. Then

$$\hat{\mathbf{A}}^{\text{tree},m} = \text{Norm}\big(W^{\text{tree},m}\big). \tag{16}$$

Fusion with the frozen knn graph is

$$\tilde{\mathbf{A}}^m = \frac{1}{2}\,\hat{\mathbf{A}}^{\mathrm{knn},m} + \hat{\mathbf{A}}^{\mathrm{tree},m}. \tag{17}$$

The multimodal item graph used by later convolution is

$$\mathbf{A}^{\mathrm{mm}} = \alpha_v \tilde{\mathbf{A}}^v + \alpha_p \tilde{\mathbf{A}}^p + \alpha_z \tilde{\mathbf{A}}^z + \alpha_c \hat{\mathbf{A}}^c, \tag{18}$$

with dataset-specific nonnegative weights summing to one. Item representations H, initialized from the CSA-decoded codes $\tilde{\mathbf{e}}^m$, are updated by ℓ multiplications $H \leftarrow \mathbf{A}^{\mathrm{mm}} H$ ($\ell = 1$ by default). RIT therefore adds a short, order-aware interest tree as a structural prior, at no extra trainable parameters.

3.4 Training Objective

We first define the preference score between user u and item i as

$$z_{u,i} = \mathbf{e}_u^\top \mathbf{e}_i. \tag{19}$$

Here $\mathbf{e}_u$ and $\mathbf{e}_i$ are the final user and item representations used for scoring, obtained by fusing the propagated representations with the ID embeddings. Based on this score, we optimize recommendation with Bayesian Personalized Ranking loss:

$$\mathcal{L}_{bpr} = \sum_{u,i^+,i^- \in D} \big(-\ln \sigma(z_{u,i^+} - z_{u,i^-})\big) + \beta \left\|\Theta\right\|_2^2, \tag{20}$$

where D is the training triplet set, $\sigma(\cdot)$ is the sigmoid, and β controls L_2 regularization.

The overall training objective is

$$\mathcal{L}_{total} = \mathcal{L}_{bpr} + \lambda_{\mathrm{fsq}} \mathcal{L}_{\mathrm{fsq}} + \lambda_{\mathrm{align}} \mathcal{L}_{\mathrm{icr}}, \tag{21}$$

where λ_{fsq} weights the FSQ reconstruction of the coarse code and λ_{align} is the alignment dose on q_0.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

We follow the MMRec protocol on three Amazon review categories McAuley et al. (2015): Baby, Sports and Outdoors (Sports), and Electronics. Each item has a visual image and a textual description; image and text features are the frozen VGG and Sentence-BERT embeddings released by MMRec. Table 1 reports dataset statistics after 5-core filtering. Electronics is an order of magnitude larger than Baby (192,403 users, 63,001 items, 1,689,188 interactions, sparsity 99.98%). It is the standard large-scale stress test for item–item graphs in this benchmark: dense cosine matrices over 63k items are memory-heavy, and several prior methods in Table 2 omit it. Many 2026 papers likewise report only Baby, Sports, or Clothing. We keep Electronics in the main table because graph noise and cross-modal neighborhood disagreement are more severe at this scale.

4.1.2 Baselines

We compare against general collaborative filtering models (BPR, LightGCN, LayerGCN) and multimodal recommenders that use the same frozen MMRec features: VBPR He & McAuley (2016), MMGCN Wei et al. (2019), DualGNN Wang et al. (2021), GRCN Wei et al. (2020), SLMRec Tao et al. (2022), BM3 Zhou et al. (2023b), FREEDOM Zhou & Shen (2023), MGCN Yu et al. (2023), SOIL Su et al. (2024), GUME Lin et al. (2024), AlignRec Liu et al. (2024b), SMORE Ong & Khong (2025), MENTOR Xu et al. (2025a), SiMGR Lian et al. (2026), MoToRec Liu et al. (2026), and IIMRec Xu et al. (2026).

Table 1: Dataset Statistics.

Dataset	#Users	#Items	#Interactions	#Sparsity
Baby	19445	7050	160792	99.88%
Sports	35598	18357	296337	99.95%
Electronics	192403	63001	1689188	99.98%

Table 2: Overall performance on the MMRec Amazon splits. Best results are in bold and second-best are underlined. "–" means the entry is not reported on this Electronics split.

Dataset	Baby				Sports				Electronics			
	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
BPR (UAI'09)	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0235	0.0367	0.0127	0.0161
LightGCN (SIGIR'20)	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0313	0.0387	0.0363	0.0540	0.0204	0.0250
LayerGCN (ICDE'23)	0.0529	0.0820	0.0281	0.0355	0.0594	0.0916	0.0323	0.0406	0.0391	0.0581	0.0220	0.0269
VBPR (AAAI'16)	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0293	0.0458	0.0159	0.0202
MMGCN (MM'19)	0.0378	0.0615	0.0200	0.0261	0.0370	0.0605	0.0193	0.0254	0.0207	0.0331	0.0109	0.0141
DualGNN (MM'21)	0.0448	0.0716	0.0240	0.0309	0.0568	0.0859	0.0310	0.0385	0.0363	0.0541	0.0202	0.0248
GRCN (MM'20)	0.0532	0.0824	0.0282	0.0358	0.0559	0.0877	0.0306	0.0389	0.0349	0.0529	0.0195	0.0241
SLMRec (TMM'22)	0.0540	0.0810	0.0285	0.0357	0.0676	0.1017	0.0374	0.0462	0.0422	0.0630	0.0237	0.0291
BM3 (WWW'23)	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0437	0.0648	0.0247	0.0302
FREEDOM (MM'23)	0.0624	0.0985	0.0324	0.0416	0.0710	0.1077	0.0382	0.0476	0.0396	0.0601	0.0220	0.0273
MGCN (MM'23)	0.0620	0.0964	0.0339	0.0427	0.0729	0.1106	0.0397	0.0496	0.0439	0.0643	0.0245	0.0298
SOIL (MM'24)	0.0680	0.1028	0.0365	0.0454	0.0786	0.1155	0.0435	0.0530	0.0492	0.0718	0.0279	0.0337
GUME (CIKM'24)	0.0673	0.1042	0.0365	0.0460	0.0778	0.1165	0.0427	0.0527	0.0458	0.0680	0.0253	0.0310
AlignRec (CIKM'24)	0.0674	0.1046	0.0363	0.0458	0.0758	0.1160	0.0414	0.0517	0.0472	0.0700	0.0262	0.0321
SMORE (WSDM'25)	0.0680	0.1035	0.0365	0.0457	0.0762	0.1142	0.0408	0.0506	0.0437	0.0650	0.0244	0.0299
MENTOR (AAAI'25)	0.0678	0.1048	0.0362	0.0450	0.0763	0.1139	0.0409	0.0511	0.0439	0.0655	0.0244	0.0300
SiMGR (AAAI'26)	0.0702	0.1052	0.0382	0.0471	0.0782	0.1160	0.0428	0.0526	0.0479	0.0701	0.0273	0.0330
MoToRec (AAAI'26)	0.0709	0.1077	0.0376	0.0473	0.0784	0.1163	0.0433	0.0529	–	–	–	–
IIMRec (MM'26)	0.0709	0.1094	0.0387	0.0485	0.0793	0.1184	0.0431	0.0537	–	–	–	–
CSA-RIT	**0.0728**	**0.1122**	**0.0402**	**0.0504**	**0.0875**	**0.1279**	**0.0485**	**0.0589**	**0.0559**	**0.0820**	**0.0316**	**0.0383**

4.1.3 Evaluation Metrics

For all methods, ID embeddings with dimensionality 64 are initialized by Xavier initialization, and model training is conducted with Adam. For CSA we use a single FSQ layer and tune $\lambda_{\text{align}} \in \{0.05, 0.1, 0.2, 0.5\}$ together with learning rates $\{5 \times 10^{-5}, 10^{-4}\}$; the reported CSA configuration is selected on validation Recall@20. RIT uses maximum order n=2, beam width b=5 with first-order neighbors unpruned, k=k_c=10, and the parameter-free enhancement in Section 3.3; the same setting is used on Baby, Sports, and Electronics. For each user, interaction records are randomly split into training, validation, and test sets with a ratio of 8:1:1. We report recommendation performance using Recall at K and Normalized Discounted Cumulative Gain at K.

4.2 Overall Performance Comparison

Table 2 compares CSA-RIT with general and multimodal baselines under the frozen MMRec feature protocol. CSA-RIT attains the best score on every metric and dataset, including the 2026 graph-based methods that report the same Baby/Sports splits. The strongest published 2026 competitor on those two sets is IIMRec; CSA-RIT remains higher on Baby (R@20 0.1122 vs. 0.1094) and by a larger margin on Sports (R@20 0.1279 vs. 0.1184). MoToRec and IIMRec do not report the large Electronics split used here, and the artifacts they release do not cover it; our reproduction did not match their reported performance, so we leave these entries empty. SiMGR's published Electronics results come from a different catalog of 20,247 users, so we re-implemented it on our split, where it reaches R@20 0.0701 and stays below SOIL. On Electronics, CSA-RIT reaches R@20 0.0820 against SOIL 0.0718. MSCA, PNGRec, and DIGEST do evaluate Electronics, but we could not verify a complete public metric table at the time of writing, so they appear only in related work.

Table 3: Ablation Study on Key Modules.

	Baby				Sports				Electronics			
Variant	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
Backbone	0.0711	0.1042	0.0392	0.0481	0.0830	0.1132	0.0461	0.0563	0.0540	0.0788	0.0298	0.0355
w/ CSA	0.0720	0.1095	0.0396	0.0487	0.0853	0.1263	0.0478	0.0582	0.0555	0.0811	0.0310	0.0381
w/ RIT	0.0723	0.1102	0.0392	0.0492	0.0867	0.1272	0.0480	0.0585	0.0553	0.0808	0.0313	0.0379
CSA-RIT	0.0728	0.1122	0.0402	0.0504	0.0875	0.1279	0.0485	0.0589	0.0559	0.0820	0.0316	0.0383

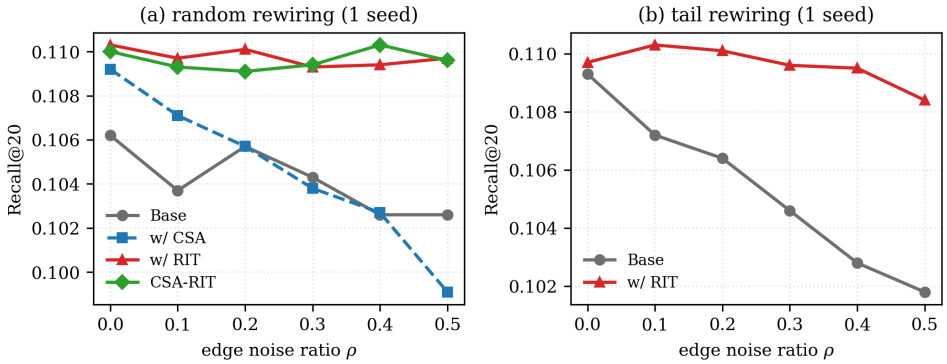


Figure 2: Recall@20 on Baby against the fraction ρ of the edges of each frozen modality top-k graph replaced before training. (a) Random rewiring draws replacement edges from uniformly random items. (b) Tail rewiring draws them from the lowest-similarity items of the same row.

4.3 Ablation Studies

We conduct ablation experiments to evaluate CSA and RIT. Table 3 reports the backbone, the backbone with CSA only, the backbone with RIT only, and the full CSA-RIT model.

Each module improves on the backbone on all three datasets, and combining them gives the best result. The gains are not additive, which is consistent with the two modules acting on the two inputs of the same aggregation. CSA regularizes the same item across modalities before propagation, while RIT adds an edge support drawn from user co-occurrence rather than from the similarity ranking that builds the modality graphs, so it also helps while those graphs are intact; the structural-noise study in Section 4.4 shows that its margin over the backbone widens as the graph degrades.

4.4 Analysis of Structural Robustness

RIT makes a structural claim, so we test it by corrupting the structure itself. Before any training begins, we replace a fixed fraction ρ of the edges of every frozen modality top-k graph. This targets exactly the failure mode that motivates RIT, because a rigid top-k neighborhood must return k neighbors even when fewer than k of them are relevant. We use two replacement rules. Random rewiring redirects an edge to a uniformly drawn item, while tail rewiring draws it from the lowest-similarity items of the same row, which is closer to how a forced edge actually originates. Row degrees are preserved, the injected edges are drawn once and then held fixed for the whole run, and every variant therefore sees exactly the same corrupted graphs.

Fig. 2 reports Baby Recall@20 against ρ. The backbone loses between 3% and 7% of its clean Recall@20 as ρ grows to 0.5, and CSA degrades at least as fast, which is what a module acting on the content side should do under purely structural noise. RIT stays within roughly 1% of its clean value over the same range, and CSA-RIT holds the highest curve at both ends of the sweep. The margin of RIT over the backbone therefore widens with ρ instead of remaining constant. This is the behavior expected of a module that repairs corrupted structure. RIT also improves the clean benchmarks (Table 3), where its co-occurrence support remains informative, but the sweep shows that its advantage grows as the graph degrades.

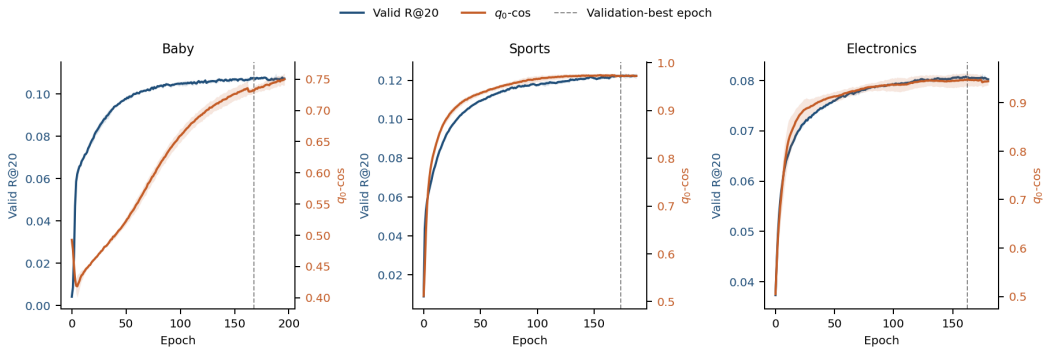


Figure 3: Training trajectories of q_0-cos and validation Recall@20 at the selected CSA dose (three-seed mean; shaded band, ± 1 sample standard deviation). The dashed line marks the mean validation-best epoch.

The mechanism follows from where the tree support comes from. Tree edges are selected by user co-occurrence, which is independent of the similarity ranking that produced the corrupted edges, so RIT restores relations that the corrupted cosine ranking can no longer rank correctly. This also separates the two modules. Structural noise is the regime in which RIT matters, and it is the regime in which CSA does not help, which is consistent with the two modules repairing the two different inputs of neighborhood averaging.

4.5 Analysis of Coarse Alignment

We next examine CSA under a matched backbone, with validation Recall@20 used for model selection.

Alignment co-moves with ranking during training. Pooling epoch-wise checkpoints from the CSA grid, the coarse-code cosine q_0-cos increases together with Recall@20: the Pearson correlation is 0.56 on Baby (3743 epoch points) and 0.75 on Sports (5140 points). Binning q_0-cos in width 0.05 yields a monotone rise in mean validation Recall@20 on both datasets. Thus, as the model improves, the same-item coarse codes become more aligned—q_0 is an observable of this process.

Selected-dose trajectories confirm co-movement and a late peak. Fig. 3 plots q_0-cos and validation Recall@20 along training at the validation-selected CSA dose, as a three-seed mean with a ± 1 sample-standard-deviation band. The two curves rise together until the mean validation-best epoch (dashed). After that checkpoint, validation Recall@20 eases off while q_0-cos stays high or continues to increase: further coarse alignment past the validation peak does not improve ranking. On Baby, q_0-cos dips in the first few epochs while ranking still rises, then both increase; this early transient is absent on Sports and Electronics.

5 Conclusion

In this paper, we presented CSA-RIT, which repairs the two inputs of neighborhood averaging over modality graphs, namely the features that are averaged and the adjacency that selects them. Coarse Semantic Alignment aligns an item's modalities in a coarse quantized code before propagation, so that ranking co-moves with the code until a validation peak. Robust Interest Tree adds a short co-occurrence tree to each frozen modality graph, selected by user co-occurrence rather than by the similarity ranking, and it adds no trainable parameters. Under injected structural noise, its Recall@20 on Baby stays within about 1% of its clean value while the backbone and the content-side module degrade several times more. The full model is also best on every metric on Baby, Sports, and large-scale Electronics. These results suggest that repairing both inputs of the averaging operator is what makes frozen-graph recommenders robust, and that robustness is better measured under injected noise than inferred from clean-benchmark accuracy.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing of this manuscript, including grammar correction, word choice, and paragraph-level editing. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate or assist in proving mathematical claims, propose or refine hypotheses, design or give feedback on research methodology or experiments, implement methods, clean or reformat datasets, or interpret results. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts.

ETHICS STATEMENT

This work studies recommendation on public, de-identified benchmark datasets and raises no human-subject or privacy concerns. We do not foresee direct negative societal impacts beyond those inherent to recommender systems.

REPRODUCIBILITY STATEMENT

The main paper specifies the backbone, the two modules, and all hyperparameters used for the reported results. The implementation and configuration files are released as anonymized supplementary material.

REFERENCES

Anthony J Bell and Terrence J Sejnowski. The "independent components" of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997.

Chen Gao, Xiang Wang, Xiangnan He, and Yong Li. Graph neural networks for recommender system. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pp. 1623–1625, 2022.

Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 639–648, 2020.

Yue He, Hongbo Chen, Jingxi Xie, Fengling Li, and Jingjing Li. Same frequency begets shared interests: Popular-niche wavelet graph learning for multimodal recommendation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2026.

Christian Jutten and Jeanny Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10, 1991.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.

Jiuqiang Li and Hongjun Wang. Multi-view semantic contrastive alignment for multimodal recommendation. In *Proceedings of the ACM Web Conference*, 2026.

Yahong Lian, Haotian Tian, Chunyao Song, and Tingjian Ge. Sign-aware multimodal graph recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Guojiao Lin, Meng Zhen, Dongjie Wang, Qingqing Long, Yuanchun Zhou, and Meng Xiao. Gume: Graphs and user modalities enhancement for long-tail multimodal recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1400–1409, 2024.

Jialin Liu, Zhaorui Zhang, and Ray C. C. Cheung. Motorec: Sparse-regularized multimodal tokenization for cold-start recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

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

Yifan Liu, Kangning Zhang, Xiangyuan Ren, Yanhua Huang, Jiarui Jin, Yingjie Qin, Ruilong Su, Ruiwen Xu, Yong Yu, and Weinan Zhang. Alignrec: Aligning and training in multimodal recommendations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1503–1512, 2024b.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 43–52. ACM, 2015.

Fanshen Meng, Zhenhua Meng, Ru Jin, Yuli Chen, Rongheng Lin, and Budan Wu. Tamer: Interest tree augmented modality graph recommender for multimodal recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5998–6006, 2025.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations*, 2024.

Rongqing Kenneth Ong and Andy WH Khong. Spectrum-based modality representation fusion graph convolutional network for multimodal recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 773–781, 2025.

Xiangyu Sai, Meysam Madadi, Sergio Escalera, and Yong Xu. DIGEST: Dynamic graph refinement with dual contrastive semantic transfer for multimodal recommendation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2026.

Hongzu Su, Jingjing Li, Fengling Li, Ke Lu, and Lei Zhu. Soil: Contrastive second-order interest learning for multimodal recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5838–5846, 2024.

Gu Tang, Jinghe Wang, Xiaoying Gan, Bin Lu, Ze Zhao, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. R2mr: Review and rewrite modality for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1337–1348, 2025.

Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:5107–5116, 2022.

Maolin Wang, Yutian Xiao, Binhao Wang, Sheng Zhang, Shanshan Ye, Wanyu Wang, Hongzhi Yin, Ruocheng Guo, and Zenglin Xu. Findrec: Stein-guided entropic flow for multi-modal sequential recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3008–3018, 2025.

Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:1074–1084, 2021.

Suhang Wang, Yilin Wang, Jiliang Tang, Kai Shu, Suhas Ranganath, and Huan Liu. What your images reveal: Exploiting visual contents for point-of-interest recommendation. In *Proceedings of the 26th international conference on world wide web*, pp. 391–400, 2017.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pp. 1437–1445, 2019.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 3541–3549, 2020.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith CH Ngai. Mentor: multi-level self-supervised learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12908–12917, 2025a.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Wei Wang, Xiping Hu, Steven Hoi, and Edith Ngai. A survey on multimodal recommender systems: Recent advances and future directions. *arXiv preprint arXiv:2502.15711*, 2025b.

Jinfeng Xu, Zheyu Chen, Ziyue Peng, Shuo Yang, Jinze Li, Zewei Liu, Shujie Li, Yipeng Du, and Edith C. H. Ngai. One graph, multiple gains: Single high-quality item-item graph for multimodal recommendation. In *Proceedings of the ACM International Conference on Multimedia*, 2026.

Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 6576–6585, 2023.

Xiaohan Zhan, Yuliang Shi, Jihu Wang, Shijun Liu, Fanyu Kong, and Zhiyong Chen. Dhmrec: Collaboration-guided multimodal disentanglement and hierarchical fusion for recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Xiangyu Zhao, Yan Gao, Yao Hu, and Enhong Chen. Notellm-2: Multimodal large representation models for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2815–2826, 2025.

Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3872–3880, 2021.

Lingzi Zhang, Xin Zhou, Zhiwei Zeng, and Zhiqi Shen. Are id embeddings necessary? whitening pre-trained text embeddings for effective sequential recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 530–543. IEEE, 2024.

Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. *arXiv preprint arXiv:2301.12097*, 2023a.

Xin Zhou and Zhiqi Shen. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 935–943, 2023.

Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM web conference 2023*, pp. 845–854, 2023b.

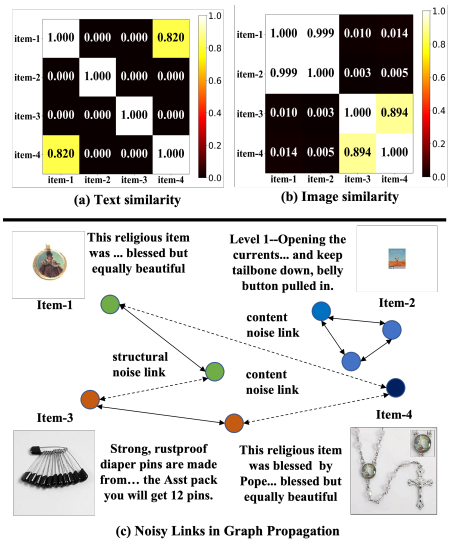


Figure 4: Motivation. (a) Text similarity and (b) image similarity between four sampled items; the two views disagree on which pairs are similar. (c) Noisy links in graph propagation, where each node carries the item's image and text description and links are marked by the noise they represent.

Algorithm 1 Order-aware beam search

Input: sparsified co-occurrence graph $\dot{\mathbf{S}}^c$; root r; order n; beam width b
Output: tree nodes T_r

$\mathcal{V} \leftarrow \{r\}$, $F \leftarrow \{(r, 1)\}$
for $o = 1$ to n **do**
  $\mathcal{C} \leftarrow \emptyset$
  for all $(u, s) \in F$ **do**
    for all $v \notin \mathcal{V}$ with $\dot{S}^c_{uv} > 0$ **do**
      if $v \notin \mathcal{C}$ **then**
        $\sigma(v) \leftarrow s \cdot \dot{S}^c_{uv}$; $\mathcal{C} \leftarrow \mathcal{C} \cup \{v\}$
      else
        $\sigma(v) \leftarrow \max\big(\sigma(v),\, s \cdot \dot{S}^c_{uv}\big)$
      end if
    end for
  end for
  if $o = 1$ **then**
    $N_r^{(1)} \leftarrow \mathcal{C}$
  else
    $N_r^{(o)} \leftarrow \text{Top-}b(\mathcal{C}, \sigma)$
  end if
  $\mathcal{V} \leftarrow \mathcal{V} \cup N_r^{(o)}$
  $F \leftarrow \big\{(v, \sigma(v)) : v \in N_r^{(o)}\big\}$
end for
return $T_r \leftarrow \bigcup_{o=1}^{n} N_r^{(o)}$

A Appendix

A.1 Motivation figure

A.2 Order-aware beam search