DEEP INTEREST MINING FOR INTENT-ENRICHED SEMANTIC IDS IN MULTIMODAL GENERATIVE RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic IDs (SIDs) define the discrete item vocabulary for generative recommendation, so the item semantics they explicitly expose are bounded by the evidence preserved before quantization. We introduce **DeepInterestGR**, an intent-enriched SID framework with three complementary components. **CMSA** preserves visual evidence through recommendation-oriented VLM captions and projected image embeddings; **DCIM** mines item-side intent descriptors grounded in product content rather than personalized user states; and **QARM** adds descriptor quality as an auxiliary policy-learning signal only when the generated SID decodes to the target item. Experiments on three primary Amazon Product Review categories show improvements over competitive generative and RL-based baselines, with relative gains of up to **15.1%** in NDCG@5 and **13.9%** in NDCG@10 over the strongest per-metric baseline. Controlled surface-only comparisons, direct SID-semantic diagnostics, multi-seed reward controls, and independent quality-label validation support a bounded conclusion: preserving visual and usage-context evidence before quantization, together with exact-target-gated descriptor-quality feedback, improves SID-based generative recommendation in the evaluated settings. The anonymized code repository is available at https://anonymous.4open.science/r/DeepInterestGR.

1 INTRODUCTION

Generative recommendation (GR) casts sequential recommendation as next-token prediction and directly generates the next item a user is likely to interact with (Li et al., 2024; Zhai et al., 2024; Wang et al., 2025b; Zhang et al., 2025; Zhou et al., 2025; Han et al., 2025). Compared with conventional multi-stage retrieval–ranking pipelines, this formulation provides a unified interface for user modeling and has shown strong empirical performance in e-commerce recommendation (Yi et al., 2025), search recommendation (Wang et al., 2025d), advertising (Zhang et al., 2025), and video recommendation. In SID-based GR, however, the semantic content directly exposed by the generated tokens is bounded by the discrete item vocabulary: if important item evidence is discarded before SID construction, it is not explicitly encoded in the SID semantics used during generation.

Semantic IDs (SIDs) are a key enabler of GR, mapping large item catalogs into compact sequences of discrete tokens (Hou et al., 2023). By compressing the item space while preserving compatibility with next-token prediction, SIDs improve the efficiency and scalability of GR (Rajput et al., 2023; Li et al., 2025). In multimodal scenarios, however, an SID is not only an index. It also defines what item semantics are exposed to the autoregressive model. If relevant visual evidence or usage intent is absent before quantization, it is not explicitly represented in the resulting code semantics, even though the generator may still learn item-level collaborative associations.

As shown in Figure 1, many SID generation methods encode item evidence into dense representations and then discretize it into token sequences (Rajput et al., 2023; Zhou et al., 2025; Zhang et al., 2025; Ye et al., 2025; Hou et al., 2025). Recent systems such as OneRec (Zhou et al., 2025), GPR (Zhang et al., 2025), and MiniOneRec (Kong et al., 2025) strengthen this paradigm through better generative training or more integrated optimization. Our goal is complementary: rather than claim-

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.

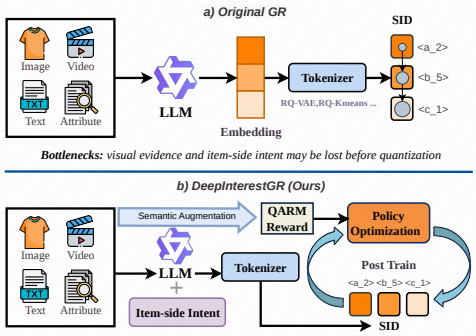


Figure 1: High-level comparison between standard SID-based generative recommendation and DeepInterestGR. Standard pipelines compress item content into semantic IDs before generation, which may discard visual evidence and item-side usage context. DeepInterestGR enriches the pre-quantization item representation with visual evidence and item-side intent descriptors, then applies exact-target-gated descriptor-quality feedback during policy optimization over generated SIDs.

ing that prior methods ignore semantics or are strictly two-stage, we study what semantic evidence is preserved in the pre-quantization representation that becomes the SID vocabulary.

Two Amazon Sports pairs make the problem concrete (Table 2). **Case 1** has similar "training" and "durable" soccer-ball text but different item-side contexts: youth recreation versus adult match preparation, exposing *surface ambiguity*. **Case 2** has short "soccer training ball" titles whose indoor-futsal versus outdoor-field distinction lies mainly in images, exposing *evidence omission*. If either distinction is lost before quantization, it is not explicit in the resulting SID semantics. The following Case 1 and Case 2 bottlenecks recur:

Weak visual preservation. Case 2's short text omits the indoor/outdoor distinction carried by its image. Captions may miss fine details, while image embeddings may be hard for a text-oriented SID pipeline to interpret; omitting both before quantization can underrepresent recommendation-relevant usage scenes.

Surface-only item semantics. Case 1's near-identical attributes can collapse youth recreation and adult match preparation into nearby prefixes. We call these content-inferred contexts item-side intent descriptors, not personalized user interests; they can later support user-history matching.

Sparse policy feedback for semantic quality. SID-format and binary matching rewards mainly supervise target matching, not whether an SID's descriptor is specific and actionable rather than vague or hallucinated.

Based on these observations, we propose **DeepInterestGR**. The central goal is to enrich item representations before SID quantization so that generated SIDs preserve coarse product categories while separating fine-grained usage intents that are ambiguous in, or omitted from, surface text. **CMSA** addresses the evidence omission exposed by Case 2 through a dual-path design: a VLM caption path produces recommendation-oriented visual descriptions, while a visual embedding path projects image features into the same pre-quantization representation used by the SID tokenizer. We use the term alignment in a modest sense: CMSA coordinates visual text, item text, and image embeddings before RQ-VAE, rather than claiming to learn a new contrastive multimodal embedding space. **DCIM** addresses the surface ambiguity exposed by Case 1 by mining item-side intent descriptors from the textualized item content. Finally, **QARM** evaluates the reliability of these descriptors and adds their quality as an LLM-derived signal during RL training, gated by exact target-item decoding so that semantic quality cannot reward an incorrect prediction by itself.

Contributions. We evaluate DeepInterestGR on three primary Amazon Product Review categories, with compact category-level checks on two additional categories, and compare it with traditional sequential models, generative SID models, and RL/preference-optimized recommendation baselines. The main contributions are:

- We formulate a pre-quantization evidence view of multimodal SID construction: the semantic evidence explicitly exposed by a fixed SID vocabulary is limited to what its construction retains.
- We construct intent-enriched SIDs with two complementary modules: **DCIM** mines grounded item-side intent descriptors, while **CMSA** preserves both caption-level visual semantics and continuous image features before quantization.
- We design **QARM**, which couples exact target-item correctness with offline descriptor reliability during GRPO while granting no quality bonus to an incorrect decoded item.
- Controlled surface-only comparisons, direct SID diagnostics, multi-seed reward controls, and independent quality-label validation complement the overall recommendation results and delimit what each module contributes.

2 Related Work

SID construction. Semantic IDs (SIDs) represent items as compact discrete token sequences for scalable recommendation. Clustering, hashing, RQ-VAE, and RQ-KMeans variants construct hierarchical codes from item embeddings (Petrov & Macdonald, 2024; Hou et al., 2023; Wang et al., 2025c; Rajput et al., 2023; Zhang et al., 2025). Later generative systems use these codes as autoregressive targets (Wang et al., 2026b; Li et al., 2025). Existing work primarily improves quantization or downstream generation; our complementary focus is the evidence contained in the representation before it is quantized.

Multimodal item representation. Multimodal recommenders combine textual metadata with visual evidence, and recent generative systems increasingly expose such features to SID construction (Zhou et al., 2025; Zhang et al., 2025; Ye et al., 2025). CMSA does not introduce a new contrastive space. It retains a MiniOneRec-compatible tokenizer while supplying both VLM-generated visual text and projected image embeddings before quantization.

LLM-derived item semantics. LLMs have been used as offline feature generators and semantic signal providers for recommendation (Chen et al., 2024; Yan et al., 2025; Yi et al., 2025). Such signals can enrich sparse metadata but may also be generic or unsupported. DCIM therefore operationalizes item-side usage context as content-grounded descriptors rather than claiming access to a user's latent psychological state; the descriptors annotate items and are subsequently matched through user-history modeling.

Policy optimization for generative recommendation. Generative recommendation casts user modeling and item prediction as next-token generation (Li et al., 2024; Wang et al., 2025a; 2026a; Zhai et al., 2024; Han et al., 2025; Zhou et al., 2025; Zeng et al., 2026). RL and preference optimization can further improve structured generation, but correctness and format rewards do not assess the reliability of item semantics. QARM adds descriptor reliability as a bounded, exact-target-gated signal over a fixed SID vocabulary; it neither reconstructs the codebook nor directly scores user–item preference.

3 Method

In this section, we present the **DeepInterestGR** framework. It enriches the item representation before SID quantization and then trains the generative policy with an exact-target-gated descriptor-quality signal. The two cases in Table 2 clarify the roles of its three components. For the short-text evidence omission in Case 2, **CMSA** (Section 3.2) injects image-derived evidence through both VLM-generated visual text and projected image embeddings before RQ-VAE quantization. For the surface ambiguity in Case 1, **DCIM** (Section 3.2) mines item-side intent descriptors from the textualized evidence so that similar titles can retain distinct usage contexts. **QARM** (Section 3.3) then assesses descriptor reliability during policy optimization without updating the constructed codebook. The data flow remains CMSA captioning, DCIM extraction, text/image fusion, RQ-VAE quantization, and finally SFT/GRPO with QARM. Captions, descriptors, embeddings, and quality labels are precomputed offline; online recommendation invokes only the trained SID generator over the constructed vocabulary, not these preprocessing models.

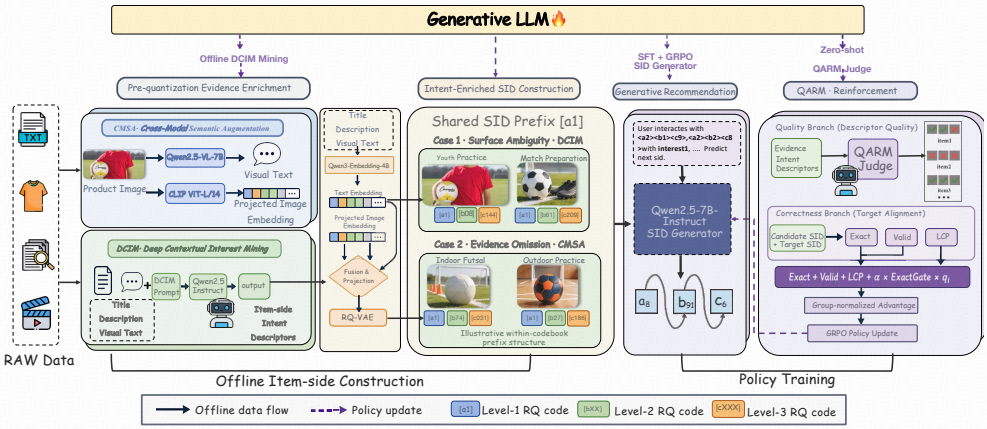


Figure 2: Architecture of DeepInterestGR. CMSA preserves caption-level and continuous visual evidence, while DCIM extracts item-side intent descriptors before text–image fusion and RQ-VAE quantization. The inset briefly illustrates how the two running cases are represented by hierarchical SIDs within a learned codebook. The SID generator is initialized with SFT and optimized using GRPO, where QARM provides an exact-target-gated descriptor-quality signal without modifying the constructed SID codebook.

3.1 Problem Formulation

We consider sequential recommendation: given user u's interaction history $\mathcal{S}_u = [i_1, \ldots, i_T]$, predict the next item i_{T+1}. Each item $i \in \mathcal{I}$ carries raw evidence $\mathbf{x}_i = (\mathbf{t}_i, \mathbf{d}_i, \mathbf{v}_i)$, which is encoded and quantized through the basic flow

$$\mathbf{x}_i \longrightarrow \mathbf{e}_i \longrightarrow \mathrm{SID}_i \,. \tag{1}$$

Under a fixed SID vocabulary, the downstream generator observes only the resulting codes; item evidence omitted from $\mathbf{e}_i$ before quantization is not explicitly encoded in, or recoverable from the SID semantics alone. We introduce an item-side intent descriptor $\mathbf{z}_i = \mathrm{DCIM}(\mathbf{t}_i, \mathbf{d}_i, \mathbf{v}_i^{\text{text}})$, where $\mathbf{v}_i^{\text{text}}$ is produced by the CMSA caption path. The descriptor is a grounded proxy for usage motivations implied by item content; it does not use future target information or personalize the item representation to a particular test user. Items are encoded via residual quantization (Rajput et al., 2023; Zheng et al., 2024): $\mathbf{s}_i = \text{RQ-VAE}(\mathbf{e}_i^{\text{mm}}) = (s_i^{(1)}, \ldots, s_i^{(H)})$, where $\mathbf{e}_i^{\text{mm}}$ fuses item attributes, visual textual evidence, projected image embeddings, and mined intent descriptors.

3.2 Intent-Enriched SID Construction

CMSA caption path. CMSA first converts each item's main image into recommendation-oriented visual text, making usage-scene and appearance evidence available to the text encoder and descriptor miner. This is one of two preservation paths, not a new contrastive representation space.

DCIM descriptor extraction. DCIM then mines grounded usage or purchase contexts from the title, description, and CMSA caption. The resulting descriptors remain item-side content proxies rather than personalized user states or verified psychological intent, and are concatenated with the source text before encoding.

Image fusion and residual quantization. In parallel, CMSA projects continuous image features and fuses them with the descriptor-enriched text embedding. A three-level RQ-VAE then quantizes this fused representation using the inherited MiniOneRec encoder–codebook–decoder design. If multiple items share all hierarchical codes, generation appends a unique collision-disambiguation token to preserve item identity; this token is excluded from LCP rewards and prefix-semantic diagnostics, which use only the hierarchical RQ codes. Full construction equations, prompts, and inherited tokenizer details appear in Appendix B.

3.3 Quality-Aware Reinforcement Mechanism (QARM)

To improve downstream generation over the constructed SIDs, we propose the **Quality-Aware Reinforcement Mechanism (QARM)**, as depicted in the right part of Figure 2. QARM uses supervised

fine-tuning (SFT) for initial alignment, precomputes descriptor-quality labels, and then applies reinforcement learning (RL) with an exact-target-gated descriptor-quality term. The RL stage optimizes the generative policy over SID sequences; it does not modify the VLM/LLM descriptor generator or retrain the RQ-VAE codebook.

Stage 1: Supervised Fine-Tuning. In the SFT stage, we train the generative model to predict target SID sequences given user histories represented by SIDs constructed from CMSA/DCIM-enriched item representations. The training objective minimizes the negative log-likelihood:

$$\mathcal{L}_{\text{SFT}} = -\sum_{(\mathbf{X},\mathbf{Y})\in\mathcal{D}} \sum_{h=1}^{L_i} \log p_\theta(y_h|\mathbf{X}, y_1, \ldots, y_{h-1}), \tag{2}$$

where $\mathcal{D}$ is the training dataset, $\mathbf{X}$ is the user-history SID sequence built from intent-enriched item representations, $\mathbf{Y}$ is the target SID sequence, and L_i includes the unique token only for a collision-disambiguated item. This stage aligns the model with collaborative patterns over the constructed SID vocabulary; the DCIM descriptors affect SFT through the item SIDs rather than as online text inputs.

Pre-RL QARM Binary Quality Labeling. Before RL training, we use Qwen2.5-7B-Instruct as a zero-shot binary classifier to assess each mined descriptor $z_i^{(j)}$ against its source item evidence:

$$l_i^{(j)} = \text{LLM}_{\text{cls}}(z_i^{(j)}, \mathbf{t}_i, \mathbf{d}_i, \mathbf{v}_i^{\text{text}}) \in \{0, 1\}, \tag{3}$$

where $l_i^{(j)} = 1$ indicates a *positive* descriptor (specific, actionable, and grounded in the supplied item content) and $l_i^{(j)} = 0$ indicates a *negative* descriptor (vague, generic, unsupported, or contradictory). When an item has multiple descriptors, we aggregate descriptor labels into an item-level quality score:

$$q_i = \frac{\sum_{j=1}^{J_i} w_i^{(j)} l_i^{(j)}}{\sum_{j=1}^{J_i} w_i^{(j)}}, \tag{4}$$

where J_i is the number of retained descriptors for item i, and $w_i^{(j)}$ is the confidence score returned by the DCIM prompt. If confidence is unavailable, we set all $w_i^{(j)} = 1$. Thus descriptor labels are binary, while the item-level $q_i \in [0, 1]$ can be continuous. The score is used only as an auxiliary signal during policy optimization. On 500 randomly sampled descriptor–item pairs, QARM agrees with adjudicated human labels with 0.844 accuracy, 0.842 Macro-F1, and Cohen's $\kappa = 0.684$; two human annotators achieve $\kappa = 0.762$. A non-Qwen Llama-3.1-8B-Instruct judge provides an independent-family check, with the full protocol and confusion matrices reported in Appendix E.1. These results validate a recognizable, though imperfect, descriptor-quality target rather than personalized relevance.

Quality-Aware Reward Function. The reward function combines standard recommendation rewards with an exact-target-gated descriptor-quality bonus. Given a generated SID sequence y and the ground-truth target item i^*, the exact-match reward is:

$$r_{\text{exact}}(y) = \mathbb{1}[\text{SID}(y) = \text{SID}(i^*)]. \tag{5}$$

Following prior RL baselines for SID generation, we also include SID-format rewards that provide partial credit when a candidate is syntactically valid or shares a correct prefix with the target SID:

$$r_{\text{format}}(y) = \lambda_{\text{valid}} \mathbb{1}[y \in \mathcal{S}] + \lambda_{\text{pref}} \frac{\text{LCP}(y, \text{SID}(i^*))}{H}, \tag{6}$$

where $\mathcal{S}$ is the valid SID vocabulary, $\text{LCP}(\cdot)$ is the longest common prefix over the $H = 3$ RQ codes, and the collision-disambiguation token is excluded. Exact+valid constitutes the standard reward, while adding LCP gives the strict no-QARM counterpart of the full method. These format rewards are standard stabilizers rather than QARM-specific contributions.

The QARM-specific term is the semantic-quality bonus. To avoid rewarding an irrelevant item only because its mined descriptor is fluent, the bonus is gated by exact target-item decoding:

$$r_{\text{quality}}(y) = \mathbb{1}[\hat{i}(y) = i^*] \cdot q_{\hat{i}(y)}, \tag{7}$$

where $\hat{i}(y)$ denotes the item decoded from SID sequence y, and $q_{\hat{i}(y)}$ is the aggregated QARM quality score. The final reward is:

$$r(y) = r_{\text{exact}}(y) + r_{\text{format}}(y) + \alpha \cdot r_{\text{quality}}(y), \tag{8}$$

Table 1: Overall performance comparison on three Amazon Product Reviews datasets. Bold indicates the best performance, and underline indicates the second best. Δ denotes the relative improvement of DeepInterestGR over the best baseline.

Method	Beauty				Sports				Instruments			
	HR@5	HR@10	N@5	N@10	HR@5	HR@10	N@5	N@10	HR@5	HR@10	N@5	N@10
Traditional Sequential Models												
GRU4Rec	0.0312	0.0518	0.0189	0.0256	0.0198	0.0324	0.0118	0.0161	0.0285	0.0467	0.0172	0.0231
Caser	0.0287	0.0483	0.0171	0.0234	0.0175	0.0291	0.0103	0.0142	0.0261	0.0432	0.0156	0.0212
HGN	0.0335	0.0549	0.0201	0.0271	0.0212	0.0348	0.0126	0.0172	0.0302	0.0495	0.0183	0.0246
Transformer-based Models												
SASRec	0.0398	0.0645	0.0241	0.0323	0.0248	0.0401	0.0149	0.0201	0.0361	0.0583	0.0218	0.0292
BERT4Rec	0.0421	0.0682	0.0256	0.0342	0.0267	0.0432	0.0161	0.0217	0.0385	0.0621	0.0233	0.0312
S^3-Rec	0.0445	0.0718	0.0271	0.0361	0.0283	0.0457	0.0171	0.0230	0.0407	0.0656	0.0247	0.0330
FDSA	0.0432	0.0698	0.0263	0.0351	0.0274	0.0443	0.0165	0.0223	0.0394	0.0636	0.0239	0.0320
Generative & LLM-based Models												
TIGER	0.0487	0.0763	0.0302	0.0395	0.0312	0.0498	0.0192	0.0256	0.0445	0.0712	0.0273	0.0362
LC-Rec	0.0523	0.0821	0.0328	0.0428	0.0341	0.0543	0.0212	0.0281	0.0478	0.0765	0.0295	0.0390
HSTU	0.0578	0.0897	0.0368	0.0476	0.0385	0.0611	0.0241	0.0318	0.0532	0.0845	0.0335	0.0438
MiniOneRec	0.0612	0.0945	0.0389	0.0502	0.0398	0.0627	0.0251	0.0330	0.0556	0.0878	0.0347	0.0455
BIGRec	0.0534	0.0839	0.0336	0.0439	0.0349	0.0556	0.0218	0.0289	0.0487	0.0779	0.0302	0.0399
D3	0.0498	0.0785	0.0312	0.0409	0.0325	0.0519	0.0202	0.0269	0.0456	0.0732	0.0282	0.0374
S-DPO	0.0589	0.0918	0.0372	0.0483	0.0385	0.0609	0.0242	0.0319	0.0537	0.0852	0.0334	0.0440
DeepInterestGR	**0.0678**	**0.1032**	**0.0436**	**0.0558**	**0.0452**	**0.0703**	**0.0289**	**0.0376**	**0.0623**	**0.0972**	**0.0394**	**0.0513**
Δ Improv.	+10.8%	+9.2%	+12.1%	+11.2%	+13.6%	+12.1%	+15.1%	+13.9%	+12.1%	+10.7%	+13.5%	+12.7%

where $\alpha = 0.5$ balances descriptor reliability with recommendation correctness. For a fixed query, q_{i^*} is constant, so QARM does not rank multiple descriptors or assign a bonus to an incorrect item. Instead, when an exact target co-occurs with candidates at different valid-prefix reward levels, the quality term changes the exact trajectory's group-relative margin. Across training instances, this allocates different learning emphasis according to offline descriptor reliability while leaving the descriptor generator and SID codebook fixed. Appendix E.3 directly audits when this change survives group normalization.

Stage 2: GRPO Optimization with QARM Rewards. We employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024) for efficient RL training. For each query, we sample a group of G SID candidates (*e.g.,* $\text{item}_1, \ldots, \text{item}_G$) and compute group-normalized advantages using the quality-aware rewards:

$$\hat{A}_j = \frac{r_j - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G) + \epsilon}, \tag{9}$$

where $\epsilon = 10^{-6}$ prevents division by zero when all sampled candidates receive identical sparse rewards. The policy update follows the standard clipped GRPO surrogate with KL regularization to prevent drift from the SFT initialization:

$$J(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{j=1}^{G} \min\big(\rho_j(\theta)\hat{A}_j, \bar{\rho}_j(\theta)\hat{A}_j\big) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\Bigg], \tag{10}$$

where $\rho_j(\theta)$ is the likelihood ratio to the behavior policy, $\bar{\rho}_j(\theta) = \text{clip}(\rho_j(\theta), 1-\varepsilon_c, 1+\varepsilon_c)$, ε_c is the GRPO clipping range, π_{ref} is the SFT reference policy, and β is the regularization coefficient. Descriptor quality is a training-only signal: it affects inference through the learned policy parameters, while inference itself receives histories of constructed SIDs rather than descriptor text. Thus CMSA and DCIM construct intent-enriched SIDs, whereas QARM modulates policy learning over this fixed vocabulary.

4 Experiments

We evaluate DeepInterestGR through four questions. **RQ1** compares overall performance with competitive baselines across benchmarks; **RQ2** isolates the contributions of DCIM, CMSA, and QARM; **RQ3** compares CMSA's caption and image-embedding paths with text-only representations; and **RQ4** compares exact-target-gated QARM with matched reward controls and tests whether its descriptor-quality labels and aggregated item-quality signal matter. Beyond aggregate ranking metrics, we include branch-level CMSA ablations and failure-oriented SID analysis to examine whether multimodal intent enrichment changes the generated token structure for visually similar but intent-different items, rather than only improving top-K accuracy.

4.1 EXPERIMENTAL SETUP

We conduct experiments on three primary real-world public datasets from Amazon Product Reviews (McAuley et al., 2015): **Beauty**, **Sports and Outdoors** (Sports), and **Musical Instruments** (Instruments). Following prior work (Rajput et al., 2023; Hou et al., 2023; Zhou et al., 2020), we apply 5-core filtering and leave-last-out evaluation. These categories were selected to cover appearance-, activity-, and specialized-preference-driven goods. They contain 22K–36K users, 10K–18K items, and 198K–296K interactions, with sparsity around 0.0004–0.0008. Industrial and Scientific and Toys and Games provide compact additional category checks under the same protocol; we do not claim full-Amazon coverage. Baselines span traditional sequential, Transformer-based, generative SID, and RL/preference-optimized recommenders. We report **HR@K** and **NDCG@K** ($K \in \{5, 10\}$) with beam size 20.

All experiments use NVIDIA A100 80GB GPUs. Qwen2.5-7B-Instruct is the DCIM model, QARM judge, and generative backbone; Qwen2.5-VL-7B-Instruct captions at most one main image per item; `openai/clip-vit-large-patch14` supplies continuous visual features; and Qwen3-Embedding-4B encodes item/descriptor text. SFT and GRPO use LoRA. SFT uses learning rate 3×10^{-4} for 3 epochs; GRPO uses 1×10^{-5} for 2 epochs, KL coefficient 0.001, group size $G = 8$, and $\epsilon = 10^{-6}$. Appendix F reports observed resource ranges, and the repository provides prompts, seeds, preprocessing, and final configurations.

4.2 FAILURE-ORIENTED CASE STUDY ON AMAZON SPORTS

We now revisit the two pre-quantization failures introduced in the Introduction. The first appendix item, Appendix A (Table 2), provides the full anonymized Amazon Sports examples that separate two questions. **Case 1** asks whether available surface text is sufficient to distinguish different item-side usage contexts when titles remain similar. **Case 2** asks whether the relevant context is present in the text at all when short titles omit distinctions carried mainly by images. The first motivates DCIM's contextual descriptors; the second motivates CMSA's preservation of visual evidence. The underlying item identifiers and reproduced case-study metadata are provided in the anonymized code repository; the product names and images in the paper are anonymized to preserve the failure patterns without exposing product identifiers.

Case 1: resolving surface ambiguity. The youth and adult products share the same high-level category and overlapping title tokens, so a surface-text representation emphasizes their common "soccer ball" and "training" attributes. Their item evidence nevertheless supports different usage contexts: beginner recreational practice versus competitive match preparation. DCIM makes this distinction explicit as item-side intent descriptors, rather than treating either descriptor as a personalized user state.

Case 2: recovering omitted visual context. The indoor and outdoor products have even shorter titles, which do not state the usage environment needed to distinguish them. CMSA exposes the indoor-court and outdoor-field evidence through both caption-level visual semantics and continuous image features; DCIM then organizes that evidence into interpretable usage-context descriptors before quantization.

Joint implication for SID construction. The two cases represent different failures—ambiguity in available surface text and omission of relevant evidence from that text—but lead to the same representational consequence. Within each independently trained codebook, the text-only pair shares a two-token prefix (*e.g.,* `a1,b33` or `a1,b42`), whereas the corresponding DeepInterestGR pair separates at the second level while retaining a common coarse prefix. Raw token identities are arbitrary across codebooks, so Table 2 illustrates within-space prefix structure rather than token alignment. These examples are qualitative illustrations; the controlled Surface-only comparison (Appendix D.1) and direct SID diagnostics across all three datasets (Appendix D.3) test the underlying pattern systematically.

4.3 MAIN RESULTS (RQ1)

Table 1 presents the overall performance comparison of DeepInterestGR against all baselines on the three primary Amazon Review datasets. We make the following observations:

DeepInterestGR achieves the best performance among compared baselines. DeepInterestGR outperforms all compared baselines across all three datasets and all evaluation metrics, achieving

relative improvements of 9.2%–15.1% over the strongest baseline for each metric in Table 1. The strongest baseline is usually MiniOneRec, while S-DPO is the strongest RL/preference-optimized baseline. For example, on the Sports dataset, DeepInterestGR achieves HR@5 of 0.0452, surpassing MiniOneRec by 13.6% and S-DPO by 17.4%. These consistent gains across Beauty, Sports, and Musical Instruments suggest that intent-enriched SID construction is beneficial in the evaluated product domains, although we do not claim full-category generalization beyond these datasets. Full pipeline comparisons and additional-category results are reported in Appendix C.

Quality-aware reinforcement learning outperforms preference optimization baselines. S-DPO, which applies direct preference optimization for recommendation, represents the strongest RL/preference-based baseline in our comparison. DeepInterestGR surpasses S-DPO by 15.1% in HR@5 on Beauty and 17.4% on Sports. Because the full systems also differ in pre-quantization representation and SID construction, this comparison establishes overall performance rather than isolating QARM. The matched strict no-QARM comparison in RQ4 isolates the contribution of the exact-target-gated descriptor-quality term.

4.4 Ablation Study (RQ2)

To understand the contribution of each core component, we conduct ablation studies by systematically removing individual modules from the full DeepInterestGR framework. Results on the Beauty dataset are presented in Table 3.

DCIM is the most critical novel component. Removing DCIM while retaining metadata and CMSA-derived visual evidence causes the largest single-component drop (-11.8% HR@5; -10.8% HR@10), supporting the hypothesis that instruction-tuned LLMs extract actionable item-side descriptors that distinguish similar surface attributes.

Table 3: Ablation study on Beauty. Each row removes one component from the full model.

Variant	HR@5	HR@10	N@5	N@10
DeepInterestGR (Full)	**0.0678**	**0.1032**	**0.0436**	**0.0558**
w/o DCIM	0.0598	0.0921	0.0378	0.0489
w/o CMSA (Text-Only)	0.0641	0.0983	0.0412	0.0529
w/o QARM (Exact + Valid + LCP)	0.0642	0.0981	0.0411	0.0528
w/o RL (SFT only)	0.0567	0.0879	0.0358	0.0463

DCIM beyond surface rewriting. Surface-only is an operational matched control: it uses the same Qwen2.5-7B-Instruct model, source evidence, output budget, downstream SID construction, and Exact+Valid+LCP reward as DCIM, but restates explicit attributes instead of extracting grounded usage context. In the descriptive single-seed full-test comparison, DCIM exceeds Surface-only, which in turn exceeds w/o DCIM, on every dataset and metric (Appendix D.1). Direct analysis of separately trained but architecturally identical SID spaces further shows higher usage-context and prefix consistency across all three datasets without a material pre-suffix collision increase (Appendix D.3). Complete rewriting and surface-similar hard-slice controls appear in Appendix D.

CMSA reduces visual-text modality mismatch. Removing CMSA causes a -5.5% drop in HR@5 and -4.7% in HR@10. This suggests that dual-path visual enrichment provides meaningful gains beyond text-only item representations in this setting. We avoid interpreting CMSA as a new contrastive representation-learning method; the result supports the narrower claim that combining VLM-generated visual text with projected image embeddings before quantization improves SID-based recommendation.

QARM adds a stable gain beyond matched format rewards. Removing only the quality term while retaining exact, validity, and LCP rewards decreases HR@5 from 0.0678 to 0.0642 (-5.3%) and HR@10 from 0.1032 to 0.0981 (-4.9%). The matched three-seed comparison in Appendix E.2 preserves this ordering. QARM does not choose among several descriptors for one correct target; it assigns training emphasis across instances according to the fixed target item's descriptor reliability, without rewarding an incorrect item.

Reinforcement learning yields the largest overall improvement. SFT-only drops -16.4% in HR@5 and -14.8% in HR@10, showing the importance of RL in this setup. QARM is only the descriptor-quality term, whose value depends on specific, grounded DCIM descriptors; exact-match and SID-format rewards remain separate standard signals.

4.5 CMSA Dual-Path Visual Enrichment Analysis (RQ3)

We investigate whether CMSA-based dual-path visual enrichment further improves over text-only representations. To isolate the two visual paths, we compare four variants under the same DCIM, QARM, RQ-VAE, and generative backbone settings: (1) *Text-Only* removes both visual paths; (2) *Caption-only* keeps only the VLM-generated visual text path; (3) *Image-embedding-only* keeps only the projected image-embedding path; and (4) *Caption + Image Embedding* uses the full CMSA module.

The two visual paths are complementary. Both individual paths improve over Text-Only: Caption-only raises HR@5 from 0.0641 to 0.0652 and NDCG@10 from 0.0529 to 0.0537, while Image-embedding-only reaches 0.0648 HR@5 and 0.0533 NDCG@10. Their combination is best on every metric (+5.8% HR@5 and +5.5% NDCG@10), supporting complementary caption semantics and continuous visual evidence before quantization.

Table 4: Fine-grained CMSA branch ablation on Beauty. All variants share DCIM, QARM, RQ-VAE, and generator settings.

Setting	HR@5	HR@10	N@5	N@10
Text-Only	0.0641	0.0983	0.0412	0.0529
Caption-only	0.0652	0.0995	0.0420	0.0537
Image-embedding-only	0.0648	0.0989	0.0417	0.0533
Caption + Image Embedding	**0.0678**	**0.1032**	**0.0436**	**0.0558**
Δ vs. Text-Only	+5.8%	+5.0%	+5.8%	+5.5%

4.6 QARM Reward Strategy and Quality Label Analysis (RQ4)

We compare the exact-target-gated QARM term with matched reward controls and test whether the item–quality association matters. Results are shown in Table 5. All rows use the same SFT initialization and GRPO recipe and differ only in their reward signal.

QARM adds information beyond format rewards. Strict no-QARM retains Exact, Valid, and LCP as the matched counterfactual. True q_i improves all metrics, but a fixed item-level permutation does not; stable three-seed means (Appendix E.2) attribute the gain to the item–quality association, not reward magnitude.

Table 5: Controlled QARM comparison on Beauty. Standard uses Exact+Valid; strict w/o QARM adds LCP; Full additionally adds the true q_i. The shuffled control preserves the marginal q_i distribution.

Setting	HR@5	HR@10	N@5	N@10
Standard (Exact + Valid)	0.0635	0.0973	0.0408	0.0526
Strict w/o QARM (+ LCP)	0.0642	0.0981	0.0411	0.0528
Strict shuffled q_i	0.0636	0.0974	0.0406	0.0522
Full QARM (true q_i)	**0.0678**	**0.1032**	**0.0436**	**0.0558**

The quality term survives group normalization in identifiable groups. In 10,000 groups (Appendix E.3), changes survive in 14.0% at 10^{-5} and 13.7% at 10^{-4}; active-group mean $|\Delta A| = 0.0370$, with exact-candidate mean $+0.094$. They concentrate with multiple LCP levels; two-level groups differ only numerically. Thus offline q_i changes policy updates without direct user–item supervision.

5 Limitations

Our category-level Amazon evidence is not universal. DCIM descriptors are content-grounded usage-context proxies, not verified psychological intent; quality depends on metadata, images, and pretrained models. Offline CMSA/DCIM/QARM and GRPO add computation. Over a fixed codebook, QARM neither supervises user–item preference nor restores omitted SID evidence. The full-test Surface-only control is descriptive and single-seed, not a significance claim.

6 Conclusion

We proposed **DeepInterestGR**: **DCIM** supplies grounded item descriptors, **CMSA** preserves visual evidence, and **QARM** adds exact-target-gated reliability. Evaluated Amazon results and controlled diagnostics support preserving pre-quantization evidence. This conclusion is bounded to these proxies and settings; broader domains and richer preference supervision remain future work.

AI USE STATEMENT

We did not use generative AI tools for tasks requiring disclosure under the ICLR 2027 AI Policy. We used generative AI tools only to edit author-written text for readability, including grammar, clarity, and style. The authors reviewed all AI-assisted edits and take responsibility for the final content of this work.

REFERENCES

Junyi Chen, Lu Chi, Bingyue Peng, and Zehuan Yuan. Hllm: Enhancing sequential recommendations via hierarchical large language models for item and user modeling. *arXiv preprint arXiv:2409.12740*, 2024.

Ruidong Han, Bin Yin, Shangyu Chen, He Jiang, Fei Jiang, Xiang Li, Chi Ma, Mincong Huang, Xiaoguang Li, Chunzhen Jing, et al. Mtgr: Industrial-scale generative recommendation framework in meituan. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5731–5738, 2025.

Yupeng Hou, Zhankui He, Julian McAuley, and Wayne Xin Zhao. Learning vector-quantized item representation for transferable sequential recommenders. In *Proceedings of the ACM Web Conference 2023*, pp. 1162–1171, 2023.

Yupeng Hou, Jiacheng Li, Ashley Shin, Jinsung Jeon, Abhishek Santhanam, Wei Shao, Kaveh Hassani, Ning Yao, and Julian McAuley. Generating long semantic ids in parallel for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 956–966, 2025.

Xiaoyu Kong, Leheng Sheng, Junfei Tan, Yuxin Chen, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. Minionerec: An open-source framework for scaling generative recommendation. *arXiv preprint arXiv:2510.24431*, 2025.

Xiaopeng Li, Bo Chen, Junda She, Shiteng Cao, You Wang, Qinlin Jia, Haiying He, Zheli Zhou, Zhao Liu, Ji Liu, et al. A survey of generative recommendation from a tri-decoupled perspective: Tokenization, architecture, and optimization. 2025.

Yongqi Li, Xinyu Lin, Wenjie Wang, Fuli Feng, Liang Pang, Wenjie Li, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. A survey of generative search and recommendation in the era of large language models. *arXiv preprint arXiv:2404.16924*, 2024.

Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, 2015.

Aleksandr V Petrov and Craig Macdonald. Recjpq: training large-catalogue sequential recommenders. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 538–547, 2024.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Hao Wang, Wei Guo, Luankang Zhang, Jin Yao Chin, Yufei Ye, Huifeng Guo, Yong Liu, Defu Lian, Ruiming Tang, and Enhong Chen. Generative large recommendation models: Emerging trends in llms for recommendation. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 49–52, 2025a.

Jinze Wang, Lu Zhang, Yiyang Cui, Tiehua Zhang, Zhishu Shen, Yuze Liu, Xingjun Ma, and Jiong Jin. Do we really need sft? prompt-as-policy over knowledge graphs for cold-start next poi recommendation. *arXiv preprint arXiv:2510.08012*, 2025b.

Jinze Wang, Tiehua Zhang, Lu Zhang, Yang Bai, Xin Li, and Jiong Jin. Hyperman: Hypergraph-enhanced meta-learning adaptive network for next poi recommendation. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025c.

Jinze Wang, Yangchen Zeng, Tiehua Zhang, Lu Zhang, Yuze Liu, Yongchao Liu, Xingjun Ma, and Zhu Sun. Agent4poi: Agentic context-conditioned affordance reasoning for multimodal point-of-interest recommendation. *arXiv preprint arXiv:2605.15203*, 2026a.

Jinze Wang, Yangchen Zeng, Tiehua Zhang, Lu Zhang, Yuze Liu, Zhishu Shen, Jiong Jin, and Zhu Sun. Meta-modal agent: Sequential evidence routing for missing-modality candidate reranking. *arXiv preprint arXiv:2605.25007*, 2026b.

Yejing Wang, Shengyu Zhou, Jinyu Lu, Ziwei Liu, Langming Liu, Maolin Wang, Wenlin Zhang, Feng Li, Wenbo Su, Pengjie Wang, et al. Nezha: A zero-sacrifice and hyperspeed decoding architecture for generative recommendations. *arXiv preprint arXiv:2511.18793*, 2025d.

Bencheng Yan, Shilei Liu, Zhiyuan Zeng, Zihao Wang, Yizhen Zhang, Yujin Yuan, Langming Liu, Jiaqi Liu, Di Wang, Wenbo Su, et al. Unlocking scaling law in industrial recommendation systems with a three-step paradigm based large user model. *arXiv preprint arXiv:2502.08309*, 2025.

Wencai Ye, Mingjie Sun, Shuhang Chen, Wenjin Wu, and Peng Jiang. Align3gr: Unified multi-level alignment for llm-based generative recommendation. *arXiv preprint arXiv:2511.11255*, 2025.

Chao Yi, Dian Chen, Gaoyang Guo, Jiakai Tang, Jian Wu, Jing Yu, Mao Zhang, Sunhao Dai, Wen Chen, Wenjun Yang, et al. Recgpt technical report. *arXiv preprint arXiv:2507.22879*, 2025.

Yangchen Zeng, Zhenyu Yu, Dongming Jiang, Wenbo Zhang, Yifan Hong, Zhanhua Hu, Jiao Luo, and Kangning Cui. Learning where to embed: Noise-aware positional embedding for query retrieval in small-object detection. *arXiv preprint arXiv:2604.15065*, 2026.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Jiayuan He, et al. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In *International Conference on Machine Learning*, pp. 58484–58509. PMLR, 2024.

Jun Zhang, Yi Li, Yue Liu, Changping Wang, Yuan Wang, Yuling Xiong, Xun Liu, Haiyang Wu, Qian Li, Enming Zhang, et al. Gpr: Towards a generative pre-trained one-model paradigm for large-scale advertising recommendation. *arXiv preprint arXiv:2511.10138*, 2025.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, and Ji-Rong Wen. Adapting large language models by integrating collaborative semantics for recommendation. In *ICDE*, 2024.

Guorui Zhou, Hengrui Hu, Hongtao Cheng, Huanjie Wang, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Lu Ren, Liao Yu, et al. Onerec-v2 technical report. *arXiv preprint arXiv:2508.20900*, 2025.

Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *CIKM*, 2020.

A Full Failure-Oriented Case Study

Table 2: Failure-oriented Amazon Sports examples. Text-only SID construction can group products by surface tokens, while dual-path CMSA and DCIM encode item-side usage-context evidence before quantization; item identifiers are anonymized.

Item	Surface Text	Visual Evidence / Text-only Failure	Descriptor and SID Effect
Case 1—Surface ambiguity: similar titles, different item-side usage contexts. Similar soccer-ball descriptions obscure youth recreational practice versus adult match preparation; DCIM makes this contextual distinction explicit before quantization.			
 *Youth / beginner ball*	**Title:** Classic Soccer Ball, Size 4, Blue/Yellow **Surface:** soccer ball, size 4, training, durable	Bright blue/yellow color; youth-oriented appearance; soft rubber texture; backyard or school-field scene. **Failure:** text-only SID may place it near generic training or match balls.	parent-purchased youth beginner soccer practice; school/recreational play. **QARM:** 1; **Conf.:** 0.89 **SIDs: Text:** `a1,b33,c192`; **Full:** `a1,b08,c144`
 *Adult / match ball*	**Title:** Classic Soccer Ball, Size 5, White/Black **Surface:** soccer ball, size 5, training, durable	Standard match-ball texture; stitched panels; grass/turf field; full-size goal and club-practice cues. **Failure:** surface text remains close to the youth ball, but usage intent differs.	competitive outdoor soccer training; adult club practice; match preparation. **QARM:** 1; **Conf.:** 0.86 **SIDs: Text:** `a1,b33,c197`; **Full:** `a1,b61,c209`
Case 2—Evidence omission: short text, usage context carried by the image. Generic training-ball titles omit the distinction between indoor futsal control and outdoor field practice; CMSA preserves this visual evidence before quantization.			
 *Indoor futsal ball*	**Title:** Soccer Training Ball, Size 4 **Surface:** soccer training ball, size 4	Indoor court/gym-floor setting; compact size; futsal-style use; low-bounce or hard-court control cues. **Failure:** short title can collapse into ordinary soccer training balls.	indoor futsal low-bounce control training; small-sided court play. **QARM:** 1; **Conf.:** 0.91 **SIDs: Text:** `a1,b42,c083`; **Full:** `a1,b74,c031`
 *Outdoor training ball*	**Title:** Soccer Training Ball, Size 5 **Surface:** soccer training ball, size 5	Grass/turf field; outdoor durability cues; water-resistant usage; team practice and full-size field context. **Failure:** surface text is close to the indoor futsal item, but context indicates outdoor match training.	outdoor field soccer practice; team drills; grass/turf match training. **QARM:** 1; **Conf.:** 0.88 **SIDs: Text:** `a1,b42,c091`; **Full:** `a1,b27,c186`

B Detailed Method Construction

B.1 CMSA Caption Path

To reduce modality mismatch before quantization, we propose **Cross-Modal Semantic Augmentation (CMSA)** (see Figure 2), which enriches the item representation through two complementary visual paths before SID construction. The first path textualizes images with a VLM so visual semantics can be consumed by the text encoder and DCIM. The second path embeds the original image with a vision encoder, projects it to the SID embedding dimension, and fuses it with the text-side representation before RQ-VAE. CMSA is intentionally compatible with the MiniOneRec-style tokenizer: the RQ-VAE architecture and downstream generative backbone remain unchanged, while the pre-quantization item representation is enriched.

Motivation. In multimodal recommendation, items are associated with both textual attributes (*e.g.,* title, description) and visual content (*e.g.,* product images). A caption path provides interpretable visual semantics, but it may omit fine-grained cues such as texture, shape, or layout. A visual embedding path preserves continuous visual features, but by itself may be less interpretable to an LLM-based descriptor miner. CMSA therefore uses both paths: visual text supports intent mining, while visual embeddings preserve image-specific information for quantization.

Path 1: VLM-Based Visual-to-Text Semanticization. For each item i with visual content $\mathbf{v}_i$, we use Qwen2.5-VL-7B-Instruct, denoted $\mathcal{V}$, to generate a recommendation-oriented textual description of the image:

$$\mathbf{v}_i^{\text{text}} = \mathcal{V}(\mathbf{v}_i, \text{Prompt}_{\text{align}}), \tag{11}$$

where $\text{Prompt}_{\text{align}}$ instructs the VLM to describe recommendation-relevant visual attributes, such as aesthetic style, color scheme, usage scenario, and lifestyle signals. We concatenate this visual text with item metadata:

$$\tilde{\mathbf{t}}_i = \text{Concat}(\mathbf{t}_i,\ \mathbf{d}_i,\ \mathbf{v}_i^{\text{text}}), \tag{12}$$

where $\tilde{\mathbf{t}}_i$ serves as the input to DCIM (Section 3.2). The caption path therefore makes visual semantics available to the LLM descriptor miner.

B.2 DCIM Prompt and Descriptor Construction

The core innovation of our framework lies in **Deep Contextual Interest Mining (DCIM)**, as illustrated in Figure 2. DCIM extracts item-side intent descriptors from item metadata and visual textual evidence. We use the word "interest" in the recommender-system sense of a plausible purchase or usage context grounded in item evidence. The descriptor remains item-side and is subsequently used by the sequential model to match user histories. This distinction avoids target leakage and clarifies that DCIM is not generating a personalized profile for each test query.

Motivation. Existing SID generation methods often encode items using surface textual features (*e.g.,* product titles and descriptions), which capture explicit attributes but may omit plausible usage contexts that help explain why users interact with the item. For example, given a product "noise-canceling headphones", a surface encoder captures attributes such as brand and specifications but may miss grounded contextual descriptors such as "focus-oriented work", "frequent travel", or "high-fidelity listening". We hypothesize that instruction-tuned LLMs can help extract these item-side contextual descriptors from metadata and visual textual evidence, thereby enriching the semantic content encoded into SIDs.

Interest Extraction via LLM Prompting. For each item i with textual metadata $(\mathbf{t}_i, \mathbf{d}_i)$ and optional visual content $\mathbf{v}_i$ (processed via CMSA, see Section 3.2), we use Qwen2.5-7B-Instruct, denoted $\mathcal{M}$, to extract intent descriptors with a structured template:

$$\mathbf{z}_i = \mathcal{M}(\text{Prompt}_{\text{DCIM}}(\mathbf{t}_i, \mathbf{d}_i, \mathbf{v}_i^{\text{text}})), \tag{13}$$

where $\text{Prompt}_{\text{DCIM}}(\cdot)$ guides the LLM through: (1) surface analysis of explicit item attributes, (2) extraction of plausible usage or purchase contexts supported by the item evidence, and (3) synthesis of interpretable descriptor tags with confidence scores. Here $\mathbf{v}_i^{\text{text}}$ denotes the textual description of the visual content produced by CMSA. **Descriptor-Enhanced Text Representation.** The mined

descriptors $\mathbf{z}_i = \{z_i^{(1)}, z_i^{(2)}, \ldots, z_i^{(J)}\}$ are concatenated with the metadata and CMSA visual text to form an enriched textual input:

$$\mathbf{x}_i^{\text{text}} = \text{Concat}(\mathbf{t}_i, \mathbf{d}_i, \mathbf{v}_i^{\text{text}}, \mathbf{z}_i). \tag{14}$$

Unlike shallow item text, $\mathbf{x}_i^{\text{text}}$ encodes explicit attributes, visual textual evidence, and item-side intent descriptors. CMSA then fuses this text-side representation with the projected image embedding before RQ-VAE quantization (Section 3.2). This encourages items with similar usage motivations and visual evidence to be mapped to nearby regions in the SID space, enabling the generative model to learn intent-aware item relationships.

B.3 CMSA Image Projection and Fusion

Path 2: Visual Embedding Preservation. In parallel, we encode the original image with frozen OpenAI CLIP ViT-L/14 (`openai/clip-vit-large-patch14`), denoted $f_{\text{vis}}(\cdot)$, and project it to the same dimensionality as the text-side representation:

$$\mathbf{e}_i^{\text{vis}} = W_v f_{\text{vis}}(\mathbf{v}_i), \tag{15}$$

where W_v is a learnable projection matrix. This path preserves continuous visual information that may not be fully captured by the caption.

Dual-Path Fusion before RQ-VAE. After DCIM produces item-side descriptors $\mathbf{z}_i$, we encode the enriched textual representation and fuse it with the visual embedding:

$$\mathbf{e}_i^{\text{text}} = f_{\text{emb}}(\mathbf{x}_i^{\text{text}}), \tag{16}$$

$$\mathbf{e}_i^{\text{mm}} = W_f[\mathbf{e}_i^{\text{text}}; \mathbf{e}_i^{\text{vis}}], \tag{17}$$

where $f_{\text{emb}}(\cdot)$ denotes Qwen3-Embedding-4B, $[\cdot;\cdot]$ denotes concatenation, and W_f projects the fused vector to the RQ-VAE input dimension. The resulting multimodal embedding $\mathbf{e}_i^{\text{mm}}$ is then quantized by the same RQ-VAE configuration used in MiniOneRec-style SID construction. Thus, CMSA preserves both interpretable visual semantics and continuous visual features without changing the tokenizer architecture.

B.4 Inherited Residual Quantization

We adopt RQ-VAE to quantize the fused item embeddings $\mathbf{e}_i^{\text{mm}} \in \mathbb{R}^d$ into hierarchical SID tokens. The quantization iteratively assigns codebook entries: $s_i^{(h)} = \arg\min_k \|\mathbf{R}_i^{(h)} - \mathbf{c}_k^{(h)}\|_2$, where $\mathbf{R}_i^{(1)} = \mathbf{e}_i^{\text{mm}}$ and residuals are computed as $\mathbf{R}_i^{(h+1)} = \mathbf{R}_i^{(h)} - \mathbf{c}_{s_i^{(h)}}^{(h)}$. The quantized SID $\mathbf{s}_i = (s_i^{(1)}, \ldots, s_i^{(H)})$ captures item semantics at multiple granularity levels, where each token (*e.g.,* `a8`, `b91`) corresponds to a discrete codebook index. We retain MiniOneRec's RQ-VAE encoder–codebook–decoder, objective, codebook and training configuration, and collision and invalid-SID handling; only the pre-quantization input representation changes. We use $H = 3$. When multiple items share all three RQ codes, a unique disambiguation token is jointly generated to preserve item identity. This token is excluded from LCP rewards and prefix-semantic diagnostics, both of which operate only on the three hierarchical RQ codes.

C Extended Experimental Protocol and RQ1 Robustness

C.1 Additional RQ1 Analyses

Deep interest mining provides substantial gains over surface-level features. Comparing DeepInterestGR with the controlled MiniOneRec implementation (Kong et al., 2025) (the same Qwen2.5-7B backbone and evaluation protocol, but without CMSA/DCIM/QARM), we observe 10.8%–13.6% improvements in HR@5 across datasets. Absolute values need not match those reported by MiniOneRec under other datasets, filtering, or candidate protocols; all comparisons here are rerun on our shared splits. The gap measures the complete pipeline rather than DCIM alone; the controlled analyses below isolate pre-quantization representation and reward effects.

Table 6: Additional category-level robustness check on two Amazon Product Review categories. We compare with MiniOneRec and S-DPO, the strongest generative and RL/preference-optimized baselines in Table 1. This compact check is not intended as a full-catalog evaluation.

Dataset	Method	HR@5	HR@10	N@5	N@10
Industrial	MiniOneRec	0.0887	0.1052	0.0752	0.0805
	S-DPO	0.0832	0.1043	0.0640	0.0707
	DeepInterestGR	**0.1141**	**0.1350**	**0.0932**	**0.0999**
Toys	MiniOneRec	0.0660	0.0849	0.0496	0.0557
	S-DPO	0.0643	0.0884	0.0456	0.0522
	DeepInterestGR	**0.0874**	**0.1111**	**0.0662**	**0.0739**

SID-based generative models perform strongly in the evaluated setting. Traditional sequential models (GRU4Rec, Caser, HGN) and Transformer-based models (SASRec, BERT4Rec, S^3-Rec) underperform the strongest SID-based generative baselines in Table 1. For instance, SASRec achieves HR@5 of 0.0398 on Beauty, while DeepInterestGR achieves 0.0678. This suggests that, in our evaluated setting, autoregressive modeling over semantic IDs can make effective use of the constructed item vocabulary, particularly when the SIDs are enriched with item-side intent descriptors from DCIM.

C.2 Additional Category Results

Additional category-level check. To reduce dependence on the three primary categories without expanding our claim to the full Amazon catalog, we additionally evaluate DeepInterestGR on **Industrial and Scientific** and **Toys and Games**. To keep this check compact, we compare against MiniOneRec and S-DPO, the strongest generative and RL/preference-optimized baselines in Table 1. As shown in Table 6, DeepInterestGR remains consistently better than the strongest per-metric baseline on both categories. These results suggest that the benefit of pre-quantization item-evidence enrichment is not limited to the three primary categories, while broader full-catalog validation remains future work.

C.3 Detailed CMSA Branch Interpretation

CMSA: Fine-grained Visual Branch Ablation. Table 4 examines whether incorporating visual information through CMSA improves over text-only item representations. The Text-Only setting is aligned with the w/o CMSA row in Table 3: both remove the VLM-generated visual text path and the projected image-embedding path while keeping DCIM and QARM unchanged. Caption-only and Image-embedding-only isolate the two CMSA paths, while the full setting tests whether the two paths provide complementary evidence before SID quantization.

D DCIM and Representation Controls

D.1 Controlled Surface-Only Comparison

Surface-only and DCIM use the same Qwen2.5-7B-Instruct model, source title/description/visual text, and output budget; only the prompt objective changes, while w/o DCIM omits this LLM-derived expansion. Surface-only restates explicit attributes, whereas DCIM extracts grounded usage contexts. All three variants use the same downstream SID construction and Exact+Valid+LCP training reward without QARM, isolating the pre-quantization representation. This is a descriptive single-seed diagnostic and is separate from the end-to-end Full-QARM main results.

D.2 Rewriting and Surface-Similar Hard-Slice Controls

DCIM beyond surface rewriting. The ablation above does not by itself distinguish DCIM from generic LLM rewriting. We therefore conduct two controlled diagnostics. On the full test sets, Surface-only and DCIM use the same Qwen2.5-7B-Instruct model, source evidence, and output

Table 7: Full-test controlled Surface-only comparison without QARM.

Dataset	Variant	HR@5	HR@10	NDCG@5	NDCG@10
Beauty	w/o DCIM	.0601	.0925	.0382	.0494
	Surface-only	.0620	.0954	.0396	.0510
	DCIM	**.0642**	**.0981**	**.0411**	**.0528**
Sports	w/o DCIM	.0406	.0628	.0257	.0334
	Surface-only	.0423	.0651	.0268	.0347
	DCIM	**.0440**	**.0674**	**.0279**	**.0360**
Instruments	w/o DCIM	.0555	.0865	.0351	.0457
	Surface-only	.0578	.0898	.0366	.0476
	DCIM	**.0601**	**.0931**	**.0381**	**.0496**

Table 8: HR@10 on the surface-similar hard slice. The slice is selected using title-only TF-IDF similarity and does not use DCIM descriptors.

Method	Beauty	Sports	Instruments
w/o DCIM	0.0832	0.0558	0.0770
LLM Summarization	0.0864	0.0582	0.0802
LLM Paraphrasing	0.0855	0.0575	0.0794
DCIM	**0.0952**	**0.0645**	**0.0889**

budget; only the prompt objective changes, while w/o DCIM omits this LLM-derived expansion. All three variants use the same downstream SID construction and Exact+Valid+LCP training reward without QARM. Surface-only restates explicit attributes, whereas DCIM extracts grounded usage context. Across every dataset and metric, DCIM exceeds Surface-only, which in turn exceeds w/o DCIM (Appendix D.1). As a complementary stress test, we select instances whose target item has high title-only TF-IDF similarity to another catalog item. This selection is independent of DCIM descriptors and tests whether the representations handle lexically similar items.

Table 8 reports HR@10 on this slice. Under the shared downstream training protocol, generic summarization and paraphrasing improve over removing DCIM, indicating that better textual evidence is useful. However, DCIM consistently performs best on all three datasets. Together with the descriptive, single-seed full-test control in Appendix D.1, this supports the bounded conclusion that usage-context formulation contributes beyond simply rewriting item text, especially when titles are lexically similar.

Diagnostic scope. These ablations support a bounded mechanism claim rather than a broad claim about general-purpose LLM rewriting. All variants share the same backbone and training protocol, and the surface-similar slice is selected only from title-level TF-IDF similarity. Direct analysis of separately trained but architecturally identical SID spaces further shows that DCIM improves independent-judge usage-context and prefix consistency over Surface-only and w/o DCIM on all three datasets, without a material collision increase (Appendix D.3). Thus the recommendation gains are accompanied by measurable changes in the constructed SID structure.

D.3 Direct SID Diagnostics

For each dataset, we sample 100 anchors and retrieve five neighbors per anchor within each variant's independently trained but identically configured three-level RQ-VAE space. Longer shared prefixes rank first, followed by RQ-code Hamming distance and deterministic item-ID tie breaking. The unique collision token is excluded. The independent Llama judge sees only original titles, descriptions, and image captions, and determines whether each pair shares a primary usage context. Top-5 context consistency is the mean binary judgment over the five retrieved pairs. Prefix-depth consistency is computed separately for pairs sharing at least one, two, and three RQ codes; Prefix AUC summarizes this three-point consistency curve. Confidence intervals are obtained by anchor-level resampling.

Table 9: Direct SID diagnostics. Brackets are 95% intervals; all collision values, including interval endpoints, are percentages measured before appending the unique disambiguation token.

Dataset	Variant	Context consistency ↑	Prefix AUC ↑	Collision ↓	Bucket size ↓
Beauty	w/o DCIM	.584 [.548,.620]	.603 [.568,.638]	4.1% [3.5%,4.8%]	2.46 [2.31,2.62]
	Surface-only	.669 [.634,.704]	.687 [.653,.721]	3.5% [2.9%,4.1%]	2.32 [2.19,2.47]
	DCIM	**.758 [.726,.790]**	**.776 [.745,.807]**	3.6% [3.0%,4.2%]	2.35 [2.21,2.50]
Sports	w/o DCIM	.556 [.519,.593]	.579 [.542,.616]	4.7% [4.0%,5.4%]	2.58 [2.41,2.76]
	Surface-only	.643 [.607,.679]	.662 [.626,.698]	4.1% [3.5%,4.8%]	2.43 [2.28,2.59]
	DCIM	**.731 [.696,.766]**	**.750 [.716,.784]**	4.2% [3.6%,4.9%]	2.45 [2.30,2.61]
Instruments	w/o DCIM	.571 [.534,.608]	.591 [.555,.627]	4.4% [3.8%,5.1%]	2.51 [2.35,2.68]
	Surface-only	.655 [.619,.691]	.675 [.640,.710]	3.8% [3.2%,4.4%]	2.37 [2.23,2.52]
	DCIM	**.744 [.710,.778]**	**.763 [.730,.796]**	3.7% [3.1%,4.3%]	2.34 [2.20,2.49]

Table 10: Descriptor-quality validation. P/R denotes positive- and negative-class precision/recall.

Comparison	Accuracy	Macro-F1	Positive P/R	Negative P/R	κ
Human A vs. Human B	0.882	0.881	0.895/0.891	0.867/0.871	0.762
QARM vs. adjudicated human	0.844	0.842	0.863/0.857	0.820/0.827	0.684
Independent judge vs. human	0.820	0.818	0.847/0.829	0.788/0.809	0.636
QARM vs. independent judge	0.856	0.854	0.863/0.876	0.847/0.832	0.709
Majority baseline vs. human	0.560	0.359	0.560/1.000	0.000/0.000	0.000

DCIM improves both semantic measures on every dataset without materially increasing pre-suffix collision. The collision rate is the fraction of catalog items assigned to a three-code tuple shared by at least two items, and bucket size is averaged over those collided items. Because codebooks are trained independently, we compare within-space neighborhoods and never interpret raw code indices across variants.

E QARM Controls and Audits

E.1 Quality-Label Validation

Two annotators label 500 random descriptor–item pairs and adjudicate disagreements under the same specificity, actionability, and grounding rubric used by QARM. For an independent-family check, `meta-llama/Llama-3.1-8B-Instruct` predicts labels at temperature zero while blinded to QARM labels, SIDs, variants, and recommendation outcomes.

The corresponding $(\mathrm{TP}, \mathrm{FP}, \mathrm{FN}, \mathrm{TN})$ counts are $(246, 29, 30, 195)$, $(240, 38, 40, 182)$, $(232, 42, 48, 178)$, $(240, 38, 34, 188)$, and $(280, 220, 0, 0)$ in table order. Agreement with humans is substantial but imperfect, so QARM is treated as a noisy auxiliary signal rather than semantic ground truth.

E.2 Multi-Seed QARM Control

For shuffled QARM, a fixed item-level permutation preserves the empirical q_i distribution while breaking its association with item identity. All variants start from the same SFT initialization within each seed.

E.3 QARM Advantage Audit

We audit 10,000 fixed-candidate GRPO groups by computing advantages twice on identical samples: strict w/o QARM uses Exact+Valid+LCP, while Full additionally uses the true q_i. Of all candidates, 95.3% are valid; 53.2% of groups contain no exact target and 46.8% activate a nonzero quality term. Among all groups, 32.5% are raw-active groups with two base-reward levels and 14.3% are raw-active groups with at least three levels. Effective changes occur in 14.0% of groups at tolerance 10^{-5} and 13.7% at 10^{-4}, or 29.9% of raw-active groups at 10^{-5}. Across all groups, mean/max $|\Delta A|$ is 0.0053/0.184; within effective groups it is 0.0370/0.184, and the exact candidate's mean

Table 11: Three-seed Beauty results for the matched QARM controls. The last column reports mean $\pm$ sample standard deviation.

Setting	Metric	Seed 1	Seed 2	Seed 3	Mean $\pm$ std
Strict w/o QARM	HR@5	.0642	.0638	.0647	$.0642 \pm .0005$
	HR@10	.0981	.0976	.0987	$.0981 \pm .0006$
	NDCG@5	.0411	.0407	.0415	$.0411 \pm .0004$
	NDCG@10	.0528	.0524	.0532	$.0528 \pm .0004$
Strict shuffled q_i	HR@5	.0636	.0640	.0632	$.0636 \pm .0004$
	HR@10	.0974	.0980	.0969	$.0974 \pm .0006$
	NDCG@5	.0406	.0409	.0402	$.0406 \pm .0004$
	NDCG@10	.0522	.0526	.0518	$.0522 \pm .0004$
True QARM	HR@5	.0678	.0671	.0683	$\mathbf{.0677 \pm .0006}$
	HR@10	.1032	.1024	.1038	$\mathbf{.1031 \pm .0007}$
	NDCG@5	.0436	.0430	.0439	$\mathbf{.0435 \pm .0005}$
	NDCG@10	.0558	.0551	.0562	$\mathbf{.0557 \pm .0006}$

Table 12: Fixed-candidate advantage audit. Rates use all 10,000 groups; magnitudes use structurally active groups.

Variant	Mean $\lvert\Delta A\rvert$	Max $\lvert\Delta A\rvert$	$> 10^{-4}$	$> 10^{-2}$
Uniform	0.0508	0.247	14.1%	12.9%
True q_i	0.0370	0.184	13.7%	11.3%
Shuffled q_i	0.0368	0.181	13.7%	11.2%

Table 13: Observed resource ranges across Beauty, Sports, and Instruments. SFT and GRPO use LoRA; "not logged" denotes unavailable peak-memory measurements.

Stage	Workload/model	A100 GPU-hours	Peak memory / latency
CMSA caption	10K–18K images; Qwen2.5-VL-7B	2–20	not logged
CMSA image encoding	CLIP ViT-L/14	0.2–3	5–25GB
DCIM	10K–18K items; Qwen2.5-7B	1–12	not logged
Text embedding	Qwen3-Embedding-4B	0.2–3	10–30GB
RQ-VAE	$H = 3$	1–10	5–30GB
QARM labeling	20K–72K labels; Qwen2.5-7B	1–12	not logged
SFT	Qwen2.5-7B LoRA, 3 epochs	4–30/run	15–45GB/GPU
GRPO	Qwen2.5-7B LoRA, 2 epochs, $G = 8$	15–120/run	25–70GB/GPU
Online inference	Qwen2.5-7B, beam 20	–	batch-1 P50 0.1–2.0s/query; 16–40GB

ΔA is $+0.094$. Target scores have mean/std 0.62/0.21 and P10/P50/P90 0.31/0.64/0.88. Among raw-active groups, the number of distinct wrong-candidate LCP levels is one, two, three, and at least four in 69.4%, 21.6%, 7.1%, and 1.9%, respectively. Thus changes concentrate exactly where multiple base-reward levels prevent the quality term from being removed by group normalization.

The true and shuffled variants have identical marginal q_i distributions, so similar aggregate magnitudes are expected; their difference is which target receives each reliability value. In two-level groups their maximum $|\Delta A|$ is below 7×10^{-7}, whereas multi-level groups reach 0.176–0.247.

F Observed Resource Profile

Table 13 reports observed min–max ranges across the three primary datasets on A100 80GB GPUs. Each item uses at most one main image. GPU-hours equal wall-clock hours times the number of GPUs. These ranges characterize the Full pipeline and exclude ablation sweeps and failed tuning runs.

Online inference uses the trained Qwen2.5-7B SID generator, but it does not invoke the offline CMSA VLM, DCIM LLM, QARM judge, CLIP/Qwen3 encoders, or RQ-VAE.