

RoFiGR: Coarse Routing, Fine Refinement via Level-Adaptive Distillation in Multimodal Generative Recommendation

Ji Dai², Zhipeng Wei¹, Zhuang Zhuang¹, Rongfeng Guo¹, Jie Chen¹, Quan Fang^{2*}, Fei Pan^{1*}

¹Kuaishou Technology, Beijing, China

²Beijing University of Posts and Telecommunications, Beijing, China

daiji@bupt.edu.cn, weizhipeng@kuaishou.com, zhuangzhuang@kuaishou.com, 2023096062@email.szu.edu.cn, chenjie20@kuaishou.com, qfang@bupt.edu.cn, panfei05@kuaishou.com

Abstract

Generative recommendation (GR) represents each item as a multi-level semantic identifier (SID) and predicts the next item through a coarse-to-fine hierarchy. Empirical analysis reveals pronounced first-SID difficulty: the first SID token lacks a narrowing prefix and is substantially more difficult to predict than subsequent levels. This difficulty reflects a broader coarse-routing gap across L1–L2. Existing methods nevertheless train all SID levels with the same next-code objective, while content-centric SID construction and hard-target supervision provide limited explicit behavioral guidance, a limitation further amplified in multimodal GR with separate textual and visual SID trees. To address this gap, we propose **RoFiGR**, which enriches the behavioral information available for coarse routing while reducing its access cost. We first redesign the SID hierarchy through Collaborative Hierarchical SID Learning, aligning coarse partitions with preference-aware geometry while preserving fine-grained content residuals. We then introduce Structure-Aware Multimodal Generation to expose item and source relations in flattened histories, enabling more effective behavior-aware routing. Finally, Anchored Level-Adaptive Distillation transfers a frozen sequential teacher’s dynamic item preferences into each SID tree through graded supervision, combining stable coarse-level anchors with an adaptively allocated residual budget. This design jointly increases the effective behavioral information available to L1–L2 and reduces the cost of retrieving it from flattened multimodal histories. Experiments on three Amazon review domains show that RoFiGR ranks first in 13 of 15 settings and improves HR@10 by up to 15.31%. Component ablations support the contribution of each module, while level-wise diagnostics show that the full model improves the coarse decisions it targets.

Introduction

Sequential recommendation aims to predict a user’s next interaction from her behavioral history; discriminative approaches address this task by learning a sequence representation and scoring each candidate item against it (Kang and McAuley 2018; Sun et al. 2019). Generative recommendation (GR) instead reformulates next-item prediction as sequence generation: each item is assigned a multi-level discrete semantic identifier (SID), typically obtained by residual quantization of item content embeddings (Rajput et al. 2023; Zhuang et al. 2026; Team et al. 2026). As illustrated in Figure 2, the successive SID levels form a hierarchical partition

*Corresponding author.

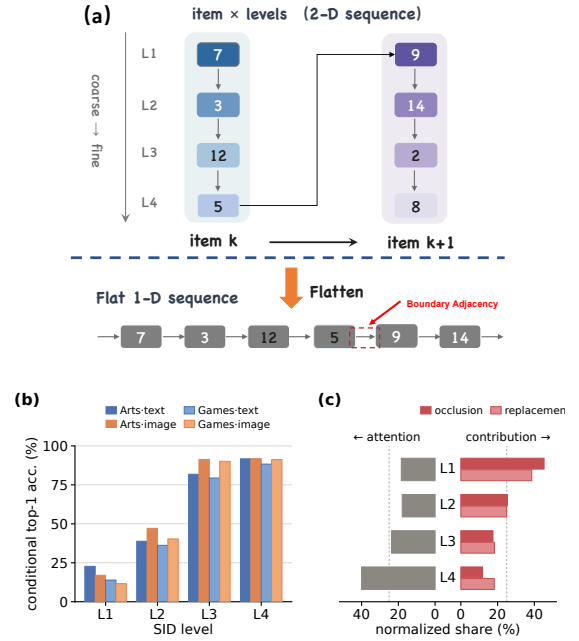


Figure 1: (a) Flattening the 2-D item $\times$ level SID structure into a 1-D token stream obscures item boundaries and source-level relations. (b) Conditional top-1 accuracy of MACRec on Arts and Games for the text and image SID trees; accuracy at level l is evaluated only when the preceding prefix $c_{<l}$ is correct. (c) MACRec diagnostics show attention concentrates on later SID levels, while prediction-change shares are larger at earlier levels.

of the item space, and autoregressively generating a target SID amounts to a coarse-to-fine search through this partition, with a complete SID uniquely identifying the target item. This frees the model from scoring the full candidate set; moreover, semantically similar items share SID prefixes, bringing generalization and cold-start benefits (Rajput et al. 2023; Singh et al. 2024). Meanwhile, as item information on modern platforms becomes increasingly multimedia, recent work constructs separate SID hierarchies for textual and visual content and models them jointly (Zhang et al. 2025a;

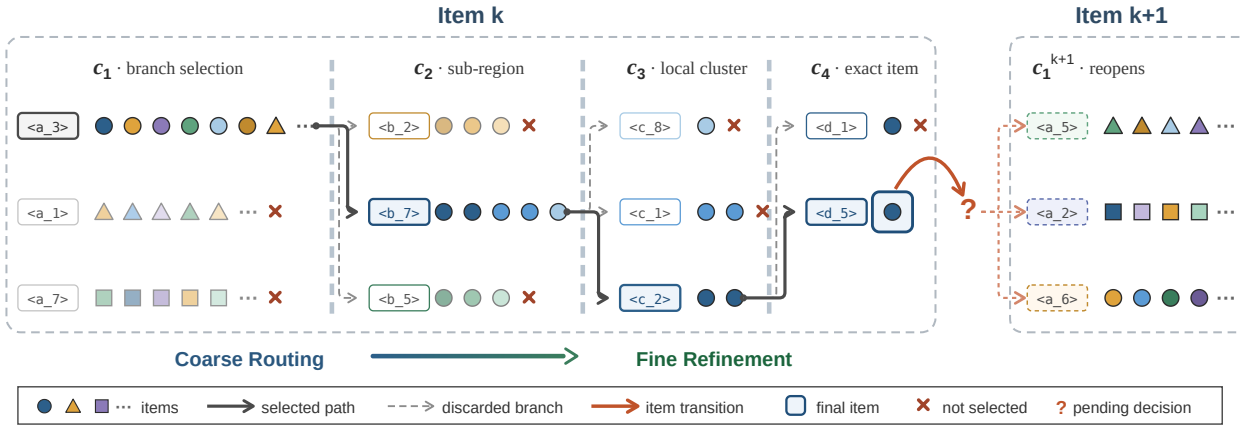


Figure 2: Hierarchical SID generation alternates between coarse routing and fine refinement: early codewords select a broad branch, later ones identify the item, and each new item restarts routing.

Liu et al. 2024; Zhai et al. 2025); our work operates in this multimodal setting. Representative implementations adopt a T5-style encoder-decoder architecture (Raffel et al. 2020; Rajput et al. 2023; Wang et al. 2024; Liu et al. 2025): the SIDs of successive historical items are concatenated into a flattened token sequence for the encoder, and the decoder autoregressively produces the target SID—in the multimodal case, the textual and visual identifier trees each perform this generation separately—with every SID position trained under the same token-level cross-entropy objective.

However, as Figure 2 illustrates, different SID levels involve fundamentally different prediction decisions. The first-level codeword determines the coarse region of the target item and relies primarily on behavioral transition patterns in the user history, whereas subsequent levels progressively focus on distinguishing among a much smaller set of similar items as the prefix becomes fixed and the feasible set rapidly shrinks. This coarse-to-fine distinction naturally arises from hierarchical residual quantization, which allows deeper levels to encode progressively finer content residuals, and is consistent with recent analyses of memorization-generalization behavior in generative recommendation (Ding et al. 2026; Singh et al. 2024). We call this the **routing-to-refinement transition**: L1 carries the primary routing decision, but L2 remains substantially harder than L3–L4 and still selects a broad prefix-constrained region; hence L1–L2 are routing levels, whereas L3–L4 refine within the routed region. The levels also differ sharply in prediction difficulty. On the MACRec backbone (Zhang et al. 2025a), L1 conditional top-1 accuracy remains below 20%, whereas L3–L4 reach 80–92% once the preceding prefix is correct across both textual and visual SID trees (Figure 1(b)). More than 90% of failed MACRec and SynGR generations make their first error at L1 (Table 4), indicating that the generation outcome is largely determined by the initial routing decision. Although recent methods reallocate information, optimization, or computation across SID levels (Wei et al. 2025; Xu et al. 2026; Yu et al. 2026; Cao et al. 2026), they still train all levels with the same hard-target objective, which does not explicitly expose

history-conditioned relative preferences among competing coarse branches.

The remaining difficulty first concerns the effective supply of behavioral guidance. At identifier construction, content reconstruction organizes branches by content similarity but does not explicitly encourage items with similar preference-aware profiles to share a coarse prefix. At generation, a hard target identifies the observed child under the current history, but it does not show how the other legal children compare. This distinction matters because a coarse-prefix probability aggregates the item probabilities inside its branch, allowing item-level teacher scores to be aggregated into a graded branch distribution that complements, rather than replaces, the hard target. Preference-aware geometry and teacher distributions therefore provide complementary forms of guidance: the former organizes coarse branches according to static behavioral structure, while the latter provides history-conditioned relative preferences among them. A separate limitation concerns access to the behavioral information already present in the history. Each item is serialized as a fixed sequence of codewords, but the standard token-distance bias does not explicitly encode whether two positions belong to the same item or which source level supplies a pairwise relation, causing within-item coarse-to-fine succession and cross-item boundary adjacency to receive the same positional treatment (Figure 1(a)). The diagnostics in Figure 1(c) show the resulting discrepancy: attention assigned to the most recent item concentrates on later SID levels, whereas occlusion and replacement assign larger prediction-change shares to earlier levels, especially L1. This mismatch suggests that the flattened sequence makes structurally relevant information costly to retrieve, even though fine codewords remain essential for item identity and within-region refinement.

Limited behavioral guidance and implicit relation structure together help explain the difficulty of coarse prediction. To close this **coarse-routing gap**, RoFiGR works in two complementary directions. Collaborative Hierarchical SID Learning aligns coarse prefixes with static preference-aware geometry, and Anchored Level-Adaptive Distillation

supplies history-conditioned relative branch preferences. Together, they increase the effective behavioral information available to routing. Structure-Aware Multimodal Generation instead exposes item-distance and source-level relations to encoder attention, making information already present in the flattened history easier to access.

Our contributions are summarized as follows. (1) **Insight:** we identify the **coarse-routing gap** in SID-based generative recommendation, showing that coarse decisions dominate prediction difficulty and account for most first errors. (2) **Method:** we propose RoFiGR, which combines Collaborative Hierarchical SID Learning, Structure-Aware Multimodal Generation, and Anchored Level-Adaptive Distillation to improve behavioral guidance and its structural readout. (3) **Evaluation:** on three Amazon review domains, RoFiGR ranks first in 13 of 15 settings. Component ablations establish contributions to final recommendation accuracy, while diagnostics on Arts show that the full model improves coarse-prefix accuracy and reduces the concentration of first errors at L1.

Reproducibility. Experimental code, execution traces, and selected training curves accompany the paper; reproducibility details are provided in the Appendix.

Related Work

Sequential Recommendation. Sequential recommenders predict the next interaction from historical behaviors, evolving from recurrent and convolutional encoders (Hidasi et al. 2016; Tang and Wang 2018a) to self-attentive architectures (Kang and McAuley 2018; Sun et al. 2019) enhanced with self-supervised objectives (Zhou et al. 2020; Qiu et al. 2022). Multimodal variants further exploit item content to enrich preference modeling (Bian et al. 2023; Hu et al. 2023; Wang et al. 2023; Zhang et al. 2025b). These discriminative models score the candidate set directly, producing dense item-level relative preferences that can be normalized into a transition distribution.

Generative Recommendation. TIGER (Rajput et al. 2023) established the two-stage paradigm that quantizes content embeddings into residual Semantic IDs (SIDs) and autoregressively generates the target identifier. Follow-ups improve identifier quality via collaborative, contrastive, or end-to-end tokenization (Wang et al. 2024; Zhu et al. 2024; Liu et al. 2025). Multimodal extensions broaden what SIDs encode (Liu et al. 2024; Zhai et al. 2025; Zhao et al. 2026): MACRec (Zhang et al. 2025a) aligns modalities at both tokenization and generation stages, and SynGR (Chen et al. 2026) promotes cross-modal synergy during generation. These studies enrich identifier content and cross-modal interaction, but do not explicitly provide history-conditioned graded supervision over the alternative branches at each level.

Hierarchical Structure of Semantic IDs. Since early hierarchical item indexing (Hua et al. 2023), residual quantization has endowed SIDs with an inherent coarse-to-fine decision hierarchy. SID-MLP (Guo et al. 2026a) finds that the first token faces the largest candidate space and dominates prediction difficulty, with later tokens reducing to local refinement once the prefix is fixed; How Well (Ding

et al. 2026) shows that prefix sharing enables compositional generalization but dilutes the memorization of specific transitions—exactly what discriminative models excel at (Wang 2026); Latte (Hou et al. 2026) proves that the fixed semantic tree couples the probabilities of prefix-sharing items regardless of user context. To address such level heterogeneity, GLASS (Cao et al. 2026) aggregates long-term interests for coarse tokens and retrieves local history for fine ones; Hi-SAM (Pan et al. 2026) assigns cross-modal consensus to coarse levels and modality-specific details to deep ones; APAO (Yu et al. 2026) reinforces vulnerable prefixes; other efforts reorganize level roles in the tokenizer (Wei et al. 2025; Wang et al. 2026; Liang et al. 2026) or allocate inference computation by position (Guo et al. 2026b; Yang et al. 2026). These approaches recognize level heterogeneity and redistribute existing information, optimization, or computation accordingly. RoFiGR instead combines behavior-aware coarse-prefix organization with history-conditioned graded branch supervision and explicit pairwise relations in the flattened history.

Methods

Preliminaries and Overview

Given a user history $\mathcal{H} = (i_1, \dots, i_n)$, the task is to predict the next item i_{n+1} . Each item i has a text or image SID $\mathbf{c}^m(i) = (c_1^m(i), \dots, c_L^m(i))$, where $m \in \{t, v\}$, $c_l^m(i) \in [K_l]$, and $L = 4$. An encoder-decoder p_ϕ consumes the serialized SIDs of the history and generates a target SID autoregressively. Its probability factorizes over the four levels as $p_\phi(\mathbf{c}^m | \mathcal{H}) = \prod_l p_\phi(c_l^m | \mathbf{c}_{<l}^m, \mathcal{H})$.

Let $\mathcal{I}^m(\mathbf{c}_{<l}^m)$ denote the items that share a prefix in tree m . A coarse code therefore selects a set of items, and its probability aggregates the item probabilities inside that branch:

$$P(\mathbf{c}_{<l}^m | \mathcal{H}) = \sum_{i \in \mathcal{I}^m(\mathbf{c}_{<l}^m)} P(i | \mathcal{H}). \quad (1)$$

A frozen discriminative recommender (Wang et al. 2023) provides a static item target $\mathbf{e}(i)$, formed by independently normalizing and equally fusing its ID, text, and image representations, and a history-conditioned score $s(i | \mathcal{H})$. These item targets are extracted once before SID training, and its history-conditioned scores are cached before generator training. Thus, the frozen recommender is not executed during student optimization or inference. Figure 3 summarizes the three components of RoFiGR. Collaborative Hierarchical SID Learning and Anchored Level-Adaptive Distillation provide static branch organization and history-conditioned relative preferences, respectively; together, they increase the effective behavioral information available to coarse routing. Structure-Aware Multimodal Generation instead makes item-distance and source-level relations explicit, facilitating access to behavioral evidence already present in the flattened history. The tokenizer and generator are trained in two separate stages, as detailed in the final subsection.

Collaborative Hierarchical SID Learning

We retain MACRec’s multimodal residual quantizer (Zhang et al. 2025a), including its text and image reconstruction,

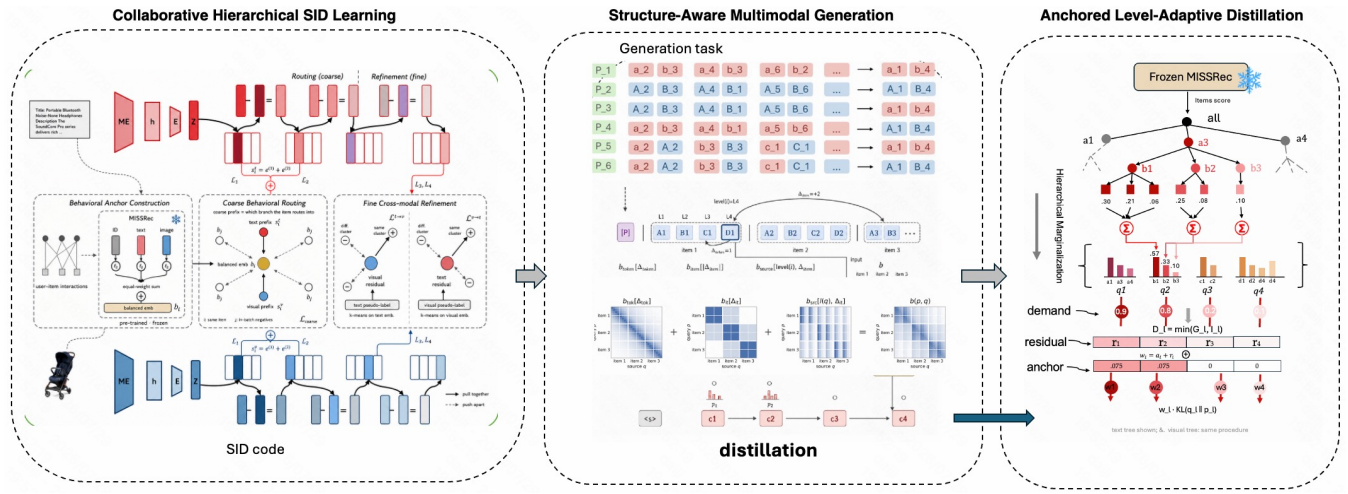


Figure 3: Overview of RoFiGR. **Left:** collaborative prefix alignment reorganizes L1–L2 of both SID trees, while the inherited cross-modal objective preserves fine-level refinement. **Middle:** eight generation tasks use flattened SID histories whose encoder attention is augmented with token-, item-distance-, and source-level biases. **Right:** teacher item scores are aggregated into level-wise branch distributions and distilled with anchored adaptive weights. Only one decoder stream is shown; the six history-conditioned tasks share the mechanism.

residual quantization, and late cross-modal interaction. Content reconstruction alone, however, does not explicitly encourage items with similar global preference profiles to share a coarse prefix. We therefore align the cumulative L1–L2 prefix of each tree with the balanced frozen MISSRec target $\mathbf{e}(i)$, which combines multimodal content with the global structure learned by the sequential recommender. For modality m , the prefix representation $\mathbf{z}_{\leq 2}^m(i)$ is the sum of the two selected code vectors. We set $u_{ij}^m = \text{sim}(g(\mathbf{z}_{\leq 2}^m(i)), \mathbf{e}(j))/\tau_c$, where g is a shared projector, τ_c is the alignment temperature, and u_{ij}^m is the positive pair.

For a minibatch $\mathcal{S}$, we use an in-batch contrastive objective (Chen et al. 2020) for tree m :

$$\mathcal{L}_{\text{coa}}^m = -\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \log \frac{\exp(u_{ii}^m)}{\sum_{j \in \mathcal{S}} \exp(u_{ij}^m)}. \quad (2)$$

The final $\mathcal{L}_{\text{coa}}$ averages the text- and image-tree objectives. The projector is shared by the two trees, whereas their codebooks remain separate. We apply the alignment once to the cumulative coarse prefix and leave L3–L4 unconstrained by it. Let $\mathcal{L}_{\text{base}}$ collect MACRec’s reconstruction, quantization, codebook, and fine-level cross-modal objectives. The tokenizer objective is $\mathcal{L}_{\text{SID}} = \mathcal{L}_{\text{base}} + \lambda_{\text{coa}} \mathcal{L}_{\text{coa}}$. After collision repair at the fine levels, the resulting trees preserve a unique complete SID for every item without changing their coarse partitions.

Structure-Aware Multimodal Generation

Multimodal training tasks. We retain MACRec’s three bidirectional task families: next-item generation, asymmetric item generation, and item-level identifier translation. Each family has a text-target and an image-target variant. The first two families contain a user history, whereas the two translation tasks map the text and image SIDs of a single item with-

out history. We add a fourth family in which every historical item supplies both complete identifiers. Its two prompts generate the next text SID and the next image SID separately. The generator thus has eight tasks in total, six of which are history-conditioned. All tasks use independent soft prompts and share the same encoder–decoder.

For a bimodal history, the four text tokens of an item are followed by its four image tokens, and all eight tokens share one item coordinate. This setup lets both trees read the same behavior sequence, but flattening still hides whether two encoder positions belong to the same item. Although level-specific vocabularies identify each token’s own level, the standard positional bias does not express the source level associated with a pairwise relation.

Structure-aware encoder. For query position p and source position q , let Δ_{tok} and Δ_{it} denote their signed token and item distances, respectively, and let $l(q)$ be the SID level of the source token. We clip item distance to $[-19, 19]$, which covers the 20-item history used in our experiments. In the first encoder self-attention layer, we augment the native token-distance bias with relation-aware coordinates (Shaw, Uszkoreit, and Vaswani 2018):

$$b(p, q) = b_{\text{tok}}[\Delta_{\text{tok}}] + b_{\text{it}}[\Delta_{\text{it}}] + b_{\text{src}}[l(q), \Delta_{\text{it}}]. \quad (3)$$

The two added tables are per-head parameters initialized to zero. The added biases apply only when both positions correspond to SID tokens and are set to zero for any pair involving a soft-prompt token. The item-distance term separates within-item relations from cross-item relations at the same token distance, while the source-level term distinguishes which level supplies the information. This modification makes the existing history structure explicit to encoder attention; it does not alter the decoder or introduce additional behavioral evidence.

Anchored Level-Adaptive Distillation

The generator’s one-hot target identifies the observed child but does not say how the remaining branches compare. Following knowledge and ranking distillation (Hinton, Vinyals, and Dean 2015; Tang and Wang 2018b; Kang et al. 2020), we obtain this missing information from the frozen teacher. Under the gold prefix $\mathbf{c}_{<l}^{m\star}$, let $\mathcal{I}_{l,k}^m$ contain the items assigned to legal child k , and let $\mathcal{C}_l^m$ denote all legal children. We aggregate the teacher’s item scores into the following softened branch distribution:

$$q_l^m(k) = \frac{\sum_{i \in \mathcal{I}_{l,k}^m} \exp(s(i | \mathcal{H})/\tau_t)}{\sum_{k' \in \mathcal{C}_l^m} \sum_{j \in \mathcal{I}_{l,k'}^m} \exp(s(j | \mathcal{H})/\tau_t)}. \quad (4)$$

This distribution has the same branch-aggregation form as Eq. (1), but is induced by discriminative scores at temperature τ_t rather than assumed to be an exact posterior. Text-target tasks use the text tree and image-target tasks use the image tree. The student distribution $p_{\phi,l}^m$ is restricted to the same legal children and is conditioned on the gold prefix during training.

The distillation weight combines a fixed routing anchor with a small adaptive residual. For gold child $k_l^\star$, we define the demand as the smaller of two positive log-probability margins. The first compares the teacher and student at the gold child; the second compares the teacher with the empirical branch-frequency prior $\pi_l^m(k_l^\star)$. This quantity is detached from the computation graph before updating its exponential moving average $\bar{D}_l^m$. The final level weight is

$$w_l^m = \mu_l + B \frac{(\bar{D}_l^m)^\gamma}{\sum_{j=1}^L (\bar{D}_j^m)^\gamma + \epsilon}. \quad (5)$$

Here μ_1 and μ_2 reserve stable supervision for routing, while $\mu_3 = \mu_4 = 0$; B is the residual budget. The exponent γ controls the concentration of the residual allocation, and ϵ ensures numerical stability. The two SID trees maintain separate running averages. The structural prior prevents a nearly collapsed child set from appearing informative, and the positive part prevents a negative demand from receiving weight. Distillation applies to the six history-conditioned tasks only; the two identifier-translation tasks remain supervised by their generation objectives.

Training Objectives and Inference

RoFiGR is trained in two stages. We first optimize $\mathcal{L}_{\text{SID}}$ and freeze both materialized SID trees. We then train the encoder-decoder with the eight generation tasks, the inherited item-level cross-modal objective, and branch distillation:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{gen}} + \lambda_{\text{cm}} \mathcal{L}_{\text{cm}} + \sum_{(d,m,l) \in \mathcal{D}} w_l^m \text{KL}(q_l^m \| p_{\phi,l}^m). \quad (6)$$

Here $\mathcal{L}_{\text{gen}}$ is the token-level cross-entropy over all generation tasks, $\mathcal{L}_{\text{cm}}$ is MACRec’s cross-modal item-alignment loss, and $d \in \mathcal{D}$ indexes valid history-conditioned task-tree-level tuples. Each KL term is averaged over the corresponding minibatch examples. The teacher remains frozen throughout training and is not used at inference, while temperatures,

loss weights, and moving-average parameters are tuned on the validation set.

At inference, trie-constrained beam search decodes four streams: text→text, image→image, bimodal→text, and bimodal→image. Valid SIDs are mapped to items; per-stream scores are normalized, averaged with equal weight, and deduplicated. The other four tasks are training-only, and the frozen teacher is absent at inference.

Experiments

Experimental setup. We evaluate RoFiGR on three domains from the Amazon review dataset (Ni, Li, and McAuley 2019)—Instruments, Arts, and Games—and compare it with representative discriminative, unimodal generative, and multimodal generative recommenders. We report HR@1/5/10 and NDCG@5/10. Hyperparameters and checkpoints are selected on the validation set; test data are used only for final reporting. Reported means use five independent runs with seeds {42, 43, 44, 45, 46}. For RoFiGR and the strongest corresponding baseline in each domain, the Appendix reports matched per-seed results, standard deviations, and two-sided paired t -tests, with $p < 0.05$ considered statistically significant. Inference uses trie-constrained beam search with beam size 20, and the four history-conditioned inference streams are combined by mean fusion. Full dataset statistics, pre-processing, baseline configurations, and training details are provided in Appendix A.

Overall Performance

Table 1 reports the overall performance on three datasets.

RoFiGR ranks first in 13 of the 15 settings. Its gains generally become more pronounced at larger cutoffs: on Instruments and Arts, the relative HR@10 improvements (6.43% and 15.31%) exceed the corresponding NDCG@10 gains (2.46% and 8.67%), whereas HR@1 decreases slightly by 2.88% and 0.46%. This pattern indicates that RoFiGR is particularly effective at keeping the relevant item in the retrieved candidate set, which is consistent with its emphasis on coarse routing. Importantly, NDCG improves in all six reported ranking settings, so the broader coverage does not come at the cost of overall ranking quality. Games shows the most uniform behavior, with clear improvements in both recall and ranking at every cutoff.

Ablation Study

We examine both the necessity of the individual components and the design choices underlying their hierarchical realization. Table 2 reports the results.

Core components. Among the core-component removals, disabling coarse-level collaborative alignment or Anchored Level-Adaptive Distillation produces the largest average drops (4.3% and 4.7%), showing that SID organization and branch-level teacher supervision provide complementary gains. The two coordinate biases contribute smaller but consistent improvements. In particular, the larger drop without the item-distance bias indicates that restoring item organization is their primary benefit, while the source-level bias

Table 1: Overall performance comparison on three datasets. The best results are shown in **bold**, and the second-best results are underlined. Δ denotes the relative improvement of RoFiGR over the strongest baseline.

Dataset	Metric	BERT4Rec	SASRec	MISSRec	P5-CID	VIP5	TIGER	MQL4GRec	MACRec	SynGR	RoFiGR	Δ (%)
Instruments	H@1	0.0450	0.0318	0.0723	0.0512	0.0737	0.0754	0.0763	0.0829	0.0897	<u>0.0872</u>	-2.88
	H@5	0.0856	0.0946	0.1089	0.0839	0.0892	0.1007	0.1058	0.1103	<u>0.1224</u>	0.1262	+3.14
	H@10	0.1081	0.1233	0.1361	0.1119	0.1071	0.1221	0.1291	0.1336	<u>0.1457</u>	0.1551	+6.43
	N@5	0.0667	0.0654	0.0797	0.0678	0.0815	0.0882	0.0902	0.0966	<u>0.1062</u>	0.1069	+0.66
	N@10	0.0739	0.0746	0.0880	0.0704	0.0872	0.0950	0.0997	0.1040	<u>0.1136</u>	0.1164	+2.46
Arts	H@1	0.0289	0.0212	0.0479	0.0421	0.0474	0.0532	0.0626	0.0631	0.0722	<u>0.0719</u>	-0.46
	H@5	0.0697	0.0951	0.1021	0.0713	0.0704	0.0894	0.0997	0.1011	<u>0.1073</u>	0.1193	+11.24
	H@10	0.0922	0.1250	0.1321	0.0994	0.0959	0.1167	0.1254	0.1290	<u>0.1348</u>	0.1555	+15.31
	N@5	0.0502	0.0610	0.0699	0.0607	0.0586	0.0718	0.0816	0.0823	<u>0.0900</u>	0.0959	+6.48
	N@10	0.0575	0.0706	0.0815	0.0662	0.0635	0.0806	0.0898	0.0913	<u>0.0989</u>	0.1075	+8.67
Games	H@1	0.0115	0.0069	0.0201	0.0169	0.0173	0.0166	0.0200	0.0190	<u>0.0209</u>	0.0224	+7.00
	H@5	0.0426	0.0587	<u>0.0674</u>	0.0532	0.0480	0.0523	0.0645	0.0668	<u>0.0623</u>	0.0746	+10.71
	H@10	0.0725	0.0985	0.1048	0.0824	0.0758	0.0857	0.1007	<u>0.1067</u>	0.1000	0.1202	+12.58
	N@5	0.0270	0.0333	0.0385	0.0331	0.0328	0.0345	0.0421	<u>0.0429</u>	0.0414	0.0484	+12.93
	N@10	0.0366	0.0461	0.0499	0.0454	0.0418	0.0453	0.0538	<u>0.0557</u>	0.0535	0.0630	+13.27

Table 2: Ablation study of RoFiGR. H and N denote HR@10 and NDCG@10; $\bar{\Delta}$ averages the relative change over the four reported metrics. Variants marked with $\dagger$ require re-materialized SID trees; full specifications are provided in the Appendix.

Variant	Arts		Games		Instr.	$\bar{\Delta}$
	H@10	N@10	N@10	N@10		
RoFiGR (full)	15.55	10.75	6.30	11.64	-	-
<i>Core component ablation</i>						
w/o coarse alignment $\dagger$	14.92	10.25	6.04	11.12	-4.3	
w/o item-distance bias	15.20	10.50	6.15	11.36	-2.3	
w/o source-level bias	15.39	10.66	6.25	11.52	-0.9	
w/o anchored distill.	14.70	10.18	6.06	11.16	-4.7	
w/o fixed anchor	14.78	10.16	5.99	11.03	-5.2	
w/o adaptive residual	15.42	10.67	6.24	11.51	-0.9	
<i>Design alternatives</i>						
align all levels $\dagger$	15.03	10.29	6.02	11.20	-4.0	
uniform KD weights	15.23	10.55	6.15	11.40	-2.1	
fine-level anchor	14.88	10.22	5.95	11.05	-5.0	
item-level KD	15.05	10.36	6.09	11.28	-3.3	
uniform teacher	14.61	10.02	5.90	10.96	-6.3	

adds a finer distinction among relations with identical token and item distances.

Hierarchical design choices. The fixed coarse-level anchor accounts for most of the distillation gain, whereas the adaptive residual supplies a smaller complementary improvement. Extending alignment to all levels, distributing the KD budget uniformly, or moving the anchor to the fine levels all perform worse than the proposed configuration. Per-level marginalization also outperforms item-level KL and a matched uniform teacher. Additional controls in Appendix F.1 indicate that access to the frozen teacher’s dynamic scores alone does not account for the full gain.

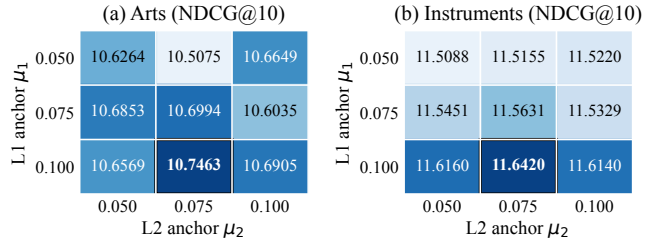


Figure 4: Sensitivity to the fixed coarse-level anchor weights on Arts and Instruments. Cells report NDCG@10 ($\times 100$); the outlined cell marks the adopted configuration, and colors are normalized within each dataset.

Together, these comparisons support stable coarse-routing supervision with a smaller level-dependent residual.

Hyperparameter Analysis

We analyze three hyperparameters that directly reflect the hierarchical design of RoFiGR: the fixed anchor allocation across the first two SID levels, the adaptive residual budget, and the capacity of the first-level codebook. In each analysis, the remaining model configuration is fixed to the final setting.

Coarse-level anchor allocation. Figure 4 varies the fixed L1 and L2 anchor weights jointly. Performance changes only mildly over the complete search region: the NDCG@10 ranges are approximately 2.2% on Arts and 1.1% on Instruments relative to their respective maxima. Nevertheless, both datasets show a consistent preference for stronger L1 anchoring, while an intermediate L2 weight performs slightly better than either boundary. We therefore adopt $(\mu_1, \mu_2) = (0.10, 0.075)$. Importantly, the surrounding configurations remain competitive, indicating that the method does not rely on a sharply tuned parameter point.

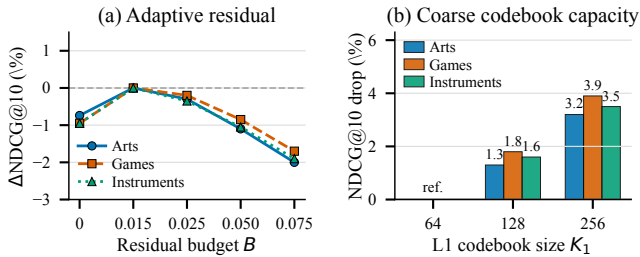


Figure 5: Sensitivity to the adaptive residual budget and the L1 codebook size. Panel (a) reports relative NDCG@10 with respect to $B = 0.015$; panel (b) reports the degradation from the $K_1 = 64$ reference.

Table 3: Single-modality inference performance; $\bar{\Delta}$ averages the relative gain over the four reported metrics.

Setting	Arts		Games	Instr.	$\bar{\Delta}$
	H@10	N@10	N@10	N@10	(%)
Text-only base	13.30	9.25	5.30	10.45	0.0
+ RoFiGR	14.15	9.85	5.66	10.93	+6.1
Image-only base	11.80	8.10	4.85	9.20	0.0
+ RoFiGR	12.55	8.55	5.15	9.70	+5.9

Residual budget. Figure 5(a) shows that the anchor-only setting ($B = 0$) is already competitive and that a small budget ($B = 0.015$) provides a consistent marginal gain across all three datasets. Larger budgets progressively reduce NDCG@10, supporting the residual as a data-dependent correction rather than the main source of distillation supervision.

Coarse-level codebook capacity. Figure 5(b) tests whether increasing the L1 codebook size alleviates the coarse-level bottleneck. Enlarging K_1 from 64 to 128 or 256 consistently reduces NDCG@10, with a larger drop at 256. Increasing coarse-level capacity is therefore not an effective remedy, motivating the compact $K_1 = 64$ design.

Single-Modality Analysis

Table 3 compares RoFiGR with modality-matched text-only and image-only bases. The gains persist across all three datasets, indicating that the overall RoFiGR design remains effective without multimodal inference inputs. The teacher is still multimodal during training, so its information enters the single-modality student only through supervision and introduces no additional inference cost.

Diagnostic Analysis

Table 4 reports coarse-level diagnostics on Arts, which exhibits the median overall gain among the three datasets. L1 Acc. measures first-level prediction accuracy, while L1:2 Acc. requires both coarse SID levels to be correct. Err.@L1 denotes the proportion of unsuccessful generations whose first error occurs at L1, and L1 Drop measures the decrease in HR@10 after masking the L1 token.

Both MACRec and SynGR exhibit limited coarse-prefix accuracy, with most failed generations making their first error

Table 4: Coarse-level diagnostic results on Arts for the text and image SID trees. All values are reported in percentage.

Method	L1 Acc.↑	L1:2 Acc.↑	Err. @L1↓	L1 Drop↑
<i>Text SID tree</i>				
MACRec	17.60	10.20	92.30	3.75
SynGR	18.70	11.15	90.80	4.10
RoFiGR	22.45	14.20	84.60	5.25
<i>Image SID tree</i>				
MACRec	16.80	9.60	93.10	3.45
SynGR	17.90	10.55	91.70	3.80
RoFiGR	21.70	13.55	85.80	4.95

Table 5: Stage-wise preparation and training time on Games (minutes). Teacher preparation includes training and score caching.

Method	Teacher Prep.	SID	Generator	Total
MACRec	—	34	226	260
RoFiGR	58	38	241	337

at L1. RoFiGR consistently improves L1 and L1–L2 accuracy on both SID trees and reduces the concentration of first errors at L1. Masking L1 also produces a larger HR@10 drop for RoFiGR, which is consistent with the coarse code making a greater functional contribution rather than merely becoming easier to predict.

Efficiency Analysis

On Games, RoFiGR increases total training time from 260 to 337 minutes, while generator training remains close to MACRec (241 vs. 226 minutes). Offline teacher preparation (training plus score caching) is a one-time, in-domain cost; full timing details are provided in the Appendix.

Conclusion

We identify a *coarse-routing gap* in SID-based generative recommendation: coarse levels require behavioral evidence over competing branches, yet hard targets do not reveal how alternative branches compare and flattened histories do not explicitly encode the pairwise relations needed to access existing behavioral information. RoFiGR addresses both limitations. Collaborative Hierarchical SID Learning and Anchored Level-Adaptive Distillation provide static and history-conditioned behavioral guidance, while Structure-Aware Multimodal Generation exposes item-distance and source-level relations. RoFiGR ranks first in 13 of 15 settings across three domains. Component ablations establish gains in final accuracy, and diagnostics on Arts show that the full model improves the coarse decisions it targets.

Branch distillation still depends on the quality of the frozen teacher’s item-level preferences. Future work can explore stronger or jointly updated teachers and test whether the routing-to-refinement asymmetry persists across other SID tokenizers, tree configurations, catalogs, and generative architectures.

References

- Bian, S.; Pan, X.; Zhao, W. X.; Wang, J.; Wang, C.; and Wen, J.-R. 2023. Multi-modal Mixture of Experts Representation Learning for Sequential Recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 110–119.
- Cao, S.; She, J.; Liu, J.; Zeng, B.; Guo, C.; Cai, K.; Luo, Q.; Tang, R.; Li, H.; Gai, K.; Li, Z.; and Yang, C. 2026. GLASS: A Generative Recommender for Long-sequence Modeling via SID-Tier and Semantic Search. *arXiv preprint arXiv:2602.05663*.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, 1597–1607.
- Chen, W.; Guo, X.; Li, S.; Zhang, F.; Yuan, M.; Fan, J.; Zhang, Z.; Wang, D.; and Zhuang, F. 2026. SynGR: Unleashing the Potential of Cross-Modal Synergy for Generative Recommendation. *arXiv preprint arXiv:2605.18920*.
- Ding, Y.; Guo, Z.; Li, J.; Peng, L.; Shao, S.; Shao, W.; Luo, X.; Simon, L.; Shang, J.; McAuley, J.; and Hou, Y. 2026. How Well Does Generative Recommendation Generalize? *arXiv preprint arXiv:2603.19809*.
- Guo, Z.; Hou, Y.; Ju, C. M.; Shah, N.; and McAuley, J. J. 2026a. MLPs are Efficient Distilled Generative Recommenders. *CoRR*, abs/2605.12617.
- Guo, Z.; Wang, J.; Zhou, R.; Liu, Y.; Guo, J.; Zhao, J.; Xu, X.; Liu, Y.; and Zhan, K. 2026b. S²GR: Stepwise Semantic-Guided Reasoning in Latent Space for Generative Recommendation. *CoRR*, abs/2601.18664.
- Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; and Tikk, D. 2016. Session-based Recommendations with Recurrent Neural Networks. In *International Conference on Learning Representations*.
- Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.
- Hou, Y.; Kim, H.; Ju, C. M.; Escoto, E.; Shah, N.; and McAuley, J. 2026. Expressiveness Limits of Autoregressive Semantic ID Generation in Generative Recommendation. *arXiv preprint arXiv:2605.06331*.
- Hu, H.; Guo, W.; Liu, Y.; and Kan, M.-Y. 2023. Adaptive Multi-Modalities Fusion in Sequential Recommendation Systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 843–853.
- Hua, W.; Xu, S.; Ge, Y.; and Zhang, Y. 2023. How to Index Item IDs for Recommendation Foundation Models. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 195–204.
- Kang, S.; Hwang, J.; Kweon, W.; and Yu, H. 2020. DE-RRD: A Knowledge Distillation Framework for Recommender System. *arXiv preprint arXiv:2012.04357*.
- Kang, W.-C.; and McAuley, J. J. 2018. Self-Attentive Sequential Recommendation. In *2018 IEEE International Conference on Data Mining*, 197–206.
- Liang, Y.; Zhang, Z.; Zhu, Y.; Zhang, K.; Guo, Z.; Zhou, W.; Yang, Z.; Wu, K.; Ni, Y.; Zeng, A.; Fu, C.; Wang, J.; and Xia, J. 2026. Rethinking Generative Recommender Tokenizer: Recsys-Native Encoding and Semantic Quantization Beyond LLMs. *arXiv preprint arXiv:2602.02338*.
- Liu, E.; Zheng, B.; Ling, C.; Hu, L.; Li, H.; and Zhao, W. X. 2025. Generative Recommender with End-to-End Learnable Item Tokenization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 729–739.
- Liu, H.; Wei, Y.; Song, X.; Guan, W.; Li, Y.-F.; and Nie, L. 2024. MMGR: Multimodal Generative Recommendation with Transformer Model. *arXiv preprint arXiv:2404.16555*.
- Ni, J.; Li, J.; and McAuley, J. 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*, 188–197.
- Pan, P.; Zhou, T.; Lu, P.; Fei, T.; Chen, H.; and Luo, C. 2026. Hi-SAM: A Hierarchical Structure-Aware Multimodal Framework for Large-Scale Recommendation. *CoRR*, abs/2602.11799.
- Qiu, R.; Huang, Z.; Yin, H.; and Wang, Z. 2022. Contrastive Learning for Representation Degeneration Problem in Sequential Recommendation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 813–823.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.
- Rajput, S.; Mehta, N.; Singh, A.; Keshavan, R. H.; Vu, T.; Heldt, L.; Hong, L.; Tay, Y.; Tran, V. Q.; Samost, J.; Kula, M.; Chi, E. H.; and Sathiamoorthy, M. 2023. Recommender Systems with Generative Retrieval. In *Advances in Neural Information Processing Systems*, volume 36.
- Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 464–468.
- Singh, A.; Vu, T.; Mehta, N.; Keshavan, R. H.; Sathiamoorthy, M.; Zheng, Y.; Hong, L.; Heldt, L.; Wei, L.; Tandon, D.; Chi, E. H.; and Yi, X. 2024. Better Generalization with Semantic IDs: A Case Study in Ranking for Recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 1039–1044.
- Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; and Jiang, P. 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.
- Tang, J.; and Wang, K. 2018a. Personalized Top-N Sequential Recommendation via Convolutional Sequence Embedding.

In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 565–573.

Tang, J.; and Wang, K. 2018b. Ranking Distillation: Learning Compact Ranking Models with High Performance for Recommender System. *arXiv preprint arXiv:1809.07428*.

Team, O.; Yang, B.; Ding, B.; Chu, C.; Zang, D.; Pan, F.; Li, H.; Jiang, H.; Bao, H.; Wang, H.; et al. 2026. OneReason Technical Report. *arXiv preprint arXiv:2606.06260*.

Wang, J.; Zeng, Z.; Wang, Y.; Wang, Y.; Lu, X.; Li, T.; Yuan, J.; Zhang, R.; Zheng, H.-T.; and Xia, S.-T. 2023. MIS-SRec: Pre-training and Transferring Multi-modal Interest-aware Sequence Representation for Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6548–6557.

Wang, S. 2026. Discrimination Is Generation: Unifying Ranking and Retrieval from a Tokenizer Perspective. *CoRR*, abs/2605.14853.

Wang, W.; Bao, H.; Lin, X.; Zhang, J.; Li, Y.; Feng, F.; Ng, S.-K.; and Chua, T.-S. 2024. Learnable Item Tokenization for Generative Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2400–2409.

Wang, Z.; Xu, L.; Deng, W.; Yan, H.; Liu, K.; and Chu, X. 2026. IntRR: A Framework for Integrating SID Redistribution and Length Reduction. *CoRR*, abs/2602.20704.

Wei, T.; Ning, X.; Chen, X.; Qiu, R.; Hou, Y.; Xie, Y.; Yang, S.; Hua, Z.; and He, J. 2025. CoFiRec: Coarse-to-Fine Tokenization for Generative Recommendation. *arXiv preprint arXiv:2511.22707*.

Xu, Z.; Chen, J.; Chen, S.; He, Y.; Yang, J.; Yuan, C.; Ding, K.; and Wang, C. 2026. Trie-Aware Transformers for Generative Recommendation. *arXiv preprint arXiv:2602.21677*.

Yang, S.; Gao, M.; Wang, Z.; and Yu, J. 2026. Where Reasoning Matters: Rethinking Latent Reasoning in Semantic ID-based Generative Recommendation. *arXiv preprint arXiv:2607.12425*.

Yu, Y.; Wang, Y.; Ma, W.; Guo, Z.; and Zhang, M. 2026. APAO: Adaptive Prefix-Aware Optimization for Generative Recommendation. *arXiv preprint arXiv:2603.02730*.

Zhai, J.; Mai, Z.-F.; Wang, C.-D.; Yang, F.; Zheng, X.; Li, H.; and Tian, Y. 2025. Multimodal Quantitative Language for Generative Recommendation. In *International Conference on Learning Representations*.

Zhang, F.; Liu, X.; Xi, D.; Yin, J.; Chen, H.; Yan, P.; Zhuang, F.; and Zhang, Z. 2025a. Multi-Aspect Cross-modal Quantization for Generative Recommendation. *arXiv preprint arXiv:2511.15122*.

Zhang, S.; Chen, L.; Shen, D.; Wang, C.; and Xiong, H. 2025b. Hierarchical Time-Aware Mixture of Experts for Multi-Modal Sequential Recommendation. In *Proceedings of the ACM Web Conference 2025*, 3672–3682.

Zhao, Y.; Shi, L.; Zhong, Y.; Kou, F.; Zhang, P.; Zhang, J.; Xu, M.; and Liu, Y. 2026. MusicRec: Multi-modal Semantic-Enhanced Identifier with Collaborative Signals for Generative Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 16459–16467.

Zhou, K.; Wang, H.; Zhao, W. X.; Zhu, Y.; Wang, S.; Zhang, F.; Wang, Z.; and Wen, J.-R. 2020. S³-Rec: Self-Supervised Learning for Sequential Recommendation with Mutual Information Maximization. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 1893–1902.

Zhu, J.; Jin, M.; Liu, Q.; Qiu, Z.; Dong, Z.; and Li, X. 2024. CoST: Contrastive Quantization based Semantic Tokenization for Generative Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 969–974.

Zhuang, Z.; Feng, S.; Qian, H.; Yang, M.; Qi, H.; Shen, Y.; and Yin, B. 2026. Think2Go: Generative Next POI Recommendation with LLM Reasoning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2112–2123.

Appendix A: Dataset, Baselines, and Evaluation Protocol

This appendix provides the dataset, baseline, and evaluation details supporting the experimental setup in the main text.

A.1 Datasets

We use the processed Amazon review data released with MQL4GRec, available at https://drive.google.com/drive/folders/1eewycbcAJ95atmF_V3bNchPIFDSw_TQC. We retain the Arts, Games, and Instruments domains and apply the chronological preprocessing protocol below.

Table A1: Statistics of the three Amazon review domains after preprocessing.

Dataset	#Users	#Items	#Inter.	Sparsity	Avg. len.
Arts	22,171	9,416	174,079	99.92%	7.85
Games	42,259	13,839	373,514	99.94%	8.84
Instruments	17,112	6,250	136,226	99.87%	7.96

We apply 5-core filtering and sort each user’s interactions chronologically. For every user, the last interaction is held out for testing, the penultimate interaction for validation, and all earlier interactions for training. The most recent 20 training items form the history; shorter histories are padded and padding positions are excluded from all position-dependent terms. Item text is encoded once with a frozen LLaMA text encoder, while the associated product image is encoded once with a frozen ViT image encoder.

A.2 Baselines

We compare against three discriminative sequential recommenders: BERT4Rec, SASRec, and the multimodal pre-trained MISSRec. P5-CID and VIP5 represent pretrained or language-model based recommendation, while TIGER and MQL4GRec are representative single-modality generative recommenders. MACRec and SynGR are the closest multimodal generative baselines. We use the authors’ released implementations where available and apply the same chronological split, full-catalog candidate set, and metric implementation to all methods. Reproduced baselines use released code and default or recommended hyperparameters; only data paths, random seeds, and minimal format adapters are changed. SynGR is trained separately on each target domain without the additional six-domain pretraining corpus used in its original study, keeping all reproduced generative models within the same target-domain data boundary.

A.3 Evaluation Protocol

We conduct full-catalog evaluation and report $HR@K$ and $NDCG@K$ for $K \in \{1, 5, 10\}$ as applicable. Reported means use five random seeds, 42–46. Appendix E gives matched per-seed statistics for RoFiGR and the strongest corresponding baseline in each domain, together with paired, two-sided t -tests under the same seeds; $p < 0.05$ is treated as statistically significant. Relative improvements are computed from unrounded means.

Appendix B: Detailed Architecture and Training

B.1 Overall Architecture

Experiments were executed on three servers, each equipped with four NVIDIA L20 GPUs with 48 GB of memory, providing 12 GPUs in total. Runs were independently scheduled across these servers according to their resource requirements. Our implementation is integrated into the released MACRec codebase and uses its software environment and package versions; the added RoFiGR modules introduce no new framework-level dependency.

RoFiGR follows a two-stage pipeline. Collaborative Hierarchical SID Learning first constructs separate textual and visual SID trees. The frozen trees are then consumed by a shared T5 encoder–decoder trained with the eight tasks in Table C1. During generator training, a frozen sequential teacher supplies cached item scores that are projected onto legal branches of the two SID trees. Inference contains only the tokenizer products, the generator, trie-constrained decoding, and fixed multi-stream fusion.

B.2 Tokenizer

We use *first-SID difficulty* for the empirical difficulty at L1, *coarse-routing gap* for the routing problem spanning L1–L2, and *refinement* for L3–L4. Each modality uses a four-level residual quantizer with codebook sizes [64, 256, 256, 512] and code dimension 32. The two modalities retain independent codebooks. For collaborative alignment, the cumulative L1–L2 code representation is projected into the space of a frozen, balanced item target and optimized with one contrastive objective. The text, image, and item components are normalized independently before equal-weight fusion, and the two modality-specific quantizers share the projection design but not the codebooks. Concretely, the shared projector is one bias-free linear map from the 32-dimensional cumulative code representation to the frozen item target dimension; no hidden layer or normalization module is added to it.

The inherited cross-modal residual interaction begins at L3, preserving modality-specific fine residuals. Complete SIDs are made one-to-one with items by repairing collisions first at L4 and, only when required, jointly at L3–L4; L1 and L2 are never modified by collision repair. The tokenizer is fully trained and frozen before generator optimization, so no generator gradient changes the SID trees.

B.3 Generator Architecture

The generator is a compact T5 encoder–decoder with four encoder layers, four decoder layers, hidden size 128, feed-forward size 1024, and six attention heads. All tasks share the encoder, decoder, and SID token embeddings, while each task has an independent soft prompt. Besides T5’s native token-distance relative bias, the first encoder self-attention layer receives two zero-initialized, per-head tables: a signed item-distance bias and a source-level-by-item-distance bias. The 20-item history bounds signed item offsets to $[-19, 19]$; the source-level table uses 11 signed coarse-distance buckets, with the outer buckets collecting distances of magnitude at

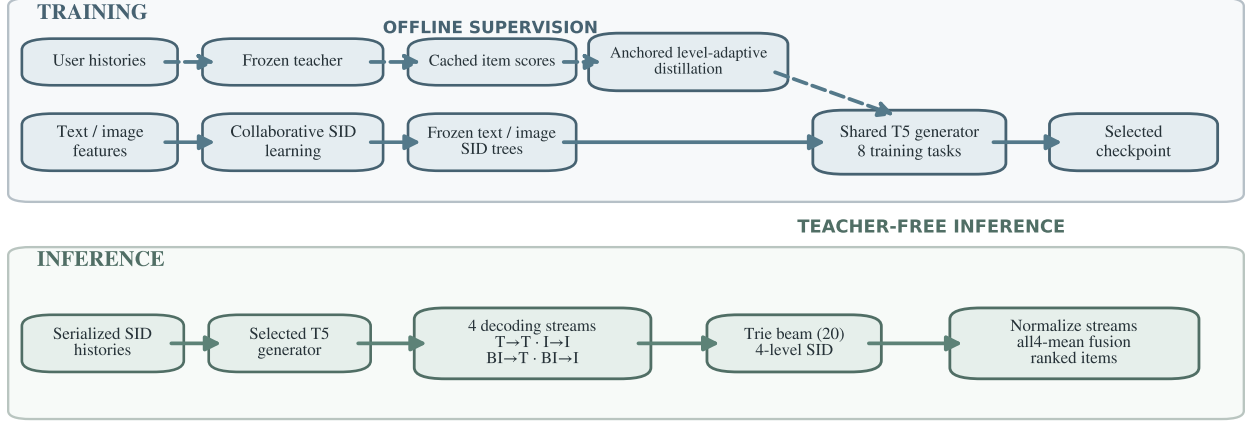


Figure B1: Training and inference workflow. SID learning and teacher-score caching precede generator training; the frozen teacher is absent at inference, where four history-conditioned streams are decoded and fused by all4-mean.

least nine. Bias entries involving soft prompts, separators, or padding are set to zero.

B.4 Anchored Level-Adaptive Distillation

Let $\mathcal{C}_l(u)$ be the valid children under prefix $u = c_{<l}$ and let $\mathcal{I}(u, c)$ denote the items whose SID extends u by child c . From the cached teacher score, we define $P_T(i | \mathcal{H}) \propto \exp(s(i | \mathcal{H})/\tau_t)$. Its level- l branch distribution is the exact marginal

$$q_l(c | u, \mathcal{H}) = \frac{\sum_{i \in \mathcal{I}(u, c)} P_T(i | \mathcal{H})}{\sum_{c' \in \mathcal{C}_l(u)} \sum_{i \in \mathcal{I}(u, c')} P_T(i | \mathcal{H})}. \quad (7)$$

The implementation computes the numerator with log-sum-exp and normalizes only over legal children. Structural branch frequency is

$$\pi_l(c | u) = \frac{N(u, c)}{\sum_{c' \in \mathcal{C}_l(u)} N(u, c')}, \quad (8)$$

where $N(u, c)$ is the number of catalog items under the child branch. This count-based null is used only by the adaptive controller and never modifies q_l . It prevents a small legal support from producing mechanically high teacher confidence that does not exceed the branch-count prior.

For the gold child c_l^* , let q_l^* , p_l^* , and π_l^* denote its teacher, student, and structural-null probabilities. The per-sample level demand used by the implementation is

$$D_l = \min\{[\log q_l^* - \log p_l^*]_+, [\log q_l^* - \log \pi_l^*]_+\}. \quad (9)$$

When $|\mathcal{C}_l(u)| = 1$, all three probabilities equal one and hence $D_l = 0$, avoiding residual budget for a deterministic decision. The detached, effective-sample-weighted minibatch demand updates a tree-specific running statistic,

$$\bar{D}_l^m \leftarrow \rho \bar{D}_l^m + (1 - \rho) D_l^m, \quad m \in \{\text{text}, \text{image}\}. \quad (10)$$

The residual and final KD weights are

$$r_l^m = B \frac{(\bar{D}_l^m)^\gamma}{\sum_{j=1}^4 (\bar{D}_j^m)^\gamma + \epsilon}, \quad w_l^m = \mu_l + r_l^m, \quad (11)$$

with $\mu = [0.10, 0.075, 0, 0]$. The loss is

$$\mathcal{L}_{\text{KD}}^m = \sum_{l=1}^4 w_l^m \text{KL}(q_l^m \| p_l^m). \quad (12)$$

Thus all four levels may receive distillation, but only L1–L2 have fixed anchors; L3–L4 receive weight only through the shared residual budget. The fixed anchors are active from the start of generator training, whereas the EMA-driven residual is enabled after a 1% controller warmup. The text and image trees maintain separate EMA controllers, and the same target tree shares its controller across three history-conditioned tasks.

B.5 Teacher Preparation

MISSRec is trained once within each domain. Its item embeddings are extracted before SID training, and its history-conditioned full-sort scores are computed and cached before generator training. SID learning and generator optimization therefore perform no teacher forward pass, and the teacher is absent at test time.

Appendix C: Generator Tasks and Inference Protocol

C.1 Generator Tasks

All eight tasks use the complete training set without task downsampling. A bimodal history serializes the four text tokens followed by the four image tokens for each item; all eight tokens share one item coordinate. The two same-item translation tasks retain MACRec’s item-level InfoNCE objective with $\lambda_{\text{cm}} = 0.01$. Text-history tasks serialize only the four text SID tokens per item, image-history tasks use only the four image SID tokens, and the two bimodal tasks concatenate both four-token SIDs.

Table C1: The eight generator tasks and their roles in training and inference. Shaded rows are retained at inference.

Task	History/input	Target tree	History-conditioned	KD	Inference
seqrec	text history	text	yes	yes	yes
seqimage	image history	image	yes	yes	yes
item2image	current text SID	image	no	no	no
image2item	current image SID	text	no	no	no
seqimage2item	image history	text	yes	yes	no
seqitem2image	text history	image	yes	yes	no
biseq2text	text+image history	text	yes	yes	yes
biseq2image	text+image history	image	yes	yes	yes

C.2 Inference Protocol

The reported protocol retains MACRec’s trie constraint and SID-to-item mapping while decoding four history-conditioned streams: seqrec, seqimage, biseq2text, and biseq2image. Each uses beam size 20 to generate a complete four-level SID. After mapping valid SIDs to items, scores are normalized per stream, averaged with equal weight (all4-mean), and deduplicated before ranking. The other four tasks, including the two asymmetric cross-modal history tasks, are training-only. Inference uses no teacher, collaborative prior, reranker, or test-time stream search.

Appendix D: Hyperparameters and Model Selection

The following settings support the hyperparameter analysis in the main text.

D.1 Search Space and Selection

Table D1: Hyperparameter search spaces and adopted values.

Parameter	Search space	Selected
μ_1	$\{.05, .075, .10\}$	.10
μ_2	$\{.05, .075, .10\}$	.075
B	$\{0, .015, .025, .05, .075\}$	.015
γ	$\{1, 1.5, 2\}$	1.5
K_1	$\{64, 128, 256\}$ (design)	64
SID depth L	4 (design)	4

The loss-allocation variables are selected by validation NDCG@10 with the data split, SID trees, teacher cache, decoding protocol, and evaluation code held fixed. K_1 and SID depth are reported as architectural comparisons rather than dataset-specific hyperparameters. One controller setting is shared by all three domains.

All tunable hyperparameters were selected on the validation set according to NDCG@10. The test set was used only for final reporting.

Both domains have broad high-performing regions, with maxima at the selected anchors. In a separate residual-budget sweep over $B = \{0, .015, .025, .05, .075\}$, relative validation changes are $\{-0.8, 0, -0.3, -1.0, -1.9\}\%$, supporting only a small adaptive correction.



Figure D1: Validation-set grid search of the fixed L1/L2 distillation anchors, with all remaining settings held constant. Cells report NDCG@10 ($\times 100$); the outlined cell is the adopted $(\mu_1, \mu_2) = (0.10, 0.075)$ setting, and shading is normalized within each domain.

D.2 Final Configuration

Table D2: Final RoFiGR configuration. Shared values span all three domains; shaded cells mark the domain-specific settings.

Stage	Parameter	Instr.	Arts	Games
Tokenizer	levels / codebooks	4/[64, 256, 256, 512]		
	code dimension	32		
	$\lambda_{\text{coa}} / \tau_c$	0.01 / 0.10		
	optimizer / learning rate	Adam / 10^{-3}		
Generator	optimizer / scheduler	AdamW / cosine		
	learning rate	10^{-3}	1.25×10^{-3}	1.25×10^{-3}
	batch / accumulation	1024 / 2		
	training steps	8,000	12,000	12,000
	weight decay / warmup	0.01/0.01		
	$\lambda_{\text{cm}} / \text{beam}$	0.01/20		
Distillation	τ_t	1.0		
	$(\mu_1, \mu_2, \mu_3, \mu_4)$	(0.10, 0.075, 0, 0)		
	$B/\gamma/\epsilon$	0.015/1.5/10 ⁻⁸		
	EMA decay ρ	0.99		

Tokenizer and distillation settings are shared across all domains. Domain-specific adjustments are limited to generator learning rate and training steps, reflecting dataset scale; the optimizer family, schedule, batch construction, beam size, and distillation controller remain fixed. This separation keeps the architecture common across domains while allowing each generator to converge under its validation-selected optimization budget.

Appendix E: Complete Five-Seed Results and Ablations

We provide the matched-seed statistics and complete ablations supporting the main results.

E.1 Evaluation Settings

We distinguish three result scopes. *Overall* results use final all4-mean fusion; *single-modality* results use independently trained restricted students; and *tree-specific* diagnostics read one output tree of the complete multimodal model. Tree-specific H@10 in Table 4 decodes only that output tree under the same trie constraint, without multi-stream fusion. In Table 3, the text-only student maps text histories to text SIDs and the image-only student analogously maps image histories to image SIDs. Each modality-matched base keeps that restricted I/O but disables the three RoFiGR additions. Comparisons are therefore made only within each modality block, not as a decomposition of Table 1’s all4-mean result. The multimodal teacher is used only during training. L1 Drop is the decrease in tree-specific H@10 after masking L1 evidence, a sensitivity measure rather than a standalone recommendation metric or causal estimate.

E.2 Overall Results

Table E1: Matched five-seed results for RoFiGR and the strongest corresponding baseline. All per-seed values, means, and standard deviations are reported on the $\times 100$ metric scale.

Metric Method	42	43	44	45	46	Mean $\pm$ std
<i>Instruments</i>						
H@10 SynGR	14.39	14.62	14.69	14.53	14.62	14.57 $\pm$.12
RoFiGR	15.29	15.67	15.48	15.62	15.49	15.51$\pm$.15
N@10 SynGR	11.23	11.44	11.34	11.48	11.31	11.36 $\pm$.10
RoFiGR	11.46	11.77	11.70	11.55	11.72	11.64$\pm$.13
<i>Arts</i>						
H@10 SynGR	13.57	13.32	13.52	13.61	13.38	13.48 $\pm$.12
RoFiGR	15.38	15.66	15.57	15.71	15.43	15.55$\pm$.14
N@10 SynGR	9.81	10.01	9.75	9.94	9.94	9.89 $\pm$.11
RoFiGR	10.88	10.66	10.81	10.59	10.81	10.75$\pm$.12
<i>Games</i>						
H@10 MACRec	10.56	10.81	10.61	10.76	10.61	10.67 $\pm$.11
RoFiGR	12.20	11.87	12.09	11.98	11.96	12.02$\pm$.13
N@10 MACRec	5.64	5.48	5.68	5.53	5.52	5.57 $\pm$.09
RoFiGR	6.18	6.38	6.45	6.23	6.26	6.30$\pm$.11

Exact paired two-sided p -values (H@10/N@10): Instruments $7.59 \times 10^{-5} / 9.63 \times 10^{-3}$; Arts $1.63 \times 10^{-5} / 7.55 \times 10^{-4}$; and Games $1.78 \times 10^{-4} / 2.31 \times 10^{-4}$.

Aggregate means over the five seeds reproduce the main comparison in Table 1; matched per-seed statistics for the strongest corresponding baseline in each domain are given in Table E1. Representative traces in the code package document execution and metric export. RoFiGR improves H@10 and N@10 over the strongest corresponding baseline on all three domains. Together with the H@1 results in Table 1, the pattern indicates that the larger benefit lies in retaining

relevant candidates through the beam rather than uniformly sharpening the first position.

E.3 Ablations

Table E2: Complete five-seed ablation statistics. Both the mean and standard deviation are reported on the $\times 100$ metric scale; no normalization is applied. * denotes a paired $p < 0.05$ comparison with the full model; † variants rematerialize the SID trees before retraining the generator.

Variant	Arts H@10	Arts N@10	Games N@10	Instr. N@10
RoFiGR (full)	15.55$\pm$0.14	10.75$\pm$0.12	6.30$\pm$0.11	11.64$\pm$0.13
w/o coarse alignment†	14.92 $\pm$ 0.23*	10.25 $\pm$ 0.16*	6.04 $\pm$ 0.10*	11.12 $\pm$ 0.17*
w/o item-distance bias	15.20 $\pm$ 0.20*	10.50 $\pm$ 0.14*	6.15 $\pm$ 0.09*	11.36 $\pm$ 0.15*
w/o source-level bias	15.39 $\pm$ 0.15	10.66 $\pm$ 0.10	6.25 $\pm$ 0.07	11.52 $\pm$ 0.11
w/o anchored distill.	14.70 $\pm$ 0.24*	10.18 $\pm$ 0.17*	6.06 $\pm$ 0.10*	11.16 $\pm$ 0.18*
w/o fixed anchor	14.78 $\pm$ 0.23*	10.16 $\pm$ 0.16*	5.99 $\pm$ 0.10*	11.03 $\pm$ 0.17*
w/o adaptive residual	15.42 $\pm$ 0.15	10.67 $\pm$ 0.11	6.24 $\pm$ 0.07	11.51 $\pm$ 0.11
align all levels†	15.03 $\pm$ 0.21*	10.29 $\pm$ 0.14*	6.02 $\pm$ 0.09*	11.20 $\pm$ 0.15*
uniform KD weights	15.23 $\pm$ 0.19*	10.55 $\pm$ 0.13*	6.15 $\pm$ 0.08*	11.40 $\pm$ 0.14*
fine-level anchor	14.88 $\pm$ 0.22*	10.22 $\pm$ 0.15*	5.95 $\pm$ 0.09*	11.05 $\pm$ 0.16*
item-level KD	15.05 $\pm$ 0.20*	10.36 $\pm$ 0.14*	6.09 $\pm$ 0.08*	11.28 $\pm$ 0.15*
uniform teacher	14.61 $\pm$ 0.26*	10.02 $\pm$ 0.18*	5.90 $\pm$ 0.11*	10.96 $\pm$ 0.19*

The component removals preserve the main-text pattern across domains. Coarse alignment and anchored distillation account for the largest individual changes, while the two co-ordinate tables provide smaller complementary gains. The design alternatives further show that extending alignment to every level, uniformly allocating KD, or moving the anchor to fine levels does not recover the full model. The † variants use newly trained tokenizers under the same data, architecture, and generator protocol; all other rows share the frozen full-model SID trees.

Variant specifications. *W/o coarse alignment* zeros its loss and rematerializes both SID trees; the two *w/o bias* variants delete only the named lookup table. *W/o anchored distillation* removes all branch KD, while *w/o fixed anchor* keeps only the residual and *w/o adaptive residual* only the anchor. *Align all levels* forms the collaborative representation from L1–L4 and rematerializes the trees. *Uniform KD weights* evenly divides the unchanged KD mass; *fine-level anchor* moves the fixed mass to L3–L4. *Item-level KD* applies one KL over cached teacher items using complete-SID likelihoods. *Uniform teacher* uses uniform legal-child targets and turns off the adaptive controller.

Scale of variation. Across the matched overall runs in Table E1, RoFiGR’s standard deviations range from 0.11 to 0.15 on the $\times 100$ scale, compared with 0.09 to 0.12 for the corresponding baselines. The absolute mean gains are 0.94/2.07/1.35 in H@10 and 0.28/0.86/0.73 in N@10 for Instruments/Arts/Games. Thus, the observed gains remain larger than ordinary run-to-run variation; the paired tests additionally exploit the matched seed design rather than treating the five runs as independent groups.

Appendix F: Extended Diagnostics and Teacher Controls

F.1 Teacher Controls

The following controls separate access to teacher information from the form of calibration. The first block gives MACRec the same frozen MISSRec teacher; the second retains RoFiGR and replaces only its dynamic branch scores.

Table F1: Teacher-control results ($\times 100$).

Setting	Arts H@10	Arts N@10	Games N@10	Instr. N@10
<i>Shared-teacher control</i>				
MACRec	12.90	9.13	5.57	10.40
MACRec + item-level KD	13.65	9.48	5.84	10.72
RoFiGR	15.55	10.75	6.30	11.64
<i>Dynamic branch-score teacher</i>				
No branch KD	14.70	10.18	6.06	11.16
SASRec scores	15.18	10.49	6.14	11.43
MISSRec scores	15.55	10.75	6.30	11.64

Teacher controls. MACRec with item-level KD recovers only part of the gain, while SASRec remains above no calibration and close to MISSRec. These controls indicate that access to the frozen teacher’s dynamic scores alone does not account for the full improvement.

F.2 Extended Diagnostics

Figure F1 shows low conditional L1 accuracy on both SID trees, while Table F2 shows that failed generations concentrate their first errors at L1. Together, these patterns are not specific to RoFiGR; the text-tree panel is reported first, followed by the image-tree panel.

Table F2: Tree-specific coarse diagnostics across domains ($\times 100$). Err. is the share of failed generations whose first error is at L1; Drop is the decrease in tree-specific H@10 after masking L1 evidence. This panel reports the text SID tree.

Text SID tree					
Dataset	Method	L1 $\uparrow$	L1:2 $\uparrow$	Err. $\downarrow$	Drop $\uparrow$
Arts	MACRec	17.60	10.20	92.30	3.75
	SynGR	18.70	11.15	90.80	4.10
	RoFiGR	22.45	14.20	84.60	5.25
Games	MACRec	14.78	8.52	93.83	3.22
	SynGR	15.47	9.08	92.38	3.47
	RoFiGR	20.12	12.23	85.64	4.83
Instr.	MACRec	19.17	11.28	91.53	3.88
	SynGR	20.42	12.13	89.94	4.23
	RoFiGR	24.28	15.38	83.82	5.38

Across all domains and both SID trees, RoFiGR improves L1 and L1–L2 accuracy, reduces the concentration of first errors at L1, and shows a larger drop when L1 evidence is masked, which is consistent with a greater functional contribution from coarse-prefix evidence. Relative to

SynGR, RoFiGR raises L1 accuracy by 3.75–4.65 points and lowers first-error concentration by 5.90–6.74 points across the six tree–domain pairs. These quantities are mechanism-aligned diagnostics rather than causal decompositions; coupled changes are expected because all three modules feed the same autoregressive prefix.

Table F2: Tree-specific coarse diagnostics across domains (*continued*): image SID tree.

Dataset	Method	L1 $\uparrow$	L1:2 $\uparrow$	Err. $\downarrow$	Drop $\uparrow$
Arts	MACRec	16.80	9.60	93.10	3.45
	SynGR	17.90	10.55	91.70	3.80
	RoFiGR	21.70	13.55	85.80	4.95
Games	MACRec	13.92	7.88	94.18	3.02
	SynGR	14.82	8.62	92.98	3.27
	RoFiGR	19.33	11.58	86.72	4.57
Instr.	MACRec	18.08	10.62	92.28	3.57
	SynGR	19.37	11.52	90.82	3.92
	RoFiGR	23.52	14.68	84.88	5.08

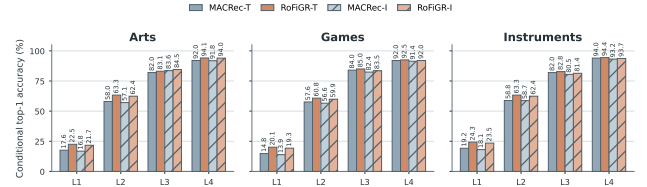


Figure F1: Conditional top-1 accuracy at L1–L4 for MACRec and RoFiGR on both SID trees across the three domains (%). Each level is evaluated only when the preceding prefix is correct; T and I denote text and image trees.

F.3 Efficiency

Total preparation and training time is 153/226/337 minutes for Instruments/Arts/Games. The respective components are teacher training (21/29/41 minutes), SID learning (22/31/38), generator training (102/155/241), and teacher-score caching (8/11/17). Teacher training and caching are one-time offline costs; inference keeps the same online model, stream count, and beam size.

Appendix G: Discussion and Limitations

Across domains and both SID trees, RoFiGR consistently improves L1 prediction, although absolute coarse-level accuracy remains moderate. In sequence-only recommendation, next-item purchases also depend on search intent, immediate context, and external stimuli absent from interaction history. The learned coarse distribution nevertheless retains sufficient probability mass for trie-constrained beam search, consistent with the HR@10 gains in Table 1. The remaining uncertainty motivates adaptive or stronger teacher signals, improved decoding, alternative tokenizers, and additional inputs such as search queries or contextual features. These extensions preserve the coarse-routing and fine-refinement interface rather than replacing it.